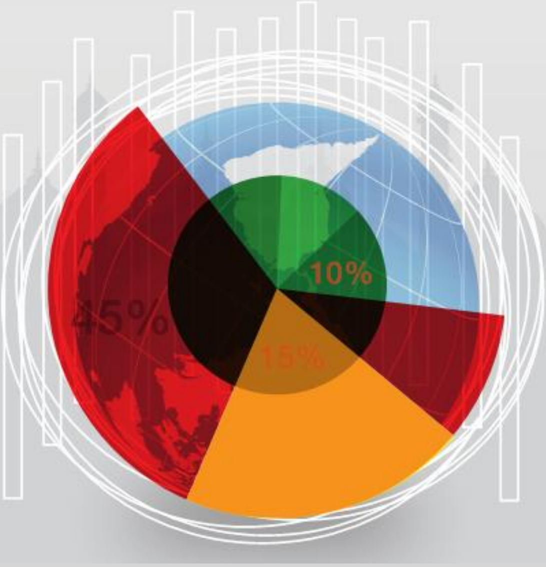
Attitudes towards old age and age of retirement across the world: findings from the future of retirement survey

Khan, Hafiz T.A. ORCID: https://orcid.org/0000-0002-1817-3730 and Leeson, George W. (2013) Attitudes towards old age and age of retirement across the world: findings from the future of retirement survey. In: 12th Islamic Countries Conference on Statistical Sciences (ICCS-12), 19-22 Dec 2012, Doha, Qatar.

ISBN 978-969-8858-11-7



PROCEEDINGS ICCS-12

Vol. 23

مؤتمر الدول الإسلامية الثاني عشر للعلوم الإحصائية

12th Islamic Countries Conference on Statistical Sciences

Statistics for Everyone and Everywhere

December 19-22, 2012
at
Qatar University, Doha, Qatar

Jointly organized by

Islamic Countries Society
of Statistical Sciences

Department of Mathematics,
Statistics and Physics
Qatar University

44-A, Civic Centre, Sabzazar, Multan Road, Lahore, Pakistan
Email: secretary@isoss.net URL: www.isoss.net

Published by: ISOSS, Lahore, Pakistan.

"All papers published

in the Proceedings of

ICCS-12

were accepted after formal peer review

by the experts in the relevant field"

Dr. Munir Ahmad
Editor

CONTENTS

FOREWORD

The 12th Islamic Counties Conference on Statistical Sciences (ICCS-12) was held at Qatar University, Doha, Qatar on December 19-22, 2012. A total of 22 technical sessions were held during the four day conference. About 120 research papers out of 239 papers received, were presented, discussed and appreciated by the conference participants. There were seven sessions on the Theory and Applications of Statistics, two sessions on survey sampling, three sessions on demography and population studies, three sessions on economic and business statistics, three on medical and bio-statistics and one session each on mathematical studies, statistical education, environmental statistics, and statistical communications.

There were eight key-note speakers and these are:

1. **Dr. Munir Ahmad** (Pakistan) -
   ISOSS: History, Challenges and Future Developments
2. **Dr. Shahjahan Khan** (Australia) -
   Linear Model Inference with Non-sample Prior Information
3. **Dr. Muhammad Hanif** (Pakistan) -
   Has Sample Survey Sampling Undergone A Scientific Revolution?
4. **Dr. Edward Wegman** (USA) -
   Big Data: Technology and Analysis
5. **Dr. Abdel H. El-Shaarawi** (Egypt/Canada) -
   Environmental Control and Economic Development
6. **Dr. Mohammad Fraiwan Al-Saleh** (Jordan) -
   Statistical Ideas that are Rarely Mentioned in Classrooms
7. **Dr. Aman Ullah** (USA) -
   Robustify Financial Time Series Forecasting with Bagging
8. **Dr. Ehsan S. Soofi** (USA) -
   When Association Indices Fail and Information Indices Succeed

Beside the above, there were four invited speakers and these are:

1. **Dr. Mohammad Ahsanullah** (USA) -
   Generalized Extreme Value Distribution
2. **Dr. Ismail Bin Mohd.** (Malaysia) -
   Biparaboloid-Ellipsoid Programming with Different Axis
3. **Dr. Hafiz T.A. Khan** (United Kingdom) -
   Attitudes Towards Old Age and Age of Retirement Across the World: Findings from the Future of Retirement Survey
4. **Dr. Mohammad Yahyah** (United Kingdom) -
   Application of Statistics in the Pharmaceutical Industry

The papers were reviewed by the Scientific and Editorial Committee with the help of many experts in the field. The papers are now available on the conference website in the shape of pre-proceedings.

ISOSS acknowledges the work put in by the members of the Committee and especially Mr. M. Imtiaz and Mr. M. Iftikhar who had composed the research papers in the Pakistan Journal of Statistics style and uploading on the website.

The Committee is grateful to the President Qatar University and her administration for the facilities provided to the conference participants and to the faculty, staff and students of the Department of Mathematics, Statistics and Physics for efforts made by them in organizing the sessions and looking after the authors of papers, chairmen and rapporteurs of the sessions.

ISOSS management appreciates the hard work put in by Dr. Ayman Bakleezi, Professor of Statistics and Coordinator of the Statistics Program, Department of Mathematics, Statistics and Physics at Qatar University and Chairman, Local Organizing Committee in making the conference a great success. ISOSS also admires the efforts made by Dr. Ali S. Hadi, President of ISOSS and Dr. Shahjahan Khan, Former President ISOSS (2005-2011) for the success of the conference.

PROFESSOR DR. MUNIR AHMAD
Founding President and Patron ISOSS
&
Chairman, Scientific Program Committee
ICCS-12, Doha – Qatar

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 1-16

EXTENDED GENERALIZED GAMMA DISTRIBUTION AND SAME ITS APPLICATIONS

Bachioua Lahcene
Department of Mathematics, Deanship of Preparatory Year, Hail University, KSA.
Email: drbachioua@gmail.com

ABSTRACT

In this paper, we are going to discuss an extended form, including the reversed extended generalized Gamma distribution as its subfamily, and refer to it as the extended generalized same distribution. Because of many difficulties described in the literature in modeling the parameters, we propose here a new extended model. The model associated to this heuristic is implemented in Splus. We validate the result of the generalized gamma distribution routine in the specific cases.

In this paper, we are examining a ‘six-parameter extended generalizations’ of the gamma distribution, and derive parameter for that distribution. These techniques, in the general case, depend upon the method of moment’s considerations, which lead to simultaneous equations for which closed form solutions are not available.

However, these models involve stronger distributional assumptions than is desirable, and inferences may not be robust for departures from these assumptions. In this paper, a mixture model is proposed using the generalized distribution family. The generalised F mixture model can relax the usual stronger distributional assumptions and allow the analyst to uncover structure in the data that might otherwise be missed. Computational problems with the model and model selection methods have also been discussed. Comparison of maximum likelihood estimates, with those obtained from mixture models under other distributions is also included.

KEY WORD

Euler gamma function, extended expression gamma functions, density function, special distributions, extended generalized gamma distribution, mixture distributions, diffraction theory.

1. INTRODUCTION

According to the links, Karl Pearson was the first to formally introduce the gamma distribution. However, the symbol gamma for the gamma function, as a part of calculus, originated far earlier, by Legrenge (1752 to 1853). In statistics, there are many family distributions, which can be used in various areas. Such distributions include the Normal, Chi-squared, Exponential, Rayleigh, Weibull, Erlang, Gamma, Extreme-Value, Lognormal and others. The relationships between various distributions are shown in [1]. It also demonstrates the relation of the above continuous family distributions.

The gamma distribution originated from Pearson's work, and was known as the Pearson type III distribution, before acquiring its modern name in the 1930s and 1940s. Pearson's 1895 paper introduced the type IV distribution, which contains student’s t distribution as a special case [23].

It’s good to have a familiarity with the Gamma distribution. The sum of exponential distributions is a gamma distribution, and the valuation function of double-bounded model involves dichotomous choice elicitation questions, resulting in the interval censoring of individual subject value. The censored survey can use the survival analysis to provide wide parametric distributions [11].

The history of this family of distributions was reviewed and its properties were discussed by Stacy, in 1962. In this paper we shall employ a simple model and statistical-mechanical method(s) to derive the three-parameter generalized gamma distribution. Subsequent work on statistical problems, associated with the distribution, has been done by Bain and Weeks. Special cases of the generalized gamma distribution include the Weibull, gamma, Rayleigh, exponential and Maxwell velocity distributions [33].

Different properties like monotonicity of the hazard functions and tail behaviors of the gamma distribution and the generalized exponential distribution are quite similar in nature, but the later has a nice compact distribution function. It is observed that for a given gamma distribution a generalized exponential distribution exists, and the two distribution functions are almost identical. Since the gamma distribution function does not have a compact form, generating gamma random numbers efficiently is known to be problematic [32].

The Johnson System of distributions is composed of three distribution families. These three families cover the entire allowable skewness and kurtosis plane. The basis for the Johnson System is that a distribution, being approximated, may be transformed in such a way that it can be considered for an even distribution. The disadvantage of the Johnson System is that the methods for fitting the distribution depend on one of the three families that are appropriate for the purpose [23].

The adequacy of the Gamma distribution (GD) for monthly precipitation totals has been reconsidered. The motivation for this study is the observation that the GD fails to represent precipitation in considerable areas of global observed and simulated data. This misrepresentation may lead to erroneous estimates of the Standardised Precipitation Index. In this study, the GD is compared to the Weibull (WD), Burr Type III (BD), exponentiated Weibull (EWD) and generalized Gamma (GGD) distribution. These distributions extend the GD in terms of possible shapes (skewness and kurtosis) and the behavior for large arguments [27].

In addition, we will study some moment properties, and derive exact and explicit formulas for the mean, variance, skewness and kurtosis. We are also going to form a theorem to characterize this distribution and discuss its limiting distributions as the shape parameters tend to zero or infinity. Finally, possible applications of this distribution in bioassays as a dose response curve will be discussed and illustrated with two examples. It is fairly commonplace in reliability analyses to encounter data which is incompatible with the extended generalized Gamma distribution as its subfamily, exponential, Weibull, and other familiar probability models. Such data motivates research to enlarge the group of distributions which are useful to the reliability analyst [16].

We are going to consider a form of the extended generalized Gamma distribution same as generalized logistic distribution, named symmetric extended generalized logistic distribution or extended type III generalized logistic distribution. The distribution is derived by compounding a two-parameter generalized Gumbel distribution with a two-parameter generalized gamma distribution.

2. EXTENDED GENERALIZED GAMMA DISTRIBUTION

The Extended Generalized Gamma Distribution function is characterized by six parameters, $r \in IR$, $p\,, k, p, m, n, \lambda \succ 0$ and is defined in its original form firstly introduced by Bachioua, in 2004 [12]. A random variable X that has a probability density function is said to have an extended generalized gamma distribution of the following form:

$$f_X(x) = \begin{cases} \dfrac{x^{k-1}\left[x^m + n\right]^{-r} e^{-\lambda\, x^{\,p}}}{\Lambda_r(k, p, m, n, \lambda)}, & \mathit{for} \quad 0 \le x < \infty, \quad \mathrm{r} \in \mathrm{IR}\,, \ \ k, p, m, n, \lambda \succ 0 \\ 0, & \mathit{otherwise}, \end{cases};$$

where

$$\Lambda_r(k, p, m, n, \lambda) = \int_0^\infty x^{k-1}\left[x^m + n\right]^{-r} e^{-\lambda\, x^{\,p}}\, dx; \quad p\,, k, p, m, n, \lambda \succ 0; \quad r \in IR$$

It implies from the definition of X that $f_X(x) \ge 0$, $0 \le x < \infty$, and $r \in IR$, $p\,, k, p, m, n, \lambda \succ 0$. Also we have $\int_0^\infty f_X(x) dx = 1$.

In terms of the model function statistical properties; reliability and hazard functions, and estimation of some parameters of the distribution are studied. Also, the form of the distribution is considered under various forms of the model of extended generalized gamma distribution [10].

Property (1): For $r \in IR$, $p\,, k, p, m, n, \lambda \succ 0$. This 6-parameter extended generalized gamma distribution can be regarded as an extension of the Kobayashi's generalized gamma distribution, of the following form:

$$f_X(x) = \begin{cases} \dfrac{x^{k-1} \;\; x + n \;\;^{-r}\, e^{-\;x}}{\Gamma_r(k, n)}, & \mathit{for} \quad 0 \le x < \infty, \quad \mathrm{r} \in \mathrm{IR}\,, \ \ k, n \succ 0 \\ 0, & \mathit{otherwise}, \end{cases};$$

where $\Gamma_r(k, n) = \int_0^\infty x^{k-1} \;\; x + n \;\;^{-r}\, e^{-\;x}\, dx; r \in IR, k, n \succ 0$.

Proof: Also, this case is reduced to the well known gamma function when $\Lambda_r(k, 1, 1, n, 1) = \int_0^\infty x^{k-1} \;\; x + n \;\;^{-r}\, e^{-\;x}\, dx = \Gamma_r(k, n)$; for $k, n, r > 0$, then the Kobayashi's generalized gamma distribution is special cases of extended generalized gamma distribution.

Property (2): For $r=0$, $p,k,m,n,\lambda \succ 0$. This 6-parameter extended generalized gamma distribution can be regarded as an extension of the generalized gamma distribution, of the following form:

$$f_X(x)=\begin{cases}\dfrac{x^{k-1}e^{-\lambda x^{p}}}{p^{-1}\ \lambda^{\frac{p-k}{p}-1}\Gamma(k/p)}, & for \quad 0\le x<\infty, \qquad k,p,,\lambda\succ 0\\ 0, & otherwise,\end{cases};$$

where $\Gamma(k)=\int_0^\infty x^{k-1}e^{-\ x}\,dx,\ k\succ 0$.

Proof: Also, this case is reduced to the well known gamma function when $\Lambda_0(k,p,m,n,\lambda)=p^{-1}\ \lambda^{\frac{p-k}{p}-1}\Gamma(k/p);\quad k/p=s\succ 0.$ And the graph of pdf with $p=1$ are:

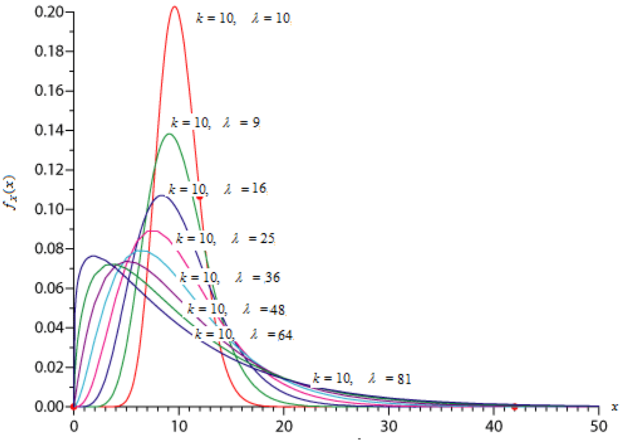


Fig. 6.6: Illustration of the Gamma pdf for parameter values over
$k=10,\quad \lambda=10,9,16,25,36,48,64,81$.

For special case $k/p=s$ then $p^{-1}\ \lambda^{\frac{p-k}{p}-1}\Gamma(k/p)=p^{-1}\ \lambda^{\ s}\,\Gamma(s);\quad s\succ 0$, then the random variable X that has a probability density function is said to have a generalized gamma distribution with three parameters of the following form:

$$f_X(x)=\begin{cases}\dfrac{x^{k-1}e^{-\lambda x^{k/s}}}{(k/s)^{-1}\ \lambda^{\ s}\,\Gamma(s)}, & for \quad 0\le x<\infty, \qquad k,s,\lambda\succ 0\\ 0, & otherwise,\end{cases};$$

where $\Gamma(s)=\int_0^\infty x^{s-1}e^{-\ x}\,dx,\ s\succ 0$.

The probability density function of the gamma distribution can be expressed in terms of the gamma function parameterized in terms of a shape parameter *k* and scale parameter *θ*. Both *k* and *θ* will be positive values. Alternatively, the gamma distribution can be parameterized in terms of a shape parameter $\alpha = k$ and an inverse scale parameter $\beta = 1/\lambda$, called a rate parameter:

$$f_X(x) = \begin{cases} \beta^{\alpha} \dfrac{x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}, & for \quad 0 \le x < \infty, \qquad \alpha, \beta \succ 0 \\ 0, & otherwise, \end{cases};$$

For special case $k/p = s = 1,$ then $\Gamma(s) = 1$, and density function is said to have a Weibull distribution with two parameters of the following form:

$$f_X(x) = \begin{cases} \dfrac{k}{\lambda} x^{k-1} e^{-\lambda x^{p}}, & for \quad 0 \le x < \infty, \qquad k, s,, \lambda \succ 0 \\ 0, & otherwise, \end{cases};$$

3.EXTENDED GENERALIZED OF GAMMA MODEL DISTRIBUTION

A continuous random variable X is said to have an extended generalized gamma distribution with 6-parameters and a model function $\alpha(x)$, defined $\theta = (k, m, n, r, \lambda, p)$, denoted by $X \sim EGG(\theta, \alpha(x))$ iff its *pdf* is given by;

$$f_X(x) \equiv f_X(\theta, \alpha(x)) = \begin{cases} \dfrac{\alpha'(x)}{\Lambda(\theta)} \alpha(x)^{k-1} \left[\alpha(x)^m + n \right]^{-r} e^{-\lambda \alpha(x)^{p}}, & for \quad x \in (a, b) \\ 0, & otherwise, \end{cases};$$

where k, p, m are shape parameters, λ is a scale parameter, n and r are, respectively, displacement and intensity parameters. And the function;

$$\Lambda(\theta) = \Lambda_r(k, p, m, n, \lambda) = \int_0^{\infty} x^{k-1} \left[x^m + n \right]^{-r} e^{-\lambda x^{p}} dx; \quad \theta = (k, m, n, r, \lambda, p);$$

From this definition it follows that the *pdf* of the random variable $Y = \alpha(X)$ will be;

$$f_Y(y) \equiv f_Y(\theta, y) = \begin{cases} \dfrac{1}{\Lambda(\theta)} y^{k-1} \left[y^m + n \right]^{-r} e^{-\lambda y^{p}}, & for \quad y \in (0, +\infty), \\ 0, & otherwise, \end{cases};$$

and the following relation is true for all, $x_1, x_2 \in (a, b); x_1 \prec x_2$ then;

$$\int_{x_1}^{x_2} f_X \;\; x \;\; dx = \int_{\alpha \;\; x_1}^{\alpha \;\; x_2} f_Y \;\; y \;\; dy.$$

The random variable $Y = \alpha(X) \sim EGG(\theta)$ will play an essential role in derivation of many statistical properties of X.

Theorem (1): The function $f_X(x)$ is a *pdf* .

Proof: It implies from the definition of X . That $f_X(x) \geq 0; \forall x \in (a,b)$. Also we have;

$$\int_a^b f_X \;\; x \;\; dx = \int_0^\infty f_Y \;\; y \;\; dy = 1$$

The distribution function of X will be defined by;

$$F_X \;\; x; \theta, \alpha \;\; x \quad = \int_0^x f_X \;\; t \;\; dt; \text{ for } x \in \;\; a,b \;\; .$$

Hence, we get for a given $x; x \in \;\; a,b$, then;

$$F_X \;\; x \;\; = \int_0^{y=\alpha \;\; x} f_Y \;\; t \; dt = F_y \;\; y \;\; = \frac{\Lambda_y \;\; \theta}{\Lambda \;\; \theta} \text{, for } y \in (0,+\infty) .$$

Theorem (2): If $Y = \alpha(X) \sim EGG(\theta)$, then the s-the moment about the origin is;

$$\mu_s = EY^s = \frac{\Lambda \;\; k+s,m,n,r,\lambda,p}{\Lambda \;\; \theta} \text{; for } s = 1,2,3,...$$

Proof: Since,

$$EY^s = \int_0^\infty \frac{1}{\Lambda \;\; \theta} y^s \cdot y^{k-1} \left[y^m + n \right]^{-r} e^{-\lambda y^p} dy$$

The proof is completed.

Theorem (3): If $Y = \alpha(X) \sim EGG(\theta)$, then its respective mean and variance are;

1. mean $= \dfrac{\Lambda \;\; k+1,m,n,r,\lambda,p}{\Lambda \;\; \theta}$

2. variance $= \dfrac{\Lambda \;\; \theta \;\; \Lambda \;\; k+2,m,n,r,\lambda,p \;\; - \Lambda^2 \;\; k+1,m,n,r,\lambda,p}{\Lambda^2 \;\; \theta}$

Proof: From theorem 3.2 for $s = 1,2$, the proof is followed.

Theorem (4): If $X \sim EGG(\theta, \alpha(x))$ then its respective reliability and hazard (or failure rate) functions are

1. $R \;\; x \;\; = R \;\; x; \theta, \alpha \;\; x \quad = \dfrac{\Lambda \;\; \theta \;\; - \Lambda_{\alpha \;\; x} \;\; \theta}{\Lambda \;\; \theta}$; for $x \in \;\; a,b$

2. $h \;\; x \;\; = h \;\; x; \theta, \alpha \;\; x \quad = \dfrac{1}{R \;\; x} \alpha' \;\; x \;\; \alpha^{k-1} \;\; x \;\; \left[\alpha^m \;\; x \;\; + n \right]^{-r} e^{-\lambda \alpha^p \;\; x}$; for $x \in \;\; a,b$.

Proof: By substitution f_X and F_X in the definitions;

$$R \;\; x \;\; = 1 - F_X \;\; x \;\; ; \; h \;\; x \;\; = \frac{f_X \;\; x}{R \;\; x}$$

4. SAME CASE OF EXTENDED GENERALIZED GAMMA MODEL DISTRIBUTION

Many distributions can be derived as special cases of the extended generalized gamma distribution. This can be done by reducing the number of parameters of X to less than 6-parameters, by assigning proper values for some parameters of θ. For example, the following are some important new types of *pdf* of 5-parameters extended generalized gamma function [14].

1. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ k,m,0,r,\lambda,p}\alpha^{k-r\ m-1}\ x\ e^{-\lambda\alpha^{p}\ x}$

2. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ 1,m,n,r,\lambda,p}\left[\alpha^{m}\ x\ +n\right]^{-r} e^{-\lambda\alpha^{p}\ x}$

3. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ k,m,n,r,0,p}\alpha^{k-1}\ x\left[\alpha^{m}\ x\ +n\right]^{-r}$

4. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ k,0,n,r,\lambda,p}\alpha^{k-1}\ x\ \ 1+n\ ^{-r} e^{-\lambda\alpha^{p}\ x}$

5. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ k,m,n,0,\lambda,p}\alpha^{k-1}\ x\ e^{-\lambda\alpha^{p}\ x}$

6. $f_X\ x = \dfrac{\alpha'\ x}{\Lambda\ k,m,n,r,\lambda,0}\alpha^{k-1}\ x\left[\alpha^{m}\ x\ +n\right]^{-r} e^{-\lambda}$

Notice that, all of these distributions and others are still defined in general, since the model function is yet unspecified. Many $\alpha\ x$ can be proposed, for example if we take;

$$\alpha\ x = \left(\frac{x-\eta}{\delta}\right)^{\beta};\ \text{for } x>\mu,\ \delta>0,\ \beta\geq 1,$$

Where μ,δ,β are, respectively, location, scale, and shape parameters, then the *pdf* of $X\sim EGG(\theta,\alpha(x))$ becomes with 9-parameters and is given by;

$$f_X\ x = \frac{\beta}{\delta\Lambda\ \theta}\left(\frac{x-\eta}{\delta}\right)^{\beta k-1}\left[\left(\frac{x-\eta}{\delta}\right)^{\beta m}+n\right]^{-r} e^{\lambda\left(\frac{x-\eta}{\delta}\right)^{\beta p}};\ \text{for } x>\eta.$$

The following distributions are some particular cases of this distribution.

Case (1): when $n=0$, then;

$$f_X\ x = \frac{\beta p\lambda^{\frac{k-rm}{p}}}{\delta\ \Gamma\left(\frac{k-rm}{p}\right)}\left(\frac{x-\eta}{\delta}\right)^{\beta\ k-rm\ -1} e^{-\lambda\left(\frac{k-\eta}{\delta}\right)^{\beta p}};\ \text{for } r<\frac{k}{m},$$

and for $k-rm=p$,

$$f_X(x) = \frac{\beta p\lambda}{\delta}\left(\frac{x-\eta}{\delta}\right)^{\beta p-1} e^{-\lambda\left(\frac{x-\eta}{\delta}\right)^{\beta p}}, \quad x>\eta,$$

This form can be considered as an extended generalized Weibull distribution, and when $\beta p = 1$, then

$$f_X(x) = \frac{\lambda}{\delta} e^{-\lambda\left(\frac{x-\eta}{\delta}\right)}, \quad x>\eta,$$

which is the generalized 3-parameters exponential distribution.

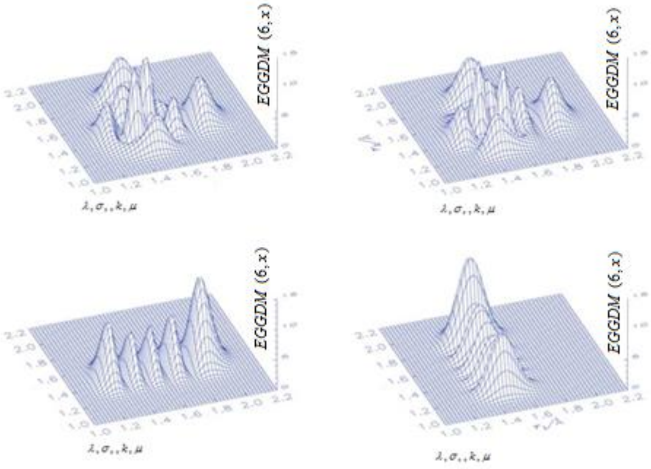


Fig. 4.1: Illustration of the the generalized 3-parameters exponential distribution

For $\beta(k-rm)=1$, then;

$$f_X(x) = \frac{\beta p\lambda^{\frac{1}{\beta p}}}{\delta\,\Gamma\left(\frac{1}{\beta p}\right)} e^{-\lambda\left(\frac{k-\eta}{\delta}\right)^{\beta p}}, \quad x>\eta,$$

This form is an extended generalized normal distribution, and when $\beta p = 2$, $\lambda = \frac{1}{2}$, then $f_X(x)$ is a half-normal distribution.

Case (2): when $\lambda = 0$, from theorem, we have

$$f_X(x) = \frac{m\beta n^{r-\frac{k}{m}}}{\delta\beta\left(r-\frac{k}{m},\frac{k}{m}\right)}\left(\frac{x-\eta}{\delta}\right)^{\beta k-1}\left[\left(\frac{x-\eta}{\delta}\right)^{\beta m}+n\right]^{-r} \text{; for } x>\eta,$$

which can be considered as an extended generalized beta distribution.

Case (3): when $r = 0$, from theorem, we have

$$f_X(x) = \frac{\beta p \lambda^{\frac{k}{p}}}{\delta\,\Gamma\left(\frac{k}{p}\right)} \left(\frac{x-\eta}{\delta}\right)^{\beta k-1} e^{-\lambda\left(\frac{x-\eta}{\delta}\right)^{\beta p}} \quad ; \text{ for } x > \eta$$

This can be considered as an extended generalized gamma distribution, and when $\beta k = 1$, then $f_X(x)$ can be regarded as extended generalized exponential distribution.

Case (4): when $m = p$, we have ;

$$f_X(x) = \frac{m\beta\lambda^{\frac{k}{m}-r}}{\delta\,\Gamma\left(\frac{k}{m}-r\right)} \left(\frac{x-\eta}{\delta}\right)^{\beta k-1} \left[\left(\frac{x-\eta}{\delta}\right)^{\beta m} + n\right]^{-r} e^{-\lambda\left(\frac{x-\eta}{\delta}\right)^{\beta m}} \quad ; \text{ for } x>\eta \text{ and } \frac{k}{m} > r,$$

and for small $n\lambda$, which is another form of the generalization of gamma distributions.

5. APPLICATIONS OF EXTENDED GENERALIZED GAMMA IN MIXTURE MODEL

The parameters we wish to estimate are: K the number of clusters, the relative weight/size of each cluster and the probability distributions for each variable/column for each cluster. From these we can partially assign the observations to the clusters. Formally, we can describe a K component mixture model as follows [5].In order to estimate the input's probability density, we will use a linear combination of *M* basis functions:

$$P(x) = \sum_{j=1}^{k} p(x/j)\, p(j)$$

$P(x)$ is the density at point x;
$p(x/j)$ is called the j-th component density;
$p(j)$'s are the mixing coefficients, i.e. the prior probability of a data vector having been generated from component j of the mixture.

Constraints:

$$\begin{cases} 0 \le p(j) \le 1 \\ \sum_{j=1}^{M} p(j) = 1 \\ \int p(x/j)\, dx = 1 \end{cases}$$

Let detection probabilities be denoted as p_{ij}, where i denote the occasion and j denotes the group. For, p_{11} is the probability of detection for the first survey in a site in the first mixture and p_{12} is the probability of detection for the first survey for a site in the second mixture, so the second number that is subscripted identifies the group[7].

Example (1): Two-component gamma mixture regression model:

Let Y_{ij} (i = 1,2,…, m; j = 1,2,…, n_i) represents the maternity LOS for the j^{th} individual in the ith hospital, where m is the number of hospitals, n_i is the number of patients within hospital i, the total number of observations being $n = \sum_{i=1}^{m} n_i$. The probability density function of Y is assumed to be a two-component gamma mixture:

$$f(y_{ij}) = pf_1(y_{ij}) + (1-p) f_2(y_{ij})$$

where 0 < p < 1 gives the proportion of patients belonging to the first component or sub-population, and $f_k(y_{ij})$ is the kth component gamma distribution with mean μ_k and shape parameter ν_k:

$$f_k(y_{ij}) = \frac{1}{y_{ij}\Gamma(v_k)}\left(\frac{v_k y_{ij}}{\mu_{k,ij}}\right)^{v_k} e^{-\frac{v_k y_{ij}}{\mu_{k,ij}}} \text{ ; } k = 1,2;\ i = 1,2,\ldots, m;\ j = 1,2,\ldots,\ n_i$$

A plausible interpretation is thus in terms of its unobserved heterogeneity, a sub-population of usual patients with relatively short LOS, and another sub-population whose members tend to stay longer due to unexpected complications after delivery.

Data were collected from n = 909 women hospitalized for vaginal delivery with multiple complicating diagnoses in Western Australia. Their maternity LOS ranged between one and 45 days with mean 6.22 days and SD 5.19 days. The empirical LOS distribution exhibits substantial heterogeneity (skewness = 3.44) because of disparities in patient characteristics and possibly hospital care and medical practice received during hospitalization. Figure 1 plots the observed LOS together with the fitted two-component gamma mixture distribution [9].

The proportion of women with relatively long LOS is estimated to be 5.6%. For this sample of women admitted to m = 26 public hospitals for child birth, their average age was 27.68 years (S.D. = 6.07), 27.3% were non-married, 34% were emergency cases, but only 5.6% had private medical insurance coverage. Aboriginals accounted for 11.8% of the sample, and the majority of women (85%) resided in urban areas. In terms of clinical factors, the average number of diagnoses recorded was 7.67 (SD 3.16), while the average or obstetrical procedures performed during hospitalization was 2.87 (SD 1.94) [9].

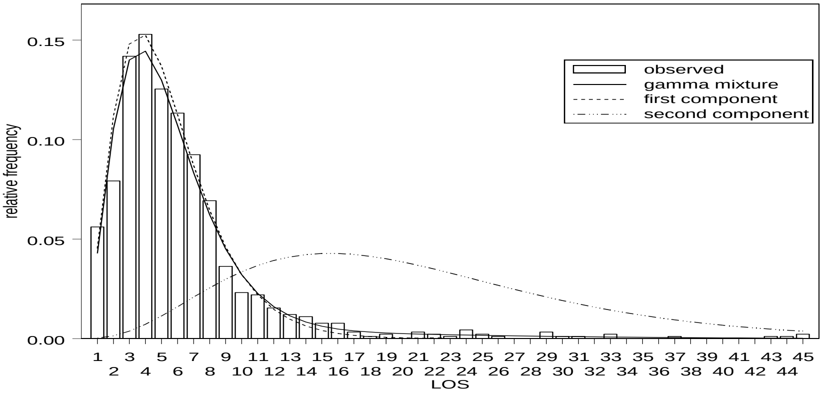


Fig. 5.1: Empirical fitted two-component Gamma mixture model

Example (2): Component Beta mixture model:

The investigator surveys 250 study sites, with each site being surveyed 4 times. The encounter histories are recorded in cells B4:B19, and the frequency of each history is recorded in cells C4:C19. The total number of sites is given in cell C20, and the number of unique histories is given in cell C21 (which you might remember indicates the number of terms in our multinomial likelihood function). To avoid over-parameterization, you can only run models with 15 or fewer parameters. The naïve estimate for occupancy (occupancy unadjusted for detection probability) is computed in cell C22 as the total number of sites which had one or more detections divided by the total number of sites [24].

In the example below, we entered beta values of -2, -1, 0, 1, and 2 for ψ, p_1, p_2, p_3, and p_4, respectively. These betas correspond to the following probabilities: ψ=0.04535, p_1 = 0.07926, p_2 = 0.5000, p_3 = 0.92074, and p_4 = 0.95465. Note that the betas can take on any value, and the link function constrains the MLE's to be between 0 and 1, which is necessary because π, p_1, p_2, p_3, and p_4 ,and ψ are probabilities, and probabilities range between 0 and 1. The figure below shows betas that range from -9 to +9, and the corresponding, sin "transformed" probability estimate. Look for the beta values of -2, -1, 0, 1, and 2, and find their corresponding probabilities on the graph [31]:

	B	C	D	E	F	G
3	History	Frequency	Parameter	Estimate	Betas	MLE
4	1111	15		1		0.50000
5	1110	17	Mixture 1			
6	1101	12	ψ	1		050000
7	1100	18	p_1	1		050000
8	1011	8	p_2	1		050000
9	1010	13	p_3	1		050000
10	1001	9	p_4	1		050000

11	1000	20	Mixture 2			
12	0111	5	ψ	0		050000
13	0110	9	p_1	1		050000
14	0101	7	p_2	1		050000
15	0100	15	p_3	1		050000
16	0011	5	p_4	1		050000
		$\ln(L(p_i / n_i, y_i)) = y_1 \ln(p_1) + y_2 \ln(p_2)$ $+y_3 \ln(p_3)\ +y_4 \ln(p_4) + ... + y_{16} \ln(p_{16})$ Sites=250, Histories =16, Naïve Estimate =0.696				
Results						
			D	E	F	G
6			ψ	1	-2.0000	0.04535
7			p_1	1	-1.0000	0.07926
8			p_2	1	0.0000	0.50000
9			p_3	1	1.0000	0.92074
10			p_4	1	2.0000	0.95465

Fig. 5.2: Empirical distribution of same component Beta mixture model

Example (3): Component Poisson mixture model:

The probability function of the k-finite Poisson mixture is given by;

$$f(x) = \sum_{j=1}^{k} p_j \frac{e^{-\lambda_j} \lambda_j{}^x}{x!}$$

We assume that $0 \le \lambda_1 < \lambda_2 < .... < \lambda_k$ in order to ensure the identifiability of the above finite mixture. Note that we allow the first component mean to be 0, thus allowing the corresponding distribution to be degenerate. The joint distribution of the observed data x and the unobserved indicator parameters z conditional on the model parameters can be written as;

$$f(x,z \mid \theta) = f(x \mid \theta, z) f(z \mid p) = \prod_{i=1}^{n} \prod_{j=1}^{k} f(x_i \mid \lambda_j)^{z_{ij}}$$

where $f(x_i \mid \lambda_j)$ is in our case the Poisson distribution with parameter λ_j. Bayesian formulation requires prior distributions for the model parameters θ. Immediate choices are the conjugate prior distributions $Gamma(a, \beta)$ for each λ_j and the $Dirichlet(d_1, ..., d_k)$ for p [30]. Denoting the priors $\pi(\theta)$, $\pi(\lambda)$ and $\pi(p)$, Bayes theorem leads to the joint posterior distribution of θ can be written as;

$$p(\theta \mid x, z) \propto f(x, z \mid \theta) \pi(\theta) = f(x \mid \lambda, p) f(z \mid p) \pi(p) \pi(\lambda)$$

The Poisson mixtures case we need to sample from the full conditional posterior distribution for the parametersλ, p and z**.** These sampling steps are readily found to be;

$$z_{ij} \sim Multinomial(1, w_{i1}, ..., w_{ik}) \text{ i=1, . . ., n, j=1, . . .,k;}$$

where $w_{ij} = \dfrac{p_j f(x_i \mid \lambda_j)}{f(x_i)}$, j=1, . . ., k

$$p \sim Dirichlet\left(d_1 + \sum_{i=1}^{n} w_{i1}, ..., d_k + \sum_{i=1}^{n} w_{ik} \right)$$

$$\lambda_j \sim Gamma(a + \sum_{i=1}^{n} z_{ij} x_i, b + \sum_{i=1}^{n} z_{ij}) I(\lambda_{j-1}, \lambda_{j+1}) \text{, j=1, . . . , k}$$

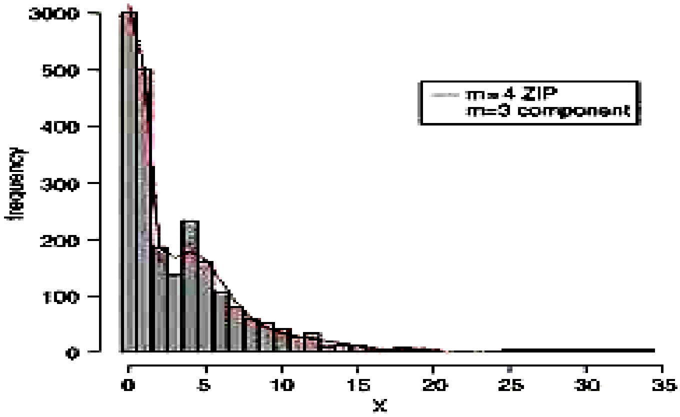


Fig. 5.3: Empirical distribution of same component Poisson mixture model

Example (4): Component Gaussian mixture model:

A Gaussian mixture model is a weighted combination of Gaussian probability density functions which are referred in this context as Gaussian components of the mixture model, describing a class (object category). The number of mixtures per class is supplied to the classifier [5]. For each class the Expectation Maximization algorithm is applied to establish the mixture/group means and covariance whereby identical group prior probabilities are estimated. The class-conditional probability density function of each class is determined by substituting the mean vector and covariance matrix of each mixture into the multivariate Gaussian distribution equation and summing all these probability values.

Even if normal distribution has been considered over many years, the raised problem by asymmetry or fat tails phenomenon leads to the concepts of other distributions, taking into account this typical feature. The class mixture probabilities are determined by the proportion of samples belonging to a specific class in the training set. To classify a new observation, the class mixture is computed by using the numbers of mixtures per class, iterated from 1 to 10, to determine the best possible number of mixtures per class. It should be noted that the Gaussian Mixture Classifier makes use of diagonal covariance matrices for the mixtures [6].

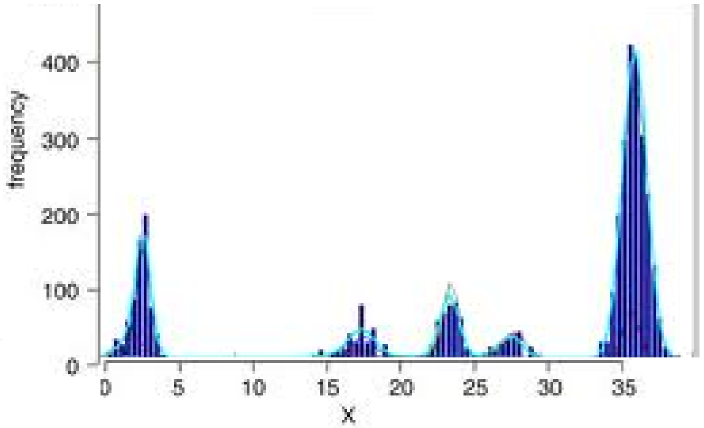


Fig. 5.3: Empirical distribution of same component Gaussian mixture model

6. REFERENCES

1. Abramowitz Milton and Irene A. Stegun (1972). *Handbook of Mathematical Functions with Formulas*, *Graphs, and Mathematical Tables*. New York, Dover.
2. Agarwal, S.K. and Al- Saleh, Jamal (2001). Generalized gamma type distribution and its hazard rate function, *Commun. in Statist., Theo. and Meth.*, 30(2), 309-318.
3. Agarwal, S.K. and Kalla, S.L. (1996). A generalized gamma distribution and its applications in relativity. *Commun. in Statist., Theo. and Meth*, 25, 201-210.
4. Agarwal, S.K. and Kalla, S.L. (1996). A generalized gamma distribution and its application in reliability, *Commun. in Statist., Theo. and Meth*.
5. Al- Saleh, Jamal and Agarwal, S.K. (2002).Finite mixture of certain distributions, *Commun. in Statist., Theo. and Meth*, 31(12), 2123-37.
6. Al- Saleh, Jamal and Agarwal, S.K. (2006). Extended Weibull Type distribution and Finite mixture of distributions. *Statistical Methodology*, 3, 224-233.
7. Al- Saleh, Jamal and Agarwal, S.K. (2007). Finite Mixture of Gamma Distribution: A Conjugate Prior, *Computational Statistics and Data Analysis*, 51(9), 4369-78.
8. Al-Musallam, F. and Kalla, S.L. (1998). Further results on a generalized gamma function ocurring in diffraction theory. *Integral Transforms and Special Function*, 7, 175-190.
9. Al-Saleh, J. and Agarwal, S. (2006). Extended Weibull type distribution and finite mixture of distributions. *Statistical Methodology*, 3, 224-233.
10. Arset, M.V. (1987). How to identify bathtub hazard rate. *IEEE Transactions on Reliability*, 36, 106–108.
11. Bachioua, Lahcene (2008). Extended Generalized Type Mixture Model and the Estimation of Hazard Rate, *International Conference on Recent Trends in Mathematical Sciences (ICRMS2008)*, Department of Mathematics, College of Science , University of Bahrain, November 10th–12th Bahrain.
12. Bachioua, Lahcene (2004). *On Extended and Reliability General Mixture Gamma Distribution Model*, A Dissertation Submitted to The College of Science / University

of Baghdad in Partial Fulfillment of The Requirements for The Degree of Doctor of Philosophy (Ph.D) of Science in Mathematics, University of Baghdad, Iraq.

13. Bachioua, Lahcene (2006). On Generalized Gamma Distribution Function. *First conference in Mathematics*, Department of mathematics, college of Applied science and Mathematics Zarqa Private University Amman-Jordan 18-20 April . Jordan.
14. Bachioua, Lahcene (2009). On Extended Generalized Gamma Distribution. *International Journal of Applied Mathematics & Statistics*, 15, D09,
15. Bachioua, Lahcene (2011). On Extended Generalized Gamma Distribution Function. *Third conference in Mathematics*, Department of mathematics, college of Applied science and Mathematics Zarqa Private University Amman-Jordan 18-20 April. Jordan.
16. Bain, L.J. and Weeks, D.L. (1965). Tolerance limits for the generalized Gamma distribution. *Journal of the American Statistical Association*, 60, 1142–1152.
17. Balakrishnan, N. and Basu, A.P. (1996). *The Exponential Distribution: Theory, Methods and Applications*, Gordon and Breach, Singapore.
18. Berrettoni, J. (1964). Practical applications of the Weibull distribution. *Industrial Quality Control*, 21, 71-79.
19. Bondesson, L., (1992). *Generalized Gamma convolutions and related classes of distributions and densities*, Lecture Notes in Statistics 76, Springer Verlag, New York.
20. Cherian, K. (1941). A bivariate correlated gamma type distribution Function. *J. Ind. Math. Soc*., 5, 133-144.
21. Coles, S.G. (1989). On goodness-of-fit tests for the two-parameterWeibull distribution derived from the stabilized probability plot. *Biometrika*, 76, 593–598.
22. Djamil Ziou, Nizar B., and Ali El-Zaart, (2007). Finite Gamma Mixture Modeling Using Minimum Message Length Inference, *Pattern Recognition Journal* (Submitted).
23. Elderton, Sir W.P, and Johnson, N.L. (1969). *Systems of Frequency Curves*. Cambridge University Press.
24. Everitt B.S, Hand D.J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London, UK, London.
25. Ghitany, M.E. (1998). On a recent generalization of gamma distribution, *Communications in Statistics-Theory and Methods*, 27, 223–233.
26. Gupta, A. K. and Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications*. NY: Marcel Dekker, Inc.
27. Hung W.L. and J.W. Wu (1999). Some Properties of the Extended Generalized Logistic-Gamma Distribution with Applications. *International Journal of Information and Management Sciences*, 10(4), 41-58.
28. Kalla, S.L., Al-Saqabi, B.N. and Khajah, H. (2001). A unified form of gamma type distributions. *Applied Mathematics and Computation*, 118, (2-3), 175-187.
29. Kececioglu, D. and Sun, F.B. (1994). Mixed-Weibull parameter-estimation for burn-in data using the Bayesian-approach. *Microelectronics and Reliability*, 34, 1657-1679.
30. Kopsinis Y., Thompson J. S., and Mulgrew B. (2007). System-Independent Threshold and BER Estimation in Optical Communications Using the Extended Generalized Gamma Distribution. *Optical Fiber Technology*, 13, 39-45.

31. Libby, D. L. and Novick, M. R. (1982). Multivariate generalized beta distributions with applications to utility assessment. *Journal of Educational Statistics*, 7, 271-294.
32. Nadarajah, S. and Gupta, A. (2006). Some bivariate gamma Distributions. *Applied Mathematics Letters*, 19, 767-774.
33. Stacy E. W. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics*, 33, 1187-1192.
34. Wu J.W., W.L. Hung and H.M. Lee (2000). Some Moments and Limit Behaviors of the Generalized Logistic Distribution with Applications, *Proc. Natl. Sci. Counc.*

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 17-32

PHYSICAL ACTIVITY AMONG DUBAI POPULATION PREVALENCE AND SOME ASSOCIATED FACTORS

Hamid Y Hussain, Nehad Hassan Mahdi, Fatma Al Attar and **Nagy Hamid**
Dubai Residency Training Program, Dubai Health Authority, Dubai
Email: hussainh569@hotmail.com

ABSTRACT

Regular practicing of physical activity considered to be one of the easiest and cost effective way of improving and maintaining health and avoiding diseases like Diabetes mellitus, cardiovascular diseases, obesity and others.

Objectives: the study, aims to study the prevalence of physical activities among Dubai population and the effect of some associated factors, it is also aiming to assess the knowledge, practice, attitudes of Dubai population.

Methodology: a cross sectional survey has been carried out upon representative random sample of adult Dubai population age rang (18-65) years, the sample was identified from schools, universities, primary health care centers visitors, governmental offices, commercial Malls and house hold families, sample size was estimated by using epi info soft ware, it was 2226 individuals of different age, sex, income, social class. The questionnaire covered variety of domains like socio-demographic data, Knowledge, attitudes, practice. Importance of physical activates, and reasons of avoidances.

Results: the study reveals that about 23.6% of the total sample showed good knowledge about the importance of physical activity and 86.6% showed positive attitude towards practicing physical activities, the study showed that about 34.6% of the total sample are practicing physical activity regularly(prevalence rate among Dubai adult population), it was appear that practicing of physical activity is significantly higher among emirates in comparison with expatriates, highly educated individuals (university and above), and high income people (10000 ED and above), the study showed that the main reason behind non practicing physical activity were lack of time 47.3%, tiredness and exhaustion 20.1%. UN availability of suitable places 17.3%, the multiple logistic regression analysis showed that there are four factors significantly affect on practicing of physical activities in Dubai, they are, Nationality odds ratio was 1.49 among Emirates compared to expatriates, Educational level, odds ratio was 2.00 among higher education compared with low education (primary school). Awareness and knowledge factor Odds Ration 3.49 and income factor showed higher practicing of physical activity among individuals with high income (10000 and above) compared to low income individuals less than 10000 ED.

Recommendations: The study recommend establishing national public health program to approach physical activity problem and developing effective strategies to deal with the causes stand behind this problem like, the time management, offering more facilities, increasing awareness and creating incentive system.

INTRODUCTION

Leisure physical activities and recreational practices has utmost importance in the present era, it is no longer a community interested in providing physical activity as much interest in providing the best ways to invest, and each community develop its own approach, where it notes that the experiences of practicing physical activities vary depending on the cultures of individuals and communities (1).

The practice of regular physical activity is one of the best and easiest ways to improve and maintain health and avoid or reduce the incidence of certain diseases such as diabetes, obesity, cardiovascular diseases and others (2). Physical activity was defined as the amount of energy utilized by the individual on daily or weekly basis, resulting from the muscles activity to sustain vital life and keep the body functioning healthy. Physical activity classified on the basis of moderate intensity by 30 minutes for five days per week and high intensity by 20 minutes two to three days a week. (3).

The results of recent research conducted both in North America or in Europe revealed a significant reduction in the level of physical activity and that this decrease occurs after the age of 12 years and continue until the age of 18, the most in girls (4).Data resulting from the health surveys collected from different parts of the world, showed striking image, where the proportion of adults of low or semi low- physical activity between 60% and 85% in all parts of the globe. (3) the practice of physical activity among young people in various parts of the world is declining, especially in poor urban areas as it was estimated that less than one third of youth people only engaged in physical activities 56% male and 36% females in North America and Canada(5).

The results of the research on the levels of physical activity and health among young Saudi made it clear that the rate of physical inactivity is high at the age of 13 years to reach its peak in adulthood, which increased from 54% in childhood from 7 years to 71% at 23 years, also coincided with increased rates of physical inactivity, high rates of obesity (6).

The national health survey on health indicators and patterns of life, which was conducted in the United Arab Emirates in 2000 survey, showed that 49.4% of the UAE population living Sedentary, and only 20.6% are people with higher physical activity level (7).

The pattern of sedentary life is a major underlying cause of death, disease, and disability. About two million deaths can be attributed to the lack of physical activity. The preliminary results of a study conducted by the WHO on the risk factors (potential exposure to risk), that the pattern of sedentary life is one of the ten leading causes of death and disability in the world. And increases the lack of physical activity of all causes of death, also doubles the risk of exposure to the risk of cardiovascular disease and type II diabetes (diabetes), and obesity. It also increases the likelihood of exposure to the risk of disease, colon cancer and breast cancer, high blood pressure, lipid disorders, osteoporosis, depression, and anxiety (8 and 9).

OBJECTIVES

1. To assess the level of information and attitudes towards physical activity and its importance for health.
2. To study the prevalence of physical activity among residents in Dubai.
3. To study the effect of some associated factors on physical activity among residents in Dubai. Social and demographic factors: Age - Sex - Nationality - educational level - social status - the status of work and monthly income

MATERIALS AND METHODS

Study design

A cross section survey among Dubai population has been carried out

Study Setting

Dubai from February to April 2009

Target population

Adult population of Dubai (18-65) years old**.**

Sample Size

Sample size was estimated by "EPI-INFO -" 6.04 "" 33% prevalence of physical activity (7), 2% accuracy and 95% limits of confidence,and had reached the minimum sample size required is 2119 Individuals

Sample design

A stratified random sample was used to representing the population of the city of Dubai, where it was sampled from each of both areas (Deira and Bur Dubai) which again divided according to geographical distribution of the city of Dubai, samples were selected from the health centers, affiliated schools and universities, families and visitors to shopping malls and household families until completing the sample.

DATA COLLECTION

Tool of Data collection:

A questionnaire has been developed, tested and used for data collection by direct interview

Content of questionnaire included four levels about:

1. Questions on Socio-demographic data, age, sex, educational level, social status
2. Questions on the knowledge and benefits of physical activity, 4 question on the benefits of physical activity and diseases which can be avoided by physical activity.
3. Questions on the Attitude towards physical activity, included 6 question on relation of physical activity with time, weight, Diabetes mellitus and stress

4. Questions on the practice of physical activities Diseases that can be avoided by physical activity number of times the practice of physical activity per week including 4 questions on frequency and intensity of practicing physical activity.

Physical activity has been categorized according to WHO criteria into the following groups:

1. Physically inactive: not practicing physical activity at all or less than 150 minutes per week.
2. Physically active (practicing different type of physical activities like rapid walking, lifting heavy weights less than 20 kg, etc for at least 150 minutes per week or more
3. vigorous physical activity (running, swimming, aerobics or lifting heavy weights more than 20 kg for 20 minutes three times per week.

STATISTICAL ANALYSIS

- The data were analyzed using the Statistical Package for Social Sciences SPSS "13"
- The use of stability coefficient Cronbach Alfa and Guttman to test the consistency of information and to identify trends.
- factor analysis statistical tool was used to test the reliability of the questionnaire
- Chi square was used to study the relationship between physical activity and demographic characteristics, information and trends.
- Multiple logistic regression statistical test was used to study the factors affecting the physical activity (dependent factor is the practice of physical activity and non-dependant factors are age, sex, nationality, social status, educational level, monthly income, the level of information and trends.
- Statistical significant level used was 95% and P value was less than 0.05.

ETHICAL ISSUE

Has been considered to the best and participants consent obtained.

RESULTS

The study sample included 2226 participants, 56.1% of males and 43.9% female, ages ranged between 18-65 years and mean age of(28.27) years, and they form (Emirates)more than half of the respondents (56.9%), about (35.4%)have a university level of education / Graduate while the percentage of uneducated or those with primary education (27.9%) . for the social status the study revealed that (47.1%) were singles and (46.8%)married, as related to the work they study showed two-thirds of participants are workers (63.7%).

Alpha Cronbach test for internal consistency was 0.73 and manner Guttmann indivisible descriptive questions-odd and even 0.78 . Reliability by test Factorial analysis as interpreted by four factors: the amount of variability was only 2%, and thus were found over the appropriate tool for the application.

The study revealed that Most of the participants (92.5%) have some knowledge about the importance of physical activity in relation to health and diseases as in table (2) . study showed that that 93% of the sample did not realize all the benefits of physical activity and 6.7% have full knowledge, more than half of the respondents (59.8%) had insufficient information about the diseases that can be avoided by physical activity, while only 31.9% had given a complete answer concerning the disease. About 42.2% of the participants the Knew the minimum requirements of physical activity per week to maintain good health. It was observed that out of the total sample 67.3% have a medium level and 23.7% had a good level of information as observed from Figure (1).

Table (3) shows the attitude of the participants towards physical activity, the tendency has been positive for the importance of physical activity to 78.6%, and attitude of the participants towards the importance of physical activity in maintaining ideal weight, and release tension, and a the importance of physical activity even when the lack of time and overloaded of responsibilities has been a positive attitude with 77.4%, 57.5% and 77.3% respectively.

For all participants, the positive attitude towards physical activity has shown by figure (2) 86.6% and neutral attitude is 12.9%. It has been shown that only 34.3% of participants have been actively engaged in physical activity as reflected by figure (3) and that is the prevalence rate of physical activities among Dubai population.

According to the table (4), the relationship between physical activity and demographic characteristics, shows that participants who more practicing of physical activity are, more than 20 - less than 30 years, followed by 35 - less than 50 years (35.5%, 33.8%, respectively), while was lowest at the age of 50-65 years (25.9%), but these differences did not statistically significant P < 0.05.

The study pointed out that the practice of physical activity have been observed at a higher rate among males than females but no statistical significant difference (35.6%, 32.5%, respectively) and observed that the Emirates are exercising more than non-Emirates in terms of statistically significant (39.8% and 30.1% respectively, chi square = 22.79 P <05 . and in reference to of educational level, it was found that the highest percentage of physical activity among university graduates / people with high graduate (41.1%) while the lowest was for the uneducated or those with primary education (26.8%) the difference was statistically significant (chi square = 32.11 P<0.05 . For martial status it was found that the highest percentage of physical activity among married couples (36.9%), followed by divorced and singles (33% and 32%, respectively) without statistically significant value .

The study pointed out there is no difference statistically significant between the practice of physical activity in individuals working and not working (34.7% and 33.6%, respectively) and concerning monthly income physical activity increased with higher monthly income participant in comparison with low,, reaching 40.7% for individuals who have a monthly income of 20.000 dirham and lowest percentage of persons with a monthly income of less than AED 5000 or (5000 - 10000) AED (32.7% and 29.9% respectively) which was statistically significant difference (chi square = 12.20 significance level less from 05 and 0).

Concerning the relationship between the level of information and physical activity Table (5) shows the highest rate of physical activity were showed among those with good level of information about physical activity (49.7%) compared to those who have average or poor (30.5%, 21.9% respectively) and this difference was statistically significant (chi square = 78.90 significance level less than 0. 05).

Regarding the relationship between attitudes and practice of physical activity (Table 6) the results show that individuals with a positive attitude have highest percentage of practicing exercises compared to those with neutral or negative attitude (35% and 29.6% and 27.3% respectively), but this difference did not prove to be statistically significant. The study shows that the reasons for non-practice of physical activity, as pointed out by figure (4) are lack of time (47.3%), followed by fatigue and stress (20.1%) and lack of places to exercise (17,3%) About 11.2% and 10.9%, respectively, due to absence of encouragement and costs of sports activities.

And by studying the effect of combined factors physical activity, by multi-stage logistic regression (Table 7) the study revealed the existence of four major factors affecting physical activity, nationality, where the odds ratio was among non- emirates 1.49 compared to the n emirates, level of education, where the individuals with primary education is about twice the lack of physical activity, the Knowledge and information level as individuals with low information are at risk of 3.49 in comparison with high knowledge people in practicing physical activity, finally the monthly income when it is less than less than 10.000 Dirham. The study pointed to the existence of statistically significant relationship between the extent to which individuals perceive their health and physical activity, explained by table (8).

About 37.7% among persons who reported practicing of physical activity have realized better health, 26.7% in excellent health comparing to those who did not engage in physical activity (34.2% and 15.8% respectively), this relationship is a statistically significant. (Chi quare = 60.80, significance level less than 5 0.0) About the participant’s perception of physical activity, about 77.2% of the participants recognized its importance and impacts.

Table 1:
Distribution of study population according to demographic data

Demographic Data	Frequency N=2226	Percentage
Age (years)		
Less than 20	471	21.7
-20	1405	64.7
-35	237	10.9
50-65	58	2.7
Standard Deviation ± Mean	8.42±28.27	
Sex		
Male	1249	56.1
Female	977	43.9
Citizenship		
National	960	43.1
Non National	1266	56.9
Marital Status		
Single	1048	47.1
Married	1041	46.8
Divorce	106	4.8
Widow	31	1.3
Educational Level		
Uneducated/Primary	619	27.9
Elementary	359	16.1
Secondary	459	20.6
University/ Higher Education	789	35.4
Work Status		
Not Working	807	36.3
Working	1419	63.7
Monthly Income (AED)		
Less than 5000	510	26.4
-5000	344	17.8
-10000	547	28.3
-15000	175	9.0
+20000	359	18.5

Table 2
Distribution of Study Sample according to the information about Physical Activity

Physical Activity Information	Frequency N=2226	Percentage
Physical Activity useful for Health:		
Wrong Answer	166	7.5
Wright Answer	2060	92.5
Benefits of Physical Activity:		
Wrong Answer	5	0.2
Incomplete Wright Answer	2071	93.0
Complete Write Answer	150	6.7
Diseases that can be avoided by doing Physical Activity:		
Wrong Answer	184	8.3
Incomplete Wright Answer	1331	59.8
Complete Write Answer	711	31.9
The number of days/week that you should do physical activity to keep you healthy:		
Wrong Answer	57.8	1287
Wright Answer	42.2	939

Table 3
Distribution of Study Sample According to the Attitudes Towards Physical Activities

Physical Activity Determinants	Not Sure	Agree	Not Agree
Do you think physical activity not important and loose your time	169 7.6	307 13.8	1750 78.6
Do you think physical activity help in reducing weight and keep you shape well?	1722 77.4	48 18.3	96 4.3
Do you think if you don't do physical activity you will be in risk for heart problems	1375 61.8	723 32.5	128 5.8
Do you think if you don't do physical activity you will be in risk for diabetic disease	1370 61.5	712 32.0	144 6.5
From your point of view physical activity help in reducing stress or help in adapt the stress	1679 75.5	457 20.5	90 4.0
From your point of view physical activity important even if you don't have the time or other responsibilities.	1720 77.3	363 16.3	143 6.4

Table 4: Distribution of Study Sample According to Practicing of Physical Activity by Demographic Variables

Demographic Data		Practicing physical activity		Total N=2226	Significant Test
		No N=1463	Yes N=763		
Age (Years)	Less than 20	314 66.7	157 33.3	471 100.0	CHI Square = 2.89
	-20	906 64.5	499 35.5	1405 100.0	
	-35	157 66.2	80 33.8	237 100.0	
	50-65	43 74.1	15 25.9	58 100.0	
Sex	Male	804 64.4	445 35.6	1249 100.0	CHI Square = 2.31
	Female	659 67.5	318 32.5	977 100.0	
Nationality	National	578 60.2	382 39.8	960 100.0	CHI Square = *22.79
	Non National	885 69.9	381 30.1	1266 100.0	
Marital Status	Single	713 68.0	335 32.0	1048 100.0	CHI Square = 6.09
	Married	657 63.1	384 36.9	1041 100.0	
	Divorce	71 67.0	35 33	106 100.0	
	Widow	22 71.0	9 29	31 100.0	
Educational Level	Uneducated/Primary	453 73.2	166 26.8	619 100.0	CHI Square = 32.11*
	Elementary	235 65.5	124 34.5	359 100.0	
	Secondary	310 67.5	149 5.32	149 100.0	
	University/ Higher Education	465 58.9	324 41.1	789 100.0	
Work Status	Working	536 66.4	271 33.6	807 100.0	CHI Square =0.27
	Not Working	972 65.3	492 34.7	1419 100.0	
Monthly Income	Less than 5000	343 67.3	167 32.7	510 100.0	CHI Square = *12.20
	-5000	241 70.1	103 29.9	344 100.0	
	-10000	341 62.3	206 37.7	547 100.0	
	-15000	108 61.7	67 38.3	175 100.0	
	+20000	213 59.3	146 40.7	359 100.0	

* Missing Data foe 291 P value <0.05

Table 5
Distribution According Information and Practice

The Information level about physical activity	Physical Activity Practicing		Total
	Yes	No	
Weak	157 78.1	44 21.9	201 100.0
Medium	1041 69.5	457 30.5	1498 100.0
Good	265 50.3	262 49.7	527 100.0
Total	1463 65.7	763 34.3	2226 100.0

*Chi square =78.9 P value <0.05

Table 6
Distribution of Study Sample According to Attitudes and Practice

The level Toward physical activity	Physical Activity Practicing		Total
	Yes	No	
Negative	8 772.	3 27.3	11 100.0
Neutral	202 70.4	85 29.6	287 100.0
Positive	1253 65.0	675 35.0	1928 100.0
Total	1463 65.7	763 34.3	2226 100.0

Chi square 3.47 P value >0.05

Table 7
Multiple Logistic Regression for Factors Affecting Physical Activity

Regression Coefficient	Independent Variables	Odd S Ratio	CI 95% Minim.-Maxim
Nationality	0.397	1.49	1.84-1.20
The Educational Level Primary/Elementary/Secondary	0.308 0.660	1.36 1.94	1.71-1.08 2.49-1.50
The Information Level Middle Weak	0.741 1.249	2.10 3.49	2.62-1.68 5.32-2.29
Monthly Income 10000 Less than 10000 Less than 15000	0.082- 0.248	0.922 1.28	1.23-0.69 1.0-1.70

Table 8
Distribution According to Recognizing Effect of Physical Activity on Health

Physical Activity Practice	**Health Perception Level**					**Total**
	Very Good	Excellent	Acceptable	Good	Weak	
No	219 15.8	475 34.2	409 29.5	246 17.7	39 2.8	1388 100.0
Yes	201 26.7	284 37.7	185 24.5	70 9.2	14 1.8	754 100.0
Total	420 19.6	759 35.4	594 27.7	316 14.8	53 2.5	• 2142 100.0

- no answer for 84 about recognizing effect of physical activity on Health
- Chi Square =60.80,P value <0.05

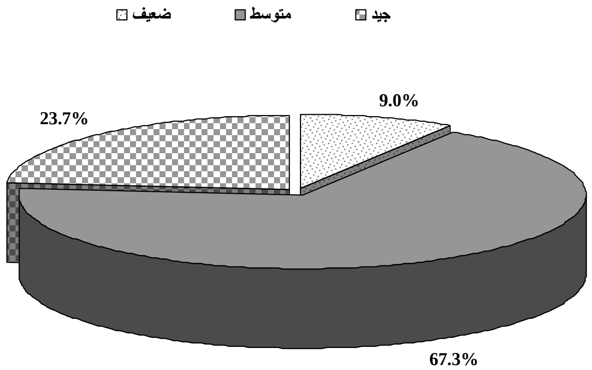


Fig. 1: Distribution of Study Population According to Level of Information about Physical Activity (67.3 Medium, 23.7 Good, 9.0 Week).

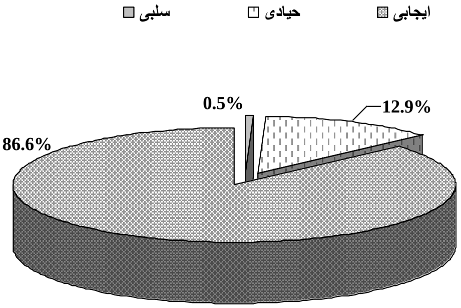


Fig. 2: Distribution According to Attitudes Towards Physical Activity (86.6 Positive, 12.9 Nutral, 0.5 Negative).

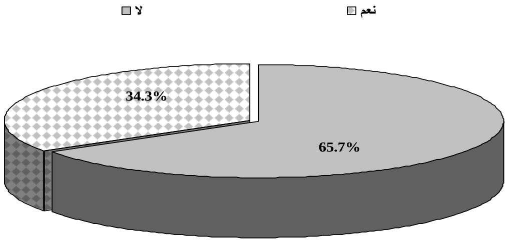


Fig. 3: Distribution of Sample According to the Practice of Physical Activity (34.3 Practising,65.7 Non Practicing).

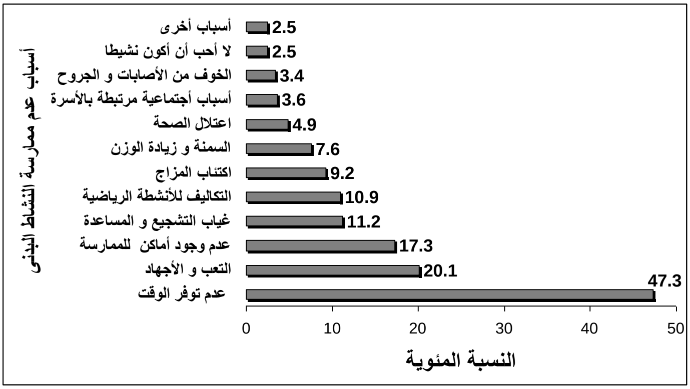


Fig. 4: Distribution According to the Causes of None Practicing (47.3 Lack of Time, 20.1 Stress and Exhaustion, 17.3 Lack of Suitable Places, 11.2 Lack of Encouragement, 10.9 Cost of Physical Activity)

DISCUSSIONS

This study has been carried out to identify the prevalence rate of physical activity among Dubai population and to identify the most important determinants of physical activity in the emirates of Dubai, where the study showed that the level of the prevalence rate of practicing of physical activity, as defined by WHO criteria among general population of the was a rate of (34.3%) which is almost equal to one third of Dubai population and this figure reflect low rate which can be explained by immersion of population in the management of development projects as a result of the rapid developmental processes witnessed by this emirate, it is low in comparison with the study conducted in Vietnam (2001-2004 ((13), which showed higher rates of exercise, reaching up to 44% and another study in Brazil showed exercise up to 49% (14) and the EU countries ranged from 91% in Finland to a minimum 40% in Portugal (15), while physical activities among Americans are up to 45% among, about 43% of women and 48% among men (16). In A similar and other study conducted in the United Arab Emirates (the National Health Survey 2000) showed that about 49.3% of the population living physical inactive lifestyle(sedentary life style) and only 20.6% of population doing higher physical activity level (5), similar to results a study in Saudi Arabia by Mohammed Al-Hazza showed that most children and adolescents in Saudi Arabia do not practice the minimum required for physical activity (16) and another study by same researcher showed that 78% of adults are physically inactive (17).

The study showed that the level of knowledge among a study population on the importance of physical activity was low in general, good answers was 23.7% only, while other responses ranged between medium and weak which indicates that real need to work on this aspect in the future and also reflects the that the majority of the public has knowledge within and intermediate levels, this result is in similar with the result concluded by Emirates national survey 2000 (5).

With regard to the assessment of attitudes toward physical activity among study population, the study showed that 86.6% of the study population showed positive attitude which reflects significantly on their healthy behavior, yet the level of the actual practice of physical activity among study population stay low which indicate the need for identifying the real determinants in a step to be managed for further improvements.

The study highlighted the possible reasons behind the non-practicing physical activity by the study population and found that was the most important factors are lack of time 47.3% which indicate the need to work with the government channels to develop regulations and legislations to manage this issue, then fatigue and stress factor, and lack of appropriate places, absence of encouragement, high costs and depressed mood, which is fully consistent with the study carried out in Armenia, which referred to the statistically significance effect of age, sex, educational level, income level factors (, 18, 17) which also showed in UAE national survey (5) Thus, drawing attention to these factors are extremely important in each intervention program to improve physical activity practicing among general population.

With respect to age, the study shows that the age group of 20-35 is the highest prevalence of physical activity level of 36% in contrast to the age group 50-65 years as the lowest category of physical activity, This is consistent with the one that studies conducted in the UAE and Saudi Arabia (, 18,5,17).

According to the sex variable, both males and females are almost showed similar prevalence rates in practicing physical activity, on the contrary with other study in Vietnam (13), where women have demonstrated superiority in numbers of males in physical activity level of 65% to 49%, while the study conducted in Emirates showed that females engaged less in physical activity than males (5).

A research conducted in the United States of America showed about 63% of all workers do not engage in physical activity at work place and sitting for a long time (20.19) which is in similar with the results of this study that shows 65.3% of workers not doping any sort of exercise at work place. When assessing the effect of some associated factors with by applying multiple logistic regression analysis, it has been shown that odds ratio was statistically significant with the factors of nationality, low educational level, level of information and monthly income, and this result is consistent with the findings of the study in Brazil (14) but differ from the study of Vietnam (13) as factors of –smoking and non working women were of the of the most important reasons for non-exercise physical activity in addition to monthly income and higher education factor was one of the determinants in other study.

CONCLUSION

The research found that about two third of Dubai's population are living a sedentary lifestyle, mainly due to lack of time required for practicing regular physical activity as the most important reason as well as other reasons such as non-availability of suitable places and increased costs of sports activities, in spite of low prevalence rate of practicing physical activities among Dubai population the majority of the study population revealed positive attitudes towards physical activity and showed a high understanding of its importance and impacts on individuals and population health.

RECOMMENDATIONS

The study recommends the following:

1. The importance of developing a national physical activity program to approach the phenomenon of lack of physical activity among community individuals and to explore effective strategies to deal with the reasons for reluctance to paucity of time, lack of adequate places and the costs of sports activities and others.
2. Providing adequate and suitable sports venues that encourage the practice of physical activity and sports in Dubai.
3. The need to provide educational interventions about the importance of physical activity for health promotion.
4. Upgrading physical education curricula at schools and universities to improve and change practice, attitude and knowledge about Physical activities.

REFERENCES

1. Ahmed Bin Mohamed Al Fathial (2008). Effect of physical education teaching during leisure time on their attitudes. *Education and psychology*, N 29, 215-237.
2. Al Myziani Khalid Bin Saleh (2003). *Physical education prescription for all ages, Arab Hournal for Nutrition 4th year*. N8 June, 48-76.
3. Hazza Mohamed Physiology of physical activity, Theoretical bases and lab finding, Textbook, for printing at College of Education, Abdul Aziz King University. Saudia Arabia Kingdom.
4. Adbul Wahab Al Najar (2008). *Physical education department college of Education*. King Saood University, Article published on 15th December, Arab website for Health and fitness.
5. WHO (2002). Health International day / WHO 7th April
   http://www.emro.who.int/whd2002/Readings.htm
6. Hazza Mohammed, M Al Ahmedy (2004). *Estimate physical activity and energy expenditure, importance and Standard methods of measurement*. Research centre, college of Education. King Saood University, Saudia Arabia Kingdom.
7. (2000). *Health Survey UAE 2000*. College of Business and Economy.
8. (2009). *Diabetes mellitus Guideline for better care*. Article
   http://www.diabetes-om.com/ar/diabetes-news/1-latest-news/53-diabetes-sport.htm
9. (2009). *Our Health our responsibility*, article /9.
   http://www.sehetna.com/pages.php?PageID=91

10. International physical activity questionnaire to measure the level of physical activity in the past seven days of youth and adults. Physical activity questionnaire shortcut. Website: www.ipaq.ki.se.
11. Briefed physical activity questionnaire, www.ipaq.ki.s.
12. WHO (2005). *Recommended amount of physical activity*. Global strategy on diet, physical activity and health WHO. Geneva.
13. Jekel, J.F. and Katz, D.L. (2001). *Elmore JG. Epidemiology, biostatistics and preventive medicine*. 2nd edition. Philadelphia, London. New York: WB Saunders Company.
14. Munter P, GU D, Wildman R, Chen J, Qan W, Welton PK, He J. (2005). Prevalence of Study of cardio vascular diseases in Asia. *American Journal of Public Health*, 95, 1631-1636.
15. Hallal, PEDRO CURI, Victoria (2003). Physical activity prevalence in Brazil, *Medicine and Science in Sports and Exercise*, 35(11), 1894-1900.
16. Angil Maratenze, Miculi, Javier (2001). Prevalence of physical activities during leisure time in European Union countries. *Medicines and science in sports and Exercise*, 33(7), 1142-1146.
17. Macera C.A. (2005). Ham SA, Prevalence of physical activities in United States. *Prevalence Chronic Diseases*, 2(2).
18. Al Hazzaa M. (2002). Physical activity fitness and fatness among children and adolescent in. 17 Saudi Arabia. *Saudi Medical Journal*, 23(2), 144-150.
19. M. Al. Hazzaa. P. (2004). Prevalence of physical activities in Saudi Arabia. *East Mediterranean Health Journal*, 10, 415.
20. Lina Hako Bayan (2008). Physical activity national survey in Armenia, research. Fellowship, Caucusus research centre, Arminea.
21. Macera C. and Jones D. (2000). Life style physical activity intervention. *Ann Epidemiol* 10(7), 4.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 33-42

A GENERALIZATION OF THE STANDARD HALF–CAUCHY DISTRIBUTION

Saleha Naghmi Habibullah
Department of Statistics, Kinnaird College for Women, Lahore, Pakistan.
Email: salehahabibullah@gmail.com

ABSTRACT

Cordiero and Lemonte (2011) utilize the generator approach suggested by Eugene et al. (2002) to obtain a generalization of the Half–Cauchy (HC) distribution with location parameter equal to zero that they refer to as the Beta-Half-Cauchy (BHC) distribution. They plot the density and hazard rate functions, provide explicit expressions for the moments, moment generating function, etc. and illustrate the usefulness of the distribution for lifetime data modeling. In this paper, we utilize the generalized differential equation presented by Habibullah et al. (2009) to obtain a generalization of the standard Half-Cauchy distribution that is self-inverse at unity and can be called the Self-Inverse at Unity Half-Cauchy (SIUHC) distribution. The distribution is positively skewed and its graph exhibits a fair amount of flexibility in terms of the extent of skewness indicating its potential for modeling lifetime data pertaining to both biological and non-biological life. Moments about the origin and about the mean are obtained in terms of trigonometric functions. The phenomenon of self-inversion facilitates the derivation of a large number of additional properties. As indicated by Habibullah and Saunders (2011), the self-inversion property of the distribution carries implications for improving the efficiency of the empirical cumulative distribution function.

1. INTRODUCTION

Cordiero and Lemonte (2011) comment that the statistics literature is full of a multitude of continuous univariate distributions that have been used extensively over the past decades for modeling data related to engineering, medicine, environmental science, demography, economics, finance and other fields. However, in the opinion of the authors, in applied areas such as lifetime analysis and insurance, there is a clear need for extended forms of these distributions i.e. new distributions that are more flexible since the data at hand may have a high degree of skewness and kurtosis and, according to these authors, additional control can be obtained over both skewness and kurtosis by adding new parameters. They refer to the generator approach pioneered by Eugene et al. (2002) as a very good example of techniques that are now being developed for building meaningful distributions and, in particular, mention the beta-normal (BN) distribution the parameters of which control skewness through the relative tail weights.

In their paper, Cordiero and Lemonte (2011) obtain a generalization of the Half–Cauchy distribution with location parameter equal to zero by adopting the generator approach suggested by Eugene et al. (2002) and refer to it as the beta-half-cauchy (BHC) distribution. They investigate some mathematical properties of the new model, discuss maximum likelihood estimation of its parameters, and derive the observed information matrix. They claim that the proposed model is much more flexible than the half-Cauchy (HC) distribution and, as such, can be used effectively for modeling lifetime data.

Habibullah (2009) presents two types of differential equations that yield an unlimited number of distributions that are invariant under the reciprocal transformation or ‘Strictly Closed Under Inversion’ (SCUI) Habibullah et al. (2009) present a generalized differential equation for generating such distributions. Habibullah and Saunders (2011) introduce the concept of self-inversion, and regard SCUI distributions as being “self-inverse at unity”.

In this paper, we utilize the generalized differential equation of Habibullah et al. (2009) for generating a density function that is self-inverse at unity and seems to carry the potential for modeling lifetime data. The density can be regarded as a generalization of the standard half-Cauchy distribution. As such, we would like to call it the Self-Inverse at Unity Half-Cauchy (SIUHC) distribution.

The graph of the density is positively skewed and exhibits a fair amount of flexibility in terms of the extent of skewness indicating its applicability for modeling lifetime data pertaining to both biological and non-biological life. A fairly large number of properties including those pertaining to the mode and modal ordinate, raw and central moments, quantiles, hazard rate, etc., are obtained. Most of the expressions obtained are in terms of trigonometric functions.

2. THE SELF-INVERSION PROPERTY

The detailed discussion of the concept of self-inversion given in Habibullah and Saunders (2011) is summarized below:

If χ is a class of non-negative (life-length) random variables which is defined on $0,\infty$, it is said to be closed under inversion (CUI) whenever $X \in \chi$ implies its reciprocal is also in the class. When $X \in \chi$ is such that $X \sim 1/X$, it can be regarded as being strictly closed under inversion (SCUI). (See Habibullah (2009).) In case of SCUI variates, we have F(x)=1-F(1/x) for all x>0

A variate Y with $Y/a \sim a/Y$ can be regarded as being ***“self-inverse at a",*** *a* being regarded as the point of reciprocity. The SCUI variate *X* with $X \sim 1/X$ is then obviously self-inverse at unity.

However, Habibullah and Saunders (2011) assert that, in general, the term ``self-inversion" will be used without specifying whether the point of reciprocity is $a \neq 1$ or unity. (This specification will be made only when there is a risk of *misinterpretation* unless the point of reciprocity is indicated.)

As demonstrated by Habibullah and Saunders (2011), the self-inversion property of a distribution carries implications for improving the efficiency of the empirical cumulative distribution function.

3. GENERALIZED DIFFERENTIAL EQUATION FOR GENERATING DISTRIBUTIONS SELF-INVERSE AT UNITY

Habibullah et al. (2009) present the following theorem that contains a ***generalized*** differential equation along with a set of conditions that yield probability functions that are Strictly Closed Under Inversion (SCUI) or, in other words, self-inverse at unity:

Theorem 3.1:

Let $g(y)$ be the *pdf* of $Y = \ln X$ where the random variable X has the *pdf* $f(x)$ defined on $(0,\infty)$. If

$$\frac{d}{dy}\left[\ln g(y)\right] = \frac{\sum_{i=0}^{n} b_i \left[w(y)\right]^i}{\sum_{i=0}^{n} a_i \left[w(y)\right]^i} \tag{3.1}$$

then $f(x)$ is SCUI provided that the following conditions hold:

Case I: $w(y)$ is an odd function of y i.e. $w(y) = -w(-y)$

(a) $a_i \neq 0$ and $b_j \neq 0$ for some *i,j*, $0 \leq i, j \leq n$, and

(b) $$\sum_{i=0}^{2j} (-1)^i a_{2j-i} b_i = 0\,, \quad j = 0,1,2,....,m\,,$$

$$\sum_{i=0}^{2j} (-1)^i a_{n-i} b_{n-2j+i} = 0\,, \quad j = 0,1,2,....,m \tag{3.2}$$

where m is $\frac{n}{2}$ or $\frac{n-1}{2}$ according as *n* is an even or odd non-negative integer,

Case II: $w(y) = \left[w(-y)\right]^{-1}$

(a) $a_i \neq 0$ and $b_j \neq 0$ for some *i,j*, $0 \leq i, j \leq n$,

(b) $$\sum_{i=0}^{j} \left(a_i b_{i+n-j} + a_{i+n-j} b_i\right) = 0, \quad j = 0,1,2,....,n-1,$$

$$\sum_{i=0}^{n} a_i b_i = 0 \tag{3.3}$$

4. THE 'SELF-INVERSE AT UNITY' GENERALIZATION OF THE HALF-CAUCY DISTRIBUTION

Letting n=2 in Eq. (3.1), we have

$$\frac{d}{dy}\left[\ln g(y)\right] = \frac{b_2\left[w(y)\right]^2 + b_1\left[w(y)\right] + b_0}{a_2\left[w(y)\right]^2 + a_1\left[w(y)\right] + a_0} \tag{4.1}$$

Now, letting $w\ \ y\ = e^{ay}$ and putting

$$b_2 = -a,\ b_1 = 0,\ b_0 = a\ \ and\ \ a_2 = 1,\ a_1 = 0,\ a_0 = 1$$

where $a > 0,$ we obtain

$$\frac{d}{dy}\ \ln g(y)\ \ = \frac{-ae^{2ay} + a}{e^{2ay} + 1} \tag{4.2}$$

We note that $w\ \ y\ = \left[w\ -y\ \right]^{-1}$ and the set of conditions (3.3) is fulfilled.

Applying the transformation $X=e^{Y}$ we obtain

$$f(x;a) = \frac{2a}{\pi}\left(\frac{x^{a-1}}{1+x^{2a}}\right),\ 0 < x < \infty,\ a > 0 \tag{4.3}$$

which is a one-parameter density self-inverse at unity. It is easy to see that for a = 1, the density reduces to the standard half-Cauchy distribution.

5. DISTRIBUTION FUNCTION

The distribution function is given by

$$F\ \ x\ \ = \begin{cases} 0, & x \le 0 \\ \dfrac{2}{\pi}\left[\tan^{-1}\ \ x^{a}\ \ \right], & 0 < x < \infty \end{cases} \tag{5.1}$$

6. MODE

Following the usual procedure for the determination of the mode, we have

$$\hat{x} = \left(\frac{a-1}{a+1}\right)^{\frac{1}{2a}} \tag{6.1}$$

The modal ordinate comes out to be

$$f(\hat{x}) = \frac{a-1\ ^{\frac{a-1}{2a}}}{\pi\ \ a+1\ ^{\frac{-a-1}{2a}}} \tag{6.2}$$

7. GRAPHICAL REPRESENTATION

Figure 7.1 contains the graph of the density (4.3) whereas Figure 7.2 presents the graph of the distribution function (5.1) for various values of the parameter a.

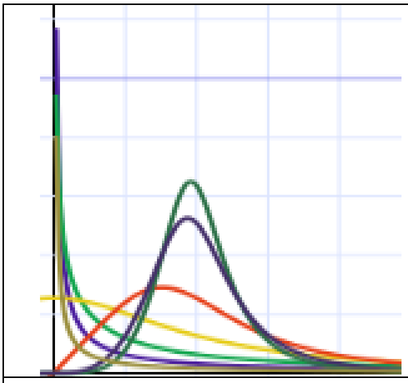


Fig. 7.1
The graph of the Density (4.3)
for *a*=1/8, ¼, ½, 1, 2, 4, 5

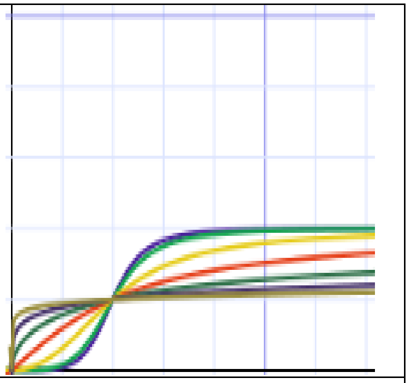


Fig. 7.2
The graph of the DF (5.1)
for *a*=1/8, ¼, ½, 1, 2, 4, 5

As seen in Figure 7.1, for a=1 (standard half-Cauchy distribution), the graph of the density begins from a height equal to $2/\pi = 0.64,$ not from zero. For all values of a<1, the graph of the distribution is “exponential-type”. On the other hand, for all values of a>1, the graph of the distribution rises from level zero, reaches a maximum and then declines gradually. As such, for all values of a>1, the distribution is unimodal and moderately positively skewed, and, as seen in Figure 7.1, the skewness decreases as a increases.

In view of the above observations, we may conclude that the parameter *a* controls the shape of the frequency curve so that we can regard it as the shape parameter.

Remark: Habibullah (2009) proves that every unimodal pdf in the class of SCUI distributions defined on $0,\infty$ is positively skewed.

8. MOMENTS ABOUT THE ORIGIN AND ABOUT THE MEAN

In this section, we derive moments about the origin and about the mean which are obtained in terms of trigonometric functions.

8.1 Moments about the Origin:

We have

$$E\ \ X^r \ = \frac{2a}{\pi}\int_0^\infty \left(\frac{x^r x^{a-1}}{1+x^{2a}}\right)dx = \frac{2a}{\pi}\int_0^\infty \left(\frac{x^{a+r-1}}{1+x^{2a}}\right)dx$$

Now, from Gradshteyn and Ryzhik (2007), we know that

$$\int_0^\infty \left(\frac{x^{\mu-1}}{1+x^{\nu}}\right)dx = \frac{\pi}{\nu}\cos e\,c\left(\frac{\mu\pi}{\nu}\right)$$

so that

$$\mu_r' = E\ \ X^r \ = \frac{2a}{\pi}\int_0^\infty \left(\frac{x^{a+r-1}}{1+x^{2a}}\right)dx = \cos ec\left(\frac{a+r\ \ \pi}{2a}\right) = \sec\left(\frac{r\pi}{2a}\right) \qquad (8.1)$$

8.2 Arithmetic Mean:

From (8.1), we have

$$E\ \ X\ \ =\sec\left(\frac{\pi}{2a}\right) \tag{8.2}$$

from which it is obvious that for $a=1$ (half-Cauchy distribution), the mean does not exist.

8.2 Recurrence Relation:

From (8.1) we have

$$E\ \ X^{r+1}\ \ =\sec\left(\frac{r+1\ \ \pi}{2a}\right)=\frac{1}{\cos\left(\frac{r\pi}{2a}\right)\cos\left(\frac{\pi}{2a}\right)-\sin\left(\frac{r\pi}{2a}\right)\sin\left(\frac{\pi}{2a}\right)}$$

or

$$E\ \ X^{r+1}\ \ =\frac{E\ \ X^{r}}{\cos\left(\frac{\pi}{2a}\right)-\sin\left(\frac{r\pi}{2a}\right)\sin\left(\frac{\pi}{2a}\right)E\ \ X^{r}} \tag{8.3}$$

8.3 Moments about the Mean:

The variance is given by

$$\mu_2=E\ \ X^2\ \ -\left[E\ \ X\ \ \right]^2=\sec\left(\frac{\pi}{a}\right)-\left[\sec\left(\frac{\pi}{2a}\right)\right]^2 \tag{8.4}$$

so that the coefficient of variation comes out to be

$$CV=\frac{\sqrt{\sec\left(\frac{\pi}{a}\right)-\left[\sec\left(\frac{\pi}{2a}\right)\right]^2}}{\sec\left(\frac{\pi}{2a}\right)}$$

Also, utilizing the relationships between moments about the origin and the central moments, we have

$$\begin{aligned}\mu_3&=\mu_3'-3\mu_2'\mu_1'+2\mu_1'^3\\&=\sec\left(\frac{3\pi}{2a}\right)-3\sec\left(\frac{\pi}{a}\right)\sec\left(\frac{\pi}{2a}\right)+2\left[\sec\left(\frac{\pi}{2a}\right)\right]^3\end{aligned}$$

and

$$\begin{aligned}\mu_4&=\mu_4'-4\mu_3'\mu_1'+6\mu_1'^2\mu_2'-3\mu_1'^4\\&=\sec\left(\frac{2\pi}{a}\right)-4\sec\left(\frac{3\pi}{2a}\right)\sec\left(\frac{\pi}{2a}\right)+6\left[\sec\left(\frac{\pi}{2a}\right)\right]^2\sec\left(\frac{\pi}{a}\right)-3\left[\sec\left(\frac{\pi}{2a}\right)\right]^4\end{aligned}$$

8.4 Coefficients of Skewness and Kurtosis:

The coefficient of skewness is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{\left[\sec\left(\frac{3\pi}{2a}\right) - 3\sec\left(\frac{\pi}{a}\right)\sec\left(\frac{\pi}{2a}\right) + 2\left[\sec\left(\frac{\pi}{2a}\right)\right]^3\right]^2}{\left[\sec\left(\frac{\pi}{a}\right) - \left[\sec\left(\frac{\pi}{2a}\right)\right]^2\right]^3}$$

and the coefficient of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\sec\left(\frac{2\pi}{a}\right) - 4\sec\left(\frac{3\pi}{2a}\right)\sec\left(\frac{\pi}{2a}\right) + 6\left[\sec\left(\frac{\pi}{2a}\right)\right]^2 \sec\left(\frac{\pi}{a}\right) - 3\left[\sec\left(\frac{\pi}{2a}\right)\right]^4}{\left[\sec\left(\frac{\pi}{a}\right) - \left[\sec\left(\frac{\pi}{2a}\right)\right]^2\right]^2}$$

8.5 Negative or Inverse Moments:

Replacing *r* by –*r* in (8.2), we have

$$E\ \ X^{-r}\ \ = \sec\left(-\frac{r\pi}{2a}\right) = \sec\left(\frac{r\pi}{2a}\right) = E\ \ X^r \quad (8.4)$$

This is in accordance with the property of SCUI random variables proved by Habibullah 2009).

9. GEOMETRIC AND HARMONIC MEANS:

From Habibullah (2009), we know that the geometric mean of any density self-inverse at unity is 1 and the harmonic mean the reciprocal of its arithmetic mean. As such, we have

$$G.M.\ \ X\ \ = 1 \tag{9.1}$$

$$H.M.\ \ X\ \ = \cos\left(\frac{\pi}{2a}\right) \tag{9.2}$$

10. QUANTILES AND MEDIAN

The q^{th} quantile of the density (4.3) is obtained as follows:

For every *q* lying in (0,1),

$$\frac{2a}{\pi}\int_0^{X_q}\left(\frac{x^{a-1}}{1+x^{2a}}\right)dx = q \Rightarrow \tan^{-1}\ \ X_q{}^a\ \ = \frac{\pi q}{2} \Rightarrow X_q{}^a = \tan\left(\frac{\pi q}{2}\right)$$

so that

$$X_q = \left[\tan\left(\frac{\pi q}{2}\right)\right]^{\frac{1}{a}} \tag{10.1}$$

It is obvious from (9.3) that, in accordance with the property proved by Habibullah (2009), the q^{th} quantile of the density is the reciprocal of its $1-q^{\ th}$ quantile.

Putting *q=0.5*, we have

$$\tilde{X} = \left[\tan\left(\frac{\pi}{4}\right)\right]^{\frac{1}{a}} = 1 \qquad (10.2)$$

implying that the median of each distribution (4.3) is unity.

11. PROPERTIES PERTAINING TO RELIABILITY THEORY

In this section, we present some properties pertaining to the survival function, hazard rate and mean residual life.

11.1 Survival Function:

The survival function is given by:

$$\bar{F}\ \ x\ \ = 1-F\ \ x\ \ = 1-\frac{2}{\pi}\left[\tan^{-1}\ \ x^{a}\ \ \right], 0<x<\infty \qquad (11.1)$$

Now

$$F(1/x) = 2\left[\tan^{-1}\ \ 1/x^{a}\ \ \right]/\pi = 2\left[\frac{\pi}{2}-\tan^{-1}\ \ x^{a}\ \ \right]/\pi$$

or

$$F(1/x) = 1-2\tan^{-1}\ \ x^{a}\ \ /\pi = \bar{F}\ \ x$$

so that the well-established result pertaining to the distribution functions of SCUI random variables is easily proved.

11.1 Hazard Rate:

The hazard rate is obtained as follows:

$$h\ \ x\ \ = \frac{f(x)}{1-F(x)} = \frac{2ax^{a-1}/\pi\ \ 1+x^{2a}}{1-2\left[\tan^{-1}\ \ x^{a}\ \ \right]/\pi}$$

or, in other words,

$$h\ \ x\ \ = \frac{2ax^{a-1}}{\left[\pi-2\tan^{-1}\ \ x^{a}\ \ \right]\ 1+x^{2a}} \qquad (11.2)$$

Figure 11.1 shows the graph of the hazard rate for various values of a. As seen in Figure 11.1, for a<1, the graph of the hazard rate is monotonically decreasing so that the density (4.3) is DHR. On the other hand, for all values of a>1, the graph of the hazard rate rises from level zero, reaches a maximum and then declines gradually. For a=1 (standard half-Cauchy distribution), the graph of the hazard rate does not start from level zero but from level $2/\pi = 0.64,$ and can be seen to rises slightly before it turns its direction and declines gradually toward the x-axis. As such, we can say that, for all values of $a \geq 1$, the hazard rate of the density (4.3) has an upside down bathtub shape.

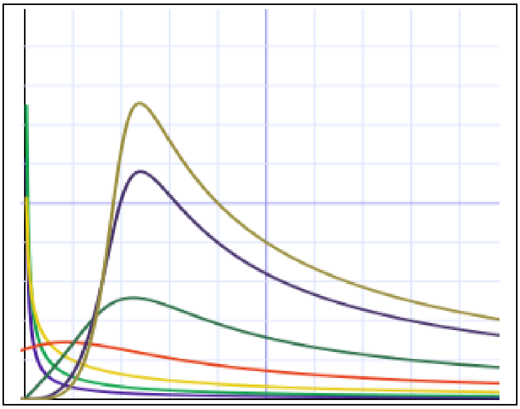


Figure 11.1
The graph of the Hazard Rate (10.2) for *a*=1/8, ¼, ½, 1, 2, 4, 5

DFR distributions have been discussed widely in the literature during the past half-century (see, for example, Barlow et al. (1963), Gurland and Sethuraman (1994) and Nesterenko et al. (1996). Among others, Erto (1989), Jiang et al. (2003), Bebbington et al. (2008), Shaban and Knopik (2011) and Palumbo and Pallotta (2012) provide an exposition/discussion and/or real-life examples of the Upside Down Bathtub-Shaped hazard rate. The ability to encompass both DFR and Upside Down Bathtub-Shaped hazard rates testifies to the flexibility of the newly derived SIUHC distribution and places it among the class of distributions that possess intuitive appeal for modeling lifetime data.

12. CONCLUDING REMARKS:

It is well-known that, in a large variety of situations, the empirical probability distribution of life-length is moderately positively skewed. In this paper, we have obtained a generalized version of the standard half-Cauchy distribution that is skewed to the right and is self-inverse at unity. The positively skewed shape of the SIUHC distribution points to its potential for modeling lifetime data and self-inversion at unity facilitates the derivation of a large number of properties. More importantly, as shown by Habibullah and Saunders (2011), the self-inversion property of the distribution carries implications for improving the efficiency of the empirical cumulative distribution function.

REFERENCES

1. Barlow, R.E., Marshall, A.W. and Proschan, F. (1963). Properties of Probability Distributions with Monotone Hazard Rate. *Annals of Mathematical Statistics*, 34(2), 375-389.
2. Bebbington, M., Lai, C-D. and Zitikis, R. (2008). A Proof of the Shape of the Shape of the Birnbaum Saunders Distribution. *Math. Scientist*, 33, 49-56.

3. Cordeiro, G.M. and Lemonte, A.J. (2011). The Beta-Half-Cauchy Distribution. *Journal of Probability and Statistics*, Vol. 2011, Article ID 904705, 18 pages, doi: 10.1155/2011/904705.
4. Erto, P. (1989). Genesis, Properties and Identification of the Inverse Weibull Lifetime Model. *Statistica Applicata* 1, 117-128.
5. Eugene, N., Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications. *Commun. in Statist., Theo. and Meth.*, 31(4), 497-512.
6. Gupta, R.D. (2001). Exponentiated Exponential Family: An Alternative to Gamma and Weibull Distributions. *Biometrical Journal*, 43(1), 117-130.
7. Gradshteyn, I.S. and Ryzhik, I.M. (2007). *Table of Integrals, Series and Products*. Academic Press, Elsevier, Seventh edition, page 322.
8. Gurland, J. and Sethuraman, J. (1994). Reversal of Increasing Failure Rates when Pooling Failure Data. *Technometrics*, 36(4), 416.
9. Habibullah, S.N. (2009). *On a Class of Distributions Closed Under Inversion*. PhD Thesis, National College of Business Administration and Economics, Lahore, Pakistan.
10. Habibullah, S.N., Memon, A.Z. and Ahmad, M., (2009). On a Generalized Differential Equation for Generating SCUI Distributions. *Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS X)*, Cairo, Egypt, December 20-23.
11. Habibullah, S.N. and Saunders. S.C. (2011). A Role for Self- Inversion, *Proceedings of the International Conference on Advanced Modeling and Simulation (ICAMS)*, Rawalpindi, Pakistan, Nov 28-30.
12. Jiang, R., Ji, P. and Xiao, X. (2003). Aging Property of Unimodal Failure Rate Models. *Reliability Engineering and System Safety*, 79, 113-116.
13. Knopik, L. (2011). Mixture of Distributions as a Lifetime Distribution of a Bus Engine, *Journal of Polish CIMAC*, Gdansk University of Technology, www.polishcimac.pl/Papers1/2011/012.pdf.
14. Nesterenko, M.V., Upton, S.J. and Kochar, S.C. (1996). Dispersive ordering of order statistics. *Statistics & Probability Letters*, 27(3), 271-274.
15. Palumbo, B. and Pallotta, G. (2012). New Approach to the Identification of the Inverse Weibull Model, Contributed Paper, *46th Scientific Meeting of the Italian Statistical Society*.
16. Shaban, S.A. and Boudrissa, N.A. (2008). Failure Rate of the Weibull-Weibull Length-Biased Mixture Model, interstat.statjournals.net/YEAR/2008/articles/0810010.pdf.
17. http://www.onlinefunctiongrapher.com accessed on July 23, 2012 for drawing the graphs contained in Fig. 7.1, Fig. 7.2 and Fig. 11.1.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 43-52

A NOTE ON JOINT INCLUSION PROBABILITIES SAMPLING WITHOUT REPLACEMENT

Muhammad Hanif and **Aftab Ahmad**
National College of Business Administration and Economics, Lahore, Pakistan
Email: drhanif@ncbae.edu.pk; aftab.stat@gmail.com

ABSTRACT

The expression of sampling variance of an estimator of a finite population total involves the first two orders inclusion probabilities π_i and π_{ij} . The expression of the second order inclusion probabilities π_{ij}. For most of the unequal probability sampling design is very complex. So a number of researchers have paid their attention in approximating π_{ij} in terms of π_i and π_j and derived some useful results. In this paper we have tried to classify the methods in to three groups and compare their performance by calculating their entropy values and obtained some important results. This gives easy to execute approximations of Horvitz-Thompson (1952) variance of population total and its estimate. The entropy of these new groups of approximations is compared empirically. The π_{ij} can be approximated using the first order inclusion probabilities of any sampling design. In Section 2 we will introduce some basic relationships relating to the first and second order inclusion probabilities to be derived in the next three sections. In Sections 3, 4, and 5 and designated as groups 1, 2 and 3 modified approximations will be derived. Numerical calculations, comparisons and concluding remarks will be discussed in the last section.

KEYWORDS

Horvitz–Thompson estimator; First order inclusion probabilities; Joint inclusion probabilities; Approximations.

1. INTRODUCTION

Consider a finite population of N units labeled i ($i = 1, 2,3,\ldots, N$) and let Y_i be an unknown value attached to unit i. Consider the problem of estimating the total $Y=\sum_1^N Y_i$ from a sample s of n units randomly drawn without replacement but with unequal probabilities. The probability of drawing a sample s is denoted by P_s so that $\sum_{s\in\Omega} P_s = 1$. Let $\pi_i = \sum_{s\ni i} P_s$ and $\pi_{ij} = \sum_{s\ni i,j} P_s$. π_i and π_{ij} are called first and second order inclusion probabilities respectively. The Horvitz-Thompson (1952) estimate of population Y is given by $\hat{Y} = \sum_{i\in s}\frac{Y_i}{\pi_i}$ and the variance $Var(\hat{Y})$ is given by

$$Var(\hat{Y})_{HT} = \frac{1}{2}\sum_{i}\sum_{j\neq i}^{N} \; \pi_i\pi_j - \pi_{ij} \; \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2. \tag{1.1}$$

The estimate of (1.1) suggested independently by Sen (1953) and Yates and Grundy (1953) is

$$\mathrm{var}(\hat{Y})_{SYG} = \frac{1}{2}\sum_{i}\sum_{\substack{i,j\in s\\ j\neq i}} \left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2, \tag{1.2}$$

where π_{ij} for all possible pairs (i ,j) when $j \neq i$ are non zero. The dependence of expression (1.2) on the π_{ij}'s sometimes makes its use problematical particularly when some of the π_{ij}'s are zero or near zero, which makes $\mathrm{var}(\hat{Y})_{SYG}$ unstable. Sometimes the evaluation of the π_{ij}'s poses special problems and complications. As an alternative we can search for approximations of joint inclusion probabilities in the literature. Some references to the families of approximations are Hartley and Rao (1962), Brewer (1963), Hájek (1964), Samiuddin and Asad (1981), Brewer and Hanif (1983), Herzel (1986), Tillé (1996), Deville (1996), Sunter (1986), Hanif and Ahmad (2001), Shahbaz and Hanif (2003) and Brewer and Donadio (2003).

Brewer and Donadio's work (2003) primarily relates to high entropy sampling procedures. It appears to us that their work is potentially of greater importance and significance. The π_{ij} can be approximated using the first order inclusion probabilities of any sampling design. We will here apply Aftab and Hanif's (2012) maximum entropy sampling design to provide the maximum entropy π_{ij}'s. In Section 2 we will introduce some basic relationships relating to the first and second order inclusion probabilities to be derived in the next three sections. In Sections 3, 4, and 5 and designated as groups 1, 2 and 3 modified approximations will be derived. Numerical calculations, comparisons and concluding remarks will be discussed in the last section.

2. SOME IMPORTANT RELATIONSHIPS

In unequal probability sampling without replacement, first and second order inclusion probabilities play an important role in the construction of these estimators of total or mean and in their variance estimators. For fixed sample sizes Brewer and Hanif (1983) listed the following relationships of first and second order inclusion probabilities:

a) $$\sum_{j\neq i}^{N}\pi_{ij} = \; n-1 \; \pi_i. \tag{2.1}$$

b) $$\sum_{j\neq i}^{N}\pi_i\pi_j = \pi_i \; n-\pi_i \; . \tag{2.2}$$

c) $$\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}\pi_{ij} = n \; n-1 \; . \tag{2.3}$$

d) $$\sum_{i=1}^{N}\sum_{j\neq i}^{N}\pi_i\pi_j = n^2 - \sum_{i=1}^{N}\pi_i^2. \tag{2.4}$$

The expressions (2.1) to (2.3) are derived by Cochran while the (2.4) is derived by Hanif (1970). Among these four relationships (2.1) is of crucial importance. Expressions (2.2) (2.3) and (2.4) are consequences of (2.1) and $\sum_{i=1}^{N} \pi_i = n$ (for fixed size of sample). Brewer and Donadio (2003) then look for an approximation of π_{ij} to be used with the Horvitz- Thompson (1952) variance expression. We now modify it in the next section and include it in the group 1 approximations.

3. THE GROUP-1 APPROXIMATIONS

We now consider the approximation (3.1), which was suggested by Hanif and Ahmad (2001) and by Brewer and Donadio (2003) as

$$\pi_{ij} \cong \left(\frac{c_i + c_j}{2} \right) \pi_i \pi_j , \; j \neq i. \qquad (3.1)$$

Brewer and Donadio (2003) suggested following three different choices for c_i :

i) $c_i = \left(\frac{n-1}{n-\pi_i} \right),$ (3.2)

ii) $c_i = c = \left(\frac{n-1}{n - \frac{1}{n} \sum_{k \in U} \pi_k^2} \right)$, and (3.3)

iii) $c_i = \left(\frac{n-1}{n - 2\pi_k - \frac{1}{n} \sum_{k \in U} \pi_k^2} \right).$ (3.4)

The (3.4) is based on an asymptotic expression for π_{ij}, which was obtained by Hartley and Rao (1962) and used for the randomized systematic πps sampling procedure also by Asok and Sukhatme (1976) and used for the Sampford’s (1967) procedure.

However we prefer to use (3.1). This enables us to write

$$\sum_{j \neq i}^{N} \pi_{ij} \cong \frac{\pi_i}{2} \sum_{j \neq i}^{N} \pi_j (c_i + c_j) = (n-1)\pi_i . \quad n-1 \;\; \pi_i = \frac{\pi_i}{2} \left[c_i \left(\sum_{j=1}^{N} \pi_j - \pi_i \right) + \sum_{j=1}^{N} \pi_j c_j - \pi_i c_i \right].$$

Hence

$$(n-1)\pi_i \cong \frac{\pi_i}{2} \left[c_i (n - \pi_i) + c - c_i \pi_i \right] \text{ where } c = \sum_{k=1}^{N} c_k \pi_k \; .$$

On simplification we get

$$c_i \cong \frac{2 \;\; n-1 \;\; - c}{n - 2\pi_i} .$$

Putting c_i in (3.1) and on simplification we get

$$\pi_{ij} \cong \frac{n-1 \ \ \pi_i\pi_j}{k}\left[\frac{1}{n-2\pi_i}+\frac{1}{n-2\pi_j}\right], \tag{3.5}$$

where $k = 1+\sum_{i=1}^{N}\frac{\pi_i}{n-2\pi_i}$.

To check that this result is correct we may write

$$\sum_{j\neq i}^{N}\pi_{ij} \cong \frac{(n-1)\pi_i}{\left[\sum_{k=1}^{N}\frac{\pi_k}{n-2\pi_k}+1\right]}\left(\frac{1}{n-2\pi_i}\sum_{j\neq i}^{N}\pi_j + \sum_{j\neq i}^{N}\frac{\pi_j}{n-2\pi_j}\right),$$

or

$$\sum_{j\neq i}^{N}\pi_{ij} \cong \frac{(n-1)\pi_i}{\left[\sum_{k=1}^{N}\frac{\pi_k}{n-2\pi_k}+1\right]}\left(\frac{n-2\pi_i}{n-2\pi_i}+\sum_{k=1}^{N}\frac{\pi_k}{n-2\pi_k}\right),$$

or

$$\sum_{j\neq i}^{N}\pi_{ij} \cong \frac{(n-1)\pi_i}{\left[\sum_{k=1}^{N}\frac{\pi_k}{n-2\pi_k}+1\right]}\left(1+\sum_{k=1}^{N}\frac{\pi_k}{n-2\pi_k}\right) = (n-1)\pi_i\,. \tag{3.6}$$

Although an approximation, π_{ij} at (3.6) is completely and uniquely determined. The need for the consideration of those three approximations in Brewer and Donadio (2003) is thus avoided. In fact if the consequences of three approximations are worked out it will reveal a lack of coherence. For example their second approximation requires $c_i = c$. This leads to $\pi_{ij} \cong c\pi_i\pi_j$.

Further $\sum_{j\neq i}\pi_{ij} \cong c\pi_i \ \ n-\pi_i \ \ = \pi_i \ \ n-1 \ \Rightarrow c = \frac{n-1}{n-\pi_i}$. If all π_i 's are not equal this leads to a contradiction.

Also it is worth mentioning that when we work out the approximation $\pi_{ij} \cong c\pi_i\pi_j\left(\frac{1}{1-a\pi_i}+\frac{1}{1-a\pi_j}\right)$ and Shahbaz and Hanif (2003) suggested approximation $\pi_{ij} \cong c_i\pi_i\pi_j + c_j\pi_i\pi_j$, we get the same result given at (3.5) so we include these two approximations in the same group and called it Group 1 approximations.

4. THE GROUP-2 APPROXIMATIONS

Consider the approximation

$$\pi_{ij} \cong A \;\; c_i + c_j \;\; + B. \tag{4.1}$$

For possible solution we have to find plausible values of A and B. To proceed further we sum both sides of (4.1) such that $j \neq i$. Now this leads to $\sum_{j\neq i}^{N} \pi_{ij} = \; n-1 \; \pi_i \cong A\left[\; N-1 \; c_i + c - c_i\right] + \; N-1 \; B$ where $c = \sum_{k=1}^{N} c_k$ (Note that this notation of c is different from the notation used in group 1 approximation). Simplifying we have

$$\begin{aligned}(n-1)\pi_i &= A\left[\; N-2 \; c_i + c\right] + \; N-1 \; B \\ &= Ac + \; N-1 \; B + A \; N-2 \; c_i\end{aligned}$$

Finally we get $A = \dfrac{(n-1)}{N-2}$, $c_i = \pi_i$ and $B = \dfrac{-n \;\; n-1}{N-1 \;\; N-2}$.

Thus (4.1) simplifies to

$$\pi_{ij} \cong \frac{n-1}{N-2} \;\; \pi_i + \pi_j \;\; - \frac{n(n-1)}{N-1 \;\; N-2}.$$

$$\pi_{ij} \cong \frac{n-1}{N-2}\left(\pi_i + \pi_j - \frac{n}{N-1}\right). \tag{4.2}$$

The above corresponds to Sen-Midzuno (1953) sampling scheme. The one by one drawing procedure is known. Notice that for $\pi_{ij} > 0$ for all pairs $j \neq i$, $\pi_i + \pi_j > \dfrac{n}{N-1}$.

Also the approximation

$$\pi_{ij} \cong A \;\; \pi_i + \pi_j \;\; + B, \tag{4.3}$$

produces the same expression (4.2). Note that the joint inclusion probability expressions for approximations (4.1), Sen-Midzuno (1953) sampling scheme and (4.3) are same. Hence we classify them as Group 2 of approximations.

To check that this result is correct we sum over both sides of (4.2) and get

$$\sum_{j\neq i}^{N} \pi_{ij} \cong \frac{(n-1)}{(N-2)} \sum_{j\neq i}^{N} \left[\pi_i + \pi_j - \frac{n}{N\text{-}1} \right],$$

or

$$\sum_{j\neq i}^{N} \pi_{ij} \cong \frac{(n-1)}{(N-2)} \left[\;\; N\text{-}1 \;\; \pi_i + \;\; n - \pi_i \;\; - \frac{n \;\; N\text{-}1}{N\text{-}1} \right] = (n-1)\pi_i .$$

5. THE GROUP-3 APPROXIMATION

Shahbaz and Hanif (2003) suggested approximation $\pi_{ij} \cong a\pi_i\pi_j$ .We slightly modify it as

$$\pi_{ij} \cong c_i\pi_i\pi_j \,. \tag{5.1}$$

Summing over both sides of (5.1) we get

$$\sum_{j\neq i}^{N} \pi_{ij} = (n-1)\pi_i \cong c_i\pi_i \sum_{j\neq i}^{N} \pi_j \,.$$

On simplification we have

$$c_i = \frac{(n-1)}{n-\pi_i}. \tag{5.2}$$

Inducting value of c_i from (5.2) in (5.1) we get

$$\pi_{ij} \cong \frac{n-1}{n-\pi_i}\pi_i\pi_j \,. \tag{5.3}$$

Check; $\sum_{j\neq i}^{N} \pi_{ij} \cong \frac{n-1}{n-\pi_i}\pi_i \sum_{j\neq i}^{N} \pi_j = \ n-1 \ \ \pi_i.$

Approximation (5.3) fulfill condition (2.1).

6. EMPIRICAL STUDY AND CONCLUSIONS

There are $\binom{N}{n}$ distinct samples of fixed sample size *n*. P_s is the probability of selecting a sample *s* such that $\sum_{s\in\Omega} P_s = 1$. The amount of uncertainty regarding the outcome *s*, if the sample is selected according to P_s is measured by the entropy, defined as $H = -\sum_{s\in\Omega} P_s \ln(P_s)$. We can compare the performance of different groups of approximations on randomness criteria i.e. with respect to their entropy values. Shannon (1948) suggested this entropy formalism for the quantitative measure of randomness.

For this purpose fifteen natural and two artificial small populations, selected from standard sampling literature, are worked out in this study. Table 6.1 displays a brief summary of some major characteristics of each population such as population size, study variable, auxiliary variable, their variability and correlation between these variables. For a sample of size 2 the population sizes vary from 4 to 20. Moreover for a sample s = (i, j) $P_s = \pi_{ij}$. We also compare the entropy of these three groups of approximations of joint inclusion probabilities with maximum entropy sampling design where

$$P_s = e^{\lambda_i+\lambda_j} \Rightarrow \pi_{ij} = e^{\lambda_i+\lambda_j}, \tag{6.1}$$

and

$$\pi_i = e^{\lambda_i}(A_1 - e^{\lambda_i}), \text{ where } A_1 = \sum_{i=1}^{N} e^{\lambda_i} , \qquad (6.2)$$

In approximations (3.6), (4.2) and (5.2) the terms π_i and π_j derived by maximum entropy sampling given at (6.2) are used.

Table 6.2 constitutes entropy values of maximum entropy sampling and three groups of approximations, abbreviated as H_{MES}, H_{G1}, H_{G2} and H_{G3} respectively.

Before proceeding further, consider the expression of joint inclusion probabilities of Group 2 approximations i.e.

$$\pi_{ij} \cong \frac{(n-1)}{(N-2)}\left[\pi_i + \pi_j - \frac{n}{N\text{-}1}\right],$$

the condition for this relation to hold is $\pi_i + \pi_j > \frac{n}{N\text{-}1}$, otherwise for some samples we may face negative values of π_{ij}, and consequently entropy cannot be calculated. Investigation reveals that to hold this condition the data for the auxiliary variable and consequently for the main variable should be well mixed; it should not involve much variability i.e. the coefficients of variation should be small.

Table 6.2 indicates that entropy values of Group 1 approximations are very close to Maximum Entropy Sampling. However it is difficult to compute Group 1 approximation (3.5). Group 2 approximations will give good results if the data fulfill the required condition i.e. $\pi_i + \pi_j > \frac{n}{N\text{-}1}$. The entropy values for populations 4 and 5 of Group 2 confirm our statement. The performance of Group 3 approximations is also reasonable. For the first two populations, where the population size is small i.e. N = 4, the entropy value is slightly larger than the corresponding entropy values for Maximum Entropy Sampling. We can conclude that the Group 3 approximation may overestimate for very small data and also that its results for the remaining fifteen populations should not be ignored either. The expression of this approximation is simple and just depend upon sample size and first order inclusion probability.

7. REFERENCES

1. Aftab, A. and Hanif. M. (2012). Maximum entropy sampling. Submitted for publication to *Pak. J. Statist.*
2. Asok, C. and Sukhatme, B. (1976). On Sampford’s procedure of unequal Probability sampling without replacement. *J. Amer. Statist. Assoc.*, 71, 912-918.
3. Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Aust. J. Statist.*, 5, 5-13.
4. Brewer, K.R.W. and Donadio, M.E. (2003). The high entropy variance of the Horvitz Thompson estimator. *Survey Methodology*, 29, 189-196.
5. Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities; Lecture Notes in Statistics*: *Volume 15*. Springer Verlag, New York.

6. Cochran, W.G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *J. Agricultural Sc., 30*, 262-275.
7. Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Statist.*, 35, 1491-1523.
8. Hanif, M. and Ahmad, M. (2001). Approximate variance formula for variance of Horvitz - Thompson estimator using first order inclusion probabilities. *Presented at ICCS-VII*, Lahore, 2001.
9. Hanif, M. and Brewer, K.R.W. (1980). Sampling with unequal probabilities without replacement: a review. *International Statistical Review*, 48, 317-335.
10. Hartley, H. and Rao, J. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.*, 33, 350-74.
11. Herzel, A. (1986). Sampling without replacement with unequal probabilities with pre-assigned joint inclusion probabilities of any order. *Metron*, 44, 1-4, and 49-68.
12. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47, 663-685.
13. Samiuddin, M. and Asad. H. (1981). A simple procedure of unequal Probability Sampling. *Biometrika*, 68(3), 728-731.
14. Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
15. Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *J. Ind. Soc. Agri. Statist.*, 5, 119-127.
16. Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technology Journal*, 27, 623-656.
17. Shahbaz, M.Q. and Hanif, M. (2003). A simple procedure for unequal probability sampling without replacement and a sample of size 2. *Pak. J. Statist.*, 19(1), 151-160.
18. Tillé, Y. (1996). Some remarks on unequal probability sampling designs without replacement. *Annales d'Economie et de Statistique*, 44, 177-189.
19. Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc., B*, 15, 235-261.

Table 6.1
Description of Populations

S#	*N*	Study Variable y	Auxiliary Variable x	ρ	C.V. (x)	C.V. (y)	Source
1	4	Small population used by Chen (1994)					
2	4	Small artificial population used Yates and Grundy (1953) and Raj (1956)		0.99	0.52	0.35	Sukhatme and Sukhatme (1970)
3	5	Small artificial population		0.99	0.50	0.68	Cochran (1977), p-268
4	10	Actual weight	Estimated weight	0.97	0.17	0.19	Cochran (1977), p-203
5	10	# of persons per block	# of rooms per block	0.65	0.14	0.15	Cochran (1977), p-325
6	10	Area under wheat in 1937	Area under wheat in 1936	0.99	0.94	0.93	Sukhatme and Sukhatme (1970)
7	10	Area under wheat in 1937	Area under wheat in 1936	0.98	0.59	0.65	Sukhatme and Sukhatme (1970)
8	10	Catch of fish in Kg	# of boats	-	-	-	Lahiri (1951)
9	14	-	-	0.98	1.11	1.46	Kish (1965)
10	15	# of people in 1930	# of people in 1920	0.94	0.69	0.67	Cochran (1977), p-152
11	16	# of inhabitants (in 1000's) of cities in 1930	# of inhabitants (in 1000's) of cities in 1920	0.99	0.98	0.98	Cochran(1977)
12	17	Oat acreage in 1957 (even units)	Total acreage in 1947	0.80	0.61	0.71	Sampford (1962), p-61
13	18	Oat acreage in 1957 (odd units)	Total acreage in 1947	0.91	0.73	0.75	Sampford (1962), p-61
14	19	Actual # of household in a block	Eye estimate of household in a block	-	-	-	Cochran(1977), p-272
15	19	Wheat acreage	# of villages	0.59	0.50	0.63	Sukhatme and Sukhatme (1970)
16	20	-	-	0.75	0.49	0.56	Yates (1960)
17	20	# of people in 1930	# of people in 1920	0.99	0.10	0.10	Cochran (1977)

Table 6.2
Comparison of Approximation with Actual Group

Popu #	H(actual)	H_{G1}	H_{G2}	H_{G3}
1	1.29682439	1.29643904	-	1.32596403
2	1.47360366	1.47331931	-	1.50752304
3	2.02829054	2.02813528	-	2.00186388
4	3.77752932	3.77752922	3.777099	3.7494449
5	3.78842290	3.78842281	3.788253	3.76553689
6	2.90343834	2.9033588	-	2.84442041
7	3.4565335	3.45652344	-	3.39213635
8	3.34027793	3.34025156	-	3.27252565
9	3.40066479	3.40062002	-	3.33361507
10	4.23737940	4.23737221	-	4.16359502
11	4.01958491	4.01952353	-	3.942271
12	4.43605934	4.43605633	-	4.36607464
13	4.52731417	4.52731215	-	4.45814155
14	4.990490682	4.99049052	-	4.94739385
15	4.87628312	4.87628269	-	4.8280029
16	5.01067405	5.01067382	-	4.96036373
17	4.35805554	4.35803694	-	4.2763613

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 53-60

ON FINITE MIXTURE OF INVERSE WEIBULL DISTRIBUTION

Kareema Abulkadhum Makhrib
Department of Mathematics, College of Education for Pure Sciences,
Bablyon University, Hilla, Iraq.
Email: kareema_kadim@yahoo.com

ABSTRACT

Mixture models play a vital role in many practical applications. So this article is considered with the mixture model of three inverse Weibull distributions (MTHIWD). Some Statistical properties of the model with some graphs of the density and hazard function are discussed.

KEY WORDS

Finite mixture; Statistical properties; inverse Weibull distribution.

1. INTRODUCTION

Everitt and Hand (1981), Titterington et al. (1985), Maclachlan and Basford (1988), Lindsay (1995), Maclachlan and Krishnan (1997) and Maclachlan and Peel (2000). Recently, AL-Hussaini and Sultan (2001) have reviewed properties and the estimation techniques of finite mixtures of some life time models. Identifiability questions of mixtures must be settled before one can meaningfully discuss the problems of estimation, testing hypotheses or classification of random variables, which are based on observations from the mixture. Identifiability gives a unique representation for a class of mixtures. Lack of identifiability is a serious problem if we intend to classify future observations into one of several classes from our knowledge of the component distributions. Identifiability of mixtures has been discussed by several authors, including Teicher (1963), Yakowitz and Spragins (1968), Balakrishnan and Mohanty (1972), AL-Hussaini and Ahmad (1981), Ahmad and AL-Hussaini (1982), and Ahmad (1988).

Jiang et al. (1999) have shown that the InverseWeibull (IW) mixture models with negative weight can represent the output of a system under certain situations.

Jiang et al. (2001) have considered the shapes of the density and failure rate functions and graphical methods to discuss the MTIWD. Jiang et al. (2003) have discussed the aging property of the unimodal failure rate models including the IW distribution. Calabria and Pulcini (1990) have discussed the maximum likelihood and least square estimates of the parameters of the IW distribution. Sultan, Ismail and Al-Moisheer (2007) investigated the mixture model of two InverseWeibull distributions (MTIWD), some properties of the model with some graphs of the density and hazard function. And they proved the identifiability property of the MTIWD proved. In addition, the estimates of the unknown parameters via the EM Algorithm were obtained.

In this paper we introduce the distribution and discuss some important of its properties and measures. Also, we prove that the MTHIWD is identifiable.

2. MIXTURE DISTRIBUTION

The mixture of $\boldsymbol{n}$ distributions has the pdf as

$$\boldsymbol{f}(\boldsymbol{x}:\boldsymbol{\Theta}) = \boldsymbol{p_1 f_1}(\boldsymbol{x}:\boldsymbol{\Theta_1}) + \cdots. + \boldsymbol{p_n f_n}(\boldsymbol{x}:\boldsymbol{\Theta_n}), \textstyle\sum_{i=1}^{n} \boldsymbol{p_i} = \boldsymbol{1} \tag{1}$$

where $\boldsymbol{\Theta} = [\boldsymbol{p_1}, \boldsymbol{\theta_{11}}, \dots \boldsymbol{\theta_{1k}}, \boldsymbol{\theta_{21}}, \dots \boldsymbol{\theta_{2k}}]$, $\boldsymbol{\Theta_i} = [\boldsymbol{\theta_{i1}}, \dots \boldsymbol{\theta_{ik}}]$,$\boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{n}$ and $\boldsymbol{f_i}(\boldsymbol{x}:\boldsymbol{\Theta_i})$ is the density function of ith component

The cdf of this distribution is given as

$$\boldsymbol{F}(\boldsymbol{x}:\boldsymbol{\Theta}) = \boldsymbol{p_1 F_1}(\boldsymbol{x}:\boldsymbol{\Theta_1}) + \cdots. + \boldsymbol{p_n F_n}(\boldsymbol{x}:\boldsymbol{\Theta_n}) \tag{2}$$

where $\boldsymbol{F_i}(\boldsymbol{x}:\boldsymbol{\Theta_i})$, is the cdf of the ith component

Now, using the matrix notation we can rewrite (1) and (2) as

$$\boldsymbol{y} = \mathrm{f}(x:\boldsymbol{\Theta}) = \boldsymbol{P}'\mathbf{f} \tag{3}$$

and

$$F(x:\boldsymbol{\Theta}) = \boldsymbol{P}'\boldsymbol{F} \tag{4}$$

where $\boldsymbol{P}' = [\boldsymbol{p_1}, \dots\dots, \boldsymbol{p_n}]$, $\mathbf{f} = [\boldsymbol{f_1}, \dots., \boldsymbol{f_n}]' = (\mathbf{y_1}, \dots, \mathbf{y_n})' = \mathbf{y}$, and $\boldsymbol{F} = (\boldsymbol{F_1}, \dots., \boldsymbol{F_n})$ each of size $\boldsymbol{nx1}$.

2.1 Properties

1. *The mean and the variance*

Since $\boldsymbol{\mu_r} = \boldsymbol{E}(\boldsymbol{X^r}) = \boldsymbol{P}'\boldsymbol{E^r}$

where $\boldsymbol{E^r} = [\boldsymbol{E_1}(\boldsymbol{X^r}), \dots \boldsymbol{E_n}(\boldsymbol{X^r})]'$

$$[\boldsymbol{p_1} \quad \cdots \quad \boldsymbol{p_n}]' \begin{bmatrix} \boldsymbol{E_1}(\boldsymbol{X}) \\ \vdots \\ \boldsymbol{E_n}(\boldsymbol{X}) \end{bmatrix} = \boldsymbol{p_1 E_1}(\boldsymbol{X}) + \quad \dots + \boldsymbol{p_n E_n}(\boldsymbol{X}) = \textstyle\sum_{i=1}^{n} \boldsymbol{p_i E_i}(\boldsymbol{X})$$

If $\boldsymbol{r} = \boldsymbol{1}$, then $\boldsymbol{E} = [\boldsymbol{E_1}(\boldsymbol{X}), \dots \boldsymbol{E_n}(\boldsymbol{X})]'$

And the mean of mixture distribution is given as

$$\boldsymbol{\mu_1} = \boldsymbol{E}(\boldsymbol{X}) = \boldsymbol{P}'\boldsymbol{E} \tag{5}$$

2. The variance of mixture distribution is given as

$$\boldsymbol{var}(\boldsymbol{X}) = \boldsymbol{\mu}_2' - \boldsymbol{\mu}_1'^{\,2}$$
$$= \textstyle\sum_{i=1}^{n} \boldsymbol{p_i E_i}(\boldsymbol{X^2}) - \left(\sum_{i=1}^{n} \boldsymbol{p_i E_i}(\boldsymbol{X})\right)^2$$

$$\boldsymbol{\sigma_X^2} = \boldsymbol{P'E^2} - \boldsymbol{P'EE'P} \tag{6}$$

$$\boldsymbol{\sigma_X^2} = [\boldsymbol{p_1} \quad \boldsymbol{p_2}] \begin{bmatrix} \boldsymbol{E_1}(\boldsymbol{X^2}) \\ \boldsymbol{E_2}(\boldsymbol{X^2}) \end{bmatrix} - [\boldsymbol{p_1} \quad \boldsymbol{p_2}] \begin{bmatrix} \boldsymbol{E_1}(\boldsymbol{X}) \\ \boldsymbol{E_2}(\boldsymbol{X}) \end{bmatrix} [\boldsymbol{E_1}(\boldsymbol{X}) \quad \boldsymbol{E_2}(\boldsymbol{X})] \begin{bmatrix} \boldsymbol{p_1} \\ \boldsymbol{p_2} \end{bmatrix}$$

$$\boldsymbol{\sigma_X^2} = \boldsymbol{p_1 \sigma_1^2} - \boldsymbol{2 p_1 p_2 E_1}(\boldsymbol{X})\boldsymbol{E_2}(\boldsymbol{X}) + \boldsymbol{p_2 \sigma_2^2}$$

3. *The moment generating function*

The moment generating function of mixture distribution is given as

$$\begin{aligned} \boldsymbol{M_X(t)} &= \boldsymbol{E(e^{tX})} \\ &= \sum_{i=1}^{n} \boldsymbol{p}_{\mathrm{i}} \boldsymbol{E}_i(e^{tX}) \\ &= \boldsymbol{P'M} \end{aligned} \tag{7}$$

where $\boldsymbol{M} = [\boldsymbol{E_1}(e^{tX}), \ldots \boldsymbol{E_n}(e^{tX})]' = [\boldsymbol{M_1}(t), \ldots \boldsymbol{M_n}(t)]$

4. *The median and the mode*

The median of mixture distribution is obtained by the solving the following equation

$$\boldsymbol{P'F = 0.5} \tag{8}$$

And the mode (modes) of mixture distribution is (are) obtained by solving the following nonlinear equation with respect to *x*

$$\boldsymbol{d(P'\mathrm{f})/dx = 0} \tag{9}$$

5. *The Reliability and Failure Rate Functions*

The reliability (survival) function of mixture distribution is given as

$$\boldsymbol{R(x)} = \boldsymbol{1} - [\boldsymbol{p_1}, \ldots\ldots, \boldsymbol{p_n}] \begin{bmatrix} \boldsymbol{F_1} \\ \vdots \\ \boldsymbol{F_n} \end{bmatrix}$$

That is

$$\boldsymbol{R(x)} = 1 - \boldsymbol{P'F} \tag{10}$$

And the failure rate function (hazard rate the failure rate function, HRF) of mixture distribution is given as

$$\boldsymbol{h(x)} = \boldsymbol{P'}\mathbf{f} \,/(1 - \boldsymbol{P'F}) \tag{11}$$

The reverse hazard function is defined as

$$\boldsymbol{r(x)} = \boldsymbol{P'}\mathbf{f} \,/\boldsymbol{P'F} \tag{12}$$

3. MAIN RESULTS

3.1 The mixture of three Inverse Weibull distribution (MTHIWD) has its pdf, according to (3), as

$$\boldsymbol{f(x;\Theta) = p_1 f_1(x;\Theta_1) + p_2 f_2(x;\Theta_2) + p_3 f_3(x;\Theta_3)}, \tag{13}$$

3

where $\boldsymbol{p_1 + p_2 + p_3 = 1, \Theta = (\Theta_1,\ \Theta_2, \Theta_3\ ), \Theta_{\mathrm{i}} = (\alpha_i, \beta_i)}$, and

$$\boldsymbol{f_i(x;\Theta_{\mathrm{i}}) = \beta_i \alpha_i{}^{-\beta_i} x^{-(\beta_i+1)} e^{-(\alpha_i x)^{-\beta_i}}}, \mathbf{x} \geq \boldsymbol{0}, \boldsymbol{\alpha_i}, \boldsymbol{\beta_i} > 0, i = 1,2,\boldsymbol{3}$$

The cdf of the MTHIWD, according to (4) is given as

$$\boldsymbol{F(x;\Theta) = p_1 F_1(x;\Theta_1) + p_2 F_2(x;\Theta_2) + p_3 F_3(x;\Theta_3)} \tag{14}$$

where

$$F_i(x;\Theta_i) = e^{-(\alpha_i x)^{-\beta_i}}, \mathbf{x} \geq \mathbf{0}, \alpha_i, \beta_i > 0, i = 1,2,3, \tag{14.1}$$

is the cdf of ith component.

3.1.1 *Properties*

1. *Mean and variance*: The mean of the MTHIWD , according to (5), is given as

$$E(X) = \sum_{i=1}^{3} p_i E_i(X)$$

$$\mu_{1=}' \frac{p_1}{\alpha_1}\Gamma\left(1-\frac{1}{\beta_1}\right)+\frac{p_2}{\alpha_2}\Gamma\left(1-\frac{1}{\beta_2}\right)+\frac{p_3}{\alpha_3}\Gamma\left(1-\frac{1}{\beta_3}\right), \beta_i > 0 \tag{15}$$

And the variance of the MTHIWD, according to (6), is given as

$$\sigma^2 = \frac{p_1}{\alpha_1^2}\left[\Gamma\left(1-\frac{2}{\beta_1}\right)-p_1\Gamma\left(1-\frac{1}{\beta_1}\right)\right]+\frac{p_2}{\alpha_2^2}\left[\Gamma\left(1-\frac{2}{\beta_2}\right)-p_2\Gamma\left(1-\frac{1}{\beta_2}\right)\right]$$
$$+\frac{p_3}{\alpha_3^2}\left[\Gamma\left(1-\frac{2}{\beta_3}\right)-p_3\Gamma\left(1-\frac{1}{\beta_3}\right)\right]-\frac{2p_1p_2}{\alpha_1\alpha_2}\left[\Gamma\left(1-\frac{1}{\beta_1}\right)\Gamma\left(1-\frac{1}{\beta_2}\right)\right]$$
$$-\frac{2p_1p_3}{\alpha_1\alpha_3}\left[\Gamma\left(1-\frac{1}{\beta_1}\right)\Gamma\left(1-\frac{1}{\beta_3}\right)\right]-\frac{2p_2p_3}{\alpha_2\alpha_3}\left[\Gamma\left(1-\frac{1}{\beta_2}\right)\Gamma\left(1-\frac{1}{\beta_3}\right)\right] \tag{16}$$

2. *Mode and median*: The mode (modes) of the MTHIWD is (are) obtained by solving the following nonlinear equation with respect to *x*

$$\sum_{i=1}^{3} p_i\beta_i{\alpha_i}^{-\beta_i}x^{-(\beta_i+2)}e^{-(\alpha_i x)^{-\beta_i}}\left[-(\beta_i+1)+\beta_i{\alpha_i}^{-\beta_i}x^{-\beta_i}\right]=0 \tag{17}$$

And the median is as

$$\sum_{i=1}^{3} p_i e^{-(\alpha_i x)^{-\beta_i}} = 0.5 \tag{18}$$

3. *Reliability and failure rate functions*: The reliability function (survival function) of the MTHIWD , according to (10) is given as

$$R(x) = p_1R_1(x;\Theta_1)+p_2R_2(x;\Theta_2)+p_3R_3(x;\Theta_3) \tag{19}$$

where $R_i(x;\Theta_i) = 1 - e^{-(\alpha_i x)^{-\beta_i}}, i = 1,2,3$ is the reliability component.

And failure rate function of the MTHIWD, according to (11) with substitution (12) and (19) in (11), is given as

$$h(x) = p_1h_1(x;\Theta_1)+p_2h_2(x;\Theta_2)+p_3h_3(x;\Theta_3) \tag{20}$$

where $h_i(x) = \frac{p_i f_i(x;\Theta_i)}{p_i R_i(x;\Theta_i)} = \frac{\beta_i{\alpha_i}^{-\beta_i}x^{-(\beta_i+1)}e^{-(\alpha_i x)^{-\beta_i}}}{1-e^{-(\alpha_i x)^{-\beta_i}}}$, is the hazard function of ith component which can be written as follows

$$h(x) = r_1(x)h_1(x)+r_2(x)h_2(x)+r_3(x)h_3(x) \tag{21}$$

where

$$\boldsymbol{r_1(x) = 1/(1+\frac{p_2R_2(x;\Theta_2)+p_3R_3(x;\Theta_3)}{p_1R_1(x;\Theta_1)})} \tag{22}$$

$$\boldsymbol{r_2(x) = 1/(1+\frac{p_1R_1(x;\Theta_1)+p_3R_3(x;\Theta_3)}{p_2R_2(x;\Theta_2)})} \tag{23}$$

$$\boldsymbol{r_3(x) = 1/(1+\frac{p_1R_1(x;\Theta_1)+p_2R_2(x;\Theta_2)}{p_3R_3(x;\Theta_3)})} \tag{24}$$

The hazard function of the MTHIWD given in (21) satisfies the following limits.

Lemma 1.

$$\lim_{x\to 0} h(x) = 0 \tag{25}$$

and $\lim_{x\to\infty} h(x) = 0$ (26)

Proof:

If we prove that $\boldsymbol{\lim_{x\to 0} h_i(x) = 0, i = 1,2,3,}$, then we get $\boldsymbol{\lim_{x\to 0} h(x) = 0}$.

$$\boldsymbol{\lim_{x\to 0} \frac{\beta_i {\alpha_i}^{-\beta_i} x^{-(\beta_i+1)} e^{-(\alpha_i x)^{-\beta_i}}}{1-e^{-(\alpha_i x)^{-\beta_i}}} = \frac{\lim\limits_{x\to 0} \beta_i {\alpha_i}^{-\beta_i} x^{-(\beta_i+1)} e^{-(\alpha_i x)^{-\beta_i}}}{\lim\limits_{x\to 0}(1-e^{-(\alpha_i x)^{-\beta_i}})}}$$

Let $\boldsymbol{y = (\alpha_i x)^{-\beta_i}}$, that is $\boldsymbol{x = \frac{y^{-\frac{1}{\beta_i}}}{\alpha}}$, then $\boldsymbol{y} \to \infty$ as $\boldsymbol{x \to 0}$, so

$$\boldsymbol{\lim_{x\to 0} h_i(x) = \frac{\beta_i {\alpha_i}^{-\beta_i} \lim\limits_{y\to\infty} y^{1+\frac{1}{\beta_i}} \lim\limits_{y\to\infty} e^{-y}}{1- \lim\limits_{y\to\infty} e^{-y}} = 0}, \text{since } \infty(\boldsymbol{0}) = \boldsymbol{0}$$

Then

$$\boldsymbol{\lim_{x\to 0} r_i(x) = \lim_{x\to 0}\left[1/(1+\frac{p_2R_2(x;\Theta_2)+p_3R_3(x;\Theta_3)}{p_1R_1(x;\Theta_1)}\right] = p_i} \ , i = 1,2,3,$$

Then (25) is proved.

And then $\boldsymbol{y \to 0}$ as $\boldsymbol{x} \to \infty$, so

$$\boldsymbol{\lim_{x\to\infty} h_i(x) = \frac{\beta_i {\alpha_i}^{-\beta_i} \lim\limits_{y\to 0} y^{1+\frac{1}{\beta_i}} \lim\limits_{y\to 0} e^{-y}}{1-\lim\limits_{y\to 0} e^{-y}} = 0},$$

by using the same assumption. It can shown that (26) is proved.

4. *Identifiability*: Chandra (1977) has proved the following: Let there be associated with each $P_i \in \Phi$ a transform ϕ_i having the domain of definition $D_{\phi_i}$and suppose that the mapping $M: P_i \to \phi_i$ is linear. Suppose also that there exists a total ordering (≤) of Φ such that

i) $P_1 \leq P_2(P_1, P_2 \in \boldsymbol{\Phi})$ implies $D_{\phi_1} \subseteq D_{\phi_2}$,

ii) For each$P_1 \in \boldsymbol{\Phi}$, there exists some t_1 in the closure of $T_1 = \{t: \phi_1(t) \neq 0\}$such that $\lim_{\substack{t\to t_1 \\ t\in T_1}}(\phi_1(t)/\phi_2(t)) = 0$, for each $P_1 \leq P_2(P_1, P_2 \in \boldsymbol{\Phi})$,

Proposition.

The class of finite mixture of the family of Inverse Weibull distribution is an identifiable.

Proof.

Let X be an IWD with the pdf and cdf respectively as

$$\boldsymbol{f(x;\Theta) = \beta {\alpha_i}^{-\beta} x^{-(\beta+1)} e^{-(\alpha x)^{-\beta}}}, \mathbf{x \geq 0}, \boldsymbol{\alpha, \beta} > 0$$

$$\boldsymbol{F(x;\Theta) = e^{-(\alpha x)^{-\beta}}}$$

Then the moment generating function of $Y = Ln(X)$ is

$$M_Y(t) = E(X^t) = \frac{1}{\alpha^t}\Gamma(1 - \frac{t}{\beta}) \text{ for } \beta < t \tag{27}$$

From (14.1), we have

$$F_1 < F_2 \text{ when } \beta_1 = \beta_2 \text{ and } \alpha_1 < \alpha_2 \tag{28}$$

$$F_1 < F_2 \text{ when } \beta_1 < \beta_2 \text{ and } \alpha_1 = \alpha_2 \tag{29}$$

Let $F_1 < F_2$ whenever $D_{M_1}(t) = (-\infty, \beta_1), D_{M_2}(t) = (-\infty, \beta_2)$ and $t_1 = \beta_1$, then from (28), (29), we have that $D_{M_1}(t) \subseteq D_{M_2}(t)$ and

$$\lim_{t_1 \to \beta_1} M_1(t) = \frac{1}{\alpha_1^{\beta_1}}\Gamma\left(1 - \frac{\beta_1}{\beta_1}\right) = \frac{1}{\alpha_1^{\beta_1}}\Gamma(0+) = \infty, \tag{30}$$

See Abramowitz and stegun (1965).

And if $\beta_1 < \beta_2$ and $\alpha_1 = \alpha_2$, then

$$\lim_{t_1 \to \beta_1} M_2(t) = \frac{1}{\alpha_1^{\beta_1}}\Gamma\left(1 - \frac{\beta_1}{\beta_2}\right) > 0, \tag{31}$$

Then from (30) and (31)we get ,

$$\lim_{t_1 \to \beta_1} \frac{M_2(t)}{M_1(t)} = 0.$$

4. CONCLUSION

In this paper we introduce the MTHIWD and discuss some important of its properties and measures. Also, we prove that the MTHIWD is identifiable.

REFERENCES

1. Abramowitz, M., Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover, New York.
2. Ahmad, K.E. and AL-Hussaini, E.K. (1982). Remarks on the non-identifiability of mixtures of distributions. *Ann. Inst. Statist. Math*, 34, 543-544.
3. AL-Hussaini, E.K. and Ahmad, K.E. (1981). On the identifiability of finite mixtures of distribution. *IEEE Trans. Inform. Theory*, 27 (5), 664-668.

4. AL-Hussaini, E.K. and Sultan, K.S. (2001). Reliability and hazard based on finite mixture models. In: Balakrishnan, N., Rao, C.R. (Eds.). *Handbook of Statistics*, vol. 20. Elsevier, Amsterdam, 139-183.
5. Balakrishnan, N. and Mohanty, N.C. (1972). On the identifiability of finite mixture of Laguerre distributions. *IEEE Trans. Inform. Theory*, 18, 514-515.
6. Calabria, R. and Pulcini, G. (1990). On the maximum likelihood and least-squares estimation in the Inverse Weibull distributions. *Statist. Appl*. 2 (1), 53-66.
7. Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distribution*. Chapman & Hall, London.
8. Jiang, R., Zuo, M.J. and Li, H., (1999). Weibull and InverseWeibull mixture models allowing negative weights. *Reliab. Eng. Syst. Safet*., 66, 227-234.
9. Jiang, R., Murthy, D.N.P. and Ji, P. (2001). Models involving two Inverse Weibull distributions. *Reliab. Eng. Sys. Safet*., 73 (1), 73-81.
10. Jiang, R., Ji, P. and Xiao, X. (2003). Aging property of unimodal failure rate models. *Reliab. Eng. Syst. Safet*., 79, 113-116.
11. Lindsay, B.G. (1995). Mixture Models: Theory, Geometry and Applications. The Institute of Mathematical Statistics, Hayward, CA.
12. Maclachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
13. Maclachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Applications to Clustering*. Marcel Dekker, New York.
14. Maclachlan, G.J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
15. Sultan, K.S., Ismail, M.A. and Al-Moisheer, A.S. (2007). Mixture of two inverse Weibull distributions: Properties and estimation. *Computational Statistics & Data Analysis,* 51, 5377-5387.
16. Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist*., 34, 1265-1269.
17. Titterington, D.M., Simth, A.F.M., Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distribution*. Wiley, Chichester.
18. Yakowitz, S.J. and Spragins, J.D. (1968). On the identifiability of finite mixtures. *Ann. Math. Statist*., 39, 209-214.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 61-70

WEIGHTED REDUCED MAJOR AXIS METHOD FOR REGRESSION MODEL

Anwar Saqr and **Shahjahan Khan**
Deptt. of Mathematics and Computing Australian, Centre for Sustainable Catchments,
University of Southern Queensland Toowoomba, Queensland, Australia
Email: saqr.anwer@yahoo.com and khans@usq.edu.au

ABSTRACT

The reduced major axis (RMA) method is widely used in many disciplines as a solution to errors in variables regression model, although it lacks efficiency. This paper provides an alternative view on the RMA estimator. Moreover, it introduces a new estimator to fit regression line when both variables are subject to measurement errors. The proposed weighted RMA (WR) estimator is derived based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. Compared to the RMA and OLS-bisector estimators the proposed WR estimator is less sensitive to the variation of the ratio of error variances (λ). The simulation results show that the WR estimator is more consistent and efficient than the RMA and OLS-bisector estimators.

Key Words: Linear regression models; Measurement error models; Reflection of points; Ratio of error variances; OLS-bisector.
2010 Mathematics Subject Classification: Primary 62J05, Secondary 62F10.

1 Introduction

The reduced major axis (RMA) method is widely used in many disciplines, and it has received much attention from the experts and some have suggested that it is more useful than other methods to deal with the measurement error model (cf Sprent and Dolby, 1980; Smith, 2009; Ludbrook, 2010).

The RMA estimator was suggested as a solution to the likelihood equations in the case of the normal functional model when there is no additional information (cf Cheng and Van Ness, 1999, p. 43). This estimator is constructed based on the geometric mean of the ordinary least squares (OLS) estimator for the regression of y on x and the reciprocal of that of x on y. Halfon (1985) and Draper and Yang (1997) pointed out that the RMA estimator minimizes the vertical and horizontal distances between the observed points and the regression line. Isobe et al. (1990) examined five different estimators, and pointed out that the OLS bisector (OLS-b) estimator is the best method to use, when there is no basis to distinguish between the explanatory and response variables.

Let x be the observable or *manifest* explanatory variable, and let ξ be the true or *latent* explanatory variable. Similarly let η be the true value of the response variable and y be the manifest response variable.

[1] On leave from Department of Statistics, Faculty of Sciences, AlJabal AlGarby University, Libya.

If the *latent* variables ξ_j and η_j are measured without error, then the simple linear regression model without ME is expressed as

$$\eta_j = \beta_0 + \beta_1 \xi_j + \epsilon_j, \quad j = 1, 2, \ldots, n, \tag{1.1}$$

where ϵ is the equation error. If there is ME in both explanatory and response variables, we can define the manifest variables as

$$x_j = \xi_j + u_j, \text{ and } y_j = \eta_j + \tau_j \qquad j = 1, 2, \ldots, n, \tag{1.2}$$

where η_j is the jth realisation of the *latent* response variable, ξ_j is the jth value of the *latent* explanatory variable, τ_j is the ME in the response variable and u_j is the ME in the explanatory variable. It is assumed that,

$$\tau_j \sim N(0, \sigma_\tau^2), \; u_j \sim N(0, \sigma_u^2), \; cov(u, \tau) = 0, \text{ and } cov(u, \epsilon) = 0. \tag{1.3}$$

Note the ME in the response variable τ_j can be absorbed in the equation error ϵ which may be expressed as $e_j = \tau_j + \epsilon_j$. The simple regression model with ME in both variables and equation error ϵ is expressed as

$$y_j = \beta_0 + \beta_1 x_j + v_j, \quad j = 1, 2, \ldots, n, \tag{1.4}$$

where $v_j = e_j - \beta_1 u_j$, and $cov(u, e) = 0$, then

$$\sigma_v^2 = \sigma_e^2 + \beta_1^2 \sigma_u^2. \tag{1.5}$$

Note that the OLS method is not valid here, because the variables x_j and v_j in equation (1.4) are not independent.

There is a common recommendation to use the RMA estimator for the ME model, but without enough justifications (Smith, 2009). Jolicoeur (1975) stated that it is difficult to interpret the meaning of the slope of the RMA regression. However, the common believe is that the RMA estimator minimizes the vertical and horizontal distances between the observed points and the fitted line (cf Halfon, 1985; Draper and Yang, 1997). But it is not quite true, because it could be demonstrated that the RMA estimator minimizes the orthogonal error of the observed points with the unfitted regression line instead of the fitted regression line.

Sections 2 and 3 provide the RMA estimator and alternative way of deriving this estimator. The proposed estimator *weighted reduced major axis estimator* (WR) is introduced in Section 3. The simulation studies are conducted to compare the performances of the proposed estimator with the RMA, OLS-b, and OLS estimators in Section 4. Some concluding remarks are included in Sections 5.

2 Reduced major axis estimator

The RMA estimator of the slope parameter is the geometric mean of the slope of y on x regression line, and the reciprocal of the slope of x on y regression line, where x and y both are random (see Leng et al. 2007). It is given by

$$\hat{\beta}_{1R} = sgn(SP_{xy}) \sqrt{SS_y SS_x^{-1}} = sgn(Sp_{xy})\, S_y S_x^{-1} \; ,$$

where $SS_x = \sum_{j=1}^{n}(x_j - \bar{x})^2$, $SS_y = \sum_{j=1}^{n}(y_j - \bar{y})^2$, $SP_{xy} = \sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$, and S_y and S_x are the standard deviations of y and x respectively.

In the literature, the RMA regression is also known as the standardized major axis (see Warton et al. 2006), and geometric mean estimator, or the line of organic correlation (cf Tessier, 1948, Kermack and Haldane, 1950, Ricker, 1973). In physics it is known as a type of standard weighting model (see Machonald and Thompson, 1992), while the astronomers call it as Strömberg's impartial line (see Feigelson and Babu, 1992).

A host of recent publications indicate that using the RMA is necessary and sufficient to fit the straight line when both the response and explanatory variables are subject to errors (see for example Levinton and Allen, 2005, Zimmerman et al, 2005, Sladek et al, 2006, and Vincent and Lailvaux, 2006). Jolicoeur (1975) and Spernt and Dolby (1980) pointed out that the RMA estimator is unbiased if and only if

$$\lambda = \beta_1^2 ,$$

where $\lambda = (\sigma_\tau^2 \sigma_u^{-2})$ is the ratio of error variances. But several other studies indicate that this assumption is unrealistic (cf Sprent and Dolby, 1980). Another competing estimator preferred by many authors is OLS-bisector (OLS-b) estimator (see e.g Isobe et al. 1990) which is given by

$$\hat{\beta}_{1OLS-b} = (\hat{\beta}_1 + \hat{\beta}_2)^{-1} \left[\hat{\beta}_1 \hat{\beta}_2 - 1 + \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)} \right] ,$$

where $\hat{\beta}_1 = S_{yx} S_x^{-2}$, and $\hat{\beta}_2 = S_y^2 S_{xy}^{-1}$.

3 Theoretical analysis

In the regression analysis with ME in both variables it is crucial to note the difference between the distance of the observed point from the fitted line, unfitted line, and unobserved point. Although, many authors use distance between the observed point and regression line without being specific about the fitted or unfitted lines. This issue is crucial when there is ME in both variables. The mathematical relationship between the vertical and orthogonal distances of the observed points and the fitted regression line is explained below. The fitted line of the true model (without equation error and ME) is given by

$$\eta_j = \beta_0 + \beta_1 \xi_j, \quad j = 1, 2, \ldots, n. \tag{3.1}$$

Let (x_j, y_j) be the observed point and (z_j, c_j) be its reflection about the fitted line (3.1), then

$$z_j = x_j \cos 2\psi + (y_j - \beta_0) \sin 2\psi, \quad \text{and } c_j = x_j \sin 2\psi - (y_j - \beta_0) \cos 2\psi + \beta_0, \tag{3.2}$$

where $\psi = tan^{-1}\beta_1$, and β_0, and β_1 are the regression parameters. For details on reflection of points please see Vaisman (1997, p. 164-169). Let the relationships between the orthogonal and vertical distance of the observed point (x_j, y_j) be explained in the context of (a) fitted line, $(\eta_j = \beta_0 + \beta_1 \xi_j)$, based on the *latent* variables and (b) unfitted line $(y_j = \beta_{0x} + \beta_{1x} x_j)$, based on the *manifest* variables.

There are potentially two orthogonal distances of any observed point, one from the *fitted line* (here represented by Υ) and the other from the *unfitted line* (Ox). In principle, the RMA method should minimise Υ, but in practice it minimises Ox. Figure 1 shows the reflection of $A = (x_j, y_j)$ about the *fitted line* $C = (z_j, c_j)$ with the orthogonal distance $\Upsilon = \overline{AB}$, and the reflection of $A = (x_j, y_j)$ about the *unfitted line* $F = (x_j^*, y_j^*)$ with the orthogonal distance $Ox = \overline{AD}$.

(a) Fitted line.

It is well known from the properties of the reflection process that the reflection line (which is the fitted line) is a bisector and orthogonal on the distance between the observed point (x_j, y_j) and its reflection point (z_j, c_j). Then the half of the square distance between the observed point (x_j, y_j) and its reflection point (z_j, c_j) will equal the orthogonal square distance (Υ_j^2) between the observed point (x_j, y_j) and the fitted line. It is given by

$$4\Upsilon_j^2 \quad = \quad (z_j - x_j)^2 + (c_j - y_j)^2. \tag{3.3}$$

Then from (3.1), (3.2), and (3.3) the square orthogonal distance (Υ_j^2) is given by

$$\Upsilon_j^2 \quad = \quad \frac{1}{4}\left[(2x_j sin^2\psi + y_j sin2\psi - \beta_0 sin2\psi)^2 + (x_j sin2\psi - 2y_j cos^2\psi + 2\beta_0 cos^2\psi)^2\ \right]$$

Since $x_j = \xi_j + u_j\ ,\ y_j = \eta_j + e_j$ and $\beta_1 = \dfrac{sin\psi}{cos\psi}$ so

$$\begin{aligned}\Upsilon_j^2 \quad &= \quad \frac{1}{4}[(-2x_j sin^2\psi + \beta_1\xi_j sin2\psi + e_j sin2\psi)^2 + (x_j sin2\psi - 2\beta_1\xi_j cos^2\psi - 2e_j cos^2\psi)^2] \\ &= \quad u_j^2 sin^2\psi - u_j e_j sin2\psi + e_j^2 cos^2\psi.\end{aligned}$$

Then $\qquad E(\Upsilon_j^2) = E(u_j^2) sin^2\psi + E(e_j^2) cos^2\psi.$

The variance of Υ_j is given by $\sigma_\Upsilon^2 = \sigma_u^2 sin^2\psi + \sigma_e^2 cos^2\psi$, where $E(z_j - x_j) = E(c_j - y) = 0$, and $\beta_1^2 = \sin^2\psi\, cos^{-2}\psi$, the variance of Υ_j becomes

$$\sigma_\Upsilon^2 \quad = \quad \left(\sigma_e^2 + \sigma_u^2 \frac{sin^2\psi}{cos^2\psi}\right) cos^2\psi = (\sigma_e^2 + \beta_1^2\sigma_u^2) cos^2\psi.$$

Then the relationship between the variance of the orthogonal distance and the variance of the vertical distance is given by formula

$$\sigma_\Upsilon^2 = \sigma_v^2 cos^2\psi = \sigma_v^2(1 + \beta_1^2)^{-1}. \tag{3.4}$$

Note that both the vertical and orthogonal distances measure the distance between the observed point (x_j, y_j) and the fitted line, but it does not measure the distance between the observed point (x_j, y_j) and the unobserved point (ξ_j, η_j). Under certain assumptions such as $\lambda = 1$ or $\beta_1 = 1$ the distance between the observed point and the unobserved point is equal to the double of the orthogonal distance, where the distance between the observed point and the unobserved point (Δ) is given by

$$\Delta^2 = (x_j - \xi_j)^2 + (y_j - \eta_j)^2 = (u_j^2 + {}_j^2),$$

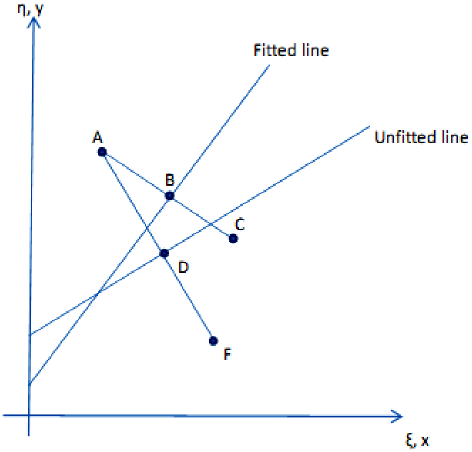


Figure 1: Graph of two orthogonal distances ($\overline{AB} = \Upsilon$, and $\overline{AD} = Ox$) between the observed point and the fitted and unfitted lines.

where u_j, and e_j are the ME in the explanatory and response variables respectively. From (1.3) the variance of the distance (Δ) is $\sigma^2_\Delta = \sigma^2_e + \sigma^2_u$. From (1.5) and if $\lambda = 1$ then $\sigma^2_\Delta = 2\sigma^2_e$.

(b) Unfitted line.

In order to find the relationship between the observed point (x_j, y_j) and the unfitted line we follow the similar procedure as in case (a) except replacing the parameters of the fitted line, $\psi = tan^{-1}\beta_1$, β_0, and β_1 by the coefficients of the unfitted line (the ME model) $\hat{\theta} = tan^{-1}\hat{\beta}_{1x}$, $\hat{\beta}_{0x}$, and $\hat{\beta}_{1x}$ respectively. Then we get

$$x_j^* \quad = \quad x_j cos2\hat{\theta} - (y_j - \hat{\beta}_{0x}) sin2\hat{\theta}, \text{ and } y_j^* = x_j sin2\hat{\theta} - (y_j - \hat{\beta}_{0x}) cos2\hat{\theta} + \hat{\beta}_{0x},$$

where (x_j^*, y_j^*) is the reflection point of the observed point (x_j, y_j) about the unfitted line.

The relationship between the sample variance of the orthogonal distance (Ox) and vertical distance (v) is given by

$$S^2_{Ox} \quad = \quad S^2_v cos^2\hat{\theta} = S^2_v(1 + \hat{\beta}^2_{1x})^{-1}. \tag{3.5}$$

Also the relationship between the observed point and unfitted population line becomes

$$\sigma^2_{Ox} \quad = \quad \sigma^2_v cos^2\theta = \sigma^2_v(1 + \beta^2_{1x})^{-1}. \tag{3.6}$$

From (3.4) and (3.6) the relationship between the variances of the orthogonal distance in cases (a) and (b) is given by

$$\sigma^2_\Upsilon \quad = \quad \sigma^2_{Ox}\ cos^2\psi cos^{-2}\theta = \sigma^2_{Ox}(1 + \beta^2_{1x})(1 + \beta^2_1)^{-1}. \tag{3.7}$$

Note that in general, $\sigma^2_\Upsilon < \sigma^2_{Ox}$, and they are equal if and only if there is no measurement error. Therefore, any method that minimizes σ^2_{Ox}, will not work well, and that is what is happening with the RMA method. The next section shows that the RMA method is minimizing σ^2_{Ox}, rather than σ^2_Υ.

From (3.5) the sum of squares of orthogonal distance (SS_{Ox}) between the observed point (x_j, y_j) and the unfitted line $(\hat{y_j} = \hat{\beta}_{0x} + \hat{\beta}_{1x}x_j)$, it can derived the RMA estimator can be derived by different procedures in order to understand its working mechanism as follows:

$$\begin{aligned} SS_{Ox} &= SS_v cos^2\hat{\theta} = \sum_{j=1}^{n}(y_j - \hat{\beta}_{0x} - \hat{\beta}_{1x}x_j)^2 cos^2\hat{\theta} \\ &= \sum_{j=1}^{n}((y_j - \bar{y}) - \hat{\beta}_{1x}(x_j - \bar{x}))^2 cos^2\hat{\theta} = \sum_{j=1}^{n}((y_j - \bar{y})cos\hat{\theta} - (x_j - \bar{x})sin\hat{\theta})^2 \quad (3.8) \end{aligned}$$

Let $Q_1 = sin\hat{\theta}$, and $Q_2 = cos\hat{\theta}$. Then

$$SS_{Ox} = \sum_{j=1}^{n}((y_j - \bar{y})Q_2 - (x_j - \bar{x})Q_1)^2.$$

Differentiating SS_{Ox} w.r.t. Q_1 and Q_2, and setting the derivatives to zero, we get

$$\begin{aligned} \frac{\partial SS_{Ox}}{\partial Q_1} &= 2\sum_{j=1}^{n}((y_j - \bar{y})Q_2 - (x_j - \bar{x})Q_1)(-(x_j - \bar{x})) = 0, \\ \frac{\partial SS_{Ox}}{\partial Q_2} &= 2\sum_{j=1}^{n}((y_j - \bar{y})Q_2 - (x_j - \bar{x})Q_1)(y_j - \bar{y}) = 0, \end{aligned}$$

or equivalently

$$Q_1 S_x^2 = Q_2 S_{yx}, \quad (3.9)$$

$$Q_2 S_y^2 = Q_1 S_{yx}. \quad (3.10)$$

From (3.9), (3.10) and $\hat{\beta}_{1x} = Q_1 Q_2^{-1}$ we get two estimators of the slope

$$\hat{\beta}_{11} = S_{yx}S_x^{-2} \text{ and } \hat{\beta}_{12} = S_y^2 S_{yx}^{-1}. \quad (3.11)$$

Then the RMA is the geometric mean of the estimators in (3.11), that is,

$$\hat{\beta}_{1RMA} = sgn\{S_{yx}\} \sqrt{S_y^2 S_x^{-2}}.$$

It is clear that the RMA estimator is derived by minimizing the orthogonal distance between the observed point (x_j, y_j) and unfitted line. Hence it does not minimizes the vertical and horizontal distances between observed points and the fitted line.

Remark: The RMA estimator is based on the minimization of the orthogonal distance between the observed points and the regression line. Since there are two regression lines, namely the fitted (latent) and unfitted (manifest) lines, it is essential to clarify the orthogonal distance between an observed point and either the fitted line or the unfitted line. Obviously the orthogonal distance of a point from the fitted line is most likely to be different from the unfitted line.

4 The weighted RMA estimator

In this section we introduce the weighted RMA (WR) estimator. The proposed estimator minimises the orthogonal distance between the observed point (x_j, y_j) and the unfitted regression line. This estimator is based on the relationship (3.5) between the vertical and orthogonal distances of the observed points and the unfitted regression line. The proposed estimator is derived as follow
Multiply equation (3.9) by S_y^2, and equation (3.10) by S_{yx}, we get

$$\begin{aligned} Q_1 S_x^2 S_y^2 &= Q_2 S_{yx} S_y^2 & (4.12) \\ Q_1 S_{yx}^2 &= Q_2 S_{yx} S_y^2, & (4.13) \end{aligned}$$

from equations (4.12) and (4.13) we get

$$\begin{aligned} Q_1(S_x^2 S_y^2 + S_{yx}^2) &= Q_2 2 S_{yx} S_y^2 & (4.14) \\ (S_x^2 S_y^2 + S_{yx}^2) sin\hat{\theta} &= 2 S_{yx} S_y^2 cos\hat{\theta}. & (4.15) \end{aligned}$$

From (4.14) and (4.15) we define an estimator

$$\hat{\beta}_{1O} = \frac{sin\hat{\theta}}{cos\hat{\theta}} = \frac{2S_{yx}S_y^2}{S_y^2 S_x^2 + S^2 yx}, \qquad (4.16)$$

which is simplified as follows

$$\hat{\beta}_{1O} \; = \; \frac{2S_y^2 S_x^{-2}}{S_y^2 S_{yx}^{-1} + S_{yx} S_x^{-2}} = \frac{2\hat{\beta}_{1R}^2}{(\hat{\beta}_1 + \hat{\beta}_2)} = \omega \;\; \hat{\beta}_{1R}, \qquad (4.17)$$

where $\omega = \hat{\beta}_{1R}\hat{\beta}_{OLS-m}^{-1}$, and $\hat{\beta}_{OLS-m}$ is the mean of the slope of the $OLS(y|x)$ and that of $OLS(x|y)$, that is, $\hat{\beta}_{OLS-m} = (\hat{\beta}_{1xy} + \hat{\beta}_{1yx})/2$. Note that the estimators $\hat{\beta}_{1O}$, $\hat{\beta}_{1RMA}$, and $\hat{\beta}_{1OLS-mean}$ are equal if $\omega = 1$, and $\hat{\beta}_{1O}$ is the estimator obtained to minimising the Ox.

It is well known that when $\sigma_{yx} > 0$ the slope of the unfitted line β_{1x} is less than the true slope β_1 of the fitted model. Then the true slope β_1 is located between $\hat{\beta}_{1O}$ and $\hat{\beta_{1x}}$ estimators. Hence the proposed WR estimator is defined as the mean of $\hat{\beta}_{1S}$ and β_{1x}, and is given by

$$\hat{\beta}_{1WR} \; = \; (\hat{\beta}_{1O} + \beta_{1x})/2.$$

Note that the WR estimator and the other estimators mentioned in this paper are used when there is no prior information on the error variances. In order to demonstrate that the performance of WR estimator is better than the RMA, OLS, and OLS-b estimators when λ is known or misspecified, we provide the results of extensive simulation studies in the next section.

5 Simulation studies

We perform extensive simulations to illustrate that the proposed WR estimator is relatively unbiased and consistent compared to the RMA, OLS, and OLS-b estimators. It is more so when λ is large. In this section we compare the WR estimator to the RMA estimator, OLS and OLS-b estimator for a wide range of values of λ $(0.1 \leq \lambda \leq 19)$. These studies demonstrate that the WR estimator is not sensitive to the ratio of error variances λ, whereas the RMA estimator grows larger as the value of λ increases. Also the WR estimator preforms consistently better than the OLS-b, and OLS estimators. The results based on $10,000$ replications of samples size $n = 100$, $\beta_0 = 0$, and $\beta_1 = 0.6$ of normal structural model, where $x \sim N(0, 100)$, are presented in the following graphs.

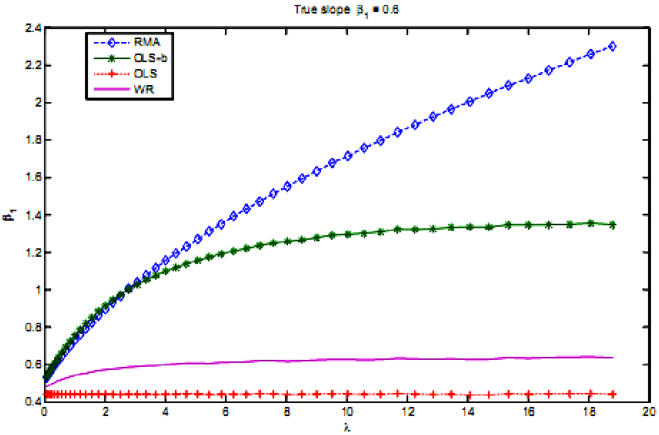


Figure 2: Plot of the mean slope of four different estimators against $0.08 \leq \lambda \leq 19$. when $\beta_0 = 0, \beta_1 = 0.6$.

It is clear from Figure 2 that the values of the OLS-bisector estimator are away from the true values of β_1. The values of the RMA estimator are far above the true values of β_1. As λ increases, the RMA estimator appears to grow large. Clearly the proposed WR estimator is much closer to the true values of β_1 than the other three estimators.

Figure 3 gives useful indications about the statistical properties of the estimators. The measurement error makes the spread of the RMA estimator the highest. While the spread of the OLS-b estimator appears to be better than that of the RMA estimator, though they are not small. The OLS estimator is consistently an under estimate of β_1. Moreover, it is inappropriate for ME models, and hence we do not compare it with the WR estimator. Sarach and Celik (2011) discussed eight different regression techniques, and pointed out that the OLS-bisector estimator is nearer to the real value than all other estimators, and the mean squares error of OLS-bisector is smaller than other estimators. The current study reveals that the WR estimator is consistently better than the OLS-b estimator in term of the closeness of $\hat{\beta}_{1WG}$ to β_1.

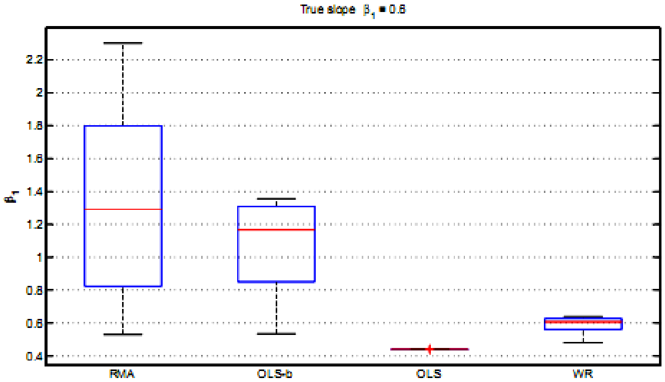


Figure 3: Graph of the distribution of the mean slope of four different estimators when $\beta_0 = 0, \beta_1 = 0.6$, and $0.08 \leq \lambda \leq 19$.

6 Concluding Remarks

This paper proposes a new estimator based on the mathematical relationship between the vertical and orthogonal distances of the observed points with fitted and unfitted lines. This estimator is appropriate for fitting straight lines when both variables are subject to measurement errors, especially when there is no basis for distinguishing between response and explanatory variables. Moreover, the WR estimator is appropriate to the normal structural model even when λ is misspecified. The graphs in Figures 2 and 3 provide clear evidence that the WR is much closer to the true slope than the other competing estimators. The values of the proposed estimator are nearer to the real value than the RMA, OLS-b, and OLS estimators. Therefore, the proposed estimator possesses better statistical proprieties than the other estimators. Moreover, the new method is stable and works well for different values of λ.

References

1. Draper, R, and Yang, Y. (1997). Generalization of the geometric mean functional relationship. Computational Statistics and Data Analysis 23, 355-372.

2. Feigelson, D, and Babu, J. (1992). Linear regression in astronomy II. Astrophys Jou 397, 5562.

3. Halfon, E. (1985). Regression method in ecotoxicology: a better formulation using the geometric mean functional regression. Notes. Environ. Sci. Technol 19, 747-749.

4. Isobe, T, Feigelson D, Akritas G, and Babu J. (1990). Linear regression in astronomy I. Astrophys. Jou. 364, 10413.

5. Jolicouer, P. (1975). Linear regressions in fishery research: some comments. Jou. Fish. Res. Board Can. 32, 14911494.

6. Kermack, A, and Haldane, S. (1950). Organic correlation and allometry. Biometrika 37, 3041.
7. Leng, L, Zhang, L, Kleinman, and Zhu, W. (2007). Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. Journal of Physics Conference Series 78, 012084-012088.
8. Levinton, S, and Allen, J. (2005). The paradox of the weakening combatant: trade-off between closing force and gripping speed in a sexually selected combat structure. Funct Ecol 19, 159165.
9. Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: But which regression?. Clinical and Experimental Pharmacology and Physiology 37, 692699.
10. Macdonald, R, and Thompson, J. (1992). Least-squares fitting when both variables contain errors: pitfalls and possibilities. Am Jou Physiol 60, 6673.
11. Ricker, E. (1973). Linear regressions in Fishery research. Jou Fish. Res. Board Can 30, 409-434.
12. Sladek, V, Berner, M, and Sailer, R. (2006). Mobility in central european late neolithic and early bronze age: femoral cross-sectional geometry. Am Jou Phys Anthropol 130, 320332.
13. Smith, J. (2009). Use and Misuse of the Reduced Major Axis for Line-Fitting. American Journal Of Physical Anthropology 140, 476486.
14. Sprent, P, and Dolby, R. (1980). Query: the geometric mean functional relationship. Biometrics 36, 547-550.
15. Teissier, G. (1948). La relation d'allometrie sa signification statistique et biologique. Biometrics 4, 14-53.
16. Vaisman, I. (1997). Analytical Geometry. World Scientific. Singapore.
17. Vincent, E, and Lailvaux, P. (2006). Female morphology, web design, and the potential for multiple mating in Nephila clavipes: do fat-bottomed girls make the spider world go round. Biol Jou Linn Soc 87, 95102.
18. Warton, I, Wright, J, Falster S, and Westoby M. (2006). Bivariate line-fitting methods for allometry. Biol Rev 81, 259291.
19. Zimmerman, F, Breitenmoser-Wu rsten, C, and Breitenmoser, U. (2005). Natal dispersal of Eurasian lynx (Lynx lynx) in Switzerland. Jou Zool 267, 381395.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 71-80

SLOPE ESTIMATOR FOR THE LINEAR ERROR-IN-VARIABLES MODEL

Anwar Saqr and **Shahjahan Khan**
Deptt. of Mathematics and Computing Australian, Centre for Sustainable Catchments,
University of Southern Queensland Toowoomba, Queensland, Australia
Email: saqr.anwer@yahoo.com and khans@usq.edu.au

ABSTRACT

It is well known that in the presence of errors-in-variable the ordinary least squares (OLS) estimator of the parameters of the regression model is inappropriate. This is true even if the ratio of error variances (λ) is known. Wald's grouping method could deal with such problem but it lacks efficiency and is subject to identifiability problem. The main aim of the paper is to introduce a reflection based grouping method to improve the efficiency of the Wald's estimator under flexible assumption on λ. We compare the relative performance of the proposed reflection grouping (RG) estimator with the OLS, ML, Wald's and Geary's estimators by simulation studies under various conditions. The simulation results show that the RG estimator is more consistent and efficient than the other estimators.

Key Words: Linear regression models, Measurement error, Reflection of points; Ratio of error variances; Method of cumulants; Instrumental variable, and method of moments.
2010 Mathematics Subject Classification: Primary 62J05, Secondary 62F10.

1. Introduction

The error-in-variables problem in the simple linear regression model is also known as the measurement error (ME) problem. The ME poses a serious problem, as it directly impacts on estimators and their standard error (see Fuller, 2006, p. 3). In practice, it is rare to measure the variables precisely, for example, medical variables such as blood pressure and blood chemistries, agricultural variables such as soil nitrogen and rainfall etc.

In the conventional notation, let x be the true explanatory variable which is called the *latent* variable. Let m be the observable, or *manifest* explanatory variable. Similarly let η be the true value of the response and y be the associated manifest variable.

If the *latent* variables x_j and η_j are measured without error then their linear relationship is expressed as

$$\eta_j = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \ldots, n. \tag{1.1}$$

If there is ME in both response and explanatory variables, the actual observed values, m and y are not the true values, and we define

$$m_j = x_j + u_j, \quad y_j = \eta_j + e_j \quad j = 1, 2, \ldots, n. \tag{1.2}$$

[1] On leave from Department of Statistics, Faculty of Sciences, AlJabal AlGarby University, Libya.

where η_j is the jth realisation of the *latent* response variable, x_j is the jth value of the *latent* explanatory variable, e_j is the ME in the response variable and u_j is the ME in the explanatory variable. It is assumed that, $e_j \sim N(0, \sigma_{ee})$, and $u_j \sim N(0, \sigma_{uu})$. The model with the fixed x is called the *functional model*, whereas, the model with independent identically distributed random variable x is called *structural model*. The simple regression model with ME in both variables and without equation error is expressed as

$$y_j = \beta_0 + \beta_1 m_j + v_j, \quad j = 1, 2, \ldots, n, \tag{1.3}$$

where $v_j = e_j - \beta_1 u_j$. Note in equation (1.3) m_j and v_j are not independent, and hence least squares method is not valid for the above model.

Wald (1940) considered the problem of error in both explanatory and response variables. He proposed an estimation method based on dividing the observations of both the response and explanatory variables into two groups, G1 and G2. The G1 contains the first half of the ordered observations and G2 contains the second half. He showed that the slope of the line joining the group means provided consistent estimator for the slope parameter of the simple linear regression model. Properties of this estimator can be found in Gupta and Amanullah (1970). Similarly, Bartlett (1949) developed another grouping method where the available observations were divided into three, instead of two, groups. Gibson and Jowett (1957) investigated optimum ways of grouping the observations, and offered advice on how to place the data into these three groups to obtain the most efficient estimate of the slope. Another approach to deal with ME the instrumental variable (IV). Fundamentally, the IV method involves finding a variable z_j that is correlated with the manifest explanatory variable m_j, but is uncorrelated with the random error component, u_j. It is difficult to obtain a good IV which meets the above criteria. The disadvantage of Wald's method is the loss of efficiency (cf. Theil and Yzeren, 1956). The purpose of the present paper is to increase the efficiency of Wald's method using a different measure to find the instrumental variable.

2. Existing estimation methods

In this section we introduce estimators based on the principle of grouping of observations, maximum likelihood, and cumulants.

2.1. Grouping method

In 1940 Wald pointed out that a consistent estimator of β_1 may be calculated if the following assumptions are met:

1. The random variables $e_1, \ldots, e_n$ have the same distribution and they are uncorrelated, that is, $E(e_i e_j) = 0$ for $i \neq j$. The variance of e_j is finite.
2. The random variables $u_1, \ldots, u_n$ have the same distribution and they are uncorrelated, that is, $E(u_i u_j) = 0$ for $i \neq j$. The variance of u_j is finite.
3. The random variables e_j and u_j are uncorrelated, that is, $E(e_j u_j) = 0$.
4. The limit inferior of $\left|\{\sum_{j=1}^{k} x_j - \sum_{j=k+1}^{n} x_j\}/n\right| > 0$, where n is even $(n = 2, 4, 6, \ldots, \infty)$, and $k = \frac{n}{2}$.

Then divide the observations into two groups based on the ranks of the manifest explanatory variable m_j, those above the median of m_j into one group G_1 and those below the median into another group G_2. Wald considers the expression

$$a_1 = \frac{(m_1 + \ldots + m_k) - (m_{k+1} + \ldots + m_n)}{n},\ b_1 = \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{n}.$$

Then Wald's estimators of β_1 and β_0 are given by

$$\begin{aligned} \hat{\beta}_1 &= \frac{a_1}{b_1} = \frac{(y_1 + \ldots + y_k) - (y_{k+1} + \ldots + y_n)}{(m_1 + \ldots + m_k) - (m_{k+1} + \ldots + m_n)} = \frac{\bar{y}_2 - \bar{y}_1}{\bar{m}_2 - \bar{m}_1} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{m}, \end{aligned}$$

where $\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$, and $\bar{m} = \frac{\sum_{j=1}^{n} m_j}{n}$.

Johnston (1972, p. 284) showed how Wald's grouping method is based an instrumental variable. If the number of sample observations is even then define a z matrix as

$$z' = \left[\begin{array}{ccccc} 1 & 1 & 1 & \ldots & 1 \\ -1 & -1 & 1 & \ldots & -1 \end{array} \right],$$

where the second row included minus or plus one according to the value of the manifest explanatory variable m_j is below or above the median of m_j. If we rewrite the model (2.1) in matrix for m as

$$\eta = x\beta,$$

where $x' = \left[\begin{array}{ccccc} 1 & 1 & 1 & \ldots & 1 \\ x_1 & x_2 & x_3 & \ldots & x_n \end{array} \right]$, and $\beta = (\beta_0, \beta_1)'$, then the instrumental variable estimator of β is defined by

$$\hat{\beta} = (z'x)^{-1} z'\eta = \left[\begin{array}{cc} n & 0 \\ 0 & \frac{n}{2}(\bar{x}_2 - \bar{x}_1) \end{array} \right]^{-1} \left[\begin{array}{c} n\bar{\eta} \\ \frac{n}{2}(\bar{\eta}_2 - \bar{\eta}_1) \end{array} \right] = \left[\begin{array}{c} \bar{\eta} \\ \frac{\bar{\eta}_2 - \bar{\eta}_1}{\bar{x}_2 - \bar{x}_1} \end{array} \right].$$

The estimator of the slope is

$$\hat{\beta}_1 = \frac{\bar{\eta}_2 - \bar{\eta}_1}{\bar{x}_2 - \bar{x}_1}.$$

According to Johnston (1972, p. 284) $\bar{\eta}$ is the estimator of $\beta_0 + \beta_1 E(x)$, and so

$$\hat{\beta}_0 = \bar{\eta} - \hat{\beta}_1 \bar{x}.$$

It is suggested that one should omit the central observation before computations if n is odd.

Wald countered the problem of consistency, the groups are not independent of the error terms if they did not base on the order of the true values. He proved that the grouping by the observed values is the same as grouping with respect to the true values. However, there are some criticisms in literature about Wald estimator, but these criticisms lacked consensus. Neyman and Scott (1951) pointed out that the Wald estimator is consistent for β_1 in the structural relation situation if and only if

$$Pr[m_{p_1} - e < x \leq m_{p_1} - \mu] = Pr[m_{1-p_2} - e < x < m_{p_1} - \mu] = 0,$$

where m_{p_1} and m_{1-p_2} are the p_1 and $(1-p_2)$ percentile points of $F(m)$, the distribution of m, and $e-\mu$ is the range of u.

This condition means that we must know the range of the error in m, and in order to satisfy the condition the range should be finite, otherwise the condition becomes $Pr[-\infty < x < \infty] = 0$ which is never satisfied. Often relies on the central limit theorem and assumes that u is normally distributed, where it has an infinite range, then the above condition be unsatisfied when the errors u_j are normally distributed (see Madansky, 1959). Wald estimator is consistent under very specific conditions except that the errors are not normally distributed (cf. Gupta and Amanullah, 1970). While Pakes (1982) claimed that the work of Gupta and Amanullah (1970) is needless, where the Wald's estimator is inconsistent. However, according to Theil and Yzeren (1956) the Wald's method is valuable, though there is loss of efficiency. Johnston (1972, p. 284) stated 'Under fairly general conditions the Wald estimator is consistent but likely to have a large sampling variance'. Moreover, Fuller (2006, p. 74) mentioned that the Wald's method was often interpreted improperly. In fact, there are many discussions on improving the efficiency of the grouping method by dividing the observations to more than two groups and groups of unequal size (see Nair and Banerjee, 1942, Bartlett, 1949, Dorff and Gurland, 1961, and Ware, 1972). Under the normality assumption the grouping estimator is the maximum likelihood estimator (see Chang and Huang, 1997). In practice, the grouping method is still important, and the grouping estimator is the maximum likelihood estimator under the normality assumption (Chang and Huang 1997, Cheng and Van Ness, 1999, p. 130).

2.2. Maximum likelihood method

The likelihood method can have one or more solutions, and might be a saddle point, a local maximum, or a local minimum of the likelihood function. Lindley (1947) mentioned that the likelihood equations are consistent if the ratio of error variances λ is known. The ML estimator of β_1 is given by

$$\hat{\beta}_{ML} = \left[(S_y^2 - \lambda S_m^2) + \sqrt{(S_y^2 - \lambda S_m^2)^2 + 4\lambda S_{ym}^2}\ \right] / 2S_{ym}.$$

The most common criticisms of this method is the misspecification of λ. This method deals only with models that do not include an equation error which means that all data should fall exactly on a straight line, which is rare to happen in practice.

2.3. Cumulants method

This method is closely related to the method of higher-order moments, and both methods lead to similar estimators. Geary (1949) wrote a series of papers on the method of moments. He introduced a treatment for the ME model under assumptions that the latent variable x is not normally distributed and all moments exist. Geary's estimators (G_a, G_b) or cumulant method estimators do not work if (x_j, η_j) are jointly normally distributed, because all cumulants of order ≥ 3 are zero in normal systems. The Geary's estimators are given by

$$\hat{\beta}_{1G_a} = \frac{k(1,3)}{k(2,2)}, \quad \text{and } \hat{\beta}_{1G_b} = \frac{k(2,2)}{k(3,1)},$$

where $k(\cdot,\cdot)$ represents an appropriate cumulant (see Fuller 2006, p. 72, for details). The cumulant estimates deals only with the non-normal structural model (cf Cheng and Van Ness, 1999, p. 127).

3. Proposed reflection method

The reflection grouping (RG) method is constructed based on the ranks of a new variable d_j which is located at the middle of the manifest explanatory variable m_j and its *reflection* m_j^*. The difference from Wald's original method is to use the ranks of the transformed variable d_j to dividing the observations into two groups instead of using the ranks of the manifest explanatory variable. Moreover, assume the additional assumption that the ratio of error variances $\lambda = \sigma_{ee}\sigma_{uu}^{-1}$ is known as $\lambda < 1$, $\lambda = 1$, or $\lambda > 1$. To avoid the unwanted and troublesome influence of the ME in the explanatory variable, the idea of *reflection* of the manifest variable is used for all the values of explanatory variable. The *reflection* of the points is taken about the fitted regression line of the manifest variables. This is essentially done by a transformation of the observed values of the explanatory variable to their *reflection* on the Euclidean plane. In the conventional notation, the *reflection* of the explanatory variable $m_j = x_j + u_j$ (with ME u_j) for $j = 1, 2, \ldots, n$, can be defined as

$$m_j^* = m_j \cos 2\psi + (y_j - \hat{\beta}_{0m}) \sin 2\psi, \tag{3.1}$$

where $\hat{\beta}_{0m}$ is the least square estimate of the intercept parameter, ψ is the angle measure defined as $\psi = \arctan \hat{\beta}_{1m}$ in which $\hat{\beta}_{1m}$ is the least square estimate of the slope parameter in the manifest model, and cos, and sin are the usual trigonometric cosine and sine functions respectively. For the definition of *reflection* of points on the Euclidean plane (see Vaisman, 1997, p. 164-169).

The general idea of using the reflection is that the true value of the latent explanatory variable x is located at the middle of the manifest variable m and its reflection m^*, if the ratio of error variances $\lambda = 1$. The reflection group estimator takes different form depending on the value of λ. We consider the following three cases for $\lambda = 1$, $\lambda < 1$, and $\lambda > 1$. Therefore we suggest a new variable $d_j = (m_j + m_j^*)/2$ when $\lambda = 1$, but if $\lambda < 1$ we use another variable $d_{1j} = (d_j + m_j^*)/2$, and $d_{2j} = (d_j + m_j)/2$ if $\lambda > 1$. The main advantage of using the proposed variable d_j for grouping is to reduce the sum squares of error. We propose modifications to the Wald's method to introduce three different forms of the reflection grouping estimator for the slope β_1 for varying values of λ.

(a) **When $\lambda = 1$**

Let the instrumental variable T_1 be based on the ranks of the variable $d_j = (m_j + m_j^*)/2$. The entries in the second row of T_1' is -1 if the value of d_j is less then the median of $d_j's$, and $+1$ otherwise. A typical representation of T_1' is

$$T_1' = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ -1 & -1 & 1 & \ldots & -1 \end{bmatrix}.$$

The RG estimator of β_1 and β_0 is given by

$$\hat{\beta}_{RG1} = (T_1'm)^{-1}T_1'y = \begin{bmatrix} n & 0 \\ 0 & \frac{n}{2}(\bar{m}_{12} - \bar{m}_{11}) \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \frac{n}{2}(\bar{y}_{12} - \bar{y}_{11}) \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\bar{y}_{12} - \bar{y}_{11}}{\bar{m}_{12} - \bar{m}_{11}} \end{bmatrix}, \text{and}$$

$$\hat{\beta}_{1RG1} = \frac{\bar{y}_{12} - \bar{y}_{11}}{\bar{m}_{12} - \bar{m}_{11}}, \quad \hat{\beta}_{0RG1} = \bar{y} - \hat{\beta}_{1RG1}\bar{m}$$

(b) **When** $\lambda < 1$

Similarly, let the instrumental variable T_2 be based on the ranks of the variable $d_{j2} = (d_j + m_j)/2$. The entries in the second row of T_2' is -1 if the value of d_{j2} is less then the median of d_{j2}, and $+1$ otherwise. The RG estimator of β_1 and β_0 becomes

$$\hat{\beta}_{RG2} = (T_2'm)^{-1}T_2'y = \begin{bmatrix} n & 0 \\ 0 & \frac{n}{2}(\bar{m}_{22} - \bar{m}_{21}) \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \frac{n}{2}(\bar{y}_{22} - \bar{y}_{21}) \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\bar{y}_{22} - \bar{y}_{21}}{\bar{m}_{22} - \bar{m}_{21}} \end{bmatrix}, \text{and}$$

$$\hat{\beta}_{1RG2} = \frac{\bar{y}_{22} - \bar{y}_{21}}{\bar{m}_{22} - \bar{m}_{21}}, \quad \hat{\beta}_{0RG2} = \bar{y} - \hat{\beta}_{1RG2}\bar{m}$$

(c) **When** $\lambda > 1$

Finally, let the instrumental variable T_3 be based on the ranks of the variable $d_{j1} = (z_j + m_j^*)/2$. The entries in the second row of T_3' is -1 if the value of d_{j1} is less then the median of d_{j1}, and $+1$ otherwise. The RG estimator of β_0 and β_1 becomes

$$\hat{\beta} = (T_3'm)^{-1}T_3'y = \begin{bmatrix} n & 0 \\ 0 & \frac{n}{2}(\bar{m}_{32} - \bar{m}_{31}) \end{bmatrix}^{-1} \begin{bmatrix} n\bar{y} \\ \frac{n}{2}(\bar{y}_{32} - \bar{y}_{31}) \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \frac{\bar{y}_{32} - \bar{y}_{31}}{\bar{m}_{32} - \bar{m}_{31}} \end{bmatrix}, \text{and}$$

$$\hat{\beta}_{1RG3} = \frac{\bar{y}_{32} - \bar{y}_{31}}{\bar{m}_{32} - \bar{m}_{31}}, \quad \hat{\beta}_{0RG3} = \bar{y} - \hat{\beta}_{1RG3}\bar{m}$$

We should omit the central observation before computations if n is odd. Although the second row of T_1', T_2', T_3' (that is the sequence of -1 and $+1$) appears to be similar, they will be different when the method is applied to any real data set.

The aims of the proposed RG method are to deal with the situation of misspecification λ which makes the maximum likelihood estimator biased, and increasing the efficiency of the Wald's method.

4. Simulation studies

We perform large scale simulation study to compare between the proposed (RG) estimator and other existing estimators for both normal model when $0.1 < \lambda < 9$, and non-normal model when $\lambda = 1$, < 1, and > 1.

4.1 Normal structural model

This study compares between the proposed RG estimator, ML, and OLS estimators of the slope parameter β_1 of the normal structural model for sample size $n = 40$, and when λ is correct and misspecified. The data is based on $10,000$ replications, $x \sim N(0, 49)$, and the parameters settings are $\beta_1 = 1$, $\beta_0 = 0$, when $0.1 < \lambda < 9$.
It is clear from Figure 1 that the ML estimator does not work well when λ is misspecified.

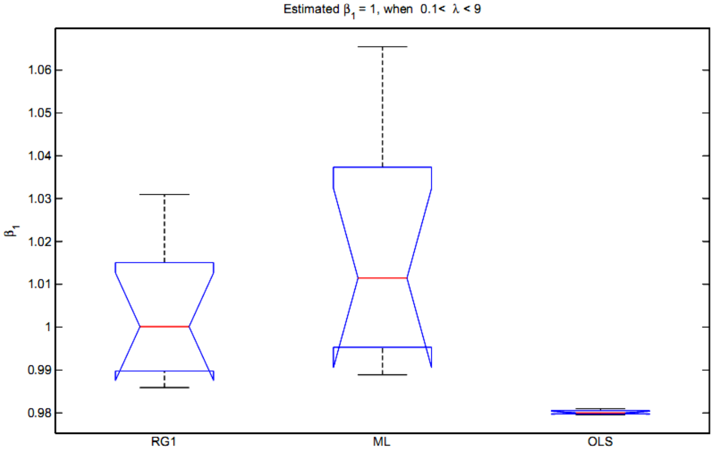


Figure 1: Graph of the estimated slope for three different estimators when $0.1 < \lambda < 9$.

That means without selecting the correct value of λ the ML method overestimates the true slope. The proposed method is better than both the ME and OLS estimators when the ratio of error variances λ is incorrect.

4.2 Non normal structural model

This study provides comparison between the proposed RG estimator and Wald's (W), Geary's (G), and OLS estimators of the slope parameter β_1 of the non-normal structural model for different sample sizes, and when $\lambda = 1$, < 1, and > 1. Also it provides the comparison in terms of the mean absolute error (MAE) of these estimators.

(a) **When** $\lambda = 1$ (d_j is used for $\hat{\beta}_{RG1}$). The data is based on $100,000$ replications, $x \sim$ uniform on $[-5, 5]$, $u \sim N(0, 1)$, $e \sim N(0, 1)$, and the parameters settings are $\beta_1 = 1$, $\beta_0 = 0$ (see Fig. 2).

(b) **When** $\lambda < 1$ (d_{1j} is used for $\hat{\beta}_{RG2}$). The data is based on $100,000$ replications, $x \sim$ uniform on $[-5, 5]$, $u \sim N(0, 1)$, $e \sim N(0, 2.25)$, and the parameters settings are $\beta_1 = 1$, $\beta_0 = 0$ (see Fig. 3).

(c) **When** $\lambda > 1$ (d_{2j} is used for $\hat{\beta}_{RG3}$). The data is based on $100,000$ replications, $x \sim$ uniform on $[-5, 5]$, $u \sim N(0, 2.25)$, $e \sim N(0, 1)$, and the parameters settings are $\beta_1 = 1$, $\beta_0 = 0$ (see Fig. 4).

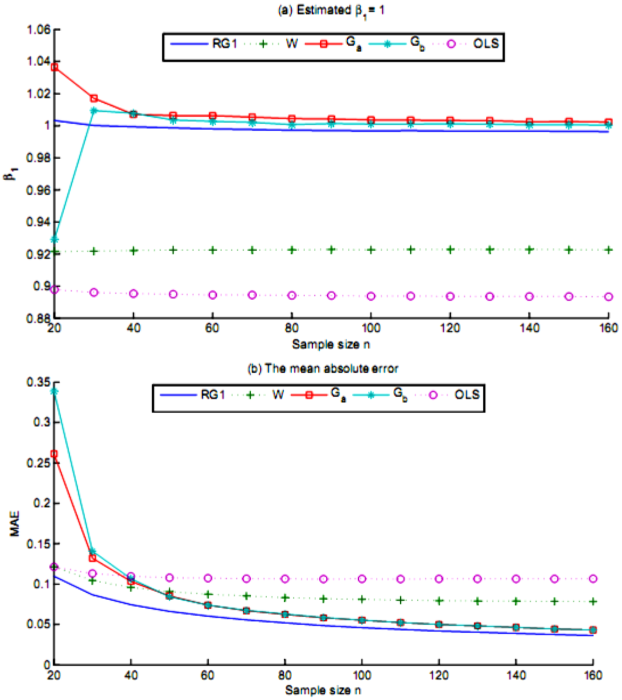


Figure 2: Graph of the estimated slope for five different estimators when $\lambda = 1$.

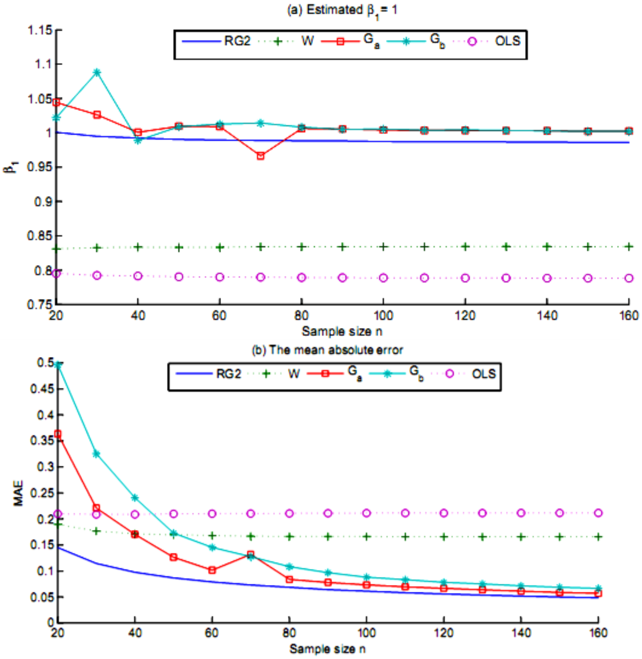


Figure 3: Graph of the estimated slope and MAE for five different estimators when $\lambda < 1$.

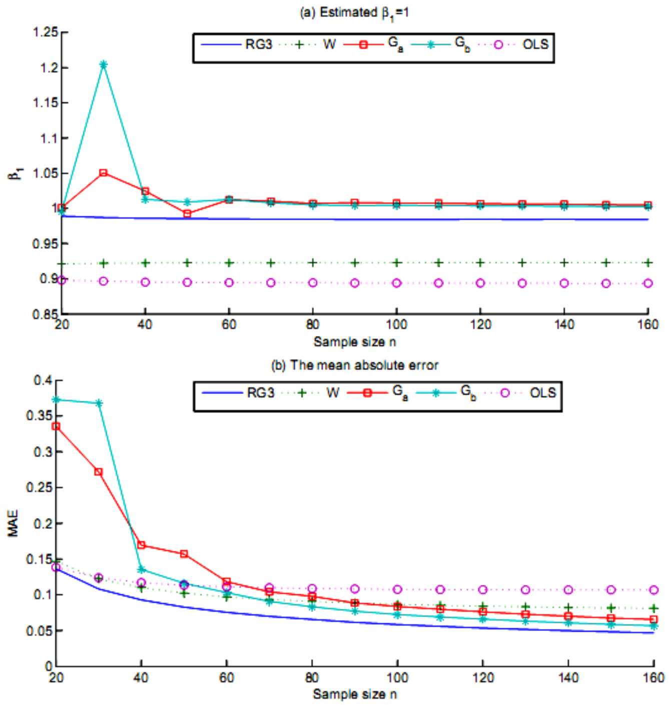


Figure 4: Graph of the estimated slope and MAE for five different estimators when $\lambda < 1$.

In Figures 2a, 3a, and 4a the values of Wald's estimator are away from the true value of β_1, but they appear to be slightly better than those of OLS. The values of the OLS are the lowest and far below the true value of $\beta_1 = 1$. Obviously, the proposed estimator are much closer to the true value of β_1 than the Geary's estimators. Figures 2a, 3a, and 4a show that if the sample size is large then the two estimators of Geary are close to the true value of the slope, but they fluctuate significantly if n is small.

Figures 2b, 3b, and 4b reveal that the MAE of the OLS estimator is the highest. The MAE of Geary's estimators or cumulant method estimators are better and smaller than that of Wald's and OLS. Whereas the MAE of the Wald's estimator appears to be better than that of the OLS, though it is not small. Clearly, the MAE of the RG estimator is better and the smallest compared to the other estimators.

5. Concluding Remarks

The proposed RG method is constructed based on a specific modifications to two grouping method. It works under a fixable and realistic assumption that the ratio of error variances equal, less, or grater than one. The proposed method is straightforward, easy to implement, and handles the ME in both normal and non-normal structural models. The simulated results show that the proposed RG method is appropriate to the normal structural model more than both the ML, and OLS methods even when λ is misspecified. Based on the forgoing discussion it is evident that the proposed RG method significantly increases the efficiency of Wald's grouping method. Moreover, it performs better than the ML method when λ is misspecified, and sample size is small.

References

1. Bartlett, M S. (1949) Fitting a stright line when both variables are subject to error. *Biometrics*, 5, 207-212.
2. Chang, Y. P. and Huang, W. T. (1997) Inferences for the linear errors-in-variables with changepoint model. *Journal of the American Statistical Association*, 92, 171-178.
3. Cheng, C L, and Van Ness, J W. (1999) *Kendall's Library Of Statistics 6, Statistical Regression With Measurement Error.* Wiley, New York.
4. Dorff, M. and Gurland, J. (1961) Small sample behavior of slope estimates in a linear functional relation. *Biometrics*, 17, 283-298.
5. Fuller, W A. (2006) *Measurement Error Models.* Wiley, New Jersey.
6. Geary, C. (1949). Sampling aspects of the problem from the error-in-variable approach. Econometrica, 17, 26-28.
7. Gibson, W M and Jowett, G H. (1957) Three-group regression analysis. Part 1: Simple regression analysis. *Appplied Statistics*, 6, 114-122.
8. Gupta, Y. P., and Amanullah. (1970) A note on the moments of the Wald's estimator. *Statistica Neerlandica*, 24, 109-123.
9. Johnson, J. (1972) *Econometric Methods.* McGraw Hill Book Company, New York.
10. Kendall, M. G., and Stuart, A. (1961) *The Advanced Theory of Statistics Volume Two.* Charles Grin and Co Ltd, London.
11. Lindley, D. V (1947). Regression lines and the linear functional relationship. Suppl.J. Roy. Statist. Soc. 9, 218-244.
12. Madansky, A. (1959) The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association*, 54, 173-205.
13. Nair, K. R., and Banerjee, K. S. (1942). A note on fitting of straight lines if both variables are subject to error. *Sankhya*, 6, 331.
14. Neyman, J. and Scott, E. L. (1951) On certain methods of estimating the linear structural relation. *Annals of Mathematical Statistics*, 22, 352-361.
15. Pakes, A. (1982) On the asymptotic bias of the Wald-type estimators of a straight line when both variables are subject to error. *International Economic Review*, 23, 491-497.
16. Theil, H. and J. Van Yzeren. (1956) The efficiency of Walds method of fitting straight lines. *Revue de Institut Internationale de Statistique*, 24, 17-26.
17. Vaisman, I. (1997) *Analytical Geormetry.* World Scientific, Singapore.
18. Wald, A. (1940) Fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics*, 11, 284-300.
19. Ware, J. H. (1972) The Fitting of Straight Lines When Both Variables are Subject to Error and the Ranks of the Means are Known. *Journal of the American Statistical Association*, 67, 891-897.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 81-88

SHADOW LEARNING IN UNDERGRADUATE MATHEMATICS: AN EXPLORATORY STUDY FROM UAE

Hana Sulieman
Department of Mathematics and Statistics, American University of Sharjah,
Sharjah, United Arab Emirates. Email: hsulieman@aus.edu

ABSTRACT

Purpose – This study aims to explore the practice of shadow education (private tutoring) in Mathematics among first year university students and to examine how this mode of learning during high school years impact students' performance in freshman mathematics course.

Methodology – Taking a case study of Business Mathematics course at the American University of Sharjah (AUS) in United Arab Emirates (UAE), the author undertook a quantitative analysis of students' final grades in the course categorized by students' responses to some private tutoring related questionnaire items.

Findings – Students who have received private tutoring in Mathematics in the form of individualized instruction for at least one year in high school are more likely to request tutoring in the subject during freshman year at the university. As for performance, students who purchased private tutoring in order to improve their performance in the course consistently underperformed their peers irrespective of whether they had tutoring experience in Mathematics in high school or not. They showed significantly higher failing rate in the course than their peers.

Research limitations – The study did not include other tutoring unrelated factors that could affect students' performance in the course such as students' motivation, household and/or school characteristics.

Implications for teaching – Given the observed negative consequences of private tutoring on students' learning, educators of higher education where tutoring practices may exist should invest time in discouraging these practices and educating their students of the negative effects tutoring can have. They should develop teaching strategies that promote independent learning, give more support to challenged students in Mathematics and encourage peer tutoring and group studying.

KEYWORDS

Tutoring; Shadow Education; UAE.

1. INTRODUCTION

Private tutoring or so-called "Shadow Education" is a growing industry in both developed and developing countries. In some countries in East Asia, private tutoring accounts for close to 3% of Gross Domestic Product(Dang and Rogers, 2008). Private supplementary tutoring in academic subjects outside school hours for remuneration is described as shadow because it mirrors the mainstream education system in its scope;

intensity, size and orientation, and public attention in almost all societies are focused more on mainstream than on its shadow (Bray, 2006). Private tutoring is not a new phenomenon; some reference to it exists principally in the nineteenth century publications. In recent years, it has been growing steadily as a parallel education sector to the mainstream education system and with its growth, there exist an increasing body of literature examining its scope, causes and impacts on learning and livelihood.

While private tutoring may have many positive effects, such as increasing human capital, providing constructive after-school activities for students, and generating additional income for tutors (often under-paid teachers), it also produces a number of negative effects. For example, private tutoring can lower the quality of education provided by the school system as teachers, who often tutor their own students, tend to focus their efforts on private tutoring rather than on their classes, put pressure on students, exacerbate social inequities, and facilitate the spread of corruption in the education system (Dang and Rogers, 2008; Bray, 2007). In addition, the literature presents no conclusive evidence that private tutoring increases aggregate student achievement at the national level. On the contrary, some argues that it can actually lower the quality of teaching and learning in the classroom if teachers focus their efforts on private tutoring rather than on their classes. Not to forget the leisure time of children contravened by too many hours of studying and may negatively affect their mental, social and physical well-being.

In this paper, we try to shed some light on the practice of private tutoring among first year university students taking a Business Mathematics course at the American University of Sharjah, United Arab Emirates (UAE). The impact of this practice on students' learning measured by their final grades in the course is examined.

2. PRIVATE TUTORING IN THE UAE

Similar to the situation in many countries around the world, the use of private tutoring in the United Arab Emirates (UAE) is increasingly wide-spreading. The UAE is classified as a high income developing economy by the International Monetary Fund (http://www.imf.org). Its population leapt over the 8 million mark in 2011, most of whom (about 88%) are expatriates who come from Asia, Arab countries and Western and European countries and reside in the country for employment purposes.

A substantial number of high school students in both public and private education schools in UAE opt for private tuitions and experts have attributed this to a peer pressure, high competition in university admission, weakness in school education system and a trend brought on by expatriate teachers from countries where this practice is common. Numerous reports have appeared in the country's media illustrating some of the negative impacts of private tutoring on students and their families. According to the Abu Dhabi Department for Economic Development, 27% of Emirati families spend on average 1,436 AED (390 USD) per month on private tuitions which works out to 4.8% of their household expenditure (Ahmed, August 24, 2010).

In a recent study conducted by the Dubai School of Government (Farah, 2011) more than 65% of Emirati students attend private tutoring lessons in the final year of high school. This figure is significantly higher than the 51% reported in 2009. 53% of the surveyed students reported participation in private tutoring in one or more earlier grades. The study found that boys are significantly more likely to take private tutoring than girls. Among all subjects tutored, mathematics and physics were the most popular across genders with boys more likely than girls to receive tutoring in mathematics. Over 80% of the participants who received private tutoring claimed that it had some positive impact upon their studies while about 85% of them agreed to the statement that taking private tutoring was the only way to graduate from school or get good education. As for the private tutors, the study reported that 82.5% of tutors were men and 52.5% of them found tutoring their classroom students. The practice of tutoring their own students has been shown in the literature to lead to favoritism and other forms of corruption in the education system.

In this exploratory study we try to shed some light on the private tutoring practices among freshman university students. The sample includes 152 students enrolled in first year Business Mathematics course at the American University of Sharjah, UAE, in Fall 2010. Nearly 90% of these students were freshmen students while the rest were at least sophomore students who were repeating the course for second or third time. The course had six sections taught be three different instructors. At the end of the semester, the students were asked about the study strategies and learning aids they utilized outside classroom hours to help them improve performance in the course. These study strategies were classified as: studying independently the course material which includes lecture notes, textbook and other supplementary material provided by instructor and not requiring the help of peer of tutor; seeking additional help from peers or/and instructor and seeking additional help from a private tutor. The students were also asked about whether they had private tutoring experience in Mathematics in high school or not and how they perceive the effectiveness of private tutoring in helping them perform well in Mathematics. Their responses were then linked to their final grades in the course so as to provide quantitative basis for the analysis of the results.

3. RESULTS AND DISCUSSION

To eliminate the instructor effect on the analysis results, the grades (100-point scale) were first standardized by subtracting the corresponding average score and dividing by standard deviation. Multiple t-tests were used to compare the standardized grades for different groups of students based on their responses to private tutoring related questions. 5% level of significance is used as a nominal value to declare significance.

Of the students surveyed, 24% were independent learners who studies course material on their own without the help of their peers or private tutor. 66% of them reported that, in addition to studying on their own, they utilized the help of their peers and some from the instructor during office hours and extra hours requested by appointment. One form of peer help is obtained from the Mathematics Learning Center which offers free-of-charge, drop-in and one-to-one tutoring service in all mathematics courses ranging from developmental to 200-level courses. The tutors in the center are qualified senior Mathematics majors who have passed certain selection criterion and attended a short training course on tutoring. 10% of the students acquired the additional help of private tutor who tutored them regularly (3%) or when needed (before exams) (7%). The form of tutoring received is one-to-one individualized instruction.

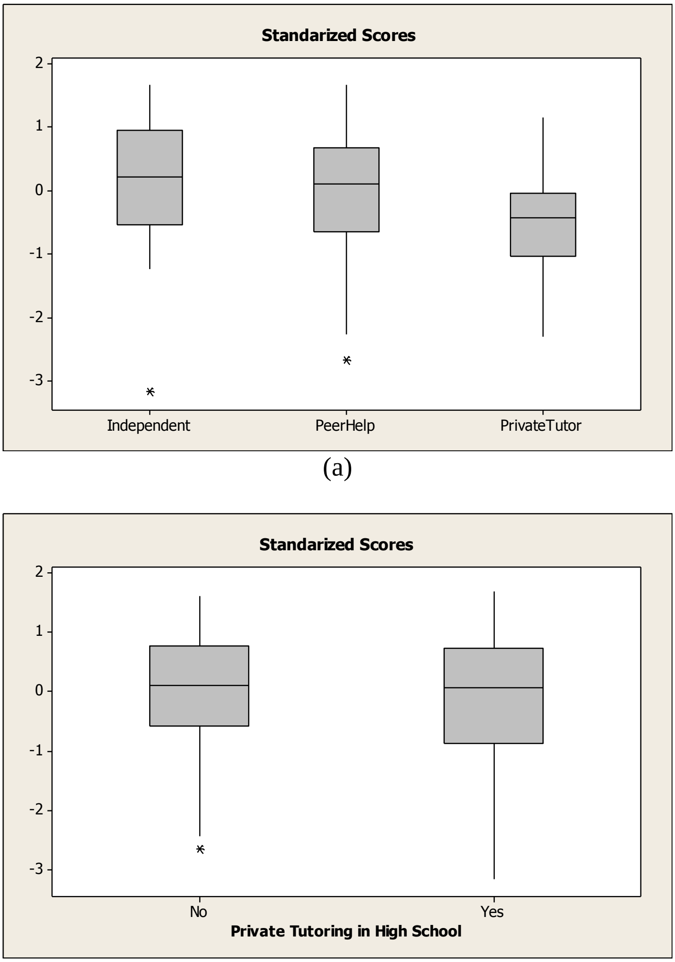
Table 1 shows the counts and percentages of students classified by the learning support method they reported (independent learning, peer help and private tutor) and whether the student had private tutoring experience in high school. Six students did not answer either one of the two questions.

Table 1: Two-way classification of students by learning method and by high school tutoring experience

Learning Method	***Tutoring experience in High school***		***Total***
	No	Yes	
Independent	24 (69%)	11 (31%)	35
Peer help	60 (62%)	37 (38%)	97
Private tutor	5 (36%)	9 (64%)	14
Total	89	57	146

The results clearly show students who received private tutoring in high school in Mathematics are more likely (64%) to continue seeking private tutoring in the subject in first year university than those who did not have private tutoring experience in high school (36%). On the other hand, the likelihood that a student who received private tutoring in high school becomes independent learner while attending university is about 31% compared to 69% for those who did not receive tutoring in high school. It should be mentioned that among the 11 independent learners who had tutoring experience in high school, 81% of them have had private tutoring in Mathematics during only the last year of high school compared to 40% for the 37 students who were seeking peer help and 25% for the 9 private tutored students. The rest of the students in each group have received private tutoring in Mathematics for two or three years of high school. This suggests that students who have longer years of experience in receiving tutoring in Mathematics in high school are more likely to request tutoring in the subject in the freshman year of university education.

The following box-plot charts depict the comparisons of the standardized final grades in the course for students from different classifications according to private tutoring related experiences. The line in the middle of the box in each plot represents the median (center) of the scores' distribution. The asterisk points at the lower end of the box indicate unusually low values.



(a)

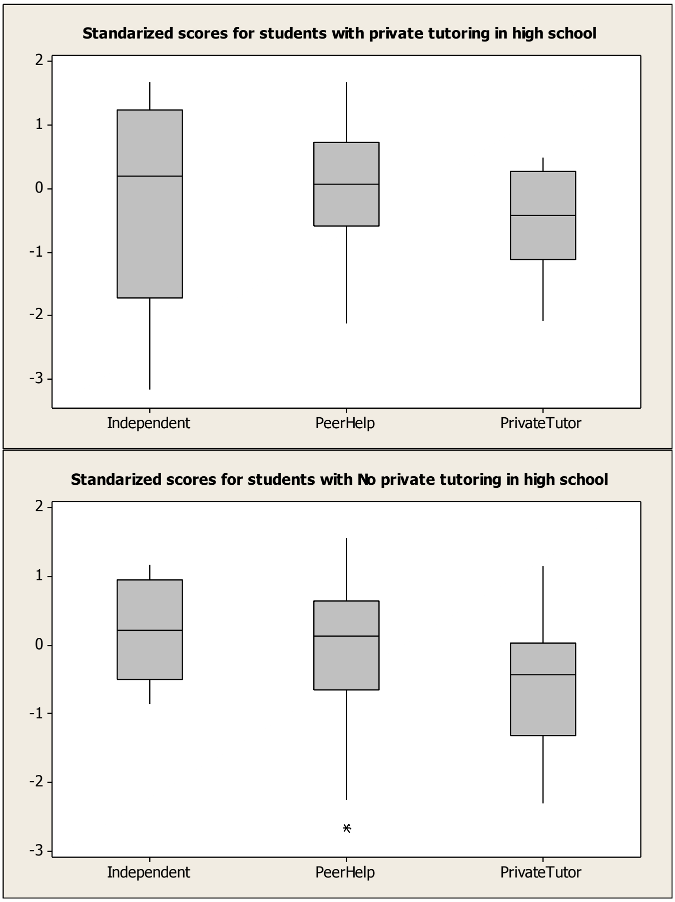
(b)

Fig. 1: Box-plots of standardized final scores categorized by
(a) learning method and
(b) tutoring experience in high school

Box-plot(a) in Figure 1 shows the comparisons between students' scores in the three learning groups (independent, peer help and private tutor). It is obvious that the private tutor group scored the least on average. Pair wise t-tests indicate that the private tutor group scored significantly less than the independent learners (*p-value=0.02*) but not significantly less than the peer help group (*p-value=0.06).* The independent learners and the peer help group did not show significant difference between their average scores (*p-value=0.28).* It should be mentioned that 58% of the private tutor students had failing grade in the course compared to 26% of the peer help students and 21% of the independent learners group. One can question the motivation of the students to purchase private tutoring in university level Mathematics and whether their grades will be worse if they did not purchase tutoring. Few studies have addressed the question of motivation for seeking tutoring in the formal school system and shown that motivation is one of the unobservable factors that could significantly affect the effectiveness of tutoring. Bray (2006) explained that some tutees join tutoring because of peer pressure while more motivated tutees join tutoring so as to master the subject and outperform their peers. While our collected data do not reveal what motivate students to seek private tutoring in the course other than improving performance, it is reasonable to believe that Business majors taking a required prerequisite Mathematics course are primarily drawn to private tutoring to achieve a pass in the course so that they can move on with their major courses in business. They do not aim at outperforming their peers in the course or mastering course topics. Majority of them seek tutoring when faced with difficulties in the course material causing them apprehension of failing the course. Purchasing private tutoring when challenged with Mathematics seems very logical thing to do for some of them as this has been the strategy adapted and encouraged by their parents during formal school education. In terms of the cost, it is much cheaper to pay for the tutoring than to pay the costs of repeating the course.

Plot (b) in Figure 1 shows the standardized scores for students who had private tutoring experience in high school and students who did not have that experience. There is no significant difference in the average score (data center) between the two groups (*p-value* ~1). While tutoring was effective in helping majority of the students graduate from high school and possibly outperform their high school peers, it did not necessarily equip them with sufficient skills that allow them to perform better in university education than those who did not receive tutoring in high school. There could be some tutoring unrelated reasons for the two groups of students to perform at the same level. This result, however, prompts one to question the tutors and their competency in tutoring Mathematics curriculum at university level. The tutors are usually high school teachers (university professors are prohibited by law from engaging in this practice) who have been teaching their tutees for some years. While they were successful in providing effective teachings that allowed their tutees to graduate from high school, outperform their peers and meet the competitive university admission standards they seem less effective in tutoring university level Mathematics.

In Figure 2 the standardized scores for students who received private tutoring in Mathematics in high school and those who did not were separated. The box-plot charts in the figure show the comparisons of the scores in each group by the learning method used in the course. It is evident that students who purchased private tutoring to help them improve performance in first year Mathematics course consistently underperformed (not significantly) their peers regardless whether they had private tutoring experience in the subject during high school years or not.


Standarized scores for students with private tutoring in high school

Standarized scores for students with No private tutoring in high school

Fig. 2: Box-plots of standardized final scores of the three learning methods for students with and without tutoring experience in high school

It is interesting to report that 55% of the students who received tutoring in Mathematics in high school agreed to the statement that this experience has been effective in facilitating their learning of the subject during first year university. It is unclear if these students truly felt that without their tutoring experience in high school they would have had more difficulty in their learning of first year university Mathematics and possibly performed worse than what the data has revealed or perhaps they simply did not want to report that tutoring did not have any impact. 17% of them also agreed that receiving private tutoring in Mathematics during first year university is necessary for better performance compared to 9% for those who did not have tutoring experience high school.

4. CONCLUSIONS

This exploratory study represents a first attempt to examine the long term impact of private tutoring in high school on students' learning of Mathematics in first year university. The findings show that students who received tutoring in Mathematics in high school are more likely to request tutoring in the subject during the freshman year and less likely to become independent learners than those who did not receive tutoring in high school. As for performance, it is shown that students who purchased private tutoring while attending a freshman Business Mathematics course performed significantly less than those who did not purchase tutoring in the course. Similar performance patterns were seen for these students irrespective to whether they had private tutoring experience in Mathematics in high school years or not.

On the whole, the effectiveness of private tutoring in either high school or first year university on students' learning of university Mathematics was found insignificant in this study. Students who received tutoring in high school and continued to receive tutoring while taking first year university Mathematics showed significantly higher failing rates in the course than their peers. The results raise the question of competency of the tutor, usually a high school teacher, in tutoring university Mathematics curriculum. However, we should mention that *there could be some other unobservable factors or tutoring unrelated reasons* for the results obtained in this study.

REFERENCES

1. Ahmed, A. (2010). Private tutoring becoming a trend in the UAE. *Khaleej Times*, August 2010.
2. Banerjee, A.V., Cole, S., Duflo, E. and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, 122(3), 1235-1264.
3. Bray, M. (2006). Private supplementary tutoring: comparative perspectives on parents and implications. *Compare*, 36(4), 515-530.
4. Bray, M. (2007). *The shadow education system: Private tutoring and its implications for planners*. Paris: International Institute for Educational Planning, UNESCO.
5. Dang, H. and Rogers, F. (2008). The growing phenomenon of private tutoring: Does is deepen human capital, widen inequalities or waste resources? *The world Bank Research Observer*, 23(2), 161-200.
6. Farah, S. (2011). *Private tutoring trends in the UAE*. Dubai School of Government, Policy Brief 26.
7. Navarra-Madsen, J. and Ingram, P. (2010). Mathematics tutoring and student success. *Procedia Social and Behavioral Sciences*, No. 8, 207-212.
8. Ono, H. (2007). Does examination hell pay off? A cost-benefit analysis of 'Ronin' and college education in Japan. *Economics of Education Review*, 26(3), 271-284.
9. Paviot, L., Heinsohn, N. and Korkman, J. (2005). The association between extra tuition and student achievement. Paper prepared for the *International SACMEQ Educational Policy Research Conference*, UNESCO International Institute for Educational Planning, Paris.
10. Tansel, A. and Bircan, F. (2006). Demand for education in Turkey: A Tobit analysis of private tutoring expenditures. *Economics of Education Review*, 25(3), 303-313.
11. UK news: (2005). Private tuition a waste of money? *Education + Training*, 47(8/9).
12. Unal, H. Ozkan., E, Milton, S., Price, K. and Curva, F. (2010). The effect of private tutoring on performance in mathematics in Turkey: A comparison across occupational types. *Procedia Social and Behavioral Sciences*, 2, 5512-5517.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 89-102

IMPACT OF HIGHER ORDER AND PRODUCT TERMS ON MEAN SQUARE ERROR FOR VARIOUS ESTIMATORS

Muhammad Ismail[1], **Muhammad Rashad**[2], **Muhammad Qaiser Shahbaz**[3] and **Muhammad Hanif** [2]

[1] Lahore Garrison University, Lahore, Pakistan, Email: drismail39@gmail.com

[2] National College of Business Administration & Economics, Lahore, Pakistan
Email: geo_stat@hotmail.com; drmainhanif@gmail.com

[3] Department of Mathematics COMSATS Institute of IT, Lahore. Pakistan.
Email: qshahbaz@gmail.com

ABSTRACT

General practice of deriving the expression for mean square error (MSE) of various estimators is to ignore product and higher order terms and including the product term when expression for bias is to be derived. We suggest inclusion of all product and higher terms to derive the expression for mean square error (MSE). We derived a general form of expression from which one can obtain the expression for mean square error (MSE) of classical ratio estimator up to any order.

KEY WORDS

Auxiliary Variable, Mean Square Error (MSE), Higher Order

1. INTRODUCTION

Statisticians always show much concern to the variability and try to overcome this problem. They made efforts to control or minimize variability by using different statistical methods.

In estimation theory, survey statisticians have developed many estimators and compared them on the basis of their variances.

The scientific development in the field of survey sampling has long history but the groundbreaking work in this field is done by Neyman (1934). Cochran (1940) appears to be the first to use auxiliary information in ratio estimator. The highly correlated auxiliary variable ‘x’ increases the efficiency and precision of the estimators of population characteristics.

A large number of estimators have been constructed in single phase and two phase sampling by modifying regression, ratio and product estimators. The mean square errors of most of the existing estimators are derived by using Taylor’s series up to first order approximation by ignoring the higher order and product terms which reduces precision of the estimators.

2. NOTATIONS

The following notations will be used for deriving the higher order mean square error of classical ratio estimator.

$$\left.\begin{array}{l}\overline{\mathrm{y}}=\overline{\mathrm{Y}}+\overline{\mathrm{e}}_{\mathrm{y}}\\ \overline{\mathrm{y}}_{2}=\overline{\mathrm{Y}}+\overline{\mathrm{e}}_{\mathrm{y}_{2}}\\ \overline{\mathrm{x}}=\overline{\mathrm{X}}+\overline{\mathrm{e}}_{\mathrm{x}}\\ \overline{\mathrm{x}}_{2}=\overline{\mathrm{X}}+\overline{\mathrm{e}}_{\mathrm{x}_{2}}\end{array}\right\}$$

N. Balakrishnan discussed the following results in Continuous Multivariate Distributions Vol. 1 Second edition page 261.

$$L=\mu_{2r,2s}=\frac{(2r)!\,(2s)!}{2^{r+s}}\sum_{t=0}^{p}\frac{(2\rho)^{2t}}{(r-t)!\,(s-t)!\,(2t)!}\qquad;\qquad r,s=0,1,2,\ldots$$

$$E\left(\overline{e}_{x}^{2r}\overline{e}_{y}^{2s}\right)=L\theta S_{x}^{2r}S_{y}^{2s}$$

$$M=\mu_{2r+1,2s+1}=\frac{(2r+1)!\,(2s+1)!}{2^{r+s}}\,\rho\sum_{t=0}^{p}\frac{(2\rho)^{2t}}{(r-t)!\,(s-t)!\,(2t+1)!}$$

$$E\left(\overline{e}_{x}^{2r+1}\overline{e}_{y}^{2s+1}\right)=M\theta S_{x}^{2r+1}S_{y}^{2s+1}$$

"p" is minimum of r & s.

$\mu_{r,s}=0$ if $r+s$ is odd where $\theta=\frac{1}{n}-\frac{1}{N}$

3. SOME EXISTING ESTIMATORS

The following estimators are developed by various survey statisticians and their mean square errors are derived by using first order approximation:

3.1 For Single Phase

- **Classical ratio estimator**

  The classical ratio estimator is as follows

  $$t_{1}=\frac{\overline{y}}{\overline{x}}\overline{X}$$

  The MSE of classical ratio estimator is

  $$MSE\left(t_{1E}\right)=\theta\overline{Y}^{2}\left[C_{y}^{2}+C_{x}^{2}-2C_{x}C_{y}\rho\right].$$

- **Bhal and Tuteja's Estimator**

  The Bhal and Tuteja's estimator is as follows

$$t_2 = \bar{y}\, exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right),$$

The MSE of Bhal and Tuteja’s estimator is

$$MSE\left(t_{2\text{E}}\right) = \theta\bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - C_x C_y \rho\right)$$

3.2 For Two Phase

- **Classical ratio estimator**
  The classical ratio estimator is as follows

$$t_{1(2)} = \frac{\bar{y}_2}{\bar{x}_2}\bar{X} \qquad (\bar{X} \text{ is known})$$

The MSE of classical ratio estimator is

$$MSE\left(t_{1(2)E}\right) = \theta_2\bar{Y}^2\left[C_y^2 + C_x^2 - 2C_x C_y \rho\right].$$

- **Bhal and Tuteja’s Estimator**
  The Bhal and Tuteja’s estimator is as follows

$$t_{2(2)} = \bar{y}_2\, exp\left(\frac{\bar{X}-\bar{x}_2}{\bar{X}+\bar{x}_2}\right), \qquad (\bar{X} \text{ is known})$$

The MSE of Bhal and Tuteja’s estimator is

$$MSE\left(t_{2(2)\text{E}}\right) = \theta_2\bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - C_x C_y \rho\right)$$

4. HIGHER ORDER EXPRESSIONS FOR MEAN SQUARE ERROR

4.1 For Single Phase

- The general expression for MSE of **classical ratio estimator** is

$$\begin{aligned} MSE_g\left(t_1\right) = {} & \bar{Y}^2 \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} a_i a_j E\left(\bar{e}_x^{i+j}\right) + \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} a_i a_j E\left(\bar{e}_x^{i+j}\bar{e}_y^2\right) \\ & + 2\bar{Y}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty} a_i a_j E\left(\bar{e}_x^{i+j}\bar{e}_y\right) + \bar{Y}^2 - 2\bar{Y}^2 \sum_{j=0}^{\infty} a_j E\left(\bar{e}_x^j\right) - 2\bar{Y}\sum_{j=0}^{\infty} a_j E\left(\bar{e}_x^j \bar{e}_y\right). \end{aligned} \tag{4.1}$$

By expanding (4.1) upto different orders we can get the expressions upto required order. We have expanded it from first to fourth order.

- **Expand up to first order**

$$MSE_1\left(t_1\right) = MSE\left(t_{1E}\right) + \theta\bar{Y}^2\left(1+2\rho^2\right)C_x^2 C_y^2.$$

- **Expand up to second order**

$$MSE_2\left(t_1\right)=MSE_1\left(t_1\right)+\theta\bar{Y}^2\left[\,3C_x^4+2\left(1+2\rho^2\right)C_x^2C_y^2+3\left(1+4\rho^2\right)C_x^4C_y^2-12C_x^3C_y\rho\right].$$

- **Expand up to third order**

$$MSE_3\left(t_1\right)=MSE_2\left(t_1\right)+\theta\bar{Y}^2\begin{bmatrix}6C_x^4+15C_x^6\;-6C_x^3C_y\rho-60C_x^5C_y\rho\\ \quad+6\left(1+4\rho^2\right)C_x^4C_y^2+15\left(1+6\rho^2\right)C_x^6C_y^2\end{bmatrix}.$$

- **Expand up to fourth order**

$$MSE_4\left(t_1\right)=MSE_3\left(t_1\right)+\theta\bar{Y}^2\begin{bmatrix}30C_x^6+105C_x^8\;-60C_x^5C_y\rho-420C_x^7C_y\rho\\ \quad+6\left(1+4\rho^2\right)C_x^4C_y^2+30\left(1+6\rho^2\right)C_x^6C_y^2\\ \quad+105\left(1+8\rho^2\right)C_x^8C_y^2\end{bmatrix}.$$

- The general expression for MSE of **Bhal and Tuteja's** estimator is

$$MSE_g\left(t_2\right)=\bar{Y}^2\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_x^{i+j}\right)+\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_x^{i+j}\bar{e}_y^{2}\right)$$

$$+2\bar{Y}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_x^{i+j}\bar{e}_y\right)+\bar{Y}^2-2\bar{Y}^2\sum_{j=0}^{\infty}a_jE\left(\bar{e}_x^{j}\right)-2\bar{Y}\sum_{j=0}^{\infty}a_jE\left(\bar{e}_x^{j}\bar{e}_y\right). \quad (4.2)$$

By expanding (4.2) upto different orders we can get the expressions upto required order. We have expanded it from first to fourth order.

- **Expand up to first order**

$$MSE_1\left(t_2\right)=MSE_1\left(t_{2E}\right)+\frac{\theta\bar{Y}^2\left(1+2\rho^2\right)C_x^2C_y^2}{4}.$$

- **Expand up to second order**

$$MSE_2\left(t_2\right)=MSE_1\left(t_2\right)+\theta\bar{Y}^2\left[\frac{3C_x^4}{16}+\frac{1}{2}\left(1+2\rho^2\right)C_x^2C_y^2+\frac{3}{16}\left(1+4\rho^2\right)C_x^4C_y^2-\frac{3}{2}C_x^3C_y\rho\right].$$

- **Expand up to third order**

$$MSE_3\left(t_2\right)=MSE_2\left(t_2\right)+\theta\bar{Y}^2\begin{bmatrix}\frac{3}{8}C_x^4+\frac{15}{64}C_x^6+\frac{3}{8}\left(1+4\rho^2\right)\mathrm{C}_\mathrm{x}^4\mathrm{C}_\mathrm{y}^2\\ \quad+\frac{15}{64}\left(1+6\rho^2\right)\mathrm{C}_\mathrm{x}^6\mathrm{C}_\mathrm{y}^2\\ \quad-\frac{3}{4}\mathrm{C}_\mathrm{x}^3C_y\rho-\frac{15}{8}C_x^5C_y\rho\end{bmatrix}.$$

- **Expand up to fourth order**

$$MSE_4\left(t_2\right) = MSE_3\left(t_2\right) + \theta\bar{Y}^2\begin{bmatrix}\frac{15}{32}C_x^6 + \frac{105}{256}C_x^8 + \frac{15}{32}\left(1+6\rho^2\right)C_x^6C_y^2 \\ + \frac{105}{256}\left(1+8\rho^2\right)C_x^8C_y^2 - \frac{15}{8}C_x^5C_y\rho - \frac{105}{32}C_x^7C_y\rho\end{bmatrix}.$$

4.2 For Two Phase

- The general expression for MSE of **classical ratio estimator** is

$$MSE\left(t_{1(2)}\right) = \bar{Y}^2\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\right) + \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\bar{e}_{y_2}^2\right)$$

$$+2\bar{Y}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\bar{e}_{y_2}\right)+\bar{Y}^2-2\bar{Y}^2\sum_{j=0}^{\infty}a_jE\left(\bar{e}_{x_2}^{j}\right)-2\bar{Y}\sum_{j=0}^{\infty}a_jE\left(\bar{e}_{x_2}^{j}\bar{e}_{y_2}\right) \quad (4.3)$$

- **Expand up to first order**

$$MSE_1\left(t_{1(2)}\right) = MSE\left(t_{1E}\right) + \theta_2\bar{Y}^2\left(1+2\rho^2\right)C_x^2C_y^2.$$

- **Expand up to second order**

$$MSE_2\left(t_{1(2)}\right) = MSE_1\left(t_{1(2)}\right) + \theta_2\bar{Y}^2\begin{bmatrix}3C_x^4 + 2\left(1+2\rho^2\right)C_x^2C_y^2 \\ +3\left(1+4\rho^2\right)C_x^4C_y^2 - 12C_x^3C_y\rho\end{bmatrix}.$$

- **Expand up to third order**

$$MSE_3\left(t_{1(2)}\right) = MSE_2\left(t_{1(2)}\right) + \theta_2\bar{Y}^2\begin{bmatrix}6C_x^4 + 15C_x^6 - 6C_x^3C_y\rho - 60C_x^5C_y\rho \\ +6\left(1+4\rho^2\right)C_x^4C_y^2 + 15\left(1+6\rho^2\right)C_x^6C_y^2\end{bmatrix}.$$

- **Expand up to fourth order**

$$MSE_4\left(t_{1(2)}\right) = MSE_3\left(t_{1(2)}\right) + \theta_2\bar{Y}^2\begin{bmatrix}30C_x^6 + 105C_x^8 - 60C_x^5C_y\rho \\ -420C_x^7C_y\rho + 6\left(1+4\rho^2\right)C_x^4C_y^2 \\ +30\left(1+6\rho^2\right)C_x^6C_y^2 + 105\left(1+8\rho^2\right)C_x^8C_y^2\end{bmatrix}.$$

- The general expression for MSE of Bhal and Tuteja's estimator is

$$MSE_g\left(t_{2(2)}\right) = \bar{Y}^2\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\right) + \sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\bar{e}_{y_2}^2\right)$$

$$+2\bar{Y}\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}a_ia_jE\left(\bar{e}_{x_2}^{i+j}\bar{e}_{y_2}\right)+\bar{Y}^2-2\bar{Y}^2\sum_{j=0}^{\infty}a_jE\left(\bar{e}_{x_2}^{j}\right)-2\bar{Y}\sum_{j=0}^{\infty}a_jE\left(\bar{e}_{x_2}^{j}\bar{e}_{y_2}\right). \quad (4.4)$$

By expanding (4.4) upto different orders we can get the expressions upto required order. We have expanded it from first to fourth order.

- **Expand up to first order**

$$MSE_1\left(t_{2(2)}\right) = MSE\left(t_{2(2)E}\right) + \frac{\theta_2\bar{Y}^2\left(1+2\rho^2\right)C_x^2C_y^2}{4}.$$

- **Expand up to second order**

$$MSE_2\left(t_{2(2)}\right) = MSE_1\left(t_{2(2)}\right) + \theta_2\bar{Y}^2\left[\begin{array}{l} \frac{3C_x^4}{16} + \frac{1}{2}\left(1+2\rho^2\right)C_x^2C_y^2 \\ + \frac{3}{16}\left(1+4\rho^2\right)C_x^4C_y^2 - \frac{3}{2}C_x^3C_y\rho \end{array}\right].$$

- **Expand up to third order**

$$MSE_3\left(t_{2(2)}\right) = MSE_2\left(t_{2(2)}\right) + \theta_2\bar{Y}^2\left[\begin{array}{l} \frac{3}{8}C_x^4 + \frac{15}{64}C_x^6 + \frac{3}{8}\left(1+4\rho^2\right)C_x^4C_y^2 \\ + \frac{15}{64}\left(1+6\rho^2\right)C_x^6C_y^2 \\ -\frac{3}{4}C_x^3C_y\rho - \frac{15}{8}C_x^5C_y\rho \end{array}\right].$$

- **Expand up to fourth order**

$$MSE_4\left(t_{2(2)}\right) = MSE_3\left(t_{2(2)}\right) + \theta_2\bar{Y}^2\left[\begin{array}{l} \frac{15}{32}C_x^6 + \frac{105}{256}C_x^8 + \frac{15}{32}\left(1+6\rho^2\right)C_x^6C_y^2 \\ + \frac{105}{256}\left(1+8\rho^2\right)C_x^8C_y^2 - \frac{15}{8}C_x^5C_y\rho - \frac{105}{32}C_x^7C_y\rho \end{array}\right].$$

5. NUMERICAL STUDY

Numerical Study for Classical Ratio Estimator at Different Level of Approximations of MSE for Single Phase Sampling.

In this section empirical study has been made by classical ratio estimator at different level of approximations of MSE for single phase sampling. The data of Khare and Sinha (2007) and Singh and Kumar (2011) have been considered. Description of the population data is given below:

The data on physical growth of upper socioeconomic group of 95 school children of Varanasi under an ICMR study, Department of pediatrics, B.H.U., during 1983 – 84 has been taken under study. Let us consider the study and auxiliary variable as follows:

y : Weight in kg of the children,
x : Skull circumference in cm of the children.

Population 1

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.03006$, $\rho = 0.328$

Order	n = 10	n = 15	n = 18	n = 21	n = 25
Existing	0.755096	0.473786	0.380016	0.313037	0.248738
1	0.756007	0.474357	0.380474	0.313414	0.249037
2	0.757346	0.475197	0.381148	0.31397	0.249479
3	0.757232	0.475126	0.381091	0.313923	0.249441
4	0.757236	0.475129	0.381093	0.313924	0.249443

Population 2

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.05860$, $\rho = 0.846$

Order	n = 10	n = 15	n = 18	n = 21	n = 25
Existing	0.41936	0.263128	0.211051	0.173852	0.138142
1	0.426283	0.267471	0.214534	0.176722	0.140422
2	0.430596	0.270178	0.216705	0.17851	0.141843
3	0.427642	0.268324	0.215218	0.177286	0.14087
4	0.427725	0.268376	0.21526	0.17732	0.140898

Artificial Population 1

N = 15, $\overline{Y} = 45.4$, $C_y = 0.2281$, $C_x = 0.1009$, $\rho = 0.2235$

Order	n = 2	n = 3	n = 4	n = 5	n = 6
Existing	46.37566	28.53887	19.62047	14.26943	10.70208
1	46.89604	28.8591	19.84063	14.42955	10.82216
2	47.67058	29.33574	20.16832	14.66787	11.0009
3	47.96658	29.5179	20.29355	14.75895	11.06921
4	48.00191	29.53964	20.3085	14.76982	11.07736

Artificial Population 2

N = 30, $\overline{Y} = 54.47$, $C_y = 0.1604$, $C_x = 0.0825$, $\rho = 0.2716$

Order	n = 2	n = 3	n = 4	n = 5	n = 6
Existing	35.09418	22.56055	16.29373	12.53364	10.02691
1	35.37241	22.73941	16.42291	12.633	10.1064
2	35.72127	22.96367	16.58487	12.7576	10.20608
3	35.90868	23.08415	16.67188	12.82453	10.25962
4	35.92092	23.09202	16.67757	12.8289	10.26312

Artificial Population 3

N = 45, $\overline{Y} = 71.57$, $C_y = 0.1294$, $C_x = 0.0627$, $\rho = 0.95$

Order	n = 2	n = 3	n = 4	n = 5	n = 6
Existing	12.87338	8.382665	6.137309	4.790094	3.891952
1	13.32526	8.676914	6.35274	4.958236	4.028567
2	13.46138	8.765547	6.417632	5.008884	4.069718
3	13.24587	8.625219	6.314892	4.928697	4.004566
4	13.25043	8.628185	6.317064	4.930391	4.005943

Numerical Study for Bhal and Tuteja's Estimator at Different Level of Approximations of MSE for Single Phase Sampling.

Population 1

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.03006$, $\rho = 0.328$

Order	n = 10	n = 15	n = 18	n = 21	n = 25
Existing	0.755096	0.473786	0.380016	0.313037	0.248738
1	0.756007	0.474357	0.380474	0.313414	0.249037
2	0.757346	0.475197	0.381148	0.31397	0.249479
3	0.757232	0.475126	0.381091	0.313923	0.249441
4	0.757236	0.475129	0.381093	0.313924	0.249443

Population 2

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.05860$, $\rho = 0.846$

Order	n = 10	n = 15	n = 18	n = 21	n = 25
Existing	0.41936	0.263128	0.211051	0.173852	0.138142
1	0.426283	0.267471	0.214534	0.176722	0.140422
2	0.430596	0.270178	0.216705	0.17851	0.141843
3	0.427642	0.268324	0.215218	0.177286	0.14087
4	0.427725	0.268376	0.21526	0.17732	0.140898

Artificial Population 1

N = 15, $\overline{Y} = 45.4$, $C_y = 0.2281$, $C_x = 0.1009$, $\rho = 0.2235$

Order	n = 2	n = 3	n = 4	n = 5	n = 6
Existing	44.15016	27.16933	18.67891	13.58466	10.1885
1	44.28026	27.24939	18.73395	13.62469	10.21852
2	44.48873	27.37768	18.82215	13.68884	10.26663
3	44.48987	27.37838	18.82264	13.68919	10.26689
4	44.4916	27.37945	18.82337	13.68972	10.26729

Artificial Population 2

N = 30, $\overline{Y} = 54.47$, $C_y = 0.1604$, $C_x = 0.0825$, $\rho = 0.2716$

Order	n = 2	n = 3	n = 4	n = 5	n = 6
Existing	33.00261	21.21597	15.32264	11.78665	9.429318
1	33.07217	21.26068	15.35494	11.81149	9.449192
2	33.17291	21.32544	15.40171	11.84747	9.477974
3	33.17203	21.32488	15.40130	11.84716	9.477724
4	33.17261	21.32525	15.40157	11.84736	9.477889

Artificial Population 3

N = 45, $\overline{Y} = 71.57$, $C_y = 0.1294$, $C_x = 0.0627$, $\rho = 0.95$

Order	**n = 2**	**n = 3**	**n = 4**	**n = 5**	**n = 6**
Existing	24.52070	15.96696	11.69010	9.12398	7.41323
1	24.63367	16.04053	11.74396	9.16602	7.44739
2	24.75601	16.12019	11.80228	9.21154	7.48438
3	24.71517	16.09360	11.78281	9.19634	7.47203
4	24.71579	16.09400	11.78311	9.19657	7.47221

Population 1

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.02478$, $\rho = 0.477$

Order	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 10$** **$n_2 = 8$**	**$n_1 = 15$** **$n_2 = 12$**	**$n_1 = 21$** **$n_2 = 19$**	**$n_1 = 25$** **$n_2 = 23$**
Existing	0.674713	0.202414	0.134943	0.040584	0.028162
1	0.675402	0.202621	0.13508	0.040626	0.028191
2	0.676389	0.202917	0.135278	0.040685	0.028232
3	0.676247	0.202874	0.135249	0.040677	0.028226
4	0.67625	0.202875	0.13525	0.040677	0.028226

Population 2

N = 95, $\overline{Y} = 19.4968$, $C_y = 0.15613$, $C_x = 0.05402$, $\rho = 0.729$

Order	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 10$** **$n_2 = 8$**	**$n_1 = 15$** **$n_2 = 12$**	**$n_1 = 21$** **$n_2 = 19$**	**$n_1 = 25$** **$n_2 = 23$**
Existing	0.475085	0.142526	0.095017	0.028577	0.01983
1	0.479734	0.14392	0.095947	0.028856	0.020024
2	0.483081	0.144924	0.096616	0.029058	0.020163
3	0.481326	0.144398	0.096265	0.028952	0.02009
4	0.481374	0.144412	0.096275	0.028955	0.020092

Artificial Population 1

N = 15, $\overline{Y} = 45.4$, $C_y = 0.2281$, $C_x = 0.1009$, $\rho = 0.2235$

Order	**$n_1 = 6$** **$n_2 = 2$**	**$n_1 = 6$** **$n_2 = 3$**	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 6$** **$n_2 = 5$**
Existing	35.67359	17.83679	8.918396	3.567359
1	36.07388	18.03694	9.01847	3.607388
2	36.66967	18.33484	9.167418	3.666967
3	36.89737	18.44868	9.224342	3.689737
4	36.92455	18.46227	9.231137	3.692455

Artificial Population 2

$N = 30,\ \overline{Y} = 54.47,\ C_y = 0.1604, C_x = 0.0825,\ \rho = 0.2716$

Order	$n_1 = 6$ $n_2 = 2$	$n_1 = 6$ $n_2 = 3$	$n_1 = 6$ $n_2 = 4$	$n_1 = 6$ $n_2 = 5$
Existing	25.067274	12.533637	6.266819	2.506727
1	25.266010	12.633005	6.316502	2.526601
2	25.515190	12.757595	6.378798	2.551519
3	25.649054	12.824527	6.412263	2.564905
4	25.657796	12.828898	6.414449	2.565780

Artificial Population 3

$N = 45,\ \overline{Y} = 71.57,\ C_y = 0.1294, C_x = 0.0627,\ \rho = 0.95$

Order	$n_1 = 6$ $n_2 = 2$	$n_1 = 6$ $n_2 = 3$	$n_1 = 6$ $n_2 = 4$	$n_1 = 6$ $n_2 = 5$
Existing	8.981427	4.490714	2.245357	0.898143
1	9.296693	4.648347	2.324173	0.929669
2	9.391657	4.695829	2.347914	0.939166
3	9.241306	4.620653	2.310326	0.924131
4	9.244483	4.622242	2.311121	0.924448

Numerical Study for Bhal and Tuteja's Estimator at Different Level of Approximations of MSE for Two Phase Sampling.

Population 1

$N = 95,\ \overline{Y} = 19.4968,\ C_y = 0.15613, C_x = 0.02478,\ \rho = 0.477$

Order	$n_1 = 6$ $n_2 = 4$	$n_1 = 10$ $n_2 = 8$	$n_1 = 15$ $n_2 = 12$	$n_1 = 21$ $n_2 = 19$	$n_1 = 25$ $n_2 = 23$
Existing	0.674713	0.202414	0.134943	0.040584	0.028162
1	0.675402	0.202621	0.13508	0.040626	0.028191
2	0.676389	0.202917	0.135278	0.040685	0.028232
3	0.676247	0.202874	0.135249	0.040677	0.028226
4	0.67625	0.202875	0.13525	0.040677	0.028226

Population 2

$N = 95,\ \overline{Y} = 19.4968,\ C_y = 0.15613, C_x = 0.05402,\ \rho = 0.729$

Order	$n_1 = 6$ $n_2 = 4$	$n_1 = 10$ $n_2 = 8$	$n_1 = 15$ $n_2 = 12$	$n_1 = 21$ $n_2 = 19$	$n_1 = 25$ $n_2 = 23$
Existing	0.475085	0.142526	0.095017	0.028577	0.01983
1	0.479734	0.14392	0.095947	0.028856	0.020024
2	0.483081	0.144924	0.096616	0.029058	0.020163
3	0.481326	0.144398	0.096265	0.028952	0.02009
4	0.481374	0.144412	0.096275	0.028955	0.020092

Artificial Population 1

N = 15, $\overline{Y} = 45.4$, $C_y = 0.2281, C_x = 0.1009$, $\rho = 0.2235$

Order	**$n_1 = 6$** **$n_2 = 2$**	**$n_1 = 6$** **$n_2 = 3$**	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 6$** **$n_2 = 5$**
Existing	33.96166	16.98083	8.49042	3.39617
1	34.06173	17.03087	8.51543	3.40617
2	34.22210	17.11105	8.55552	3.42221
3	34.22298	17.11149	8.55574	3.42230
4	34.22431	17.11216	8.55608	3.42243

Artificial Population 2

N = 30, $\overline{Y} = 54.47$, $C_y = 0.1604, C_x = 0.0825$, $\rho = 0.2716$

Order	**$n_1 = 6$** **$n_2 = 2$**	**$n_1 = 6$** **$n_2 = 3$**	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 6$** **$n_2 = 5$**
Existing	23.57330	11.78665	5.89332	2.35733
1	23.62298	11.81149	5.90574	2.36230
2	23.69493	11.84747	5.92373	2.36949
3	23.69431	11.84716	5.92358	2.36943
4	23.69472	11.84736	5.92368	2.36947

Artificial Population 3

N = 45, $\overline{Y} = 71.57$, $C_y = 0.1294, C_x = 0.0627$, $\rho = 0.95$

Order	**$n_1 = 6$** **$n_2 = 2$**	**$n_1 = 6$** **$n_2 = 3$**	**$n_1 = 6$** **$n_2 = 4$**	**$n_1 = 6$** **$n_2 = 5$**
Existing	17.10746	8.55373	4.27687	1.71075
1	17.18628	8.59314	4.29657	1.71863
2	17.27164	8.63582	4.31791	1.72716
3	17.24314	8.62157	4.31078	1.72431
4	17.24357	8.62179	4.31089	1.72436

6. CONCLUSION

The results suggest that in calculating the MSE, there existed significant difference when we used different levels of approximation. It is further observed that when we increase the order of approximation at different levels, the value of MSE, after certain levels of approximation, is also iterated. The existing practice to calculate MSE ignores second and higher order terms completely which may lead us to unsatisfactory conclusions.

The study also justified the basic principle of sampling theory i.e. when we increase the sample size, MSE is reduced.

7. REFERENCES

1. Bahl, S. and Tuteja, R.K. (1991). Ratio and Product type exponential estimator. *Information and Optimization sciences*, XII(I), 159-163.
2. Beale, E.M.L. (1962). Some uses of computers in operations research. *Industrielle Organisation,* 31, 51-52.

3. Bowley, A.L. (1913). Working class households in reading. *J. Roy. Statist. Soc.*, 76, 672-701.
4. Bowley, A.L. (1926). Measurements of precision attained in sampling. *Bull. Inst. Inte. Statist.* 22, 1-62.
5. Breidt, J. and Fuller, W.A. (1993). Regression weighting for multiplicative samples. *Sankhya,* 55, 297-309.
6. Brewer, K.R.W. (1963). Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process. *Austral. J. Statist*, 5, 93-105.
7. Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977). *Foundations of inference in survey sampling.* Wiley-Interscience.
8. Chaudhuri, A., and Roy, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika,* 41, 355-362.
9. Cochran, W.G. (1977). *Sampling Techniques.* J. Wiley New York.
10. Cochran, W.G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *J. Agri. Sc.* 30, 262-275.
11. Cochran, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *J. Amer. Statist. Assoc.,* 37, 199-212.
12. Deming, W.E. (1960). *Sample Design in Business Research*. John Wiley and Sons, Inc., New York.
13. Dupont, F. (1995). *Redressement alternatifs en presence de plusieurs niveaux d'information auxillaire*. Internal report from INSEE, Paris, France.
14. Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika,* 46,477-480.
15. Hansen, M.H. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.,* 14, 333-362.
16. Hansen, M.H. and Hurwitz, W.N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*. Vol. 2, John Wiley and Sons, New York.
17. Hartley, H.O. and Ross, A. (1954). Unbiased ratio estimators. *Nature,* 174, 270-271.
18. Hidiroglou, M.A. and Sarndal, C.E. (1995), Use of auxiliary information for two-phase sampling. *Proceedings of the section on Survey Research Methods, American Statistical Association*, Vol. II. 873-878.
19. Hidiroglou, M.A. and Sarndal, C.E. (1998), Use of auxiliary information for two phase sampling. *Survey Methodology*. 24(1), 11-20.
20. Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika.* 27, 217-223.
21. Kiregyera, B. (1984). Regression-type estimator using two auxiliary variables and model of double sampling from finite populations. *Metrika*. 31, 215-226.
22. Kish, L. (1965). *Survey Sampling*. Jhon Wiley.
23. Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates. *Bull. Inter. Statist. Inst.*, XX XIII, Book 2, 133-140.
24. Mickey, M.R. (1959). Some finite population unbiased ratio and regression estimators. *J. Amer. Statist. Assoc.*, 54, 594-612.
25. Midzuno, H. (1950). An outline of the theory of sampling systems. *Ann. Inst. Statist. Math*., 1, 52-58.
26. Mohanty, S. (1967) Combination of regression and ratio estimate. *J. Ind. Statist.*, 5, 16-19.

27. Mukerjee, R., Rao, T.J. and Vijayan, K. (1987). Regression type estimators using multiple auxiliary information. *Austral. J. Statist.* 29(3), 244-254.
28. Murthy, M.N. (1964). Product method of estimation. *Sankhya*, 26, 294-307.
29. Murthy, M.N. and Nanjamma, N.S. (1959). Almost unbiased ratio estimates based on interpenetrating sub-sample estimates. *Sankhya*, 21, 381-392.
30. N. Balakrishnan, S. Kotz and N.L. Johnson (2000). Continuous Multivariate Distributions, Vol. 1, Second edition. John Wiley & Sons, New York.
31. Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.*, 97, 558-606.
32. Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika,* 45,154-165.
33. Prasad, B. (1989). Some improved ratio-type estimators of population mean and ratio in finite population sample surveys. *Comm. Statist., Theory and Methods*. 18, 379-392.
34. Quenouille, M.H. (1956). Note on Bias in Estimation. *Biometrika*, 43, 353-360.
35. Raj, D. (1968). *Sampling Theory.* Tata McGraw Hill.
36. Rao, J.N.K. (1973). On double sampling for stratification and analytic surveys. *Biometrika,* 6, 125-133.
37. Rao, P.S.R.S. and Rao, J.N.K. (1971). Small sample results for ratio estimators. *Biometrika,* 58, 625-630.
38. Rao, P.S.R.S. (1969). Comparison of four ratio type estimators under a model. *J. Amer. Statist. Assoc.,* 64, 574-580.
39. Rao, T.J. (1983). *A new class of unbiased product estimators.* Stat Math. Tech. Report No.15/83, I.S.I., Calcutta India.
40. Rao, T.J. (1987). On a certain unbiased product estimators. *Comm. Statist. Theo. Meth.,* 16, 3631-3640.
41. Robson, D.S. (1957). Applications of multivariate polykays to the theory of unbiased ratio-type estimation. *J. Amer. Statist. Assoc.*, 52, 511-522.
42. Roy, D.C. (2003). A regression type estimates in two-phase sampling using two auxiliary variables. *Pak. J. Statist.* 19(3), 281-290.
43. Royall, R.M. (1970). On finite population theory under certain linear regression models. *Biometrika,* 57, 377-387.
44. Sahoo, J. and Sahoo, L.N. (1993). A class of estimators in two-phase sampling using two auxiliary variables. *J. Ind. Soc. Agri. Statist.* 31, 107-114.
45. Sahoo, J. and Sahoo, L.N. (1994). On the efficiency of four chain-type estimators in two-phase sampling under a model. *Statistics.* 25, 361-366.
46. Samiuddin, M. and Hanif,M. (2007). Estimation of population mean in single phase and two phase sampling with or without additional information. *Pak. J. Statist.* 23(2), 99-118.
47. Sarndal, C.E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *Inter. Statist. Rev.,* 55,279-294.
48. Searls, D.T. (1964). The utilization of a known coefficient of variation in the estimation procedure. *J. Amer. Statist. Assoc*. 59, 1225-1226.

49. Sen, A.R. (1953). On the estimator of the variance in sampling with varying probabilities. *J. Ind. Soc. Agri. Statist*, 5, 119-127.
50. Singh, H.P. (1987). Class of almost unbiased ratio and product type estimators for finite population mean applying Quenouille's method. *J. Ind. Soc. Agri. Statist*., 39, 280-288.
51. Singh, G.N. and Upadhyaya, L.N. (1995). A class of modified chain type estimators using two auxiliary variables in two phase sampling. *Metron.* 1. III, 117-125.
52. Singh, H.P. and Espejo, M.R. (2003). On linear regression and ratio-product estimation of a finite population mean. *The Statistician.* 52, 59-67.
53. Singh, H.P. and Espejo, M.R. (2007). Double sampling ratio-product estimator of a finite population mean in sample surveys. *J. Appl. Statist.* 34, 71-85.
54. Singh, R., Chauhan, P. and Sawan, N. (2007). On the bias reduction in linear variety of alternative to ratio-cum-product estimator. Statistics in Transition, 8(2), 293-300.
55. Singh, H.P., Upadhyaya, L.N. and Chandra, P. (2004). A General family of estimators for estimating population means using two auxiliary variables in two phase sampling. *Statistics in Transition.* 6, 1055-1077.
56. Singh, H.P., Singh, S. and Kim, J. (2006). General families of chain ratio type estimators of the population mean with known coefficient of variation of the second auxiliary variable in two phase sampling. *J. Korean Statist. Soc.* 35(4), 377-395.
57. Srivastava, S.K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statist. Assoc. Bull.* 16, 121-132.
58. Srivenkataramana, T. and Tracy, D.S. (1979). On ratio and product methods of estimation in sampling. *Statist. Neerlandica*, 33, 37-49.
59. Srivenkataramana, T. (1980). A dual to ratio estimator in sample surveys. *Biometrika,* 67, 199-204.
60. Srivenkataramana, T. and Tracy, D.S. (1981). Extending product method of estimation to positive correlation case in surveys. *Austral. J. Statist.,* 23, 95-100.
61. Sukhatme, P.V. and Sukhatme, B.V., Sukhatme, S. and Ashok, C. (1984). *Sampling Theory of Surveys with Applications*. University Press, Ames, Iowa, USA.
62. Tin, M. (1965). Comparison of some ratio estimators. *J. Amer. Statist. Assoc.*, 60, 294-307.
63. Tripathi, T.P. (1970). *Contributions to the sampling theory in multivariate information*. Ph.D. Thesis submitted to Punjabi University, Patiala, India.
64. Tripathi, T.P. (1973). Double sampling for inclusion probabilities and regression method of estimation. *J. Ind. Statist. Assoc.,* 10, 33-46.
65. Tripathi, T.P. (1976). On double sampling for multivariate ratio and difference method of estimation. *J. Ind. Statist. Assoc.,* 33, 33-54.
66. Tripathi, T.P., Singh, H.P. and Upadhyaya, L. N (1988). A generalized method of estimation in double sampling. *J. Ind. Statist. Assoc.,* 26, 91-101.
67. Watson, D.J. (1937). The estimation of lead areas. *J. Agri. Sci.*, 27, 475.
68. Williams, W.H. (1963). The precision of some unbiased regression estimators. *Biometrics*, 19(2), 352-361.
69. Yates, F. (1960). *Sampling Methods for Censuses and Surveys*, 2nd Edition (London: Charles Griffin).

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 103-108

SOME EXPONENTIAL CHAIN PRODUCT TYPE ESTIMATOR FOR POPULATION MEAN IN TWO PHASE SAMPLING

Muhammad Noor-ul-Amin and **Muhammad Hanif**
National College of Business Administration and Economics, Lahore, Pakistan
Email: nooramin.stats@gmail.com; drhanif@ncbae.edu.pk

ABSTRACT

In this paper, exponential product type estimator has been proposed for estimating the finite population mean of the study variable using the information of two auxiliary variables. The expression for mean square error of the proposed estimator has been derived. It is shown that the proposed estimator is more efficient as compared to the sample mean estimator and Singh and Vishwakarma's (2007) estimator. Empirical study has also been carried out to demonstrate the performance of proposed estimators.

KEY WORDS AND PHRASES

Ratio-cum-ratio; product-cum-product; exponential type estimators; population mean; auxiliary variables; mean square errors; biases.

AMS (1991) *subject classification.* 62D05

1. INTRODUCTION

The efficiency of estimator can be improved considering the auxiliary information. In Double sampling we usually obtained the estimates for the parameter of auxiliary variable at the first-phase and estimates for variable of interest is obtained at second-phase, such as Sukhatme (1962), Cochran (1963), Mohanty (1967), Khare and Srivastava (1981), Mukerjee et al. (1987), Chand (1975), Kiregyera (1980, 84), Sahoo et al. (1993), Hidiroglou (2001), Roy (2003), Singh and Vishwakarma (2007), Singh et al. (2007), and Samiuddin and Hanif (2007).

Let us consider the finite population of size N and let $\bar{Y} = \sum_{i=1}^{N} Y_i \Big/ N$, $\bar{X} = \sum_{i=1}^{N} X_i \Big/ N$ and $\bar{Z} = \sum_{i=1}^{N} Z_i \Big/ N$ are the population means of the variable y, x and z respectively. The sample at the first phase is drawn by simple random sampling without replacement of size 'n_1' from the population and we assume that $\bar{x}_1 = \sum_{i=1}^{n_1} x_i \Big/ n_1$ be the sample mean of variable 'x' for the first phase sample, where $\bar{y}_2 = \sum_{i=1}^{n_2} y_i \Big/ n_2$, $\bar{x}_2 = \sum_{i=1}^{n_2} x_i \Big/ n_2$ and

$\bar{z}_2 = \sum_{i=1}^{n_2} z_i \Big/ n_2$ are the means of variable 'y', 'x' and 'z' respectively for the sample obtained at second phase using simple random sampling without replacement of size 'n_2'. We have the following assumptions which have also been used by Singh et al. (2008).

- If second phase sample is not independent from the first phase sample.

$$\left.\begin{array}{l} e_{\bar{y}_2} = \dfrac{\bar{y}-\bar{Y}}{\bar{Y}} \,,\; e_{\bar{x}_1} = \dfrac{\bar{x}-\bar{X}}{\bar{X}} \\ \text{E}(e_{\bar{y}_2})=\text{E}(e_{\bar{x}_1})=\text{E}(e_{\bar{z}_2})=0, \text{E}(e^2{}_{\bar{y}_2}) = \theta_2 C_y^2 \,, \text{E}(e_{\bar{x}_1}^2) = \theta_1 C_x^2 \,, \\ \text{E}(e_{\bar{y}_2}\; e_{\bar{x}_1}) = \theta_1\; \rho_{xy} C_x C_y \;\;, \text{E}(e_{\bar{y}_2}\; e_{\bar{z}_2}) = \theta_2\; \rho_{yz} C_z C_y \,, \\ \theta_1 = \dfrac{1}{n_1} - \dfrac{1}{N} \,, \theta_2 = \dfrac{1}{n_2} - \dfrac{1}{N} \;\,, C_y = \dfrac{S_y}{\bar{Y}} \,, \rho_{xy} = \dfrac{S_{xy}}{S_x S_y} \,,\; H_{ij} = \rho_{ij} \dfrac{C_i}{C_j} \\ S_x = \sqrt{\dfrac{\sum(x-\bar{x})^2}{n}} \,, S_y = \sqrt{\dfrac{\sum(y-\bar{y})^2}{n}} \,, S_{xy} = \sqrt{\dfrac{\sum(x-\bar{x})(y-\bar{y})}{n}} \end{array}\right\} \quad (1.1)$$

- If the second phase sample is independent from the first phase sample. Then the only difference is in covariance terms, i.e.

$$\text{E}(e_{\bar{y}_2}\; e_{\bar{x}_1}) = 0 \;, \text{E}(e_{\bar{z}_2}\; e_{\bar{x}_1}) = 0 \;, \text{E}(e_{\bar{y}_2}\; e_{\bar{z}_2}) = \theta_2\; \rho_{xy} C_z C_y \;\; \text{etc.} \quad (1.2)$$

2. SOME AVAILABLE EXPONENTIAL-TYPE ESTIMATORS

Some popular Exponential-Type estimators for the population mean are available as:

- Bahl and Tuteja's (1991) suggested exponential product-type estimator for single phase sampling as:

$$t_1 = \bar{y}\;\; \exp\left[\frac{\bar{x}-\bar{X}}{\bar{X}+\bar{x}}\right]. \quad (2.1)$$

The mean square error of t_1 is

$$\text{MSE}(t_1) \approx \bar{Y}^2 \theta \left[C_y^2 + \frac{C_x^2}{4}(1+4H_{yx}) \right]. \quad (2.2)$$

The following exponential type estimators of Double sampling are reproduced

- Singh and Vishwakarma's (2007) suggested exponential product-type estimator as

$$t_2 = \bar{y}_2 \exp\left[\frac{\bar{x}_2 - \bar{x}_1}{\bar{x}_1 + \bar{x}_2}\right]. \quad (2.3)$$

The mean square error of t_2 is

$$\text{MSE}\,(t_2) \approx S_y^2\left[\theta_2 + \frac{C_x}{4C_y}(\theta_2-\theta_1)\left\{\frac{C_x}{C_y} - 4\rho_{xy}\right\}\right]. \tag{2.4}$$

t_2 is modified form of t_1, Singh and Vishwakarma in (2007) has proved that the performance of t_2 is better than t_1.

3. PROPOSED ESTIMATORS

In this section, an product exponential type estimator for two auxiliary variables has been proposed. The mean square error of proposed estimators will be derived as

$$t_3 = \bar{y}_2 \exp\left[\frac{\bar{x}_1 - \bar{X}}{\bar{X}+\bar{x}_1} + \frac{\bar{z}_2 - \bar{Z}}{\bar{Z}+\bar{z}_2}\right]. \tag{3.1}$$

Using the notations given in (1.1), t_4 (by neglecting the terms with power two or greater) may be written as

$$t_3 = \bar{Y}(1+e_{\bar{y}_2})\exp\left[\frac{e_{\bar{x}_1}}{2}\left(1-\frac{e_{\bar{x}_1}}{2}\right) + \frac{e_{\bar{z}_2}}{2}\left(1-\frac{e_{\bar{z}_2}}{2}\right)\right]. \tag{3.2}$$

Expanding the right hand side and neglecting the terms with power two or greater, we get

$$t_3 - \bar{Y} = \bar{Y}\left[e_{\bar{y}_2} + \frac{e_{\bar{x}_1}}{2} + \frac{e_{\bar{z}_2}}{2}\right]. \tag{3.3}$$

Squaring both sides and taking the expectation we get MSE of t_4

$$\text{MSE}\,(t_3) = E\left[t_3 - \bar{Y}\right]^2 = \bar{Y}^2 E\left[e_{\bar{y}_2} + \frac{e_{\bar{x}_1}}{2} - \frac{e_{\bar{z}_2}}{2}\right]^2. \tag{3.4}$$

Squaring both sides, taking the expectation and on simplification we get

$$\text{MSE}\,(t_3) \approx \bar{Y}^2\left[\theta_2\left(C_y^2 + \frac{C_z^2}{4}\left(1+4H_{yz}\right)\right) + \frac{\theta_1 C_x^2}{4}\left(1+4H_{yx}+2H_{zx}\right)\right]. \tag{3.5}$$

4. EMPIRICAL STUDY

The following population has been used for the purpose of comparison. The percent relative efficiencies of t_2 and t_3 (estimators for two-phase sampling) based on the sample mean estimator, $\bar{y}$, are presented in Table 1. The description of populations in this table is as follows:

Population I: [Source: Gujarati (2004, Page 433)]

Y: Average miles per gallon

X: Top speed, miles per hour

Z: Cubic feet of cab space

The required parameters of the population are:

$$\bar{Y} = 33.83457, \quad C_y^2 = 0.088324, \quad \rho_{yx} = -0.69079, \quad N = 81.$$
$$\bar{X} = 112.4568, \quad C_x^2 = 0.015765, \quad \rho_{yz} = -0.36831, \quad n_1 = 35.$$
$$\bar{Z} = 98.76543, \quad C_z^2 = 0.050987, \quad \rho_{zx} = -0.04265, \quad n_1 = 4.$$

Population II: [Source: Gujarati (2004, Page 433)]

Y: Average miles per gallon

X: Engine horsepower

Z: Cubic feet of cab space

The required parameters of the population are:

$$\bar{Y} = 33.83457, \quad C_y^2 = 0.088324, \quad \rho_{yx} = -0.79445, \quad N = 81.$$
$$\bar{X} = 117.47, \quad C_x^2 = 0.236391, \quad \rho_{yz} = -0.36831, \quad n_1 = 31.$$
$$\bar{Z} = 98.76543, \quad C_z^2 = 0.050987, \quad \rho_{zx} = 0.831815, \quad n_2 = 4.$$

Table 1
Percent relative efficiencies of different estimators for the population mean with respect to the sample mean

Estimators	Populations	
	I	II
$\bar{y}$	100	100
t_2	96.84	61.84
t_3	115.28	230.87

5. CONCLUSION

The efficiency comparison of estimators is given in Table 1. From Table 1, it is observed that the suggested estimators are more efficient than sample mean estimator and Singh and Vishwakarma’s (2007) estimator.

REFERENCE

1. Bahl, S. and Tuteja, R.K. (1991). Ratio and product type exponential estimator. *Information and Optimization Sciences*, 12, 159-163.
2. Chand, L. (1975). *Some Ratio-type Estimators based on two or more Auxiliary Variables.* Unpublished Ph.D. Dissertation. Iowa State University, Iowa.
3. Cochran, W.G. (1977). *Sampling Techniques.* John Wiley, New York.
4. Hidiroglou, M.A. (2001). Double Sampling. *Survey Methodology*, 27, 143-154.
5. Khare, B.B. and Srivastava, S.R. (1981). A general regression ratio estimator for the population mean using two auxiliary variables. *Alig. J. Statist.*, 1, 43-51.
6. Kiregyera, B. (1980). A Chain Ratio-Type Estimator in Finite Population Double Sampling using two Auxiliary Variables. *Metrika*, 27, 217-223.
7. Kiregyera, B. (1984). A Regression-Type Estimator using two Auxiliary Variables and Model of Double sampling from Finite Populations. *Metrika*, 31, 215-226.
8. Mohanty, S. (1967). Combination of Regression and Ratio Estimate. *Jour. Ind. Statist. Assoc.,* 5, 16-19.
9. Mukerjee, R. Rao, T.J. and Vijayan, K. (1987). Regression type estimators using multiple auxiliary information. *Austral. J. Statist.,* 29(3), 244-254.
10. Roy, D.C. (2003). A regression type estimator in Double sampling using two auxiliary variables. *Pak. J. Statist*., 19(3), 281-290.
11. Sahoo, J., Sahoo, L.N. and Mohanty, S. (1993). A Regression Approach to Estimation in Double Sampling using two Auxiliary Variables. *Current Sciences*, 65, 1, 73-75.
12. Samiuddin, M. and Hanif, M. (2007). Estimation of population mean in single phase and Double sampling with or without additional information. *Pak. J. Statist.*, 23(2), 99-118.
13. Singh, P. and Vishwakarma, K. (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. *Austral. J. Statist*., 36, 217-225.
14. Sukhatme, B.V. (1962). Some ratio type estimators in Double sampling. *J. Amer. Statist. Assoc.*, 57, 628-632.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 109-120

ON BAYESIAN RELIABILITY FUNCTION ESTIMATION OF BURR-XII DISTRIBUTION

Dhwyia Salman Hassan[1], **Nashaat Jaisam Al-Anber**[2]
and **Abdulmunem Kadhim Hammadi**[3]

[1] Department of Statistics, College of Economic and Administration,
University of Baghdad, Baghdad, Iraq. Email: dhwyia.salman@yahoo.com

[2] Department of Information Technology, Technical College of Management,
Baghdad Foundation of Technical Education, Baghdad, Iraq.
Email: nashaatg74@yahoo.com; nashaatg74@hotmail.com

[3] Ministry of Higher Education and Scientific Research, Baghdad, Iraq

ABSTRACT

The Burr-XII distribution has been widely used especially in the modeling of life time event data. It provides a statistical model which has a wide variety of application in many areas and the main advantage is its ability in the context of life time event among other distributions. The conventional maximum likelihood method is the usual way to estimate the parameters of a distribution .Bayesian approach has received much attention in contention with other estimation methods. In this study we consider the bayes estimation of the unknown parameters, when both parameters (α,β) are unknown. It is assumed that the unknown parameters have gamma prior distributions.

It is not possible to compute analytically the double integration used to find bayes estimator under squared error loss function, so we use lindley's approximation to approximate the double integration and find the bayes estimate of unknowns parameter and reliability function. We also explore the performance of bayes estimator of reliability function with maximum likelihood estimator numerically under varying values of parameters, sample sizes and hyper parameters. Through the simulation study a comparison is made on the performance of these estimators with respect to the integrated mean square error (IMSE) and integrated mean absolute percentage error (IMAPE).on the results of this simulation study the Bayesian approach used in the estimating of reliability function for the two parameter Burr-XII distribution is found to be superior compared to the conventional Maximum Likelihood Method with respect to both IMSE and IMAPE values.

KEY WORDS

Burr-XII distribution;bayes Method; lindley's approximation; Maximum likelihood Method; Reliability Function.

1. INTRODUCTION

Reliability theory is mainly concerned with the determination of the probability that a system, consisting possibly of several components, will operate adequately for a given period of time in its intended application. The Reliability function $R(t)$ is defined as the probability that system will operate at time period t. The two parameter Burr-XII distribution was first introduced in the literature by Burr (1942)[2],and has gained special attention in the last two decades due to the importance of using it in practical situations especially in reliability studies and failure time modeling since the Burr-XII distribution has a non-monotone hazard function, which can accommodate many shapes of hazard function. Lewis[9] stated the fact that many standard theoretical distributions, including the two common failure time distributions, the Weibull and the exponential are special cases or limiting cases of the Burr-XII distribution. Evans and Ragab[7] obtained Bayes estimates for shape parameter and reliability function on type-II censored samples. Mousa[11]obtained empirical Bayes estimation of the shape parameter and reliability function based on accelerated type-II censored data. Moore and Papadopoulos[3] obtained bayes estimates for shape parameter and the reliability function when the other shape parameter was assumed to be known. Mousa and Jaheen[10] obtained bayes approximate estimates for the two parameters and reliability function based on progressive type-II censored samples. Based on the same progressive samples as above, Soliman[1] obtained the bayes estimate using both the symmetric and asymmetric loss functions. The usefulness and properties of the Burr-XII distribution as a failure model were discussed in many papers[5],[4],[6]. Nasir and AL-Anber[12] carried out a comparison between maximum likelihood estimator and Bayesian estimator of the parameter and Reliability function for the Burr-XII distribution under Jeffrey prior and extension of Jeffrey prior.

The probability density function of two parameter Burr-XII distribution becomes:

$$f(t;\alpha,\beta) = \alpha\beta t^{\alpha-1}(1+t^{\alpha})^{-\beta-1} \ \ ; \ t,\alpha,\beta > 0 \tag{1}$$

where α,β are the shape parameters of the distribution. The reliability function is given by:

$$\begin{aligned} R(t) &= \Pr \ \ T > t \ \ = 1 - F(t) = \int_{t}^{\infty} f(u;\alpha,\beta)du \\ &= (1+t^{\alpha})^{-\beta} \end{aligned} \tag{2}$$

And the hazard function become

$$h(t) = \frac{f(t)}{R(t)} = \frac{\alpha\beta t^{\alpha-1}}{(1+t^{\alpha})} \tag{3}$$

which is obviously depends non-monotonically on failure time.

The rest of the study is arranged as follow. In Methods, maximum likelihood estimator and Bayes estimator of reliability function are presented. In Results, simulation study is discussed and the results are presented and followed by conclusion.

2. MATERIALS AND METHODS

Maximum likelihood Estimation (M.L.E): we introduce the concept of maximum likelihood estimation with two parameter Burr-XII distribution. We have set of random failure times ($t_1,t_2,...,t_n$) and vectors of unknown parameters $\theta=(\theta_1,...,\theta_k)$, then the likelihood function is:

$$L(t;\theta)=\prod_{i=1}^{n} f(t_i;\theta) \tag{4}$$

The ith elements of the score vector with respect to θ is:

$$U_i(\theta)=\frac{\partial \ln L(t,\theta)}{\partial\theta_i} \qquad ,i=1,...,k \tag{5}$$

Now we can find the maximum likelihood estimator by using two parameter Burr-XII distribution with Parameters α,β .The probability density function of two parameters Burr-XII distributions is given by:

$$f(t;\alpha,\beta)=\alpha\beta t^{\alpha-1}(1+t^{\alpha})^{-\beta-1} \quad ; \ t,\alpha,\beta>0 \tag{6}$$

The likelihood function is:

$$L(t_1,t_2,...,t_n;\alpha,\beta)=\prod_{i=1}^{n} f(t_i;\alpha,\beta) \tag{7}$$

$$=\alpha^n\beta^n\prod_{i=1}^{n} t_i^{\alpha-1}\prod_{i=1}^{n}(1+t_i^{\alpha})^{-\beta-1}$$

The scores vector is:

$$U_1(\theta)=\frac{\partial \ln L(t;\alpha,\beta)}{\partial\beta} \ =\frac{n}{\beta}-\sum_{i=1}^{n}\log(1+t_i^{\alpha}) \tag{8}$$

$$U_2(\theta)=\frac{\partial \ln L(t;\alpha,\beta)}{\partial\beta} \ = \ \frac{n}{\alpha}+\sum_{i=1}^{n}\log(t_i)-(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}\log(t_i)}{(1+t_i^{\alpha})} \tag{9}$$

Let $U(\theta)=0$, then the maximum likelihood estimator of α and β are the solution of the two nonlinear equation below :

$$\frac{n}{\beta}-\sum_{i=1}^{n}\log(1+t_i^{\alpha})=0 \tag{10}$$

$$\frac{n}{\alpha}+\sum_{i=1}^{n}\log(t_i)-(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}\log(t_i)}{(1+t_i^{\alpha})}=0 \tag{11}$$

Since the maximum likelihood estimator is invariant and one to one mapping. The maximum likelihood estimator of Reliability function is:

$$\hat{R}_{m.l.e}(t)=(1+t^{\hat{\alpha}_{m.l.e}})^{-\hat{\beta}_{m.l.e}} \tag{12}$$

Bayes estimation (Bay): In this section we consider the bayes estimation of the unknown parameters .when both α and β are unknown .it is assumed that α and β have the following gamma prior distributions:

$$\pi_1(\alpha) \propto \alpha^{a-1} e^{-b\alpha} \tag{13}$$

$$\pi_2(\beta) \propto \beta^{c-1} e^{-d\beta} \tag{14}$$

where $a, b, c, d, \alpha, \beta > 0$

Here all the hyper parameters a, b, c, d are assumed to be known and non-negative. Suppose $(x_1, x_2, ..., x_n)$ is a random sample from Burr-XII distribution. then based on the likelihood function of the observed data the joint posterior density function of α and β can be written as

$$L(\alpha, \beta \setminus t_1, t_2, ..., t_n) = \frac{L(t_1, t_2, ..., t_n; \alpha, \beta).\pi_1(\alpha)\pi_2(\beta)}{\int_0^\infty \int_0^\infty L(t_1, t_2, ..., t_n; \alpha, \beta).\pi_1(\alpha)\pi_2(\beta) d\alpha d\lambda} \tag{15}$$

Therefore the bayes estimator of reliability function under squared error loss function is:

$$\hat{R}_{bay}(t) = E_{\alpha,\beta\setminus t}\left[(1+t^\alpha)^{-\beta}\right] = \frac{\int_0^\infty \int_0^\infty (1+t^\alpha)^{-\beta} L(t_1, t_2, ..., t_n; \alpha, \beta).\pi_1(\alpha)\pi_2(\beta) d\alpha d\lambda}{\int_0^\infty \int_0^\infty L(t_1, t_2, ..., t_n; \alpha, \beta).\pi_1(\alpha)\pi_2(\beta) d\alpha d\lambda} \tag{16}$$

It is not possible to compute analytically the above double-integration. In this case we use lindley's approximation[8] to approximate the above integration.

Lindley proposed his procedure to approximate the ratio of the two integrals such as above integral .Based on lindley's approximation, the approximate Bayes estimators of reliability function under squared error loss function is:

$$\hat{R}_{bay}(t) = (1+t^{\hat{\alpha}_{mle}})^{-\hat{\beta}_{mle}} + \frac{1}{2}\left[A + l_{30}B_{12} + l_{03}B_{21} + l_{21}C_{12} + l_{12}C_{21}\right] + P_1 A_{12} + P_2 A_{21} \tag{17}$$

where

$$A = \sum_{i=1}^{2} \sum_{j=1}^{2} W_{ij} \tau_{ij}$$

Let $\lambda_1 = \alpha, \lambda_2 = \beta$

$$l_{ij} = \frac{\partial^{i+j} L(\alpha, \beta)}{\partial \lambda_1{}^i \partial \lambda_2{}^j} \quad ; \ i, j = 0, 1, 2, 3 \quad ; \ i + j = 3$$

$$P_i = \frac{\partial \ln \pi(\alpha, \beta)}{\partial \lambda_i}$$

$$w_i = \frac{\partial R(t)}{\partial \lambda_i}$$

$$W_{ij} = \frac{\partial^2 R(t)}{\partial \lambda_i \partial \lambda_j}$$

$$A_{ij} = w_i\tau_{ii} + w_j\tau_{ji}$$
$$B_{ij} = (w_i\tau_{ii} + w_j\tau_{ij})\tau_{ii}$$
$$C_{ij} = 3w_i\tau_{ii}\tau_{ij} + w_j(\tau_{ii}\tau_{jj} + 2\tau_{ij}^2)$$

Here:

$L(.,.)$: is the log likelihood function of the observed data.

$\pi(\lambda_1,\lambda_2)$:is the joint prior density function of (λ_1,λ_2)

τ_{ij} : is the $(i, j)^{th}$ element of the inverse of fisher information matrix.

$\hat{\lambda_1},\hat{\lambda_2}$: are the maximum likelihood estimator of λ_1 and λ_2 respectively and all the quantities are evaluated at $(\hat{\lambda_1},\hat{\lambda_2})$

Now we have

$$P_1 = \frac{a-1}{\alpha} - b$$
$$P_2 = \frac{c-1}{\beta} - d$$
$$w_1 = -\beta(1+t^\alpha)^{-\beta-1}(t^\alpha \log(t))$$
$$w_2 = -(1+t^\alpha)^{-\beta}\log(1+t^\alpha)$$
$$W_{11} = \left[-\beta(1+t^\alpha)^{-\beta-1}\right]\left[t^\alpha(\log t)^2\right] + \left[t^\alpha \log t\right]\left[\beta(\beta-1)(1+t^\alpha)^{-\beta-2}t^\alpha \log t\right]$$
$$W_{12} = \left[\beta(1+t^\alpha)^{-\beta-1}\right]\left[\log(1+t_i^\alpha)\right]\left[t^\alpha \log t\right]$$
$$W_{21} = -(1+t^\alpha)^{-\beta}\frac{t^\alpha \log t}{(1+t^\alpha)} + \log(1+t^\alpha)(\beta(1+t^\alpha)^{-\beta-1}t^\alpha \log t)$$
$$W_{22} = (1+t^\alpha)^{-\beta}(\log(1+t^\alpha))^2$$
$$A = W_{11}\tau_{11} + W_{12}\tau_{12} + W_{21}\tau_{21} + W_{22}\tau_{22}$$
$$B_{12} = (w_1\tau_{11} + w_2\tau_{12})\tau_{11}$$
$$B_{21} = (w_2\tau_{22} + w_1\tau_{21})\tau_{22}$$
$$C_{12} = 3w_1\tau_{11}\tau_{12} + w_2(\tau_{11}\tau_{22} + 2\tau_{12}^2)$$
$$C_{21} = 3w_2\tau_{22}\tau_{21} + w_1(\tau_{22}\tau_{11} + 2\tau_{21}^2)$$
$$A_{12} = w_1\tau_{11} + w_2\tau_{21}$$
$$A_{21} = w_2\tau_{22} + w_1\tau_{12}$$
$$l_{30} = \frac{\partial^3 L(\alpha,\beta)}{\partial\alpha^3} = \frac{2n}{\alpha^3} - (\beta+1)\sum_{i=1}^{n}\frac{(\log t)^3 t^\alpha(1-t^\alpha)}{(1+t^\alpha)^3}$$
$$l_{03} = \frac{\partial^3 L(\alpha,\beta)}{\partial\beta^3} = \frac{2n}{\beta^3}$$

$$l_{12}=\frac{\partial^3 L(\alpha,\beta)}{\partial\alpha\partial\beta^2}=0$$

$$l_{21}=\frac{\partial^3 L(\alpha,\beta)}{\partial\alpha^2\partial\beta}=-\sum_{i=1}^{n}\frac{t_i^{\alpha}(\log t_i)^2}{(1+t_i^{\alpha})^2}$$

$$\tau_{11}=\frac{(n/\beta^2)}{\left[(\frac{n}{\alpha^2})+(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}(\log t_i)^2}{(1+t^{\alpha})^2}\right]\left[\frac{n}{\beta^2}\right]-\left[\sum_{i=1}^{n}\frac{t_i^{\alpha}\log t_i}{(1+t_i^{\alpha})}\right]^2}$$

$$\tau_{12}=\tau_{21}=\frac{\sum_{i=1}^{n}\frac{t_i^{\alpha}\log t_i}{(1+t_i^{\alpha})}}{\left[(\frac{n}{\alpha^2})+(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}(\log t_i)^2}{(1+t^{\alpha})^2}\right]\left[\frac{n}{\beta^2}\right]-\left[\sum_{i=1}^{n}\frac{t_i^{\alpha}\log t_i}{(1+t_i^{\alpha})}\right]^2}$$

$$\tau_{22}=\frac{(\frac{n}{\alpha^2})+(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}(\log t_i)^2}{(1+t^{\alpha})^2}}{\left[(\frac{n}{\alpha^2})+(\beta+1)\sum_{i=1}^{n}\frac{t_i^{\alpha}(\log t_i)^2}{(1+t^{\alpha})^2}\right]\left[\frac{n}{\beta^2}\right]-\left[\sum_{i=1}^{n}\frac{t_i^{\alpha}\log t_i}{(1+t_i^{\alpha})}\right]^2}$$

3. RESULTS

In this simulation study, we have chosen $n=10,15,25,50$ and 100 to represent small, moderate and large sample size, several values of parameter $\alpha=0.5,1,2$ and $\beta=0.5,1,2$, three sets of Hyber parameters $a,b,c,d=0,0.5,1$.The number of replication used was (L=1000).The simulation program was written by using matlab-R2010b program. After the Reliability function was estimated, integrated mean square error (IMSE) and integrated mean absolute percentage error (IMAPE) was calculated to compare the methods of estimation, Where:

$$IMSE(R^{\wedge}(t))=\frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{n_t}\sum_{j=1}^{n_t}(R_i\,(t_j)-R_i^{\wedge}(t_j))^2\right\}$$

$$=\frac{1}{L}\sum_{i=1}^{L}MSE(R_i^{\wedge}(t)) \qquad (18)$$

$$IMAPE(R^{\wedge}(t))=\frac{1}{L}\sum_{i=1}^{L}\left\{\frac{1}{n_t}\sum_{j=1}^{n_t}\left|\frac{R_i\,(t_j)-R_i^{\wedge}(t_j)}{R_i\,(t_j)}\right|\right\}$$

$$=\frac{1}{L}\sum_{i=1}^{L}MAPE(R_i^{\wedge}(t)) \qquad (19)$$

The results of the simulation study are summarized and tabulated in table 1 of the two estimators for all sample size and parameters values respectively.

Table 1
IMSE and IMAPE for different estimators of reliability function for Burr-IXX distribution

N	α ha	beta	A	b	c	d	IMSE		Best method	IMPE		Best method
							mle	bay		mle	bay	
10	0.5	0.5	0	0	0	0	0.015323	0.118233	mle	0.133317	0.152939	mle
15							0.013307	0.012391	bay	0.107765	0.10226	bay
25							0.011622	0.011269	bay	0.098454	0.096391	bay
50							0.002295	0.00226	bay	0.052959	0.052828	bay
100							0.000999	0.001002	mle	0.03604	0.035989	bay
10	0.5	1	0	0	0	0	0.013257	0.011732	bay	0.175394	0.165461	bay
15							0.01061	0.009928	bay	0.162724	0.157361	bay
25							0.005005	0.004784	bay	0.109951	0.107637	bay
50							0.002718	0.002656	bay	0.086705	0.085682	bay
100							0.001771	0.00175	bay	0.065613	0.065232	bay
10	0.5	2	0	0	0	0	0.011969	0.010775	bay	0.338636	0.318468	bay
15							0.009082	0.008475	bay	0.302343	0.290769	bay
25							0.004385	0.004183	bay	0.202922	0.198698	bay
50							0.002927	0.002873	bay	0.174365	0.172415	bay
100							0.001175	0.001159	bay	0.10595	0.105243	bay
10	1	0.5	0	0	0	0	0.010985	0.0621	mle	0.121765	0.146029	mle
15							0.007065	0.006595	bay	0.097334	0.093321	bay
25							0.004242	0.004086	bay	0.07657	0.074871	bay
50							0.001823	0.001808	bay	0.04835	0.048003	bay
100							0.00104	0.001007	bay	0.037269	0.036664	bay
10	1	1	0	0	0	0	0.016631	0.014717	bay	0.218698	0.204881	bay
15							0.011225	0.010389	bay	0.173998	0.166801	bay
25							0.00481	0.004594	bay	0.117917	0.115192	bay
50							0.002369	0.002303	bay	0.081603	0.080575	bay
100							0.001397	0.001376	bay	0.061326	0.060968	bay
10	1	2	0	0	0	0	0.009369	0.008615	bay	0.320123	0.311034	bay
15							0.009981	0.009341	bay	0.326497	0.311051	bay
25							0.004533	0.004361	bay	0.214948	0.210151	bay
50							0.002312	0.002269	bay	0.156088	0.153913	bay
100							0.001244	0.001232	bay	0.12058	0.119638	bay
10	2	0.5	0	0	0	0	0.00997	0.389034	mle	0.122387	0.244156	mle
15							0.006978	0.050494	mle	0.098469	0.132092	mle
25							0.00413	0.004025	bay	0.076653	0.075168	bay
50							0.001688	0.001585	bay	0.049071	0.047172	bay
100							0.000855	0.00085	bay	0.034797	0.034844	mle

N	α ha	beta	A	b	c	d	IMSE		Best method	IMPE		Best method
							mle	bay		mle	bay	
10	2	1	0	0	0	0	0.014133	0.198476	mle	0.238188	0.323526	mle
15							0.007585	0.006943	bay	0.167951	0.1591	bay
25							0.004441	0.004187	bay	0.134856	0.130177	bay
50							0.001981	0.00195	bay	0.08884	0.087864	bay
100							0.001002	0.000984	bay	0.062696	0.062182	bay
10	2	2	0	0	0	0	0.011628	0.010514	bay	0.461215	0.430348	bay
15							0.006732	0.00635	bay	0.361062	0.349492	bay
25							0.003956	0.003791	bay	0.293496	0.281582	bay
50							0.001726	0.001707	bay	0.19246	0.191744	bay
100							0.000798	0.000788	bay	0.123368	0.122489	bay
10	0.5	0.5	1	1	1	1	0.011665	0.126838	mle	0.122245	0.14944	mle
15							0.010159	0.022154	mle	0.111279	0.106135	bay
25							0.009244	0.008563	bay	0.08778	0.083044	bay
50							0.002317	0.002279	bay	0.053854	0.053687	bay
100							0.001036	0.001023	bay	0.036257	0.035965	bay
10	0.5	1	1	1	1	1	0.014194	0.011193	bay	0.193298	0.171785	bay
15							0.012471	0.010666	bay	0.16862	0.156608	bay
25							0.006238	0.0057	bay	0.12333	0.118023	bay
50							0.002597	0.002488	bay	0.083492	0.081715	bay
100							0.001833	0.001795	bay	0.067257	0.066549	bay
10	0.5	2	1	1	1	1	0.013937	0.010774	bay	0.376321	0.326998	bay
15							0.00695	0.005643	bay	0.251804	0.222818	bay
25							0.005208	0.004829	bay	0.224717	0.218688	bay
50							0.002884	0.002756	bay	0.172491	0.167794	bay
100							0.00103	0.001002	bay	0.097789	0.096217	bay
10	1	0.5	1	1	1	1	0.009888	0.167748	mle	0.110245	0.149499	mle
15							0.006178	0.005791	bay	0.090207	0.086517	bay
25							0.00383	0.003458	bay	0.072124	0.068931	bay
50							0.002007	0.001954	bay	0.051436	0.050667	bay
100							0.00107	0.001021	bay	0.038129	0.03722	bay
10	1	1	1	1	1	1	0.014757	0.011076	bay	0.204438	0.177531	bay
15							0.009682	0.007875	bay	0.160752	0.145623	bay
25							0.005309	0.004807	bay	0.117658	0.111826	bay
50							0.002362	0.002245	bay	0.081856	0.079862	bay
100							0.001329	0.001291	bay	0.059427	0.058706	bay

N	α ha	beta	A	b	c	d	IMSE		Best method	IMPE		Best method
							mle	bay		mle	bay	
10	1	2	1	1	1	1	0.012148	0.009755	bay	0.36011	0.324969	bay
15							0.00759	0.006539	bay	0.285209	0.264352	bay
25							0.004852	0.004371	bay	0.214054	0.202695	bay
50							0.002395	0.002349	bay	0.161065	0.15779	bay
100							0.001231	0.001199	bay	0.120972	0.118406	bay
10	2	0.5	1	1	1	1	0.010848	0.326406	mle	0.121229	0.254122	mle
15							0.006951	0.404502	mle	0.095601	0.179931	mle
25							0.003689	0.003289	bay	0.072892	0.067242	bay
50							0.001621	0.001478	bay	0.0483	0.045946	bay
100							0.00086	0.000838	bay	0.035058	0.034877	bay
10	2	1	1	1	1	1	0.013344	0.195509	mle	0.230578	0.302803	mle
15							0.007814	0.00638	bay	0.180881	0.162389	bay
25							0.004927	0.004308	bay	0.137034	0.127621	bay
50							0.001906	0.001841	bay	0.087061	0.085259	bay
100							0.000975	0.000936	bay	0.06205	0.060883	bay
10	2	2	1	1	1	1	0.008865	0.006462	bay	0.392739	0.346959	bay
15							0.006125	0.005323	bay	0.356582	0.339641	bay
25							0.003821	0.003373	bay	0.267855	0.247899	bay
50							0.001882	0.001825	bay	0.202117	0.201894	bay
100							0.000816	0.000792	bay	0.125694	0.124278	bay
10	0.5	0.5	0.5	0.5	0.5	0.5	0.011665	0.126694	mle	0.122245	0.151377	mle
15							0.010159	0.022537	mle	0.111279	0.109137	bay
25							0.009244	0.008726	bay	0.08778	0.0841	bay
50							0.002317	0.002293	bay	0.053854	0.05382	bay
100							0.001036	0.001026	bay	0.036257	0.036033	bay
10	0.5	1	0.5	0.5	0.5	0.5	0.014194	0.011961	bay	0.193298	0.177485	bay
15							0.012471	0.011073	bay	0.16862	0.159309	bay
25							0.006238	0.005845	bay	0.12333	0.119437	bay
50							0.002597	0.002514	bay	0.083492	0.082134	bay
100							0.001833	0.001803	bay	0.067257	0.066694	bay
10	0.5	2	0.5	0.5	0.5	0.5	0.013937	0.011561	bay	0.376321	0.339766	bay
15							0.00695	0.005945	bay	0.251804	0.229707	bay
25							0.005208	0.00491	bay	0.224717	0.219651	bay
50							0.002884	0.002787	bay	0.172491	0.168967	bay
100							0.00103	0.001009	bay	0.097789	0.096566	bay

N	α ha	beta	A	b	c	d	IMSE		Best method	IMPE		Best method
							mle	bay		mle	bay	
10	1	0.5	0.5	0.5	0.5	0.5	0.009888	0.16649	mle	0.110245	0.149983	mle
15							0.006178	0.005873	bay	0.090207	0.087288	bay
25							0.00383	0.003542	bay	0.072124	0.069661	bay
50							0.002007	0.001967	bay	0.051436	0.050833	bay
100							0.00107	0.001032	bay	0.038129	0.037425	bay
10	1	1	0.5	0.5	0.5	0.5	0.014757	0.011989	bay	0.204438	0.184222	bay
15							0.009682	0.008331	bay	0.160752	0.149421	bay
25							0.005309	0.004941	bay	0.117658	0.113334	bay
50							0.002362	0.002274	bay	0.081856	0.080369	bay
100							0.001329	0.001299	bay	0.059427	0.058895	bay
10	1	2	0.5	0.5	0.5	0.5	0.012148	0.010313	bay	0.36011	0.332022	bay
15							0.00759	0.0068	bay	0.285209	0.26843	bay
25							0.004852	0.004493	bay	0.214054	0.205012	bay
50							0.002395	0.002354	bay	0.161065	0.158525	bay
100							0.001231	0.001207	bay	0.120972	0.119043	bay
10	2	0.5	0.5	0.5	0.5	0.5	0.010848	0.325277	mle	0.121229	0.250471	mle
15							0.006951	0.40414	mle	0.095601	0.178854	mle
25							0.003689	0.003354	bay	0.072892	0.068219	bay
50							0.001621	0.001504	bay	0.0483	0.046266	bay
100							0.00086	0.000842	bay	0.035058	0.034908	bay
10	2	1	0.5	0.5	0.5	0.5	0.013344	0.1966	mle	0.230578	0.311031	Mle
15							0.007814	0.006772	bay	0.180881	0.167085	Bay
25							0.004927	0.004473	bay	0.137034	0.129904	Bay
50							0.001906	0.00186	bay	0.087061	0.08576	Bay
100							0.000975	0.000947	bay	0.06205	0.061186	Bay
10	2	2	0.5	0.5	0.5	0.5	0.008865	0.007128	bay	0.392739	0.350493	Bay
15							0.006125	0.005542	bay	0.356582	0.340216	Bay
25							0.003821	0.00351	bay	0.267855	0.251569	Bay
50							0.001882	0.001839	bay	0.202117	0.201368	Bay
100							0.000816	0.000799	bay	0.125694	0.124506	Bay

4. COMMENTS AND CONCLUSION

In table 1, when we compared reliability estimation for Burr-XXI distribution by using mentioned methods we find the best estimator is Bayes estimator, and this result is true for all values of parameter and especially in large sample sizes used in the study. Also we note that when the number of sample size increases the integrated mean square error (IMSE) and integrated mean square error (IMAPE) decreases in all cases.

REFERENCES

1. Soliman, A.A. (2005). Estimation of parameter of life from progressively censored data using Burr-XII model. *IEEE. Trans. Rel*, 54(1), 34-42.
2. Burr, I.W. (1942). Cumulative frequency Function. *The Annals of Mathematical Statistics*, 13, 215-232.
3. Moore, D. and Papadopoulos, A.S. (2000). The Burr Type XII distribution as a failure model under various loss functions. *Microelectronics Reliability*, 40, 2117-2122.
4. Wingo, D.R. (1993). Maximum likelihood estimation of Burr type XII distribution parameters under type-II Censoring. *Microelectron. Reliab*., 33,77-81.
5. Al-Hussaini, E.K. and Jaheen, Z.F. (1992). Bayesian estimation of the parameters, reliability and failure rate functions of the Burr type XII failure model. *J. Statistical and Simulation,* 41, 31-40.
6. Wang, F.K., Keats, J.B. and Zimmer, W.J. (1996). The Maximum likelihood estimation of Burr type XII parameters with censored and uncensored data, *Microelectron. Reliab*., 36, 362-395.
7. Evans, I.G. and Ragab, A.S. (1983). Bayesian Inference given a type-2 censored sample from Burr distribution. *Commun. Statist.-Theor. Math*, 2, 1569-1580.
8. Kundu, D. and Gupta, R.D. (2008). Generalized exponential distribution; Bayesian inference. *A computational statistis and Data Analysis*, 52, 4, 1821-1836.
9. Lewis, A.W. (1981).*The Burr distribution as a general parametric family in survivorship and reliability theory applications*. Ph.D., Thesis, Department of Biostatistics, University of North Carolina, Chapel Hill.
10. Ali Mousa, M.A.M. (1995). Empirical Bayes Estimators for the Burr type XII accelerated life testing model based on type-2 censored data, *J. Statist .Comput. Simul*, 52, 95-103.
11. Ali Mousa, M.A.M. and Jaheen, Z.F. (2002). Statistical inference for the Burr model based on progressively censored data. *Computers and Mathematics with Applications*, 43, 1441- 1449.
12. Nasir, S.A and AL-Anber, N.S (2012). A Comparison of the Bayesian and Other Methods for Estimation of Reliability Function for Burr-XII Distribution. *J. Math& stat* 8(1), 42-48.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 121-132

ATTITUDES TOWARDS OLD AGE AND AGE OF RETIREMENT ACROSS THE WORLD: FINDINGS FROM THE FUTURE OF RETIREMENT SURVEY

Hafiz T.A. Khan[1] and **George W. Leeson**[2]
[1] Middlesex University, London, UK. Email: h.khan@mdx.ac.uk
[2] Oxford Institute of Ageing, University of Oxford, UK
Email: george.leeson@ageing.ox.ac.uk

ABSTRACT

The 21st century has been described as the first era in human history when the world will no longer be young and there will be drastic changes in many aspects of our lives including socio-demographics, financial and attitudes towards the old age and retirement. This talk will introduce briefly about the Global Ageing Survey (GLAS) 2004 and 2005 which is also popularly known as "The Future of Retirement". These surveys provide us a unique data source collected in 21 countries and territories that allow researchers for better understanding the individual as well as societal changes as we age with regard to savings, retirement and healthcare. In 2004, approximately 10,000 people aged 18+ were surveyed in nine counties and one territory (Brazil, Canada, China, France, Hong Kong, India, Japan, Mexico, UK and USA). In 2005, the number was increased to twenty-one by adding Egypt, Germany, Indonesia, Malaysia, Poland, Russia, Saudi Arabia, Singapore, Sweden, Turkey and South Korea). Moreover, an additional 6320 private sector employers was surveyed in 2005, some 300 in each country with a view to elucidating the attitudes of employers to issues relating to older workers. The paper aims to examine the attitudes towards the old age and retirement across the world and will indicate some policy implications.

INTRODUCTION

Population ageing is now a global reality. In global terms, only 8.1 per cent of the population was aged 60 years and over in 1950. This figure had increased to just 10 per cent by the end of the 20th century but is expected to increase even further to 21.8 per cent by the year 2050 (United Nations, 2011) by which time the number of older people (aged over 60 years) will outnumber the number of young persons (aged under 15 years) globally. At the turn of the 21st century, there are approximately 600 million people aged 60 and over in the world corresponding to three times the number in 1950, and by 2050, the absolute figure is expected to reach almost 2 billion - again a tripling over just 50 years. Growth in older age group is massive, and it is fast, and the opportunities and the challenges to societies around the world are significant. The potential effects of ageing on the family are huge and varied (Harper, 2000, 2004; Lesson, 2006; Leeson and Khan, 2012). Although of a more modest nature, growth in the total global population is dramatic too. By mid-2000, world population had reached 6.1 billion and is increasing by 1.2 per cent corresponding to 77 million people per year, and by 2050 the world's population according to the UN medium variant is expected to be 9.1 billion (United

Nations, 2011). While the population of the more developed regions of the world is expected to change only modestly over the course of the coming 50 years that of the less developed regions is expected to increase steadily over that same period from 4.9 billion today to 7.8 billion. Asia

Life expectancy in the more developed regions at the turn of the 21st century was 76 years compared with 63 years in the less developed regions. Continued declines in mortality are expected to push these figures to 82 and 74 years respectively by the year 2050 thereby reducing the gap between developed and less developed regions. Moreover, a longevity dividend will emerge in many countries and it is expected that about 33 countries in the world will enjoy life expectancy at birth their over 80 years (UN, 2012). Global fertility levels are predicted to decline from 2.65 to 2.05 by the period 2045- 2050, which results in slowing population growth - the annual growth rate would by then have declined to just 0.38 per cent - still approximately 40 million persons per year. This growth will be exclusively in the less developed regions of the world. Most developed countries will experience a trajectory of below replacement levels of fertility (UN, 2012).

By 2030 a demographic shift will prevail as a quarter of the population of the developed world will be over 65 years and the same time a quarter of the population of Asia will be over 60 years. Demographic development at the global level is intricately linked as the flow of economic and human capital forms a high level of interdependency (Harper, 2006). In 2005, more than half of 194 reporting countries globally viewed the ageing of population as a major concern - the proportion as high as 75 per cent in the developed economies of the world (United Nations, 2006). Government responses include increasing the age of retirement, limiting early retirement and encouraging more women to enter the workplace. No country or government can or should hide from the ageing of its population. Nor from the challenges and opportunities such mature societies present in every aspect of human life. It is obvious that the ageing process will enforce companies to follow age discrimination legislation and to allow older employees in work place and to increase the retirement age logically. Today we see in many corporations in the US the average age of the employees is steadily increasing (US Bureau of Labour Statistics, 2008). Allowing increased retirement age benefit younger generation to learn from older generation and share ideas *vice-versa*. For line and HR – managers it will be decisive to understand the effects of age diversity on the overall organization (e.g. in terms of performance, productivity and turnover). Given the practical importance of an ageing society some countries have already considered their pension and retirement issues seriously. They have taken policy actions either with regard to increase the retirement age or to reform the pension policy. Example may be given for the UK and France where pension and retirement age options are changed recently.

The old age cannot be defined exactly because it does not have the same meaning in all societies. In many parts of the world people are considered old because of certain changes in their activities or social role. For example, people may be considered old when they become grandparents or when they have grey hair, or when they begin to do less or different types of work, or take retirement from job. On the other hand, demographic age may not be a factor to determine old age for the poorest part of the world. Some Poor vulnerable people may look like old due to suffering from long term chronic poverty and malnutrition and Bangladesh may be an example (Eusuf, 2012).

They may not be able to think of an extended retirement age. On the other hand, higher longevity is considered to be a triumph of development in most developed countries and people remains active at their age 80s. In that case people deserve to continue and stay in labour market and the government should think about the various options of the retirement too. Study shows that retirement age is directly linked with onset of dementia and early retirement often leads to increase the risk of dementia (Sorensen, 2009).

Ageing issues are not getting priority in policy agenda to some counties due to lack of understanding the consequences of ageing or people’s stereotype attitudes towards ageing. Research is needed to understand peoples’ attitudes towards the meaning of old age as well as age of retirement. Therefore, using the Global Ageing Survey (GLAS) (which is also popularly known as HSBC the Future of Retirement Survey), this study aims to investigate peoples’ attitudes towards old age and age of retirement across the world. All in all, these comprehensive data provide us with a unique opportunity to investigate global trends in attitudes and expectations to ageing and late life work and retirement, and to compare these attitudes and expectations across countries and cultures. This paper presents some of the findings of the survey relating to retirement age, examining cross national and regional differences in people's attitudes to when men and women should retire from the labour force. These attitudes will be compared with employer attitudes and practices.

DATA AND METHODS

In the first wave of the HSBC Future of Retirement Global Survey in 2004, some 11,000 persons aged 18 years and over in 10 countries and territories across four continents (Brazil, Canada, China, France, Hong Kong, India, Japan, Mexico, UK and USA) were surveyed on their attitudes and expectations to ageing and late life work and retirement. That first wave showed that these attitudes and expectations were predominantly positive across the globe. In addition, people's expectations in respect of withdrawal from the labour market proved to be more flexible and forward-looking than labour market infrastructures often allow. The second wave of the Global Survey has surveyed around 24,000 persons aged 18 years and over in 20 countries and territories across five continents. In 2005 the number of countries was increased to twenty by adding Egypt, Germany, Indonesia, Malaysia, Poland, Russia, Saudi Arabia, Singapore, Sweden and Turkey. The total population of the countries and territories covered comprises almost 62 per cent of the world's population. Again the focus has been on attitudes to ageing and late life work and retirement, both in respect of the family and the workplace, but also in relation to government financing and supporting older people. In addition, the second wave has surveyed 6018 private sector employers in these same countries and territories with a view to elucidating the attitudes of employers to issues relating to older workers. The interviews were mostly conducted by telephone and on some occasions by face-to-face. The data collection, editing, coding and final data-entry were performed by Harris Interactive. Details of the survey methodology and research reports can be found on their website: http://www.hsbc.com/hsbc/retirement_future/research-summary (HSBC, 2006a, b and c).

Both waves are used as source of data collection. Statistical analyses were performed using cross tabulations and chi-squared tests between selected variables. In some cases F-tests are used to verify the homogeneity across various groups. Geographical regions are constituted for the selected countries as North America (US and Canada), Europe (France, Germany, Poland, Russia, Sweden, Turkey and UK), Latin America (Brazil and Mexico), Middle East (Egypt, Saudi Arabia) and finally Asia (China, Hong Kong, India, Indonesia, Japan, Malaysia, and Singapore).

RESULTS AND DISCUSSION

Perceptions of the age of retirement - the retirement age gap?

The emergence of retirement at a broadly fixed age for all workers arose in the second half of the 20th century to cope with the specific health and socio-economic needs of the then older population, and in response to the changing administrative and personnel management demands of growing corporations (Harper, 2006). This has been extended in western developed countries through the spread of early retirement so that large numbers of healthy active men and women in their 50s and early 60s are choosing to *retire* or are forced to *retire* by their employers. There is now considerable evidence that early retirement practices are motivated by retirement incentives in current pension schemes (Gruber & Wise, 2000; Gould & Solem, 2000; Firbank, 1997; Luchak, 1997). Similarly, accumulated wealth, savings behaviour, and the availability of other sources of income in later life, both state-financed and private, are also influential (Banks et al., 2002). There is some evidence that direct and indirect age discrimination by employers encourages early withdrawal from the labour market (McKay & Middleton, 1998; Scales & Scase, 2000) and that push factors, such as redundancy or fixed retirement ages, are responsible for a large percentage of early retirements (Arrowsmith & McGoldrick, 1997). Finally, there is evidence that the current older cohorts have internalized the notion of retirement, including early retirement, as a period of funded leisure, and expectations of this are considerably entrenched, not only for the employee, but also for his or her partner and wider family (Scales & Scase, 2000). This is compounded by the growing responsibilities that many of these cohorts have for kin care and support, especially for their parents.

The increasing level of pension provision due to the introduction of occupational and private schemes, the increasing opportunities to purchase leisure goods appropriate to late life, and the market pressure on employers, which lead to rationalisation of the workforce, have all led to the reduction in the number and percentage of older workers in the labour market. This emphasis on finances and leisure means that ill health, once the main catalyst of retirement, has become a conditioning variable (Harper, 2006). Therefore, so long as an individual remains in good health, health status is largely irrelevant.

The concept of *retirement* has thus evolved. Both the Bismarkian and Beveridge systems were founded on a notion of older people requiring a short period of rest as they endured the frailty of old age. The rhetoric behind the Anglo-American programmes of the 1960s and 1970s in particular, such as US Social Security and UK pre-retirement planning, included a notion of reward for contribution to society, which may have reflected the arrival of the 2nd World War cohort as new retirees. By the 1980s, the internalisation of the idea of a period of funded leisure at the end of one's working life had become firmly established. The developed world had the social and cultural structures and most importantly the economic affluence that enabled societies to allow healthy active men and women to retire early and enjoy an extended period of leisure. The notion of retirement has thus been redefined from one of *Rest* in the 1940s and 1950s, to *Reward* in the 1970s to a *Right* by the 1980s (Harper, 2004a).

However, by the beginning of the 21st century in these societies, retiring to enjoy an extended period of leisure is no longer sufficient. Individuals want to take control of their lives and continue to be active. In the developing world, many societies do not have the social and cultural structures and most certainly not the economic affluence that can enable a period of leisure after retirement.

So what is people's perception of retirement age? When should retirement begin? What would delay their retirement? How do their aspirations compare with the practices of employers? Indeed, for many, retirement is still unheard of (Harper, 2006). As urban employees are now given the possibility of a period of funded retirement after their working lives, perhaps we are seeing these societies moving into active contributory retirement without first experiencing the phase of leisure. Thus, these societies may be leapfrogging from retirement as a period of rest due to ill health and frailty in old age straight into the new definition of retirement as a time of continued activity, while we shall see in this paper that perceptions of old age and age of retirement in these societies is not in line with this perception of retirement itself.

In the first wave, the survey asked a question on “At what age do you think of a person as becoming old”. It has been found that there is a significant difference across countries in terms of reporting old age (F=421, df=9, p<0.0001). People in France, Canada and Japan reported average old ages are 70.7, 68.0 and 67.0 years respectively. By contrast, people in India, Mexico and China reported 55.9, 55.7 and 49.8 years respectively. The lowest figure in China suggests investigating the attitudes of ordinary Chinese towards their old age as well as the timing for retirement. On the other hand, while comparing mean old age with the life expectancy of each country China is found to have a largest gap of 23.6 years and France has the lowest 10.2 years. The gap reveals a clear picture of the difference between mean old age and the mean life expectancy. However, it would be more meaningful to evaluate the gap between the retirement and the life expectancy. Age is found to be an important factor in reporting old age. The higher the age the higher the old age is reported in the survey. It also reveals that a significant difference exists in reporting old age across three age cohorts 18-39, 40-59 and 60+ years (F=683, df=2, p<0.0001). There also gender difference in terms of reporting old age, however females reported old age is higher than males (62.1 vs 59.9 years). Although a similar pattern is prevailed across all three age cohorts in developed countries, however, a reverse scenario is seen for developing countries. Women in developing countries are not confident about their old age which requires further investigation. Overall, a significant difference is also observed for developed and developing countries.

Perceived best age of retirement

In the second wave a question was asked on “At what age do you think males and females should retire”. At the global level, the survey reveals that both men and women think that women should retire earlier than men, namely average perceived best retirement ages of 57 years for women and 61.1 years for men. This is a significant difference between the average ages (p<0.0001). However, men feel that both men and women should retire at an earlier age (average 60.6 years for men and 56.5 years for women) than the retirement ages stated by women (average 61.6 years for men and 57.4 years for women). Again, these differences are significant (p<0.0001 in both cases).

As appears from figure 1, the average perceived best retirement age for men/women increases significantly with increasing age of the respondent (p<0.0001 in the case of the average retirement age for both men and women). From age 20 years, the average age at which respondents feel men should retire increases from 60.1 to 64.5 years; and in the case of retirement age for women it increases from 56 to 61.5 years.

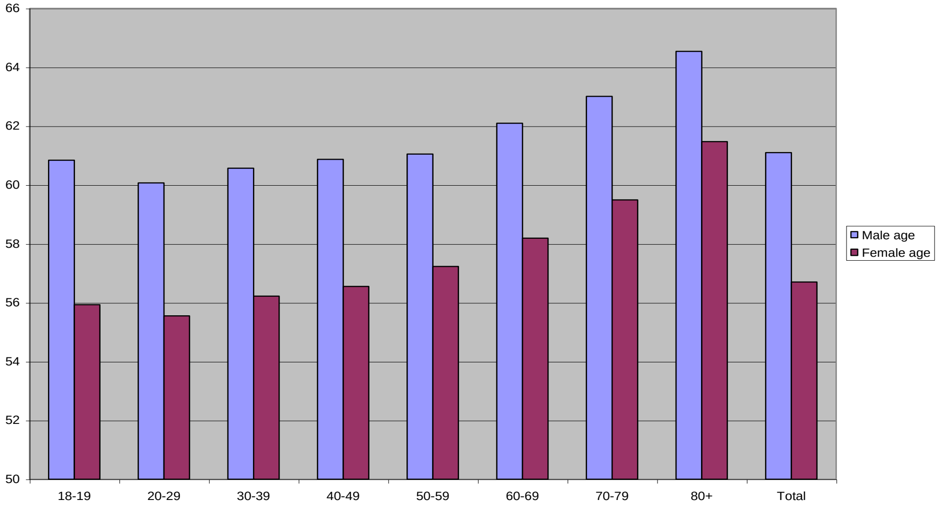


Fig. 1: The average perceived best retirement age for men and women respectively, according to age of respondent.

At the global level, those who are unemployed have the lowest perceived best age of retirement for both males (60.4 years) and females (55.7 years), while those in part-time employment have the highest (61.7 years for males and 57.8 years for females). These differences are in both cases significant (p<0.0001).

In addition, those living in the developed world have a perceived best age of retirement for both males and females that is significantly higher than that of those living in the developing world (p<0.0001). For males the stated ages are 62.4 and 60.1 years respectively, and for females they are 59.7 and 55 years (Figure 2).

If we combine age, gender and developed/developing region in a multivariate analysis, we still find the same underlying patterns in perceived best age of retirement. The perceived age for males is everywhere higher than the perceived age for females; the perceived age in the developed world is everywhere higher than the perception in the developing world; female perception is everywhere higher than male perception; and perceived ages increase with increasing age everywhere after age 20 years. Only in the oldest 80 years and over age group do we find differences in these respects that are not significant.

However, while people in Western Europe, most of Asia and Latin America have broadly similar views on the perceived retirement age, respondents from the United States and Japan quote higher average retirement ages for both men and women, while those from Egypt, Turkey and Saudi Arabia quote lower average ages. At either end of the spectrum lie 53 years in Saudi Arabia rising to 65 years in Japan and the United States for male retirement, and 48 years in Turkey rising to 64 years in the United States for female retirement.

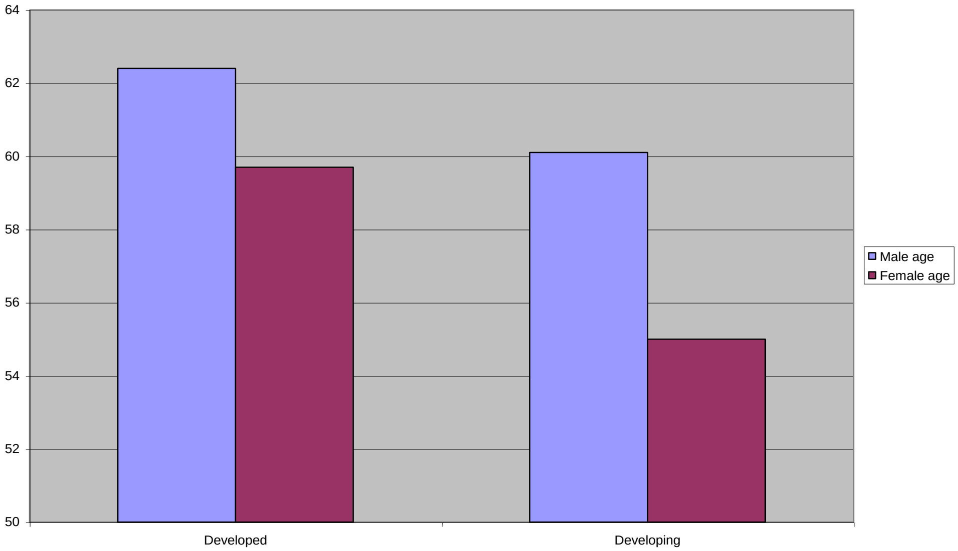


Fig. 2: The average perceived best retirement age for men and women

Continuing at the national level, we find that gender differences in the average perceived best retirement age for men are not significant in Brazil, Canada, China, France, Malaysia, Saudi Arabia, Sweden and Turkey, while they are not significant in respect of the average perceived best retirement age for women in Brazil, Canada, China, France, Hong Kong, Malaysia, Russia, Saudi Arabia, Sweden, Turkey and the United States.

Retirement age gap

In all of the countries surveyed, the *perceived best* retirement age stated by individuals and the *practised* retirement age (taken as the typical age of retirement in the workplaces surveyed) differ. In the developed economies of the United States and Japan, and the developing economies of Egypt, India, Malaysia, China and Brazil, the perceived best age for both males and females is older than the practised age. The largest difference is in India where the best retirement age for men at 63.4 years is some 6.2 years higher than the practised male retirement age of 57.2.

However, in Canada, and the European countries of Sweden, France, the UK, Poland, Russia and Turkey, along with Saudi Arabia and Mexico, the perceived best age for both males and females is younger than that practised. In France the perceived best retirement age for women is 3.5 years younger than the practised female retirement age of 60.5. Thus it would appear that the perceived best age for retirement is older than that currently practised in the US and Japan, despite these countries already having relatively late practised retirement ages. On the other hand, in Europe, with its already youngish early retirement ages in the private sector, the perceived best age for retirement is seen to be even younger than the practised retirement age in the private sector.

Old age in the workplace

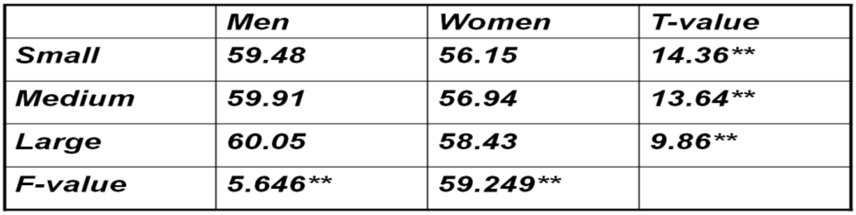
The survey also provides data on the age at which private employers view their workers as old. Only in Egypt and Brazil was the average age at which private employers perceived a worker as old higher than the average practised age of retirement stated by these same employers, while in Japan these two average ages were equal. In every other country, the average practised retirement age was higher than the age at which private employers perceive a worker as old, this being the case in Germany by as much as 11 years. Thus, while workers are leaving the labour market in Egypt and Brazil on average *before* their employers perceive them as old workers, elsewhere, and in Germany in particular, those perceived as "older workers" by employers are still active in the workplace.

For the 10 countries surveyed in the first wave of the HSBC Future of Retirement Global Survey, we are able to compare the average practised age of retirement in the private sector, with both the employer's average perceived age at which an individual becomes an older worker, and the perceived start of "old age". In six of these 10 countries - the United States, Canada, France, the United Kingdom, Japan and India - the average age at which a person is considered to be old is older than both the average practised age of retirement and the average perceived age of older workers. However, in Hong Kong, China, and Mexico, the age at which a person is considered to be old is younger than the practised age of retirement, though still older than the perceived age of older workers.

In these same 10 countries, we can compare people's perceived best age for retirement with their perception of when old age begins (Figure 5). In the more developed countries of Canada, France, Japan, the United Kingdom and the United States, people think individuals should retire before they become old. In the developing countries people think individuals should retire after or at the same time as they reach old age. There is no significant correlation (p=0.4) between the perception of retirement age and old age in the material. There seems then to be evidence that developed economies have reached a stage where the perceived age of retirement and onset of old age coincide with aspirations for activity in retirement. Developing economies meanwhile have *leapfrogged* into a stage where they associate retirement with positivism but where the perceptions of retirement age and the onset of old age do not quite match this aspiration.

The analysis of employers’ report towards practised retirement age by gender and size of the firm is presented in Table 1. Three types of companies are considered according to their size of employees and they are small, medium and large companies. It has been found that although average retirement age is higher for men than women irrespective of the size of companies however there exists a significant difference in retirement age by company size. Large company employees retire later compare to smaller companies in the market. This may be partly due to more job freedom, flexibility and confidence in large companies.

Table 1:
At what age do men/women typically retire from your organization?



	Men	Women	T-value
Small	59.48	56.15	14.36**
Medium	59.91	56.94	13.64**
Large	60.05	58.43	9.86**
F-value	5.646**	59.249**	

Note: Small (10-99 employees), Medium (100-499 employees), and Large (over 500 employees)

The determinants of retirement

Having considered the perceptions of best retirement age and practised retirement age, we shall conclude this paper by presenting the views of respondents in respect of the factors they feel should determine retirement. The survey reveals more or less universal support for individual desire and ability rather than age being the determinant for retirement. Globally, 35 per cent of respondents feel that people should retire *when they feel the time is right,* while a further 25 per cent feel retirement should be *when they are no longer able to work at what they did.* Only 22 per cent support *age* as the determinant.

This corresponds well with the feelings expressed by employers in as much as only 25 per cent feel that their organization should be able to enforce a fixed retirement age, while 72 per cent feel the employees should be able to go on working to any age if they are capable of doing the job well. The proportion stating that individual desire should be the determinant of retirement increases with increasing age of the respondents from 31 per cent of the 18-19 year olds to 38 per cent of the 50-59 year olds.

The whole move to a new concept of retirement and withdrawal seems to be rooted in societies across generations. The only national exemptions to this are Russia and Turkey, where the largest proportion support an *age* determinant; and Egypt, Saudi Arabia, India and Indonesia, where the largest proportion of respondents supports *ability* as the determinant of retirement.

This would seem to indicate that the developed world has moved significantly away age as a determinant of retirement and towards individual choice. This is perhaps not surprising in Europe and North America with in-place and forthcoming legislation on age discrimination in employment (Lees on & Harper 2006). Elsewhere, perhaps the support for age and ability determinants reflect the feeling for a need to have a cut off in an environment, where retirement is not yet viewed as a natural transition to an active phase of life.

ACKNOWLEDGEMENTS

This research arises as a part of the strategic alliance between the Oxford Institute of Population Ageing and the HSBC Bank Plc to promote the understanding of ageing issues across the globe. Authors gratefully acknowledge the financial support of HSBC to carry out research at the Oxford Institute of Population Ageing, University of Oxford, UK.

REFERENCES

1. Arrowsmith J. and McGoldrick, A.E. (1997). HRM SERVICE practices: Flexibility, Quality and Employee Strategy, *International Journal of Service Industry Management,* 7(3), 46-62.
2. Banks J et al (2002). Retirement, Pensions and the Adequacy of Saving: A Guide to the Debate, *Briefing Note No.* 29, Institute for Fiscal Studies, London.
3. Eusuf, MA. (2012). Dynamics of Urban Poverty in Bangladesh. PhD thesis, IDPM, Manchester University, UK.
4. Firbank 0 (1997). Early retirement incentive programs: organizational strategies and individual choices, *Arbete och Halsa,* 16, pp. 121-137.
5. Gould R & Solem PE (2000). Change from early exit to late exit, COSTA Meeting, Rome.
6. Gruber J & Wise DA (2000). *Social Security and Retirement around the World,* University of Chicago Press, Chocago.
7. Harper S (2000). Ageing Update: Ageing 2000 – questions for the 21st century. *Ageing and Society* 20:111-122.
8. Harper S (2004). Changing Families as European Societies Age, *European Journal of Sociology,* XLIV, 2, 155-184.
9. Harper S (2004a). The Implications of Ageing Societies, in Oberg BM (ed) *Changing Worlds and the Ageing Subject: Dimensions in the Study of Ageing and Later Life,* Ashgate, Burlington, VT.
10. Harper S (2006). *Ageing Societies,* HodderArnold, London.
11. HSBC (2006a). The future of retirement: what the world wants, London: HSBC Group Head Office. *www.thefutureofretirement.com*
12. HSBC (2006b). The future of retirement: what the people want, London: HSBC Group Head Office. *www.thefutureofretirement.com*
13. HSBC (2006c). The future of retirement: what business want, London: HSBC Group Head Office. *www.thefutureofretirement.com*.
14. Leeson GW (2006). The effect of HIV/AIDS on Intergenerational Relationships in Zimbabwe - the story of Sophie, Oxford Institute of Ageing, University of Oxford.
15. Leeson GW and Khan, HTA (2012). Levels of Welfarism and Intergenerational Transfers within the Family: *Evidence from the Global Ageing Survey (GLAS),* in: Global Ageing in the Twenty-First Century: Challenges, Opportunities and Implications, edited by McDaniel, S. and Zimmer, Z., Ashgate Publishing Ltd, UK, ISBN: 978-1-4094-3271-5.
16. Leeson GW & Harper S (2006). *Examples of International Case Law on Age Discrimination in Employment,* Department for Work and Pensions, London.
17. Luchak A (1997). Retirement plans and pensions: an empirical study, *Industrial*

Relations, 52, pp. 865-66.

18. McKay S & Middleton S (1998). Characteristics of Older Workers: Secondary Analysis of the Family and Working Lives Survey, *Research Report No. RR45,* Department for Education and Skills, Suffolk.
19. Scales J & Scase R (2000). *Fit and Fifty, A report prepared for the Economic and Social Research Council,* ESRC, Swindon.
20. Sorensen, S (2009). Yahoo News, May 18.
21. United Nations (2006). *World Population Policies 2005,* New York.
22. United Nations (2011). *World Population Prospects The 2010 Revision,* Population Division, Department of Economic and Social Affairs, New York: UN.
23. United Nations (2012). Ageing in the Twenty-First Century: A Celebration and a Challenge, UNFPA, New York.
24. US Bureau of Labour Statistics (2008). Older Workers. Retrieved Nov 14, 2012, from www.bls.gov/spotlight/2008/older_workers/

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 133-152

CONTRACEPTIVE NEEDS AND USE OF CONTRACEPTIVE SERVICES BY PRE-MENOPAUSAL WOMEN AGE 50 YEARS AND OVER IN BOTSWANA

Njoku Ola Ama
Department of Statistics, University of Botswana, Gaborone
Email: amano@mopipi.ub.bw; njoku52@gmail.com

ABSTRACT

The paper examines the contraceptive needs and use of contraceptive services by a stratified sample of older women (50 years and above) who have not attained menopause, from four selected sites in Botswana. The paper analyzes how some of the determinants of family planning impact on the uptake of contraceptive services/methods among the older women. The paper shows that 8.5% of the older women still want to have children while contraceptive prevalence among this group of older women is 59.3% and unmet need for family planning is 37.3%. The contraceptives mainly used by the older women are condom, breastfeeding, abstinence and barrier methods. A binary logistic regression analysis of the data reveals that having information, knowledge, availability and accessibility of specific methods do not always lead to the use of the method. For instance, although information and availability significantly predicts contraceptive use, knowledge and accessibility of the methods do not significantly predict contraceptive use. Thus, while availability and accessibility of condom is negatively correlated with contraceptive use, information and knowledge of condom is positively correlated with use. Older women who are employed, are in marriage and have some education are more likely to use contraceptives.

The study recommends increased effort by public healthcare programme planners, policy makers and NGOs to increase information, education and communications (IEC) interventions to boost uptake of family planning services among the older women. Such interventions should be service specific, specifically provide information and knowledge on the available methods and how best to access the methods; establish an environment of trust and respect; and explain how to use the chosen methods, benefits, and possible side effects.

KEYWORDS

Information, knowledge, accessibility, availability, family planning, older women.

1. INTRODUCTION

The government of Botswana in collaboration with other stakeholders involved in the provision of family planning services e.g. UNFPA have put in place various strategies and policies to increase uptake of family planning services in Botswana. These are aimed at increasing contraceptive prevalence rate (CPR), reduction in both total fertility rate (TFR) and unmet need for family planning services. However, these services are known only to target men and women of the reproductive age group (15-49 years) with little or no emphasis on older women 50 years and above, especially those who may not have attained menopause. Yet the women in this age group have their specific needs which because most of them are still sexually active and are therefore vulnerable to unwanted pregnancy, HIV and STIs, need special attention. Those older women who live in violent relationships are often unable to make family planning choices and are at greater risk of unwanted pregnancy, HIV/AIDS infection and STIs. The purpose of the study is to examine the utilization of family planning services by pre-menopausal women aged 50 and over and to analyze how these determinants impact on the uptake of family planning services/methods.

The paper accomplished these objectives using information provided by 61 older women from a a stratified sample of 444 older women drawn from four selected sites in Botswana. The paper made use of the binary logistic regression methods to examine the odds of family planning use and its determinants. The results enhance knowledge of the family planning needs of this significant but overlooked group and highlight the problems of older women in meeting their family planning needs. Recommendations that can enhance policy formulation by family planning programme managers and policy makers to address the needs of the older women are also made.

2. BACKGROUND

Family planning is critical in preventing unwanted pregnancies and unsafe abortion and in reducing maternal mortality as well as reducing poverty, maternal and child mortality. The service also empowers women to choose when and with whom to have children. Family planning means deliberately choosing when to have children by spacing the number of pregnancies, as well as avoiding unintentional pregnancy. It involves the use of some form of contraception or natural family planning methods until a couple is ready for a child[1]. According to[1], having another baby within a year of giving birth physically strains a mother by depleting necessary iron and vitamins in her body. Also attempts at breastfeeding two children who are closely spaced may be less successful. Caring for children can be stressful, and women who choose when and how many children to have may be better able to work or further their educations before committing to the potentially demanding tasks of raising children.

Fertility naturally declines with age, so when a woman is heading towards the end of her fertile years, she has a lower chance of becoming pregnant. However, there is still need to think about using contraceptives because the woman can still become pregnant. An unplanned pregnancy at an older age can be devastating for the individual woman and can present difficult choices. It is therefore advisable for the older women to continue contraception until there is no further chance of ovulation and risk of pregnancy. Women in their late reproductive years may also have heavy, irregular or painful periods which must be taken into account when choosing a method of contraception The author[3] estimated that the possibility of pregnancy in women between 45-49 years is two to three per cent, while the risk of pregnancy without contraception after the age of 50 years is minimal - less than one percent[2]

The family planning methods that can be used in either delaying or stopping pregnancy include oral contraceptives (the “Pill”); hormonal injectables; subdermal implants; intrauterine devices (IUDs); male and female sterilization; and barrier methods such as male and female condoms, diaphragms, and spermicides. Other modern methods include the Lactational Amenorrhea Method (LAM); fertility awareness methods such as methods that involve keeping track of when the fertile time of the menstrual cycle starts and ends (the Standard Days Method); and symptoms-based methods, which depend on observing signs of fertility (cervical secretions, basal body temperature)[4]. The natural family planning methods include breastfeeding, withdrawal and abstinence. Women of all ages need to remember that condoms are the only contraceptive choice to protect against sexually transmissible infections.

Studies have shown that if women had only the number of pregnancies they wanted, at the intervals they wanted, maternal mortality would drop by about one-third [5]. Women with birth-to-pregnancy intervals of less than five months experienced a risk of maternal death that was 2.5 times higher than women with birth-to-pregnancy intervals of 18 to 23 months[6]. In the developing world, an estimated 137 million women who want to avoid a pregnancy are not using a family planning method[7]. These women have an “unmet need” for family planning. Women with unmet need fall into two groups: women who wish to wait at least two years until their next pregnancy, and those who want to stop childbearing altogether. Globally, an estimated 55 percent of those with unmet need for family planning have a need for spacing and 45 percent for limiting[7-9]. Women may have an unmet need for family planning for a variety of reasons: lack of knowledge about the risk of becoming pregnant; fear of side effects of contraceptives; perceptions that their husbands, other family members, or their religion opposes family planning; or lack of access to family planning services[10]. Many of these barriers could be overcome through better information and counseling for both women and men.

2.1 The healthcare situation in Botswana

There is substantial unmet need for family planning in sub-Saharan Africa. In other regions of the world, unmet need is generally lower because more women in those regions are using family planning. Nevertheless, unmet need remains an important component of the total potential demand for family planning[11].

Primary health care is provided in all the 24 health districts of Botswana and the health services are very accessible in the urban and rural areas with 84% of the general population living within five kilometers of a primary health care health facility[12]. There is little or no difference between urban and rural dwellers as the health care facilities are easily accessible[13]. The hospitals are open 24 hours a day and the clinics are open from 7:30 a.m. to 4:30 p.m. (with someone on call to attend to emergencies)[14]. Health services are virtually free at the public facilities, requiring only a nominal charge of 5 Botswana Pula (US$ 0.70 at the exchange rate of 1 US$=7.2 Pula)[15]. It is worth noting that the maternal child health and family planning services are exempted from the nominal fee. Paved roads connect most villages, making referrals relatively easy even in the rural areas. Ambulances are available at the lower-level health facilities for transfers to hospitals. The government also has a contract with Netcare 911 to provide air ambulance services in emergencies[15].

The Primary health care programmes that are in place in Botswana include, amongst others: epidemiology and disease control services; occupational health services; environmental health services; food sciences laboratory services; maternal and child health/family planning services; expanded programme on immunization; food and nutrition; health education services; HIV/AIDS and sexually transmitted diseases; oral health services; and rehabilitation for persons with disabilities, as well as curative services[15]. Because of Botswana's strong family planning program, use of modern contraceptives among all women 15–49 years of age increased during the last four decades from 16 percent in 1984 to 29 percent in 1988, 40 percent in 1996, and 51 percent in 2007[16-17]. Use of traditional methods of contraception decreased from 7.5 percent in 1984 to 2.6 percent in 2007[16]. Male condoms are the most commonly used method of contraception (42 percent), followed by injectables (7 percent) and oral contraceptives (6 percent). Use of long-term methods such as intrauterine device and implants is negligible. The use of male condoms increased from one percent in 1984 to 11 percent in 1996 and 42 percent in 2007, and this increase has been attributed to an effective multimedia dual protection HIV campaign[15-16].

The Government of Botswana is thus highly ensuring that all people enjoy adequate health as enshrined in the Vision 2016 of Botswana. Howbeit, these services target women, men and youths within the primary sexually reproductive years (15-49 years). No specific health care programme has been designed to specifically target the older women (50 years and above), notwithstanding that many of them are still sexually active and very vulnerable to incidence of rape, HIV/AIDS and STIs. Very little, if any, information is available on the older women's family planning attitudes and behaviour, sexual behaviour and how sexual activities change with aging and illness in Botswana. A study by author(s)[18] reported sexual activity among the older women of varying ages and noted that women were significantly less likely than men at all ages to report sexual activity and but that many older women were sexually active although such activities may have declined;: the prevalence of sexual activity declined with age (73% among respondents who were 57 to 64 years of age, 53% among respondents who were 65 to 74 years of age, and 26% among respondents who were 75 to 85 years of age); Among respondents who were sexually active, about half of both men and women reported at least one bothersome sexual problem. Author(s)[19] have shown that a substantial percentage of the older women still enjoy sex with their partners, are very reluctant to attend the clinics with the younger women and have unmet need for family planning (72%)[20]. These women are very vulnerable to HIV/AIDS and STIs and many of them perceive their sexuality problems as normal and hardly consult the medical experts. Reducing the unmet need for family planning, stigma and discrimination associated with older women seeking family planning services from the public healthcare hospitals and clinics is important for the design of family planning programs because it affects the potential demand for family planning services and has important implications for future population growth. The good health of older women, as well as equal access and quality services across the life course, are very critical and should be of practical concerns to healthcare delivers and policy makers in the implementation of healthcare programme, particularly for HIV/AIDS, as most of the care given to orphan children, elderly members of their families and HIV infected persons is provided by the older women.

2.2 Correlates of uptake of family planning services

Analysis of DHS surveys data from 13 out of 27 developing countries showed that lack of knowledge, fear of side effects, and husband's disapproval were the principal reasons for nonuse among women who were otherwise motivated to use family planning[21]. A study by[22] using DHS-II data from sub-Saharan African countries indicated that lack of information about family planning, opposition to family planning, and ambivalence about future childbearing were the principal factors responsible for unmet need for family planning. Health problems and opposition to use are major reasons women not currently using contraceptives do not intend to use them in future. Twenty-three percent not intending to use contraception cited health concerns as the main reason while 5 percent expressed opposition to use, 5 percent expressed that the husband/partner disapproved, and 4 percent cited religion[23]. In the case of Botswana such factors as cost and access are much lesser concerns, indicating further need to strengthen demand for family planning services as they might be other causes for non-use of family planning services. With respect to pre-menopausal women aged 50 and over in Botswana: this paper aims to 1) determine their family planning needs and use of family planning services; 2) assess their main sources of contraceptive information and knowledge of contraceptive services. ; and 3) establish correlates of contraceptive service use, including demographic characteristics, information, knowledge, availability, and accessibility of contraceptive methods.

2.3 Operationalization of variable being studied

2.3.1 Availability, accessibility and knowledge

Availability of family planning services and methods in this study was measured by the older women's assessment of the method-mix (number of methods available) in the various clinics and hospitals; whether a particular method can be found in the clinics and hospitals and how regularly the older women were able to obtain their desired services from the clinics and hospitals. Accessibility, on the other hand, was measured by the closeness of the healthcare facilities to old women's homes, affordability of the services, the healthcare services being sensitive to social and cultural considerations such as gender, language and religion and also the quality of the services[24].

Knowledge of family planning methods and of places to obtain family planning services is crucial in the decision on whether to use a method and which method to use. It is presumed that more widespread knowledge of family planning method will result in greater use of contraceptives[25-26]. Acquiring knowledge about fertility control is an important step toward gaining access to contraceptive methods and using a suitable method in a timely and effective matter. Knowledge of family planning methods was measured by asking the respondent to name the ways that a couple can use to delay or avoid a pregnancy or birth. If the respondent did not spontaneously mention a particular method, the interviewer described the method and asked the respondent if she knew it. All the modern family planning methods, namely: female sterilization, male sterilization, the pill, intrauterine device (IUD), injectables, condom, combined oral contraceptives, and emergency contraception, diaphragm, vasectomy, spermicides, Norplant as well as the natural methods, periodic abstinence, breastfeeding, observation of safe periods and withdrawal were included in the list.

2.3.2 Information on family planning services

It is internationally agreed that individuals and couples should have informed and voluntary choice in using family planning and choosing the most appropriate methods[27]. Item 7.12 of United Nations ICPD programme of action reiterates "The aim of family-planning programmes must be to enable couples and individuals to decide freely and responsibly the number and spacing of their children and to have the information and means to do so and to ensure informed choices and make available a full range of safe and effective methods. The success of population education and family-planning programmes in a variety of settings demonstrates that informed individuals everywhere can and will act responsibly in the light of their own needs and those of their families and communities. The principle of informed free choice is essential to the long- term success of family-planning programmes"[28].

For individuals and couples to have informed and voluntary choice of family planning, information on available family planning services and methods, and the risk of using these methods ought to readily available. It requires that the healthcare providers are adequately trained on how to communicate this information to the older women. Furthermore, the use of several media to reach users of the family planning services need to be exploited. Appropriate information on family planning services and methods is therefore very critical in the successful implementation of the family planning programmes at all stages and for all groups of people including the older women.

3. METHODS

3.1 Setting and sample

The paper is derived from a study that was conducted between February and October 2011 and funded by the University of Botswana. Four health districts of Botswana namely Gaborone, Selibe Phikwe (predominantly urban areas), Kweneng East, and Barolong (predominantly rural areas) were purposively selected for the study. The 2011 population projection of women 50 years and above is 139,915 women, comprising 15.2% of the total female population and 12.1% of the total country's population [28]. The estimated sample size for the study was calculated at 454, using the sample size calculator programme[29] that allows for 95% confidence (and an error margin of 5%), and that posits that the response from the sampled population would be the same as that of the entire population. This number was allocated to the four sampled districts using probability proportional to size (PPS), where the size is the number of older women 50 years and above from each district[28]. The snow ball technique, a non-probability sampling method, was used to identify the older women from each of the districts because of the sparse nature of the population and the difficulty in obtaining an updated sampling frame of older women. The snowball technique is advantageous over the house-to-house survey as the latter is associated with a largely quantitative tradition of measuring the rare event that often suffers from a lack of response from the particular rare event, whereas snowball sampling involves locating the household with the rare event through key informant approach. Snowball sampling was found to be economical, efficient and effective for this study[30-31].

3.2 Instruments used for the study/Data collection

A questionnaire containing questions on demographic characteristics, sexual activities and family planning needs of older women, limitations, biases and stigma related to accessing family planning services was developed as research instrument. The questionnaire was of mixed structure. Some responses were provided on a five-point likert scale while some open ended questions gave the older women an opportunity to express their own opinions on the issues under investigation.

Trained research assistants administered the questionnaires to the participants at their homes or workplace. In some cases, where the respondent did not have time for a face-to-face interview, the questionnaire was self-administered and collected on a convenient agreed time. Potential respondents were informed that participation was voluntary and confidential in nature and there was no financial incentive. To maintain anonymity no personal identifiers were attached to the survey data. The purpose of the study was explained to the older women. Participants were not obliged to answer all questions and could terminate the interview at any time. A signed informed consent form was completed prior to participation but the questionnaire was completed anonymously. At the end of the data collection, 444 of the older women who were approached for the study completed the questionnaire giving a response rate of 98%.

3.3 Ethical consideration

Experts in public health and ageing reviewed the questionnaire prior to submission to the ethical committee of the University of Botswana. The Ethical Committee of the Ministry of Health also provided approval for the study. District Health Management Teams provided permission to conduct the study in each of the health districts. Research assistants, who were all females, were trained during a two-day training workshop on the content and administration of the research instrument as well as principles of attitude and behaviour towards older women participants.

3.4 Exclusion criteria

The study excluded older women who had physical or mental disabilities. This was due to the difficulty in providing special equipment such as facilities for hearing-impaired - signage, loops, and disability awareness training for interviewers, preparing large print materials for participants with visual impairments, and providing material in easy-to-read format for participants with intellectual disabilities.

3.5 Data analysis

The Statistical Package for the Social Sciences (SPSS) computer programme was used to capture and analyse the data. All variables, including the responses to the open-ended questions, were coded before being captured. Data were first analyzed using descriptive measures, such as percentages, means, and standard deviation. Multivariate binary logistic regression models were then fitted to the data to examine the impact of determinants of family planning on family planning service use by the older women.

4. RESULTS

The analysis and results in this paper are derived from the responses of 61 older women, out of a total sample of 444 respondents, who answered ‘No’ to the question, ‘Have you attained menopause?’

4.1 Demographic characteristics

The demographic characteristics of the older women (n=61) who have not attained menopause are shown in Table 1.

Age of women: 98.4% of the older women were between 50 and 59 years of age while 1.6% (n = 1) was between 70 and 79 years of age.

Highest educational qualification of older women: 26.2% of the older women had no schooling, 19.7% had primary certificate, 8.2% attempted secondary school certificate while 11.5% had secondary school certificate; 18% had diploma certificate; 4.9% had attempted degree and 9.8% had university degree while 1.6% had professional certificate.

Employment status: Majority (52.5%) of the older women were employed, 11.5% were unemployed but not seeking any employment; 18% were unemployed but seeking for some employment while 3.3% were retired civil servants.

Marital Status: 32.8% were single (never married); about 50.8% were married while 3.3% were widowed; 9.8% were cohabiting and 3.3% were divorced
Knowledge about menopause: A little over half of the older women (52.5%) had good knowledge of menopause while the rest had poor knowledge of menopause.

Table 1: Demographic Characteristics of the Older Women

Demographic characteristics of older women	**Response**	**Number of older women**	**Percent**
Knowledge of menopause	Good	32	52.5
	Poor	29	47.5
Age of older women	50-59	60	98.4
	70-79	1	1.6
Education qualification	No schooling	16	26.2
	Primary Certificate	12	19.7
	Secondary attempted	5	8.2
	Secondary school certificate	7	11.5
	Diploma certificate	11	18
	Degree attempted	3	4.9
	University Degree	6	9.8
	Professional Certificate	1	1.6
Employment status	Employed	32	52.5
	Retired civil servant	2	3.3
	Unemployed (but not seeking employment)	7	11.5
	Unemployed (but seeking employment)	11	18
	House wife	9	14.8
Marital Status	Never married (Single)	20	32.8
	Cohabiting	6	9.8
	Married	31	50.8
	Divorced	2	3.3
	Widowed	2	3.3

4.2 Family planning needs of the older women

The family planning needs of the older women were measured in this study by the methods that the women are currently using either to stop pregnancy or to delay having a baby. The questions required them to state whether they were using any family planning methods to either stop being pregnant or to delay pregnancy and which family planning methods they were using now. The results show that 8.5% of the older women still wanted to have children, 59.3% were using some family planning methods while 37.3% have unmet need for family planning.

Figure 1 shows the top ten family planning methods that the older women (n= 61) are using now. The top four preferred methods currently used now are condom (79%), abstinence (29%), breast feeding (16%) and barrier methods (13%).

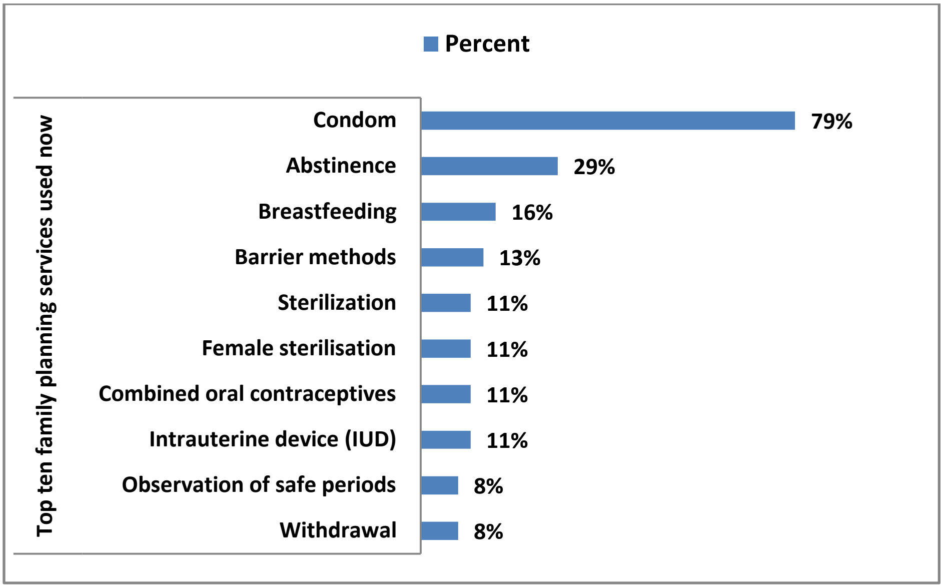


Figure 1: Top ten family planning methods being used now (n = 61)

Source of information about family planning needs

The main sources of information on available family planning were the nurses (55.8%), medical doctors (21.2%), radio and television (17.3%), medical journals (3.8%) and friends/neighbours/relatives (1.9%).

Availability, accessibility and knowledge of family planning services and use

The results of the analysis shows that condom (94%), intrauterine device (IUD) (76%), abstinence (66%), breastfeeding (62%) and combined oral contraceptives (62%) were the most available and accessible family planning services to the older women. Knowledge of condom was very high (95%), followed by the IUD (79%), abstinence (74%) and combined oral contraceptives (70%) (Table 2).

Table 2: Older Women's Assessment of Their Knowledge, Availability, and Accessibility of Family Planning Services

Family Planning services	Availability	Accessibility	Knowledge
Condom	94%	93%	95%
Intrauterine device (IUD)	76%	71%	79%
Combined oral contraceptives	62%	53%	70%
Progestogen-only pills	58%	49%	61%
Combined injectable contraceptives	60%	53%	65%
Female sterilization	46%	29%	46%
Breastfeeding	62%	62%	61%
Spermicides	24%	13%	28%
Barrier methods	28%	18%	33%
Sterilization	28%	20%	32%
Emergency contraception	24%	13%	28%
Diaphragm	48%	38%	51%
Vasectomy	42%	33%	51%
Norplant	44%	33%	46%
Abstinence	66%	69%	74%
Withdrawal	36%	24%	46%
Observation of safe periods	36%	27%	47%

In order to verify how availability, accessibility knowledge and information on these services influenced the use of the services, four multiple binary logistic regression analyses were carried out with the family planning services as predictors and the log odds of use of these services as the response variable.

The procedure defines the odds of in favour of using a family planning services, as

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \qquad 4.2.1$$

With $$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \qquad 4.2.2$$

where, π is the probability of using the family planning services or methods and $1-\pi$ is the probability of not using the family planning services. The probability of using the family planning services π is defined as

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} \qquad 4.2.3$$

The vector of covariates $(x_1, x_2, ..., x_p)$ is the predictor variable while $(\beta_0, \beta_1, \beta_2, ..., \beta_p)$ is the vector of regression coefficients which are unknown and need to be estimated and is defined as B in Table 3, 4, and 5. The quantity $Exp(B)$ in the table defines the multiplicative increase in the odds of a positive response for a unit change in any predictor variable holding the other variables fixed. Positive values of B indicate that the predictor variable increases the odds of using the family planning method or service, while a negative value indicates a decrease in the odds of use of the method.

The null hypothesis to be tested is "the model is a good fit to the data" versus the alternative hypothesis that "the model is not a good fit to the data". The null hypothesis is rejected if the p-value is less than α , the level of significance.

The four top contraceptives in terms of their availability and accessibility to the older women (Table 1) were extracted and further analysed using the multivariate binary logistic regression method and results shown in Table 3, 4, 5 and 6.

Table 3 shows that although the availability of condom, IUD, breastfeeding and abstinence adequately predicts use family planning (Chi-square = 3.801; p > 0.05). However, availability of individual services negatively correlates with use of the service (the beta values are all negative). Thus, the fact that condom, IUD, breastfeeding and abstinence are available in the healthcare system does not translate to the services being used by the older women (Exp (B) are less than one in each case)). On the contrary, accessibility of condom, IUD, breastfeeding and abstinence in the public health facilities do not adequately predict usage of family planning (Chi-square =13.042; p < 0.05). However, accessibility of the natural family planning methods, namely, abstinence and breastfeeding, although not significant (p > 0.05), correlates positively with the odds in favour of use of family planning. Older women who have access to breastfeeding and abstinence were respectively 1.5 and 1.7 times more likely to use the services (Table 4). The results show that knowledge of these four family planning services do not significantly predict use of family planning (Chi-square = 12.307; p < 0.05). While older women who had knowledge of condom were about 1.5 times more likely to use it, knowledge of IUD, abstinence and breastfeeding did not increase use of the services (they were negatively correlated with the odds in favour of use of the services). Those who had knowledge of abstinence were almost as likely to use it as those who did not have the knowledge (Table 5). Accessibility and knowledge significantly (p < 0.05) predicts use of the service.

Table 3: Logistic Regression to Measure the Impact of Availability of Top Four Family Planning Services on Usage

Availability of family planning services	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Condom	-0.765	0.85	0.81	1	0.368	0.465	0.088	2.462
IUD	-0.413	0.846	0.238	1	0.626	0.662	0.126	3.474
Breastfeeding	-0.200	0.772	0.067	1	0.795	0.818	0.18	3.717
Abstinence	-0.076	0.816	0.009	1	0.925	0.927	0.187	4.584
Constant	0.633	0.641	0.976	1	0.323	1.883		

Model adequacy: Chi-square =3.801, p > 0.05

Table 4: Logistic Regression to Measure the Impact of Accessibility of Top Four Family Planning Services on Usage

Accessibility of family planning service	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Condom	-1.377	0.848	2.637	1	0.104	0.252	0.048	1.33
IUD	-1.688	0.862	3.831	1	0.05	0.185	0.034	1.002
Breastfeeding	0.41	0.823	0.248	1	0.618	1.507	0.3	7.564
Abstinence	0.531	0.869	0.373	1	0.541	1.7	0.309	9.344
Constant	1.013	0.599	2.862	1	0.091	2.754		

Model adequacy: Chi-square = 13.042; p < 0.05

Table 5: Logistic Regression to Measure the Impact of Knowledge of Family Planning Services on Usage

Accessibility of family planning service	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Condom	0.413	0.388	1.133	1	0.287	1.511	0.707	3.23
IUD	-0.958	0.287	11.152	1	0.001	0.384	0.219	0.673
Breastfeeding	-0.242	0.292	0.689	1	0.406	0.785	0.443	1.391
Abstinence	-0.068	0.292	0.054	1	0.816	0.934	0.527	1.657
Constant	1.532	0.327	21.896	1	0	4.626		

Model adequacy: Chi square =12.307, p < 0.05

4.3 Information on family planning services versus use

The older women were asked if they had full information on family planning services from the health care systems in Botswana. Only 75% of them indicated that they had the information while 18.3% said they hadn't any information and 6.7% did not know of any information. The main sources of the information on the contraceptive needs were nurses (83.3%), medical doctors (70%), radio and television (53.3%) and the information reached the older women mainly in the form of printed materials (68%), audio-visual materials (49%), informative sessions (36%) and sessions for men and women (32%).

The study then explored how the usage of family planning services is affected by the extent of iinformation on the family planning services or methods using the multivariate binary logistic regression method (Table 6). The results of the analysis shows that full information on condom, IUD, breastfeeding and abstinence jointly predicts contraceptive usage (Chi-square = 2.379; p > 0.05). However, older women who had full information on condom, IUD and breastfeeding (although not individually significant, p > 0.05) were respectively 2.9 times, 1.9 times and 1.4 times more likely to use those services than those who did not have the full information.

Table 5: Logistic Regression to Measure the Impact of Information of Family Planning Services on Usage

Full information on family planning services	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Condom	1.073	1.368	0.615	1	0.433	2.923	0.2	42.699
IUD	0.618	0.813	0.576	1	0.448	1.854	0.377	9.133
Breastfeeding	0.304	0.73	0.173	1	0.678	1.355	0.324	5.668
Abstinence	-0.198	0.805	0.06	1	0.806	0.821	0.169	3.977
Constant	-0.671	0.408	2.702	1	0.1	0.511		

Model adequacy: Chi-square = 2.379; p > 0.05

Demographic variables versus family planning use

Table 7 shows the binary logistic regression to explore how demographic characteristics of the older women affect their use of family planning services. The table shows that although education and marital status do not significantly affect use of family planning (p > 0.05), but they are positively correlated with use of contraceptives. Older women who have some education and are married are respectively 2.6 times and 1.1 times, respectively, more likely to use family planning services than those are not educated or never married. Poor menopausal knowledge and being unemployed are negatively correlated with use of contraceptive (Beta values are negative). The older women who have poor knowledge of menopause are almost as likely to use contraceptives as those who have good knowledge of menopause (Exp (B) = 0.9) while the unemployed are less likely to use contraceptives than those employed.

Table 7: Logistic Regression to Measure the Impact of Demographic Characteristics of Older Women on Usage

Demographic characteristics of older women	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Unemployed	-0.399	0.567	0.494	1	0.482	0.671	0.221	2.04
Married	0.124	0.587	0.045	1	0.832	1.132	0.359	3.576
Poor knowledge of menopause	-0.057	0.583	0.01	1	0.922	0.944	0.301	2.96
Some education	0.946	0.63	2.255	1	0.133	2.577	0.749	8.862
Constant	-0.503	0.637	0.624	1	0.429	0.604		

5. DISCUSSION OF RESULTS AND CONCLUSIONS

This study set itself to determine the family planning needs of the older women who have not yet attained menopause, their knowledge, availability and accessibility of contraceptives from the healthcare services. Furthermore, the study sought to determine how the demographic characteristics of the older women, knowledge, availability, adequate information and accessibility of the contraceptive services impact on the usage of available services. The study revealed that the older women’s family planning needs are mainly condom (79%), abstinence (29%), breast feeding (16%) and barrier methods

(13%). The concentration on the natural family planning can be viewed from the perspective of the women not having appropriate information or knowledge of the modern methods. The author[32] outlined the primary obstacles to contraceptive use as lack of knowledge (arising from the lack of information) about contraception, its use or its availability; concerns about contraception's health effects; or cultural or familial objections. It was shown by[19] showed that the older women were using mainly the traditional methods before menopause.

Information dissemination about family planning services among the older women was low: about 25% of them had no information about the family planning services that were available in the healthcare systems in Botswana. Information about existing family planning services and their utility including associated risks of using them is very critical to the older women opting for the use of these services. Various media have been outlined for disseminating information concerning family planning services. The analysis shows that the older women got information about their family planning needs mainly through the nurses, medical doctors, radio/television in that order (see[19]). In addition, the results of this analysis indicate that generally older women with information about a particular family planning service were more likely to use the service than those who did not have the information. These findings are supported by[33], who found out that contraceptive prevalence was nearly 50% among women who recalled hearing or seeing messages in three media (radio, print and television) compared to 14% among those who did not recall any family planning messages. The authors[34] also showed that women exposed to radio messages about family planning were 1.9 times more likely to be current users. It is pertinent from these findings that a lot need to be done to bring information about available family planning services to the older women. Judging from the age of these women and the possibility that many of them may be incapacitated by poor health, special intervention needs to be put in place to inform them about existing family planning services and the risks of using the methods. Use of leaflets containing information on family planning services and the risks associated with them can be dropped at the homes of these older people.

While acquiring knowledge about fertility control is an important step toward gaining access to contraceptive methods and using a suitable method in a timely and effective manner, the relationship is not always positive. The results of the binary logistic regression analyses showed that while knowledge of condom do not significantly ($p > 0.05$) predict use of the method, those who had knowledge of condom were more likely to use it than those who did not have any knowledge of it. This result is supported by[35], who showed that although contraceptive knowledge is universal in India, with 96 percent awareness of at least one family planning method, yet usage of methods across the studied sample was less than 50%, showing a gap between knowledge and contraception adoption. Lack of knowledge of family planning was found by[8] to be another important reason for nonuse among women with unmet need.

Availability of family planning services in the healthcare systems implies that the services can be easily acquired both in terms of required number and mix. The study has shown that the most highly rated contraceptives in terms of availability to the older women, namely, condom, IUD, breastfeeding and abstinence, are negatively correlated with use of the services. Condom is the most widely used family planning method in

Botswana[36], a method that may prevent sexually transmitted infections (STDs) including HIV/AIDS if used correctly all the time. Condom is available free-of-charge in Botswana. The male condoms were made widely available in bars, gas stations, hotels, markets, pharmacies, private clinics, and salons, toilet ends, Government offices, market places, in the malls and any other strategic locations[16] and information about its use is widely spread. It is worrying that the older women are not utilizing the opportunity. This result point to the fact that there might be other hindrances to its use, namely, spousal disagreement and taboo associated with condom use[19]. Analysis of data from 13 DHS surveys by[21] showed that lack of knowledge, fear of side effects, and husband's disapproval were the principal reasons for nonuse among women who were otherwise motivated to use family planning. A study by[21] using DHS-II data indicated that lack of information about family planning, opposition to family planning, and ambivalence about future childbearing were the principal factors responsible for unmet need for family planning. The study, therefore, calls for interventions to break these bottlenecks, some of which might be culturally related.

Although access to four most perceived accessible family planning services/methods by the older women does not significantly predict family planning usage, access to the natural family planning methods, namely, abstinence and breastfeeding, on the other hand, are positively correlated with usage (beta values are positive). Thus, the direction of relationship between the accessibility of services and their usage is not generally conclusive but depends on the particular service. These results are in agreement with the findings from a World Fertility Survey (WFS) by[37-38] who indicated that level of contraceptive use has a strong positive association with availability and accessibility of family planning services. Rodriguez's analysis[36] showed further that inaccessibility of family planning services were important contributing factors to the very low level of contraceptive use in Nepal. The author[38] found out that that the proportion of current users of family planning services decreases as travel time to reach an outlet increases. The same pattern was found among both rural and urban women. The proportion of current users rises from 16 percent to 33 percent among rural women when travel time decreases from a half-hour or more to less than a half-hour (accessibility increases).

The results of this study can further be explained by the fact that although availability, accessibility, knowledge and information are essential in advancing utility of the contraceptive services, there are other sides of the coin. According to the Health Behavioural Model[39] the likelihood that someone will take action to prevent illness depends upon the individual's (i) evaluation of chances of getting a condition; (ii) evaluation of how serious a condition, its treatment, and its conseuqences would be; (iii) evaluation of how well an advised action will reduce risk or moderate the impact of the condition and (iv) evaluation of how difficult an advised action will be or how much it will cost, both psychologically and otherwise. Thus although condom is available everywhere and can be accessed free-of-charge the elderly women are not seen to be availing themselves of it. Changes in behaviour towards contraceptive use need to be adequately pursued and any intervention to enhance increased use of contraceptive methods should be directed towards behavioural change.

Demographic characteristics are other variables that can affect usage of family planning services. The study has shown that although having some education (primary up to university) and having been in marriage (married, divorced, widowed) are not significant predictors ($p > 0.05$) of family planning use, older women who have some education and have ever been married are more likely to use contraceptives than those who do not have any education and have never been married, respectively. This result however is in line with[40] findings from a multivariate analysis that increased education was significantly associated with the greater likelihood of using method of contraception. Married women were more likely to use contraception than married women who had been in more than one union, unmarried women who had been previously married, and never married women. Employment which is a proxy for income status in this analysis is positively correlated with contraceptive use. This result, however, agree with [40] that self-employed women and employees had much higher predicted probabilities of contraceptive use. Mediating variables (e.g., educational level, employment) are believed to indirectly affect behavior by influencing an individual's perceptions of susceptibility, severity, benefits, and barriers in the using contraceptive[41].

The author feels that emphasis should be on education for awareness creation of the utility, creating knowledge of where to get the FP services and risk involved in using the family planning methods so as to increase uptake of family planning services.

In the light of the study findings and discussions, it is clear that increased information and knowledge of the contraceptive services is key to usage of the services. The study, therefore, recommends increased effort by public healthcare programme planners, policy makers and NGOs to increase information, education and communications (IEC) interventions to boost uptake of family planning services among the older women. Such interventions should be service specific and specifically provide information on all available methods; establish an environment of trust and respect; explain how to use the chosen method, the benefits, and possible side effects; and be able to answer any questions that the older women might have.

It is, however, not very clear from the analysis why the older women are not using the condom despite efforts by public health planners and service providers to ensure availability, accessibility of the services, and to provide information leading to knowledge of condom. Further studies are therefore necessary to understand why the older women, particularly those who are still highly sexually active and can still be pregnant are not using condom, judging from the high prevalence of HIV/AIDS and STIs in Botswana. A further study is recommended that will target only older women (50 years and over) who have not yet attained menopause including those with disabilities such as the mentally ill and intellectual disabled as this group may be especially prone to sexual abuse.

ACKNOWLEDGEMENT

The author wishes to thank the Office of Research and Development, University of Botswana for funding the original research that gave rise to this paper.

REFERENCES

1. M. Brannagan (2010). Why is family planning important? University of Liverpool Available at http://www.livestrong.com/article/133732-why-is-family-planning-important/. Accessed 25 August 2012 (see also http://www.livestrong.com/article/ 133732-why-is-family-planning-important/#ixzz24b3wtMLu).
2. Australasian Menopause Society (2008). The Jean Hailes Foundation for Women's Health, "Fact Sheet: Contraception when you are approaching menopause", http://www.menopause.org.au/images/stories/education/docs/2008contraception_menopause_col.pdf
3. A. Gebbie (2011). Contraception: For Older Women . http://www.menopausematters.co.uk/contra1.php Accessed 01 October 2012
4. INFO Project, "Johns Hopkins Bloomberg School of Public Health, Family Planning: A Global Handbook for Providers," Baltimore: Johns Hopkins Bloomberg School of Public Health, 2007. Accessed online at www.infoforhealth.org/globalhandbook/index.shtml.
5. M. Collumbien. M. Gerressu, and J. Cleland, (2004). Non-use and use of ineffective methods of contraception. In 'Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors' Ezzati M, Lopez AD, Rodgers A, Murray CJL (2004) WHO (Geneva): 1255-1320
6. A. Conde-Agudelo, and JM. Belizán, (2000). Maternal morbidity and mortality associated with interpregnancy interval: cross sectional study. *BMJ,* 321(7271), 1255-9.
7. S. Singh, J E. Darroch, L.S. Ashford, and M. Vlassoff, "Adding It Up: The Costs and Benefits of Investing in Family Planning and Maternal and Newborn Health", Guttmacher Institute, New York. (no date). Available at http://www.unfpa.org/ webdav/site/global/shared/documents/publications/2009/adding_it_up_report.pdf. Accessed 25 August 2012
8. A. Korra (2002). Attitudes toward Family Planning, and Reasons for Nonuse among Women with Unmet Need for Family Planning in Ethiopia, Calverton, Maryland USA: ORC, Macro.
9. Central Statistical Authority (CSA) [Ethiopia] and ORC Macro (2001). Ethiopia Demographic and Health Survey 2000, Addis Ababa, Ethiopia, and Calverton, Maryland,USA: Central Statistical Authority and ORC Macro.
10. G. Sedgh, R. Hussain, A. Bankole, S. Singh, (2007). Women With an Unmet Need for Contraception in Developing Countries and Their Reasons for Not Using a Method. Occasional Report, 37 (New York: Guttmacher Institute).
11. RAND Corporation (1998). The Unmet Need for Contraception in Developing Countries, 1998. Available at http://www.rand.org/pubs/research_briefs/RB5024/ index1.html#. Accessed 25 August 2012
12. Central Statistics Office (CSO). Botswana Ministry of Health (MOH): Health Statistics Report 2004", Gaborone, Botswana, 2007: December
13. T.T. Langeni-Mndebele, (1997). Sociocultural Determinants of Fertility in Botswana. Dissertation, University of Alberta, Published in Dissertation Abstracts International, A: The Humanities and Social Sciences, 58(10): 4077 (April 1998).
14. K.D Mogobe, W. Tshiamo, and B. Motsholathebe. (2007). Monitoring Maternity Mortality in Botswana, Reproductive Health Matters, 15(30), 163-171.

15. Republic of Botswana, (2009). The Contribution of the Botswana Family Planning Program to the Largest Fertility Decline in Sub-Saharan Africa," A report prepared for presentation at the International Conference on Family Planning: Research and Best Practices, Kampala, Uganda.
16. S. Mills, V.Leburu, S. El-Halabi, L.Mokganya, and S.Chowdhury, (2010). Fertility Decline in Botswana 1980–2006: A Case Study", The World Bank. Available at http://siteresources.worldbank.org/INTPRH/Resources/376374-1278599377733 /Botswana61810PRINT.pdf. Accessed 28 August 2012.
17. World Bank (2011). Reproductive health at a glance-Botswana, Available at *www.worldbank.org/population*. Accessed 26 June 2012.
18. S.T. Lindau, L.P Schumm, E.O. Laumann, W. Levinson, C.A. O'Muircheartaigh, and L.J. Waite, (2007). A Study of Sexuality and Health among Older Adults in the United States. *The New England Journal of Medicine*, 357, 762-774.
19. N.O. Ama, and E. Ngome (2012a). Sexual and reproductive health of older women (50+ years) from selected sites in Botswana, A research report submitted to the Office of Research and Development, University of Botswana, Gaborone.
20. N.O. Ama, and E. Ngome, (2012b). Challenges older women have to access services addressing sexual and reproductive health including family planning needs in Botswana, Accepted for publication in South African Journal of Family Practice (SAFPJ).
21. J. Bongaarts, and J. Bruce, (1995). The causes of unmet need for contraception and the social content of services, *Studies in Family Planning*, 26(2), 57-75.
22. C.F. Westoff, and A. Bankole, (1995). The potential demographic significance of unmet need, *International Family Planning Perspectives*, 22, 16-20.
23. Central Statistics Office and UNICEF (2009). 2007 Botswana Family Health Survey IV Report, Gaborone, Botswana.
24. D. Clifton, (2010). Expanding access to family planning, Population Reference Bureau, Washington, USA.
25. Ministry of Health (1998). Indonesia Demographic and Health Survey 1997, Central Bureau of Statistics, Jakarta, Indonesia, 1998, p.49. Available at http://www.measuredhs.com/pubs/pdf/FR95/00FrontMatter.pdf
26. F. El-Zanaty, A. Way, (2006). Egypt Demographic and Health Survey 2005, Ministry of Health and Population, Cairo, p 55.
27. Central Statistics Office (CSO) (2003). The 2001 Population and Housing Census, The Government Printers, Gaborone.
28. United Nations 1994). Report on the International Conference on Population and Development. Cairo, United Nations Publication.
29. Creative Research Systems (2003). The Survey Systems: Sample Size Calculator, Available at: http://www.surveysystem.com/sscalc.htm Accessed 20 April 2010
30. S. Jejeebhoy, M. Koenig, and C. Elias, (2003). Community interaction in studies of gynaecological morbidity: experiences in Egypt, India and Uganda, In: Jejeebhoy S, Koenig M, Elias C, eds. Reproductive tract infections and other gynecological disorders. Cambridge, Cambridge University Press.
31. P. Singh, A. Pandey, and A. Aggarwal, (2007). House-to-house survey vs. snowball technique for capturing maternal deaths in India: A search for a cost-effective

method, *Indian J. Med. Res.*, 125, 550-556 National Institute of Medical Statistics (ICMR), New Delhi, India.

32. A. Bayer, (2002). Unmet Need for Contraception in the 21st Century, The Population Resource Center, Washington, DC 20006. Available at http://www.prcdc.org/files/Unmet_Need.pdf. Accessed 20 August 2012
33. C.F. Westoff, and G. Rodriguez, (1995). The Mass Media and Family Planning in Kenya International Family Planning Perspectives, 21(1), 26-36.
34. M.N. Jato, C. Simbakalia, J.M. Tarasevich, D.A. Awasum, C.N.B. Kihunga, and E. Ngirwamungu, (1999). The impact of multimedia family planning promotion on contraceptive behavior of women in Tanzania, International Journal of Family Planning Perspectives, 25(2), 60-67.
35. R. Deb, (2010). Knowledge, Attitude and Practices Related to Family Planning Methods among the Khasi Tribes of East Khasi hills Meghalaya. Kamla-Raj, Anthropologist, 12(1), 41-45.
36. V.M. Leburu, S. El-Halabi, L. Mokganya, S. Mills. (2009). The Contribution of the Botswana Family Planning Program to the Largest Fertility Decline in Sub-Saharan Africa October 2009 A report prepared for presentation at the International Conference on Family Planning: Research and Best Practices, Kampala, Uganda.
37. G. Rodriguez, (1978). Family planning availability and contraceptive practice, *International Family Planning Perspectives and Digest*, 4, 100-115.
38. J.W. Brackett, (1980). The role of family planning availability and accessibility in family planning use in developing countries, in World Fertility Survey Conference, 1980: Record of Proceedings, Volume 2 (Voorburg: International Statistical Institute, pp. 19-49.
39. C A. Redding, J S. Rossi, S R. Rossi, WF. Velicer, JO. Prochaska, (2000). Health Behavior Models, *The International Electronic Journal of Health Education*, 3(Special Issue), 180-193.
40. Shapiro D, Tambashe BO. (1994). The impact of women's employment and education on contraceptive use and abortion in Kinshasa, Zaire. Studies in Family Planning, 25(2), 96-110.
41. I.M. Rosenstock, (1990). The health belief model: explaining health behavior through expectancies. In: Glanz K, Lewis FM, Rimer BK, eds. *Health Behavior and Health Education: Theory, Research, and Practice.* San Francisco, CA: Jossey-Bass, 39-62.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 153-164

SURVEY SOFTWARE USE IN THE MAGHREB: A COMPARATIVE STUDY BETWEEN ALGERIA, MOROCCO AND TUNISIA

Oula Bayarassou[1], **Younès Boughzala**[2], **Le Sphinx MENA**[2],
Laila El Harouchi[3] and **Jean Moscarola**[4]

[1] École Supérieure de Commerce de Tunis - La Mannouba – 2010 – Tunisia.
Email: Oula.bayarassou@gmail.com
[2] Cité Errihab -Rue du Lac Malaren - 1053 Les berges du Lac – Tunisia.
Email: Yboughzala@lesphinx.eu
[3] Le Sphinx Développement - Casablanca – Morocco.
Email: Lelharouchi@lesphinx.eu
[4] IREGE – IAE Savoie Mont-Blanc – Université de Savoie -4 Chemin de Bellevue
74016 Annecy-le-Vieux, France. Email: jean.moscarola@univ-savoie.fr

ABSTRACT

This study seeks to discover the extent of the use of survey and data-analysis software in Tunisia, Algeria and Morocco. The study was conducted among Tunisians, Algerians and Moroccans operating in different fields. The questionnaire was administered online. This paper presents the results of said survey, as well as a comparison of the usage of survey and data-analysis software between three countries of the Maghreb. The survey cannot be deemed representative, it did afford us an overview of the use of survey and data-analysis software in three rather similar consumption environments?

This study was conducted with the technical and software support of Le Sphinx (www.lesphinx.eu), for which we wish to express our gratitude.

KEYWORDS

Data analysis, Software, Survey, Comparative study, Maghreb.

1. INTRODUCTION

Before one can affect a market, it is essential that one should understand it well and have a solid grasp of the behaviors of its players, whence the importance of marketing surveys and studies, which nowadays comprise one of the main success factors for any company seeking to further its development and the satisfaction of its customers (Jolibert and Jourdan, 2006). Indeed, they allow the manager to solve a specific problem, to check his performance, and to shed light on and assist in the planning of his decisions (Evrard et al., 2003) by collecting and processing data, whether quantitative (numbered) or qualitative (related to the behaviors and opinions of individuals). Furthermore, there is now a panoply of statistical tools, viz. survey and data-analysis software (Danaguezian, 2002) that facilitate conducting marketing surveys and studies by offering those responsible for carrying out the studies a number of functionalities depending on the projected needs (Moscarola, 1990). The use of such tools is becoming inevitable, which has led us to wonder how much importance do professional and academic users assign to survey and data-analysis software.

The goal of this study is to obtain an overview of the use of survey and data-analysis software within the context of Tunisian, Algerian and Moroccan consumption (Bayarassou et al., 2012). In the end, this study will allow us to know, on the one hand, expectations users of the survey and/or data-analysis software may have and, on the other hand, the use to which this software is put. The decision to focus this study on three countries of the Maghreb is justified by the fact that they present similar socioeconomic structures, share a common history, and are part of the same geographical entity.

2. PRESENTATION OF THE SURVEY

2.1 Design of the questionnaire

The questionnaire design stage is the time for one of the more delicate decisions that arise when setting up an survey. Our questionnaire consists mainly of closed questions (single, multiple and scales). There are two major sections: the first one deals with the activity of the surveys and studies, while the second focuses the opinion and practice of survey and data-analysis software users. Finally, we asked a number of questions relating to the respondent's profile, *viz.* sex, educational level, specialty, and socio-professional status. The questionnaire was pre-tested by approximately ten Tunisians, Moroccans and Algerians to evaluate the time needed to respond to it and to verify the relevance of the questions and the questionnaire's overall consistency prior to its distribution.

2.2 Collection of data

To collect the information, the questionnaire was administered online[1]. For Tunisia, we sent the link to the questionnaire to 7,800 contacts (companies, government agencies, researchers, students…). After a follow-up, we received 420 returns, which corresponds to a response rate of 5.38%. We then randomly selected 115 responses which we incorporated in this survey. As regards Algeria, we sent the link to the questionnaire to 691 addresses (companies, researchers, professionals…). After a follow-up, we received 45 returns. The response rate stands at 6.51%. Finally, in the case of Morocco, we called on 1,750 Moroccans working in various fields of activity. We gathered 90 online returns. The response rate was 5.14%.

Response rates for the three aforementioned countries are deemed acceptable for an online survey. Indeed, in France, the average return rate of online surveys is between 5% and 10% (Ganassali, 2007).

Table 1: Summary table describing the collection of data in Tunisia, Algeria and Morocco

Country	People contacted	Returns	Response Rate
Tunisia	7,800	420	5.38%
Algeria	691	45	6.51%
Morocco	1,750	90	5.14%

[1] http://www.sphinxonline.net/StageTunisie/sphinxtunisie/questionnaire.htm.

3. RESULTS OF THE SURVEY

First, we will present the general characteristics of respondents. We will then present the results of a descriptive analysis, to provide an overview of the activity of the surveys and studies in Tunisia, Algeria and Morocco, as well as the tools used. These analyses were performed using the Sphinx iQ software, version V.6.2.7.1.

3.1 Profile of respondents

- *Tunisian respondents:* The results of this survey revealed that 61.7% of the surveyed Tunisian population has a Master's or higher degree in marketing (37.4%), data processing (11.3%), and management of organizations (11.3%). Slightly over 25% are senior executives, mostly active in the following sectors: studies and consultancy (20.5%) and industry (20.5%). Research-teachers represent 24.3% of all respondents. We also noted that nearly 90% of them conduct surveys, over 50% of which calculate that survey activities account for three quarters of their work.
- *Algerian respondents:* The results of the descriptive analyses revealed that 50% of Algerian respondents have a Master's or higher degree, and their specialties consist mainly of: marketing (17.4%) and management of organizations (10.9%). 23.9% of them are senior executives operating in various fields, such as studies and consultancy (19.6%) and administration (6.5%). Research-teachers account for about 23.9%. We noted that 82.6% of all respondents conduct surveys and studies, and that 26.1% of them consider that survey activities only represent between 50% to 75% of their work.
- *Moroccan respondents:* The descriptive analysis revealed that 65.2% of Moroccan respondents have a Master's or higher degree, and that they specialize mostly in marketing (17.8%), economics (11.1%), and management of organizations (10%). Over 30% of all Moroccan respondents are senior executives operating in the sectors of study and consultancy (15.8%), banks and insurances (18.4%), government agencies and ministries (13.2%), and trade and distribution (10.5%). Research-teachers only account for about 17.8%. The results showed that 86.7% have already conducted surveys or studies, and that 20.8% of them consider that survey activities represent between 50 to 75% of their work.

3.2 Place of the survey and data-analysis software in the professional and academic sectors in Tunisia, Algeria and Morocco

In this section, we focus on the degree of importance that Tunisian, Algerian and Moroccan users lend to the survey and data-analysis software.

3.2.1. Type of survey according to respondent's profile

The results revealed that the satisfaction survey is the most common one in countries of the Maghreb, with response rates of 64.1% in Morocco, 48.4% in Tunisia, and 45.7% in Algeria, followed by the investigative survey, for which the response rate stands at nearly 50% in all three countries. In Algeria and Morocco, the opinion poll came in third place, with an average response rate slightly higher than 45%, whereas in Tunisia, it is the customer survey that ranked third, with a percentage of 32.2% (Figure 1).

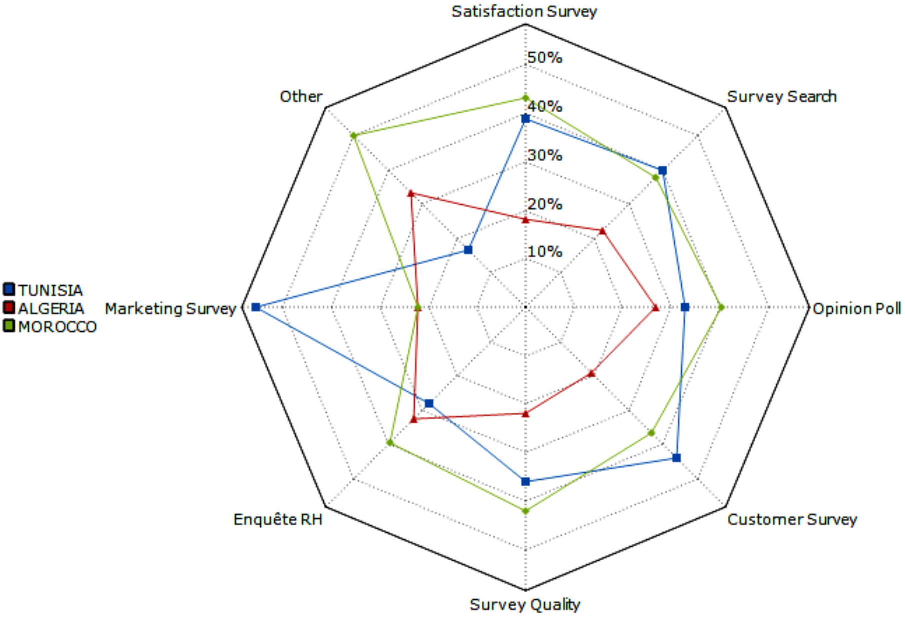


Fig. 1: Types of surveys in Tunisia, Algeria and Morocco

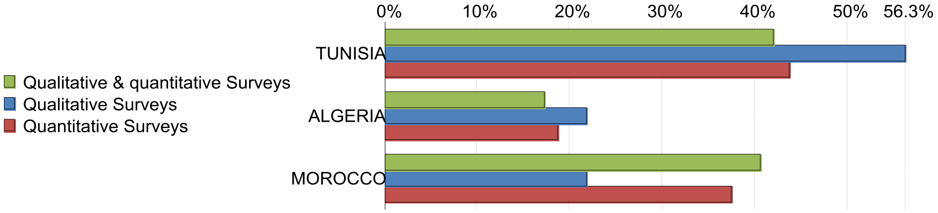


Fig. 2: Types of surveys in Tunisia, Algeria and Morocco: qualitative or quantitative?

It bears noting that the analysis of the data reveals that over 65% of all respondents – Tunisian, Algerian and Moroccan – perform both qualitative and quantitative surveys (75.6% in Morocco, 65.8% in Algeria, and 65.5% in Tunisia). Slightly over 15% of all respondents in the three countries perform quantitative surveys. Regarding the use of qualitative methods, we note that the response rate in Tunisia and in Algeria is twice as high as the one found in Morocco (9%) (Figure 2).

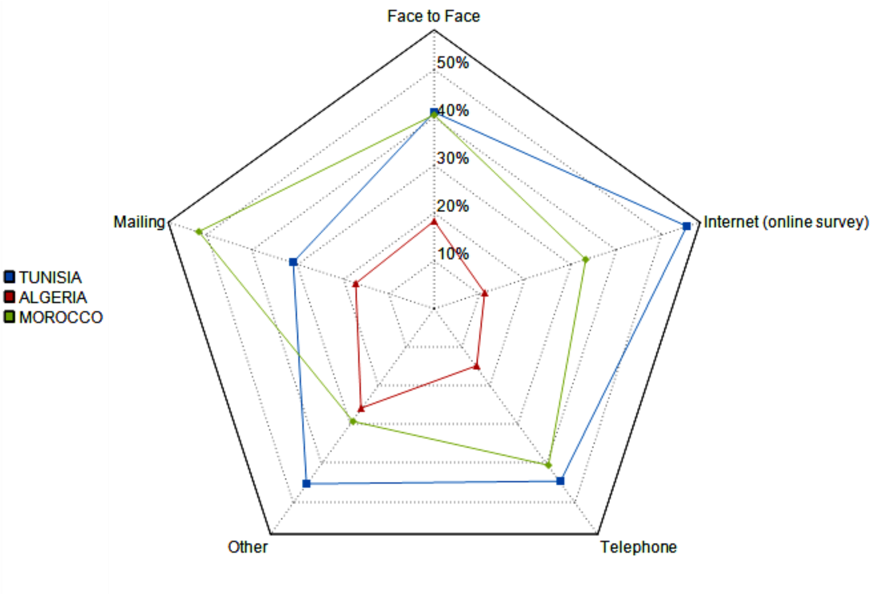


Fig. 3: Supports used for the collection of data in Tunisia, Algeria and Morocco

A very large majority of respondents, Tunisians, Algerians and Moroccans alike, use the "Face-to-Face" technique to conduct all kinds of surveys: 82.1% in Morocco, 69.9% in Tunisia, and 63% in Algeria. The second most popular method is the online survey which, according to the results of this survey, is more commonly used in Tunisia (59.1%) and Morocco (42.3%) than in Algeria (23.9%) (Figure 3).

3.2.2. Survey and data-analysis software

- *Ranking of software by fame*[2] *by Tunisian academics and professionals*

In this section, the first issue that was raised was the fame of the survey and data-analysis software.

[2] The question asked was, "*Which of the following survey and data-analysis software do you know?"*

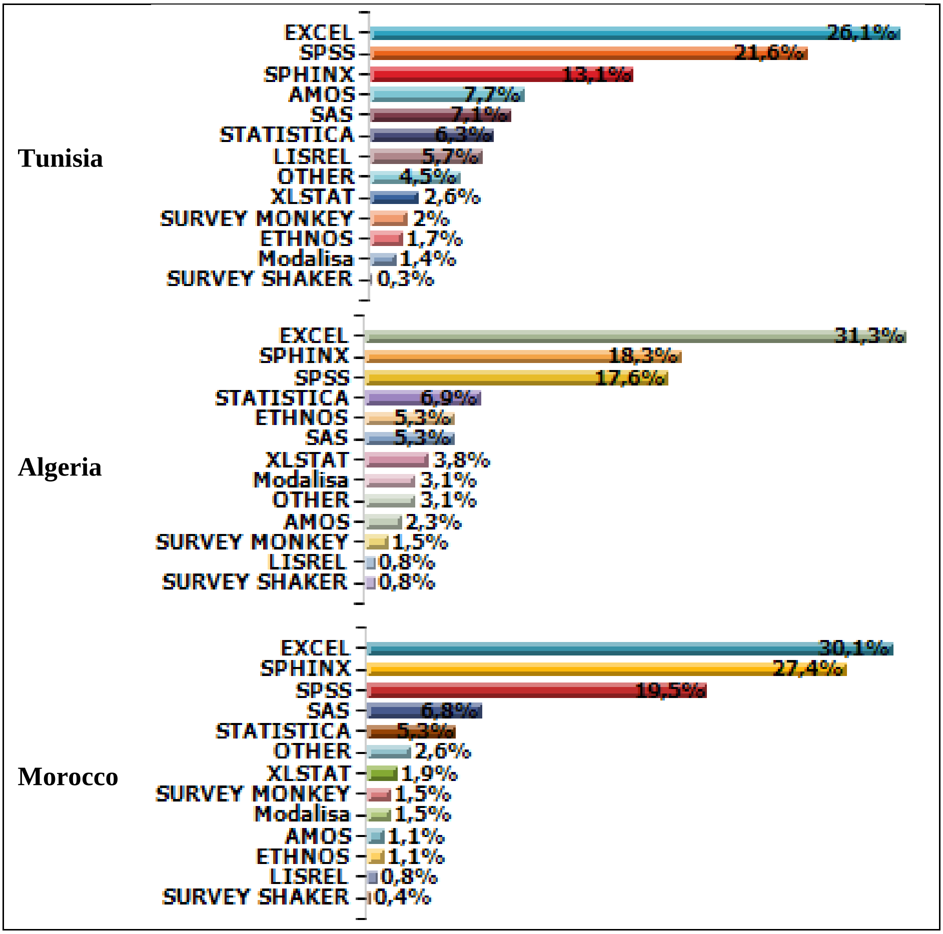


Fig. 4: Software fame in Maghreb countries

We highlighted the top 4 software mentioned by Maghreb users. Excel, Sphinx and SPSS are, respectively, the top three survey and/or data-analysis software most quoted by Algerians and Moroccans. For Tunisians, the SPSS software comes in first position, followed by Sphinx. We also noted that Maghrebians use modelling software such as Amos, (quoted by Tunisians) Statistica (quoted by Algerians) and XStat (quoted by Moroccans) (Figure 5).

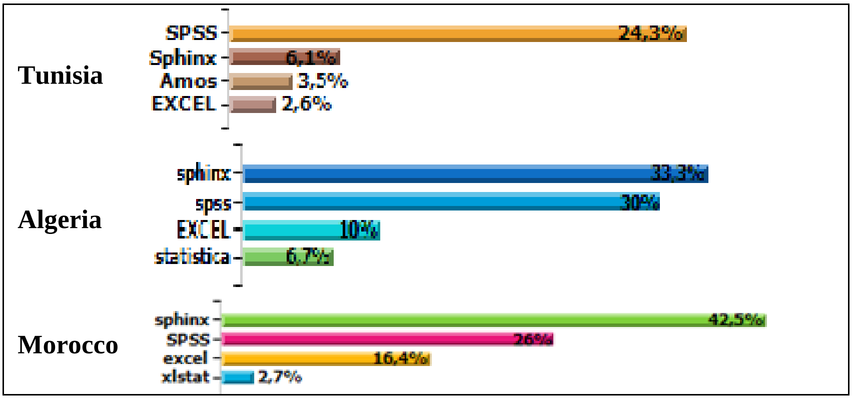


Figure 5: Software used in Maghreb countries

- *Reasons for selecting a survey and/or data-analysis software*

The results obtained indicate that users state that the choice of a survey and/or data-analysis software is not made on the basis of sales communications (0%). Indeed, most respondents – *i.e.* 70.7% of academics and 50% of professionals – based their choice on prior expertise in use of the software obtained during an academic course. Others (22.2% of academics and 20.5% of professionals) relied on the opinion of a friend or colleague who recommended the use of a given software. Finally, both professionals and academics confirm that Internet searches also had an impact on the decision-making process when selecting a survey and/or data-analysis software (7%).

- *The use of different data-analysis methods employed by academics and professionals*

This section relates to the data-analysis methods used, we asked respondents which methods they used most often in their activities. The results revealed that most Algerian and Moroccan respondents use descriptive data-analysis methods, *viz.* cross tabulations and flat tabulations, to describe the population being studied. As for Tunisians, they rather resort to more advanced methods, such as factorial analyses (PCA and CA), linear regressions, and ANOVA (Analysis of Variance). As regards qualitative analyses (lexical analysis, content analysis, and textual analysis), they are not used as frequently in any of these countries. This might be explained by the nature of the socio-professional category of the respondents, where those who work in research tend to favour the more sophisticated analyses while those who belong to the professional environment often use the classic analyses (Figure 6).

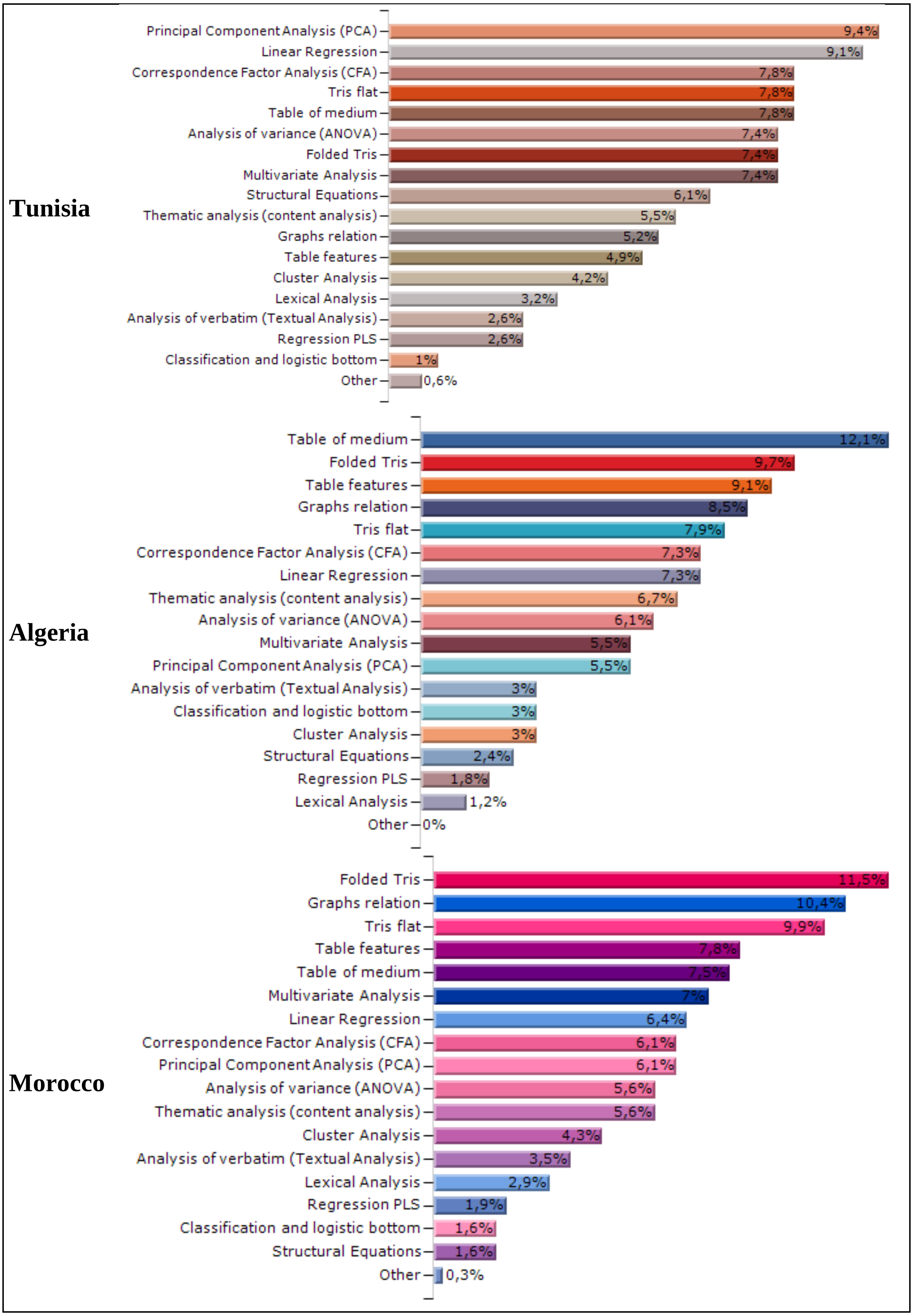
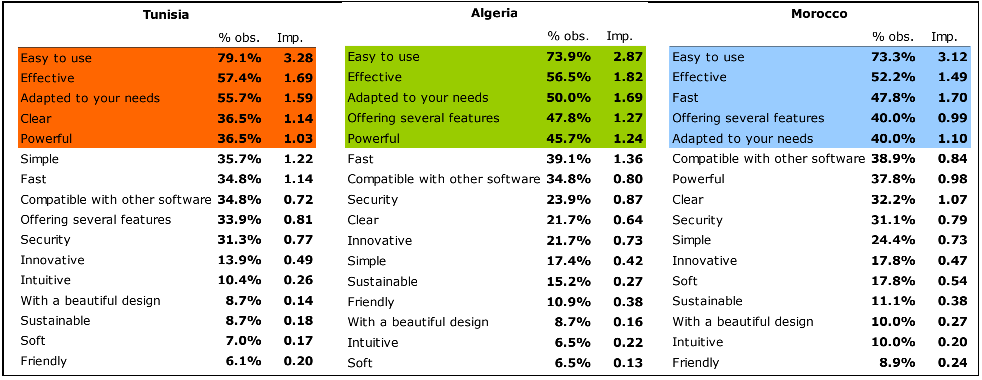


Figure 6: Data-analysis methods used in the Maghreb countries

- *Main expectations of professionals and academics regarding survey and data-analysis software*

Through this survey, we attempted to collect the respondents' opinions concerning the qualities a good survey and data-analysis software should have. To do so, we presented respondents with an exhaustive list of adjectives which might be used to describe the tool, subject of our investigation, and we asked them to choose and rank the five adjectives which, according to them, best describe a good survey and data-analysis software.

The results of this investigation reveal that users of all three countries have similar opinions, albeit with certain nuances. Actually, Tunisians, Algerians and Moroccans agree that a good survey and data-analysis software must be, first and foremost, easy to use and effective. For Algerians and Tunisians, suitability to needs ranks third; therefore, it is deemed more important than speed, which ranks third in the Moroccan classification. Too, Algerians and Moroccans place the quality "offers several functionalities" in fourth position, whereas Tunisians believe the fourth position must be assigned to the clarity of the survey and/or data-analysis software. Finally, performance ranks as the fifth trait a good survey and/or data-analysis software must have for Tunisians and Algerians, while Moroccans rank suitability to needs fifth (Figure 7).



Tunisia			Algeria			Morocco		
	% obs.	Imp.		% obs.	Imp.		% obs.	Imp.
Easy to use	79.1%	3.28	Easy to use	73.9%	2.87	Easy to use	73.3%	3.12
Effective	57.4%	1.69	Effective	56.5%	1.82	Effective	52.2%	1.49
Adapted to your needs	55.7%	1.59	Adapted to your needs	50.0%	1.69	Fast	47.8%	1.70
Clear	36.5%	1.14	Offering several features	47.8%	1.27	Offering several features	40.0%	0.99
Powerful	36.5%	1.03	Powerful	45.7%	1.24	Adapted to your needs	40.0%	1.10
Simple	35.7%	1.22	Fast	39.1%	1.36	Compatible with other software	38.9%	0.84
Fast	34.8%	1.14	Compatible with other software	34.8%	0.80	Powerful	37.8%	0.98
Compatible with other software	34.8%	0.72	Security	23.9%	0.87	Clear	32.2%	1.07
Offering several features	33.9%	0.81	Clear	21.7%	0.64	Security	31.1%	0.79
Security	31.3%	0.77	Innovative	21.7%	0.73	Simple	24.4%	0.73
Innovative	13.9%	0.49	Simple	17.4%	0.42	Innovative	17.8%	0.47
Intuitive	10.4%	0.26	Sustainable	15.2%	0.27	Soft	17.8%	0.54
With a beautiful design	8.7%	0.14	Friendly	10.9%	0.38	Sustainable	11.1%	0.38
Sustainable	8.7%	0.18	With a beautiful design	8.7%	0.16	With a beautiful design	10.0%	0.27
Soft	7.0%	0.17	Intuitive	6.5%	0.22	Intuitive	10.0%	0.20
Friendly	6.1%	0.20	Soft	6.5%	0.13	Friendly	8.9%	0.24

Figure 7: Qualities of a good survey and data-analysis software in Maghreb countries

- *Academic Training Versus Professional Training*

The results revealed that Tunisian, Algerian and Moroccan academic training offers students more theoretical courses than practical ones in statistics. Indeed, most respondents, regardless of nationality, stated they had taken more courses on statistical data-analysis methods (over 30%) and survey techniques (nearly 30%) than training courses on the use of statistical tools (21.1% of Tunisians, 17.8% of Algerians, and 23.4% of Moroccans). At first sight, this suggests that Maghrebian organizations invest little in the training of their executives and civil servants in the use of survey and data-analysis software.

3.3 Comparative analysis of uses and practices in the three Maghreb countries

Considering what has been stated previously regarding the type of studies carried out, the analysis methods used, and the fame of survey software, we can infer that the fame of a software depends on the type of studies and the methods used. Thus, in Tunisia, where investigative surveys are the most frequent ones, meaning that modelling is required and the analyses methods used are multivariate and explanatory, the best-known software are AMOS, LISREL, STATISTICA and SPSS. Meanwhile, in Algeria and Morocco, where the surveys which are conducted tend to be more professional, the analysis methods used are descriptive (univariate analyses, ranking), bivariate explanatory (cross analysis, data-mining), or textual, and familiar solutions are significantly related to survey software (cf. Figure 8). It bears reminding that these allow the user to manage the entire survey process, from questionnaire design to result analysis, and including the data-collection phase.

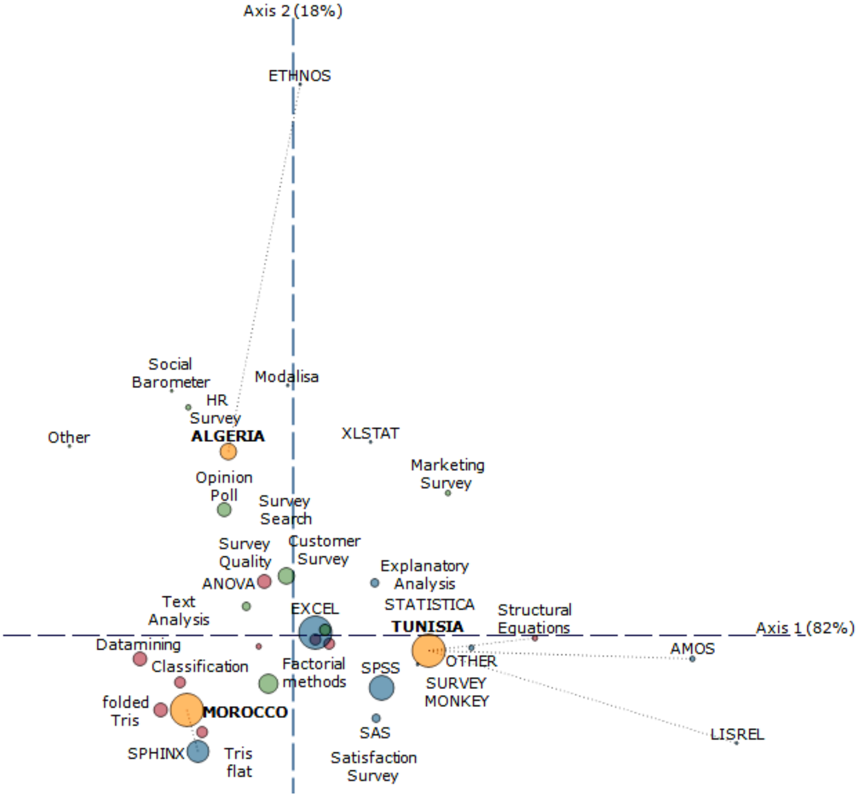


Figure 8: Synthesis of survey types, analysis methods and known software

4. COMMENTS AND CONCLUSION

This work was intended to study the importance professionals and academics in Tunisia, Algeria and Morocco assign to survey and data-analysis software. To do so, we prepared an online survey. We were able to collect responses in all three aforementioned countries (115 in Tunisia, 45 in Algeria, and 90 in Morocco). Although the survey cannot be deemed representative, it did afford us an overview of the use of survey and data-analysis software in three rather similar consumption environments, therefore allowing us to draw a comparison between the three countries in terms of use of solutions that permit designing the questionnaire and analysing the results. What is revealed is a diversity in choices of survey and/or data-analysis software. For example, despite the appearance of many other survey and data-analysis software (questionnaire design and distribution, quantitative and qualitative analyses), SPSS remains the software relatively most used in Tunisia regardless of the user's field of activity, even though it only covers the data-analysis phase, whereas in Algeria and Morocco, Sphinx enjoys greater renown and is most used. Most respondents, whether Tunisian, Algerian or Moroccan, based their choice on having participated in a course at university. They agree that the first two qualities a good survey and data-analysis software should have are, in order: "Easy to handle", and "Effective". While most Tunisians adopt more advanced data-analysis methods, Algerians and Moroccans tend to use more "classic" analysis methods. Finally, we noted that training, whether at the university or during one's professional career, is insufficient where the survey and/or data-analysis software are concerned. A survey that would reveal this deficiency in the educational system of the three target countries would prove most useful.

REFERENCES

1. Bayarassou O., Boughzala Y., El Harouchi L. (2012), Quel usage des logiciels d'enquêtes au Maghreb ? Une étude comparative entre l'Algérie, le Maroc et la Tunisie », 1ère Rencontre de Carthage sur la Statistique, 18-20 octobre 2012, Hammamet, Tunisia.
2. Danaguezian, G. (2002). La nouvelle génération de logiciels d'enquêtes et de reporting (The next generation in survey and reporting software), Survey-Magazine, 2nd quarter, 13-8.
3. Evrard, Y., Pras B., E Redhead (2003). Market : Etudes et Recherches en Marketing (Market: Marketing Studies and Research), Dunod, 3rd edition, Paris.
4. Ganassali, S. (2007). Les enquêtes par questionnaire avec Sphinx (Surveys by questionnaire with Sphinx), Pearson Edition, Paris.
5. Jolibert A. and Jourdan, P. (2006). Marketing Research: Méthodes de recherche et d'études en Marketing (Marketing research: Investigation and study methods in Marketing), Dunod, Paris.
6. Moscarola, J. (1990). Enquête et analyse de données (Survey and data analysis), Vuibert, Paris.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 165-174

A STATE SPACE FRAME WORK FOR MODELING AND FORECASTING TIME SERIES

Muhammad Kashif and **Muhammad Inayat Khan**
Department of Mathematics & Statistics, University of Agriculture, Faisalabad, Pakistan.
Email: mkashif@uaf.edu.pk; biometry@hotmail.com

ABSTRACT

This study determines that state space frame work is a better representation of autoregressive moving average models and its parameter estimation. For this purpose unknown parameters are estimated under the maximum likelihood (ML) method by using kalman filter algorithm. We compared the results obtained by state space model with the results of conventional Box-Jenkings models and found that state space model is more effective as compared to autoregressive models. The proposed state space representation is used for one step ahead forecasting. The result indicates that the state space representation has the good capacity of forecasting while the other model underestimates the series. This paper demonstrates the potential of the some generated series from the state space analysis that proclaims state space frame work model is more practical for converting autoregressive models on the behalf of its several advantages.

KEYWORDS

State space; Kalman filter; Forecasting; Modeling; Estimation.

1. INTRODUCTION

The modeling and forecasting of time series data was very popular and widely used in almost every field of study. Model selection is very interesting and difficult aspect in time series analysis. Following conventional time series methods, most series can be modeled from the class of autoregressive moving average (ARMA) models. There is a large literature on time series modeling and forecasting techniques but most influential work was carried out by Box and Jenkings[1]. Although the Box Jenkings methodology is well established and provides a lot of forecasting models. But the choice of good forecast model is an arbitrary one.

For the last decade, many researchers [2-5] have studied the state space models and argued that these models have had a profound impact on time series and related area. In start, this approach is used to capturing unobserved components of the models and its prediction [6-7]. Nowadays, State Space Models is frequently used for handling stationary and non-stationary time series [8-10]. In spite of fact that state space modeling is so extended, we are going to emphasis in our study on state space representation of ARMA models. As large number of time series can be represented by state space models so this presentation permits simple analysis of process. The behavior of the original series is easily determined by state equation with the help of observation process. Now state space modeling is widely used in place of ARMA and ARCH/GARCH models. The beauty of this approach is that it allows random variation in its components and overcomes the deficiency of classical decomposition models and structural time series models[3]. Many scientists [8,10-11] concluded that the state space muddling approach is very useful technique for converting the ARMA process and any general linear models into a form suitable for recursive estimation and forecasting. In relevance to the application of state space models, the objectives of the study are two fold: first to determine a link between the state space and ARMA models and second, to use this model to forecast the series. At the end the forecasting accuracy of proposed model compared with conventional Box-Jenkins models.

2. STATE SPACE MODELS

The state space model consists of two processes: the unobserved process and the observation process. A state space model applies to a time series is given by

$$Y_t = Z_t\alpha_{t-1} + \eta_t. \tag{1}$$

$$\alpha_t = T_t\alpha_{t\text{-}1} + \varepsilon_t. \tag{2}$$

where Y_t represents the given series and α_t is a state vector. The first equation is called measurement equation and second equation is known as transition equation. In measurement equation the term $Z_t\alpha_{t-1}$ describes the effect of past on the given series. Where as in transition equation, the term $T_t\alpha_{t\text{-}1}$ shows the effect of the past on the current state α_t. The disturbances in both equation (ε_t and η_t) are iid random vectors satisfying [$\varepsilon_t \sim N(0,H_t)$] & [$\eta_t \sim N(0,Q_t)$]. To make the model more applicable, let the disturbances { ε_t } and { η_t } are correlated with each other then we consider the above model as

$$Y_t = Z_t\alpha_{t-1} + \varepsilon_t \tag{1}$$

$$\alpha_t = T_t\alpha_{t\text{-}1} + g\varepsilon_t. \tag{2}$$

The error term in eq 1 describes the unpredictable part of Y_t and $g\varepsilon_t$ shows the unpredictable change in state variable.

3. STATE SPACE REPRESENTATION OF ARMA MODELS

In this section we examine the links between the state space and autoregressive integrated moving average (ARIMA) models. The basic relations are; an ARMA model can be put into state space form in infinite ways; and for a given state space models, there exits an ARMA model.

3.1 State Space Models to ARMA Model

The ARMA (p,q) model can be written as

$$Y_t = \sum_{i=1}^{p}\phi_i Y_{t-i} + \varepsilon_t - \sum_{j=1}^{q}\theta_j\varepsilon_{t-j}. \tag{3}$$

With the of back shift operator L, the ARMA (p, q) model is commonly written as

$$\phi(L)Y_t = \theta(L)\varepsilon_t. \tag{4}$$

where $\varphi(\mathrm{L}) = 1 - \varphi_1\mathrm{L}\text{-}\varphi_2\mathrm{L}^2 \text{-} \cdots\cdots \varphi_p\mathrm{L}^p$ and $\theta(\mathrm{L}) = 1 - \theta_1\mathrm{L}\text{-}\theta_2\mathrm{L}^2 \text{-} \cdots\cdots \theta_q\mathrm{L}^q$

In practice, most series are non-stationary. To achieved the stationary process, the commonly used operation is differencing. For a difference series, the above ARMA model can be extended as

$$\phi(L)(1-L)^d Y_t = \theta(L)\varepsilon_t. \tag{5}$$

In the above ARIMA process, the autoregressive part is divided into two parts: the standard AR polynomial and differencing operations. The appropriate model for the original series can be writing as ARIMA (p,q,d);

$$\eta(L)Y_t = \theta(L)\varepsilon_t. \tag{6}$$

Where $\eta(L) = \phi(L)(1-L)^d$

The transition equation in the above state space models can be writing as

$$\begin{aligned} \alpha_t &= T\alpha_{t\text{-}1} + g\varepsilon_t. \\ (I - TL)\alpha_t &= g\varepsilon_t \end{aligned} \tag{7}$$

Multiply both sides of equation 7 by its ad joint we get

$$\det(I - TL)\alpha_t = Adj(I - TL)g\varepsilon_t$$

$$\det(I - TL)\alpha_{t-1} = Adj(I - TL)g\varepsilon_{t-1} \tag{8}$$

Now consider the measurement equation . By applying operator $\det(I - TL)$ we get

$$\det(I - TL)Y_t = Z\det(I - TL)\alpha_{t-1} + \det(I - TL)\eta_t.$$

By substituting in the value from eq (8) we get

$$\begin{aligned} \det(I - TL)Y_t &= Z\ adj(I - TL)g\eta_{t-1} + \det(I - TL)\eta_t. \\ \det(I - TL)Y_t &= [Z\ adj(I - TL)gL + \det(I - TL)]\eta_t. \end{aligned}$$

which is clearly as ARIMA models of the type $\eta(L)Y_t = \theta(L)\varepsilon_t.$,where

$$\eta(L) = \det(I - TL) \text{ and } \theta(L) = [Z\ adj(I - TL)gL + \det(I - TL)]\eta_t.$$

3.2 ARMA Model to State Space Models

It will now be shown that any ARMA model can be reformulated as state space model. For the above general ARMA model, Let m=max{p,q}.

Define $\phi_j = 0$ for $j > p$ and $\theta_j = 0$ for $j > q$. the model can be written as

$$Y_t = \sum_{j=1}^{m} \phi_j Y_{t-j} + \varepsilon_t - \sum_{j=1}^{m} \theta_j \varepsilon_{t-j}$$

Using, $\psi_{j,t-j}$ be weights and calculated with information available at periods *t-j*, and is defined as

$$\psi_{j,t-j} = \sum_{j=i}^{m} (\phi_j Y_{t-j} - \theta_j \varepsilon_{t-j}) \qquad (9)$$

Note that $\psi_{j,t} = 0 \vee j > m$. So

$$Y_t = \psi_{1,t-1} + \varepsilon_t \qquad (10)$$

Combining equation (9) and (10) we get

$$\psi_{j,t-j} = \sum_{j=i}^{m} (\phi_j \psi_{1,t-i-1} + (\phi_j - \theta_j)\varepsilon_{t-j})$$

So that

$$\psi_{j,t-j} = \psi_{j+1,t-j-1} + \phi_j \psi_{1,t-j-1} + (\phi_j - \theta_j)\varepsilon_{t-j}$$

or

$$\psi_{j,t} = \psi_{j+1,t-1} + \phi_j \psi_{1,t-1} + (\phi_j - \theta_j)\varepsilon_t$$

Consequently, the ARMA model can be written as

$$Y_t = \psi_{1,t-1} + \varepsilon_t$$

$$\psi_{i,t} = \phi_j \psi_{1,t-1} + \psi_{j+1,t-1} + (\phi_j - \theta_j)\varepsilon_t, \vee \mathrm{j} = 1, 2, ...m$$

Thus following (5), the above ARMA process can be written in state space form as

$$Y_t = Z\ \alpha_{t-1} + \varepsilon_t.$$

$$\alpha_t = T\ \alpha_{t\text{-}1} + g\varepsilon_t.$$

$$Z = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \ \mathrm{T}= \begin{bmatrix} \phi_1 & I_{m-1} \\ \vdots & \\ \phi_m & 0 \end{bmatrix} \text{ and g=} \begin{bmatrix} \phi_1 - \theta_1 \\ \vdots \\ \phi_m - \theta_m \end{bmatrix}$$

where

4. ESTIMATION OF STATE SPACE MODELS

For the estimation of the unknown parameters, the basic requirement is to calculate the Loglikelihood function. To maximize the loglikelihood function, kalman filtering is mostly used in state space models. Kalam filter [11-13] consists of recursive equation that updates the information. The function of kalman filter is to decompose an observation into conditional mean and predictive residual sequentially. It should be noted that the kalman filter used the least square principle not normality. The following set of equations respectively constitute the celebrated Kalman Filter equations (see [2] for details):

$$\hat{\alpha}_{t-1} = T_t \hat{\alpha}_{t-1}$$

$$P_{t-1} = T_t P_{t-1} T_t' + H_t$$

$$\hat{\alpha}_t = \hat{\alpha}_{t-1} + K_t (Y_t - Z \hat{\alpha}_{t-1})$$

$$P_t = (I - K_t Z) P_{t-1}$$

$$K_t = P_{t-1} Z'(Z P_{t-1} Z' + Q_t)^{-1}$$

The first two set of equations is used as prediction equations. The Kalman filter predicts the state and uncertainty by using these two equations. The last three equations are known as updating or correction equations

5. APPLICATION TO PAKISTAN'S RGDP DATA

As an illustration, we use the data of gross domestic product of constant prices of FY-2000 in billion rupees (RGPD). The sample consists of 31 values over the period 1978 to 2008. The data was obtained from State Bank of Pakistan. The data used is annually calculated based on fiscal year system of Pakistan. For analysis purpose we used the log of RGPD data. In the first stage of analysis, the series was analyzed by using Box-Jenkings methodology. The purpose of this analysis is to select the appropriate ARMA model that best describes the series and provide accurate forecasting. The original series is found non-staionary after observing autocorrelations and partia autocorrelations graphs up to 20th lag. The stationary is achieved after taking first difference i.e. d=1 and verified by observing the patterns of correlogram. However to save space, the correlogram is not presented. Finally ARIMA(1,1,1) is defined as the best models which is given by

$$(1-L)(1-\phi_1 L)Y_t = (1-\theta_1 L)\varepsilon_t.$$

The polynomial operators for this models are

$$\eta(L) = 1-(1+\phi_1)+\phi_1 L^2,$$

$$\theta(L) = 1-\theta_1 L - 0L^2.$$

Thus the state space representation of this models is given by

$$\begin{pmatrix} \alpha_{1t} \\ \alpha_{2t} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{1(t-1)} \\ \alpha_{2(t-2)} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \varepsilon_{2t} \end{pmatrix}$$

$$Y_t = \begin{pmatrix} Y_{t-1} & 1 \end{pmatrix} \alpha_t + \eta_t$$

where $\alpha_{1t} = \phi_{1(t)}$ and $E(\eta_t) = E(\varepsilon_{2t}) = 0 \,\&\, E(\varepsilon_{2t})^2 = q2, E(\eta_t)^2 = R$ and $Q = \begin{pmatrix} 0 & 0 \\ 0 & q2 \end{pmatrix}$

For the estimation of above state space model, the kalman filter technique is used to calculate the loglikelihood function via prediction error decomposition. With the Eview software[14], results obtained from the state space representation of ARIMA(1,1,1) model is reported in Table 1.

Table 1: Estimation of the Parameters of ARIMA and State Space Model (SSM).

	ARIMA[1,1,1]	SSM
ϕ_1	0.9866	0.9940
θ_1	-0.7362	-0.6711
C(3)	-	-8.005
SV1	-	0.1683
SV2	-	0.1692
LogLikelihood	75.258	76.048
S. E of Regression	0.018	-
AIC	5.053	4.869
SC	4.958	4.729

By using Maquardt Optimization algorithm, state space model has been fitted on 31 observations. To achieve a converging solution, there required 10 iterations. The maximum value of log likelihood is 76.048 at convergence. The coeifficient c (3) denotes the variance of error terms. The estimate of the standard deviation of the error terms is $sqrt(\exp(-0.0187)) = 0.018$, which is closed to Standard Error of Regression reported in ARIMA(1,1,1) model. Further the estimate of the fitted ARIMA model was compared with state space models. The results shows that MLE's produced by state space models are closed to the conditional MLE's. It is observed that the maximum log-likelihood value form SSM is slightly higher than ARIMA log-likelihood value. The values of AIC and SC shows that state space modeling technique performs better than ARIMA approach. Some of the important series for the analysis of the data are summarized below. The plot of the standardized auxiliary residuals of the state variable is depicted in figure 1. According to (11, 15) these residuals are useful for detecting outlier and structural changes in components of the model. We cannot observe any strong or weak structural breaks and outliers in the RGPD series analysis.

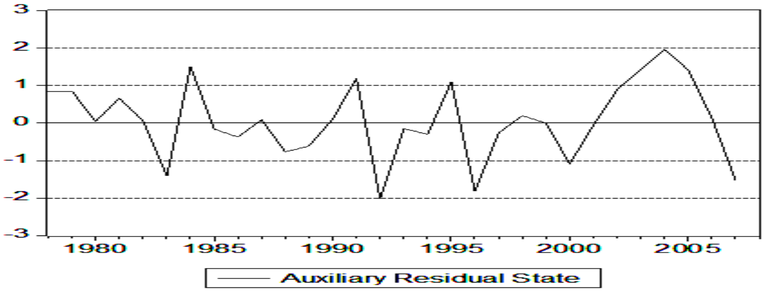


Fig. 1: Graph of Auxiliary Residual of the State Variable

The filtered state estimates of the state vector from the state space model are given below in figure 2(a-b).

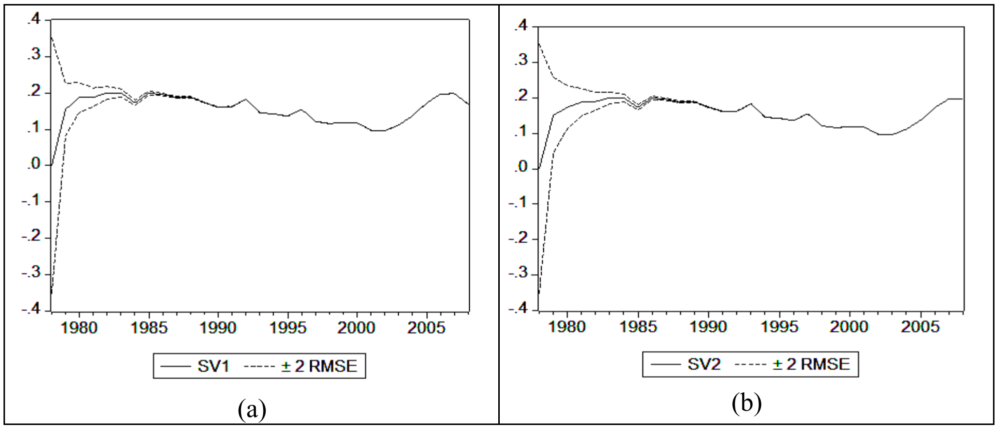


Fig. 2: Graph of Filtered State Estimate from State space models

The filtered state estimate are very similar to the fitted estimate of the ARIMA(1,1,1) models. The figure 3 shows the one-step ahead response (signals). From this figure it is very clear that state space model track the actual output fairly well.

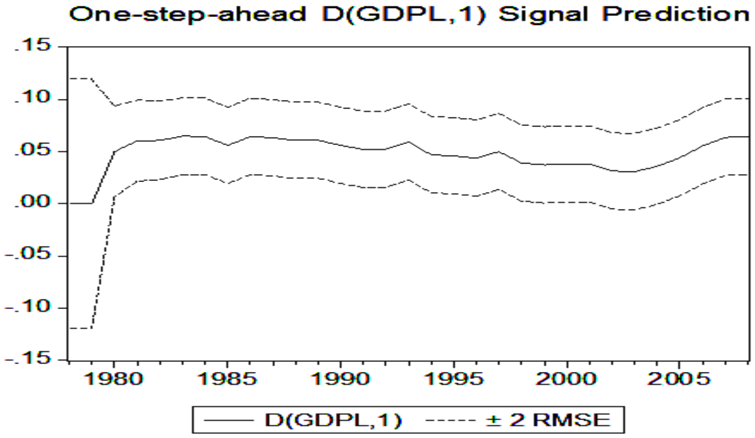
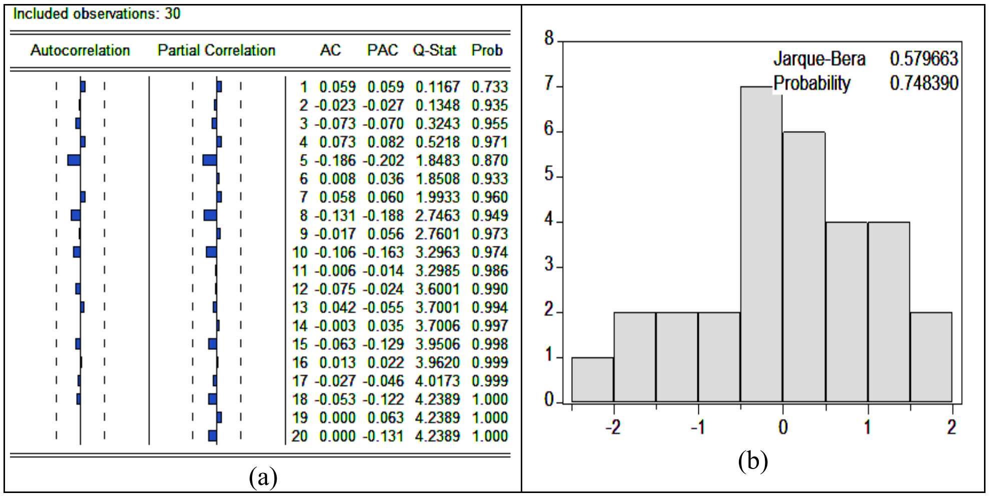


Fig. 3: Graph of One-Step Ahead Response (Signal)

At the end the figure 4(a-b) shows the correllogram with Box-Lung statistics and Jarque-Bera normality test for the standardized prediction error. No serial correlation observed from this figure.



Included observations: 30

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.059	0.059	0.1167	0.733
		2	-0.023	-0.027	0.1348	0.935
		3	-0.073	-0.070	0.3243	0.955
		4	0.073	0.082	0.5218	0.971
		5	-0.186	-0.202	1.8483	0.870
		6	0.008	0.036	1.8508	0.933
		7	0.058	0.060	1.9933	0.960
		8	-0.131	-0.188	2.7463	0.949
		9	-0.017	0.056	2.7601	0.973
		10	-0.106	-0.163	3.2963	0.974
		11	-0.006	-0.014	3.2985	0.986
		12	-0.075	-0.024	3.6001	0.990
		13	0.042	-0.055	3.7001	0.994
		14	-0.003	0.035	3.7006	0.997
		15	-0.063	-0.129	3.9506	0.998
		16	0.013	0.022	3.9620	0.999
		17	-0.027	-0.046	4.0173	0.999
		18	-0.053	-0.122	4.2389	1.000
		19	0.000	0.063	4.2389	1.000
		20	0.000	-0.131	4.2389	1.000

(a)

(b)

Fig. 4: Normality Test for the Standardized Prediction Error

Finally, the one-step ahead forecast (1980-2015) from the state space models, the identified ARIMA(1,1,1) model and the actual values are plotted in figure 5 below. The graphs clearly shows that state space model outperform the ARIMA model and shows good potential of the forecasting.

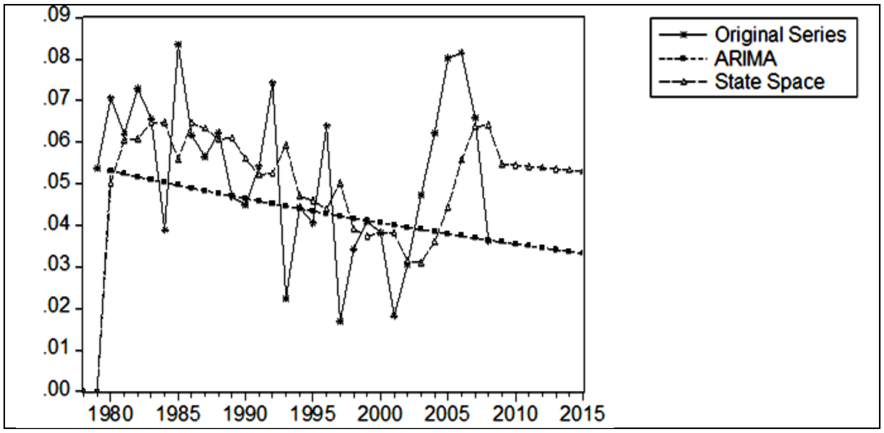


Fig. 5: Original and Forecast Series.

5. CONCLUSION

The paper weighs against the state space model and autoregressive models. It is concluded that state space approach is very effective and useful technique for representing ARIMA models. The results showed that state space model have a good potential of the forecasting and structural analysis of the problems. Moreover state space methodology can easily observe and model the different inherent features in a series such as trend, seasonal, cyclic variation. Whereas Box-Jenkings methodology is based on the elimination of these component for the purpose of stationary. The missing observations easily incorporated into state space models, however, in the box-jenkings models; it is relatively difficult to handle missing observations. In wrapping up the research, the state space models are more general, flexible and more transparent than Box-Jenkings ARIMA modeling strategy.

REFERENCES

1. Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control* (3rd edn). Englewood Clifs, New Jersey: Prentice-Hall.
2. A.C. Harvey. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
3. Brockwell, P.J. and R.A. Davis. (2000). *Time Series: Theory and Models*. Springer-Verlag, New York.
4. Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and Its Applications*, Springer- Verlag, New York.
5. Hydman, R.J., A.B. Koehler, J.K. Ord, R.D. Snyder. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer-Verlag Berlin Heidelberb.
6. H. Akaike, (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals of the Institute of Statistical Mathematics*.26, 363.
7. G. Kitagawa, and W. Gersch, (1984). A smoothness priors-stat space modeling of time series with trend and seasonality. *J. Amer. Statist. Assoc.,* 79, 378.
8. Sastri, T. (1985). Astate space Modelling Approach for Time Series Forecasting. *Management Sciences*, 13(11), 1451-1470.
9. Adnan, S.H., Bukhari, A.S. and Khan, S. (2008). Estimating Output Gap for Pakistan Economy: Structural and Statistical Approach. *SBP Research Bulletin*, 4(1), 31-60.
10. Kashif, M. Shehzad, I. Bokhari, S.M. and Niyyar, M. (2008). Results of Interbank Exchange Rates Forecasting using State Space Model. *Pak. J. Statist. and Opera. Res.*, IV(2), 111-119.
11. Asemota , O.J. 2010. Understanding the Kalman Filter: A Classical Approach with an Application. *Euro. J. Scientific Res.*, 45(1), 006-015.
12. Welch, G. and Bishop, G. (2001). *An Introduction to the Kalman Filter*, Chapel Hill SIGGRAPH.
13. Ozbek, L. and Ozelale, U. (2005). Employing the Extended Kalman Filter in Measuring the Output Gap. *Journal of Economic Dynamics & Control*, 29, 1611-1622.
14. Bossche, F.A.M. (2011). Fitting State Space Models with Eviews. *Journal of Statistical Software*, 41(8), 1-16.
15. Harvey, A.C. and Koopman, S.J. (1992). Diagnostic Checking of Unobserved Components Time Series Models. *Journal of Business and Economics Statistics*, 10(4), 337-389.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 175-182

CURE FRACTION ESTIMATION FOR MIXTURE CURE MODEL WITH CENSORED DATA

Noor Akma Ibrahim[1,2], **Fauzia Ali Taweab**[1], **Jayanthi Arasan**[2]
and **Mohd Rizam Abu Bakar**[2]

[1] Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia
Email: nakma@putra.upm.edu.my; fauziataweab@yahoo.com

[2] Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia, Malaysia

ABSTRACT

This article investigates the cure rate estimation based on mixture cure model (MCM) when left, right and interval censored data are observed. The MCM is proposed based on lognormal distribution that incorporates the effects of covariates on the probability of cure. Maximum likelihood estimators of the model parameters are obtained by implementing the EM algorithm. The performance of the estimates is assessed through simulation study under various conditions

KEYWORDS

Cure fraction, interval censoring, mixture cure model, lognormal distribution, MLE method, EM algorithm.

1. INTRODUCTION

Due to advances in cancer treatment and health care, in clinical studies there are patients insusceptible to the occurrence of the interested event (relapse or date). The proportion of such patients is considered as cured fraction, and survival models that account for cure is known by cure models. These models can be broadly classified into standard mixture cure model or bounded cumulative hazard model. The mixture model was constructed by Berkson and Gage (1952) and has been widely discussed by several authors, including, Farewell (1982), Kuk and Chen(1992), Sy and Taylor(2000), Peng and Dear(2000), Price and Manatunga(2001), Banerjee and Carlin(2004), and many others. In this model the survival function for entire population is given by:

$$S(t) = p + (1 - p)S_u(t), \tag{1}$$

where p is the probability of cure and $S_u(t)$ is the survival function for uncured subjects.

The other class of cure models is Bounded Cumulative Hazard (BCH) model. This model was proposed as an alternative class of cure modelsby Chen, et al., in 1999, to handle some drawbacks of mixture model which have been discussed also by the same authors. The BCH model was developed based on the assumption that the patients suffering from cancer entering a clinical trial and after treatment a patient is left with N cancer cells capable of metastasizing, which may grow rapidly and replace the normal tissue (i.e. cancer relapse), (Tsodikovet al.(2003)). N assumed to follow a Poisson distribution with mean θ. The survival function $S(t)$ for the entire population is given by:

$$S(t) = P(no\ cancer\ by\ time\ t\ ) \ = exp\big(-\theta F(t)\big) \tag{2}$$

The cure models can be considered to be parametric or nonparametric and this depends on whether the distribution of failure time for uncured patients is specified or not. Both parametric and semi parametric mixture models have been discussed by a number of authors. Berkson and Gage (1952) used a mixture of exponential distributions with a constant cure fraction. This parametric method was extended by Farewell (1982) to Weibull regression for modelling survival function of uncured patients and logistic regression to model probability of cure.Kuk and Chen(1992) proposed a proportional hazard cure model, where they used the proportional hazard regression to model the survival function of uncured patients and logistic regression for modeling the probability of cure. Peng and Dear(2000) and Sy and Taylor(2000) developed alternative methods for computing the semi parametric likelihood function in the proportional hazard cure model. Recently, Aljawadi et al.(2011) proposed parametric and nonparametric approach to estimate the cure fraction based on the BCH model. The works for both the mixture and the BCH models are largelybased on right-censored data, and a few of the studies focus on interval-censored.

Interval-censored lifetime data arise when an event occurs within some interval of time but the exact time of the event is unknown (Kalbfleisch and Prentice, 2002). Also, extensive statistical studies have been done on interval data, but without involving the possibility of cure. The parametric models with interval censoring were first studied by Flygare et al. (1982), where they used a two parametric Weibull distribution, not depending on covariates and based on maximum likelihood estimates and midpoints. Odell et al. (1992) described a Weibull regression model for interval data with covariates. Lindsey (1998) proposed a variety of parametric models that can be used to obtain smooth representations of the survival BCH model. There are few studies that focused on both interval censored and possibility of cure. Kim and Jhun (2008) have proposed an approach to analyse the mixture cure model for interval censored. Aljawadi et al. (2011) proposed parametric BCH model to estimate the cure fraction based on Weibull regression for modelling survival time of uncured group considering no covariates related to the probability of cure. (with only “a constant fraction of cure” estimated for the cure parameter). The main purpose of this article is to present an approach which analyses the cure model in the presence of left, interval and right censoring.

2. METHODOLOGY

In parametric maximum likelihood estimation method the survival function $\boldsymbol{S(t)}$ and probability density function $\boldsymbol{f(t)}$ for the entire population are known.In this paper, we used the standard cure model based on lognormal distribution. We further assume that the probability of being cured depends on a set of covariates $\boldsymbol{p(\beta, X)} = \frac{\boldsymbol{exp\ (\beta' X)}}{\boldsymbol{1+exp\ (\beta' X)}}$. The related functions are,

$$\boldsymbol{f_u(t)} = \frac{\boldsymbol{1}}{\boldsymbol{t\sigma\sqrt{2\pi}}} \boldsymbol{exp}\left[-\frac{\boldsymbol{(ln\, t-\mu)^2}}{\boldsymbol{2\,\sigma^2}}\right], \text{and} \boldsymbol{S_u(t)} = \boldsymbol{1-\varphi}\left(\frac{\boldsymbol{ln\, t-\mu}}{\boldsymbol{\sigma}}\right), \tag{3}$$

where $\boldsymbol{\mu}$ is a location parameter, $\boldsymbol{\sigma}$ is a scale parameter, and $\boldsymbol{\varphi}$ is the standard normal distribution function.

Usually the likelihood function for right censored data under the models that account for the possibility of cure ($\boldsymbol{p} > 0$) can be written as

$$\boldsymbol{l_c} = \boldsymbol{log} \prod_{i=1}^{n} [\{\boldsymbol{f_u}(\boldsymbol{t_i})(\mathbf{1} - \boldsymbol{p})\}^{c_i}]^{\delta_i} \, [\{\boldsymbol{p}\}^{\mathbf{1}-c_i}\{(\mathbf{1} - \boldsymbol{p})\boldsymbol{S_u}(\boldsymbol{t_i})\}^{c_i}]^{\mathbf{1}-\delta_i} \, , \qquad (4)$$

where, $\boldsymbol{f_u}(\boldsymbol{t_i})$ and $\boldsymbol{S_u}(\boldsymbol{t_i})$ are the density and survival function for uncured patients respectively, $\boldsymbol{\delta_i}$ is an indicator of censoring with zero if $\boldsymbol{t_i}$ is censored and one otherwise, and $\boldsymbol{c_i}$ is the cure indicator with zero if the patient is cured and one otherwise, (Peng and Dear(2000) and Lu.(2008)).

Old *et al*. (1992), describe a general likelihood function that could contain left, right, and interval censoring, as follows. Interval censoring occurs if instead of observing $\boldsymbol{t_i}$, only an interval $(\boldsymbol{t_{Li}}, \boldsymbol{t_{Ri}}]$ is observed where $\boldsymbol{t_i} \in (\boldsymbol{t_{Li}}, \boldsymbol{t_{Ri}}]$. The *i*th subject is left censored when his survival time is below of certain time $\boldsymbol{l_i}$; $\boldsymbol{t_i} \in (\mathbf{0}, \boldsymbol{l_i}]$. The *i*th subject is right censored when $\boldsymbol{t_i} \in (\mathbf{r_i}, \infty]$. Let us define the following indicator variables in order to identify whether the subject is interval, left or right censored data.

$\boldsymbol{\delta_L} = \mathbf{1}$ if subject is left censored, $\boldsymbol{\delta_I} = \mathbf{1}$if subject is interval and $\boldsymbol{\delta_R} = \mathbf{1}$ if subject is right censored. Note that $\boldsymbol{\delta_R} = \mathbf{1} - (\boldsymbol{\delta_L} + \boldsymbol{\delta_I})$.Then the likelihood function for a sample of $\boldsymbol{n}$ independent observations without considering cure rate is

$$\boldsymbol{l_c} = \boldsymbol{log} \prod_{i=1}^{n} \quad [\mathbf{1} - \boldsymbol{S}(\boldsymbol{l_i})]^{\delta_{Li}} [\boldsymbol{S}(\boldsymbol{t_{Li}}) - \boldsymbol{S}(\boldsymbol{t_{Ri}})]^{\delta_{Ii}} [\boldsymbol{S}(\boldsymbol{r_i})]^{\delta_{Ri}} \qquad (5)$$

In the cure model, another indicator, $\boldsymbol{c_i}$, is defined for demonstrating that the patient is cured or not, that is$\boldsymbol{c_i}$ is zero if the patient is cured and one otherwise. Now, we can reformat the censoring indicator as follows, $\boldsymbol{\delta} = \boldsymbol{I}(\boldsymbol{t_{Ri}} < \infty)$ and then the likelihood of (5) for the cure model with $\boldsymbol{p}$, can be written as

$$l_c = log \prod_{i=1}^{n} [\{(1 - p)F_u(l_i)\}^{c_i}]^{\delta_i} [\{(1 - p)(F_u(t_{Ri}) - F_u(t_{Li}))\}^{c_i}]^{\delta_i}$$
$$\times \, [\boldsymbol{p}^{\mathbf{1}-c_i}\{(\mathbf{1} - \boldsymbol{p})(\, \mathbf{1} - \boldsymbol{F_u}(\boldsymbol{r_i}))\}^{c_i}]^{\mathbf{1}-\delta_i}, \qquad (6)$$

where $\boldsymbol{f_u}(\boldsymbol{t_i})$ and $\boldsymbol{F_u}(.)$ are density and cumulative distribution function of uncured group respectively. Obviously, if $\boldsymbol{\delta} = \mathbf{1}$, then $\boldsymbol{c_i} = \mathbf{1}$, but if $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{c_i}$is not observable and it can be one or zero. Therefore $\boldsymbol{c_i}$ are partially observed and this will lead to the implementation of the EM algorithm to estimate unknown parameters.

3. PARAMETERS ESTIMATION

Our problems now is to estimate the unknown parameters $\mu, \sigma,$ and β_j. For simplicity, we assume that $t_1, \ldots, t_{n1}$ are left censored, $t_{n1+1}, \ldots, t_{n2}$ are interval censored, and $t_{n2+1}, \ldots, t_n$ are right censored. Based on the above observations and lognormal distribution the complete log-likelihood function takes the following form,

$$l_c = \sum_{i=1}^{n1}\left[log(1-p_i) + log(\varphi\left(\frac{ln\, l_i - \mu}{\sigma}\right))\right] + \sum_{i=n1+1}^{n2}\left[log(1-p_i) + log(\varphi\left(\frac{ln\, t_{Ri} - \mu}{\sigma}\right) - \varphi\left(\frac{ln\, t_{Li} - \mu}{\sigma}\right))\right] + \sum_{i=n2+1}^{n3}\left[c_i(log(1-p_i) + log(1-\varphi\left(\frac{ln\, r_i - \mu}{\sigma}\right))) + (1-c_i)logp_i\right] \quad (7)$$

The MLE's are obtained by considering this as a missing data problem and then used the EM algorithm. In the E step of the EM algorithm, we compute the expected value of the log likelihood function based on the missing observations, which can be obtained by assigning g_i as the expectation of c_i conditional on the current estimates of β and the survival function of uncured patients $S_u(t_i)$, see Peng and Dear(2000). Then the expected value of the log likelihood can be written as follows

$$l_c = \sum_{i=1}^{n1}\left[log\left(\frac{1}{1+e^{\beta x_i}}\right) + log(\varphi\left(\frac{ln\, l_i - \mu}{\sigma}\right))\right] + \sum_{i=n1+1}^{n2}\left[log\left(\frac{1}{1+e^{\beta x_i}}\right) + log(\varphi\left(\frac{ln\, t_{Ri} - \mu}{\sigma}\right) - \varphi\left(\frac{ln\, t_{Li} - \mu}{\sigma}\right))\right] + \sum_{i=n2+1}^{n3}\left[g_i(log\left(\frac{1}{1+e^{\beta x_i}}\right) + log(1-\varphi\left(\frac{ln\, r_i - \mu}{\sigma}\right))) + (1-g_i)\log\left(\frac{e^{\beta x_i}}{1+e^{\beta x_i}}\right)\right] \quad (8)$$

where,

$$g_i = \delta_i + (1-\delta_i)\left[\frac{(1-p(\beta,X))S_u(t_i)}{p(\beta,X)+(1-p(\beta,X))S_u(t_i)}\right] \quad (9)$$

The M step of the EM algorithm involves maximising (8) with respect to the unknown parameters for fixed g_i. Therefore, if $\mu^k, (\sigma^k)^2, \beta_j^k$ are estimates of $\mu, \sigma,$ and β_j at the *k*th iteration, we compute g_i^k and then we can obtain $\mu^{k+1}, (\sigma^{k+1})^2, \beta_j^{k+1}$ by maximizing (8) with respect to $\mu, \sigma,$ and β_j for fixed g_i . Repeat until convergence. The maximization can be obtained by solving the following nonlinear equations iteratively using Newton Raphson method.

$$\frac{\partial L_c}{\partial \mu} = -\sum_{i=1}^{n1}\left[\frac{\emptyset\left(\frac{\ln l_i - \mu}{\sigma}\right)}{\sigma\,\varphi\left(\frac{\ln t - \mu}{\sigma}\right)}\right] - \sum_{i=n1+1}^{n2}\left[\frac{\emptyset\left(\frac{\ln t_{Ri} - \mu}{\sigma}\right) - \emptyset\left(\frac{\ln t_{Li} - \mu}{\sigma}\right)}{\sigma\,(\varphi\left(\frac{\ln t_{Ri} - \mu}{\sigma}\right) - \varphi\left(\frac{\ln t_{Li} - \mu}{\sigma}\right))}\right] + \sum_{i=21+1}^{n}\left[\frac{g_i}{\sigma}\frac{\emptyset\left(\frac{\ln r_i - \mu}{\sigma}\right)}{1-\varphi\left(\frac{\ln r_i - \mu}{\sigma}\right)}\right] \quad (10)$$

$$\begin{aligned}\frac{\partial \mathrm{L_c}}{\partial \sigma^2} &= \sum_{i=1}^{n1}\left[\frac{(\ln l_\mathrm{i} - \mu)}{2\sigma^3}\frac{\emptyset\left(\frac{\ln l_\mathrm{i} - \mu}{\sigma}\right)}{\varphi\left(\frac{\ln l_\mathrm{i} - \mu}{\sigma}\right)}\right] \\ &- \sum_{i=n1+1}^{n2}\left[\frac{\frac{(\ln t_{iR} - \mu)}{\sigma}\emptyset\left(\frac{\ln t_{iR} - \mu}{\sigma}\right) - \frac{(\ln t_{iL} - \mu)}{\sigma}\emptyset\left(\frac{\ln t_{iL} - \mu}{\sigma}\right)}{2\sigma^2(\varphi\left(\frac{\ln t_{iR} - \mu(\gamma,x)}{\sigma}\right) - \varphi\left(\frac{\ln t_{iL} - \mu(\gamma,x)}{\sigma}\right))}\right] \\ &+ \sum\nolimits_{i=n2+1}^{n}\left[\frac{g_i\ (\ln r_i - \mu)}{2\sigma^3}\frac{\emptyset\left(\frac{\ln r_i - \mu}{\sigma}\right)}{1 - \varphi\left(\frac{\ln r_i - \mu}{\sigma}\right)}\right]\end{aligned} \tag{11}$$

$$\frac{\partial L_c}{\partial \beta_j} = -\sum\nolimits_{i=1}^{n}\left[x_{ij}\frac{e^{\sum \beta_j x_{ij}}}{1+e^{\sum \beta_j x_{ij}}}\right] + \sum\nolimits_{i=n2+1}^{n}[\,(1-g_\mathrm{i})x_{ij}] \tag{12}$$

4. SIMULATION STUDY

In this section, study is conducted to investigate the performance of theproposed estimator. Here, we generate a binary covariate from Bernoulli with probability 0.5. The cure indicator is generated from Bernoulli using a logistic regression, $p(\beta, x) = 1/[1 + exp\ (-(\beta_0 + \beta_1 x_i))]$.The survival time for each individual is generated from lognormal distribution with scale parameter μ and shape parameter σ. The interval censored data mechanism followsthat of Kim and Jhun (2008)considering different rate of right censoring, where with right censoring $(\mathrm{r_i}, \infty)$ there are two possibilities for the ith subject: the subject is cured, or the event of interest for that subject occurs after the last examination time. Table1 presents the simulation results based on 500 replication for two different rates and two different values of σ.

Table 1:
Parameters estimates for different censoring proportions and 500 replications with *n* =100 subjects.

		True	EST	*Bias*	SE
Moderate(10-20%) censoring					
$\sigma = 0.6$	β_0	0.3	0.296	0.004	0.097
	β_1	0.2	0.217	-0.017	0.111
	μ	2	2.093	-0.093	0.071
	σ	0.6	0.63	-0.03	0.068
$\sigma = 1$	β_0	0.3	0.278	0.022	0.112
	β_1	0.2	0.147	0.053	0.12
	μ	2	2.157	-0.157	0.116
	σ	1	1.071	-0.071	0.138
Heavy rate(30-40%) censoring					
$\sigma = 0.6$	β_0	0.3	0.318	-0.018	0.206
	β_1	0.2	0.258	-0.058	0.141
	μ	2	2.099	-0.099	2.075
	σ	0.6	0.631	-0.031	0.074
$\sigma = 1$	β_0	0.3	0.434	-0.134	0.504
	β_1	0.2	0.091	0.109	0.276
	μ	2	2.175	-0.175	0.128
	σ	1	1.056	-0.056	0.145

Note: EST is estimate of the parameters. SE is standard deviation of estimating parameter.

Cleary, the proposed estimates have small biases and also the results indicate that the estimation of parameter are better for data sets that have fewer censored individuals and small value of shape parameterσ.

5. CONCLUSION

In this article,we have proposed a cure mixture model for left, interval and right censored data and assuminglognormal distribution for modeling the survival time and logistic regression for the probability of cure.The parameter estimates are obtained by the MLE method with implementation of the EM algorithm, where the performance of the proposed method was investigated via a simulation study. The results of the simulation showed that in the proposed model, the bias and standard errors of the parameters estimateare smaller as censoring rate decreases.

REFERENCES

1. Aljawadi, B.A., Abu Bakar, M. R., and Ibrahim, N. A. (2011). Parametric cure rate estimation based on Exponential distribution which incorporates covariates. *Journal of statistical modeling and analytics*, 2(1), 11-20.
2. Banerjee S, Carlin BP. (2004). Parametric spatial cure rate models for interval censored time-to-relapse data. *Biometrics*, 60, 268-275.

3. Berkson J, Gage RP. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical*, 47, 501-515.
4. Farewell VT. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38, 1041-1046.
5. Flygare , M.E. and Austin , J.A. and Buckwalter, R.M. (1985). Maximum likelihood estimation for the 2-parameter Weibull distribution based on interval data. *IEEE Transactions on Reliability*, 34(1), 57-59.
6. Kalbfleisch, J.D. AND Prentice, R. L. (2002).The Statistical Analysis of Failure Time Data. 2nd ed. New York: John Wiley and Sons, Inc.
7. Kuk AYC and Chen CH. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79, 531-541.
8. Lindsey. J.K. (1998). A study of interval censoring in parametric regression models *Life Data Anal*, 4, 329-345.
9. Peng Y, and Dear KBG. (2000). An on parametric mixture model for cure rate estimation. *Biometrics*, 56,237-243.
10. Odell, P.M., Anderson, K.M. and Agostion R.B.D. (1992).Maximum likelihood estimation for interval censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48, 951-959.
11. Price D.L. and Manatunga A.K. (2001). Modeling survival data with a cured fraction. *Statistics in Medicine*, 20, 1515-1527.
12. Sy J.P, and Taylor J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56, 227-236.
13. Tsodikov, A.D., Ibrahim J.G. and Yakolev, A.Y. (2003).Estimating cure rates from survival data. *J. Amer. Statist. Assoc.*, 98, 1063-1078.
14. Wenbinlu (2008). Maximum likelihood estimation in the proportional hazards cure model. Ann *Inst Stat Math, 60, 545-574.*
15. Yang-Jin Kim and Myoungshic Jhun (2008). Cure rate model with interval censored data. *Statistics in medicine*, 27, 3-14.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 183-192

THE GROWTH PATTERN OF MALAYSIAN SCHOOL CHILDREN AND ADOLESCENTS

Bong Yii Bonn[1], **Asma Ahmad Shariff**[2], **Abdul Majid Mohamed**[2]
and **Amir Feisal Merican**[3]

[1] Institute of Graduate Studies University of Malaya,
Kuala Lumpur, Malaysia. Email: yiibonn@siswa.um.edu.my
[2] Center for Foundation Studies in Science, University of Malaya,
Kuala Lumpur, Malaysia. Email: asma@um.edu.my; ammajid@um.edu.my
[3] Institute of Biological Sciences, Faculty of Science, University of Malaya,
Kuala Lumpur, Malaysia. Email: merican@um.edu.my

ABSTRACT

Background: Growth references are used widely by medical practitioners as well as the public to assess the growth status of children. Although studies have been carried out on height and weight of school children in the 1970s to 1980s in Malaysia, sampling was restricted to urban populations of major cities and not representative of the whole Malaysian population.

Method: A comprehensive nationwide cross-sectional study has been conducted on 27,182 school children (6 to 18 years old) in year 2009 and 2010 using the two-stage stratified random sampling technique. Height and weight values were reported in survey forms based on the measurements taken in accordance with the National Physical Agility Standards for Malaysian School Students programme. The Cole's LMS method was used for calculation of growth percentiles.

Results: The height differed considerably as school children entered adolescence (13 years old and above). Males were taller and heavier than female for most age groups. The 50th percentile for height upsurge with the age of male. However, height values tended to stabilize after the age of 15 years old for female.

Conclusions: There are some differences between the distributions of height and weight of Malaysian school children and adolescents from the CDC growth patterns. Thus, it might be necessary to define new national representative growth references for local evaluation purpose.

KEYWORDS

Growth references, school children, adolescents, Malaysia, Cole's LMS method.

1. INTRODUCTION

Reference of growth has become one of the most adequate guidelines to evaluate the well-being of an individual. Regular assessment of growth in children is important for child health's monitoring. It is a standard component of community pediatric services throughout the world (Panpanich & Garner, 2009). Growth monitoring aims at improving the nutritional status, reducing the risk of death, helping to educate those who concern and leading to early detections of growth disorders (Garner, Panpanich & Logan, 2000).

Doctors in Malaysia have been using the Centres for Disease Control and Prevention growth charts or the World Health Organiztion growth charts in clinical practices to assess the growth of local children. Dugdale (1969), Dugdale et al. (1972) and Chen and Dugdale (1970) have published the height and weight of Malaysian infants and school children in the form of growth charts. The Department of Paediatrics, University of Malaya, reported on the development of growth charts for school children as well in 1987. Although studies have been carried out on such norms in the 1970s to 1980s in Malaysia, sampling was restricted to urban populations of major cities and not representative of the whole Malaysian population. Concerns arose as well for the out-of-datedness of these data.

This paper presented the physical growth of Malaysian school children and adolescents. Besides, their growth patterns were compared to that of American children as documented by the CDC 2000 in order to highlight the importance of this nationwide study.

2. MATERIALS AND METHODS

The design of this study was multi-centric and cross-sectional due to high costs and time constraints. During the planning and preparation period, formal permissions was obtained from the Ministry of Education, Malaysia and all the State Education Departments. This study was conducted in public school both from rural and urban areas encompossing all major regions in Malaysia.

The two-stage stratified random sampling technique was used for sample recruitment. In the first stage, 98 schools (about 1%) from the total number of schools in Malaysia were selected. Two classes were chosen from each grade randomly in the second stage. Lastly, the cluster sampling technique was used whereby all healthy students from the selected classes participated. This comprehensive nationwide study has been conducted on 27,182 school children (6 to 18 years old) in year 2009 and 2010.

The data for height and weight values were reported in study forms based on the measurements taken in accordance with the National Physical Agility Standards for Malaysian School Students programme. This programme was made compulsory during physical education classes by the Ministry of Education since the year 2008 and covers both primary and secondary students. Teachers were to assist lower primary school students to complete the study forms.

Upon completion of the study forms from each study site, data entry was done using MS Excel for Windows. Samples were divided into respective age categories in increments of half-year. Following data cleansing, the Cole’s LMS method was used for calculation of growth percentiles. The Cole’s LMS method suggested construction of reference percentiles adjusted for skewness of the data, and summarizes the changes in height and weight distributions by three curves representing the skewness (L), median (M) and coefficient of variation (S) (Cole, 1990). The resulting LMS quantities provide information to draw any centile curve, and was expressed by the following formula

$$C_{100\alpha}(t) = M(t)[1 + L(t)S(t)Z_{\alpha}]^{\frac{1}{L(t)}} \tag{1}$$

The LOWESS method, which uses locally weighted linear regression for smoothing was opted for curves smoothing. At each point, a local polynomial was fitted to a local region of the data using linear least squares regression. The MINITAB software is used to perform this calculation. We adopt most application and choose the degree of the local polynomial approximately to be 2 while the degree of smoothing is around 0.5 - 0.6.

3. RESULTS

The sample size for each age group and their mean and standard deviations for all anthropometric measurements are summarized in Table 1. Female school children in the study outnumbered male participants.

Table 1: The mean and standard deviation for body height (cm) and weight (kg) of school children and adolescents aged 6 to 18 years

Gender	Age (years)	No.	Mean ± SD	
			Height	Weight
Male	6.0	49	113.98 ± 7.37	20.33 ± 4.82
	6.5	334	118.33 ± 8.12	21.75 ± 4.75
	7.0	421	119.55 ± 7.19	22.35 ± 4.71
	7.5	475	123.08 ± 8.32	24.55 ± 5.90
	8.0	459	124.26 ± 8.24	24.72 ± 5.44
	8.5	472	129.70 ± 8.63	29.06 ± 6.83
	9.0	482	130.04 ± 9.48	29.04 ± 6.64
	9.5	615	134.71 ± 9.90	31.63 ± 8.08
	10.0	653	136.26 ± 11.50	32.20 ± 7.41
	10.5	644	141.07 ± 10.81	35.14 ± 8.54
	11.0	663	141.83 ± 10.15	36.00 ± 8.76
	11.5	640	145.43 ± 11.07	39.66 ± 10.31
	12.0	654	147.09 ± 11.62	41.59 ± 10.42
	12.5	631	150.93 ± 11.76	44.11 ± 11.70
	13.0	571	153.27 ± 12.02	44.69 ± 11.36
	13.5	543	156.58 ± 13.33	47.58 ± 12.68
	14.0	538	159.52 ± 12.03	49.36 ± 12.65
	14.5	569	161.53 ± 10.82	53.25 ± 14.03
	15.0	609	163.09 ± 10.55	53.05 ± 12.20
	15.5	587	166.25 ± 9.60	56.81 ± 13.24
	16.0	592	166.74 ± 9.50	57.12 ± 13.03
	16.5	647	168.88 ± 9.11	58.99 ± 12.98
	17.0	584	169.04 ± 8.76	59.96 ± 12.98
	17.5	88	169.42 ± 7.64	64.33 ± 15.16
	18.0	62	170.24 ± 6.88	60.40 ± 11.41

Gender	Age (years)	No.	Mean ± SD	
			Height	Weight
Female	6.0	79	113.40 ± 7.35	18.67 ± 3.67
	6.5	354	118.19 ± 7.11	21.28 ± 4.28
	7.0	434	119.51 ± 6.85	21.47 ± 4.07
	7.5	494	123.30 ± 8.11	24.47 ± 5.34
	8.0	529	123.14 ± 7.95	24.09 ± 4.83
	8.5	543	129.54 ± 9.01	27.82 ± 6.06
	9.0	584	129.65 ± 9.12	28.27 ± 6.22
	9.5	710	134.68 ± 9.71	30.50 ± 7.45
	10.0	715	136.14 ± 10.36	31.42 ± 7.40
	10.5	728	140.34 ± 10.92	33.39 ± 7.37
	11.0	810	142.56 ± 11.13	34.73 ± 7.62
	11.5	801	146.55 ± 10.27	38.36 ± 8.98
	12.0	803	148.26± 9.27	40.22 ± 9.57
	12.5	655	151.61 ± 10.09	43.24 ± 9.60
	13.0	641	151.87 ± 10.43	44.47 ± 10.46
	13.5	656	153.44 ± 9.24	44.51 ± 9.68
	14.0	613	154.51 ± 8.57	45.09 ± 9.13
	14.5	749	156.23 ± 7.51	48.07 ± 9.42
	15.0	725	155.93 ± 7.89	48.54 ± 9.32
	15.5	716	157.58 ± 7.14	49.53 ± 8.64
	16.0	618	157.54 ± 6.86	50.28 ± 9.85
	16.5	798	157.42 ± 6.66	50.62 ± 10.17
	17.0	699	157.74 ± 6.60	50.01 ± 9.14
	17.5	70	157.25 ± 6.82	51.53 ± 10.06
	18.0	75	156.84 ± 5.62	49.22 ± 8.47

Conversely, Table 2 and 3 displayed the calculated LMS and percentiles values for height and weight respectively according to Cole's method. Percentiles for height and weight (3rd, 10th, 25th, 50th, 75th, 90th and 97th) were obtained. The percentile curves were shown in Figure 1 and 2.

Table 2: The LMS values and height percentiles for both male and female school children

Gender	Age (years)	L	M	S	Percentiles						
					P3	P10	P25	P50	P75	P90	P97
Male	6.0	2.71	114.37	0.06	98.76	104.17	109.21	114.37	119.16	123.22	127.00
	6.5	1.95	118.59	0.07	102.23	107.70	112.99	118.59	123.96	128.61	133.04
	7.0	3.04	119.98	0.06	104.68	110.03	114.97	119.98	124.60	128.49	132.09
	7.5	0.32	122.89	0.07	107.94	112.55	117.38	122.89	128.58	133.86	139.21
	8.0	0.08	124.01	0.07	109.42	113.88	118.59	124.01	129.66	134.95	140.36
	8.5	-0.71	129.22	0.07	114.76	119.05	123.70	129.22	135.18	140.97	147.10
	9.0	-0.57	129.51	0.07	113.66	118.36	123.45	129.51	136.04	142.39	149.12
	9.5	-1.40	133.89	0.07	118.46	122.89	127.83	133.89	140.68	147.58	155.22
	10.0	0.06	135.81	0.08	115.88	121.90	128.32	135.81	143.71	151.21	158.96
	10.5	-0.14	140.60	0.08	121.99	127.59	133.58	140.60	148.04	155.14	162.52
	11.0	0.30	141.57	0.07	123.40	129.00	134.86	141.57	148.51	154.98	161.54
	11.5	1.43	145.62	0.08	124.02	131.06	138.05	145.62	153.02	159.56	165.90
	12.0	1.87	147.49	0.08	123.99	131.87	139.47	147.49	155.15	161.79	168.11
	12.5	1.49	151.16	0.08	128.08	135.63	143.10	151.16	159.01	165.94	172.62
	13.0	2.77	154.08	0.08	128.08	137.32	145.69	154.08	161.73	168.12	174.02
	13.5	2.64	157.49	0.08	128.51	138.83	148.16	157.49	165.98	173.08	179.64
	14.0	3.01	160.38	0.07	134.35	143.69	152.06	160.38	167.91	174.17	179.92
	14.5	2.94	162.22	0.07	138.96	147.16	154.66	162.22	169.16	174.97	180.35
	15.0	3.48	163.87	0.06	141.27	149.41	156.68	163.87	170.35	175.72	180.63
	15.5	3.04	166.80	0.06	146.68	153.68	160.17	166.80	172.93	178.10	182.91
	16.0	4.06	167.52	0.06	146.46	154.16	160.92	167.52	173.41	178.25	182.64
	16.5	2.01	169.13	0.05	151.05	157.03	162.88	169.13	175.15	180.41	185.44
	17.0	1.56	169.16	0.05	152.18	157.70	163.19	169.16	175.02	180.21	185.24
	17.5	3.87	169.89	0.04	153.96	159.54	164.68	169.89	174.67	178.69	182.41
	18.0	6.56	170.96	0.04	154.65	160.89	166.09	170.96	175.16	178.53	181.53
Female	6.0	3.35	113.94	0.06	97.72	103.55	108.77	113.94	118.61	122.48	126.04
	6.5	1.11	118.22	0.06	104.72	109.04	113.40	118.22	123.01	127.32	131.55
	7.0	0.98	119.51	0.06	106.63	110.72	114.89	119.51	124.13	128.30	132.42
	7.5	0.61	123.19	0.07	108.31	112.96	117.77	123.19	128.71	133.76	138.81
	8.0	1.24	123.21	0.06	107.98	112.88	117.81	123.21	128.55	133.32	137.99
	8.5	0.64	129.42	0.07	112.87	118.05	123.40	129.42	135.55	141.17	146.78
	9.0	1.32	129.75	0.07	112.16	117.85	123.54	129.75	135.86	141.30	146.60
	9.5	0.23	134.41	0.07	117.13	122.44	128.01	134.41	141.06	147.27	153.59
	10.0	-0.08	135.72	0.08	117.80	123.21	128.98	135.72	142.83	149.60	156.60
	10.5	0.33	140.06	0.08	120.55	126.55	132.84	140.06	147.53	154.50	161.57
	11.0	0.73	142.45	0.08	121.89	128.34	134.98	142.45	150.02	156.93	163.83
	11.5	2.59	147.11	0.07	125.63	133.07	140.01	147.11	153.70	159.28	164.49
	12.0	1.41	148.38	0.06	130.45	136.26	142.06	148.38	154.58	160.09	165.44
	12.5	3.04	152.26	0.06	130.86	138.43	145.32	152.26	158.61	163.91	168.82
	13.0	3.37	152.66	0.07	130.17	138.29	145.53	152.66	159.08	164.39	169.25
	13.5	3.26	154.04	0.06	134.34	141.30	147.65	154.04	159.88	164.76	169.27
	14.0	2.25	154.81	0.06	137.54	143.32	148.91	154.81	160.44	165.31	169.94
	14.5	1.83	156.38	0.05	141.68	146.49	151.25	156.38	161.38	165.78	170.02
	15.0	2.40	156.21	0.05	140.30	145.63	150.77	156.21	161.39	165.87	170.12
	15.5	3.23	157.93	0.04	143.13	148.21	152.99	157.93	162.55	166.47	170.15
	16.0	2.75	157.80	0.04	143.88	148.56	153.07	157.80	162.29	166.17	169.84
	16.5	1.21	157.45	0.04	144.79	148.85	152.94	157.45	161.93	165.95	169.89
	17.0	1.34	157.79	0.04	145.18	149.23	153.31	157.79	162.22	166.18	170.06
	17.5	-3.01	156.69	0.04	145.95	149.05	152.48	156.69	161.40	166.19	171.51
	18.0	-1.79	156.56	0.04	146.95	149.83	152.92	156.56	160.46	164.21	168.16

Table 3: The LMS values and weight percentiles for both male and female school children

Gender	Age (years)	L	M	S	Percentiles						
					P3	P10	P25	P50	P75	P90	P97
Male	6.0	-0.36	19.60	0.23	13.07	14.78	16.84	19.60	23.03	26.86	31.53
	6.5	-1.02	20.85	0.20	15.17	16.61	18.38	20.85	24.09	28.04	33.46
	7.0	-0.45	21.70	0.20	15.33	17.02	19.03	21.70	24.93	28.49	32.77
	7.5	-1.06	23.35	0.21	16.72	18.37	20.43	23.35	27.29	32.23	39.34
	8.0	-0.80	23.76	0.21	16.95	18.70	20.83	23.76	27.52	31.96	37.76
	8.5	-0.56	27.93	0.22	19.22	21.47	24.21	27.93	32.65	38.06	44.90
	9.0	0.51	28.68	0.22	17.88	21.05	24.52	28.68	33.17	37.48	41.99
	9.5	-0.62	30.14	0.24	20.24	22.74	25.84	30.14	35.74	42.40	51.18
	10.0	-0.83	30.83	0.21	21.84	24.12	26.93	30.83	35.91	41.99	50.12
	10.5	-0.21	33.94	0.24	22.10	25.23	28.98	33.94	39.98	46.56	54.37
	11.0	-0.57	34.52	0.23	23.53	26.35	29.80	34.52	40.52	47.47	56.34
	11.5	-0.46	37.87	0.25	24.83	28.16	32.24	37.87	45.06	53.41	64.05
	12.0	-0.51	39.77	0.24	26.51	29.90	34.05	39.77	47.08	55.58	66.45
	12.5	-0.78	41.79	0.24	28.43	31.76	35.91	41.79	49.64	59.31	72.70
	13.0	-0.24	43.05	0.24	27.91	31.88	36.66	43.05	50.88	59.51	69.87
	13.5	-0.30	45.56	0.26	29.02	33.29	38.49	45.56	54.41	64.39	76.68
	14.0	-0.43	47.30	0.24	31.30	35.42	40.44	47.30	55.95	65.85	78.25
	14.5	-0.76	50.46	0.24	34.18	38.23	43.30	50.46	60.01	71.75	87.92
	15.0	-0.27	51.43	0.22	34.85	39.27	44.52	51.43	59.76	68.80	79.47
	15.5	-0.52	54.74	0.22	37.77	42.19	47.53	54.74	63.75	73.96	86.65
	16.0	-0.95	54.67	0.21	39.34	43.22	48.00	54.67	63.43	74.05	88.53
	16.5	-0.58	56.99	0.21	40.23	44.61	49.89	56.99	65.83	75.83	88.26
	17.0	-1.07	57.48	0.20	42.16	46.04	50.81	57.48	66.26	76.98	91.79
	17.5	-0.92	61.46	0.21	43.73	48.20	53.72	61.46	71.66	84.09	101.14
	18.0	-0.34	59.06	0.18	42.64	47.11	52.33	59.06	67.00	75.45	85.23
Female	6.0	-0.11	18.28	0.19	12.77	14.30	16.05	18.28	20.87	23.55	26.57
	6.5	-1.32	20.46	0.18	15.50	16.76	18.30	20.46	23.31	26.83	31.80
	7.0	-1.01	20.80	0.17	15.68	17.01	18.62	20.80	23.56	26.77	30.93
	7.5	-0.67	23.58	0.21	16.65	18.44	20.61	23.58	27.32	31.64	37.14
	8.0	-1.44	23.08	0.18	17.49	18.89	20.63	23.08	26.41	30.64	36.90
	8.5	-0.16	27.07	0.22	18.28	20.66	23.45	27.07	31.36	35.93	41.19
	9.0	-0.41	27.36	0.21	18.86	21.10	23.79	27.36	31.73	36.58	42.44
	9.5	-0.41	29.32	0.24	19.52	22.06	25.14	29.32	34.54	40.44	47.74
	10.0	-0.50	30.23	0.22	20.63	23.12	26.14	30.23	35.36	41.20	48.48
	10.5	-0.44	32.30	0.21	22.35	24.97	28.11	32.30	37.45	43.19	50.17
	11.0	-0.23	33.76	0.21	22.94	25.85	29.28	33.76	39.11	44.85	51.54
	11.5	-0.34	37.06	0.23	24.94	28.14	31.96	37.06	43.31	50.22	58.55
	12.0	-0.54	38.64	0.23	26.38	29.54	33.39	38.64	45.26	52.87	62.46
	12.5	-0.37	41.90	0.21	28.91	32.36	36.47	41.90	48.52	55.79	64.50
	13.0	-0.90	42.46	0.22	30.07	33.19	37.05	42.46	49.60	58.29	70.20
	13.5	-0.48	43.09	0.21	30.18	33.59	37.66	43.09	49.77	57.19	66.22
	14.0	-0.49	43.83	0.19	31.39	34.70	38.64	43.83	50.13	57.05	65.36
	14.5	-1.15	46.44	0.17	35.27	38.15	41.65	46.44	52.60	59.91	69.65
	15.0	-1.00	47.00	0.17	35.39	38.41	42.06	47.00	53.26	60.53	69.92
	15.5	-1.27	48.05	0.16	37.31	40.10	43.47	48.05	53.90	60.79	69.91
	16.0	-1.45	48.31	0.17	36.94	39.80	43.34	48.31	54.95	63.27	75.34
	16.5	-1.69	48.50	0.17	37.64	40.36	43.72	48.50	55.01	63.43	76.30
	17.0	-1.72	48.25	0.15	38.14	40.70	43.85	48.25	54.09	61.40	72.02
	17.5	-1.43	49.63	0.17	38.20	41.10	44.66	49.63	56.20	64.32	75.83
	18.0	-2.00	47.58	0.14	38.41	40.74	43.59	47.58	52.92	59.68	69.75

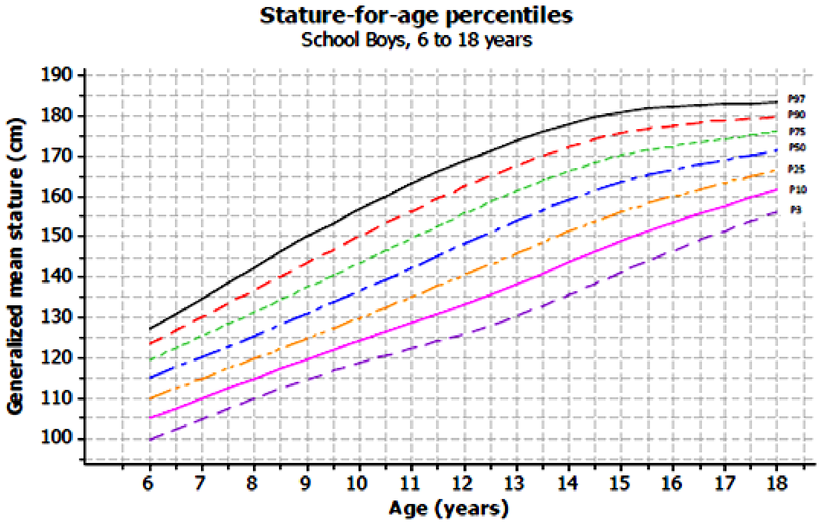


Fig.1 (a): Height-for-age percentiles for male school children aged 6 to 18 years

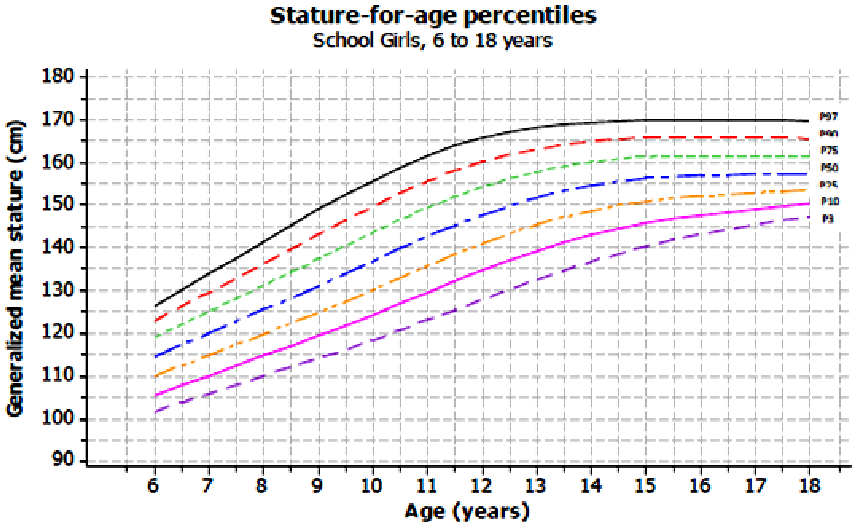


Fig. 1(b): Height-for-age percentiles for female school children aged 6 to 18 years

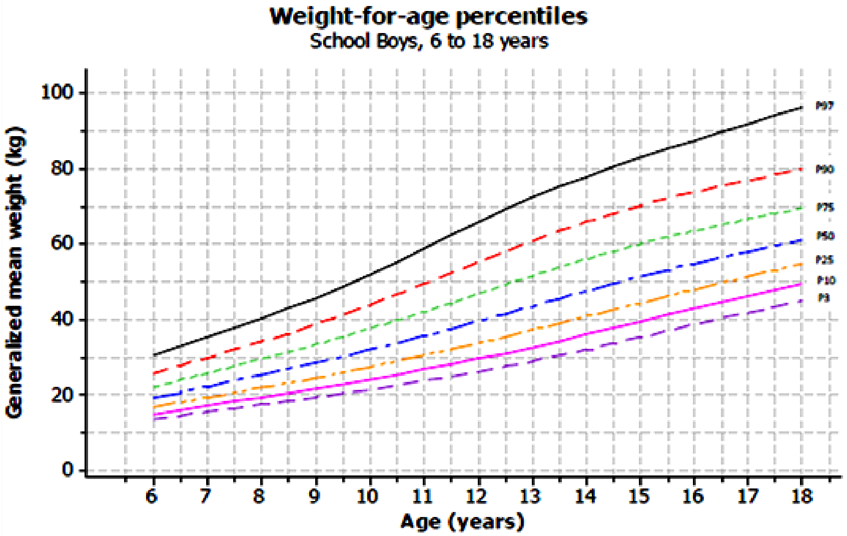


Fig. 2(a): Weight-for-age percentiles for male school children aged 6 to 18 years

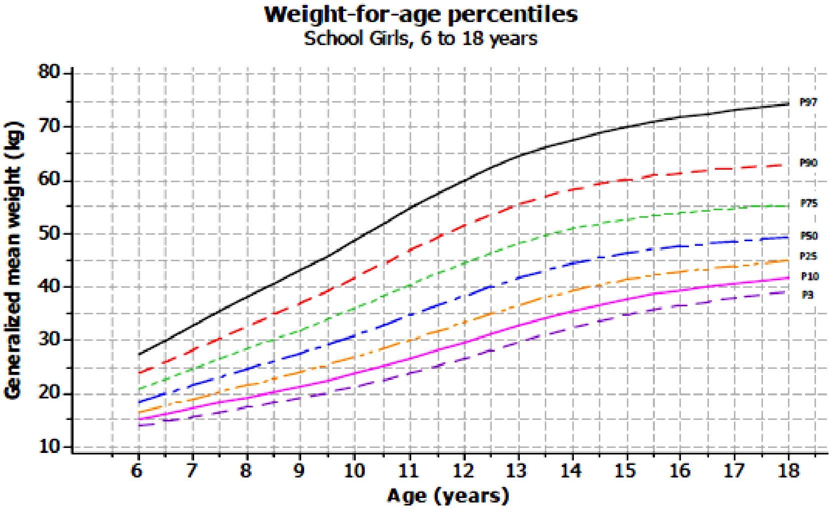


Fig. 2(b): Weight-for-age percentiles for female school children aged 6 to 18 years

The height of both gender were nearly similar at the beginning of school-going ages. Nonetheless, the height differed considerably as school children entered adolescence (13 years old and above). Male school children were taller than their female counterparts during the youth phase. The 50th percentile for height upsurge with the age of male. However, height values tended to stabilize within the range 156cm and 157cm after the age of 15 years old for female.

Male school children weighed heavier than female school children for all age groups except at age 7.5 years old and 12.5 years old. The growth in weight for male and female school children followed a similar pattern to that of height. The curve for female school children weight reached plateau at age 17.0 years old. Centile plots suggested that the weight distribution tended to be skewed. The L values (skewedness) in both gender indicated a negative skew, which verified the earlier suggestion.

4. DISCUSSIONS

In this study, the heights and weights for male and female school children at the start of school-going ages were almost similar. However, the differences between them became considerably bigger as they grew older. Male school children tend to be taller and heavier than female school children. This general tendency was recorded in other similar studies in Turkey and Qatar (Özgüven et al., 2010; Bener & Kamal, 2005).

Female school children reach puberty earlier than the male school children ~~did.~~ This rapid height increase suggests the earlier onset of pubertal growth spurts in girls than boys (Rosario et al., 2011; Silventoinen et al., 2008). However, the growth for female school children slows down after the age of adolescence (11-12 years old). The height percentiles of both gender differed noticeably especially after they entered the adolescence phase.

Growth in weight for the male and female school children followed a pattern similar to that of height. The 97th percentiles for both male and female school children weight curves were much higher, suggesting that school children are getting heavier. The general tendency for male school children to weigh heavier than female school children is similar to the growth patterns found in other countries (Bener & Kamal, 2005).

The median centile curves of CDC 2000 chart showed that the male and female school children in the United States were taller and heavier than those in Malaysia. For height, the CDC 2000 median centile curves are superior to those obtained for males and females from Malaysia as it is known that Asians are generally shorter than Americans of the same age groups. Despite this fact, the growth patterns of West Malaysian school children have improved due to the improvement in nutrition, healthcare and other factors. For weight, the same pattern of growth was observed. Malaysian males and females school children were lighter than their Western counterparts. The 97th percentile for male school children in this study showed that they were significantly heavier than those presented in the CDC 2000 curves from 6 to 14.5 years old. The present increase in consumption of fast food among school children may contribute to the unreasonable weight gain.

The findings from this study indicate that the growth patterns among Malaysian school children have improved as a result of better nutrition and health care, although their heights and weights, on average, were still lower than those of school children from the Western countries as indicated by the CDC 2000 growth charts.

5. CONCLUSIONS

The growth in weight for male and female followed a similar pattern to that of height, but they tended to be skewed. The median centile curves of CDC showed that the male and female school children in the US were taller and heavier than those in Malaysia. There are some differences between the distributions of height and weight of Malaysian school children and adolescents from the CDC growth patterns. Such differences should be taken into considerations when applying growth references. Thus, it might be necessary to define new national representative growth references for local evaluation purpose. The proposed reference growth charts may well be acceptable for assessment of local children because the height and weight centiles show small discrepancies in comparison with the international references currently used.

REFERENCES

1. Bener, A. and Kamal, A.A. (2005). Growth patterns of Qatari school children and adolescents aged 6-18 years. *Journal of Health Population Nutrition*, 23(3), 250-258.
2. Chen, S.T. and Dugdale, A.E. (1970). Weight and height curves for Malaysian schoolchildren. *Med. J. Malaya*, 25, 99-101.
3. Cole, T.J. (1990). The LMS method for constructing normalized growth standards. *Eur. J. Clin. Nutr.*, 44(1), 45-60.
4. Dugdale, A.E. (1969). The weights of Malaysian infants up to one year. *Med. J. Malaya*, 23(4), 244-246.
5. Dugdale, A.E., MacKay, D.A., Lim, R.K. and Notaney, K.H. (1972). Growth charts based on measurements of Malay pre-schoolchildren. *Med. J. Malaya*, 27(2), 85-88.
6. Garner, P., Panpanich, R. and Logan, S. (2000). Is routine growth monitoring effective? A systematic review of trials. *Archives of Disease in Childhood*, 82, 197-201.
7. Özgüven, I., Ersoy, B., Özgüven, A.A. and Erbay, P.D. (2010). Evaluation of nutritional status in Turkish adolescents as related to gender and socioeconomic status. *Journal of Clinical Research in Pediatric Endocrinology*, 2(3), 111-116.
8. Panpanich, R. and Garner, P. (2009). *Growth monitoring in children (Review)* (Publication no. 10.1002/14651858.CD001443) from John Wiley & Sons, Ltd.
9. Rosario, A.S., Schienkiewitz, A. and Neuhauser, H. (2011). German height references for children aged 0 to under 18 years compared to WHO and CDC growth charts. *Ann Hu Biol.*, 38(2), 121-130.
10. Silventoinen, K., Haukka, J., Dunkel, L., Tynelius, P. and Rasmussen, F. (2008). Genetics of pubertal timing and its associations with relative weight in childhood and adult height: The Swedish Young Male Twins Study. *Pediatrics*, 121, 885-891.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 193-198

FORECASTING AGE-SPECIFIC FERTILITY RATES OF PAKISTAN USING FUNCTIONAL TIME SERIES MODELS

Farah Yasmeen and **Zahid Mahmood**
Department of Statistics, University of Karachi, Karachi, Pakistan
Email: riazfarah@yahoo.com

ABSTRACT

The purpose of this study is to model and forecast the trend and patterns of fertility rates in Pakistan. We use functional time series (FTS) models to obtain the forecast for the next twenty years. The secondary data of age-specific fertility rates from 1984 to 2005, obtained from Pakistan Demographic Surveys (PDS) are used. These data are available for 1984-1986, 1988-1992, 1995-1997, 1999-2001, 2003 and 2005. Age-specific fertility rates are missing for rest of the years, so we estimate them using interpolation splines. Before applying FTS models, the data are smoothed using non-parametric smoothing methods. We use weighted regression B-splines with a concavity constraint. Finally we fit a model with four basis functions, which shows the major modes of variations. We obtain the forecasts for the next twenty years (2006-2025).

KEY WORDS

Age-specific fertility rates, fertility pattern, functional time series, forecast, forecast error.

1. INTRODUCTION

In many fields of statistical research, data arise in the form of curves or surfaces rather than collection of simple points. Such data are termed "functional data". Functional data analysis (FDA) has attained substantial development in recent years. Functional time series (FTS) encompasses data in the form of curves that are observed at regular intervals in time. The applications of functional time series include Hyndman and Ullah 2007, Erbas etal 2007 and Yasmeen et al 2010.

In this paper, we will apply the functional time series (FTS) models to the age-specific fertility rates of Pakistan. A variety of mathematical models have been proposed to describe the reproductivity and fertility pattern. They include Pollard et al. (1990), and Peristera and Kostaki (2007). The other useful references are Sathar and Casterline (1998) and Sathar and Kazi (1990).

However to study the fertility, the existing literature is lacking the modelling approach. Also, neither of these studies has considered the forecasting of fertility curve for future. Current piece of work mainly concentrates on the modelling and forecasting age-specific fertility rates for Pakistan. The main objectives of this paper are:

- To study the behaviour and patterns of the age specific fertility rates (ASFR) for available years;

- To estimate the ASFR for missing years using interpolation splines
- To observe and to model the trends of ASFR using FTS models
- To obtain the forecasts of fertility curve for the next twenty years.

This paper is divided into seven sections. Section 1 is introductory, with some basic concepts of describing functional time series (FTS) are explained in section 2. In section 3, fertility patterns for Pakistan and estimation of fertility data for missing years are given, whereas section 4 comprises the application of FTS models to the age-specific fertility rates. The results of the statistical analysis are discussed in section 5, and finally some concluding remarks are given in section 6.

2. FORECASTING USING FTS MODELS

Hyndman and Ullah 2007 first proposed a functional time series model and presented a method for demographic forecasting. Ramsay and Silverman 2005 described the functional data paradigm more generally. The Hyndman-Ullah technique uses several principal components obtained by applying singular value decomposition (SVD) to the matrix of log mortality/fertility rates. The main model is

$$y_t(x)=f_t(x)+\sigma_t(x)\varepsilon_t, \tag{1}$$

where $y_t(x)$ denotes the observed log fertility rate at age 'x' in year 't'. We assume that is a smooth function of 'x', allows the amount of noise to vary with 'x', and ε_t, are considered to be independent and identically distributed random variables with zero mean and unit variance. The first step is to estimate these smooth functions from the discrete noisy data. This is possible using nonparametric smoothing methods. Hence the smoothed curves are decomposed using

$$f_t(x)=\mu(x)+\sum_{j=1}^{J}\beta_{j,t}\phi_j(x)+e_t(x) \tag{2}$$

where $\mu(x)$ is the mean log fertility rate across years and $\{\phi_j(x)\}$ is a set of orthogonal basis functions. The values $\{\beta j_{,1},\ldots,\beta j_{,}n\}$ form a univariate time series for $j=1,\ldots,J$. The basis functions $\{\phi_j(x)\}$ are computed using functional principal components applied to the smooth curves $ft(x)$.

The h-step ahead forecast of $\hat{y}_{,n+h,}(x)$ can be obtained as

$$\hat{y}_{n+h}(x)=\hat{\mu}(x)+\sum_{j=1}^{J}\hat{\beta}_{j,n+h}\hat{\phi}_j(x) \tag{3}$$

where $\hat{\mu}(x)$ and $\hat{\phi}_j(x)$ are the estimates of the mean function and basis functions, respectively and denotes the h-step ahead forecast of $\beta_{j,n+h}$. Hyndman and Ullah 2007 showed that the forecast variance can be obtained by adding the variances of all individual terms.

3. THE PATTERN OF FERTILITY IN PAKISTAN

Pakistan is the world's sixth most-populous country, behind Brazil and ahead of Bangladesh. In 2011, the estimated population of Pakistan is over 187 million (CIA: The world fact book 2011). Pakistan's total population increased by over fourfold during 1950–2011. Estimates from different sources imply decline in fertility in Pakistan particularly after 1990s (Nasir et al 2009).

3.1 Available Data and Estimation of Missing Values

The secondary data of age-specific fertility rates (ASFR) from 1984 to 2005, obtained from Pakistan Demographic Surveys are used. These data are available for 1984-1986, 1988-1992, 1995-1997, 1999-2001, 2003 and 2005. Age-specific fertility rates are missing for rest of the years, so we estimate them using interpolating splines.

For each age-group, we consider a fixed number of knots as available ASFR and estimate the function for missing years (i.e. 1987, 1993, 1994, 1998, 2002 and 2004). The splines are basically the functions of polynomials. An R function is available for interpolation in R package '*stats*'. It performs cubic (Hermite) spline interpolation of given data points. The available and estimated age-specific fertility rates (ASFR) using spline function are given in table 1.

Table 1: Age Specific Fertility Rates (ASFR) of Pakistan (per 1000 woman)

Age group	15-19	20-24	25-29	30-34	35-39	40-44	45-49
1984	65.76	268.33	367.57	314.42	226.07	109.56	37.88
1985	59.15	272.98	350.79	326.98	235.29	108.57	47.88
1986	54.31	265.75	360.26	303.12	226.22	125.98	52.16
1987*	57.06	263.27	350.85	286.31	212.27	124.59	46.92
1988	66.00	263.60	333.00	278.30	203.30	111.20	41.80
1989	75.70	265.80	323.40	274.30	197.10	102.00	41.60
1990	75.50	274.80	313.20	276.00	175.90	97.00	30.50
1991	69.00	258.20	315.4	259.00	186.50	82.30	27.40
1992	73.30	261.40	312.90	254.50	162.60	74.50	27.80
1993*	72.95	254.4	310.02	244.6	147.4	83.82	30.67
1994*	66.58	241.52	308.03	234.9	147.3	95.93	32.87
1995	59.10	243.40	305.10	241.90	148.10	90.10	29.60
1996	54.7 0	258.2 0	295.90	255.40	143.00	65.50	23.20
1997	52.3 0	231 00	273.20	211.20	142.90	68.40	30.70
1998*	44.13	211.31	263.11	195.37	132.11	69.28	31.11
1999	36.20	205.60	256.90	203.60	118.30	61.70	25.80
2000	32.90	195.10	244.20	203.80	114.50	54.40	22.90
2001	24.20	162.00	242.90	197.20	118.50	57.90	21.90
2002*	21.54	153.06	238	192.73	117.66	55.66	20.45
2003	23.70	163.10	229.60	190.00	112.70	49.00	18.80
2004*	24.7	169.4	224.06	186.33	107.99	45.39	17.8
2005	20.30	157.60	225.50	179.90	106.60	50.10	18.1

Source: Pakistan Demographic Surveys 1984-2005

Available Years: 1984-1986, 1988-1992, 1995-1997, 1999-2001, 2003 and 2005.
*Estimated Years: 1987, 1993, 1994, 1998, 2002, 2004.

4. FITTING FTS MODELS TO AGE SPECIFIC FERTILITY RATES

At first, we plot ASFR for the years 1984-2005 and consider them as functional observation. The ASFR for selected years are plotted in figure 1. We first smooth these rates using non-parametric methods. We use weighted regression B spline to obtain smooth curves. To analyse these curves, we use functional time series models introduced by Hyndman and Ullah 2007. An adequate fit is obtained using a functional regression model with four basis functions.

The time series coefficients are then forecast using univariate time series models. These forecasts are then multiplied with the estimated basis functions resulting in forecasts of the entire fertility curve. This technique also provides computation of prediction intervals (for details, see Yasmeen et al 2010). All statistical analyses were performed in R version 2.15.0.

5. RESULTS

The graph of ASFR of Pakistan (figure 1) shows that during the study period, the fertility rates have been decreased for all age-groups. A greater decline occurred in the middle age group i.e. 20-24, 25-29 and 30-34, and relatively slower decline in the other age groups (15-19, 35-39, 40-44 and 45-49). Another import point is that the pattern is almost the same for the study period (1984-2005) and we did not find any significant change in this pattern.

The first step in using functional time series models is to obtain smoothed log fertility rates. We use regression B splines for smoothing.

For fertility data of Pakistan, the percentage variation due to basis functions are 89.9% 7.3% 1.9% 0.4%,. Hence only four basis functions are sufficient to explain about 99% of the total variability in rates of Pakistan women.

Figure 2 depicts the plots of the mean function, the first four basis functions and corresponding coefficients for the FTS models. The major changes in fertility rates amongst Pakistani women occur at the ages 23 and 35 years, while their respective coefficients show a decline in future years. We do not attempt to interpret the other basis functions as they involve second-and higher-order effects.

By multiplying the forecasts of the coefficients with the basis functions and summing the results, we obtain forecasts of the entire fertility curves. Figure 3 shows the forecast for the next twenty years (2006-2025).

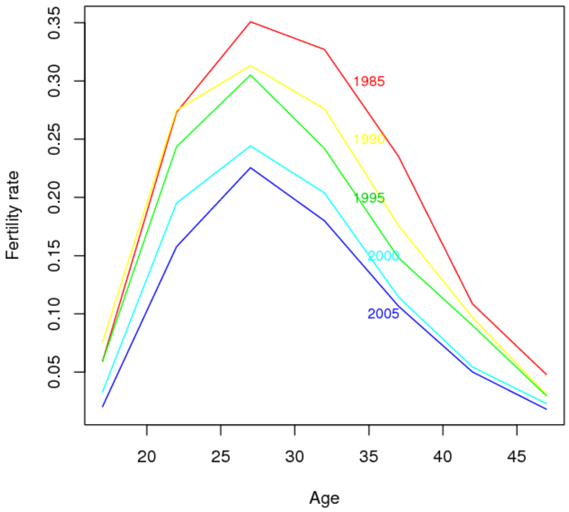


Figure 1: Fertility rates of Pakistan

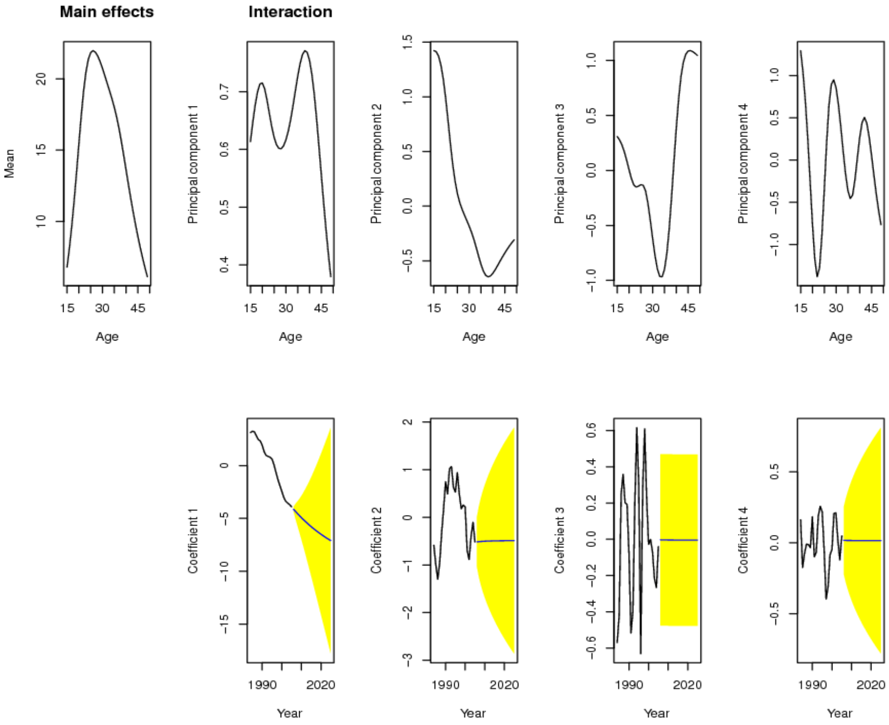


Figure 2: Forecasts of First Four Coefficients of FTS model

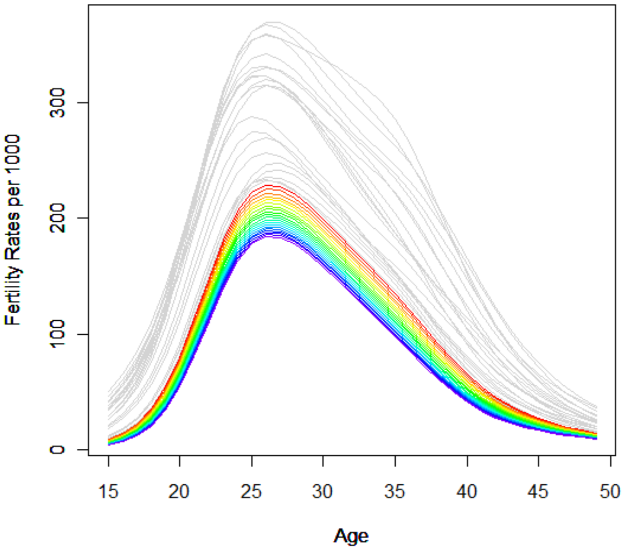


Figure 3: Twenty-year forecasts for Pakistan (2006-2005)

6. CONCLUSION

In this paper, we applied the functional time series (FTS) models to the age-specific fertility rates of Pakistan. Fertility is the second principle vital event in demography. Fertility gives researchers a glimpse into the future of society but our main focus/research is on the indicator ASFR. Forecasting ASFR gives better picture for family planning schemes. It's also has socio economic importance in society.

By plotting ASFR of Pakistan, we found that their distribution has a typical shape. For the fertility data of Pakistan, this non-linear pattern of fertility curve can be considered as the reciprocal of the V-shapes .In literature, we found that some polynomial models are fitted to the ASFR. However, neither of the study forecast the future age-specific fertility rates of Pakistan. To our knowledge, this is the first study which shows the fertility-age projections.

Estimates from different sources imply decline in fertility in Pakistan particularly after 1990s. Our results also confirm the results of the previous research. In addition, we obtained 20-years predictions for the ASFR. These plots exhibit that the future fertility rates will decline for all ages, but that the decline will be greatest for the middle age women, and relatively slower for the youngest and oldest women.

REFERENCES

1. CIA: The world fact book: 2011, Population Growth Rate
2. Erbas, B., Hyndman, R.J. and Gertig, D.M. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine,* 26, 458-470.
3. Government of Pakistan (2005). *Population Demographic Survey- 2005.* Federal Bureau of Statistics, Statistics Division, Islamabad. Pakistan.
4. Hyndman, R.J. and Ullah, M.S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics and Data Analysis,* 51(10), 4942-4956.
5. Nasir J.A., Akhtar, M. and Tahir, M.H. (2009). Rreproductivity and age-specific fertility rates in Pakistan after 1981. *Pakistan Journal of Statistics*, 25(3), 251-263.
6. Peristera, P. and Kostaki, A. (2007). Modeling fertility in modern populations. *Demographic Resear*ch, 16, 141-194.
7. Pollard, A.H., Yusuf, F. and Pollard, A.H. (1990). *Demographic Techniques*. Pergamon Press, New York, Inc.
8. Sathar, Z.A. and Casterline, J.B. (1998). The onset of fertility transition in Pakistan. *Population and Development Review*, 24, 773-796.
9. Sathar, Z.A. and Kazi, S. (1990). Women, work and reproduction in Karachi. *Inter. Family Plan. Perspectives*, 16, 66-80.
10. Yasmeen, F. (2011). Functional linear models for mortality forecasting. Unpublished PhD thesis, Monash University, Australia.
11. Yasmeen, F., Hyndman, R.J. and Erbas, B. (2010). Forecasting age-related changes in breast cancer mortality among white and black US women: A functional data approach. *Cancer Epidemiology,* 34(5), 542-549.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 199-206

DISCOVERING DIVERSITY PROFILES BY PARAMETRIC FUNCTIONAL ANALYSIS

T. Di Battista and **P. Cappola**
Department of quantitative methods and economic theory, University G. d'Annunzio,
Chieti-Pescara, Italy, Viale Pindaro 42, 65127 Pescara, Email: dibattis@unich.it

ABSTRACT

A new approach to evaluate the biodiversity is developed in this Paper. In our contest the biological diversity is seen as a function of the relative abundances of species in a community of animals or plants. In this setting several researchers have suggested using parametric families of indices of diversity for obtaining more information from the data. Patil and Taillie (1982) introduced the concept of intrinsic diversity ordering which can be determined by using the diversity profile. It may be noted that the diversity profile is a non-negative and convex curve which consists of a sequence of measurements as a function of a given parameter. Thus, diversity profiles can be explained through a process that is described in a functional setting. Considering the data as a parametric family of function we obtain statistics that belong of functional diversity profiles. This approach can be useful in order to consider the biodiversity simultaneous without losing any information on the data.

KEYWORDS

Biodiversity profile, Functional data analysis, monotonic dependence $L^p - space$, functional mean, functional variability.

1. INTRODUCTION

In statistics the diversity concept relies on the variety of a phenomenon, which is generally related to the apportionment of some quantity into a number of categories. For example in ecology the researcher is interested to classify a biological population composed by N units, into their spices by means the counting of the number of the organism belonging to each species. The objective is to evaluate the variety of living organisms in a delineated study area.

The environmental changes such as deforestation and pollution have strongly modified the ecosystem in time. Thus, the evaluation of the biodiversity distribution provides a correct interpretation of the pollution effects.

In this setting the evaluation of biodiversity has become a crucial element of environmental monitoring programs (McCann, 2002).

In order to highlight the issue of preserving the diversity of ecological communities. Because of this the statistic has tried introduce several indexes of the diversity (Gove *et al.*, 1994).

However they have the shortcoming that they are scalars *i.e*. only one single value of the diversity, therefore the analysis depends on the measure of diversity that is adopted. We can assert that ecological diversity is a multidimensional concept including both the species richness (the number of different species) and the species evenness (the relative abundance of different species). The use of a single index greatly reduces the complexity of the ecological systems. In fact Single measures of diversity could lead to different ordering of communities in terms of their diversity. In fact, each index of diversity incorporates a particular degree of sensitivity to rare and common species.

In order to give a general solution to this problem, Patil and Taillie, (1982), Tòthmérész, (1995) and Liu et al., (2007), introduced a parametric family of indexes of diversity, named diversity profiles, which consist of a sequence of measurements respect to a parameter. In this way, different aspects of community can be evaluated.

Therefore, the measure of biodiversity becomes a curve unlike an index that is a scalar.

This new suggestion allows us to implement a suitable statistical approach (Gattone, S.A. and Di Battista, T. 2009).

The idea is to use the functional data analysis (FDA) (Ramsay and Silverman, 2005, Ferraty, F. and Vieu, P. 2006, Di Battista, T., Gattone, S.A., and Valentini, P., 2007).

Thus, the functional approach allows us to study the biodiversity by referring to the entire structure of the data.

The standard approach of FDA is essentially to smooth the data on a prefixed domain, say R^m (in one-dimensional space m=1), by means of some technique such as basis functions.

In fact functional data are often observed as a sequence of point data then the function denoted by $y = f(x)$ reduces to a record of discrete observations that can be labelled by the T pairs (x_j, y_j) where $x \in R$ and y_j are the values of the function computed at the points $x_j, j = 1, 2, \dots, T$.

The use of basis functions converts the values y_{i1}, y_{i2}, … , y_{iT} for each unit $i = 1, 2, \dots, n$ to a functional form computable at any desired point $x \in R$ the statistics are simply those evaluated at the functions pointwise across replications.

The aims of FDA are fundamentally the same as those of any area of statistics in a descriptive setting one may want to investigate some essential aspects such as the mean and variability of the functional data; whereas in inference, assuming a suitable functional probability distribution, one may want to obtain test of hypothesis and confidence intervals for one or more characteristics of the population.

However, in our framework, as better explained later, the diversity profile is a function known in a fixed domain, in this cases we say that these functions belong to a parametric family of functional data.

Therefore in this paper we implement (see section 3) a suitable procedure that takes into account this specific issue.

2. ECOLOGICAL DIVERSITY MEASURES

Considering a biological population partitioned into s species than with N_l we denote the number of population units belonging to the lth species $(l = 1,2, \ldots, s)$. Hence $\mathbf{N} = (N_1, N_2, \ldots, N_s)^T$ denotes the abundance vector whereas $\mathbf{p} = (p_1, p_2, \ldots, p_s)^T$ represents the relative abundance vector where $p_l = N_l / \sum_{l=1}^{s} N_l$ represents the proportion of biological units belonging to the lth species such that $0 \leq p_l \leq 1$ $\sum_{l=1}^{s} p_l = 1$.

The requirements of an index of diversity are:

1. it must be greater than or equal to *0* and equal to *0* only when $p_l = 1$ for one species;
2. it takes its maximum when $p_l = \frac{1}{s}$ for $l = 1,2, \ldots, s$;
3. it must be an increasing function of s.

The simplest measure of diversity is species richness, *i.e.* the number s of species in a community. Widely applied in ecological studies is the entropy or Shannon index that was derived within the framework of information theory (Shannon, 1948) defined as

$$I_{sh} = -\sum_{l=1}^{s} p_l \log p_l \tag{1}$$

It is possible to show that the range of I_{sh} is 0 and log p_l.

In literature several measures of diversity have been introduced, one of most famous is the Simpson index

$$I_s = 1 - \sum_{l=1}^{s} {p_l}^2 \tag{2}$$

which range is 0 and $\frac{s-1}{s}$.

Very useful for our purpose is the general expression for the measure of diversity introduced by Patil and Taillie (1982) as the average species rarity, *i.e.*

$$I(\mathbf{p}) = \sum_{l=1}^{s} p_l R(p_l) \tag{3}$$

where $R(p_l)$ is a measure of rarity for species l. Some of the most frequently used indices of diversity are special cases of equation (3).

In the same work, Patil and Taillie (1982) proposed also a general measure of rarity given by

$$R(p_l) = \frac{1-p_l^{\beta}}{\beta} \; for \; \beta \geq -1 \tag{4}$$

substituting the expansion (4) in (3) we obtain the β-diversity profile as follow

$$I(\beta) = \frac{1-\sum_{l=1}^{s} p_l^{\beta+1}}{\beta} \tag{5}$$

We point out that the expression (5) is a function of β; so, the range of the measure of diversity becomes the domain of the function $I(\beta)$.

As it has been built the β-diversity function varies for $-1 \leq \beta \leq 1$.

Following this approach and considering that the measure of diversity has become a curve in a prefixed domain; a suitable statistical method is the functional data analysis (FDA). However the functions are well known. Hence, the standard procedure of FDA must been suitably adopted at our case.

In the next section some theoretical results of our approach are discussed (De Sanctis, A. and Di Battista, T., 2012).

2. PARAMETRIC FUNCTIONAL DATA ANALYSIS.

First of all let us introduce some Mathematical tools:
let X be an arbitrary measure space with a positive measure than we denote with $L^p(\mu)$ $0 < p < \infty$ the set of real or complex measurable functions on X for which is verify the follows conditions

$$\|f\|_p = \left\{\int_X |f|^p \, d\mu\right\}^{1/p} < \infty \qquad (6)$$

We call $\|f\|_p$ $L^p - norm$ of f (Rudin 2006).

In particular when μ is a Lebesgue measure on real space R^m we write $L^p(R^m)$ instead of $L^p(\mu)$.

If μ is the counting measure on a set A, it is customary to denote the corresponding $L^p - space$ by $L^p(A)$ or simply by c.

An element of L^p may be regarded as a sequence $X = \{\varepsilon_A\}$ and

$$\|X\| = \{\textstyle\sum_{i=1}^{\infty} |\varepsilon_i|^p\}^{1/p} \qquad (7)$$

The following results are well known:

***Theorem 1*:**

i. $L^p(\mu)$ is a (real or complex) vector space;
ii. The relation $f \sim g$ if and only if $f\,(x) = g(x)$ for almost all *x* is an equivalence relation in $L^p(\mu)$. The set of equivalence classes (which we continue to denote $L^p(\mu)$) is a metric space with respect to the distance $d(f\,, g) = \|f - g\|_p$;
iii. $L^p(\mu)$ is a complete metric space *i.e*. every Cauchy sequence in $L^p(\mu)$ converge to an element of $L^p(\mu)$

Theorem 2:

If $\{f_n\}$ has limit *f* in $L^p(\mu)$ then $\{f_n\}$ has a subsequence which converges pointwise almost everywhere (*a.e.*) to *f*.

In general we do not have an orthogonality notion in $L^p(\mu)$ because its norm is not induced by an inner product.

The only case where we have an *Hilbert space* (that is a vector space with an inner product whose induced metric space is complete) is $L^2(\mu)$ with inner product

$$(f, g) = \int_X fg \, d\mu \qquad (8)$$

In this paper we do not use the scalar product and then we can consider the *Banach-space* for every L^p space, $p > 0$, as for example $p = 1$.

we assume that all the sets of functions are subsets of some $L^p(\mu)$. In particular a subset of functions S is a subspace if it is itself a vector space that is:

1. whenever $f \in S$ and $g \in S$ we have $f + g \in S$
2. whenever $f \in S$ and α scalar we have _ $\alpha f \in S$.

Inspired by mathematical tools, and assuming a monotonic dependence we use the parameters space in order to transfer the statistics of the parameters to the functional space. In particular we make the following assumptions

1) Let the parameter space Θ be a convex subset of R^k that is $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_n)^t$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{\ ik})$; let α_{ij} be a scalar with $0 < \alpha_{ij} < 1$ and $\sum_i \alpha_{ij} = 1$ for each $j = 1, 2, \dots, k$ then

   $$\sum_i \alpha_{ij}\, \theta_{ij} = \bar{\theta}_j \in \Theta \tag{9}$$

   for each $j = 1, 2, \dots, k$.

   The extension to the general case in which $\bar{\theta}_{\ j} = h(\boldsymbol{\theta}_j)$ is a generic linear function of $\boldsymbol{\theta}_j = \left(\theta_{1j}, \theta_{2j}, \dots \theta_{nj},\right)^t$ is straightforward.

2) We suppose that there is a bi-univocal correspondence between a convex parameter space Θ and the family S, *i.e.* every functional datum $f(\boldsymbol{\theta}, x)$ of S is unequivocally defined by the parameter $\boldsymbol{\theta}$.

Let us suppose the assumptions 1) and 2) to be true, then for a matrix of parameters $\boldsymbol{\theta}$ where $\boldsymbol{\theta}_1 \leq \boldsymbol{\theta}_2 \leq, \dots, \leq \boldsymbol{\theta}_n$ we have that

$$f(\boldsymbol{\theta}_1, x) \leq f(\overline{\boldsymbol{\theta}}, x) \leq f(\boldsymbol{\theta}_n, x) \tag{10}$$

where $\overline{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$ is a linear function of the parameters.

At this point we can give the following general definition:

let the functional data be $f(\boldsymbol{\theta}_1, x), f(\boldsymbol{\theta}_2, x), \dots, f(\boldsymbol{\theta}_n, x)$ univocally defined by the set of parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_k\ )$ then a functional statistic for the set of the functional data can be obtained from a statistic of the parameters, say

$$\boldsymbol{\theta}^{-} = (h(\boldsymbol{\theta}\_1\ ), h(\boldsymbol{\theta}\_2\ ), \dots, h(\boldsymbol{\theta}\_k\ )) \tag{11}$$

The functional statistic will be an element of S which has as parameter the statistic $\overline{\boldsymbol{\theta}}$. This approach is advantageous because it is possible to require the same properties for the functional mean and variability as for the mean and variance of the parameters.

In order to study the functional variability, we first introduce the functional quantity represented by the *r*th-order algebraic deviation between the observed functional data $f(\boldsymbol{\theta}_i, x)$ with $i = 1, 2, \dots, n$ and the functional statistic $f(\overline{\boldsymbol{\theta}}, x)$

$$v_i^r(x) = \left| f(\boldsymbol{\theta}_i, x) - f(\overline{\boldsymbol{\theta}}, x) \right|^r \tag{12}$$

Then the *r*- order functional deviation can be measured pointwise by the *r*th order functional moment

$$V^r(x) = \frac{1}{n}\sum_{i=1}^{n} v_i^r(x) \qquad (13)$$

for $r = 2$ we obtain the functional variability.

It is possible to prove that the function $V^r(x)$ has the following property:

if $f(\boldsymbol{\theta}_i, x) = f(\overline{\boldsymbol{\theta}}, x)$ *a.e.*, for $i = 1, 2, \dots, n$ then $V^r(x) = 0$ *a.e.*

3. THE SIMULATION STUDY

Suppose we want to study a living organisms population in a delineated study area. Let us assume that the biological population under study is fixed and has got a list frame of *N* known units. In the environmental scenery is more frequent the case in which the study area is partitioned into a frame of *N* sub-areas such as plots or lines.

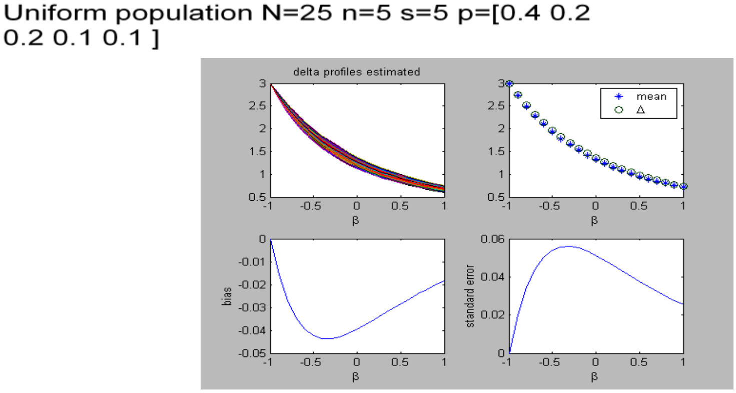
In the FDA context, it is convenient considering the β-diversity profile as a parametric function of β = [-1; 1].

Thus we can write $y_i = I(\beta)$ for *i* = 1,2,…,*N*. Where $I(\beta)$ has been expressed in (5).

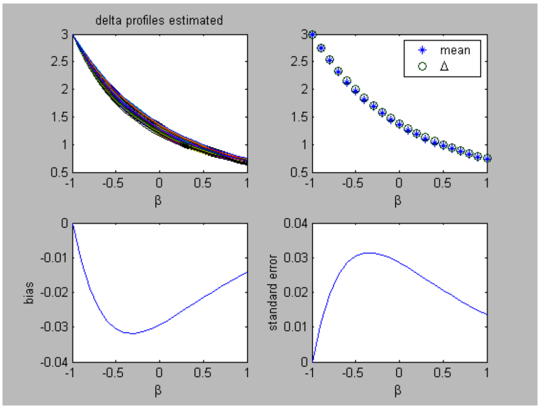
We will consider the diversity datum such a process naturally described as a function varying in the β dominium.

In this paper we simulate different biological populations by assigning to each component of $\boldsymbol{\theta}$ different distributions such as the Uniform, the Poisson and the Multinomial distribution. From each population we sample $J = 5000$ samples with different sample sizes. In this case, the function $y_i = I(\beta)$ is observed without error. For each sample of size *n*, we can evaluate the estimates $\widehat{\boldsymbol{\theta}}$ from the observed $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_n$. The criterion explicated in (11) has been applied in order to obtain the profile of functional diversity average.

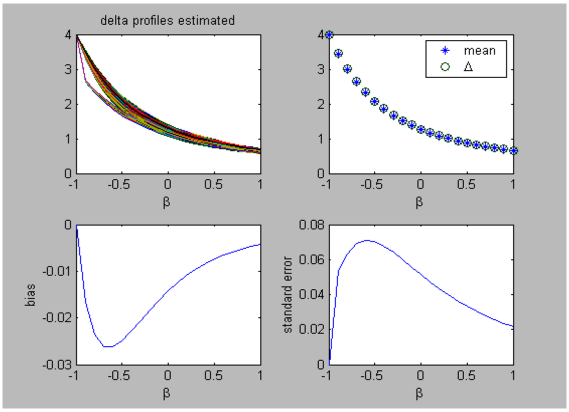
Figure 1: Functional mean diversity profiles $\widehat{I(\beta)} = \frac{1-\sum_{l=1}^{s} p_l^{\beta+1}}{\beta}$ and standard error for a sample size $n = 5$



poisson population N=25 n=5 s=5 p =[0.4 0.2 0.2 0.1 0.1]



multinomial population N=25 n=5 p =[0.4 0.2 0.2 0.1 0.1]



In Figure 1, we show the results for 3 populations with $s = 5$ species and with different level of diversity. From each population we randomly choose samples of size $n = 5$. The functional mean together the bias and the estimated standard error are plotted in three different pictures. As desired, all the functional statistics belong to the family of diversity profile explained in (5). Furthermore, the functional mean satisfies the internality property in all the simulation runs.

In particular, in the first line of each picture, we show the sampling distribution and the comparison between the functional profile estimate with the profile of the population. While in the second line, we explain the behavior of functional standard error and the functional bias. As we showed the simulation gives good results.

REFERENCES

1. De Sanctis, A. and Di Battista, T., (2012). Functional Analysis for parametric families of functional data. *International Journal of Bifurcation and Chaos* (*IJBC*), 22(09).
2. Di Battista, T., Gattone, S.A. and Valentini, P., (2007). Functional data analysis of GSR signal. *In: S. Co.2007*, Venice,
3. Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer-Verlag, New York.
4. Gattone, S.A. and Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society, Series C*, 58, 267-284.
5. Gove, J. H., Patil, G. P., Swindel, D.F. and Taillie, C. (1994). *Ecological diversity and forest management" In Handbook of Statistics*, vol. 12, Environmental Statistics (eds G.P. Patil and C.R. Rao), pp. 409-462. Amsterdam: Elsevier.
6. Liu, C., Whittaker, R.J., Ma, K. and Malcolm, J.R. (2007). Unifying and distinguishing diversity ordering methods for comparing communities. *Popln. Ecol*., 49, 89-100.
7. McCann, K.S. (2002). The diversity-stability debate. *Nature*, 405, 228-233.
8. Patil, G.P. and Taillie, C. (1982). Diversity as a concept and its measurements. *Journal of the American Statistical Association*, 77, 548-561.
9. Ramsay, J.O. and Silverman, B.W. (2005). *Functional data Analysis*. Springer, New York.
10. Rudin, W. (2006). *Real and Complex Analysis*. McGraw-Hill.
11. Tòthmérész, B. (1995). Comparison of different methods for diversity ordering. *J. Vegetn Sci*., 6, 283-290.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 207-216

ON RESISTING OUTLIERS IN BIVARIATE DATA FOR SMALL AND MODERATE SAMPLES

Ezz Hassan Abdelfattah
Statistics Department, King Abdulaziz University, Jeddah, Saudi Arabia
Email: ezzhassan@hotmail.com

ABSTRACT

This paper aims to compare some well known rank correlation measures (Spearman's rho, Kendall's tau) with Spearman's Footrule or simply Footrule (Diaconis & Graham, 1977; Franklin, 1988; Sen et al., 2011), the Greatest deviation (Gideon & Hollister, 1987) and Symmetric Footrule (Abdelfattah 1996 & Salama et al., 2001). The stability of type I error and the stability of power for these measures is introduced through a simulation study applied on some bivariate distributions containing outliers.

KEYWORDS

Rank correlation, Outliers, robustness of correlation Coefficients**.**

1. INTRODUCTION

In this paper, we will introduce a rank correlation denoted by R. This new rank correlation is to be compared with other measures of correlation such as Kendal tau $R_{k,}$ Spearman rho R_s, Spearman footrule R_f and Gideon and Hollister's R_g. The comparison will be in resisting outliers through a simulation study to compare the stability of type I and to compare the stability of the Power for these measures. The simulation study was of size 10,000 and was done for 13 different bivariate distributions for different sample sizes. The study shows a stability of type I error for the rank correlation introduced R and the stability for the power, specially as sample size becomes larger.

Here, is a brief definition for the mentioned measures:

$$R_k = \frac{2\sum_{i<j} \operatorname{sgn}(s_j - s_i)}{n(n-1)} \tag{1.1}$$

$$R_s = 1 - \frac{6\sum_{i=1}^{n}(i - s_i)^2}{n(n^2-1)} \tag{1.2}$$

$$R_g = 2\frac{\max_{1\le i\le n}\sum_{j=1}^{n} h(s_j^* \ge i) - \max_{1\le i\le n}\sum_{j=1}^{n} h(s_j \ge i)}{n(n-1)}, \tag{1.3}$$

*s** is the reverse order of *s*, h is the greatest integer function

$$R_f = 1 - \frac{3\sum_{i=1}^{n} | i - s_i |}{n^2 - 1} \tag{1.4}$$

2. DEFINITION OF R

Let S_n be the set of all n! permutations of the first n integers. Let (x_k,y_k), k=1,...,n be a sample from a continues pdf F(X,Y), $r(x_k)$ be the rank of x_k among the x data and similarly define $r(y_k)$. Let $\boldsymbol{\sigma}_1$, $\boldsymbol{\sigma}_2$ be two elements in S_n representing the rank of $(x_1,..., x_n)$ and $(y_1, ...,y_n)$. If the x data are ordered from the smallest value to the largest one, then we will let σ_k to be the rank of the y datum that corresponds to the kth smallest x value.

Let $I_k = \{1,2,\ldots,k\}$ be the first k numbers and $J_k=\{n\text{-}k+1,n\text{-}k+2,..,n\}$ be the last k numbers of the set {1,...,n}. Let #{A} be the number of elements in the set A. For any permutation $\boldsymbol{\sigma} \in S$ define

$$T_{n,k} = T_{n,k}(\sigma) = \#\{\{\sigma_1,...,\sigma_k\} \cap I_k\}, \tag{2.1}$$

$$B_{n,k} = B_{n,k}(\sigma) = \#\{\{\sigma_1,...,\sigma_k\} \cap J_k\} \tag{2.2}$$

and

$$D_{n,k}(\sigma) = T_{n,k}(\sigma) - B_{n,k}(\sigma)\ . \tag{2.3}$$

Then a statistic $F_S^{(n)} = F_S^{(n)}(\sigma)$ will be defined as follows

$$F_S^{(n)} = \sum_{k=1}^{n} D_{n,k} = \sum_{k=1}^{n} (T_{n,k} - B_{n,k}) \tag{2.4}$$

Let o be the operation representing the composition of two elements of S_n, then the correlation coefficient between $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ is

$$R = R(\sigma_1,\sigma_2) = R(1_n,\sigma_n) = \frac{F_S^{(n)}}{m} \tag{2.5}$$

where $1_n = \{1,2,\ldots,n\}$, $\sigma = \sigma_1^{-1} o \sigma_2$ and m= $\begin{matrix} n^2/4 & \text{if } n \text{ is even} \\ (n^2-1)/4 & \text{if } n \text{ is odd} \end{matrix}$.

As example, for the data given by Kendall & Gibbons (1990):

$r(x_k)$: 1 2 3 4 5 6 7 8 9 10

$r(y_k)$: 8 9 3 7 4 1 5 2 6 10

we get R= -0.12.

Note that for the same data we find Kendall’s tau = -0.07, Spearman’s rho=-0.103, Spearman’s Footrule = -0.3 while Gideon and Hollister’s R_g =0.0.

The mean and variance of R was proved to be:

$$E(R) = 0 \tag{2.6}$$

$$V(R) = \frac{2(n^2+2)}{3n^2(n-1)} \text{ if } n \text{ is even} \tag{2.7}$$

$$= \frac{2(n^2+3)}{3(n^2-1)(n-1)} \text{ if } n \text{ is odd.} \tag{2.8}$$

The derivation of equations (2.1) to (2.8) is introduced in Abdelfattah (1996).

3. PROPERTIES OF R

Since R is defined as a rank correlation between $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ if $\boldsymbol{\sigma}^*$ =(n, n-1,…, 2,1) then $\boldsymbol{\sigma}^*$ o $\boldsymbol{\sigma}_1$ is the reverse of $\boldsymbol{\sigma}_1$ and the following properties hold :

Property 1: $R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ is well defined.

Property 2: $-1 \leq R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) \leq +1$.

Property 3: $R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = R(\boldsymbol{\sigma}_2, \boldsymbol{\sigma}_1)$.

Property 4: $R(\boldsymbol{\sigma}^*, \boldsymbol{\sigma}) = - R(\mathbf{1}_n, \boldsymbol{\sigma})$.

Property 5: $R(\boldsymbol{\sigma}^* \text{ o } \boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2) = R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$

Property 6: The null distribution of $R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ is symmetric about 0.

Property 7: m $R(\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2)$ can assume 2m+1 values:
m,m-1,…,2,1,0,-1,-2,…,-m+1,-m.

The proofs of the properties can be found in Abdelfattah (1996).

4. OUTLIERS

Here, we will discuss the existing of the outliers in both the univariate and multivariate data. A simulation study for discussing the problem of resisting the outliers can be found also in this section.

4.1 Outliers in Univariate Data:

In almost every true series of observations, some are found, which differ so much from the others as to indicate some abnormal source of error not contemplated in the theoretical discussions, and the introduction of which into the investigations can only serve to perplex and mislead the inquirer. The intuitive definition of an outlier would be an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism. We need to distinguish between : extreme observations, outliers and contaminants.

Suppose we have a random sample $x_1, x_2, \ldots, x_n$ of size n from a distribution whose form we will denote by F. Suppose that $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the ordered sample (from the smallest to largest). The observations $x_{(1)}$ and $x_{(n)}$ are the sample extremes.

Whether we declare either of them to be an outlier depends on (at least an informal) consideration of how they appear in relation to the postulated model, F.

It is possible that $x_{(1)}$ be a lower outlier when $x_{(n)}$ is not, $x_{(n)}$ be an upper outlier while $x_{(1)}$ is not an outlier, both of $x_{(1)}$ and $x_{(n)}$ be outliers and neither $x_{(1)}$ nor $x_{(n)}$ appears to be outlying. So, we can see that “extreme values may or may not be outliers. Any outliers, however, are always extreme (or relatively extreme) values in the sample”.

Suppose now that all the observations come from the distribution F, but one or two come from a distribution G which has `slipped' upward relative to F (i.e. it has a larger mean). The observations from G termed contaminants. Such contaminants may appear as extremes, but not need to do so. Some of which may appear to be upper extreme, the other may be in the midset of the sample. We may find $x_{(n)}$ to an extreme and a contaminant, but it is not an outlier. However, in contrast, a non-extreme contaminant which have a less outlying than $x_{(n)}$ may be an outlier. So, we can say that “outliers may or may not be contaminants, contaminants may or may not be outliers”. Of course, we have no way of knowing whether or not any observation is a contaminant. All we can do is concentrate attention on outliers as `the possible manifestation of contamination'.

4.2 Outliers in Multivariate Data:

As Gnanadesikan & Kettenring (1972) remark, a multivariate outlier no longer has a simple manifestation as an observation which `sticks out at the end' of the sample. The sample has no `end'. But notably in bivariate data, we may still perceive an observation as suspiciously aberrant from the data mass, particularly so if the data are represented in the form of a scatter diagram. An observation may happens to be an extreme in one direction (y direction for example), but it is not extreme in the other direction. A multivariate outlier need not to be an extreme in any of its components.

In bivariate data, for example, the extreme value ($x_{(n)}$ $y_{(n)}$), which is the pair of marginal extremes, may not even be one of the observations in the sample. In particular, we note the relative sacrifice of any direct reflection of the dependence or correlational structure in the data.

4.3. Robustness of Correlation Coefficients:

Box and Anderson (1955) had introduced the following notion :

To fulfil the needs of the experimenter, statistical criteria should be

1) *sensitive to change* in the specific factors tested, (powerful).
2) *insensitive to changes* of a magnitude likely to occur in practice, in extraneous factors (robust)

Huber (1972) add, if one wants to choose in a rational fashion between different robust competitors to a classical procedure, one has to make precise the goals one wants to achieve.

A number of studies have recently appeared of the robustness of tests of the correlation coefficient, ρ, in a bivariate normal distribution under the prospect of contamination. Tiku and Balakrishnan (1986) introduced robust test for the correlation coefficient. The test introduced, was compared with other different known correlation coefficient, like Kendall's tau and Spearman's rho, and they consider the *robustness* in the sense that it is *the stability of Type-I error*, beside being powerful. In their paper of 1986, they had tested such stability of Type-I error and the power, for the correlation coefficients, numerically, through generating data from different bivariate distributions containing outliers, for different sample sizes.

Gideon and Hollister's new rank correlation coefficient R_g. They had made a simulation study to compare R_g with other rank correlation coefficients like Kendall's tau R_k and Spearman's rho R_s and they had indicate that the power of their correlation is not as good as Kendall's tau or Spearman's rho when the sample was derived from bivariate normal population. When the sample was derived from a bivariate exponential population, the power becomes better, and it overtook the power of Spearman's rho, specially when the sample size increase. Also they mentioned that when the samples were bivariate normal with biased outlier contamination, the powers of the correlation coefficients were ordered as they were for the pure bivariate normal case when the sample was quit small. However, the power of R_g increased relative to the others as the sample size increased. That is R_g had the most power for larger samples.

In fact, our simulation study show that R_g always has the weakest power relative to the rank correlation coefficients we had included which are Kendall's tau R_k, Spearman's rho R_s Spearman's Footrule R_f and our rank correlation coefficients R.

To verify the robustness of the mentioned correlation coefficients, a simulation study was done to discuss :

a] The stability of Type I error for these correlation coefficients.

b] The stability of the Power for these correlation coefficients.

Table 1 is given for Type-I error, while Table 2 is given for the Power at ρ=0.6.

Our simulation study of size 10,000 for testing the independence was done for the following bivariate distributions; for ρ =0 (Type I error) and for ρ =.6 (Power) both for different sample sizes n=8,10,20,30 and 40 :

Dis(1) : BN(0,0,1,1, ρ)

Outlier model:

Dis(2) : (n-1) BN(0,0,1,1, ρ) & 1 BN(0,0,4,4, ρ)
Dis(3) : (n-2) BN(0,0,1,1, ρ) & 2 BN(0,0,4,4, ρ)
Dis(4) : (n-1) BN(0,0,1,1, ρ) & 1 BN(0,0,10,10, ρ)
Dis(5) : (n-2) BN(0,0,1,1, ρ) & 2 BN(0,0,10,10, ρ)
Dis(6) : (n-1) BN(0,0,1,1, ρ) & 1 Bivariate Exponential with correlation ρ
Dis(7) : (n-2) BN(0,0,1,1, ρ) & 2 Bivariate Exponential with correlation ρ

Mixture model:

Dis(8) : 0.95 BN(0,0,1,1, ρ) + .05 BN(0,0,4,4, ρ)
Dis(9) : 0.90 BN(0,0,1,1, ρ) + .10 BN(0,0,4,4, ρ)

Bivariate Distributions with outliers only on one variable:

Dis(10) : (n-1) BN(0,0,1,1, ρ) & 1 BN(0,0,1,4, ρ)
Dis(11) : (n-1) BN(0,0,1,1, ρ) & 1 BN(0,0,4,1, ρ)
Dis(12) : (n-1) BN(0,0,1,1, ρ) & 1 BN(0,0,1,10, ρ)
Dis(13) : (n-1) BN(0,0,1,1, ρ) \& 1 BN(0,0,10,1, ρ)

Table 1: Type I error for the Correlation Coefficients

n=8	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.0648	.0650	.0681	.0792	.0660
Dis(2)	.0890	.0878	.0916	.0945	.0693
Dis(3)	.0729	.0946	.0957	.0772	.0729
Dis(4)	.1028	.0960	.0994	.1013	.0710
Dis(5)	.1262	.1160	.1057	.1073	.0720
Dis(6)	.0360	.0560	.0613	.0553	.0569
Dis(7)	.0751	.0788	.0836	.0832	.0775
Dis(8)	.0424	.0614	.0686	.0550	.0636
Dis(9)	.0424	.0614	.0686	.0550	.0636
Dis(10)	.0614	.0656	.0694	.0906	.0621
Dis(11)	.0618	.0636	.0688	.0774	.0681
Dis(12)	.0592	.0654	.0682	.0780	.0649
Dis(13)	.0664	.0653	.0704	.0794	.0630
n=10	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.0733	.0715	.0758	.0835	.0693
Dis(2)	.0893	.0830	.0916	.0900	.0776
Dis(3)	.0980	.0921	.0886	.0853	.0666
Dis(4)	.0993	.0936	.0963	.0956	.0741
Dis(5)	.1231	.1120	.1060	.1080	.0790
Dis(6)	.0656	.0600	.0685	.0698	.0650
Dis(7)	.0683	.0722	.0830	.0775	.0768
Dis(8)	.0733	.0715	.0758	.0835	.0698
Dis(9)	.0561	.0638	.0723	.0538	.0668
Dis(10)	.0633	.0640	.0670	.0756	.0698
Dis(11)	.0670	.0691	.0740	.0815	.0713
Dis(12)	.0645	.0648	.0681	.0743	.0708
Dis(13)	.0690	.0688	.0775	.0846	.0718
n=20	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.0633	.0636	.0770	.0683	.0743
Dis(2)	.0826	.0766	.0896	.0723	.0843
Dis(3)	.0923	.0886	.0926	.0820	.0920
Dis(4)	.0860	.0776	.0823	.0763	.0786
Dis(5)	.1083	.0933	.1016	.0956	.0790
Dis(6)	.0583	.0566	.0653	.0596	.0683
Dis(7)	.0666	.0646	.0743	.0716	.0883
Dis(8)	.0633	.0636	.0770	.0683	.0743
Dis(9)	.0633	.0636	.0770	.0683	.0743
Dis(10)	.0673	.0630	.0770	.0670	.0836
Dis(11)	.0703	.0763	.0806	.0716	.0870
Dis(12)	.0650	.0630	.0726	.0643	.0846
Dis(13)	.0603	.0613	.0710	.0686	.0793

Table 1 (continued)

n=30	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.0740	.1275	.0845	.0610	.0750
Dis(2)	.0920	.1485	.0910	.0715	.0755
Dis(3)	.0865	.1395	.0895	.0685	.0765
Dis(4)	.0920	.1435	.0920	.0755	.0720
Dis(5)	.1010	.1580	.0830	.0765	.0665
Dis(6)	.0650	.1115	.0780	.0600	.0630
Dis(7)	.0810	.1310	.0850	.0805	.0875
Dis(8)	.0740	.1275	.0845	.0610	.0750
Dis(9)	.0740	.1275	.0845	.0610	.0750
Dis(10)	.0710	.1225	.0865	.0615	.0650
Dis(11)	.0665	.1165	.0760	.0555	.0670
Dis(12)	.0710	.1245	.0835	.0595	.0725
Dis(13)	.0665	.1200	.0780	.0555	.0660
n=40	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.0653	.0700	.0833	.0906	.0873
Dis(2)	.0733	.0753	.0913	.0993	.0840
Dis(3)	.0733	.0786	.0926	.1006	.0833
Dis(4)	.0800	.0846	.0920	.1046	.0866
Dis(5)	.0826	.0800	.0920	.1066	.0813
Dis(6)	.0653	.0660	.0893	.0993	.0886
Dis(7)	.0606	.0593	.0940	.1026	.0980
Dis(8)	.0653	.0700	.0833	.0906	.0873
Dis(9)	.0653	.0700	.0833	.0906	.0873
Dis(10)	.0633	.0713	.0720	.0893	.0806
Dis(11)	.0633	.0700	.0860	.0986	.0866
Dis(12)	.0640	.0706	.0706	.0880	.0797
Dis(13)	.0646	.0686	.0786	.0920	.0806

Table 2: The Power for the Correlation Coefficients at $\rho=0.6$

n=8	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.3005	.3012	.2848	.3008	.1975
Dis(2)	.2709	.3205	.3081	.3308	.1954
Dis(3)	.3389	.3309	.3125	.3440	.2016
Dis(4)	.3542	.3353	.3202	.3500	.1937
Dis(5)	.3686	.3504	.3325	.3634	.2060
Dis(6)	.3280	.3292	.3153	.3276	.2133
Dis(7)	.3906	.3606	.3429	.3622	.2289
Dis(8)	.3017	.3052	.2861	.3038	.1968
Dis(9)	.3017	.3052	.2861	.3038	.1986
Dis(10)	.2680	.2744	.2621	.2701	.1862
Dis(11)	.2790	.2809	.2714	.2742	.1910
Dis(12)	.2598	.2625	.2561	.2601	.1809
Dis(13)	.2602	.2673	.2578	.2590	.1840

Table 2 (continued)

n=10	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.3886	3852	.3636	.3417	.2480
Dis(2)	.4306	.4065	.3818	.4286	.2618
Dis(3)	.4261	.3916	.3715	.4186	.2565
Dis(4)	.4393	.4105	.3845	.4291	.2626
Dis(5)	.4401	.4013	.3831	.4430	.2756
Dis(6)	.4385	.4266	.4013	.4205	.2656
Dis(7)	.4725	.4665	.4383	.4515	.2882
Dis(8)	.4066	.3946	.3656	.3860	.2516
Dis(9)	.4066	.3946	.3656	.3860	.2516
Dis(10)	.3706	.3653	.3460	.3571	.2385
Dis(11)	.3660	.3560	.3411	.3491	.2386
Dis(12)	.3598	.3518	.3380	.3415	.2408
Dis(13)	.3650	.3606	.3448	.3460	.2353
n=20	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.3886	.3852	.3636	.3417	.2480
Dis(1)	.7642	.7447	.6781	.6715	.5099
Dis(2)	.7366	.7073	.6686	.6863	.5083
Dis(3)	.7403	.6963	.6653	.6933	.5176
Dis(4)	.7436	.6910	.6616	.6923	.5220
Dis(5)	.7360	.6773	.6613	.7160	.5196
Dis(6)	.7816	.7730	.7076	.6990	.5390
Dis(7)	.7970	.7830	.7303	.7280	.5683
Dis(8)	.7680	.7486	.6910	.6783	.5223
Dis(9)	.7680	.7486	.6910	.6783	.5223
Dis(10)	.7440	.7326	.6686	.6560	.5060
Dis(11)	.7433	.7276	.6640	.6640	.4996
Dis(12)	.7396	.7230	.6676	.6600	.5096
Dis(13)	.7396	.7153	.6640	.6630	.5140
n=30	**R_k**	**R_s**	**R**	**R_f**	**R_g**
Dis(1)	.9229	.9471	.8501	.8523	.6657
Dis(2)	.9085	.9310	.8375	.8515	.6615
Dis(3)	.9025	.9230	.8335	.8575	.6605
Dis(4)	.9145	.9260	.8500	.8680	.6820
Dis(5)	.8960	.9105	.8370	.8690	.6745
Dis(6)	.9330	.9550	.8760	.8810	.6935
Dis(7)	.9365	.9630	.8780	.8815	.7085
Dis(8)	.9275	.9555	.8505	.8470	.6595
Dis(9)	.9275	.9555	.8505	.8470	.6595
Dis(10)	.9240	.9515	.8575	.8555	.6890
Dis(11)	.9315	.9575	.8645	.8640	.6790
Dis(12)	.9130	.9390	.8490	.8420	.6725
Dis(13)	.9200	.9525	.8530	.8560	.6650

Table 2 (continued)

n=40	R_k	R_s	R	R_f	R_g
Dis(1)	.9784	.9718	.9390	.9352	.7818
Dis(2)	.9660	.9546	.9260	.9240	.7633
Dis(3)	.9646	.9593	.9253	.9333	.7806
Dis(4)	.9673	.9546	.9300	.9280	.7586
Dis(5)	.9640	.9486	.9193	.9386	.7766
Dis(6)	.9800	.9713	.9373	.9373	.7880
Dis(7)	.9746	.9726	.9406	.9346	.8060
Dis(8)	.9746	.9666	.9313	.9313	.7680
Dis(9)	.9747	.9667	.9313	.9313	.7680
Dis(10)	.9680	.9606	.9306	.9313	.7733
Dis(11)	.9693	.9626	.9313	.9253	.7800
Dis(12)	.9680	.9606	.9220	.9213	.7780
Dis(13)	.9700	.9626	.9293	.9286	.7766

5. COMMENTS AND CONCLUSION

From our previous study, we may have the following two comments:

Stability of Type-I error. Through the bivariate distributions we had simulate from, Gideon & Hollister's R_g seems to be the more stable correlation coefficients between the correlation coefficients we are comparing with, in the sense of stability of Type I error - may be due to being insensitive. Following R_g, is the correlation coefficient R, although being sensitive. Kendall's R_k has the less stability specially for small sample sizes. Spearman's Footrule R_f gains reasonably stability and Spearman's R_s has the less stability, specially for large sample sizes.

Stability of the Power. For small *n*, R_g has the more stability for the power, while it loses such property for large *n* and it is the less power values in both cases. The reverse can be said about R_k which seems to have the less power stability for small *n* and begins to have a good power stability for large *n*. R_f seems to follow R_k's behavior. R has a reasonably power stability for all *n*, while R_s has the less power stability for all *n*.

In the sense mentioned above, the correlation coefficient R gains a good *resistance* or stability *(robustness) to outliers*, beside its *high power* comparing with all the other mentioned correlation coefficients (specially when the correlation between the two variates appears to be very small, were it has the most Power), that we recommend it to be used as a good measure for the correlation, specially if the outliers exists in the data.

REFERENCES

1. Abdelfattah Ezz H. (1996). *Measures of rank correlation coefficient resistant to outliers*. Unpublished Dissertation, Helwan University.
2. Box, G.E. and Anderson, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption (with discussion). *J. Roy. Statist. Soc.*, Ser. B 17 1-34.
3. Diaconis, P. and Graham, R.L., (1977). Spearman's footrule as a measure of disarray. *J. Roy. Statist. Soc*., *Series B Statistical Methodology*, 39, 262-268.

4. Franklin, L. (1988). Exact tables of Spearman's footrule for N=11(1)18 with estimate of convergence and errors for the normal approximation. *Statistics and Probability Letters*, 311(6), 399-406.
5. Gideon R.A. and Hollister A. (1987). A rank correlation coefficient resistant to outliers. *J. Amer. Statist. Assoc.*, 82, 656-666.
6. Gnanadesikan, R. and Kettenring, J.R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*. 28, 81-124.
7. Huber J. (1972). Robust statistics: A review. *Ann. Math. Statist.*, 43, 1041-1067.
8. Kendall M.G. (1938). A new measure of rank correlations. *Biometrika*, 30, 91-93.
9. Kendall M.G. and Gibbons (1990). Hafner Pub. Co., 1962
10. Salama I.A. and Quade, D. (2001). The symmetric footrule. *Comm. Statist.*, *Part A-Theory and Methods*, 30, 1099-1109.
11. Sen, P.K., Salama I.A. and Quade D. (2011). Spearman's footrule: asymptotics in applications. *Chilean Journal of Statistics*. 2(1), 1-18.
12. Spearman C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72-101.
13. Tiku and Balakrishnan (1986). A robust test for testing the correlation coefficient. *Comm. Stat - Simulation and Computation*, 15(4), 945-971.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 217-224

A CAKE BAKING EXPERIMENT USING FRACTIONAL FACTORIAL SPLIT PLOT (FFSP) DESIGN

Farah Yasmeen and **Asim Jamal Siddiqui**
Department of Statistics, University of Karachi, Karachi, Pakistan
Email: riazfarah@yahoo.com; asimjs2000@yahoo.com

ABSTRACT

In performing two-level factorial and fractional factorial experiments, it is usually assumed that the design structure is completely randomized. However, in some multifactor experiments, it is not feasible to run all the factors in a completely random. In industrial and agricultural experiments with many factors, a split-plot design is often useful and provides a practical design option. Recently, these designs have been applied to two-level fractional factorials.

In this paper, we have applied two-level fractional factorial split-plot(FFSP) design to a cake baking experiment and have determined the important factors that influence the response (e.g. height). We have also discussed the analysis of these designs Some important aspects of these designs including the distribution into whole-plot and split-plot factors and the selection of appropriate fraction have also been discussed.

KEY WORDS

Two-level Factorials, Fractional Factorials, Design structure, Split-plot Designs, Resolution, Minimum Aberration.

1. INTRODUCTION

In some multifactor experiments with factorial arrangement of treatments, it may not be possible to randomize completely the order of experimentation. Hence we consider restrictions on randomization and both the randomized block and Latin-square designs are used. There are still many practical situations in which it is not possible to even randomize within a block. Hence, a split-plot design is resulted. A split-plot design is simply the generalization of the randomized block design. In this design, a main treatment is applied to large plots in the blocks of an experiment and these plots are then subdivided into smaller sub-or split-plots to study additional factors or treatments.

2. TWO-LEVEL FRACTIONAL FACTORIAL SPLIT-PLOT (FFSP) DESIGNS

Consider an experiment with `n' two-level factors. It may be possible that the levels for some of the factors say ‘n_1’ are difficult to change or it is very expensive or time consuming to change the levels of these factors. To save cost or time, one may randomly choose the factor-level setting of one of these n_1 factors and then run all of the treatment combinations of the remaining $n_2 = n\text{-}n_1$ factors in a random order. Due to this restriction on randomization, a fractional factorial split-plot (FFSP) design is resulted. Here n_1 and n_2 factors are called the whole-plot (WP) and Split-plot (SP) factors respectively. In general, we define a $2^{(n1+n2)-(p1+p2)}$ design, a two-level FFSP design with n_1 factors in the WP and fractionation element p_1 and n_2 factors in the SP with fractionation element p_2.

2.1 Selection of best possible FFSP Design

The best possible FFSP designs can often be constructed by using all the factors involved in the subplot and some of the factors from the whole plots when creating the generators for the split plot factors. In order to decide which FFSP design to be used, the minimum aberration (MA) criterion (Fries and Hunter 1980) is often useful. This criterion has recently been applied to two-level split-plot designs (see Bingham and Sitter 1999). The criteria are described below.

Definition: Suppose that D_1 and D_2 are two FFSP designs. Let W= ($w_{1,}$ w_2, …., w_{n1+n2}) is the word length pattern (WLP) of the design where Wi: Number of words of length 'i' in the complete set of defining relations

Let 'r' be the smallest 'i' such that two word lengths are not equal. Then D_1 is said to have less aberration than D_2 if Wr(D1) < Wr(D_2). A $2^{(n1+n2)-(p1+p2)}$ FFSP design is said to be MA FFSP design if no other $2^{(n1+n2)-(p1+p2)}$ design has less aberration. Franklin 1984 used the idea of generating matrix to represent the defining relation for a 2^{k-p} fractional factorial design. For split-plot experiments, this idea has been extended by Huang et al 1998. They have created a catalog of MA FFSP designs. Yasmeen 2004 also collected a complete catalog of all possible 8-run, 16 run and 3-run two-level FFSP designs with minimum aberration.

2.2 Analyzing FFSP designs

For analyzing a two-level FFSP design, let us first consider the standard model for a full factorial split-plot design

$$y = f_1(WP) + € + f_2(SP) + \delta \qquad (1)$$

where f_1(WP) and f_2(SP) are the functions of WP and SP factors respectively and € are δ are the corresponding error terms. The following usual assumptions are made

(i) $€ \sim N(0, \sigma_€^{\,2})$, (ii) $\delta \sim N(0, \sigma_\delta^{\,2})$ (iii) $\sigma_€^{\,2} > \sigma_\delta^{\,2}$

Table 1:
Effects and Corresponding Error Terms

Effects to be tested	Appropriate Error terms
1. Main effects and interaction effects of the WP factors	WP error term €
2. Main effects and interaction effects of the SP factors that are aliased with WP main effects or interaction effects	WP error term €
3. Main effects and interaction effects of the SP factors involving at least one split-plot factor aliased with the interactions of WP and SP factors	SP error term δ

The issues related to the errors are summarized in table 1.

3. CAKE BAKING EXPERIMENT

Ahmed and Suboohi 1993 discussed a cake baking experiment where an unreplicated two-level fractional factorial design with completely randomized design structure was performed.

In this paper, we are going one step further. We are applying a two-level FFSP design. The first important thing is to divide the factors into WP and SP factors. The next step is to obtain a fraction from the WP and SP deigns and to combine them to obtain required FFSP design with good design properties. For this purpose, we considered the following eight factors in our experiment. The factors and their levels are given in table 2.

Selection of FFSP design

We decided to run a 16-run design with eight factors. Since among them temperature is difficult to change every time as compare to other factors, temperature and beating time are considered whole plot (WP) factors whereas the other six factors are considered as the split- plot(SP) factors.

Table 2
Factors and their levels for cake baking experiment using FFSP design

Factors	Low Level (-)	High Level (+)
Liquid (Milk)	1/8 cup	1/ 4 cup
Baking Powder	1 /2 teaspoon	1 teaspoon
Eggs	1	2
Sugar	1 /4 cup	1/ 2 cup
Butter	1/ 4 cup	1/ 2 cup
Beating time	5 minutes	7 minutes
Oven temperature	200o C	225 0 C
Pan size	Small	Large

We take a half fraction of WP factors (2^{2-1}) design and one-eighth fraction form SP runs (2^{6-3}) design. From the catalog of 16 run MA FFSP design (see Yasmeen 2004), only one $2^{(2+6)-(1+3)}$ design is possible. So we decided to run this design with the factor are labelled as

WP Factors: Temperature (A), Beating Time (B)
SP Factors: Liquid (p), Baking Powder (q), Eggs (r), Sugar (s), Butter (t) and Pan size (u)

The design generator and the complete defining relations are

$$r = Abq,\ s=ABp,\ t = Apq, \text{ and } u= Bpq \tag{2}$$

$$\begin{aligned}\text{Hence}\quad I &= Abqr = ABps = Apqt = Bpqu = pqrs = Bprt = Apru = Bqst \\ &= Aqsu = Abtu = Arst = Brsu = pstu = qrtu = Abpqrstu\end{aligned} \tag{3}$$

The WLP of this design is W = (0, 14, 0, 0, 0, 1). Hence this is a resolution IV design with all main effects are free from the two-factor interaction aliasing. However the two-factor interactions are aliased with each other. The modified alias structure is shown in table 3.

The actual treatment combinations run in the experiment and the corresponding responses are given in table 4 (Here, response variable is size of the cake in cms^3).

Table 3
Modified Alias Structure

A,	B,	p,
q,	r,	s,
t,	u,	AB+ps+qr+tu
Ap+Br+pt+su	Bp+As+qu+rt	Aq+Br+pt+su
Bq+Ar+pu+st	pq+At+Bu+rs	Au+Bt+pr+qs

To see how the experiment was conducted, the same data of table 4 can be arranged in the form of table 5. This table shows the contrasts corresponding to the WP and SP factors.

In order to conduct the experiment, a treatment combination from the WP treatments e.g (-, +) is randomly chosen. It means we have to fix the factors (temperature at low level 200^{0}C) and beating time should be at high level (i.e. 7 minutes). Now four different treatment combinations from the other 2^6 = 64 treatments of SP are selected.

Table 4
Treatments and response for FFSP design in Cake Baking Experiment

Run	A	B	p	q	r	s	t	u	Response
1	200^{0}C	5 min.	1/8 cup	1/2 ts	1	¼ cup	¼ cup	small	230.9
2	225^{0}C	5 min.	1/8 cup	1/2 ts	2	½ cup	¼ cup	Small	461.81
3	200^{0}C	7 min.	1/8 cup	1/2 ts	2	1/2 cup	¼ cup	large	340.47
4	225^{0}C	7 min.	1/8 cup	1/2 ts	1	¼ cup	¼ cup	large	226.98
5	200^{0}C	5 min.	1/4 cup	1/2 ts	1	1/2 cup	¼ cup	large	567.45
6	225^{0}C	5 min.	1/4 cup	1/2 ts	2	¼ cup	¼ cup	large	226.98
7	200^{0}C	7 min.	1/4 cup	1/2 ts	2	¼ cup	¼ cup	small	384.84
8	225^{0}C	7 min.	1/4 cup	1/2 ts	1	1/2 cup	¼ cup	small	**384.8**
9	200^{0}C	5 min.	1/8 cup	1 ts	2	¼ cup	¼ cup	large	453.96
10	225^{0}C	5 min.	1/8 cup	1 ts	1	1/2 cup	¼ cup	large	295.07
11	200^{0}C	7 min.	1/8 cup	1 ts	1	1/2 cup	¼ cup	small	307.87
12	225^{0}C	7 min.	1/8 cup	1 ts	2	¼ cup	¼ cup	small	384.8
13	200^{0}C	5 min.	1/4 cup	1 ts	2	1/2 cup	¼ cup	small	307.87
14	225^{0}C	5 min.	1/4 cup	1 ts	1	¼ cup	¼ cup	small	230.9
15	200^{0}C	7 min.	1/4 cup	1 ts	1	¼ cup	¼ cup	large	226.98
16	225^{0}C	7 min.	1/4 cup	1 ts	2	1/2 cup	¼ cup	large	680.94

The ingredients corresponding to these four combinations are placed into four bowls. Now, one bowl is randomly selected and the ingredients are beaten for 7 minutes and the mixture is placed into a large pan. Then select another bowl and do the same process. After beating the ingredients from the four bowls, we would have two small sized pans and two large sized pans.

Table 5: Contrast Table for the WP and SP factors

WP units	WP factors		SP factors						
	Temp. (A)	Time (B)	Liquid (p)	Baking powder (q)	Eggs (r)	Sugar (s)	Butter (t)	Pan Size (u)	Response
1	-	-	-	-	-	-	-	-	230.90
			+	-	-	+	+	+	567.45
			-	+	+	-	+	+	453.96
			+	+	+	+	-	-	307.87
2	+	-	-	-	+	+	+	-	461.81
			+	-	+	-	-	+	226.98
			-	+	-	+	-	+	295.07
			+	+	-	-	+	-	230.90
3	-	+	-	-	+	-	-	+	340.47
			+	-	+	+	+	-	384.84
			-	+	-	+	+	-	307.87
			+	+	-	-	-	+	226.98
4	+	+	-	-	-	+	+	+	226.98
			+	-	-	-	-	-	384.80
			-	+	+	-	-	-	384.84
			+	+	+	+	+	+	680.94

Next, we put these four pans simultaneously into the oven and bake these cakes at 200^0 C temperature. Then we perform the same experiment for the other treatment combinations. In this way, the responses corresponding to the 16 treatment combinations are obtained which are showed in table 5.

Analysis of Split-plot Design

Here we have four WP units so the factors A, B, and AB interactions will be tested against the WP error, whereas the other factors and their interactions will be tested against the SP error. We perform ANOVA to confirm our results and use the added effects of A and B as an estimate of error. The other factors p, q, r, s, t and u and some of their two-factor interactions are tested against the SP error. The effects and their estimates are given in table 6.

Table 6: Estimated Average Effects and Alias Structure

Effects	Estimates	Effects	Estimates
Average	356.98	q (Baking powder)	7.90
A (temerature)	8.87	Aq+Br+pt+su	64.65
B (Beating Time)	20.22	Bq+Ar+pu+st	57.74
AB + ps + qr + tu	95.22	R (Eggs)	96.22
p (Liquid)	38.74	pq+At+Bu+rs	-37.26
Ap+Br+pt+su	0.26	t(Butter)	114.71
Bp +As +qu+rt	65.88	u (Pan size)	40.73
s (Sugar)	122.63	Au+Bt+pr+qs	-48.61

There are 12 effects that can be tested against the split-plot error, so we may construct a normal probability plot of these effects,which is shown in Figure 1. From this figure, the significant effects from the split plot factors are s, t and r. The ANOVA is summarized in table 7. The effect corresponding to the contrast AB+ps+qr+tu appears to be significant. Since neither factor A (temperature) nor the factor B (beating time) is significant, it is most likely that the interaction effects ps+qr+tu is significant. Among the SP treatments, the largest absolute effects are r, s and t and the two sets of two-factor interactions Bp+As+qu+rt and Aq+Br+pt+su. These effects are tested against the SP error, whose estimate is obtained by pooling the other contrasts that are negligible.

From both the normal probability plot and ANOVA table, it is clear that the significant factors are sugar(s), butter(t) and eggs(r). The size of the cake is considered as response variable, as the cake will be soft and fluffy if its size will be large, so the objective here is to maximize the cake. For this, the experimenter has to use the high level of sugar (i.e. ½ cup), high level of butter (½ cup) and the number of eggs should be large (i.e. 2) because these three factors have positive main effects. Effects of the other factors like pan size, temperature and beating time are found to be insignificant.

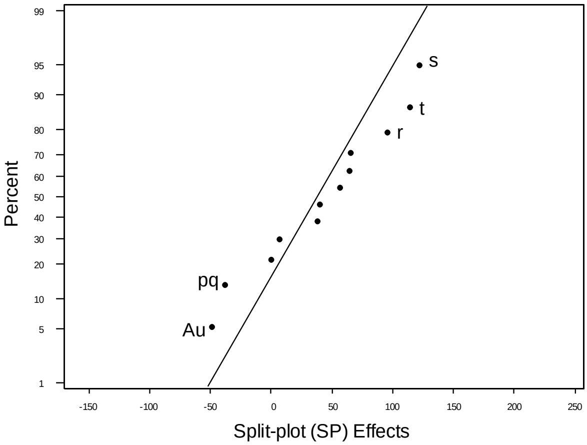


Fig. 1: Normal Probability Plot of the Effects

4. CONCLUSION

In this paper, we applied a FFSP design to a cake baking experiment. We run a 16-run design with eight factors. Since the temperature is difficult to change every time, the temperature and beating time are considered whole plot (WP) factors whereas the other six factors are considered as the split-plot(SP) factors. We select the best possible FFSP design in 16-runs. From this experiment, it is found that the significant factors are sugar(s), butter(t) and eggs(r). The size of the cake is considered as response variable. Our analysis suggests that the experimenter has to use the high level of sugar (i.e. ½ cup), high level of butter (½ cup) and the number of eggs should be large (i.e. 2). Effects of the other factors are appeared to be insignificant.

Table 7
ANOVA Table for the FFSP Design for Cake Baking Experiment

Source of Variation	Degrees of Freedom	Sum of squares	Mean Squares	F-ratio
Whole-plot Analysis				
AB+ps+qr+tu	1	36277.9	36277.9	37.168*
WP Error	2	1952.1	976.05	
Split-Plot Analysis				
s (sugar)	1	60141.1	60141.1	10.2127*
r (Eggs)	1	37028.3	37028.3	6.2878*
t (Butter)	1	52646.2	52646.2	8.9399*
pq + A t + Bu +rs	1	17357.4	17357.4	2.9475
Au + Bt + pr + qs	1	16712.7	16712.7	2.8380
Split-plot Error	7	41222	5888.8	
Total	15	263291		

*Significant at 5% L.O.S.

REFERENCES

1. Ahmed, E. and Suboohi, A. (1993). On determining Sparse Factors for cake baking using unreplicated fractional factorial experiments. *Quality Engineering*, 5(4), 571-581.
2. Bingham, D. and Sitter, R.R. (1999). Minimum Aberration Two-Level Fractional Factorial Split-plot designs. *Technometrics,* 41, 62-70.
3. Franklin, M.F. (1984). Constructing Tables of Minimum Aberration p^{n-m} Designs. *Technometrics*, 26, 225-232.
4. Fries, A. and Hunter, W.G. (1980). Minimum Aberation 2^{k-p} designs. *Technometrics*, 22, 601-608.
5. Huang, P.; Chen, D. and Voelkle, J. (1998). Minimum Aberration two-level split-plot designs. *Technometrics*, 40, 314-326.
6. Yasmeen, F. (2004). *Study of two-level fractional factorial designs with different design structures*. Unpublished M.Phil thesis, University of Karachi.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 225-232

GROWTH CHARTS OF MALAYSIAN PRESCHOOL CHILDREN

Asma Ahmad Shariff[1], **Bong Yii Bonn**[2], **Abdul Majid Mohamed**[1],
and **Amir Feisal Merican**[3]

[1] Center for Foundation Studies in Science, University of Malaya,
Kuala Lumpur, Malaysia. Email: asma@um.edu.my; ammajid@um.edu.my

[2] Institute of Graduate Studies, University of Malaya, Kuala Lumpur, Malaysia.
Email: yiibonn@siswa.um.edu.my

[3] Institute of Biological Sciences, Faculty of Science, University of Malaya,
Kuala Lumpur, Malaysia. Email: merican@um.edu.my

ABSTRACT

The assessment of growth is crucial for health care provider and paediatricians to evaluate a child's well-being in terms of nutritional status and physiological needs and to identify any discrepancies of growth failure. Malaysian doctors and child health professionals have been using the western norms in assessing the growth and development of Malaysian infants and children. Thus, this study is adopted in order to provide growth reference data and hence the growth reference curve or chart for Malaysian infants and preschool children of 0-6 years of age. Height and weight data were collected in a cross-sectional study of 15,350 children around the country. Health clinic-based data were opted by two-stage stratified random sampling technique based on stratification for states and clinics. Records for healthy infants without any prenatal diseases were taken into consideration. The Cole's LMS method was used for calculation of growth percentiles. From the study it was found that boys are significantly taller and heavier than girls ($p<0.05$). Newborns demonstrated similar weight despite their genders. Both genders showed high growth velocity in stature from birth to 1.5 years old. The proposed growth charts can be used as growth reference for growth monitoring as they were developed based on the mix-feeding practice (breastfed and formulae-fed) of the community presently.

KEYWORDS

Growth references, infants, preschool children, Malaysia, Cole's LMS method.

1. INTRODUCTION

Growth references have become one of the most adequate guidelines to evaluate the well-being of a child. These reference data are central to growth monitoring as they help doctors to diagnose growth-related conditions. The assessment of growth is crucial for health care provider and paediatricians to evaluate a child's well-being in terms of nutritional status and physiological needs and to identify any discrepancies of growth failure.

Malaysian doctors and child health professionals have been using the western norms in assessing the growth and development of Malaysian infants and children. The international growth charts might allow comparison between countries, but the regional or national references are most useful in the assessment of local changes in nutrition status of a population (Eveleth & Tanner, 1976).

The need to develop a growth reference for screening, surveillance and monitoring had been stirred by the lack of local reference for growth evaluation. Thus, this study is adopted in order to provide growth reference data and hence the growth reference curve or chart for Malaysian infants and preschool children of 0-6 years of age.

2. MATERIALS AND METHODS

Study protocol for the infant group was first submitted to the National Medical Research Registrar (NMRR) committee for review. The research conforms to the conditions as stated in all approval letters. Health clinics-based data were opted for infant group. A discussion was arranged with the Matron / Head Nurse before data for the infant was retrieved from records archive.

States (primary unit) were randomly selected from each region in the first stage of the two-stage stratified cluster sampling technique. Health Clinics and Maternal Child Health Clinics (MCH) were identified from each state. Then, one Health clinic or MCH clinic was drawn from each state in compliance with the state's matron / head nurse instructions (second stage sampling). All the records for healthy infants were taken into consideration as subjects for this study. Both breastfed and formula-fed babies were therefore included in the development of growth charts to better represent the combined growth patterns in the general population.

Data collection for infant category took approximately six months (February until July 2011). These anthropometric measurements were collected by qualified and well-trained nurses and health care personnel. Height and weight data were collected in a cross-sectional study of 15,350 children around the country, 51.0% of whom were boys (n=7,824) and 49.0% were girls (n=7,526).

The Cole's LMS method was used for calculation of growth percentiles. Height tends to be normally distributed, but weight does not strictly follow a normal distribution. The Cole's LMS method assumes that data can be normalized by using a power transformation, which stretches one tail of the distribution and shrinks the other, thus removing the skewness (Cole, 1990).

The LMS method summarizes each standard with three smooth curves, where L curve represents the power needed to normalize the data, M curve (median) and S curve is the coefficient of variation of the distribution at each age. In brief, if "Anth" is the measured anthropometry for a child at t age (year), then the SD score is calculated as (Cole, 1990)

$$Z = \frac{\left[{Anth}/{M(t)}\right]^{L(t)} - 1}{L(t)S(t)} \tag{1}$$

Finally, after estimating L (t), M (t) and S (t) for each half-year age (t), the 100αth centile could be derived from

$$C_{100\alpha}(t) = M(t)[1 + L(t)S(t)Z_{\alpha}]^{\frac{1}{L(t)}} \qquad (2)$$

3. RESULTS

Calculations were made in a three-month interval range from birth to 24 months and half-year period between the age group of two to six years old for the rationale of presentation. Data from both breastfed and formula-fed infants, regardless of their geographic or socioeconomic backgrounds, were included as the sample in this study.

Table 1 displays the LMS values and crude percentiles for infants' recumbent length / height whereas Table 2 presents the LMS values for infants' weight. Smoothed centile curves by gender were drawn in accordance to the selected percentiles (Figure 1 and 2).

The mean values for recumbent length / height increased with the age of infants for both genders. Infant boys were longer / taller than infant girls, except at age 4.5 and 5.5 years. Both gender showed high growth velocity in stature from birth to 1.5 years old.

Newborns demonstrated similar weight for both genders, with an average of 3.06 kg for boys and 3.09 kg for girls. However, the weight percentiles dispersed as they grow older. Boys were heavier than girls for almost all age groups, except at birth (newborns), 4.5 years old (15.37 kg ± 3.43 versus 15.39 kg ± 3.16) and 5.5 years old (18.35 kg ± 4.57 versus 18.62 kg ± 3.70).

Table 1
LMS values and crude percentiles for infants' length / height

Gender	Age (years)	L	M	S	Percentiles						
					P3	P10	P25	P50	P75	P90	P97
Boys	0.0	3.68	50.42	0.08	41.03	44.62	47.61	50.42	52.87	54.84	56.61
	0.5	1.16	60.32	0.10	48.55	52.34	56.15	60.32	64.45	68.14	71.74
	1.0	-0.32	73.38	0.06	66.10	68.31	70.65	73.38	76.24	78.96	81.76
	1.5	2.03	78.57	0.06	68.82	72.08	75.23	78.57	81.78	84.56	87.21
	2.0	1.65	85.85	0.06	75.19	78.69	82.14	85.85	89.46	92.64	95.70
	2.5	1.08	89.31	0.06	79.90	82.91	85.95	89.31	92.66	95.67	98.63
	3.0	2.52	94.67	0.06	83.18	87.09	90.81	94.67	98.31	101.43	104.36
	3.5	1.98	97.55	0.08	82.09	87.30	92.30	97.55	102.53	106.83	110.91
	4.0	2.55	101.77	0.06	88.12	92.81	97.22	101.77	106.02	109.65	113.04
	4.5	-0.18	103.83	0.06	93.51	96.66	99.98	103.83	107.85	111.63	115.51
	5.0	-0.52	107.59	0.06	96.06	99.52	103.23	107.59	112.23	116.69	121.35
	5.5	0.66	109.76	0.07	95.17	99.74	104.45	109.76	115.16	120.11	125.05
	6.0	2.71	114.37	0.06	98.76	104.17	109.21	114.37	119.16	123.22	127.00
Girls	0.0	0.38	49.97	0.05	45.22	46.70	48.23	49.97	51.74	53.37	55.01
	0.5	0.62	59.01	0.10	48.59	51.82	55.19	59.01	62.93	66.55	70.19
	1.0	1.62	72.39	0.06	64.55	67.11	69.64	72.39	75.07	77.44	79.74
	1.5	1.37	77.91	0.07	67.84	71.11	74.36	77.91	81.40	84.50	87.52
	2.0	1.46	84.37	0.06	74.59	77.76	80.93	84.37	87.75	90.75	93.66
	2.5	0.33	88.11	0.05	79.33	82.06	84.90	88.11	91.41	94.46	97.53
	3.0	2.21	93.66	0.06	81.99	85.91	89.69	93.66	97.44	100.70	103.80
	3.5	0.63	97.15	0.06	86.62	89.92	93.33	97.15	101.04	104.59	108.13
	4.0	2.04	101.35	0.06	90.04	93.79	97.45	101.35	105.09	108.36	111.47
	4.5	0.31	104.00	0.05	93.78	96.96	100.25	104.00	107.84	111.40	114.97
	5.0	2.37	107.56	0.06	93.22	98.10	102.74	107.56	112.11	116.00	119.67
	5.5	1.81	111.96	0.07	95.79	101.17	106.39	111.96	117.31	121.97	126.43
	6.0	3.35	113.94	0.06	97.72	103.55	108.77	113.94	118.61	122.48	126.04

Table 2
LMS values and crude percentiles for infant's weight

Gender	Age (years)	L	M	S	Percentiles						
					P3	P10	P25	P50	P75	P90	P97
Boys	0.0	1.73	3.09	0.15	2.10	2.44	2.76	3.09	3.39	3.64	3.88
	0.5	0.51	5.65	0.27	3.11	3.84	4.66	5.65	6.74	7.80	8.92
	1.0	0.16	8.68	0.14	6.62	7.22	7.89	8.68	9.55	10.38	11.27
	1.5	0.17	9.54	0.15	7.16	7.85	8.62	9.54	10.53	11.51	12.54
	2.0	-0.35	10.97	0.15	8.45	9.16	9.97	10.97	12.12	13.30	14.62
	2.5	-0.70	11.51	0.15	8.89	9.60	10.43	11.51	12.78	14.16	15.78
	3.0	-0.70	12.87	0.18	9.51	10.40	11.46	12.87	14.60	16.53	18.89
	3.5	-0.16	13.43	0.19	9.57	10.64	11.87	13.43	15.24	17.12	19.23
	4.0	-1.24	14.50	0.18	10.87	11.79	12.92	14.50	16.60	19.19	22.85
	4.5	-1.45	14.73	0.17	11.31	12.18	13.24	14.73	16.72	19.20	22.77
	5.0	-1.24	16.75	0.23	11.91	13.08	14.57	16.75	19.83	23.97	30.62
	5.5	-1.14	17.37	0.22	12.40	13.62	15.16	17.37	20.41	24.33	30.21
	6.0	-0.36	19.60	0.23	13.07	14.78	16.84	19.60	23.03	26.86	31.53
Girls	0.0	-1.33	3.02	0.13	2.44	2.59	2.78	3.02	3.32	3.66	4.08
	0.5	0.31	5.23	0.26	3.05	3.66	4.36	5.23	6.22	7.21	8.30
	1.0	0.43	8.17	0.13	6.26	6.84	7.45	8.17	8.92	9.64	10.37
	1.5	0.61	9.17	0.16	6.63	7.40	8.22	9.17	10.15	11.08	12.02
	2.0	-0.22	10.44	0.15	7.98	8.68	9.46	10.44	11.53	12.64	13.87
	2.5	-0.77	11.09	0.15	8.57	9.25	10.05	11.09	12.34	13.70	15.31
	3.0	-1.56	12.36	0.17	9.55	10.25	11.13	12.36	14.00	16.08	19.14
	3.5	-1.24	13.16	0.18	9.91	10.74	11.75	13.16	15.01	17.28	20.45
	4.0	-1.25	14.44	0.17	11.05	11.92	12.98	14.44	16.33	18.59	21.65
	4.5	-1.85	14.68	0.17	11.37	12.19	13.21	14.68	16.75	19.55	24.21
	5.0	-0.68	16.12	0.19	11.64	12.81	14.22	16.12	18.48	21.16	24.51
	5.5	-0.47	18.11	0.19	12.92	14.31	15.95	18.11	20.74	23.62	27.07
	6.0	-0.11	18.28	0.19	12.77	14.30	16.05	18.28	20.87	23.55	26.57

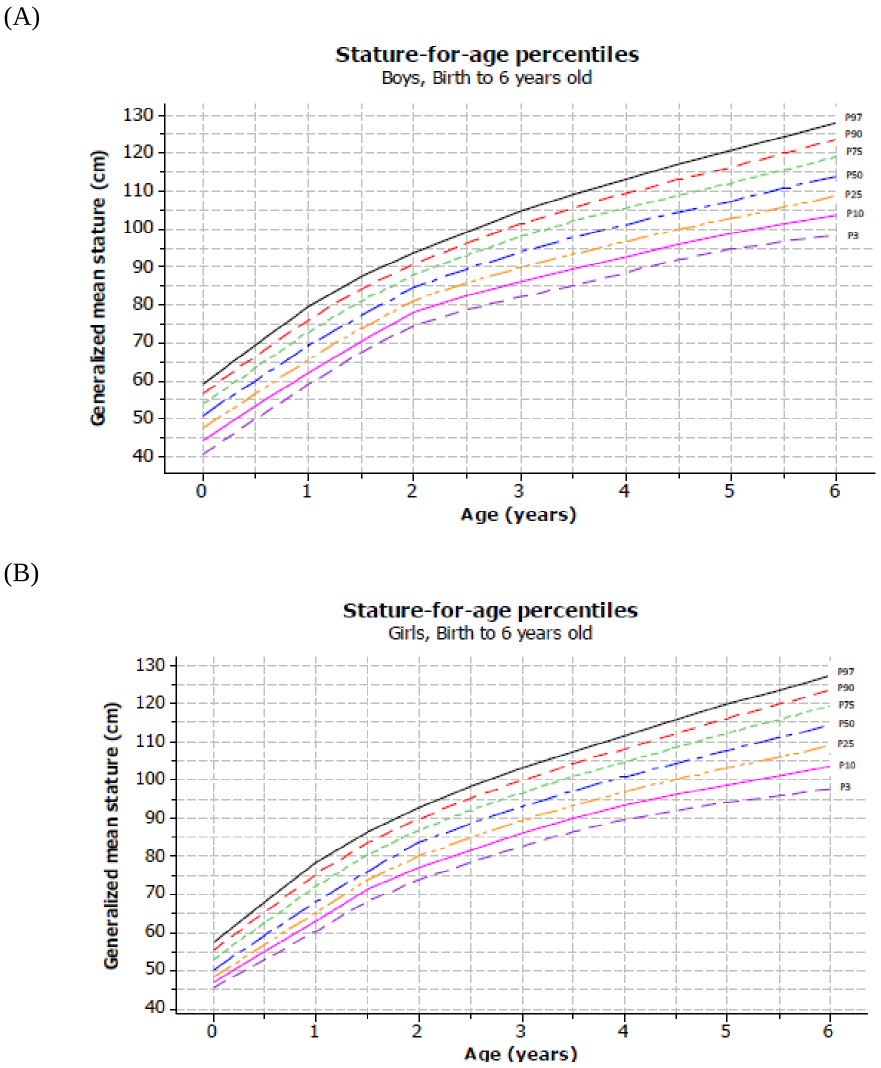


Figure 1: Length-for-age / Height-for-age smooth centile for Malaysian infants from birth to 6 years of age (A) Infant boys (B) Infant girls

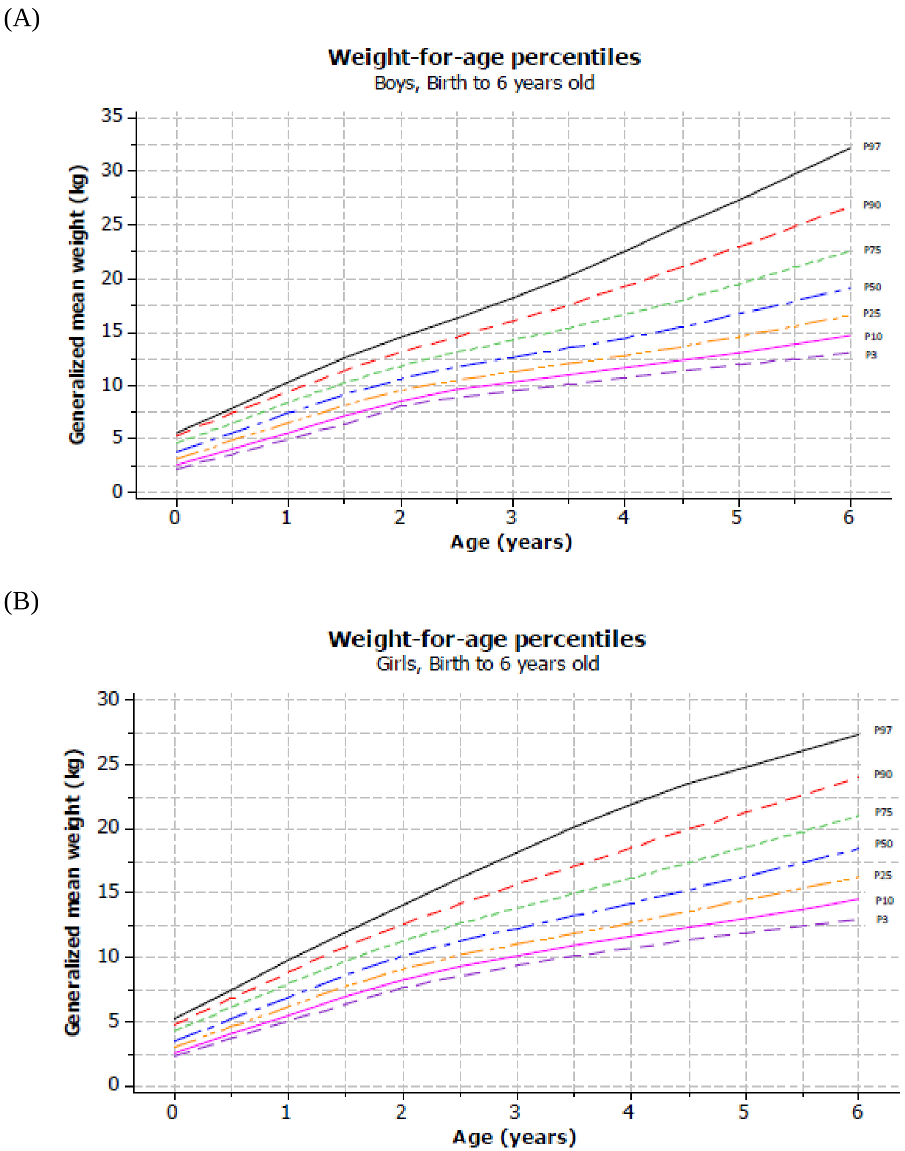


Figure 2: Weight-for-age smooth centile for Malaysian infants from birth to 6 years of age (A) Infant boys (B) Infant girls

4. DISCUSSIONS

The smooth centiles produced are based on data from both breastfed and formula-fed infants from various geographic and socioeconomic backgrounds. Infant boys were bigger in size as compared to infant girls. This finding correlates well with the results by Guaran et al. (1994), whereby they reported that male infants were larger than female infants.

The findings from this study showed that infant boys were taller than infant girls, except at age 4.5 and 5.5 years. Both genders showed high growth velocity in stature from birth to age 1.5 years old, which illustrates that growth in children is at its peak for the age range as mentioned.

Newborns demonstrated similar birth weight for both genders. The mean birth weight for newborns from this study was 3075g. The rate of increment in weight was higher from birth to six months of age. Despite this, the weight percentile dispersed as they grow older. This could be due to the food intake as they no longer depend on either breastfeeding or formulae milk-feeding.

5. CONCLUSIONS

The establishment of a local curve is very much warranted as some children might be considered thinner or shorter for no reason if the international measurements were used for assessment of growth. The proposed growth charts can be used as growth reference for growth monitoring as they were developed based on the mix-feeding practice (breastfed and formulae-fed) of the community presently.

Medical and paramedical staff will be able to use this reference curves for therapeutic purposes and as a screening tool while parents can refer the references for improving feeding and nutrition.

REFERENCES

1. Cole, T.J. (1990). The LMS method for constructing normalized growth standards. *Eur J Clin Nutr*, 44(1), 45-60.
2. Eveleth, P.B. and Tanner, J.M. (1976). *Worldwide variation in human growth*. London: Cambridge University Press.
3. Guaran, R.L., Wein, P., Sheedy, M., Walstab, J. and Beischer, N.A. (1994). Update of growth percentiles for infants born in an Australian population. *Aust. N.Z. J Obstet. Gynaecol.*, 34(1), 39-50.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 233-240

THREE APPROACHES FOR FORECASTING THE TRAFFIC ACCIDENTS AND TRAFFIC FATALITIES IN QATAR

Adil E. Yousif
Department of Mathematics, Statistics and Physics
Qatar University, Doha, Qatar
Email: aeyousif@qu.edu.qa

ABSTRACT

Statistics showed that each year about one million people are killed and fifty millions injured on roads around the world. In Qatar, the mortality rate caused by road accidents is between four to five deaths on a weekly basis. The main objective of this study is to estimate the road traffic accidents and fatalities in Qatar using three different approaches, time series models, Smeed's equation and Multilayer Perceptron from neural network.

This study also investigated the traffic accidents and traffic fatalities trend in Qatar and the effect of some predictors that may play major roles in the direction of the traffic accidents trend, in particular the economic factor. The data was obtained from three sources: Traffic Department – Qatar, Qatar Statistics Authority and Hamad Medical Corporation- Qatar, for the period between 1990 to 2011.

The number of fatalities in Qatar has been consistently rising over the past two decades. The correlation coefficient matrix as well as the figures show that there is a positive association between the economic growth and traffic accidents and traffic fatalities. In conclusion, the current study has shown that the neural network gave better and more reliable road traffic fatalities estimates followed by ARIMA and the Smeed's estimates are way high.

KEYWORDS

Multilayer Perceptron; Traffic Fatalities; ARIMA; Smeed; Qatar; Traffic accidents; Neural Network.

INTRODUCTION

Each year about one million people are killed and fifty millions injured on roads around the world. In Qatar, the mortality rate caused by road accidents is between four to five deaths on a weekly basis. This study aimed to investigate traffic accidents and traffic fatalities trend in Qatar and the effect of some predictors that may play major roles in the direction of the traffic accidents trend, in particular the economic factor.

Other objective of this study is to estimate the road traffic accidents and fatalities using time series models, Smeed's equation and Multilayer Perceptron from neural network.

Road traffic fatality rates of a country are known to depend upon factors such as the population, the number of motor vehicles in use, the total length of roads, the population density and the economic conditions (Bener and Crundall, 2005, Paulozzi et al. 2007). With over 1 million killed by car crashes annually, traffic injuries are projected to become the 3rd leading cause of disability adjusted life years lost by 2020 (Murray and Lopez, 1996). In general, the total costs of road accidents and fatalities are a burden for the country and cost over 2.8 % of the gross national product (Bener et al. 2003; Elvik 2000).

Qatar has experienced a rapid transition in its socio-economic status since after the discovery of oil. There has been a dramatic rise in the national economy expressed in terms of gross domestic product. The gross domestic product for the state of Qatar in the year 2010 was $ 463.492.

The data was obtained from three sources: Traffic Department Qatar, Qatar Statistics Authority and Hamad Medical Corporation, for the period between 1990 to 2011. There are several possible variables and risk factors can be thought of when analyzing traffic accident phenomenon. For fulfilling the research objectives, the following proposed variables were considered:

Accidents: Represents the numbers of traffic accidents in Qatar.
Fatalities: Represents the number of traffic fatalities in Qatar.
Population: Represents the number of population in Qatar.
Vehicles: Represents the number of vehicles registered in Qatar from.
GDP: Represents the value of gross domestic product indictor in Qatar.

METHODOLOGY

This study aimed to investigate traffic accidents and traffic fatalities trend in Qatar and the effect of some predictors that may play major roles in the direction of the trend. The main objective of this study is to estimate the road traffic accidents and fatalities using time series models, Smeed's equation and Multilayer Perceptron from neural network. Other objective is to examine the relationship between economic growth and traffic fatalities.

TIME SERIES

Several time series models relevant to the data were used such as, simple trend model, exponential smoothing model, and autoregressive integrated moving average model ARIMA. Since ARIMA requires a stationary set of data and the original data is violating this condition transformation was used. However the data is relatively small and difference transformation will shrink it more.

New approach of ratio transformation $Z_t = \frac{Y_t}{Y_{t-1}}\ , where,\ Y_{t-1} \neq 0$ was used. When using this transformation with first order autoregressive model

$$Z_1 = \delta + \emptyset_1 Z_{t-1} + a_t$$

It becomes:

$$\frac{Y_t}{Y_{t-1}} = \delta + \emptyset_1 \frac{Y_{t-1}}{Y_{t-2}} + a_t$$

Providing that, $Y_{t-1} \neq 0\ and\ Y_{t-2} \neq 0$ which eliminate the constant terms and the random shock term will be a multiple of the response variable. The final model will be

$$\boldsymbol{Y_t = Y_{t-1}(\delta + \emptyset_1 \frac{Y_{t-1}}{Y_{t-2}} + a_t)}$$

There could be a controversial issue about this ration since it is a division of two normal variables, however when the normality tested for this data it wasn't violated

SMEED's FORMULA

Smeed derived a formula that estimates the road traffic fatalities of a country by using the population and the number of registered vehicles of a particular country (Smeed, 1949; Smeed 1964). Smeed's formula is been used for many years in several countries including many European countries, USA., Canada, Australia and New Zealand and it is given by:

$$\frac{F}{P} = \alpha\left(\frac{V}{P}\right)^{\beta},$$

F is number of fatalities, P is the total population, V number of registered vehicles and α&β are parameters to be estimated

MULTILAYER PERCEPTRON

The multilayer perceptron (MLP) is a feed-forward, supervised learning network with up to two hidden layers. The MLP network is a function of one or more predictors (also called inputs or independent variables) that minimizes the prediction error of one or more target variables (also called outputs). Predictors and targets can be a mix of categorical and scale variables.

For Multilayer Perceptron the activation function used in this study for both Accidents and Fatalities is the hyperbolic Tangent function given by:

$$Y(\theta) = tanh(\theta) = \left(\frac{e^{\theta}-e^{-\theta}}{e^{\theta}+e^{-\theta}}\right)$$

RESULTS DISCUSSION

In Qatar, the vast majority of traffic accidents are of types compromised and hit-and-run. These two types form 97% to 98% of total accidents every year. This implies that traffic police in Qatar attend only 2% to 3% of the total traffic accident scenes. Throughout the report, we defined these 2% accidents as traffic accidents requiring police field intervention.

Figure [1] expresses the ratios of traffic accidents categories. Verifying the traffic accidents requiring police field intervention, 50%-70% of these traffic accidents are of property damage accident type. Minor injuries accidents range between 14% - 38%, while severe accidents is between 8% - 17%. It also appear that the severe accidents ratios decrease in the last six years. The fatal accidents did not exceed 6% of the total number of accidents requiring police field intervention.

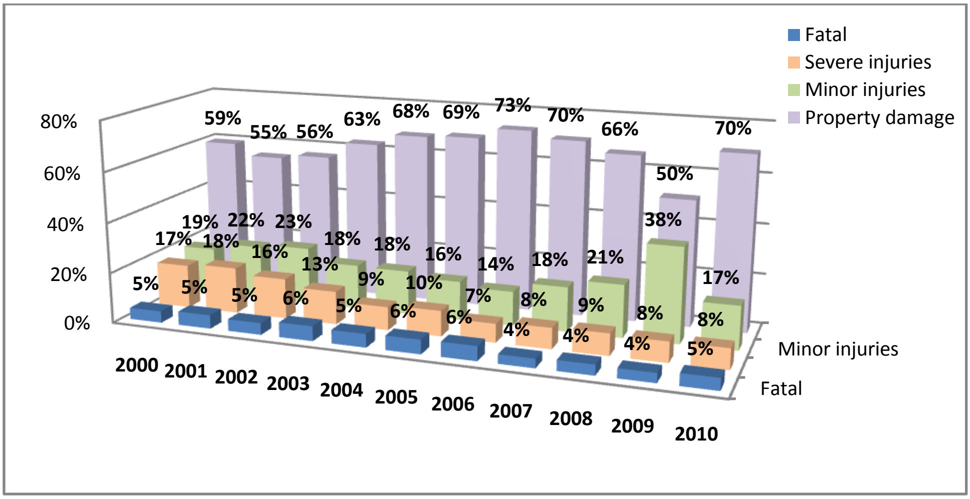


Figure 1

Figure [2] below represents the trends of the five main variables under investigation.

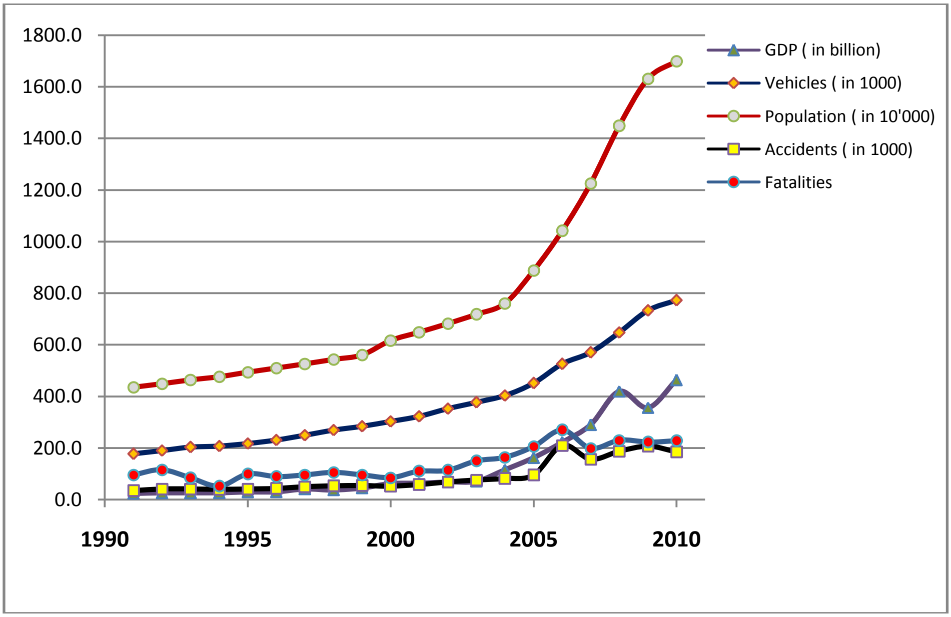


Figure 2

Analyzing the graphs in Figure [2] we found that the population in Qatar from 1991 to 2010, witnessed an increase into two major phases. The phase 1 was from 1991 to 2004, during which the general increase rate of population was about twenty-five thousands per year. For the period that followed 2004, the population rate shifted over 170 thousand per year and this is due to many projects launched in Qatar and required more manpower and expatriates from abroad. In general all the five variables indicated an increasing trend.

Pearson correlation coefficients indicated strong correlation between accidents and fatalities versus the other three variables: GDP, population, and vehicles. The p-value for each pair is less than 0.0001 (Appendix) which means the economic growth has a direct impact on the traffic accidents and fatalities.

TIME SERIES ANALYSIS

Three time series models were used to for data analysis and forecasting equations and they are trend model. Holt's Double Exponential Trend model, and Autoregressive Integrated Moving Average (ARIMA) model. The trend model that best fit the number of accidents data was the quadratic trend model

$$Y_t = 5037 - 6422t + 739t^2$$

Similarly for the fatalities the best trend model is the quadratic one:

$$Y_t = 92.7 - 3.93t + 0.621t^2$$

As traffic accidents time series as well as the fatalities exhibit an increasing trend, it is better to explore a double exponential smoothing model (Holt's corrected model). The lowest MAD for smoothing corrected model was obtained when Alpha and Gama are equal to 0.7 and 0.10003 respectively.

Since the autoregressive integrated moving average (ARIMA) models require stationary time series difference as well as ratio transformation were used to form a stationary set of data. By examining the autocorrelation function (ACF) for the number of accidents, it cuts off at lag 1. On other hand, the partial autocorrelation function (PACF) it dies down.

Therefore for the number of accidents, the more adequate ARIMA model based on ACF and PACF in addition to Ljung statistics is as following:

$$\mathrm{Z_t} = 1.0001\mathrm{Z_{t-1}} + \mathrm{a_t}$$

As shows in the following table, the standard error estimate is 32140.5 which is not the best comparing to the previous models. In addition, in Table 6, below, we can notice that the p-values increase as lag increases which means that we cannot reject the accuracy of the model.

For the number of fatalities both ACF and PACF cuts of after lag one and which yields in ARIMA (1,1) and the fitted equation is given by:

$$\mathrm{Z_t} = 97.77457 + 0.98832 * \mathrm{Z_{t-1}} + \mathrm{a_t}$$

SMEED's FORMULA

When Smeed's formula was used for the data of this study the estimate of the parameters are ($\hat{\alpha} = 0.000256, \hat{\beta} = 0.482$) which gives an estimate for the number of fatalities as follows

$$\frac{F}{P} = 0.000256\left(\frac{V}{P}\right)^{0.482}$$

NEURAL NETWORK

For neural network approach a Multilayer Perceptron technique used and the activation function used for both Accidents and Fatalities is the hyperbolic tangent function given by:

$$Y(\theta) = tanh(\theta) = \left(\frac{e^{\theta} - e^{-\theta}}{e^{\theta} + e^{-\theta}}\right)$$

Network Information for the two variables (accidents and fatalities) are as following: Input Layer contain three covariates GDP, Vehicles, and Population, number of units is three. The rescaling method for covariates is standardized one, and only one hidden layers is used with one unit in it. For the output layer with one unit and standardized rescaling method for scale dependents and activation function the identity. The error function is the sum of squares .

The following tables display the mean square error for the five models versus the number of accidents and number of fatalities.

Fatalities Mean Square Errors for Six Models:

	Trend Model	**Holt's Model**	**ARIMA**	**Smeed's Formula**	**MLP**
Accidents	629173	608083	32.64792	865434	0.151
Fatalities	664.1	19.5696	3.884185	1110.5	0.172

Forecast for the number of fatalities:

Model	**Forecast**				
	2011	**2012**	**2013**	**2014**	**2015**
Trend model	236	245	254	263	272
Holt's Model	240	248	257	266	275
ARIMA	226	224	223	222	220
Smeeds	282	299	317	333	345
MLP	225	228	231	235	236

CONCLUSION

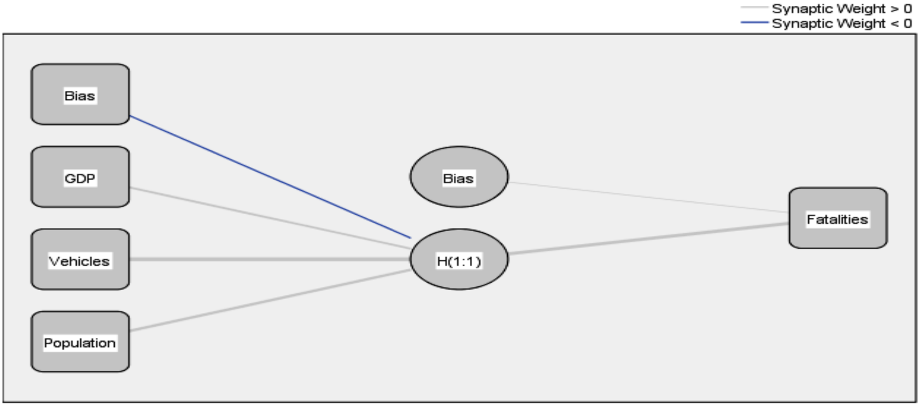
The fatalities estimate has been consistently rising over the past two decades. Although economic growth believed to help in reducing road fatalities however, the correlation coefficient matrix as well as the figures show that there is a positive association between the economic growth and traffic accidents and traffic fatalities. On the other hand comparing the three forecasting techniques, the current study has shown that the neural network gives a better and more reliable road traffic fatalities estimates followed by ARIMA whereas the Smeed's estimates are way high.

REFERENCES

1. Bener, A., Hussain, S.J., Al-Malki, M.A., Shotar, M.M., Al-Said, M.F. and Jadaan, K.S. (2010). Road traffic fatalities in Qatar, Jordan and the UAE: estimates using regression analysis and the relationship with economic growth. *Eastern Mediterranean Health Journal*.
2. Adil Yousif and Abdulbari Bener (2007). Road Traffic Fatalities and Economic Growth. Proceeding of the *9th Islamic Countries Conference on Statistical Sciences*, Malaysia.
3. Al-Ghamdi, A.S. (1996). *Road Accidents in Saudi Arabia: a Comparative and Analytical Study*. Barcelona, Spain.
4. Ali Hassan Al Marzooqi, Mohamed Badi and Aizeldin El Jack (2010). *Road Traffic Accidents in Dubai*, 2002-2008.
5. Al-Madani, D.H. (2009). *International experiences in traffic planning* (translated, Arabic). Algeria: Center for Studies and Research Department of seminars and scientific meetings.
6. Authority, Q.S. (2011). *Population structure*. Retrieved from WELCOME TO Qatar Statistics Authority WEBSITE: http://www.qsa.gov.qa/eng/PopulationStructure.htm
7. Bong-Min Yang and Jinhyun Kim (2003). *Road traffic accidents and policy interventions in Korea*. Swets & Zeitlinger.
8. Bruce L. Bowerman, Richard T.O, Connell, and Anne B. Koehler (2005). *Forecasting, Time Series, and Regression*. US.
9. Committee, P.P. (2010). *Qatar Population Situation 2010*. Doha.
10. DC, E.O. (2010). *Fifteenth meeting of the Conference of the Parties to the Convention on International Trade in Endangered Species of Wild Fauna & Flora*.
11. Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. Switzerland: Springer Science and Business Media, LLC.
12. Peter Ljubič, Ljupčo Todorovski, Nada Lavrač. (2002). *Time-Series Analysis of UK Traffic Accident Data*.
13. http://www.qatartourism.gov.qa/discover/index/1/175
14. Qatar Tourism Authority (n.d.). Economy. Retrieved from Discover Qatar: http://www.qatartourism.gov.qa/discover/index/1/211
15. Zaid, B.H. (2009). The roles of traffic organizing in road safety (translated, Arabic). Algeria: Center for Studies and Research Department of seminars and scientific meetings.

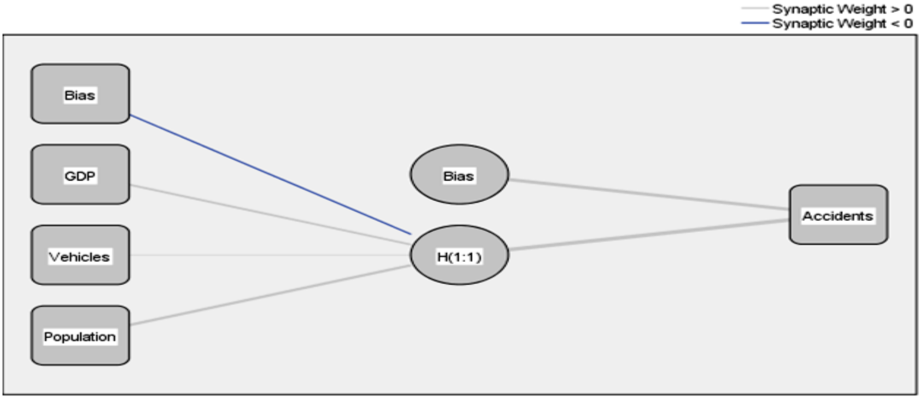
APPENDIX

	Accidents	Fatalities	GDP	Vehicles	Population
Accidents	1	0.93694	0.92689	0.94239	0.93677
		<.0001	<.0001	<.0001	<.0001
Fatalities	0.93694	1	0.85757	0.88921	0.85956
	<.0001		<.0001	<.0001	<.0001
GDP	0.92689	0.85757	1	0.96783	0.98525
	<.0001	<.0001		<.0001	<.0001
Vehicles	0.94239	0.88921	0.96783	1	0.99049
	<.0001	<.0001	<.0001		<.0001
Population	0.93677	0.85956	0.98525	0.99049	1
	<.0001	<.0001	<.0001	<.0001	



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity



Hidden layer activation function: Hyperbolic tangent

Output layer activation function: Identity

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 241-256

SPATIAL RELATIONSHIP BETWEEN ILLITERACY AND UNEMPLOYMENT AT GOVERNORATE LEVEL IN EGYPT-2006

Faisal G. Khamis
Faculty of Economics and Administrative Sciences,
AL-Zaytoonah University of Jordan, Jordan
Email: faisal_alshamari@yahoo.com

ABSTRACT

Most but not all studies found spatial clustering in each of illiteracy rate (IR) and unemployment rate (UR) in developing and developed countries. Question is raised whether the spatial pattern of each of IR and UR are existed in Egypt? If so, are the spatial patterns of IR and UR spatially correlated at governorate level?

The objective is to investigate the spatial structure of IR and its spatial correlation with spatial structure of UR. The present study utilizes a cross-sectional census data conducted in 2006 for 27 governorates. Mapping is used as the first step to conduct visual inspection for IR and UR. Several spatial econometric techniques are available in the literature, which deal with the spatial autocorrelation in geographically referenced data. Two statistics of spatial autocorrelation, based on sharing boundary neighbours, known as global and local Moran's *I*, are carried out. Wartenberg's measure is used to detect bivariate spatial correlation between IR and UR based on simulation study.

The hypothesis of spatial clustering for IR was not confirmed by a positive global Moran's *I* of .03. While for UR was confirmed by a positive global Moran's *I* of .25 with $p = .008$. Bivariate spatial correlation between IR and UR was found -.19 and $p = .978$.

In conclusion, based on visual inspection of mapping global clustering was found for IR in western-northern and middle parts and for UR in eastern-northern and southern parts. Global clustering was found for UR but was not found for IR. Out of 27 governorates, two were found as local clusters in each of IR and UR. The bivariate spatial correlation between IR and UR was not found significant.

KEYWORDS

Spatial autocorrelation, unemployment, illiteracy, mapping, global and local Moran statistics, governorates of Egypt.

1. INTRODUCTION

Regions, independent of their geographic level of aggregation, are known to be interrelated partly due to their relative locations. Similar economic performance among regions can be attributed to proximity. Consequently, a proper understanding and accounting of spatial liaisons are needed in order to effectively forecast regional economic variables. Spatial statistics relates to the analysis of the spatial aspect of data sets. All data have, implicitly or explicitly, a spatial component. The reasons why spatial statistics are of import for many areas are threefold: (i) Data that are spatially close, are usually more similar than those that are further apart. Hence, there is spatial and temporal dependence for most data sets. (ii) Spatial models to explain and make inference about data structures have important implications in many fields. (iii) The literature on spatial statistics is substantial, but there is still a lot to uncover and many questions to answer. This study showed visual picture for each of IR and UR, investigated the global and local clusters, examined the spatial relationship between UR and IR, and provide simplications for policy makers.

In recent years, a growing interest has been seen in examining the existence of spatial autocorrelation for unemployment rate (UR) and its spatial relationship to several indicators such as illiteracy rate (IR), level of education, etc. across geographic areas in developing countries. Historically, UR was slightly increased in Egypt from 10.6% in 2006 compared to 9.7% in 1996. While IR was dramatically decreased to 29.64% in 2006 compared to 39.36% in 1996.

The spatial clustering problem of UR and IR exists in several developing and even developed countries. It is very difficult to control and to find potential solutions for this problem. Due to the inequality of UR and IR is affected by several factors. For example, people moves continuously from such governorate to other, the biasedness of the distribution of economic resources across geographical locations, the job and education opportunities are much differed across geographical locations, etc. We hope to achieve some improvement in reducing the inequality in UR and IR in Egypt. Also, shows policy makers the location of clusters problem in UR and IR.

UR and IR were studied in several countries using different statistical measures. Based on spatial regression analysis, Hooghe et al. (2011) demonstrated that unemployment figures have a strong and significant impact on crime rates in Belgium. According to the study in Jordan by Amerah, (1993), health was affected negatively by unemployment. Amerah stated that since the mid-1980s unemployment had become a serious problem in Jordan, manifesting a widening gap between the demand for and supply of labour. Lack of well-paying jobs, little education and illiteracy are all associated with poverty (Arab Republic of Egypt, 2007). Elhorst (2003) proposed several reasons that make the study of spatially uneven distribution of unemployment worthwhile. One of these reasons is the wide unemployment differentials imply inefficiency in the economy as a whole and reduces growth. Jin et al. (1995) found a strong positive association in Canada between unemployment and many adverse health outcomes. El-Gamal (2012) stated that illiteracy has been a major issue for the government as it prevents several millions of Egyptian citizens from contributing effectively to the economy of the country. The negative impact of illiteracy on individuals and society makes it a key obstacle that threatens the efforts towards achieving integral and comprehensive development. Osman, Zakareya, and Mahrous (2006) emphasized the importance of education for the alleviation of poverty and the adversities that are particularly associated with illiteracy on poverty in Egypt. Some 68 million people in Arab countries are illiterate, an alarming proportion that threatens development, an Arab League culture official said (New York Times, 2002).In their study in Egypt, Wahbaand Zenou (2005) found that the predicted probability to find a job can decrease for the illiterate and the less educated workers. Unemployment and illiteracy are a major concern in Egypt. In 2006, they were 10.6% and 29.64% respectively. The report conducted by United Nations Development Programme and the Institute of National Planning (2003) stated that the relatively high UR among females was influenced by their higher rates of illiteracy. Unemployed and/or illiterate persons have an increased risk of death. These persons are facing financial problems regarding the quality of living conditions.

To understand the linkages between socioeconomic variables, investigation should focus on features of the areas rather than on the compositional characteristics of residents of the area, which cannot fully describe the social environment in which people live (Macintyre, Maciver & Sooman, 1993). So, the aim of the research is to study geographical mapping and spatial autocorrelation regarding UR and its spatial relationship to IR. Spatial autocorrelation is the term used for the interdependence of the lattice data over space. It is argued that lattice data are spatially correlated. We used exploratory spatial data analysis (ESDA) using lattice data. The ESDA quantifies the spatial pattern in order to increase the analyst's knowledge of the spatial system. As well as mapping plays an important role in monitoring disadvantaged people. Maps can reveal spatial patterns that is neither recognized previously nor suspected from the examination of statistics table. It reveals high risk communities or problem areas (Lawson & Williams, 2001). The purpose of spatial analysis is to identify pattern in geographical data and attempting to explain this pattern. Findings are expected to enhance monitoring disadvantaged people and policy interventions across governorates of Egypt. Findings also enable the decision maker to reoriented sources towards sectors and governorates suffering from unemployment and illiteracy.

In this research some hypotheses were tested. Are UR and IR indicators having global clustering across Egypt’s governorates? Are there local clusters? If so, how many clusters are there and where are they located? Then, is clustering in IR can be associated with clustering in UR? Governorates are tightly linked by migration, commuting, and inter-governorate trade. These types of spatial interaction are exposed to the frictional effects of distance, possibly causing spatial dependence of governorate labour market conditions. Some of socioeconomic indicators such as education basically offer greater chance of getting a job. The parents viewed poverty as illiteracy and the inability to earn better incomes (UNICEF, 2010).

Research relevance stems from a statement states that reducing UR and IR inequality is not a primary objective but emergent prosperity. The importance of research objective emanates that health was affected negatively by UR and IR. The report of UNICEF (2010) stated that factors that increase vulnerability to HIV include a rise in mobility, the high IR especially among women, poverty, and UR. It is very necessary for policy makers to know in which area the problem of UR and IR inequality is existed? Also, to authors' knowledge, no studies used spatial analysis techniques and geographical mapping in studying the inequality in UR and IR in Egypt.

Importance of mapping was stated by Koch (2005): why make the map if detailed statistical tables carry the same results? Perhaps the most important reason for studying spatial statistics is not only interested in answering the "how much" question, but the "how much is where" question (Schabenberger & Gotway, 2005). In light of these: (1) the existence of spatial global clustering, (2) spatial local clusters for each of UR and IR were investigated, (3) mapping was applied for each of UR and IR and for their local Moran’s I_i values, and (4) bivariate spatial correlation between UR and IR was examined based on Wartenberg's (1985) measure. The study design was a cross-sectional analysis in a census survey conducted in Egypt in 2006. Findings make a significant contribution by moving beyond the investigation of a single socioeconomic resource. However, findings push us to more fully consider where and why UR and IR matter.

The paper was structured as follows: Section one review the literature of UR and IR inequality generally in several developing countries and particularly in Egypt. Materials and methods including data details and statistical analysis are presented in second Section. Third section shows the results with some details. Discussion is explained in fourth Section. Final section is closed with several conclusions.

2. MATERIALS AND METHODS

2.1 Data

Data were collected from the book of Egypt's description by information (2009), based on census conducted in Egypt in 2006. For each of 27 governorates, UR and IR were studied. UR is an average number of unemployed individuals at the age category (15-64) years who are capable to and looking for work but unable to find a job. It is estimated as a percentage of labor force. UR varied geographically in 2006. In urban, lower, upper and frontier governorates were 12.1%, 11.0%, 9.4%, and 10.0% respectively. UR fell from round 12% of the labour force in 1998 to 8% in 2006. IR is defined as the percentage of persons who cannot read or write (10 years old or more, not in school). In Egypt, the IR was drastically decreased from 39.36% in 1996 to 29.64% in 2006.

2.2 Analysis

Data analysis involved six steps. In step 1, UR and IR were tested for normal distribution. They were found to follow approximately normal distribution. In step 2, visual inspection based on the quantified gradients for each of UR and IR using quartiles were conducted. Step 3 included the calculation of global Moran's I -statistic for each of UR and IRto detect the global clustering. The significance of I -statistic using permutation test was examined. Step 4 involved the calculation of local Moran's I_i for *ith* governorate and it's p-value using Monte Carlo simulation to detect the local clusters for each of URand IR. In step 5, using quartiles, visual inspection for the gradients of local Moran values was inspected based on choropleth mapping. In Step 6, the bivariate spatial correlation between UR and IR was examined based on Wartenberg's (1985) measure.

The UR and IR were categorized to four intervals. These intervals were used for all maps using darker shades of gray to indicate increasing values. Such approach enables qualitative evaluation of spatial pattern. In the neighbourhood researches, neighbours may be defined as governorates which border each other or within a certain distance of each other. In this research neighbouring structure was defined as governorates which share a boundary. The *second order* method (queen pattern) which included both the first-order neighbours (rook pattern) and those diagonally linked (bishop pattern) was used. A neighbourhood system of Egypt's governorates is explained in Figure 1, where ID neighbours for each governorate are shown.

ID	Governorate	ID neighbours
1	Dakahleyia	2,3,4,6,10,11
2	Gharbeyia	1,3,4,11,12
3	Menofya	2,4,12,15
4	Qalyubiya	1,2,3,5,6,15
5	Cairo	4,6,7,15,16
6	Sharkeya	1,4,5,7
7	Ismailia	5,6,8,9,16
8	North Sinai	8,16
9	Port Said	7,8
10	Dumyat	1
11	Kafr ash Shaykh	1,2,12
12	El Buhayrah	2,3,11,13,14,15
13	Alexandria	12,14
14	Matruh	12,13,15,27
15	Giza	3,4,5,14,16,17,18,19,27
16	Suez	5,7,8,15,20,21
17	El Fayyum	15,18,19
18	BaniSwaif	15,17,19,21
19	Al Minya	15,18,21,22,27
20	Southern Sinai	8,16
21	Red Sea	15,16,18,19,22,23,24,26
22	Asyut	19,21,23,27
23	Sohaj	21,22,24,27
24	Qina	21,25,26,27
25	Luxor	24,26,27
26	Aswan	21,24,27
27	New Valley	14,15,19,22,23,24,25,26

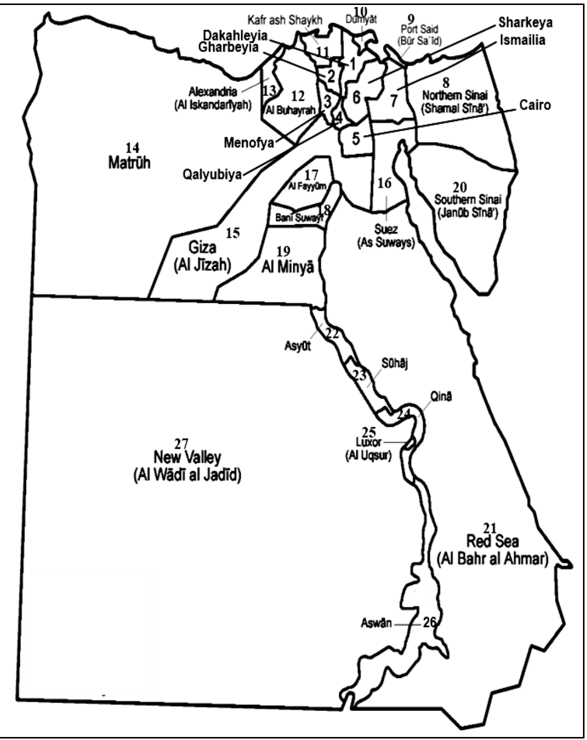


Figure 1: Study area shows all governorates with its ID and ID neighbours of each governorate

To construct a choropleth map, data for enumeration governorates are typically grouped into classes and a gray tone was assigned to each class. Although maps allow visual assessment for spatial pattern, they have two important limitations: their interpretation varies from person to person, and there is the possibility that a perceived pattern is actually the result of randomness, and thus not meaningful. For these reasons, it makes sense to compute a numerical measure of spatial pattern, which can be accomplished using spatial autocorrelation.

2.1.1. Identification of global spatial clustering:

The goal of a global index of spatial autocorrelation is to summarize the degree to which similar observations tend to occur near to each other in geographic space. The spatial autocorrelation using standard normal deviate (z-statistic) of Moran's I under normal assumption was tested. Moran's *I* is a coefficient used to measure the strength of spatial autocorrelation in regional data. The null hypothesis of no spatial autocorrelation or spatially independent versus the alternative of positive spatial autocorrelation is as follows:

H_0 : No clustering exists (no spatial autocorrelation)

H_1 : Clustering exists (positive spatial autocorrelation)

Moran's *I* is calculated as follows (Cliff and Ord, 1981):

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{S_0\sum_{i=1}^{N}(x_i-\bar{x})^2} \text{ and } S_0 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}, i \neq j$$

where, $N = 27$ is the number of governorates, the $w_{ij} = 1$ is a weight denoting the strength of the connection between two governorates i and j, otherwise, $w_{ij} = zero$, and the x_i and x_j represents UR or IR in the ith and jth governorate respectively.

A significant positive value of Moran's I indicates positive spatial autocorrelation, showing the overall pattern for the governorates having a high/low level of UR or IR similar to their neighbouring governorates. A significant negative value indicates negative spatial autocorrelation, showing the governorates having a high/low level of UR or IR unlike neighbouring governorates. To test the significance of global Moran's I , z -statistic that follows a standard normal distribution was applied. It is calculated as follows (Weeks, 1992):

$$z = \frac{I - E(I)}{\sqrt{\text{var}(I)}}$$

Permutation test was applied. A permutation test tells us that a certain pattern in data is or is not likely to have arisen by chance. The observations of each variable was randomly reallocated 1000 times with 1000 of spatial autocorrelations were calculated in each time to test the null hypothesis of randomness. The hypothesis under investigation suggests that there will be a tendency for a certain type of spatial pattern to appear in data, whereas the null hypothesis says that if this pattern is present, then this is a pure chance effect of observations in a random order. The analysis suggests an evidence of clustering if the result of the global test is found significant; though it does not identify the location of any particular clusters. Besides, the clustering that represents global characteristic of each of UR and IR, the existence and location of localized spatial UR and IR clusters are of interest in geographic sociology. Accordingly, local spatial statistic was advocated for identifying and assessing potential clusters.

2.1.2. Identification of local spatial clusters:

A global index can suggest *clustering* but cannot identify individual *clusters* (Waller and Gotway, 2004). Anselin (1995) proposed the local Moran's I_i statistic to test the local autocorrelation. Local spatial clusters, sometimes referred to as hot spots, may be identified as those locations or sets of contiguous locations for which the local Moran's I_i is significant. However, Moran's I_i for ith governorate may be defined by Waller and Gotway (2004) as:

$$I_i = \frac{(x_i-\bar{x})}{S}\sum_{j=1}^{N}\left(w_{ij}\Big/\sum_{j=1}^{N_i} w_{ij}\right)\frac{(x_j-\bar{x})}{S},\ i = 1, 2, ..., 27$$

where, analogous to the global Moran's I, x_i and x_j represent the UR or IR in the ith and jth governorate respectively, N_i = number of neighbours for the ith governorate, and S is the standard deviation. It is noteworthy to mention that the number of neighbours for the ith governorate were taken into account by the amount: $\left(w_{ij} \Big/ \sum_{j=1}^{N_i} w_{ij} \right)$, where w_{ij} was measured in the same manner as in Moran's I statistic. Local Moran statistic was used to test the null hypothesis of *no clusters*.

Cluster could be due to either aggregation of high values, aggregation of low values, or aggregation of moderate values. Thereby, high values of I_i suggesting a cluster of similar (but not necessarily large) values across several governorates, and low value of I_i suggesting an outlying UR or IR cluster in a single governorate i (being different from most or all of its neighbours). A positive local Moran value indicates local stability, such as governorate that has high/low UR or IR surrounded by governorates that has high/low UR or IR. A negative local Moran value indicates local instability, such as governorate having low UR or IR surrounded by governorates having high UR or IR or vice versa. Each governorate's I_i value was mapped to provide insight into the location of governorates with comparatively high or low local association with their neighbouring values.

2.2.3. Bivariate spatial association:

So far, only univariate spatial correlation is presented. It quantifies the spatial structure of one variable at a time. There is much discussion about what is an appropriate measure for bivariate spatial association. However, spatial dependence or spatial clustering causes losing in the information that each observation carries. When N observations are made on a variable that is spatially dependent and that dependence is positive so that nearby values tend to be similar, the amount of information carried by the sample is less than the amount of information that would be carried, if the N observations are independent. Due to a certain amount of information carried by each observation is duplicated by other observations in the cluster. A general consequence of this is that the sampling variance of statistics is underestimated. As the level of spatial dependence increases, the underestimation increases. Spatial autocorrelation coefficient can be modified to estimate the bivariate spatial correlation between two variables (Wartenberg, 1985):

$$I_{xy} = \frac{1}{S_0} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \overline{x})(y_j - \overline{y})}{\left[\sqrt{\sum_{i=1}^{N} (x_i - \overline{x})^2 \Big/ N} \right] \left[\sqrt{\sum_{j=1}^{N} (y_j - \overline{y})^2 \Big/ N} \right]}$$

where x and y represents UR and IR variables respectively. Although the mathematics is quite straightforward, very few software packages offer the option of computing, I_{xy}. Thus, programming was used to find I_{xy}. To test the significance of I_{xy}, z -statistic was applied: $z = I_{xy} \sqrt{N-1}$, which follows approximately standard normal distribution.

2.2.4. Simulation study:

Simulated data are useful for validating the results of spatial analysis. Using Monte Carlo simulation, 9 999 random samples (27 values for each sample) were simulated. The process of simulation was conducted under standard normal distribution to calculate p-value for local Moran value. When the word simulation is used, it is referred to an analytical method meant to imitate a real-life system, especially when other analyses are too mathematically complex or too difficult to reproduce.

In Monte Carlo testing, test statistic is calculated based on the data observed. Then the same statistic is calculated for a large number (say, *Nsimu*) of data sets, simulated independently under the null hypothesis of interest (e.g., simulated under complete spatial randomness). The proportion of test statistic values based on simulated data exceeding the value of a test statistic observed for the actual data set provides a Monte Carlo estimate of the upper-tail *p*-value for one sided hypothesis test (Waller & Gotway 2004).

Specifically, suppose that $I_{i(obs)}$ denotes the test statistic for the data observed and $I_{i(1)} \geq I_{i(2)} \geq \cdots \geq I_{i(Nsimu)}$ denote the test statistic values (ordered from largest to smallest) for the simulated data set. If $I_{i(1)} \geq I_{i(2)} \geq \cdots \geq I_{i(l)} \geq I_{i(obs)} > I_{i(l+1)}$ (i.e., only the l largest test statistic values based on simulated data exceed $I_{i(obs)}$), the estimated $p-value$ is given as follows:

$$p-\text{value} = \widehat{\Pr}(I_i \geq I_{i(obs)} \,|\mathrm{H}_0 \text{ is true}) = \frac{l}{Nsimu+1}, i = 1, 2, ..., N \ ,$$

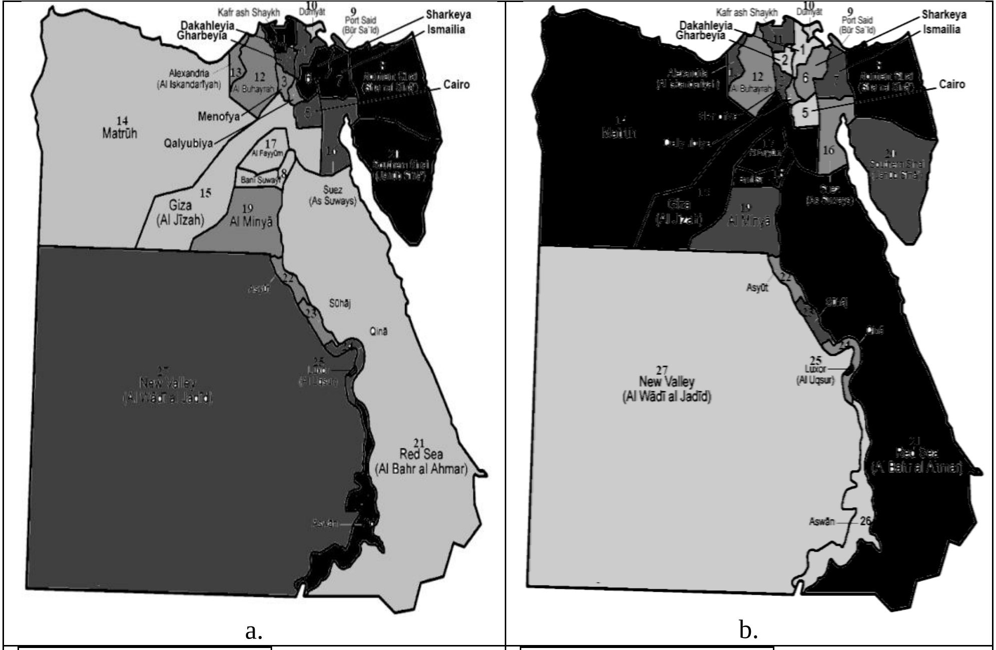
where one is added to the denominator since the estimate is based on ($Nsimu+1$) values of $(\{I_{i(1)}, \ldots, I_{i(Nsimu)}, I_{i(obs)}\})$. While results were specific to these data, the case study helps identify general concepts for future study. In the statistical analysis, all programs performed in SPLUS8 Software.

3. RESULTS

Descriptive statistics were calculated for each of UR and IR. The mean for UR and IR were found 10.89 and 27.55 respectively. The standard deviation for UR and IR were found 3.94 and 8.91 respectively. Skewness for UR and IR were found .84 and -.01 respectively. Kurtosis for UR and IR were found 2.37 and -1.05 respectively. The five-number summary of UR data set consisted of minimum, maximum and quartiles written in increasing order: Min=4.10, Q1=8.70, Q2=11.00, Q3=13.50 and Max=23.00. From the five-number summary, the variations of the four quarters were found 4.6, 2.3, 2.5 and 9.5 respectively, where the 4th. Quarter has the greatest variation of all. The five-number summary of IR data set was found: Min=11.63, Q1=19.47, Q2=27.44, Q3=35.08 and Max=41.29. From the five-number summary, the variations of the four quarters were found 7.84, 7.97, 7.64 and 6.21 respectively.

The values of spatial measures indicate how much and to what extent global clustering in studied variables is existed, how many local clusters are there, and where are they located. Figures 2a and 2b show visual insight for UR and its local Moran values respectively. Figures 3a and 3b show visual insight for IR and its local Moran values respectively. Darkest shade corresponds to the highest quartile. These maps display geographical variation across governorates of Egypt. Based on visual inspection of UR and IR taken from Figure 2a and Figure 3a respectively, an overall worsening pattern (higher scores) was found in the eastern-northern and southern parts and in the western-northern and middle parts respectively. The suggestion of spatial clustering of UR that follows a visual inspection of mapping was confirmed by a positive significant global Moran's I of .25 with an associated $z-$statistic of 2.63, $p = .008$, and permutation p =.005. Accordingly, the null hypothesis of no spatial autocorrelation was rejected. While, The suggestion of spatial clustering of IR that follows a visual inspection of mapping was not confirmed by a positive global Moran's I of .03 with an associated z-statistic of .59, $p = .558$ and permutation p=.252. This finding was not expected for IR. Accordingly, the null hypothesis of no spatial autocorrelation was not rejected.

Two significant clusters in each of UR and IR were detected. Clusters of UR were located in the middle (Bani Swaif) and in the southern (Luxor) as shown in Figure 2b. Clusters of IR were located in the eastern-northern (Port Said) and in the middle (Bani Swaif) as shown in Figure 3b. Table 1 shows the name of governorate, UR, IR, local Moran's I_i for each of UR and IR and its corresponding p-value based on simulation study under standard normal distribution. The p-value in boldface was considered significant at .05 level.



Quartiles	
░	(4.10) – (8.70)
▒	(8.70) – (11.00)
▓	(11.00) – (13.50)
█	(13.50) – (23.00)

Quartiles	
░	(-0.68)– (0.01)
▒	(0.01) – (0.12)
▓	(0.12) – (0.33)
█	(0.33) – (1.94)

Figure 2: Choropleth maps show: a. UR variable and b. its local Moran values

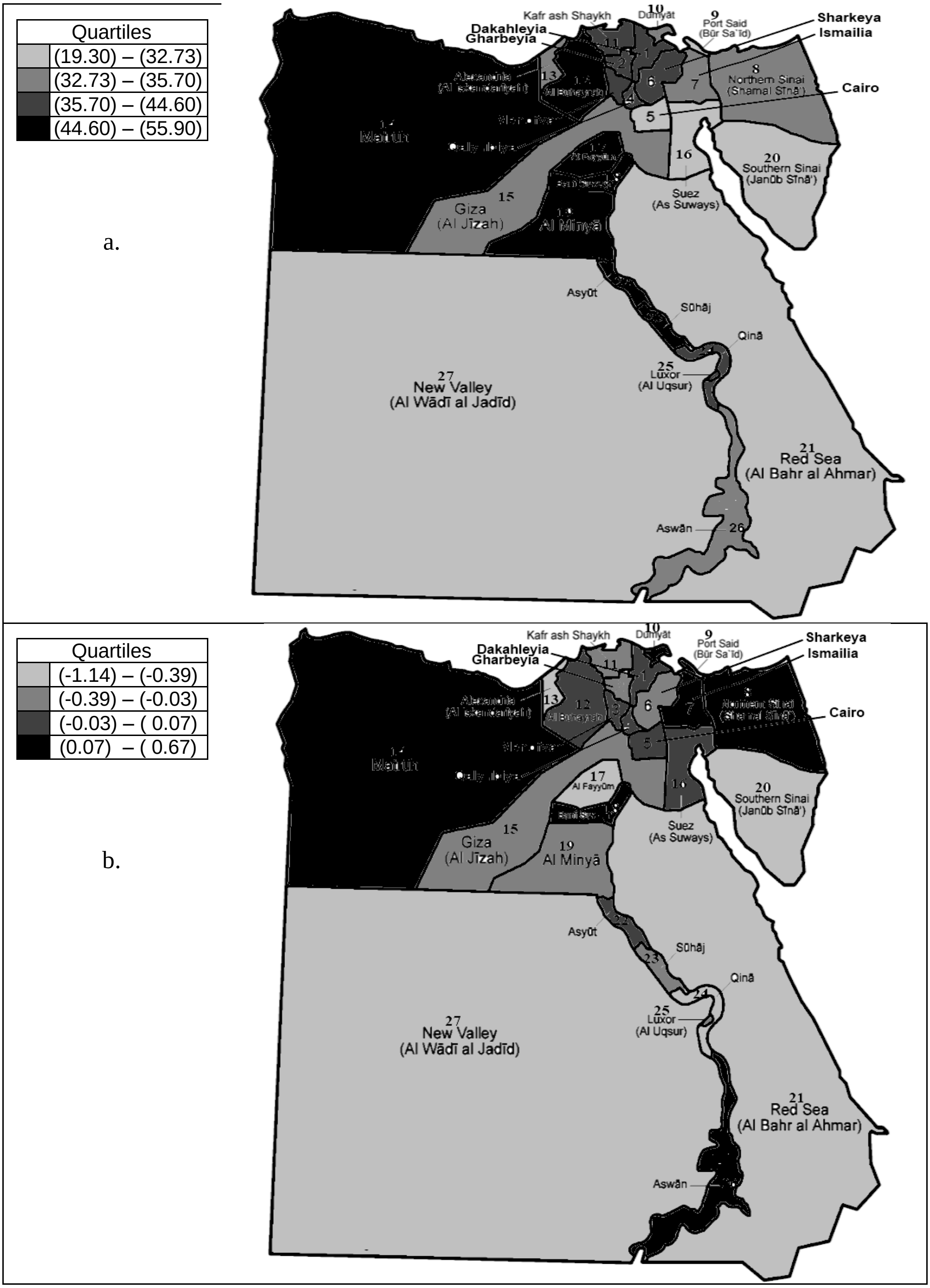


Figure 3: Choropleth maps show: a. IR variable and b. its local Moran values

Table 1
Explains governorates, UR,IR, local Moran's I_i for each of UR and IR and its corresponding p-value (p-value in boldface was considered significant at .05 level)

ID	Governorate	UR	I_i	p	IR	I_i	p
1	Dakahleyia	12.00	-.02	.540	27.91	.00	.443
2	Gharbeyia	11.00	.00	.477	25.85	-.07	.617
3	Menofya	8.70	.21	.200	27.44	.00	.492
4	Qalyubiya	8.90	.01	.417	27.52	.00	.467
5	Cairo	11.00	.00	.460	19.10	.23	.159
6	Sharkeya	13.70	.11	.293	32.16	-.19	.740
7	Ismailia	14.00	.29	.130	22.83	.34	.106
8	North Sinai	14.10	.33	.204	24.22	.33	.204
9	Port Said	11.00	.02	.451	16.39	1.85	**.008**
10	Dumyat	7.50	-.68	.830	22.42	.31	.257
11	Kafr ash Shaykh	13.50	.31	.212	34.31	-.07	.586
12	El Buhayrah	9.30	.03	.513	36.66	.08	.423
13	Alexandria	10.20	.18	.287	19.47	-.36	.795
14	Matruh	5.60	.78	.130	35.08	.42	.243
15	Giza	8.30	.38	.101	27.11	.00	.602
16	Suez	11.80	.04	.349	17.14	.28	.120
17	El Fayyum	7.10	.40	.133	40.89	-1.75	.992
18	BaniSwaif	4.20	1.52	**.025**	40.54	1.62	**.020**
19	Al Minya	9.90	.30	.270	41.29	-.12	.801
20	Southern Sinai	15.40	.12	.327	11.63	-.21	.710
21	Red Sea	4.10	.41	.068	12.69	-.58	.944
22	Asyut	10.50	.01	.437	39.06	.06	.351
23	Sohaj	9.10	.22	.346	38.50	-.06	.787
24	Qina	11.70	.10	.470	34.77	-.38	.965
25	Luxor	23.00	1.94	**.002**	27.80	-.01	.496
26	Aswan	16.30	-.02	.737	23.00	.55	.213
27	New Valley	12.10	-.08	.772	18.17	-.71	.993

The UR was not found associated with IR which is not expected. This is consistent with weak correlation (.10) between UR and IR found by Allam (2003) in Egypt. Non-spatial Pearson correlation coefficient between UR and IR was found -.23, which is not significant with $p = .249$. The bivariate spatial correlation between UR and IR was found ($I_{xy} = -.19$), which is not significant having $z = -.97$ and $p = .978$ based on simulation study under standard normal distribution. Although, the correlation values are not significant, most probably its negative direction attributed to the quality of data set and to the measurement error. It was seen that Pearson coefficient is always over estimated when it is used in finding the spatial correlation. That's why, in investigating the bivariate spatial correlation, it is recommended to use spatial measures such as Wartenberg's (1985) measure.

4. DISCUSSION

This study, undertaken in Egypt, investigated the spatial autocorrelation of each of UR and IR, and the bivariate spatial correlation between them. Spatial global clustering and local clusters for UR and IR were examined based on global Moran's I and local Moran's I_i respectively. Bivariate spatial correlation between UR and IR was examined based on Wartenberg's measure. Such findings allow policy makers to better identify what types of resources are needed and precisely where they should be employed. The above framework revealed some noteworthy findings. After rejecting the null hypothesis for UR, concluding that there is a form of global clustering, it is of course, of interest to know the exact nature of this clustering. Are there hot-spot clusters? If so, how many hot-spots are there and where are they located? Also is this clustering associated with the clustering of IR? The population size in the governorates is not equal .i.e., the rate of UR and IR did not express the absolute size of the problem. Several governorates were not observed visually as hot spots. But after considering the information of their neighbours, the pattern of their hot spots can be obviously seen. For example, governorates 7, 8, and 26 were found in IR; and governorates 14, 15, 17, 18, and 21 were found in UR.

Maps provide powerful means to communicate data to others. Unlike information displayed in graphs, tables, and charts; maps also provide bookmarks for memories. In this way, maps were not passive mechanism for presenting information. Usually, in the spatial analysis and geographical mapping, small areas should be used such as districts, counties…etc. But, in this research governorates were used which considered somewhat larger than for example the districts because the data were not available for smaller areas. Most often the word 'neighbourhood' suggests a relatively of small area surrounding individuals' homes. But researchers commonly make use of larger spatial area such as census tracts (Coulton et al., 2001). Often, the choice about neighbourhood spatial definition was made with respect to convenience and availability of contextual data rather than study purpose (Schaefer-McDaniel et al., 2010). Schaefer-McDaniel et al. stated that, researchers might utilize census data and thus rely on census-imposed boundaries to define neighbourhoods even though these spatial areas may not be the best geographic units for the study topic.

As noted by Waller and Jacques (1995), the test for spatial pattern employs alternative hypotheses of two types; the omnibus not the null hypothesis or more specific alternatives. Tests with specific alternatives include focused tests that are sensitive to monotonically decreasing risk as distance from a putative exposure source (the focus) increases. Acceptance of either types (the omnibus or a more specific alternatives) only demonstrates that some spatial pattern exist, and does not implicate a cause (Jacques, 2004). Hence, the existence of a spatial pattern alone cannot demonstrate nor prove a causal mechanism.

Anselin (1995) stated that indication of local pattern of spatial association may be in line with a global indication, although this is not necessarily be the case. It is quite possible that the local pattern is an aberration that the global indicator would not pick up, or it may be that a few local patterns run in the opposite direction of the global spatial trend. Local values that are very different from the mean (or median) would indicate locations that contribute more than their expected share to the global statistic. These may be outliers or high leverage points and thus would invite closer scouting. However, this is found in this research. Although global clustering in IR was not found significant, two local clusters were found significant. Spatial units were arbitrary subdivisions of the study region and people could move around from one area to another. That could be affected by UR and IR variability in areas other than the area they live in. i.e., the variability of UR and IR in the *ith* governorate was thought to be influenced and explained by the variability of UR and IR not just in the *ith* governorate but also in the neighbouring governorates.

The application of statistical techniques to spatial data faces an important challenge, as expressed in the first law of geography: “everything is related to everything else, but closer things are more related than distant things” (Tobler, 1979). The quantitative expression of this principal is the effect of *spatial dependence*. i.e., when the observed values are spatially clustered, the samples are not independent. UR and IR growth in governorate i generates UR and IR growth in governorate j. This mechanism of transmission causes a spatial autocorrelation of UR and IR growth. The obvious question after finding significant clusters in UR and IR is-why? Could this pattern associated by the spatial pattern of other socioeconomic indicators such as the levels of household income or by the limitation of economic resources? However, further research regarding this bivariate spatial association between UR and IR and other socioeconomic indicators is required. It will be of our interest in the near future. This paper adds to the global body of knowledge on the utilization of spatial analysis to strengthen the research–policy interface in the developing countries such as Egypt.

It should be emphasized that UR and IR inequality problem cannot be overcome in the short-run; but long-term efforts are needed to tackle this problem. In turn, enabling the economy to create more job opportunities and establish new projects, especially in the governorates that found as hot spot clusters. It means that the place of the problem is now clearly shown. Lack of investment in all levels of education and other life skills result in permanent life-long inequality problem. Finally, UR and IR studies should be conducted periodically in light of the changing of socioeconomic and political conditions.

Details presented in this research enables development actors to identify priorities, select fields and locate trends in designing national, regional, or sectorial policies. Moreover, this study provides civil society organizations, researchers, and citizens with a rich knowledge base of facts and analysis, offering information easy to understand and to apply. This knowledge can be used either in targeted initiatives, or as a tool for monitoring and assessing policies and methods. Most importantly, the research provides the information needed to both examine development strategies’ consistency with the actual situation, and to align priorities and develop them in accordance with the Egyptian people’s development needs.

As we not expected, spatial correlation between UR and IR was not found significant as was hypothesized by some studies. Although, this work was conducted as part of a wider study, its immediate implications are more for policy makers and practitioners than for researchers. Policy which pays attention to area characteristics will reduce UR and IR variability. Consequently improves the prosperity which in turn will improve health status. Indeed, the authors agree with the statement stated by Allam (2003) that further efforts are required to meet the challenges related to participation in development, particularly in the areas of illiteracy and unemployment.

To reduce the risk of unemployment in Egypt like other developing countries, efforts should focus on generating growth and deregulating labor markets rather than on expanding current labor market programs. Interventions that create opportunities for the poor people and reduce vulnerability should have priority over traditional social assistance interventions.

5. CONCLUSIONS

Conclusions are comprehensive in at least five aspects. First, based on visual inspection of mapping, high level of UR and IR were concentrated in eastern-northern and southern parts and in western-northern and middle parts respectively. Second, several governorates were not observed visually as hot spots. But after considering the information of their neighbours, the pattern of their hot spots can be obviously seen. Third, global clustering was found in UR, but was not found in IR. Two governorates were found to be local clusters in each of UR and IR in eastern-northern, middle, and southern parts. The opposite was the case for those with low UR and IR were seen in some northern and southern parts. Forth, from negative local Moran values, looking at local variation, some governorates represented as areas of dissimilarity in each of UR and IR. These governorates with low UR or IR were surrounded by governorates with high UR or IR, or vice versa. Fifth, spatial correlation between UR and IR was not found significant. In summary, the present study supports the hypothesis of a spatial inequality in each of UR and IR at governorate level that probably reflects the inequality distribution in several socioeconomic indicators across governorates of Egypt. While results were specific to these data, the case study helps to identify general concepts for future studies.

REFERENCES

1. Allam, S.T. (2003). *Egypt: human development report*. pp. 1-166.
2. Amerah, M. (1993). *Unemployment in Jordan: dimensions and prospects*. Center for international studies. pp. 1-56. (http://library.fes.de/pdf-files/bueros/vifa-nahost/b93_00075.pdf).
3. Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27(2), 93-115.
4. Arab Republic of Egypt (2007). 2007-2011 country strategy paper. African Development Bank.
5. Cliff, A.D., and Ord, J.K. (1981). Spatial processes: Models & Applications. London: Page Bros.
6. Coulton, C.J., Korbin, J., Chan, T., and Su, M. (2001) Mapping residents' perceptions of neighbourhood boundaries: A methodological note. *American Journal of Community Psychology*, 29(2), 371-383.
7. El Gamal, Y. (2012).National Strategic Plan for Pre-University Education Reform in Egypt. pp. 1-575.
8. Elhorst, J.P. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanations. *Journal of Economic Survey*, 17(5), 709-740.

9. Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.
10. Hooghe, M., Vanhoutte, B., Hardyns, W., and Bircan, T. (2011) Unemployment, Inequality, Poverty and Crime: Spatial Distribution Patterns of Criminal Acts in Belgium, 2001-06. *British Journal of Criminology*, 51(1), 1-20.
11. Jin, R. L., Shah, C.P. and Svoboda, T.J. (1995). The impact of unemployment on health: A review of the evidence. *Canadian Medical Association Journal*, 153(5): 529-540.
12. Koch, T. (2005) *Cartographies of disease: maps, mapping, and medicine*, 1st. Ed. California: Guilford Press.
13. Lawson, A.B. and Williams, F.R. (2001). *An introductory guide to disease mapping*. New York: Wiley & Sons.
14. Macintyre, S., Maciver, S. and Sooman, A. (1993) Area, class, and health: Should we be focusing on places or people. *Journal of Social Policy*, 22, 213-234.
15. New York Times (2002). *World briefly / Middle East*: 'Alarming' illiteracy among Arabs.
16. Osman, M., Zakareya, E., and Mahrous, W. (2006). Targeting the Poor in Egypt: A ROC Approach. pp. 1-43. http://www.idsc.gov.eg/Upload/Documents/63/EN/taregeting%20the%20poor.pdf.
17. Schabenberger, O. and Gotway, C.A. (2005). *Statistical methods for spatial data analysis*. Boca Raton: Chapman & Hall.
18. Schaefer-McDaniel, N., Caughy, M. O'B., O' Campo, P. and Gearey, W. (2010). Examining methodological details of neighbourhood observations and the relationship to health: A literature review. *Social science & medicine*, 70(2), 277-292. doi: 10.1016/j.socscimed.2009.10.018.
19. Tobler, W.R. (1979). *Cellular geography, In: Philosophy in Geography*. Holland: DReidel Publishing Company.
20. UNICEF (2010). *Child Poverty and disparities in Egypt: Building the Social Infrastructure for Egypt's Future*. 1-80.
21. Wahba, J. and Zenou, Y. (2005). Density, social networks and job search methods: Theory and application to Egypt. *Journal of Development Economics*, 78, 443-473. doi: 10.1016/j.jdeveco.2004.11.006.
22. Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data*. New Jersey: Wiley & Sons.
23. Waller, L.A., and Jacques, G.M. (1995). Disease models implicit in statistical tests of disease clustering. *Epidemiology*, 6, 584-590.
24. Wartenberg, D. (1985). Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis. *Geographical Analysis*, 17(4), 263-282.
25. Weeks, J.R. (1992) *Population: An Introduction to concepts and Issues*, 5th. Ed. Wadsworth, Inc.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 257-260

SOME INFINITE SERIES OF NON-BINARY NEIGHBOR DESIGNS IN CIRCULAR BLOCKS OF SIZE FOUR AND OF FIVE

Munir Akhtar[1], **Rashid Ahmed**[2] and **Fariha Yasmin**[3]
[1] COMSATS Institute of Information Technology, Wah Cantt, Pakistan.
Email: munir_stat@yahoo.com
[2] Government Higher Secondary School Mitroo, Vehari, Pakistan.
Email: rashid701@hotmail.com
[3] FC University, Lahore, Pakistan. Email: fariha_stat@yahoo.com

ABSTRACT

Neighbor designs are more useful to remove the neighbor effects in experiments where the performance of a treatment is affected by the treatments applied to its adjacent plots. Considering the neighbor effects, non-binary designs have their own importance. In this article, therefore, some infinite series are developed to generate non-binary neighbor designs in circular blocks of size four and of five.

MSC(2000): 05B05, 62K10

KEY WORDS

Circular blocks; Neighbor effects; Neighbor designs, Non-binary neighbor designs.

1. INTRODUCTION

Experiments in agriculture, horticulture and forestry often show neighbor effects. In such experiments neighbor balanced designs are useful to remove the neighbor effects. Rees (1967) introduced neighbor designs in serology and constructed these designs for v odd. A design for v treatments in b circular blocks (each with k plots) in which each treatment is a neighbor of every other treatment exactly λ' times is said to be neighbor design. A block in which each treatment appears at most once is called binary block. A blocks design in which all blocks are not binary is called non-binary block design. The construction of neighbor designs by Rees (1967) and Hwang (1973) is not necessarily binary. Several more references are discussed in detail by Ahmed et al. (2011). Bailey and Druilhet (2004) stated that ignoring neighbor effects, a block design is inefficient if any treatment occurs more than once in a block. But non-binary neighbor designs, where no treatment is neighbor to itself have their own importance to remove the neighbor effects. Naqvi et al. (2010) constructed non-binary neighbor designs in circular blocks and developed no series. The practitioners, therefore, have to generate the required neighbor designs by some hit and trial. In this article, some infinite series of non-binary neighbor designs in circular blocks of size 4 and 5 are developed. In next Section, construction method is explained with examples.

2. CONSTRUCTION METHOD

Non-Binary neighbor designs proposed in this article are constructed using method of cyclic shifts which is explained below.

Rule I:

Let S = [q_1, q_2, …, q_{k-1}] be a set of shifts, where 1≤ $q_i \le v$-1. Then S* be a set which consists each element of S and ($q_1+q_2+\ldots+q_{k-1}$) mod v along with their complements. In rule I, complement of q_i is v-q_i . A design is Non-Binary neighbor design if:

(i) S* consists of 1, 2, …, (v-1) an equal number of times, say λ' .
(ii) Sum of any two or three, …, or (k-2) consecutive elements of S is zero (mod v).

If S = [q_1, q_2, …, q_{k-1}] satisfies the two conditions (i) and (ii) then required design is generated by developing the following initial block cyclically mod v.

(0, q_1, (q_1+q_2), …, (q_1+ $q_2+\ldots+$ q_{k-1})) mod v.

Rule II:

Let S = [q_1, q_2, …, q_{k-2}]t be a set of shifts, where 1≤ $q_i \le v$-2. Then S* be a set which consists each element of S and ($q_1+q_2+\ldots+q_{k-2}$) mod (v-1) along with their complements. In rule II, complement of q_i is v-1-q_i . A design is Non-Binary neighbor design if:

(a) S* consists of 1, 2,…, (v-2) an equal number of times, say λ' .
(b) Sum of any two or three, …, or (k-3) consecutive elements of S is zero mod (v-1).

If S = [q_1, q_2, …, q_{k-2}]t satisfies the two conditions (a) and (b) then required design is generated by developing the following initial block cyclically mod (v-1).

(0, q_1, (q_1+q_2), …, (q_1+ $q_2+\ldots+$ q_{k-2}), ∞) mod (v-1), where ∞ = v-1.

Example 2.1.

If v = 8 and k = 7 then S = [2,1,3,4,3,2] provide the initial block (0,2,3,6,2,5,7) which generates the following non-binary neighbor design.

B_1 = (0,2,3,6,2,5,7), B_2 = (1,3,4,7,3,6,0), B_3 = (2,4,5,0,4,7,1), B_4 = (3,5,6,1,5,0,2),
B_5 = (4,6,7,2,6,1,3), B_6 = (5,7,0,3,7,2,4), B_7 = (6,0,1,4,0,3,5), B_8 = (7,1,2,5,1,4,6).

Here S* = [2,1,3,4,3,2,1,6,7,5,4,5,6,7] which contains 1, 2, … , (v-1) twice and sum of 2nd, 3rd and 4th elements of S is zero (mod 8), therefore, given design is non-binary neighbor design.

Example 2.2.

If v = 14 and k = 7 then S = [2,1,3,4,5,6]+[1,2,3,4,6]t provides the two initial blocks (0,2,3,6,10,2,8) and (0,1,3,6,10,3, ∞) which generate the following non-binary neighbor design.

$B_1 = (0,2,3,6,10,2,8)$, $B_2 = (1,3,4,7,11,3,9)$, $B_3 = (2,4,5,8,12,4,10)$,
$B_4 = (3,5,6,9,0,5,11)$, $B_5 = (4,6,7,10,1,6,12)$, $B_6 = (5,7,8,11,2,7,0)$,
$B_7 = (6,8,9,12,3,8,1)$, $B_8 = (7,9,10,0,4,9,2)$, $B_9 = (8,10,11,1,5,10,3)$,
$B_{10} = (9,11,12,2,6,11,4)$, $B_{11} = (10,12,0,3,7,12,5)$, $B_{12} = (11,0,1,4,8,0,6)$,
$B_{13} = (12,1,2,5,9,1,7)$, $B_{14} = (0,1,3,6,10,3, \infty)$, $B_{15} = (1,2,4,7,11,4, \infty)$,
$B_{16} = (2,3,5,8,12,5, \infty)$, $B_{17} = (3,4,6,9,0,6, \infty)$, $B_{18} = (4,5,7,10,1,7, \infty)$,
$B_{19} = (5,6,8,11,2,8, \infty)$, $B_{20} = (6,7,9,12,3,9, \infty)$, $B_{21} = (7,8,10,0,4,10, \infty)$,
$B_{22} = (8,9,11,1,5,11,\infty)$, $B_{23} = (9,10,12,2,6,12,\infty)$,$B_{24} = (10,11,0,3,7,0,\infty)$,
$B_{25} = (11,12,1,4,8,1,\infty)$, $B_{26} = (12,13,2,5,9,2,\infty)$, where $\infty = 13$

3. CONSTRUCTION OF NON-BINARY NNBD FOR k = 4

Series 3.1.

Non-Binary NNBD with $\lambda' = 2$ can be generated for $v = 4t+1$; t integer and k = 4 through the following t sets of shifts.

$S_{j+1} = [2j+1, v-(2j+1), 4j+2]$; $j = 0,1, \ldots, t-1$.

Example 3.1.

Non-Binary NNBD is generated for $v = 25$ and k = 4 through the following six sets of shifts.

$S_1 = [1, 24, 2]$, $S_2 = [3, 22, 4]$, $S_3 = [5, 20, 6]$,
$S_4 = [7, 18, 8]$, $S_5 = [9, 16, 10]$, $S_6 = [11, 14, 12]$

Series 3.2.

Non-Binary NNBD with $\lambda' = 4$ can be generated for $v = 4t+3$; t integer and k = 4 through the following $(v-1)/2$ sets of shifts.

$S_{j+1} = [j+1, v-(j+1), j+2]$; $j = 0,1, \ldots, (v-5)/2$.
$S_{(v-1)/2} = [1, v-1, (v-1)/2]$

Series 3.3.

Non-Binary NNBD with $\lambda' = 2$ can be generated for $v = 4t$; t integer and k = 4 through the following t sets of shifts.

$S_{j+1} = [2j+1, v-(2j+1), 2j+2]$; $j = 0,1, \ldots, t-2$.
$S_t = [(v-2)/2, (v-2)/2]t$

4. CONSTRUCTION OF NON-BINARY NNBD FOR k = 5

Series 4.1.

Non-Binary NNBD with $\lambda' = 10$ can be generated for $v = 2m+2$; $m > 3$ and k = 5 through the following v sets of shifts.

$S_j = [j, j, j, j](2)$; $j = 1, 2, \ldots, m-2$.

$S_{2\text{m}-1} = [m-1, m-1, m-1, m-1]$,
$S_{2m-2} = [2, 2, 6]t$,
$S_{2m-1} = [m-1, m-1, m+2]t$, $S_{2m} = [m-1, m+1, m]t$,
$S_{2m+1} = [m, m+1, m-1]t$, $S_{2m+2} = [m, m, m+1]t$

Series 4.2.

Non-Binary NNBD with $\lambda' = 5$ can be generated for $v = 2t+1$; $t(>4)$ in circular blocks of five units through the following t sets of shifts, where v is divisible by three.

$S_1 = [1, 2, 1, 4]$,
$S_{j\text{-}1} = [j, j, j, j]$; $j = 3, 4,\ldots, t$ *and* $j \neq (v/3)$.
$S_{t\text{-}1} = [2v/3, v/3, 1, 1]$,
$S_t = [v/3, v/3, v/3, 2]$

REFERENCES

1. Ahmed, R., Akhtar, M. and Yasmin, F. (2011). Brief review of one dimensional neighbor balanced designs since 1967. *Pakistan Journal of Commerce and Social Sciences*, 5(1), 100-116.
2. Bailey, R.A. and Druilhet, P. (2004). Optimality of neighbor-balanced designs for total effects. *Annals of Statistics*, 32, 4, 1650-1661.
3. Naqvi, H., Yab, M.Z. and Hanif, M. (2010). Non-binary neighbor balance circular designs for $v = 2$n and $\lambda = 2$. *Journal of Statistical Planning and Inference*, 140, 11, 3013-3016.
4. Hwang, F.K. (1973). Constructions for some classes of neighbor designs. *Annals of Statistics*, 1, 4, 786-790.
5. Rees, D.H. (1967). Some designs of use in serology, *Biometrics,* 23, 779-791.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 261-268

MEAN RESIDUAL AND MEAN PAST LIFETIME OF MULTI STATE CONSECUTIVE K-OUT-OF-n:F SYSTEMS

Elnaz Karimian
Department of Statistics, University of Isfahan, Isfahan, 81744, Iran
Email: elnaz.karimian@gmail.com

ABSTRACT

Recently, traditional reliability theory where the system and the components are always described simply as functioning or failed is on the way to being replaced by a theory for the multi state systems. In a multi state system, both the system and its components are allowed to experience more than two possible states. Multi state consecutive k-out-of-n:F systems, have attracted the attention of many engineers and researchers. On the other hand, the concepts of mean residual and mean past lifetime have been of much interest in the literatures. These measures have been successfully used in binary state reliability analysis and recently, much attention has been paid to mean residual and mean past lifetime of multi state systems. This paper studies the mean residual and mean past lifetime of the multi state linear consecutive k-out-of-n:F systems with a constant k value, which is a special case of the general multi state consecutive k-out-of-n:F system, under the assumption that the degradations in systems and components follow an acyclic Markov process which has a discrete state space.

Keywords: Consecutive k-out-of-n:F system, multi state, reliability, mean residual lifetime, mean past lifetime.

1. Introduction

In binary context, both the system and its components are allowed to take two possible states: either working or failed. Many researcher have studied the reliability evaluation of binary consecutive k-out-of-n system, for example the reader is referred to Chiang and Niu (1981), Hwang (1982), Kuo and Zuo (2003) and Lambiris and Papastavridis (1985). The dual relationship between the consecutive k-out-of-n:F and G systems is investigated by Kuo et al. (1990) and Zuo (1993).

In a discrete multi state system, it's assumed that both the system and its components are allowed to be in one of $M+1$ ($M > 0$) possible states: for example, completely working, partially working, and completely failed. See, Lisnianski and Levitin (2003) for the details of this theory.

Recently, the definition of the binary consecutive k-out-of-n system is extended to the multi state case, that the system state remains binary and its components have more than two possible states, for example see, Zuo and Liang (1994) and Malinowski and Preuss (1995, 1996).

But sometimes, it's necessary that the system and it's components in a consecutive k-out-of-n system to have more than two possible states, see Koutras(1997). Haim and Porat (1991) provide a Bayes reliability model of the consecutive k-out-of-n system, while k is supposed to be constant. When k is constant, the system has the same reliability structure at all system state levels. However, a multi-state system may have different structures at different system levels. for the details we refer to Huang et al. (2003).

Dynamic reliability analysis of consecutive k-out-of-n systems has been studied in various papers, for example see, Eryilmaz (2007) and Salehi et al. (2011). And nowadays the reliability measures of multi state consecutive k-out-of-n system has attracted attention of many researchers.

One of the most important measure that can be used to evaluate the stochastic behavior of survival over time, is mean residual lifetime (MRL) and also another useful measure is mean past lifetime (MPL). The reader is referred to Some contributions on MRL and MPL concepts, Guess and Proschan (1988), and Lai and Xie (2006), Asadi and Goliforushani (2008), Poursaeed and Nematollahi (2008), Zhao and Balakrishnan (2009), Sun and Zhang (2009), Shen et al. (2010). Recently, many researchers are interested on MRL and MPL of multi state system, for example, Eryilmaz (2010) is one of the recent study on MRL and MPL of multi states system.

In this paper MRL and MPL functions for multi state consecutive k-out-of-n:F systems are obtained and also we study the multi state parallel, and multi state consecutive k-out-of-n:F systems which have three states.

2. Model

2.1. Assumptions

We assume the followings for the system and the components in this paper.

1. The system and it's components have the state set, $\{0, 1, ..., M\}$, where the states "0", and "M" represent respectively the worst (completely failed), and the best (completely functioning) states.

2. The systems and components degrade with time t from the perfect state "M" to lower states.

3. The system and it's components are nonrepairable.

4. The components are s-independent and identical.

5. The degradation in systems and components follow an acyclic Markov process which has a discrete state space.

2.2. Notation

n number of components in the system.

k minimum number of consecutive failed components, that cause system failure.

$\phi_{k,n}(t)$ the state of a multi state consecutive k-out-of-n:F systems at time t.

$X_i(t)$ the state of component i at time t, $i = 1, 2, \ldots, n$.

$T_i^{\geq j}$ the lifetime of component i in the subset $\{j, j+1, \ldots, M\}$ of states $\{0, 1, \ldots, M\}$, $j = 1, 2, \ldots, M$.

$N(j, r, m)$ the number of ways to place j identical balls in r district urns, while any urn contains at most m balls.

$T_{k,n}^{\geq j}$ the lifetime of a multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \ldots, M\}$ of states $\{0, 1, \ldots, M\}$, $j = 1, 2, \ldots, M$.

$m_{k,n}^{\geq j}(t)$ the mean residual lifetime of a multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \ldots, M\}$ of states $\{0, 1, \ldots, M\}$, $j = 1, 2, \ldots, M$.

$m{^*}_{k,n}^{\geq j}(t)$ the mean past lifetime of a multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \ldots, M\}$ of states $\{0, 1, \ldots, M\}$, $j = 1, 2, \ldots, M$.

3. Mean Residual and Mean Past Lifetime

Definition 1 (Eryilmaz (2010)). *The MRL vector of a multi state system with states $\{0, 1, \ldots, M\}$ is defined as*

$$\boldsymbol{M}(t) = \left(m^{\geq 1}(t), m^{\geq 2}(t), \ldots, m^{\geq M}(t)\right),$$

where the jth component, $m^{\geq j}(t)$,of $\boldsymbol{M}(t)$, represents the MRL of a multi state system in state j or above under the condition that the system is in state j or above at time t. It can be computed by,

$$\begin{aligned} m^{\geq j}(t) &= \frac{1}{P\left\{T^{\geq j} > t\right\}} \int_0^{\infty} P\left\{T^{\geq j} > t + x\right\} dx \\ &= \frac{1}{P\left\{\phi(t) \geq j\right\}} \int_0^{\infty} P\left\{\phi(t+x) \geq j\right\} dx. \qquad (1) \end{aligned}$$

Definition 2 (Eryilmaz (2010)). *The MPL vector of a multi state system with states $\{0, 1, \ldots, M\}$ is defined as*

$$\boldsymbol{M^*}(t) = \left(m^{*\geq 1}(t), m^{*\geq 2}(t), \ldots, m^{*\geq M}(t)\right),$$

where the jth component, $m^{*\geq j}(t)$,of $\boldsymbol{M^*}(t)$, represents the mean time elapsed from degradation of the system from state j to a lower state given that the system is below state j at time t. It can be computed by,

$$\begin{aligned} m^{*\geq j}(t) &= \frac{1}{P\left\{T^{\geq j} \leq t\right\}} \int_0^t P\left\{T^{\geq j} \leq x\right\} dx \\ &= \frac{1}{P\left\{\phi(t) < j\right\}} \int_0^t P\left\{\phi(x) < j\right\} dx. \end{aligned} \tag{2}$$

3.1. Multi State Consecutive k-out-of-n:F Systems

Definition 3 (Huang et al. (2003)). *$\phi(x) < j\,(j = 0, 1, \ldots, M)$ iff at least k consecutive components are in state below j. An n component system with such a property is called a multi state consecutive k-out-of-n:F system.*

In this paper,we suppose that k in above definition is constant, this means that the structure of the system is the same for all state levels, such a system is called constant multi state consecutive k-out-of-n:F system, but the structure of the multi state system may be different for different system state levels. We obtain the survival function corresponding to the multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \ldots, M\}$ from Derman et al. (1982) and Definition 3, as

$$\begin{aligned} P\left\{T_{k,n}^{\geq j} > t\right\} &= P\left\{\phi_{k,n}(t) \geq j\right\} \\ &= \sum_{i=0}^{n} N(i, n-i+1, k-1) p^{n-i} q^{i}, \\ &= \sum_{i=0}^{n} (-1)^{i} \omega_{i,k}(n) P(\{X(t) < j\})^{i}, \\ &= \sum_{i=0}^{n} (-1)^{i} \omega_{i,k}(n) P\left\{T_{i,i}^{\geq j} \leq t\right\}, \end{aligned} \tag{3}$$

where $T_{i,i}^{\geq j}$ is the lifetime of a multi state parallel system of n components in the subset $\{j, j+1, \ldots, M\}$, $\omega_{i,k}(n) = \sum_{l=0}^{i} \binom{n-l}{i-l} (-1)^{l} N(l, n-l+1, k-1)$ and $N(i, n-i+1, k-1) = \sum_{\lambda=0}^{n-i+1} \binom{n-i+1}{\lambda} \binom{n-\lambda k}{i-\lambda k} (-1)^{\lambda}$. So, multi state consecutive k-out-of-n:F system can be obtained via multi state Parallel systems. By using (3) and (2), the MPL of a multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \ldots, M\}$ is obtained as

$$m_{k,n}^{*\geq j}(t) = \frac{t - \sum_{i=0}^{n} (-1)^{i} \omega_{i,k}(n) \int_0^t P\left\{T_{i,i}^{\geq j} \leq x\right\} dx}{1 - \sum_{i=0}^{n} (-1)^{i} \omega_{i,k}(n) P\left\{T_{i,i}^{\geq j} \leq t\right\}}$$

$$m^{*\geq j}_{k,n}(t) = \frac{t - \sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) P\left\{T^{\geq j}_{i,i} \leq t\right\} m^{*\geq j}_{i,i}(t)}{1 - \sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) P\left\{T^{\geq j}_{i,i} \leq t\right\}}, \qquad (4)$$

where $m^{*\geq j}_{i,i}(t)$ represents the MPL of multi state consecutive n-out-of-n:F system (multi state parallel system) in the subset $\{j, j+1, \dots, M\}$. By replacing (3) in (1), the MRL function of a multi state consecutive k-out-of-n:F system in the subset $\{j, j+1, \dots, M\}$ is obtained as

$$\begin{aligned} m^{\geq j}_{k,n}(t) &= \frac{\sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) \int_0^\infty P\left\{T^{\geq j}_{i,i} \leq t + x\right\} dx}{\sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) P\left\{T^{\geq j}_{i,i} \leq t\right\}} \\ &= \frac{\sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) [1 - P\left\{T^{\geq j}_{i,i} \leq t\right\} m^{*\geq j}_{i,i}(t)]}{\sum_{i=0}^{n}(-1)^i \omega_{i,k}(n) P\left\{T^{\geq j}_{i,i} \leq t\right\}}. \qquad (5) \end{aligned}$$

3.2. Multi State Parallel Systems with Three States

Consider a system with n identical components, and assume that the degradation in components follows an acyclic Markov process, and the system and the components have three states: 0 (failed), 1 (partially failing), 2 (perfect functioning). If we take the repair rate as zero, then from Pham and et al. (1996) and Eryilmaz (2010), we have the following state probabilities for the components.

$$\begin{aligned} P\{X_i(t) = 1\} &= \frac{\lambda_{21}}{\lambda_{21} + \lambda_{20} - \lambda_{10}} \left(e^{-\lambda_{10} t} - e^{-(\lambda_{21} + \lambda_{20})t}\right), \\ P\{X_i(t) = 0\} &= 1 - \left(\frac{\lambda_{21}}{\lambda_{21} + \lambda_{20} - \lambda_{10}} e^{-\lambda_{10} t} + \frac{\lambda_{20} - \lambda_{10}}{\lambda_{21} + \lambda_{20} - \lambda_{10}} e^{-(\lambda_{21} + \lambda_{20})t}\right), \end{aligned}$$

and therefore,

$$\begin{aligned} P\left\{T^{\geq 1}_i \leq t\right\} &= P\{X_i(t) < 1\} \\ &= 1 - \left(\frac{\lambda_{21}}{\lambda_{21} + \lambda_{20} - \lambda_{10}} e^{-\lambda_{10} t} + \frac{\lambda_{20} - \lambda_{10}}{\lambda_{21} + \lambda_{20} - \lambda_{10}} e^{-(\lambda_{21} + \lambda_{20})t}\right), \\ P\left\{T^{\geq 2}_i \leq t\right\} &= P\{X_i(t) < 2\} \\ &= 1 - e^{-(\lambda_{21} + \lambda_{20})t}, \end{aligned}$$

for $t \geq 0$, and $i = 1, 2, \dots, n$.
So, state probabilities of a parallel system of n components are obtained as

$$P\left\{T_{n,n}^{\geq 1} \leq t\right\} = \left[1-\left(\frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-\lambda_{10}t} + \frac{\lambda_{20}-\lambda_{10}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-(\lambda_{21}+\lambda_{20})t}\right)\right]^{n}, \tag{6}$$

and

$$P\left\{T_{n,n}^{\geq 2} \leq t\right\} = \left[1-e^{-(\lambda_{21}+\lambda_{20})t}\right]^{n}$$

for $t \geq 0$.

In the following, we have a mixture representation for the state probability of a multi state parallel system.

Lemma 1. *For $M = 2$, the distribution function of an n component multi state parallel system in state "2" or above can be found as*

$$P\left\{T_{n,n}^{\geq 2} \leq t\right\} = \sum_{i=0}^{n} \alpha_i(n) e^{-it(\lambda_{21}+\lambda_{20})}, \tag{7}$$

where $\alpha_i(n) = \binom{n}{i}(-1)^i, i = 1, \ldots, n.$

Proof. The proof is immediately derived from

$$\begin{aligned} P\left\{T_{n,n}^{\geq 2} \leq t\right\} &= \left[1-e^{-(\lambda_{21}+\lambda_{20})t}\right]^{n} \\ &= \sum_{i=0}^{n}\binom{n}{i}(-1)^i\left(e^{-t(\lambda_{21}+\lambda_{20})}\right)^i \end{aligned}$$

□

Using Lemma 1, the components of the MPL vector of multi state parallel systems are computed as

$$m_{n,n}^{*\geq 2}(t) = \frac{\sum_{i=0}^{n}\frac{\alpha_i(n)}{i(\lambda_{21}+\lambda_{20})}\left[1-e^{-it(\lambda_{21}+\lambda_{20})}\right]}{\sum_{i=0}^{n}\alpha_i(n)e^{-it(\lambda_{21}+\lambda_{20})}}, \tag{8}$$

and

$$m_{n,n}^{*\geq 1}(t) = \frac{\sum_{i=0}^{n}\binom{n}{i}(-1)^i\int_0^t P\left\{T_{i,i}^{\prime\geq j} > x\right\}dx}{\left[1-\left(\frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-\lambda_{10}t}+\frac{\lambda_{20}-\lambda_{10}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-(\lambda_{21}+\lambda_{20})t}\right)\right]^{n}};$$

Where $T_{i,i}^{\prime\geq j}$ is the lifetime of a multi state n-out-of-n:G system (Multi state series system)that is presented by Eryilmaz (2010) as,

$$P\left\{T_{i,i}^{\prime\geq j} > x\right\} = \left[\frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-\lambda_{10}t} + \frac{\lambda_{20}-\lambda_{10}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}\,e^{-(\lambda_{21}+\lambda_{20})t}\right]^{i}$$

So

$$m_{n,n}^{*\geq 1}(t) = \frac{\sum_{i=0}^{n}\sum_{j=0}^{i}\frac{\theta_{i,j}(n)}{j\lambda_{10}+(i-j)(\lambda_{21}+\lambda_{20})}\left[1-e^{-t(j\lambda_{10}+(i-j)(\lambda_{21}+\lambda_{20}))}\right]}{\sum_{i=0}^{n}\sum_{j=0}^{i}\theta_{i,j}(n)\,e^{-t(j\lambda_{10}+(i-j)(\lambda_{21}+\lambda_{20}))}}, \qquad (9)$$

for $t \geq 0$.
Where $\theta_{i,j}(n) = \binom{n}{i}(-1)^i\alpha'_j(i)$, $\alpha'_j(i) = \binom{i}{j}\omega^j(1-\omega)^{i-j}$ and $\omega = \frac{\lambda_{21}}{\lambda_{21}+\lambda_{20}-\lambda_{10}}$.

Now, the MPL and MRL vectors of multi state consecutive k-out-of-n:F system with three states can also be computed by using (6), (9) and (7), (8) in (4), (5).

REFERENCES

1. Asadi, M. and Goliforushani, S. (2008). On the mean residual life function of coherent systems. *IEEE Trans. Reliability*, **57**, 574-580.

2. Chiang, D.T. and Niu, S. (1981). Reliability of consecutive k-out-of-n:F system. *IEEE Transactions on Reliability*, **30(1)**, 87-89.

3. Derman, G. J. Lieberman, S. M. Ross, (1982). On the consecutive k-out-of-n:F system. *IEEE Trans. Reliability*, **R-31**, 57-63.

4. Eryilmaz, S. (2010). Mean Residual and Mean Past Lifetime of Multi State Systems With Identical Components. *IEEE Trans. Reliability*, **59(4)**, 644-649.

5. Eryilmaz, S. (2007). On the lifetime distribution of consecutive k-out-of-n:F system. *IEEE Trans. Reliability*, **56**, 35-39.

6. Guess, F. and Proschan, F. (1988), Mean residual life: theory and applications. *In Handbook of Statistics, P. R. Krishnaiha, Ed. et al.*, **7**, 215-224.

7. Haim, M. and Porat, Z. (1991). Bayes reliability modeling of a multi state consecutive k-out-of-n:F system. *In Proceedings of the Annual Reliability and Maintainability Symposium*, 582-586.

8. Huang, J., Zuo, M. J. and Fang, Z. (2003). Multi state consecutive k-out-of-n systems. *IEEE Trans. Reliability*, **35**, 527-534.

9. Hwang, F.K. (1982). Fast solutions for consecutive k-out-of-n:F system. *IEEE Transactions on Reliability*, **31(5)**, 447-448.

10. Koutras, M.V.(1997). Consecutive-k, r-out-of-n: DFM systems. *Microelectronics and Reliability*, **37(4)**, 597-603.

11. Kuo, W., Zhang, W. and Zuo, M. (1990). A consecutive k-out-of-n:G system:the mirror image of a consecutive k-out-of-n:F system. *IEEE Transactions on Reliability*, **39(2)**, 244-253.

12. Kuo W., Zuo, M. J. (2003). Optimal reliability modeling, principles and applications. *Wiley*, New York.

13. Lai, C. D. and Xie, M. (2006). Stochastic Ageing and Dependence for Reliability.*Springer*, New York.

14. Lambiris, M., Papastavridis, S. (1985). Exact Reliability Formulas for Linear and Circular Consecutive k-out-of-n:F Systems. *IEEE Trans. Reliability*, **R-34(2)**, 124-126.

15. Lisnianski, A. and Levitin, G. (2003). Multi-state System Reliability: Assessment. *Optimization and Applications*. Singapore: World Scientific Pub. Co, Inc..

16. Malinowski, J. and Preuss, W. (1996). Reliability of reverse tree structured systems with multi-state components. *Microelectronics Reliability*,**36(1)**, 1-7.

17. Malinowski, J. and Preuss, W. (1995). Reliability of circular consecutively connected systems with multi state components. *IEEE Transaction on Reliability*, **44(3)**, 532-534.

18. Pham, H., Suprasad, A. and Misra, R. B. (1996). Reliability analysis k-out-of-n systems with partially repairable multi-state components. *Microelectronics and Reliability*, **36**, 1407-1415.

19. Poursaeed M. H. and Nematollahi, A. R. (2008). On the mean past and the mean residual life under double monitoring. *Communications in Statistics-Theory and Methods*, **37**, 1119-1133.

20. Salehi, E. T., Asadi, M., Eryilmaz, S. (2012). On the mean residual lifetime of consecutive k-out-of-n system. *TEST*, **21(1)**, 93-115.

21. Shen, Y. , Xie, M. and Tang, L. C. (2010). On the change point of the mean residual life of series and parallel system. *Australian and New Zealand Journal of Statistics*, **52**, 109-121.

22. Shen, Y. , Tang, L. C. and M. Xie (2009). A model for upside down bathtub shaped mean residual life and its properties. *IEEE Trans. Reliability*, **58**, 425-431.

23. Sun, L. Q. and Z. G. Zhang (2009). A class of transformed mean residual life models with censored survival data. *Journal of the American Statistical Association*, **104**, 803-815.

24. Zhao, M. J. and Balakrishnan, N. (2009), Mean residual life order of convolutions of heterogeneous exponential random variables. *Journal of Multivariate Analysis*, **100**, 1792-1801.

25. Zuo, M. (1993). Reliability and component importance of a consecutive k-out-of-n system. *Microelectronic Reliability*, **33(2)**, 243-258.

26. Zuo, M. J. and Liang, M. (1994). Reliability of multi state consecutively connected systems. *Reliability Engineering and System Safety*, **44**, 173-176.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 269-276

STATISTICAL ANALYSIS OF THE FACTORS AFFECTING THE PROFITABILITY OF COMMERCIAL BANKS IN PAKISTAN

Salahuddin and **M. Zubair Khan**
Department of Statistics, University of Peshawar, Peshawar, Pakistan
Email: salahuddin_90@yahoo.com

ABSTRACT

An important role is played by banks in the operation of an economy of a country. During the period 2005-2009, the world banking sector showed an enormous declining trend in profitability due to worldwide economic recession. Consequently, many well-known banks worldwide filing for bankruptcy and the Pakistani banks also showed a declining trend in both public and private sectors.

The aim of this work is to study and examine various factors, which has an influence on the profitability of commercial banks in Pakistan during this period.

The data used in this study is the quarterly bank level data of 5 years (2005-2009). To determine the important factors in getting maximum profitability, panel data regression technique is used. The methodology of data analysis is that of panel data regressions in the line with Anna and Chan (2009) and Indranarain (2009) in which Return on Assets (ROA) is used as dependent variable. Six variables including interest income, non-interest income, bank size, expense management, deposits and credit risks are used as independent variables (bank specific factors). All the analysis of this study is carried out using the Statistical package “Eviews”. The stationarity of data is tested by applying the unit root test. Correlation matrix approach is used to detect multicollinearity in the data. Durbin-Watson test is used to detect autocorrelation.

To measure the individual impact of each of these independent variables on ROA as well as their pair wise impact on ROA and further joint impact of three variables on ROA, the technique of all possible regression is used to reach the best panel regression model. The analysis suggests that three variables: bank size, deposits and non-interest income have individual significant affect on the profitability of commercial bank of Pakistan.

KEYWORDS

Bank-Specific factors, Multicollinearity, Autocorrelation, All Possible Regression.

1. INTRODUCTION

In Pakistan, an essential role is played by banks in the operation of an economy. Internal factors used by contemporary researchers as determinants of bank profitability include interest income, non-interest income, bank size, expense management, deposits, credit risks, etc. There are a number of studies on determinants of profitability in the banking sector [for example, see Indranarain (2009), Valentina et al. (2009), Panayiotis et al.(2005), Kyriaki et al. (2005), Krunakar et al. (2008)], there is hardly any such study in the context of Pakistan.

State Bank of Pakistan is the controlling authority of the banking sector in Pakistan. Banking sector of Pakistan comprises of banking and non-banking financial institutions. The banking institutions mainly comprises of commercial banks which may be further classified as domestic and foreign commercial banks. Domestic commercial banks are further classified as public sector and private sector commercial banks.

The aim of this study is to examine the contribution of various factors on the profitability of commercial banks in Pakistan. The quarterly data of 5 years for the period 2005-2009, is used to determine the important factors in getting maximum profitability. The study used Return on Assets (ROA) as a measure of profitability in line with Indranarain (2009), Valentina et al. (2009), Panayiotis (2005), Kyriaki et al. (2002), and Krunakar et al. (2008). Return on Assets is defined as the total net income divided by total assets. In other words, ROA show how well a bank's administration is using the bank's real investment resources to get maximum profits.

2. MATERIAL

There are 29 domestic (4 public sectors and 25 private sectors) commercial banks in Pakistan. A sample of 15 banks is drawn by simple random sampling using Goldfish bowl method. Sample size of 15 banks from 29 banks is a moderate size (n/N > 50%), and the sample results can better describe the target population. From the quarterly financial statements of each selected bank, data for six variables: Bank Size (BS), Deposits (DEP), Interest Income (II), Credit Risk (CR), NON-Interest Income (NONII) and Expense Management (EM) is collected for further analysis.

3. METHOD OF ANALYSIS

This study involves panel data, because bank (cross-sectional unit) is surveyed over time. Panel data refers to "pooling of observations on a cross-section of firms (say banks) over several time periods" (Baltagi, 2005). In short, panel data have space as well as time dimensions.

To obtain the best model, the technique of all possible regression is applied to panel data. The best model is selected using criteria of maximum R^2. A panel data regression differs from an ordinary time-series regression or cross-section regression in the way that it has two subscripts on its variables, i.e.

$$\mathbf{Y_{it} = \boldsymbol{\alpha} + \boldsymbol{\beta} X_{it} + \boldsymbol{\varepsilon}_{it}} \qquad \text{where,} \quad i = 1,\dots N; \quad t = 1, \dots T$$

"i" denotes the bank and "t" denotes the period or the time. In other words "i" denotes the cross-section and "t" denotes the time series. "$\boldsymbol{\alpha}$" is the intercept, "$\boldsymbol{\beta}$" is K x 1 vector and "X_{it}" is the ith bank on kth independent variable at time "t".

$$\boldsymbol{\varepsilon_{it} = u_i + v_{it}}$$

"u_i" is called unobserved effect and "v_{it}" is the remainder disturbance term. For example in an equation measuring the profitability of bank, "Y_{it}" measures the profit of the ith bank, whereas X_{it} contains the set of independent variables like bank size, deposits, interest rate etc. In our study "u_i" is banks unobserved ability and "v_{it}" varies with banks and time and therefore called as usual disturbance term in the regression. Alternatively in profitability equation Y_{it} measures the output and X_{it} measures the inputs. The unobserved bank specific effect is measured by "u_i" and we can think of this as unobserved managerial skills etc (see Baltagi, 2005).

Following Baltagi approach (Baltagi, 2001), cross-section weights is used for every bank '*i* ' at time *t*, and the true variance components. This gives a matrix-weighted average of within and between estimators obtained by regressing the cross section averages across time.

Moreover, all the panel regression models are fitted by using Fixed Effect model. The reason is that under Hausman test, null hypothesis is that there is no substantial difference between Fixed Effect and Random Effect models. In case if the null hypothesis is rejected, then Fixed Effect model perform better than Random Effect model (see, Gujarati 2004).

Some preliminary tests are performed before running the panel regression models. These tests includes unit root test for checking stationarity of data, correlation matrix for checking multicollinearity and Durbin-Watson test for testing autocorrelation. White's transformation is used to control for cross-section heteroscedasticity (see, Baltagi, 2005).

To test the time series assumption of stationarity of data, the use of unit root test among the applied researchers is not new, however the application of unit root test in panels is recent (Baltagi, 2005). There are plenty of tests available in Statistical Package "Eviews" for testing the stationarity of data like, LLC, IPS, Breitung, Fisher ADF, Hadri, Fisher PP etc. Each of these tests has some advantages over the others. Among all the available unit root tests, Fisher's tests are appropriate and suitable tests to almost every type of data and in particular the data which does not have a balance panel. These tests perform well as compared to other tests for unit roots in panel data (see, Maddala and Wu, 1999). In order to have a better check of the stationarity assumption we applied five different tests (Fisher PP, Fisher ADF, IPS, LLC and Breitung) in our study for testing the stationarity of six bank specific variables. All the tests have been run at level, if a variable is found non stationary then the same test is run at 1st order difference. Keeping Individual trend and Intercept in the equation, makes the test more powerful.

4. PRELIMINARY TESTS

The correlation matrix, given in Table 1, shows that variables: deposits, expense management and interest income have high correlation (coefficient of correlation greater or equal to 0.90). To overcome the problem of multicollinearity, these three variables are removed from the final regression models. Before their removal, their individual significance was also tested but all the three variables proved insignificant in affecting the profitability of commercial banks of Pakistan. Table (2) shows the results of unit root test for the six bank specific variables. P-values of five bank specific variables are suggesting stationarity of data at 10% level of significance. Only interest income is non stationary at 10% level of significance because null hypothesis of unit root (non-stationary) is rejected, as its p-value is greater than 0.10 each time. For stationarity of interest income, the unit root test is run again at first order difference and the variable becomes stationary even at 5% level of significance (see Table-3) as its p-value is smaller than 0.05 each time. Durbin-Watson statistic is given in each table of regression model suggesting no autocorrelation problem at all.

Table 1
The Correlation Matrix

	BS	**CR**	**DEP**	**EM**	**II**	**NONII**
BS	1					
CR	-0.098	1				
DEP	0.9968*	-0.098	1			
EM	0.9401*	-0.082	0.938*	1		
II	0.9679*	-0.09	0.961*	0.9441*	1	
NONII	0.8772	-0.084	0.872	0.8476	0.87051	1

Note: “*” indicates high correlation.

Table 2
Panel Unit Root Test For bank specific variables
(Individual trend & Intercept are included in equation)

S#	**Variable**	**Fisher ADF Chi-Square**	**Fisher PP**	**Im Pesaran Shin**	**Breitung t-test**	**Test at**	**Stationary/ Non-Stationary**
1	Bank Size (BS)	47.096* (0.0244)	69.2738* (0.0001)	-1.2909* (0.0984)		Level	Stationary
2	Deposits (DEP)	50.5016* (0.0110)	79.2669* (0.0000)	-1.4722* (0.0705)		Level	Stationary
3	Interest Income (II)	24.3689* (0.7551)	40.0015* (0.1048)	1.6932* (0.9548)		Level	Non-Stationary
4	Non-Interest Income (NONII)	40.6529* (0.0928)	74.8448* (0.0000)	-1.1948* (0.1161)	-3.1363* (0.0009)	Level	Stationary
5	Expense Management (EM)	40.9740* (0.0873)	117.485* (0.0000)	-1.5534* (0.0602)		Level	Stationary
6	Credit Risk (CR)	54.7906* (0.0037)	106.405* (0.0000)	-2.5394* (0.0056)		Level	Stationary

Note: ‘*’ shows the value of the statistic used.

1. In the parenthesis p-values are given.
2. Null hypothesis of unit root (non-stationary) at $\alpha = 0.10$ is tested for the variables above. Interest income is non-stationary as ‘p-values’ suggest to accept the null hypothesis of unit root.
3. For non-interest income ‘Breitung t-test’ is also used to check the stationarity because ‘Im Pesaran Shin’ test suggested non stationarity.

Table 3
Panel Unit Root Test At 1st Difference for bank specific variables
(Individual trend & intercept are included in equation)

S#	Variable	ADF Fisher Chi-square	Im Pesaran Shin	PP-Fisher chi-square	Stationary/ Non-Stationary
1	Interest Income(ii)	47.9789* (0.0199)	-1.8392* (0.0329)	106.239* (0.0000)	Stationary

Note: ‘*’ shows the value of the statistic used.
P-values are given in parenthesis.

5. ANALYSIS AND RESULTS:

To test significance of individual variable, six models are obtained using the technique of “all possible regression. Results are given in Table (4). These results suggest that three out of six variables are significantly affecting the profitability on individual basis.

Table 4
Individual impact of variables on ROA

Variable	Coefficient	Standard Error	t-statistic	Probability	R-Squared	Durbin-Watson Statistic
Bank size **(BS)***	3.75E-10	6.05E-11	6.199745	0.0000	0.223709	1.342691
Deposits **(DEP)***	4.96E-10	7.43E-11	6.672958	0.0000	0.271503	1.347800
Credit risk **(CR)**	0.000534	0.003609	0.148032	0.8824	0.082974	1.306233
Interest income **(II)**	1.17E-08	2.32E-09	5.058392	0.7798	0.280960	1.371047
Expense management **(EM)**	5.77E-08	1.23E-08	4.692318	0.6324	0.073634	1.219504
Non-interest income **(NONII)***	6.07E-08	1.15E-08	5.255284	0.0000	0.181930	1.328367

Note: **‘*’** shows the significant variables at 10% level of significance.

The best model is: **ROA = 0.086278 + 4.96E-10 (DEP)**.

Table (5) shows pair wise results. All the three models are significantly affecting ROA but individually no variable is significant.

Table 5
All possible regressions with two predictor variables

Predictors	F-Stat	individually Significant variable	P-Value	R-Squared	Durbin-Watson Statistic
BS, CR*	7.88E+00	NIL	0	0.432558	1.643372
BS, NONII	7.89388	NIL	0	0.428880	1.651717
CR, NONII	7.895721	NIL	0	0.424304	1.647568

Note: **‘*’** shows best model among all.

The best model is: **ROA = 0.202947 + 2.61E-11 (BS) + 0.000785 (CR)**.

Table (6) shows result for regression model with three variables. In this case, the entire model is significantly affecting ROA but individually no variable is significant.

Table 6
Final regressions with three predictor variables

Variable	Coefficient	Standard Error	t-statistic	Prob.
Constant	0.202593	0.022855	8.864361	0.0000
BS	2.92E-11	1.16E-10	0.252092	0.8012
CR	0.000796	0.002962	0.268941	0.7882
NONII	-3.95E-10	1.77E-08	-0.022342	0.9822

The final model is:

ROA = 0.202593 + 2.92E-11 (BS) + 0.000796 (CR) – 3.95E-10 (NONII).

In all the regression models, Durbin-Watson Statistic is suggesting no autocorrelation problem. Non-interest income is most of the times negatively affecting the profitability (ROA) of domestic commercial banks of Pakistan.

6. SUMMARY AND CONCLUSIONS

This study is all about analyzing the important and significant factors, affecting the profitability of domestic commercial banks of Pakistan. For this purpose, 15 banks were selected from 29 commercial banks of Pakistan by simple random sampling using Goldfish bowl method. The study involves panel data for the period 2005-2009 (quarterly bank level data). The internal factors affecting the profitability are six (bank-specific variables). Return on Assets (ROA) is used to measure profitability. All the analysis is carried out by using the statistical software “Eviews”.

From the quarterly financial statements of each selected bank, data for six variables (internal factors) is collected for further analysis. These six variables are Bank Size (BS), Deposits (DEP), Interest Income (II), Credit Risk (CR), NON-Interest Income (NONII) and Expense Management (EM).

Various statistical techniques are applied to analyze the collected data. Before running the regressions some preliminary tests were carried out. To test the multicollinearity of data, the correlation matrix given in Table (1), showed high correlation among few variables, thereafter, three variables expense management, interest income and deposits were removed from the panel regression model. Stationarity of data is tested with the help of "unit root test", because the panel data is obtained by combining the cross-sectional data with time series data.

The technique of "All possible regressions" is used to estimate the impact of each individual predictor variable, pair wise predictor variables and three predictor variables on ROA. Three variables out of six predictor variables were found individually significantly affecting the profitability of commercial bank of Pakistan. These variables are bank size, deposits and non interest income. The analysis suggests that three variables: bank size, deposits and non-interest income have individual significant affect on the profitability of commercial bank of Pakistan.

All the three panel regression models of pair wise bank specific variables are significantly affecting the ROA. Co-efficients of bank size and credit risk are positive while co-efficient of non-interest income is negative in all the three regressions. The panel regression model with all three selected variables is also significantly affecting ROA. Co-efficients of bank size and credit risk are positive while co-efficient of non-interest income is negative in this model.

In all the regression models Durbin-Watson statistic is suggesting no autocorrelation problem. Non-interest income is most of the times negatively affecting the profitability (ROA) of domestic commercial banks of Pakistan.

7. REFERENCES

1. Anna P.I.V. and Hoi Si Chan (2009). Determinants of Bank Profitability in Macao. *Macau Monetary Research Bulletin*, 93-113.
2. Baltagi, B.H. (2001). *Econometric Analysis of Panel Data* (second edition). John Wiley & Sons, Chichester.
3. Baltagi, B.H. (2005). *Econometric Analysis of Panel Data* (third edition). John Wiley & Sons, Chichester.
4. Gujarati, D. (2004). *Basic Econometrics* (4th edition). New York: McGraw-Hill.
5. Indranarain Ramlall (2009). Bank-Specific, Industry-Specific and Macroeconomic Determinants of Profitability in Taiwanese Banking System: Under Panel Data Estimation, International Research Journal of Finance and Economics, Issue 34, 160-167.
6. Kyriaki Kosmidou, Sailesh Tanna and Fotios Pasiouras (2005). Determinants of Profitability of Domestic UK Commercial Banks: Panel Evidence from the Period 1995-2002, Money Macro and Finance Research Group, 37th conference, 1-27.
7. Karunakar M., K. Vasuki and S. Saravanan (2008). Are Non - Performing Assets Gloomy or Greedy From Indian Perspective? *Research Journal of Social Sciences*, 3, 4-12.
8. Maddala, G.S. and Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test, *Oxford Bulletin of Economics and Statistics*, 631-652.
9. Panayiotis P.A. Sophocles N.B. and Matthaios D.D. (2005). Bank-Specific, Industry-Specific and Macroeconomic Determinants of Bank Profitability, Working paper No. 25, Bank of Greece 5-35.
10. Valentina Flamini, Calvin McDonald and Liliana Schumacher (2009). The Determinants of Commercial Bank Profitability in Sub-Saharan Africa, IMF working Paper, WP/09/15, 2-32.

APPENDIX-A

Scheduled Domestic Banks Operating in Pakistan, as on 30th June, 2010

S#	Name of Bank	Branches	Website
A	**Public Sector Commercial Banks**	**1621**	
1	First Women Bank Ltd.	39	www.fwbl.com.pk
2	National Bank of Pakistan^	1267	www.nbp.com.pk
3	The Bank of Khyber	42	www.bok.com.pk
4	The Bank of Punjab ^	273	www.bop.com.pk
B	**Local Private Commercial Banks**	**6,850**	
1	Allied Bank Ltd.	786	www.abl.com.pk
2	Arif Habib Bank Ltd.* ^	36	www.summitbank.com.pk
3	Askari Bank Ltd.^	204	www.askaribank.com.pk
4	Atlas Bank Ltd.* ^	40	www.atlasbank.com.pk
5	Bank Al-Falah Ltd.^	309	www.bankalfalah.com
6	Bank Al-Habib Ltd.^	267	www.bankalhabib.com
7	BankIslami Pakistan Ltd	70	www.bankislami.com.pk
8	Dawood Islamic Bank Ltd.	42	www.dawoodislamic.com
9	Dubai Islamic Bank Pakistan Ltd	36	www.dibpak.com
10	Emirates Global Islamic Bank Ltd.	58	www.egibl.com
11	Faysal Bank Ltd.	136	www.faysalbank.com.pk
12	Habib Bank Ltd.^	1457	www.habibbankltd.com
13	Habib Metropolitan Bank Ltd.^	120	www.hmb.com.pk
14	JS Bank Ltd. ^	40	www.jsbl.com
15	KASB Bank Ltd.^	70	www.kasbbank.com
16	MCB Bank Ltd.^	1085	www.mcb.com.pk
17	Meezan Bank Ltd.	180	www.meezanbank.com
18	Mybank Ltd. ^	80	www.mybankltd.com
19	NIB Bank Ltd. ^	204	www.nibpk.com
20	Samba Bank Ltd.	28	www.samba.com.pk
21	Silk Bank Ltd.	85	www.silkbank.com.pk
22	Soneri Bank Ltd.	156	www.soneri.com
23	Standard Chartered Bank Ltd.	162	www.standardchartered.com
24	The Royal Bank of Scotland Ltd.	79	pwkww.rbs.com.pk
25	United Bank Ltd. ^	1120	www.ubl.com.pk

Ref. website: www.osec.ch

* Since December 2010, Atlas Bank Ltd. and Arif Habib Bank Ltd. have been merged and formed Summit Bank Ltd.
"^" indicates the banks used in this study.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 277-284

A GENERAL CLASS OF REGRESSION TYPE ESTIMATORS WHEN AUXILIARY VARIABLE IS AN ATTRIBUTE

Giancarlo Diana[1], **Saba Riaz**[2] and **Javid Shabbir**[2]

[1] Department of Statistical Sciences, University of Padova (Italy)
Email: sabaqau@gmail.com; diana@stat.unipd.it

[2] Department of Statistics, Quaid-i-Azam University,
Islamabad, Pakistan. Email: jsqau@yahoo.com

ABSTRACT

In the following paper, we are taking motivation from Rao (1991) and Diana et al. (2011) and using their idea of biased estimators, we are proposing two general classes of biased estimators to estimate the finite population mean $\overline{Y}$ with known population proportion of an auxiliary variable. Mathematical calculation has been done and linear regression estimator is used for the efficiency comparison, numerical results are displayed in Table 1. It is shown that the proposed classes are more efficient than Naik & Gupta (1996), Jhajj et al. (2006), Singh et al. (2008) and Koyuncu (2012).

1. INTRODUCTION

In sample surveys it is well known that the use of auxiliary information increases the efficiency when we estimate an unknown population parameter. In many practical situations, there exist some auxiliary attributes ϕ (say) which are highly correlated with the study variable Y such as, weight of the persons (y) and sex (ϕ), amount of milk produced (y) and a particular breed of the cow (ϕ) etc. Many authors have suggested estimators based on information about auxiliary attributes in a simple random sampling or two phase sampling see Naik& Gupta (1996), Jhajj et al. (2006), Singh et al. (2008), Shabbir & Gupta (2007, 2010), Abd-Elfattah et al. (2010)and Koyuncu (2012).

Let, we have N distinct finite population units and a sample of size n is taken by simple random sampling without replacement (SRSWOR) from the N population units. Let Y be the study variable having values y_i, $i=(1,\ldots,N)$. Let ϕ denotes the auxiliary attribute having values ϕ_i, $i=(1,\ldots,N)$.We consider $\phi_i=1$, if the i^{th} unit of the population possesses attribute ϕ and $\phi_i=0$, otherwise. Let, $A=\sum_{i=1}^{N}\phi_i$ and $a=\sum_{i=1}^{n}\phi_i$ denote the total number of units in the population and in the sample possessing attribute ϕ. Let, $\overline{Y}=\frac{\sum_{i=1}^{N}y_i}{N}$ be the unknown population mean and $\overline{y}=\frac{\sum_{i=1}^{n}y_i}{n}$ be the sample mean of the study variable Y. Also we have,

$$C_y^2 = \frac{S_y^2}{\bar{Y}^2},\ S_y^2 = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2}{N-1}, \quad \theta = \left(\frac{1}{n} - \frac{1}{N}\right)$$

$$C_p^2 = \frac{S_\phi^2}{P^2},\ S_\phi^2 = \frac{\sum_{i=1}^{N}\left(\phi_i - P\right)^2}{N-1},$$

$$\rho_{pb} = \frac{S_{y\phi}}{S_y S_\phi},\ S_{y\phi} = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(\phi_i - P\right)}{N-1}, \quad C_{yp} = \rho_{pb} C_y C_p$$

The variance of the usual unbiased estimator $\bar{y}$ under SRSWOR is

$$Var\left(\bar{y}\right) = \theta \bar{Y}^2 C_y^2 \tag{1}$$

Naik & Gupta (1996) proposed the following ratio estimator for $\bar{Y}$

$$\bar{y}_{NG} = \bar{y}\left(\frac{P}{\hat{P}}\right) \tag{2}$$

The MSE of the $\bar{y}_{NG}$ to the first order of approximation is

$$MSE\left(\bar{y}_{NG}\right) = \theta \bar{Y}^2\left(C_y^2 + C_p^2 - 2C_{yp}\right) \tag{3}$$

Jhajj et al. (2006) have suggested the following class of estimators

$$\bar{y}_{JSG} = h\left(\bar{y}, u\right) \tag{4}$$

where, $u = \frac{\hat{P}}{P}$ and $h\left(\bar{y}, u\right)$ is a function that satisfy regularity conditions.

The min MSE of the proposed class to the first order of approximation is

$$\min MSE\left(\bar{y}_{JSG}\right) = \theta \bar{Y}^2 C_y^2\left(1 - \rho_{pb}^2\right) \tag{5}$$

which is equal to that of the linear regression estimator defined as

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}_\phi\left(P - \hat{P}\right) \tag{6}$$

where, $\hat{\beta}_\phi = \frac{s_{y\phi}}{s_\phi^2}$ is the sample counterpart of the population regression coefficient $\beta_\phi = \frac{S_{y\phi}}{S_\phi^2}$.

Singh et al. (2008) suggested the regression-cum-ratio type estimators using known parameters of the auxiliary attribute as

$$\bar{y}_{SCSS1}=\left[\bar{y}+\hat{\beta}_{\phi}\left(P-\hat{P}\right)\right]\left(\frac{P+\eta}{\hat{P}+\eta}\right) \tag{7}$$

$$\bar{y}_{SCSS2}=\left[\bar{y}+\hat{\beta}_{\phi}\left(P-\hat{P}\right)\right]\left(\frac{\eta P+\psi}{\eta\hat{P}+\psi}\right) \tag{8}$$

where, η and ψ are either real numbers or functions of the known parameter associated with an auxiliary attribute such as C_p, $\beta_2(\phi)$, and ρ_{pb}.

The MSE of the ratio estimator up to first order approximation is

$$MSE\left(\bar{y}_{SCSSi}\right)=\theta\left[S_y^2\left(1-\rho_{pb}^2\right)+\tau_i^2S_{\phi}^2\right] \tag{9}$$

where, $\tau_1=\dfrac{\bar{Y}}{P+\eta}$, $\tau_2=\dfrac{\eta\bar{Y}}{\eta P+\psi}$, $i=1,2$.

Koyuncu (2012) suggested the following class of estimator

$$\bar{y}_k=\left[w_1\bar{y}+w_2\left(P-\hat{P}\right)\right]\left(\frac{\eta P+\psi}{\eta\hat{P}+\psi}\right) \tag{10}$$

where, w_1 and w_2 are suitable weights, η and ψ are either real numbers or the functions of the known parameter associated with an auxiliary attribute.

The min MSE of $\bar{y}_k$ to the first order approximation is given by,

$$\min MSE\left(\bar{y}_k\right)=\frac{\left(1-\theta\tau^2C_p^2\right)MSE\left(\bar{y}_{reg}\right)}{\left(1-\theta\tau^2C_p^2\right)+\dfrac{MSE\left(\bar{y}_{reg}\right)}{\bar{Y}^2}} \tag{11}$$

where, $\tau=\dfrac{\eta P}{\eta P+\psi}$.

2. SUGGESTED CLASSES

By analogy to the approach of Rao (1991) and Diana et al. (2011) we are introducing two different general classes of biased estimators based on an auxiliary attribute.

2.1 CLASS 1

We are taking motivation from Rao (1991) and proposing a general class

$$\bar{y}_{S1}=w_1\bar{y}+w_2\left(P-\hat{P}\right) \tag{12}$$

where w_1 and w_2 are constants to be chosen properly.

$$\bar{y}_{S1}=w_1\bar{Y}+w_1\bar{Y}\delta_y-w_2P\delta_{\phi} \tag{13}$$

By using δ method, with

$$\delta_y = \frac{\bar{y}-\bar{Y}}{\bar{Y}}, \quad \delta_\phi = \frac{\hat{P}-P}{P}, \quad E\left(\delta_y\right) = E\left(\delta_\phi\right) = 0,$$
$$E\left(\delta_y^2\right) = \theta C_y^2, \quad E\left(\delta_\phi^2\right) = \theta C_p^2 \text{ and } E\left(\delta_y \delta_\phi\right) = \theta C_{yp} = \theta \rho_{pb} C_y C_p$$

It is easy to write the Bias and the MSE,

$$Bias\left(\bar{y}_{S1}\right) = \bar{Y}\left(w_1 - 1\right) \tag{14}$$

$$MSE\left(\bar{y}_{S1}\right) = \bar{Y}^2\left(w_1 - 1\right)^2 + \theta\left[w_1^2 \bar{Y}^2 C_y^2 + w_2^2 P^2 C_p^2 - 2 w_1 w_2 \bar{Y} P C_{yp}\right] \tag{15}$$

The MSE of the proposed class will be minimum when,

$$w_1^* = \frac{C_p^2}{C_p^2 + \theta\left(C_y^2 C_p^2 - C_{yp}^2\right)}, \quad w_2^* = \frac{\bar{Y} C_{yp}}{P\left[C_p^2 + \theta\left(C_y^2 C_p^2 - C_{yp}^2\right)\right]}$$

and min MSE of $\bar{y}_{S1}$ is

$$\min MSE\left(\bar{y}_{S1}\right) = \frac{\theta \bar{Y}^2\left(C_y^2 C_p^2 - C_{yp}^2\right)}{C_p^2 + \theta\left(C_y^2 C_p^2 - C_{yp}^2\right)}$$

(16)

Also,

$$\min MSE\left(\bar{y}_{S1}\right) = \frac{\theta \bar{Y}^2 C_y^2\left(1 - \rho_{pb}^2\right)}{1 + \theta C_y^2\left(1 - \rho_{pb}^2\right)} \tag{17}$$

which can also be expressed as

$$\min MSE\left(\bar{y}_{S1}\right) = \frac{MSE\left(\bar{y}_{reg}\right)}{1 + \frac{MSE\left(\bar{y}_{reg}\right)}{\bar{Y}^2}} \tag{18}$$

2.2 CLASS 2

Now taking motivation from Diana et al. (2011), we are defining a more general class for population mean $\bar{Y}$, also considering a function g as

$$\bar{y}_{S2} = \left(w_1 \bar{y} + w_2 u\right) g\left(u\right) \tag{19}$$

where, $u = P - \hat{P}$ and g is a generic function that satisfy mild conditions.

Expanding $g\left(u\right)$ in Taylor's series up to including terms which are $O_p\left(u^2\right)$, the resulting expression will be

$$\bar{y}_{S2} \cong \left(w_1 \bar{y} + w_2 u\right)\left[g\left(0\right) + g'\left(0\right) u + \frac{1}{2} g''\left(0\right) u^2\right] \tag{20}$$

where $g(0)$ is a constant term, $g'(0)$ is the first order partial derivative in zero and $g''(0)$ is second order partial derivative in zero. For simplicity we can write $g(0)=a_0$, $g'(0)=b_0$, $\frac{1}{2}g''(0)=c_0$.

$$\overline{y}_{S2} \cong \left[w_1\overline{Y}+w_1\overline{Y}\delta_y-w_2P\delta_\phi\right]\left[a_0-b_0P\delta_\phi+c_0P^2\delta_\phi^2\right] \tag{21}$$

where, $\delta_{\bullet}$ are defined as in subsection (2.1).

Therefore,

$$\overline{y}_{S2} \cong w_1\overline{Y}\left[a_0-b_0P\delta_\phi+c_0P^2\delta_\phi^2\right]+w_1\overline{Y}\delta_y\left[a_0-b_0P\delta_\phi\right]-w_2P\delta_\phi\left[a_0-b_0P\delta_\phi\right] \tag{22}$$

The Bias and MSE of $\overline{y}_{S2}$ can be written to the first order of approximation as

$$Bias\left(\overline{y}_{S2}\right)=\overline{Y}\left(a_0w_1-1\right)+\theta\left[\left(w_1\overline{Y}c_0+w_2b_0\right)P^2C_p^2-w_1\overline{Y}b_0PC_{yp}\right] \tag{23}$$

and

$$MSE\left(\overline{y}_{S2}\right)=\overline{Y}^2\left(a_0w_1-1\right)^2+\theta\begin{bmatrix} w_1^2a_0^2\overline{Y}^2C_y^2-2w_1\overline{Y}PC_{yp}\left(w_2a_0^2+2w_1a_0b_0\overline{Y}-b_0\overline{Y}\right) \\ +P^2C_p^2\begin{Bmatrix} w_2^2a_0^2+2w_1a_0\overline{Y}\left(2w_2b_0+w_1c_0\overline{Y}\right) \\ +\overline{Y}\left(w_1^2b_0^2\overline{Y}-2w_2b_0-2w_1c_0\overline{Y}\right) \end{Bmatrix}\end{bmatrix} \tag{24}$$

Minimizing $MSE\left(\overline{y}_{S2}\right)$ to achieve the optimum values of the constants w_1 and w_2

$$w_1^*=\frac{C_p^2\left[a_0^2+\theta\left(a_0c_0P^2C_p^2-2b_0^2P^2C_p^2\right)\right]}{a_0\left[a_0^2\left\{C_p^2+\theta\left(C_y^2C_p^2-C_{yp}^2\right)\right\}+\theta P^2C_p^4\left(2a_0c_0-3b_0^2\right)\right]}$$

$$w_2^*=\frac{\overline{Y}\begin{bmatrix} a_0^3C_{yp}+a_0^2P\left\{b_0\left(\theta C_y^2C_p^2-\theta C_{yp}^2+C_p^2\right)+\theta c_0PC_p^2C_{yp}\right\} \\ +\theta b_0^3P^3C_p^4-2\theta a_0b_0^2P^2C_p^2C_{yp}\end{bmatrix}}{a_0^2\overline{Y}\left[a_0^2\left\{C_p^2+\theta\left(C_y^2C_p^2-C_{yp}^2\right)\right\}+\theta P^2C_p^4\left(2a_0c_0-3b_0^2\right)\right]}$$

$$\min MSE\left(\overline{y}_{S2}\right)=\frac{\theta\overline{Y}^2\begin{bmatrix} a_0^4C_y^2\left(1-\rho_{pb}^2\right)-\theta a_0^2P^2C_p^2\left\{b_0^2C_y^2\left(1-\rho_{pb}^2\right)+c_0^2P^2C_p^2\right\} \\ +\theta b_0^2P^4C_p^4\left(2a_0c_0-b_0^2\right)\end{bmatrix}}{a_0^2\left[a_0^2\left\{1+\theta C_y^2\left(1-\rho_{pb}^2\right)\right\}+\theta P^2C_p^2\left(2a_0c_0-3b_0^2\right)\right]} \tag{25}$$

For comparison reasons, we can write $\min MSE\left(\bar{y}_{S2}\right)$ as

$$\min MSE\left(\bar{y}_{S2}\right)=\frac{a_0^4 MSE\left(\bar{y}_{reg}\right)-\theta a_0^2 P^2 C_p^2\left\{b_0^2 MSE\left(\bar{y}_{reg}\right)+\theta c_0^2 \bar{Y}^2 P^2 C_p^2\right\}+\theta^2 b_0^2 \bar{Y}^2 P^4 C_p^4\left(2a_0c_0-b_0^2\right)}{a_0^2\left[a_0^2\left\{1+\frac{MSE\left(\bar{y}_{reg}\right)}{\bar{Y}^2}\right\}+\theta P^2 C_p^2\left(2a_0c_0-3b_0^2\right)\right]} \qquad (26)$$

The proposed class $\bar{y}_{S2}$ depends on the choice of function g. Many possible choices can be considered theoretical and practical point of view. Before going for the selection of function g, first we evaluate the efficiency of the considered estimator compared to $\bar{y}_{reg}$. It is well known that linear regression estimator $\bar{y}_{reg}$ is always more efficient than $\bar{y}$ and for this reason we have chosen it as competitor.

For the first proposed class,

$$MSE\left(\bar{y}_{reg}\right)-\min MSE\left(\bar{y}_{S1}\right)\geq 0$$

This result is always true, since

$$MSE\left(\bar{y}_{reg}\right)>0$$

For the second proposed class, after some mathematical computation we get

$$MSE\left(\bar{y}_{reg}\right)-\min MSE\left(\bar{y}_{S2}\right)=\frac{\left[a_0^4 MSE\left(\bar{y}_{reg}\right)-\theta \bar{Y}^2 P^2 C_p^2\left(b_0^2-a_0c_0\right)\right]^2}{a_0^2\left[a_0^2\left\{\bar{Y}^2+MSE\left(\bar{y}_{reg}\right)\right\}+\theta \bar{Y}^2 P^2 C_p^2\left(2a_0c_0-3b_0^2\right)\right]}$$

This expression will be certainly positive if $\left(2a_0c_0-3b_0^2\right)\geq 0$ is satisfied then our proposed class is more efficient than the linear regression estimator.

For simplification we can also write the condition as

$$c_0\geq\frac{3b_0^2}{2a_0} \qquad (27)$$

Consider an exponential function g in which assuming that the information about the proportion P of population units possessing the auxiliary attribute ϕ is known in advance.

$$\bar{y}_1^*=\left(w_1\bar{y}+w_2u\right)g\left(u\right) \qquad (28)$$

where,

$$g\left(u\right)=\exp\left(\frac{u}{2P-u}\right)$$

when we expand (28) by Taylor's theorem, we get

$$a_0 = g(0) = 1, \quad b_0 = g'(0) = \frac{1}{2P} \text{ and } c_0 = \frac{1}{2}g''(0) = \frac{3}{8P^2}$$

We can see that the condition $c_0 \geq \frac{3b_0^2}{2a_0}$ is satisfied here.

3 NUMERICAL STUDY AND COMPARISONS

In this section we compare the efficiency of the proposed classes with linear regression estimator using two population data sets as considered before by Shabbir & Gupta (2007), Abd-Elfattah et al. (2010) and Koyuncu (2012).

Population I [Source: Sukhatme & Sukhatme (1970), p. 256]

y = Number of villages in the circles.

ϕ = A circle consisting of more than five villages.

$N = 89$, $n = 23$, $\bar{Y} = 3.36$, $P = 0.124$,

$C_y = 0.601$, $C_p = 2.678$, $\rho_{pb} = 0.766$, $\beta_2(\phi) = 6.612$

Population II [Source: Sukhatme & Sukhatme (1970), p. 256]

y = Area (in acres) under the wheat crop within the circles.

ϕ = A circle consisting of more than five villages.

$N = 89$, $n = 23$, $\bar{Y} = 1102$, $P = 0.124$,

$C_y = 0.65$, $C_p = 2.678$, $\rho_{pb} = 0.624$, $\beta_2(\phi) = 6.612$

Here we are taking $\eta = C_p$ and $\psi = \beta_2(\phi)$ to estimate $\bar{y}_{SCSS1}$, $\bar{y}_{SCSS2}$ and $\bar{y}_k$.

Table 1
PRE of the estimators with respect to $\bar{y}_{reg}$.

Estimator	Efficiency (Pop I)	Efficiency (Pop II)
$\bar{y}_{reg}$	100	100
$\bar{y}_{NG}$	2.94	4.77
$\bar{y}_{JSG}$	100	100
$\bar{y}_{SCSS1}$	91.44	94.88
$\bar{y}_{SCSS2}$	88.67	93.14
$\bar{y}_k$	100.48	100.83
$\bar{y}_{S1}$	100.48	100.83
$\bar{y}_{S2}$	130.78	119.76

The comparison is performed in terms of Percent Relative Efficiency (PRE)

$$PRE\left(\bar{y}_{(*)}\right)=\frac{MSE\left(\bar{y}_{reg}\right)}{\min MSE\left(\bar{y}_{(*)}\right)}\times 100$$

In Table 1, we can see that the estimators proposed by Naik & Gupta (1996) and Singh et al. (2008) are less efficient than linear regression estimator, but Jhajj et al. (2006) proposed class is equivalent to linear regression estimator, see (5). Also from (5) and (9), it is clearly shown that $MSE\left(\bar{y}_{SCSSi}\right)$ is greater than $MSE\left(\bar{y}_{reg}\right)$. From (5), (11) and (18) we can conclude that $\bar{y}_k$ and $\bar{y}_{S1}$ are more efficient than $\bar{y}_{reg}$. It can also be noted from numerical study that $\bar{y}_k$ and $\bar{y}_{S1}$ are showing almost same results which concludes that there may be no difference with or without using ratio type estimator with biased regression estimator. The second proposed class is showing better performance in all other estimators. If we see the efficiency of $\bar{y}_k$ and $\bar{y}_{S2}$, we can conclude that exponential function may be a good choice for function g as it fulfills the optimal condition (28) whereas ratio and regression-cum-ratio estimators may not fulfill this condition see Singh et al. (2008) and Koyuncu (2012).

REFERENCES

1. Abd-Elfattah, A.M., El-Sherpieny, E.A., Mohamed, S.M. and Abdou, O.F. (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. *App. Math. and Comput.*, 215, 4198-4202.
2. Diana, G., Giordan, M., and Perri, P.F. (2011). An improved class of estimators for the population mean. *Statistical Methods and Applications*, 20, 123-140.
3. Jhajj, H.S., Sharma, M.K. and Grover, L.K. (2006). A family of estimators of population mean using information on auxiliary attribute. *Pak. J. Statist.*, 22, 43-50.
4. Koyuncu, N. (2012). Efficient estimators of population mean using auxiliary attributes. *Applied Mathematics and Computation,* 218, 10900-10905.
5. Naik, V.D., and Gupta, P.C. (1996). A note on estimating of mean with known population of an auxiliary character. *Journal of Indian Society of Agricultural Statistics*, 48, 151-158.
6. Rao, T.J. (1991). On certain methods of improving ratio and regression estimators. *Commun. in Statist.-Theo. and Meth.,* 20, 3325-3340.
7. Shabbir, J. and Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pak. J. Statist.*, 23, 1-9.
8. Shabbir, J. and Gupta, S. (2010). Estimation of the finite population mean in two-phase sampling when auxiliary variables are attributes. *Hacettepe Journal of Mathematics and Statistics*, 39, 121-129.
9. Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. (2008). Ratio estimators in simple random sampling using information on auxiliary attribute. *Pakistan Journal of Statistics and Operation Research*, 4, 47-53.
10. Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. Asia Publishing House, New Delhi, India.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 285-296

SOME PROPERTIES OF GENERALIZED LOG-PEARSON DISTRIBUTION

Zafar Iqbal and **Munir Ahmad**
National College of Business Administration and Economics, Lahore, Pakistan
Email: zafariqbal75@yahoo.com; drmunir@ncbae.edu.pk

ABSTRACT

In this paper, a Generalized Log-Pearson Distribution (GLPD) is proposed. Log – Pearson VII Distribution derived by Habibullah (2010) is the special case of GLPD. Some properties of new distribution are investigated. Characterization of GLPD is presented through conditional expectation. Finally, some Compound Mixtures for GLPD are shown.

1. INTRODUCTION

In past century, Pearson probability distributions have involved the attention of researchers in all areas of study. Pearson distributions determine applications in engineering sciences, biological, econometrics, survey sampling and in life-testing. The Pearson system is initially worked out in an effort to model evidently skewed observations.

In 1895 Pearson mentions four types of distributions (numbered I through IV), moreover the normal distribution which is actually known as type V.

Some distributions are recognized in statistics to have the property of reciprocal symmetry. Gumbel and Keeney (1950) and Seshadri (1965) appear to be the only two research works that examine the intrinsic characteristics of such distributions.

Ahmad (1985a) discusses the inverted class of distributions. Ali and Ahmad (1985) present some properties of the inverted inverse Gaussian distribution. Ahmad and Sheikh (1986) discuss a two-parameter Bernstein probability distribution. Ahmad and Kazi (1987) discuss the moments of the Bernstein reliability model. Habibullah (1987) investigates the modes of the inverted bivariate normal distribution.

Some researchers introduce Log Pearson Type III and their applications in real data. Singh and Singh (1988) give an idea of the principle of maximum entropy to derive an alternative method of parameter estimation of Log Pearson Type III Distribution and compare it with method of moments and method of maximum likelihood estimation.

In the United States, the Log-Pearson III is the default distribution for flood frequency analysis. Pilon and Adamowski (1993) develop the Log Likelihood function of Log Pearson Distribution Type III. Cheng et al (2006) present a frequency factor based method for random generation of five distribution (normal, lognormal, extreme value type 1, Pearson Type III and Log Pearson Type III) commonly used in Hydrological frequency analysis.

Habibullah (2010) introduces Log Pearson VII Distribution as

$$f(x)=\frac{k}{x}\left(1+\alpha\left(\ln x\right)^{2}\right)^{-\nu},k=\frac{\alpha^{\frac{1}{2}}}{\beta\left(\nu-\frac{1}{2},\frac{1}{2}\right)},\ x>0,\alpha>0,\nu>\frac{1}{2}. \tag{1.1}$$

She (2010) characterizes it through hazard rate function. In literature there is no further work seems on this distribution till today.

In this research we will propose a Generalized Log Pearson (GLPD) Type VII Distribution as

$$f(x)=\frac{k}{x}\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu},k=\frac{p\alpha^{\frac{1}{2p}}}{\beta\left(\nu-\frac{1}{2p},\frac{1}{2p}\right)},$$

$$x>0,\alpha>0,p\geq 1,\nu>\frac{1}{2p} \tag{1.2}$$

which is more flexible than Log Pearson of Type VII Distribution in data fitting. The goal of curve fitting is to find out the parameter values that most closely match the data. So different choices of combinations of parameters make GLPD Type VII more useful in curve fitting than Log Pearson VII. GLPD Type VII is closed under inversion which produces many distributions after replacing some special transformations.

In this paper, a Generalized Log-Pearson Distribution (GLPD) is proposed. Some properties of new distribution are investigated. Characterization of GLPD is presented through conditional expectation. Some Compound Mixtures for GLPD are shown.

2. DIFFERENTIAL EQUATION FOR GENERATING GENERALIZED LOG PEARSON DISTRIBUTION

Pearson (1895) notes that, in the limiting case, the hyper geometric distribution can be expressed in the form $\frac{df}{dx}=\frac{\left(x-a\right)f}{b_0+b_1x+b_2x^2}$ and utilizes this fact to obtain the Pearson system of continuous distribution functions.

Cobb (1980) discusses a differential equation of the form $\frac{df}{dx}=\frac{g\left(y\right)}{h\left(y\right)}f\left(y\right)dy$, $h\left(y\right)>0$ for all values of y in the domain of f, where $g\left(y\right)$ and $h\left(y\right)$ are polynomials such that the degree of $h\left(y\right)$ is one higher than the degree of $g\left(y\right)$. Ahmad (1985) uses $\frac{d}{dx}\left[\ln g\left(x\right)\right]=-\frac{\left(a_2x^2+a_1x+a_0\right)}{x\left(B_0x^2+B_1x+B_2\right)}$ where the coefficients a_0, a_1, a_2 are given by $a_0=2B_2-1$, $a_1=2B_2-a$, $a_2=2B_0$, and generates the Inverted Pearson

System of probability distributions. This class includes the Inverted Normal as well as the Type I, II, III, V Inverted distribution. Habibullah et al. (2009) use differential equation (3.1) with constraints in (3.2) and differential equation (3.3) with constraints (3.4) to provide a general class of the distributions which are strictly closed under inversion.

Theorem 2.1

If we replace

$$n = 2p, b_i = 0, i = 0,1,2....2p., b_i \neq 0, i = 2p-1,$$
$$a_j = 0, j = 1,2,....,2p-1, a_0 \neq 0, a_{2p} \neq 0$$

in (3.1) differential equation we have a density function of

$$f(x) = \frac{k}{x}\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu}, x > 0, \alpha > 0, p \geq 1, \nu > \frac{1}{2p}.$$

Proof:

The differential equation

$$\frac{d\left[\ln g\left(y\right)\right]}{dy} = \frac{\sum_{i=0}^{2p} b_i y^i}{\sum_{j=0}^{2p} a_j y^j}, \quad b_i = 0, i = 0,1,2....2p., b_i \neq 0, i = 2p-1,$$
$$a_j = 0, j = 1,2,....,2p-1, a_0 \neq 0, a_{2p} \neq 0$$

$$g\left(y\right) = k\left(a_0 + a_{2p}\left(y\right)^{2p}\right)^{-\nu}, \ \nu = \frac{-b_{2p-1}}{a_{2p}.2p}$$

Let $\ln x = y$

$$f(x) = \frac{k}{x}\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu}, k = \frac{p\alpha^{\frac{1}{2p}}}{\beta\left(\nu-\frac{1}{2p}, \frac{1}{2p}\right)},$$
$$x > 0, \alpha > 0, p \geq 1, \nu > \frac{1}{2p} \tag{2.1}$$

as required.

2.1 Graph of GLPD

The following graphs are shown for different values of parameters.

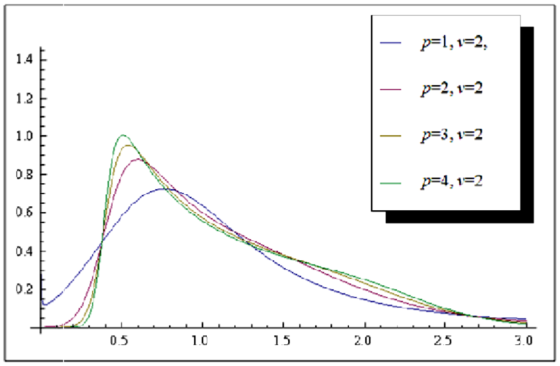
Case 1

i) α =1, *p*=1, *v*=2,

ii) α =1, *p*=2, *v*=2

iii) α =1, *p*=3, *v*=2

iv) α =1, *p*=4, *v*=2



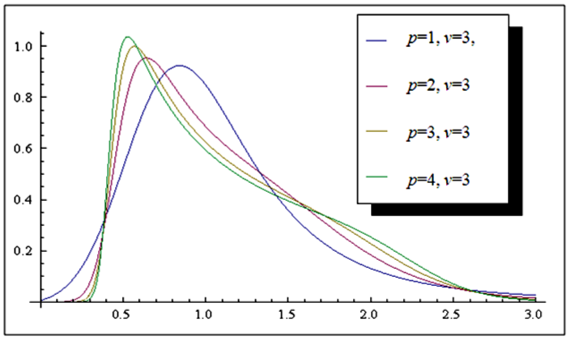
Case 2

i) α =1, *p*=1, *v*=3,

ii) α =1, *p*=2, *v*=3

iii) α =1, *p*=3, *v*=3

iv) α =1, *p*=4, *v*=3



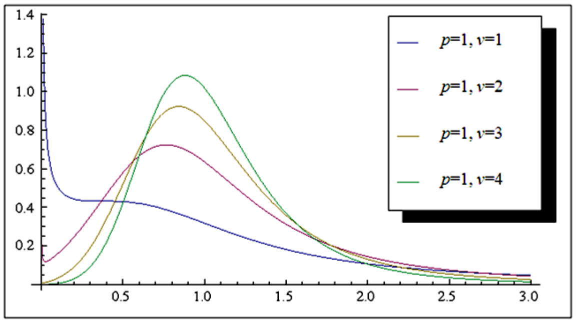
Case 3:

i) α =1, *p*=1, *v*=1,

ii) α =1, *p*=1, *v*=2

iii) α =1, *p*=1, *v*=3

iv) α =1, *p*=1, *v*=4



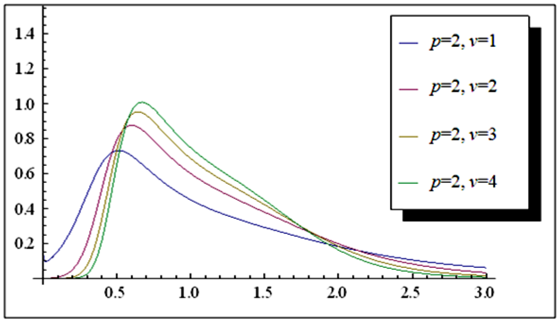
Case 4

i) α =1, *p*=2, *v*=1,

ii) α =1, *p*=2, *v*=2

iii) α =1, *p*=2, *v*=3

iv) α =1, *p*=2, *v*=4



2.2 Distribution Function of GLPD and its Graph

Distribution function of a density function is defined as

$$F\left(x\right)=k\int_{0}^{x}f\left(t\right)\ dt$$

$$F\left(x\right)=k\int_{0}^{x}\frac{\left(1+\alpha\left(\ln t\right)^{2p}\right)^{-\nu}}{t}dt$$

$$\left(1+\alpha\left(\ln t\right)^{2p}\right)=\frac{1}{z},$$

After some simplification, we have

$$F(x)=\begin{Bmatrix}\frac{1}{2}I_{\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-1}}\left(\nu-\frac{1}{2p},\frac{1}{2p}\right), & if \quad x\leq 1\\ 1-\frac{1}{2}I_{\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-1}}\left(\nu-\frac{1}{2p},\frac{1}{2p}\right), & if \quad x>1\end{Bmatrix} \tag{2.2}$$

where $I_x\left(a,b\right)=\frac{1}{B\left(a,b\right)}\int_{0}^{x}t^{a-1}\left(1-t\right)^{b-1}dt$.

2.3 The Hazard Rate of GLPD

The hazard rate of the function is defined as

$$h\left(x\right)=\begin{Bmatrix}\frac{2.k.\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu}}{x.I_{\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-1}}\left(\nu-\frac{1}{2p},\frac{1}{2p}\right)} & if \quad x\leq 1\\ \frac{k.\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu}}{x.-\frac{x}{2}I_{\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-1}}\left(\nu-\frac{1}{2p},\frac{1}{2p}\right)} & if \quad x>1\end{Bmatrix}, \tag{2.3}$$

$$h\left(x\right)=\frac{f\left(x\right)}{S\left(x\right)}$$

2.4 The Mills Ratio of GLPD

The Mills Ratio of the function is defined as

$$m(x) = \frac{S(x)}{f(x)}$$

$$h(x) = \left\{ \begin{array}{ll} \dfrac{x.I_{\left(1+\alpha(\ln x)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)}{2.k.\left(1+\alpha(\ln x)^{2p}\right)^{-\nu}} & if\ x \le 1 \\ \dfrac{2.x. - x.I_{\left(1+\alpha(\ln x)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)}{2.k.\left(1+\alpha(\ln x)^{2p}\right)^{-\nu}} & if\ \ x > 1 \end{array} \right\}, \tag{2.4}$$

2.5 Mean Failure Rate Function of GLPD

The Mean Failure Rate function is defined as

$$H(x) = \int_0^x h(x)dx$$

$$H(x) = \left\{ \begin{array}{ll} \displaystyle\int_0^x \frac{2.k.\left(1+\alpha(\ln u)^{2p}\right)^{-\nu} du}{t.I_{\left(1+\alpha(\ln t)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)} & if\ t \le 1 \\ \displaystyle\int_0^x \frac{k.\left(1+\alpha(\ln u)^{2p}\right)^{-\nu}}{t - \frac{t}{2}I_{\left(1+\alpha(\ln t)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)} du & if\ \ t > 1 \end{array} \right\},$$

After some simplification we have

$$H(x) = \left\{ \begin{array}{l} \dfrac{I_{\left(1+\alpha(\ln x)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)}{I_{\left(1+\alpha(\ln t)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)} ift \le 1 \,\&\, x \le 1 \\ \dfrac{2 - I_{\left(1+\alpha(\ln x)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)}{2 - I_{\left(1+\alpha(\ln t)^{2p}\right)^{-1}}\left(\nu - \frac{1}{2p}, \frac{1}{2p}\right)} ift > 1 \,\&\, x > 1 \end{array} \right\}, \tag{2.5}$$

2.6 The Function $\eta(x)$ of GLPD

$$\eta(x) = -\frac{f'(x)}{f(x)}$$

$$= \frac{1+\alpha\ (\ln x)^{2p} + 2pv\alpha(\ln x)^{2p-1}}{x\left(1+\alpha\ (\ln x)^{2p}\right)}$$

2.7 Shannon Entropy of GLPD

The Shannon Entropy $h(X)$ of a continuous random variable X with a density $f(x)$ is defined as

$$h(X) = -\int_S f(x)\ \log f(x)\, dx$$

where S is the support set of random variable.

$$= -\log k + v.k\int_0^\infty \left(\log\left(1+\alpha.(\log x)^{2p}\right)\right).\frac{\left(1+\alpha.(\log x)^{2p}\right)^{-v}}{x}dx$$

Let

$$1+\alpha.(\log x)^{2p} = \frac{1}{t}$$

$$= -\log k + \frac{v.k}{2.p.\alpha^{\frac{1}{2p}}}\int_0^1 \log t.\ t^{v-\frac{1}{2p}-1}.(1-t)^{\frac{1}{2p}-1}\, dt$$

Using the following formula 4.253 from Table of Integral and Series

$$\int_0^1 t^{\mu-1}.(1-t)^{\nu-1}.\log t\ dt = \mathrm{B}(\mu,\nu)\left[\psi(\mu)-\psi(\mu+\nu)\right].$$

$$= -\log k + \frac{v.k}{2.p.\alpha^{\frac{1}{2p}}}\mathrm{B}\left(\nu\text{-}\frac{1}{2p},\frac{1}{2p}\right)\left[\psi\left(\nu\text{-}\frac{1}{2p}\right)-\psi(\nu)\right]$$

Theorem 2.2

Let X be a r.v with differentiable density on its support $(0,\infty)S$.

Then the following conditions are equivalent for characterizing Generalized Log Pearson Distribution

1) $m_k(y) = \dfrac{d(y).r(y)}{2(k+1)\alpha(v-1)}$ (2.6)

2) $\dfrac{f'(y)}{f(y)} = \dfrac{-d'(y) - 2p(\nu-1)\alpha \ln^{2p-1} y}{d(y)}$

where

$$m_k(y) = E\left((\ln x)^k \mid X > y\right)$$
$$d(y) = y.\left(1+\alpha.(\ln y)^{k+1}\right)$$
$$k = 2p-1$$

Proof:

Suppose X follow the Generalized Log Pearson Type VII distribution. Then

$$m_k(y) = E\left((\ln x)^k \mid X > y\right)$$
$$k = 2p-1$$
$$= \frac{C}{\bar{F}(y)} \int_y^\infty \frac{\ln^{2p-1} x}{x} \left(1+\alpha(\ln x)^{2p}\right)^{-\nu} dx \,.$$

Letting

$z = \left[1+\alpha(\ln x)^{2p}\right]^{-1}$ we have

$$m_k(y) = \frac{C\left(1+\alpha \ln^{2p} y\right)^{1-\nu}}{2p\alpha \bar{F}(y)(\nu-1)}$$
$$yf(y) = C\left(1+\alpha \ln^{2p} y\right)^{-\nu}$$

and hence

$$E\left(\ln^{2p-1} X \mid X \geq y\right) = \frac{y\left(1+\alpha \ln^{2p} y\right)}{2p\alpha(q-1)} . \frac{f(y)}{\bar{F}(y)} = \frac{y\left(1+\alpha \ln^{2p} y\right)}{2p\alpha(\nu-1)} .r_X(y)$$
$$m_k(y) = \frac{d(y)}{(k+1)\alpha(\nu-1)} .r_X(y)$$

where

$$d(y) = y.\left(1+\delta.(\ln y)^{k+1}\right)$$

Now, let $f(x)$ be an unknown pdf of a random variable X defined on $(0,\infty)$ such that 5.1 holds.

Equation (5.1) implies that

$$\int_y^\infty \left(\ln^{2p-1} x\right) f(x)\,dx = \frac{1}{2p\alpha(\nu-1)} \left[y\left(1+\alpha \ln^{2p} y\right) f(y)\right] .$$

Differentiating both sides with respect to y , we have

$$-\left(\ln y\right)^{2p-1} f\left(y\right)=\frac{1}{2p\alpha\left(\nu-1\right)}$$
$$\left[\left(1+\alpha\ln^{2p} y\right)\left(y.f'\left(y\right)+f\left(y\right)\right)+2p\alpha\ln^{2p-1} y.f(y)\right]$$

implying that

$$\frac{-\left[1+\alpha\ln^{2p} y\right]-2p\nu\alpha\ln^{2p-1} y}{y\left(1+\alpha\ln^{2p} y\right)}=\frac{f'\left(y\right)}{f(y)}$$

$$\frac{f'\left(y\right)}{f\left(y\right)}=\frac{-d'\left(y\right)-2p\left(\nu-1\right)\alpha\ln^{2p-1} y}{d\left(y\right)}$$

or

$$\frac{d}{dy}\left[\ln f(y)\right]=\frac{-2p\alpha\nu\left(\ln y\right)^{2p-1}}{y\left(1+\alpha.\ln^{2p} y\right)}-\frac{1}{y}$$

Integrating both sides with respect to y , we have

$$\ln f(y)=-\nu\left[\ln\left(1+\alpha.\ln^{2p} y\right)\right]-\ln y+\ln k$$

$$f(x)=\frac{k}{x}\left(1+\alpha\left(\ln x\right)^{2p}\right)^{-\nu},x\succ 0,\alpha\succ 0,p\geq 1,\nu\succ\frac{1}{2p}$$

with normalizing factor, $k=\dfrac{p\alpha^{\frac{1}{2p}}}{\beta\left(\nu-\frac{1}{2p},\frac{1}{2p}\right)}$

so that

$$f\left(x\right)=\frac{p\alpha^{\frac{1}{2p}}}{\beta\left(\nu-\frac{1}{2p},\frac{1}{2p}\right).x}\left(1+\alpha\ln^{2p} x\right)^{-\nu},\ 0<x<\infty,$$

which is the Generalized Log Pearson Type VII distribution where $r_X\left(y\right)=\dfrac{f\left(y\right)}{\bar{F}\left(y\right)},\ \bar{F}\left(y\right)=1-F\left(y\right).$

2.8 Compound scale mixture of Generalized Log Normal Distribution (GLND) and Fractional Moment Exponential Distribution generate Generalized Log Pearson Distribution (GLPD) Type VII.

Theorem (2.3)

Let $f(x)$ be Generalized Log Normal Distribution

$$f(x)=\frac{p.\beta^{\frac{1}{2p}}}{2^{\frac{1}{2p}}\Gamma\left(\frac{1}{2p}\right).x}e^{-\frac{\beta.(\ln x)^{2p}}{2}}\quad ,\ x>0$$

and Fractional Moment Exponential Distribution

$$g(u\,|\,\delta,\alpha)=\frac{u^{\delta}\exp\left(-\frac{u}{\alpha}\right)}{\alpha^{1+\delta}\Gamma(1+\delta)}\quad ,\ u>0$$

Then GLPD is the scale mixture of GLND with Fractional Moment Exponential Distribution where

$$\delta=v-\frac{1}{2p}-1,\alpha=2$$

and GLPD is

$$p(x)=\frac{p}{\beta\left(\nu-\frac{1}{2p},\frac{1}{2p}\right).x}\left(1+(\ln x)^{2p}\right)^{-\nu}\quad x>0,\alpha>0,p\geq 1,\nu>\frac{1}{2p}$$

Proof:

The scale mixture of GLND with Gamma Dist is

$$p(x)=\int_{0}^{\infty}f(x\,|\,u)\;g\left(u\,|\,v-\frac{1}{2p}-1,2\right)du$$

$$p(x)=\int_{0}^{\infty}\frac{p.\exp\left(-\frac{(\log x)^{2p}u}{2}\right)u^{\frac{1}{2p}}}{2^{\frac{1}{2p}}.\Gamma\left(\frac{1}{2p}\right).x}.\frac{u^{v-\frac{1}{2p}-1}.\exp\left(-\frac{u}{2}\right)}{2^{v-\frac{1}{2p}}\Gamma\left(v-\frac{1}{2p}\right)}du$$

$$p(x)=\frac{p}{\beta\left(m-\frac{1}{2p},\frac{1}{2p}\right).x}\left(1+(\log x)^{2p}\right)^{-m}$$

is required.

4. REFERENCES

1. Ahmad, M. (1985a). Inverted Class of Distributions. Proceedings of the *45th Session of the International Statistical Institute*, 1-3, ISI, Netherlands.
2. Ahmad, M. and Kazi, M.H. (1987). On the moments of Bernstein Reliability Model. *Pak. J. Statist.*, 3(1) A, 1-9.
3. Ahmad, M. and Sheikh, A.K. (1981). Reliability Computation for Bernstein Distribution Strength and Stress. Submitted to the *10th Pak. Statistical Conference*, Islamabad, Pakistan.
4. Ahmad, M. and Sheikh A.K. (1986). On a two-parameter Bernstein probability distribution. *J. Natural Science and Mathematics*, 26(2), 19-36.
5. Ali, Z. and Ahmad, M. (1985). Inverted inverse Gaussian distribution: some properties and characterizations. Proc. *International Statistics Institute*, (Preprint), ISI, Netherlands.
6. Chen, B.E., Kondo, M., Garnier, A., Watson, F.L., Puettmann-Holgado, R., Lamar, D.R., Schmucker, D. (2006). The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila. 125(3), 607-620.
7. Gumbel, E.J. and Keeney, R.D. (1950). The Extremal Quotient. *Ann. Math. Statist.*, 21(4), 523-538.
8. Habibullah, S.N. (1987). On the Modes of the Inverted Bivariate Normal Distribution. *Pak. J. Statist.,* Series A, 3(1), 49-62.
9. Habibullah, S.N. Ahmad, M. Memon, A.Z (2010). Strictly Closed Under Inversion. Ph.D. Thesis. National College of Business Administration and Economics, Lahore, Pakistan.
10. Pearson, K. (1895). Contribution to the Mathematical Theory of Evolution II. Skew variation in Homogeneous Materials. *Phil. Trans. RSS*, London, Series A, 186, 343-414.
11. Pilon, P.J. and Adamowski, K. (1993). Asymptotic variance of flood quantile in Log–Pearson Type III distribution with historical information. *Journal of Hydrology*. 143(3), 481-503.
12. Seshadri, V. (1965). On Random Variables which have the Same Distribution as their Reciprocals. *Can. Math. Bull.*, 8(6), 819-824.
13. Singh V. P. and Singh K. (1988). Parameter Estimation for Log-Pearson Type III. Distribution by POME. *Journal of Hydraulic Engineering*, 114(1), 112-122.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 297-328

WHEN ASSOCIATION INDICES FAIL AND INFORMATION INDICES SUCCEED*

Nader Ebrahimi[1], **Nima Y. Jalali**[2] and **Ehsan S. Soofi**[2]
[1] Division of Statistics, Northern Illinois University, DeKalb, USA
[2] Lubar School of Business, University of Wisconsin-Milwaukee,
Milwaukee, USA. Email: esoofi@uwm.edu

ABSTRACT

The mutual information, denoted here as M, measures departure of the joint distribution from the independent model and identifies and orders stochastic dependence ways beyond the normal correlation coefficients, its nonparametric counterparts, and the fraction of variance reduction. We view M as an expected utility of variables for prediction. This view integrates ideas from the general dependence literature and the Bayesian perspectives. We illustrate the success of this index as a "common metric" for comparing the strengths of dependence within and between families of distributions in contrast with the failures of the popular traditional indices. For the location-scale family of distributions, an additive decomposition of M gives the normal distribution as the unique minimal dependence model in the family. An implication for practice is that the popular association indices underestimate the dependence of elliptical distributions, severely for models such as t distributions with low degrees of freedom. A useful formula for M of the convolution of random variables provides a measure of dependence when the predictors and the error term are normally distributed jointly or individually, as well as under other distributional assumptions. Finally, we draw attention to a caveat: M is not applicable to continuous variables when their joint distribution is singular, due to the fact that a functional relationship with positive probability. For an indirect application of M to singular models, we propose a modification of the mutual information index, which retains the important properties of the original index and show some potential applications.

Keywords. Entropy; location-scale family, Marshall-Olkin, mutual information, predictability, Student's t, utility.

1 Introduction

This article is an overview of the foundations and efficacy of a well known information measure of dependence, referred to as the mutual information and denoted here as M (there is no universally accepted notation). We draw ideas and cite evidence from several fields, including statistics (classics and contemporary), information theory and information science, physics and medical physics, geoscience,

*This document is an abridged version of a manuscript entitled "The Information Index of Dependence: Virtues with a Caveat", and Presence of a Probable Cause", which is currently under revision for a journal, prepared for presentation at the ICCS-12. Please do not quote without permission of the authors.

computational science, signal processing, reliability, econometrics and economics, behavioral research and marketing, and decision analysis. This measure is used in traditional problems such as categorical data analysis, regression analysis, time series, design of experiment, clustering, signal process, various Bayesian problems, as well as in more modern problems such as the importance of stochastic predictors (Theil and Chung 1988, Retzer et al. 2009), graphical models and networks (Bedford and Cooke 2002), feature selection in medical imaging (Tourassi et al. 2001), genetic and evolutionary computations (Card and Mohan 2009). An important development is a recent paper (Sims 2010) by this year's Nobel Laureate in Economics, Chris Sims, where he formulates an economic theory of "rational inattention" which includes the mutual information as a constraint for the optimal amount of information retained by economic agents from the environment.

The growing use of M in various fields calls for a revisit of this measure with eyes on its current developments and applications in statistics. In the recent upsurge of modeling dependence in statistics, M has been considered by Drouet-Mari and Kotz (2001) and Bedford and Cooke (2002), yet it is not a mainstream measure in spite of the following facts: (a) Because these studies often involve different families of models (copulas), "it is important to measure dependence on a common metric" (Smith et al. 2010); and (b) Currently the popular choices are Kendall's τ and Spearman's rank correlation ρ_s. The popularity of these indices is due to their invariance under monotone transformations which makes them applicable to copula, a widely used approach for modeling dependence (Drouet-Mari and Kotz 2001, Nelson 2006, Frees and Valdez 2008, Balakrishnan and Lai 2009). As will be shown these indices are not "common metrics" for comparing the strength of dependence even for the well known families of distributions. Applications of M to copula has appeared in other fields: image processing (Mercier et al. 2006 and Mercier and Inglada 2008), physics (Calsaverini and Vicente 2009), and decision analysis (Kotz and van Dorp 2010)). In the Bayesian statistics, application of M as the measure of expected utility is commonplace; see Ebrahimi et al. (2010a) for references. We give a synthesis of some classics and recent developments about M and highlight its theoretical foundations and its efficacy for practice.

The suitability of M as a 'common metric" for measuring the strength of dependence stems from its foundations and properties rooted in the information theory of Shannon (1948) and the information divergence between two probability distributions of Kullback and Leibler (1951). We bring a perspective that ties dependence with predictability. Dependence refers to the negation of the sharp state of independence where one variable X_i provides no probabilistic information for predicting another variable X_j. In contrast, the *complete dependence*, coined by Lancaster (1963) and interpreted by Kimeldorf and Sampson (1978) as the *perfect predictability*, refers to the sharp state of a functional relationship between two variables $P[X_i = g_i(X_j)] = 1, i \neq j$ where g_i is a one-to-one function. The independence and complete dependence define the opposite ends of the utility scale for the strength of dependence in terms of predicting one random variable by the other. This view integrates the basic connection of dependence and predictability with the utility ideas of sample information for Bayesian inference from Lindley (1956), DeGroot

(1962), and Bernardo (1979). Within this formal utility theory framework for the notion of dependence, the expected information utility in terms of Shannon entropy gives the same measure of dependence as the Kullback-Leibler (KL) information divergence between a model for dependence (joint distribution) and the independent model (product of marginals). The consequent unique measure, M, provides a robust index $\delta^2 \in [0, 1]$ for practice which, in more general terms, extends the interpretations of the squared Pearson correlation ρ_p^2 of the bivariate normal model to all absolutely continuous distributions. Analogous indices are available for discrete and categorical variables, which we do not discuss and refer the reader to Golden et al. (1990) for applications to categorical variables.

The road map of the paper is as follows. Section 2 gives an overview of the theoretical foundations of M, presents its various representations and interpretations, and its normalized index δ^2. It also illustrates that the counterparts of M based on their immediate generalizations by Rènyi's measures (Rènyi 1961) do not provide such a unique measure of dependence and gives a brief comparison with the L_p norms. In Section 3, we illustrate the efficacy of the mutual information index as a viable "common metric" for comparing the strengths of dependence within and between families of models, in contrast with the failure of four traditional popular indices: ρ_p, ρ_s, τ, and fraction of variance reduction due conditioning η^2, also known as the correlation fraction. This is accomplished by comparing the strengths of dependence within and between four widely used families of models (Gaussian, t and Cauchy, Pareto, and Farlie-Gumbel-Morgenstern (F-G-M) copula) along with four models from the recent literature. The information index also provides fresh insights about dependence within the t family, as well as for the Gaussian, Cauchy, Clayton, F-G-M copula, and a couple of recent families. Formulas for the M of these models are available in the literature; a multivariate M of the t family has appeared disjointly in statistics and physics (Gurrero 1998, Abe and Rajagopal 2001, Calsaverini and Vicente 2009).

Section 4 presents M for the location-scale (L-S) family where M decomposes into two parts, one for the dependence of the orthogonalized (scale-free) distribution and one for the dependence induced by the rotation (scale matrix), an ANOVA type decomposition for dependence. The latter is M for the normal distribution, the minimal dependence model in the entire L-S family (Ebrahimi et al. 2010b). (The decomposition for M of the t family was shown by Calsaverini and Vicente, 2009).

Section 5 pertains to singular distributions. These distributions appear in applications such as shock models and competing risk, the exponential autoregressive (EAR) processes, and Bayesian testing of sharp hypothesis. A requirement for M as a measure of departure from independence is that the joint distribution must be absolutely continuous relative to the product of the marginals. This does not hold for singular distributions, so M is not applicable to these models. However, the eagerness to utilize the versatility of M has reached to its improper use for the Marshall-Olkin copula (Mercier et al. 2006). As will be shown, M of this model can be zero when the variables are not independent and can be negative. We propose a modification of the information index allowing indirect application of M to singular models. We show its application through comparing the Marshall-Olkin bivariate exponential and the EAR models.

The final Section 6 summarizes the paper. An Appendix provides the proofs. The mathematical calculations are available in a Supplementary file.

2 Information Approach to Dependence

Consider the bivariate normal model F with correlation parameter ρ where $\rho_p^2 = \rho^2$; the subscript makes distinction between the correlation as an index and a model parameter. The perfect predictability is by a linear function $P(X_i = a + bX_j) = 1$ if and only if $\rho_p^2 = 1$, and the variables are independent - one variable has no predictive power about the other, if and only if $\rho_p^2 = 0$. The interpretation of ρ_p^2 in terms of the strength of predictability of one variable for the other is more clear from the fraction of variance reduction due to conditioning,

$$\rho_p^2 = 1 - \frac{\mathrm{Var}(X_i|x_j)}{\mathrm{Var}(X_i)} \geq 0, \quad j \neq i = 1, 2, \tag{1}$$

where the equality holds if and only if the two variables are independent. That is, for the normal model, ρ_p^2 measures predictability in terms of the concentration of the conditional distribution relative to the marginal distribution, and it is a measure of departure of F from the independence, though only invariant under linear transformations up to the sign.

The interpretations of ρ_p^2 for the bivariate normal model hinge on the conditional variance being constant and the variance and conditional variance being finite. The traditional extension of (1) when the conditional variance is not constant is the fraction of expected variance reduction due to regression, also known as the correlation fraction,

$$\eta_{i|j}^2 = 1 - \frac{E_{x_j}[\mathrm{Var}(X_i|x_j)]}{\mathrm{Var}(X_i)} \geq 0, \quad j \neq i = 1, 2. \tag{2}$$

Finite variances are still required and unlike ρ_p^2 for the normal case, $\eta_{i|j}^2$ is not symmetric in X_i and X_j. In general $\eta_{i|j}^2 = 0$ does not imply that the two variables are independent, so it does not measure departure from independence; see Table 1 in Section 3. This index is also invariant under linear transformations only.

The information notion of dependence builds on the interpretations of the bivariate normal ρ_p^2 and the expected variance reduction in (2) as follows:

- The perfect linear predictability generalizes to any functional relationship $P[X_i = g_i(X_j)] = 1$.
- The variance reduction generalizes in terms of a more general uncertainty function $\mathcal{U}(f)$ which measures the concentration of the probability density function (pdf), f. (The variance does not always satisfy this condition (e.g., beta family); see Ebrahimi et al. 2010b).

- The departure from independence is formalized by a divergence function between two probability distributions $\mathcal{D}(f, g) \geq 0$ where the equality holds if and only if $f(x) = g(x)$ for almost all x, scalar or a vector.
- The invariance under linear transformations generalizes to the invariance under all one-to-one transformations.

These ingredients provide a robust measure of dependence endowed by the same interpretations of ρ_p^2 of the normal model for all absolutely continuous distributions.

2.1 Information Notion of Dependence

For the formulation of dependence in terms of the expected uncertainty reduction, $\mathcal{U}(f)$ must satisfies two desirable properties. First, $\mathcal{U}(f) \leq \mathcal{U}(c)$ where c is a constant, a uniform proper or improper density on the support of F (Shannon 1948). This condition allows ranking distributions according to the concentration of probability. Second, $\mathcal{U}(\cdot)$ is a concave function. This condition ensures that, on average, the information utility of using one random variable for predicting another random variable is nonnegative and the worst case is when the two variables are independent. This desirable property is satisfied if and only if $\mathcal{U}(f)$ is concave (DeGroot 1962).

The unpredictability of outcomes of X_i without using $X_j, j \neq i = 1, 2$, depends solely on the concentration of its marginal distribution measured by $\mathcal{U}(f_i)$. Given an outcome x_j of X_j, the unpredictability of outcomes of X_i is measured by $\mathcal{U}[f_{i|j}(x_i|x_j)]$, the uncertainty of a conditional distribution. The utility (worth) of an outcome x_j of X_j for predicting outcomes of X_i, $i \neq j$ is the difference, $\mathcal{U}(f_i) - \mathcal{U}[f_{i|j}(x_i|x_j)]$, known as the observed information provided by x_j for predicting X_i. This is the information generalization of the variance reduction due to conditioning in the bivariate normal case. In general, the utility is a function of x_j. The *expected utility* of X_j for prediction of X_i is

$$\Delta(X_i|X_j) = E_{x_j}\Big\{\mathcal{U}[f_i(x_i)] - \mathcal{U}[f_{i|j}(x_i|X_j)]\Big\} \geq 0, \quad i \neq j; \tag{3}$$

the inequality is from the concavity condition and by Jensen inequality, $\mathcal{U}[f_i(x_i)] \geq E_{x_j}\{\mathcal{U}[f_{i|j}(x_i|x_j)]\}$, where the equality holds if and only if the two variables are independent. Thus, the expected utility is a measure of departure from the independence, as well. The expected information utility $\Delta(X_i|X_j)$ is the generalization of the reduction of the expected variance (2).

Equation (3) is a comparison between $f_{i|j}(x_i|x_j)$ and $f_i(x_i)$ based on the predictability relative to the global maximum uncertainty of the uniform distribution. Alternatively, the utility of an outcome x_j of X_j for predicting outcomes of X_i, $i \neq j$ can be measured by a divergence function $\mathcal{D}[f_{i|j}(x_i|x_j) : f_i(x_i)] \geq 0$ with equality if and only if the two pdf's are identical. Then the expected utility is the expected gain for predictability of X_i by using the conditional distributions relative to the marginal:

$$\mathcal{D}(X_i|X_j) = E_{x_j}\Big\{\mathcal{D}[f_{i|j}(x_i|x_j) : f_i(x_i)]\Big\} \geq 0, \quad i \neq j = 1, 2, \tag{4}$$

where the equality holds if and only if the two variables are independent almost everywhere.

As a third alternative, dependence in terms of the departure from the independence is measured directly by divergence between the joint pdf $f(x_1, x_2)$ and the product of marginal pdf's $f_1(x_1)f_2(x_2)$,

$$\mathcal{D}(X_1, X_2) = \mathcal{D}[f(x_1, x_2), f_1(x_1)f_2(x_2)] \geq 0, \tag{5}$$

where the equality holds if and only if $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ almost everywhere. This measure most clearly shows the necessary and sufficient condition for identifying the independence between two variables.

2.2 Mutual Information

The most well-known and widely-used measure of uncertainty is Shannon entropy (Shannon 1948). The entropy of a continuous distribution is

$$H(X) = H(F) = -\int_{\mathcal{S}} f(x) \log f(x) dx, \tag{6}$$

where $\mathcal{S}$ is the support of F, provided that the integral is finite. Finiteness of variance implies $H(X) < \infty$, but converse does not hold. The most well-known and widely-used measure of information divergence is the KL function defined as

$$K(F : G) = \int_{\mathcal{S}} f(x) \log \frac{f(x)}{g(x)} dx, \tag{7}$$

provided that F is absolutely continuous with respect to G, denoted as $F \ll G$; see Kullback (1959) for details. It is also known as cross-entropy and relative entropy. Using the Shannon entropy in (3) and KL divergence in (4) and (5) provides a unique measure of dependence for a bivariate distribution F, namely the mutual information,

$$\begin{aligned} M(F) \equiv M(X_1, X_2) &= \Delta(X_i|X_j), \quad j \neq i = 1, 2 & (8)\\ &= \mathcal{D}(X_i|X_j), \quad j \neq i = 1, 2 & (9)\\ &= K(F : F_1F_2), & (10)\end{aligned}$$

provided that $F \ll F_1F_2$.

Representation (8) is usually written as

$$\Delta(X_i|X_j) = H(X_i) - \mathcal{H}(X_i|X_j), \quad j \neq i = 1, 2, \tag{11}$$

where $\mathcal{H}(X_i|X_j) = \int f_j(x_j)H(X_i|x_j)dx_j$ is referred to as the conditional entropy.

In Shannon's information theory, X_j represents signals transmitted from a source through a noisy channel which transmits X_i, and the maximum of (8) is the capacity of the channel (Shannon 1948). In Lindley's (1956) Bayesian adaptation of

Shannon's measure, the signal is a parameter θ, the noisy channel transmits the sample X, and $\Delta(\theta|X)$ is the expected information provided by the sample about the parameter. Bernardo (1979) provided the expected utility interpretation of (8) and (9) in the Bayesian context which is now commonplace for the signal being a parameter, a future outcome, or both (Ebrahimi et al. 2010a). We interpret (8) and (9) in the more general context of the expected information utility of dependence in terms of predictability as in (3) and (4).

Often M is defined by (10) as it clearly shows that $M(X_1, X_2) \geq 0$ and the equality holds if and only if X_1 and X_2 are independent, the case of perfect unpredictability. It is well known that $M(X_1, X_2) = \infty$ when one of the variables is functionally dependent on the other, $P[X_1 = g_1(X_2)] = 1$ or $P[X_2 = g_2(X_1)] = 1$, for some measurable functions g_1 and g_2, the cases of perfect predictability (Lancaster 1963, Kimeldorf and Sampson 1978). However, the absolute continuity requires $P[X_1 = g_1(X_2)] = 0$ and $P[X_2 = g_2(X_1)] = 0$, the case of improbable cause.

The unique additive property of Shannon entropy also gives the following representation

$$M(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2). \tag{12}$$

This representation is referred to as the *shared or redundant information* by Golden et al. (1990); the latter term reflects a coding theory interpretation of (7). Formula (12) is particularly useful for calculating M by the entropy expressions of univariate and bivariate distributions available, e.g., in Cover and Thomas (1991) and Nadarajah and Zografos (2003, 2005). From (12) it is clear that finiteness of the joint and marginal entropies are necessary. However, this is not sufficient, for $M(X_1, X_2) < \infty$; this will be illustrated in Section 5.

From (10), it is clear that M is invariant under all one-to-one transformations of each $X_i, i = 1, 2$. This property is very useful since transformations are used in many statistical and modeling applications. In particular, let $U_i = F_i(X_i),\ i = 1, 2$. Then,

$$C(u_1, u_2) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right), \quad (u_1, u_2) \in [0, 1]^2,$$

is the copula of F, given by Sklar's Theorem (Sklar 1959). For applications of M to copulas see Drouet-Mari and Kotz (2001), Bedford and Cooke (2002), Mercier et al. (2006), Mercier and Inglada (2008), Calsaverini and Vicente (2009), and Kotz and van Dorp (2010). The mutual information of F can be represented in terms of the copula information:

$$M(F) = M(C) = K(C : C_0) = -H(C) = I(C) \geq 0, \tag{13}$$

where C_0 denotes the product copula $C_0(u_1, u_2) = u_1 u_2$ and $I(C)$ is referred to as the information measure of the distribution (Lindley 1956, Zellner 1971), here the copula. The first equality is due to the invariance property of M under one-to-one transformations of the components, the second inequality is from (10), and the third equality is from (12) since the marginal distributions of C are uniform on $[0, 1]$ so $H(U_i) = 0,\ i = 1, 2$. The uniform distribution is the least concentrated distribution and maps the most unpredictable (non-informative) situation. The inequality in (13) becomes equality if and only if C is the product copula.

Furthermore, we can assess the effect of transforming by univariate distributions G_i, $i = 1, 2$ different from the marginals F_i of F. Transformations $V_i = G_i(X_i), i = 1, 2$ produce a bivariate distribution

$$C^*(v_1, v_2) = F\left(G_1^{-1}(v_1), G_2^{-1}(v_2)\right), (v_1, v_2) \in [0,1]^2,$$

where the marginal distributions, $C_i^*(v_i), i = 1, 2$ are no longer uniform. However, still by the invariance of M,

$$M(F) = M(C^*) = I(C),$$

but $I(C) \neq I(C^*) = -H(C^*)$. In fact, since the marginal distributions of C^* are nonuniform on $[0, 1]$, and the uniform distribution is the maximum entropy, $H(V_i) < 0$ and $M(C^*) \leq -H(C^*) = I(C^*)$. The difference is given by the following decomposition

$$I(C^*) - I(C) = K(F : G_1G_2) - K(F : F_1F_2) = \sum_{i=1}^{2} K(F_i : G_i) \geq 0, \tag{14}$$

where the inequality changes to equality if and only if $g_i(x) = f_i(x)$, $i = 1, 2$ almost everywhere; proof is given in the Appendix. For example, when $G_i = \hat{F}_{i,n}$ is an empirical estimate based on a sample size n, then $I(\hat{C}_n^*) - I(C)$ is given in terms of $K(F_i : \hat{F}_{i,n})$, known as the KL loss in the estimation. (The literature for this topic is very rich and beyond the scope of this paper). When $\hat{F}_{i,n}$ is a parametric empirical distributions using consistent estimates, the KL loss vanishes as $n \to \infty$. For example, for a bivariate normal distribution F with correlation parameter ρ,

$$I(C) = M(F) = M(X_1, X_2) = -\frac{1}{2}\log(1 - \rho^2) \equiv M(\rho). \tag{15}$$

For $G_i = \hat{F}_i = N(\hat{\mu}_i, \hat{\sigma}_i^2)$,

$$\begin{aligned} I(C^*) - I(C) &= \sum_{i=1}^{2} K(F_i : \hat{F}_i) \\ &= \frac{1}{2}\sum_{i=1}^{2} \frac{(\mu_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} + \frac{1}{2}\sum_{i=1}^{2}\left(\frac{\sigma_i^2}{\hat{\sigma}_i^2} - \log\frac{\sigma_i^2}{\hat{\sigma}_i^2} - 1\right) \geq 0, \end{aligned}$$

which with consistent parameter estimates vanishes as $n \to \infty$. More generally, using a large sample and suitable empirical distributions the effect is negligible.

2.3 Information Index

Normalized information indices of dependence are defined by mapping $M(X_1, X_2)$ onto the unit interval. The following index is used for dependence of *absolutely continuous* distributions:

$$\delta^2(F) = \delta^2(X_1, X_2) = 1 - e^{-2M(X_1,X_2)}. \tag{16}$$

This is a well-known index. Linfoot (1957) originally proposed δ as a generalization of the bivariate normal correlation and referred to it as the *informational coefficient of correlation*. Joe (1989) presented δ in terms of the relative entropy (10). The copula representations of the information index is

$$\delta^2(F) = \delta^2(C) = 1 - e^{-2I(C)} = 1 - e^{2H(C)}. \tag{17}$$

We refer to (16) as the *mutual information index* or the *copula information index*, or the *dependence information index*.

The following representations of (16) provide further insights:

$$\begin{aligned} \delta^2(X_1, X_2) &= 1 - \frac{\exp\{E_{x_j}[H(X_i|x_j)]\}^2}{\exp\{H(X_j)\}^2} && (18) \\ &= 1 - \frac{\exp\{H(X_1, X_2)\}^2}{\exp\{H(X_1) + H(X_2)\}^2}. && (19) \end{aligned}$$

Representation (18) is from (8) and gives the fraction of entropy reduction on the exponential-scale, the information theoretic counterpart of the fraction of variance reduction (2). The total entropy in the denominator of the fraction in (19) is from (12) and measures the uncertainty of the independent model, the maximum uncertainty among all distributions with given marginals. Thus, $\delta^2(X_1, X_2)$ is the squared fraction of the reduction of maximum uncertainty due to dependence measured on the exponential-scale. Representations (18) and (19) provide more precise and meaningful interpretations for Linfoot's (1957) generalization of the normal correlation. For a bivariate normal distribution,

$$\delta^2(F) = \eta^2(F) = \rho_p^2(F) = \rho^2.$$

The calibration of δ^2 with the (Gaussian) correlation parameter is based on this relationship.

2.4 Rényi Measures and L_p Norms

The Shannon-Kullback-Leibler root of M that allows representations as a measure of the expected utility (uncertainty reduction) (3), as the divergence between a model and the independence (5), and as the shared information (12), naturally extends the same interpretations of the bivariate normal ρ_p^2 for all absolutely continuous distributions. In general, among the known divergence measures and generalizations of (6) and (7), only the KL information admits the expected utility representation (3) with $\mathcal{U}(\cdot)$ being Shannon entropy as well as the expected relative uncertainty representation (4). Next we illustrate that the immediate generalizations of (6) and (7) do not produce a unique measure like M.

Rènyi entropy and divergence information for measuring dependence are

$$H_r(X) = \frac{1}{1-r} \log \int f^r(x)dx, \quad r \neq 1, \quad r > 0,$$

$$K_r(F : F_1F_2) = \frac{1}{r-1} \log \int\int f^r(x_1, x_2)[f_1(x_1)f_2(x_2)]^{1-r} dx_1 dx_2, \quad r \neq 1, \quad r > 0,$$

and $H_1(X) = H(X)$, $K_1(F:G) = K(F:G)$ by continuity; for the latest developments see Principe (2010). The divergence with $r = \frac{1}{2}$ is symmetric $K_{1/2}(F:G) = K_{1/2}(G:F)$ and has representations in terms of the Hellinger and Bhattacharya distances, and a divergence information measure proposed by Tsallis (1988). These are viewed by Hirschberg, et al. (1991) and Granger, et al. 2004) as being appealing properties for measuring dependence.

With Rènyi's entropy, the uncertainty reduction measure (3) for the bivariate normal is free from r and equals to $M(\rho)$ in (15). Although in general Rènyi entropy does not admit representation (12), for the normal model it gives the same measure $M(\rho)$. However, Rènyi divergence in (4) and (5) give a different measure from $M(\rho)$:

$$\mathcal{D}_r(X_1|X_2) = \mathcal{D}_r(X_2|X_1) = \mathcal{D}_r(X_1, X_2) = M(\rho) + \frac{1}{2(1-r)} \log\left(1 - (1-r)^2\rho^2\right).$$

Thus for a bivariate normal distribution, Rènyi's divergence measures with $r > 1$ and $r < 1$ indicate, respectively, stronger or weaker dependence than the uncertainty reduction measures. In general, with Rènyi's measures, (3)-(5) can be all different and representation (12) is not nonnegative and its zero value does not identify independence.

Measures of dependence have been also constructed based on L_p norms. For example, consider the L_1 norm,

$$L_1(f, f_1 f_2) = \int\int \left| f(x_1, x_2) - f_1(x_1) f_2(x_2) \right| \geq 0.$$

The equality holds if and only if the integrand is zero, so the variables are independent. That is, for any dependent model $L_1(f, f_1 f_2) > 0$. A known inequality for the KL divergence gives $L_1^2(f, f_1 f_2) \leq 2M$, so small values of M implies that $L_1(f, f_1 f_2) \approx 0$. Dependence measures based on the L_2 norm also have appeared in the literature; see Seth et al. (2011) for the latest developments. Because of the log-ratio of the pdf's in (10), M is more able to detect small differences in tail behavior (tail dependence) while retaining sensitivity to the regions where the pdf is not small. This makes M a stronger measure of dependence than those measures that are based on the difference such as L_1 and L_2 norms. Recently, Szèkely, et al. (2007) proposed an index based on a weighted L_p-norm for testing independence, referred to as "distance correlation". Because the weight function uses moments this measure, unlike M, it is not invariant beyond linear transformations.

3 Comparison with Traditional Indices

We compare the efficacy of δ^2 as a viable "common metric" for comparing the strengths of dependence within and between families of models with four popular indices ρ_p, ρ_s, τ, and η^2. The last two indices are invariant under monotone transformations and their copula representations are as follows:

$$\tau = E_c\Big\{4C(U_1, U_2) - 1\Big\}, \qquad \rho_s = E_c\Big\{12U_1U_2 - 3\Big\}.$$

Note that these indices are the expected values of the copula and product copula, respectively, but (17) is a normalized version of the expected value of log copula density.

These indices range in $[-1, 1]$ and their signs indicate the direction of the association between two variables. When $\rho_p = \tau = \rho_s = 0$ and $\delta^2 > 0$, the variables are dependent, but *non-directional*. When the association indices are not zero, their signs can be used along with δ^2 to indicate the direction of departure from independence.

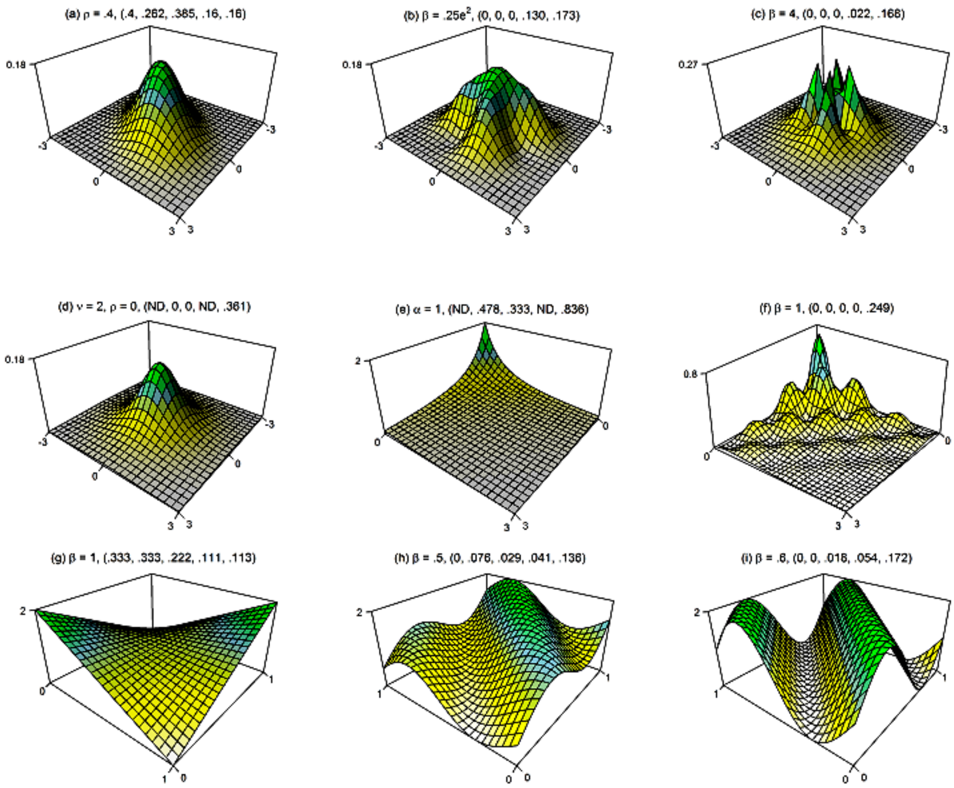
We compare the efficacy of δ^2 as a "common metric" with ρ_p, τ, ρ_s, and η^2 for comparing the strengths of dependence of distributions between and within nine families. In addition to four classic families (Gaussian, Student-t, Pareto and the F-G-M copula), we have included five families which recently have appeared in the literature. Examples of pdf's from these families are shown in Figure 1. The pdf's of these families will be given below and the expressions for their indices are shown in Table 1.

These nine models in Figure 1 are selected to illustrate some issues related to comparison of dependence between families. For each model, its indices are shown as a vector $(\rho_p, \rho_s, \tau, \eta^2, \delta^2)$, where a zero element indicates the index is zero for the entire family and "ND" means undefined for the distribution shown. The top row includes the bivariate normal with $\rho = .4$ and two recent models with normal $N(0, 1)$ marginals. The departures of (b) and (c) from the independent model (strength of dependence) are about the same level as (a), but they are non-directional, $\rho_p = \rho_s = \tau = 0$, and their η^2's are different. The middle row includes a Student-t and a Pareto for which the moment-based indices are undefined, and a recent model with lognormal marginals. The dependence of t model is non-directional, $\rho_p = \rho_s = \tau = 0$ and so is for model (f) for which $\rho_p = \rho_s = \tau = \eta^2 = 0$. The strengths of dependence of models (d)-(f) are about the levels of bivariate normal models with correlation .6, .91, and .5, respectively. The last row includes the F-G-M copula and two recent models with pdf's on the unit square which are examples of uncorrelated dependent models and uncorrelated copulas where ρ_s is also zero. These models are arranged in increasing order of δ^2 which measures their departures from the independence (product copula), but the association indices ρ_p, ρ_s, τ and η^2 rank them inversely with δ^2. Overall, δ^2 is the only "common metric" enabling us to compare the strengths of dependence between the models in Figure 1.

The bivariate normal pdf is a household item and we only mention it as the limiting distribution of the bivariate Student-t distribution with pdf

$$f(x_1, x_2|\rho, \nu) = \frac{1}{2\pi(1-\rho^2)^{1/2}}\left(1 + \frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{\nu(1-\rho^2)}\right)^{-\nu/2-1}, \quad (x_1, x_2) \in \Re^2,$$

where $\nu = 1, 2, \cdots$ is the degrees of freedom and ρ is the orientation parameter.



Legends:

(a) Bivariate normal; (b) & (c) Non-directional dependent;
(d) Student-*t*; (e) Pareto; (f) Uncorrelated polynomial;
(g) F-G-M copula; (h) Uncorrelated dependent; (i) Uncorrelated copula

Fig.1. Nine dependence models with indices $(\rho_p, \rho_s, \tau, \eta^2, \delta^2)$.

The Cauchy pdf is given with $\nu = 1$ and the normal model is the limit of $\nu \to \infty$. The association indices ρ_p, ρ_s, and τ for the Gaussian and t are well known. The conditional variance $V(X_i|x_j)$ of the t distribution is quadratic in x_j and is a function of both ρ and ν, but not defined for $\nu = 1$ (Balakrishnan and Lai, 2009) and the fraction of variance reduction is decreasing in ν. The information index (16) for the t family is calculated using the entropies of univariate and bivariate t distributions in (12). The entropy of the d-dimensional t distribution with ν degrees

Table 1. Expressions for indices used for the comparison of the strengths of dependence within and between families of models.

Family	ρ_p	ρ_s	τ	$\eta^2_{i\|j}$	δ^2
Gaussian, $\|\rho\| < 1$	ρ	$\frac{2}{\pi}\sin^{-1}(\rho)$	$\frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)$	ρ^2	ρ^2
Student-$t(\rho,\nu), \nu > 2, \|\rho\| < 1$	ρ	$\frac{2}{\pi}\sin^{-1}(\rho)$	$\frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)$	$\rho^2 + (1-\rho^2)\frac{2\nu-1}{\nu^2-1}$	$\rho^2 + (1-\rho^2)\delta^2(\nu,0)$
Cauchy, $t(\rho,1), \|\rho\| < 1$	Undefined	$\frac{2}{\pi}\sin^{-1}(\rho)$	$\frac{6}{\pi}\sin^{-1}\left(\frac{\rho}{2}\right)$	Undefined	$\rho^2 + (1-\rho^2)\delta^2(1,0)$
Pareto, $\alpha > 0$	$\frac{1}{\alpha}, \alpha > 2$	Complicated	$\frac{1}{2\alpha+1}$	$\frac{1}{\alpha^2}, \alpha > 2$	$1 - \frac{\alpha^2}{\alpha^2+1}e^{-2/(\alpha+1)}$
F-G-M Copula, $\|\beta\| \le 1$, $q(x_1,x_2) = (1-2x_1)(1-2x_2)$	$\frac{\beta}{3}$	$\frac{\beta}{3}$	$\frac{2\beta}{9}$	$\frac{\beta^2}{9}$	$1 - e^{-2m(\beta,q)}$
Uncor. dependent, $\|\beta\| \le .5$, $q(x_1,x_2) = \frac{\sin(2\pi(x_1-x_2))}{.5+x_1}$	0	$\frac{3\beta}{2\pi^2}$	$\frac{\beta(\pi\beta+2)}{2\pi^3}$	$\frac{18\beta^2}{11\pi^2}$	$1 - e^{-2m(\beta,q)}$
Uncor. Copula, $\|\beta\| \le 1$, $q(x_1,x_2) = \sin(2\pi(x_1-x_2))$	0	0	$\frac{\beta^2}{2\pi^2}$	$\frac{3\beta^2}{2\pi^2}$	$1 - e^{-2m(\beta,q)}$
Non-directional dependent, $\|\beta\| \le .25e^2$ $q(x_1,x_2)$ is in Eq. (22) (unimodal)	0	0	0	$\frac{11\beta^2}{144\sqrt{3}}$	$1 - e^{-2m(\beta,q)}$
Uncor. polynomial, $\|\beta\| \le 1$, $q(x_1,x_2) = \sin(2\pi\log x_1)\sin(2\pi\log x_2)$	0	0	0	0	$1 - e^{-2m(\beta,q)}$

NOTE: $m(\beta, q)$ is given by equation (23).

of freedom and scale matrix Σ is

$$H(\boldsymbol{X}) = \log\frac{(\nu\pi)^{d/2}\Gamma(\nu/2)}{\Gamma((\nu+d)/2)} + \frac{\nu+d}{2}\left[\psi\left(\frac{\nu+d}{2}\right) - \psi\left(\frac{\nu}{2}\right)\right] + \frac{1}{2}\log|\Sigma|, \qquad (20)$$

where $\psi(z) = \frac{d\log\Gamma(z)}{dz}$ is the digamma function and $|\Sigma|$ is determinant (Zografos and Nadarajah 2005). For M, the marginal scales cancels out and $.5\log(1-\rho^2)$ remains from the determinant: M is increasing in ρ^2 and decreasing in ν.

The pdf of bivariate Pareto Type II family is

$$f(x_1, x_2|\alpha) = \alpha(\alpha+1)(1+x_1+x_2)^{-\alpha-2}, \quad x_1, x_2 \ge 0, \quad \alpha > 0.$$

For the Pareto family, the second moments are not defined for $\alpha \le 2$. The Kendall's τ is well-known and the expression for ρ_s is too complicated, but for values of α can be easily computed by a software. The information index is computed using expression for M of the Pareto Type II distribution, which is decreasing in α.

For traditional indices of the F-G-M copula are known. The F-G-M copula and the four other families in Table 1 have pdf's in the following form:

$$f(x_1, x_2|\beta) = f_1(x_1)f_2(x_2)[1 + \beta q(x_1, x_2)], \quad (x_1, x_2) \in \Re^2, \quad \beta \leq B^{-1}, \tag{21}$$

where $f_i(x_i)$, $i = 1, 2$ are marginal pdf's and $q(x_1, x_2)$ is a measurable function on $\Re^2$ bounded as $|q(x_1, x_2)| \leq B$. For $\beta = 0$, the distributions are the independent bivariate models. For $q(x_1, x_2) = q_1(x_1)q_2(x_2)$, (21) gives the Sarmanov family (Balakrishnan and Lai, 2009), so it can be referred to as the *Generalized Sarmanov family*. In Figure 1, the multimodal pdf (c) is also in this family. (It is not included in Table 1 since the expression for its η^2 is complicated.) This and the three families below the F-G-M copula in Table 1 are in a class of bivariate models for uncorrelated variables proposed by Ebrahimi et al. (2010c).

The uncorrelated dependent model in Table 1 (Figure 1 (h)) are $f_1(x_1) = .5 + x_1$ and $f_2(x_2) = U[0, 1]$. The traditional indices of this and the uncorrelated copula (Figure 1 (i)) can be computed analytically or using a software (we used MAPLE). The pdf's of the two non-directional dependent models (Figure 1 (b) and (c)) are given by (21) with

$$\begin{aligned} &\text{Unimodal:} \quad f_i(x) = N(0, 1), \ |\beta| \leq .25e^2, \ q(x_1, x_2) = x_1x_2(x_1^2 - x_2^2)e^{-\frac{1}{2}(x_1^2+x_2^2)}, \\ &\text{Multimodal:} \quad f_i(x) = N(0, 1), \ |\beta| \leq 4, \ q(x_1, x_2) = \frac{x_1x_2(x_1^2 - x_2^2)}{(x_1^2 + x_2^2)^2}e^{-\frac{1}{2}(x_1^2+x_2^2)}. \end{aligned} \tag{22}$$

These two models belong to a construction where $\rho_p = \rho_s = \tau = 0$ (Ebrahimi et al. 2010c). The marginal distributions of the family of the uncorrelated polynomial model (f) are log-normal $f_i(x) = LN(0, 1)$. De Paula (2008) showed that for $\beta = 1$, $E(X_i^m|X_j) = E(X_i^m), i \neq j$ for all $m = 1, 2, \cdots$, yet the two variables are not independent. That is, all polynomial functions of the two variables are uncorrelated, but the two variables are not independent. Consequently, $\eta^2_{i|j} = 0$. De Paula's proof holds for all β.

The mutual information index for (21)has the following series representation:

$$m(\beta, q) = M_\beta(X_1, X_2) = \beta E_2\left\{E_1[q(X_1, X_2)]\right\} + \sum_{n=2}^{\infty} \frac{(-\beta)^n}{n(n-1)} E_2\left\{E_1[q(X_1, X_2)]^n\right\}, \tag{23}$$

where E_i, $i = 1, 2$ denotes the expectation with respect to f_i, which is increasing in $|\beta|$, (Ebrahimi et al. 2010c).

Next we illustrate that within a family where a traditional index is not identically zero, the information index is an increasing function of the magnitude of that index. Figure 2 shows this for τ. The information index for the t family can be represented as

$$\delta^2(\nu, \tau) = \sin^2\left(\frac{\pi\tau}{2}\right) + \cos^2\left(\frac{\pi\tau}{2}\right)\delta^2(\nu, 0), \tag{24}$$

which is increasing in $|\tau|$. The plots at the left side are $\delta^2(\nu, \tau)$ with $\nu = 1, 2, 3, 5$, and the limiting case of $\nu = \infty$, which corresponds to the Gaussian copula, against

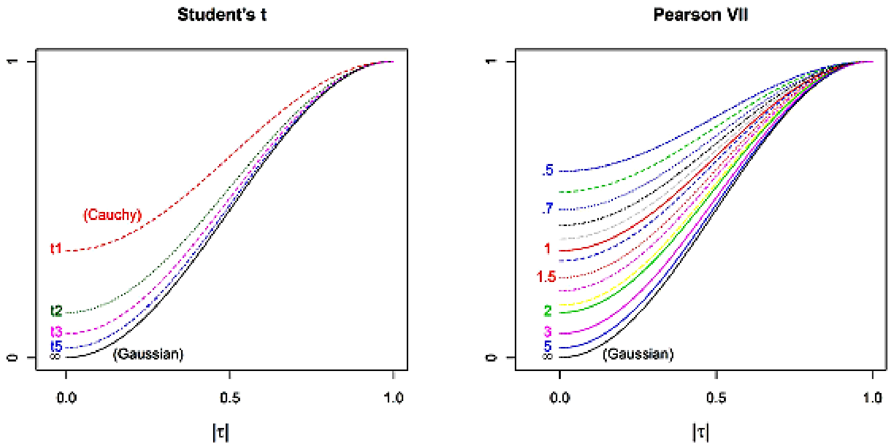


Fig. 2. Plots of information indices for the Gaussian, t, and Pearson VII families

τ. Further insights gained through the information about the t family is that when τ and the degrees of freedom are low, the range of dependence is low and there are substantial gaps between the levels of dependence of the t distributions. The spectrum of dependence of the t family can be substantially widen and refined by replacing $\nu/2$ with a parameter $\alpha > 0$. This modification gives the more general bivariate Pearson Type VII distribution. The dashed and dotted plots at the right side are $\delta^2(\alpha, \tau)$ for several fractional values of α. We note that the range of dependence is now widen and the gaps between dependence are recovered, particularly for small values of τ.

Moreover, by the invariance property of M, the indices shown in Table 1 are applicable to the copulas of these models. Figure 3 illustrates two comparisons between two sets of copulas. The plots at the left side are information indices of the bivariate Gaussian, Cauchy, and Clayton copulas, where the latter also known as Pareto survival copula given by

$$C(u_1, u_2) = \left(u_1^{-1/\alpha} + u_2^{-1/\alpha} - 1\right)^{-\alpha}, \quad \alpha > 0.$$

The marginal survival function of the bivariate Pareto Type II is $\bar{F}(x) = (1+x)^{-\alpha}$ and $Ui = \bar{F}_i(X_i), i = 1, 2$ give this copula. The information index of the Clayton copula is the same as δ^2 for Pareto shown in Table 1 and can also be expressed in terms of τ. As seen in Figure 3, the information indices for all three models are increasing τ and, for each τ, δ^2 of the Clayton copula is bounded by the indices of the Gaussian and Cauchy models. The plots at the right side gives similar comparison for the copulas of the families of the three models shown in the last row of Figure 1.

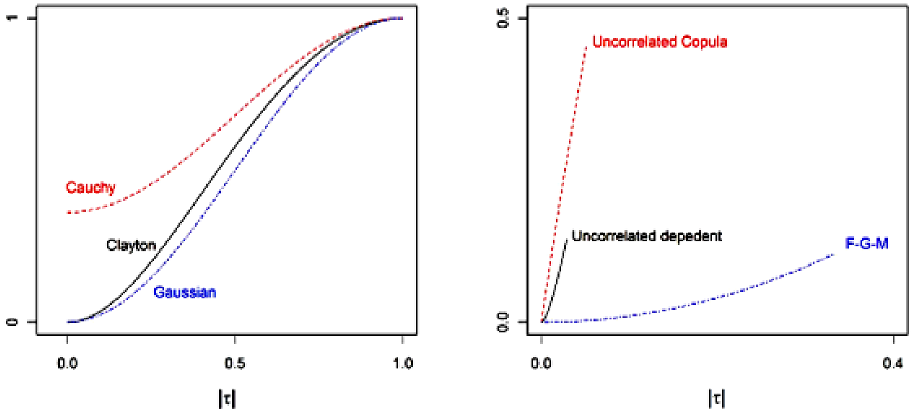


Fig. 3. Plots of information indices for six copulas.

Figure 4 shows a visual display for a glance of comparison between the five indices for serving as a "common metric" to compare the dependence of distributions within and between the nine *families* represented in Table 1 and Figure 1. The following facts are evident.

- *Correlation coefficient*: The magnitude of ρ_p measures the strengths of dependence, its sign measures the direction of association for the Gaussian model and F-G-M copula, and it is only defined for Pareto when $\alpha > 2$. It is not defined for t with $\nu \leq 2$ and since it remains the same for all t distributions with $\nu > 2$, it cannot be used to compare the strengths of dependence within this family and between a t distribution and the members of other families under comparison. It is identically zero for all members of the last four families shown in Table 1. It is noteworthy to mention that the correlation fails to capture dependence for the t family when $\nu > 2$ even though the regression function is linear.

- *Spearman's index*: This index is applicable to comparison within and between four families. It remains the same for all t distributions including Cauchy, so is useless for comparing a t distribution with other models within the family and between the families.

- *Kendall's index*: This index works like ρ_s, but for one more family where it is nonnegative, and also fails like ρ_s, but for one less family.

- *Fraction of variance reduction*: This index measures the strengths of dependence for five families. It is not defined for t with $\nu \leq 2$ and Pareto $\alpha \leq 2$ and is identically zero for all members of the last family.

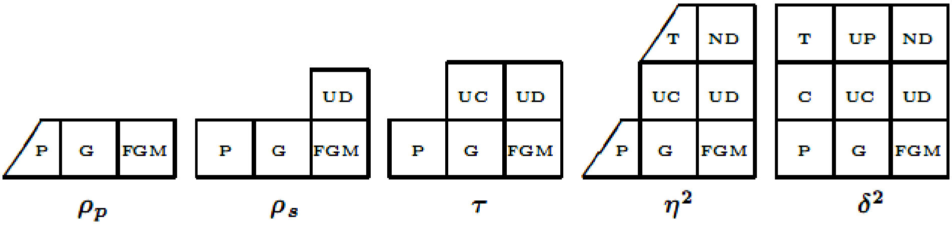


Fig.4. Applicability of indices to families: Gaussian (G); F-G-M copula (FGM); Pareto (P); Uncorrelated dependent (UD); Uncorrelated copula (UC); Non-directional dependent(ND); Uncorrelated polynomial (UP); Student-t (T); Cauchy (C); a trapezoid indicates partial applicability.

- *Information index*: This index remains the only viable measure as a "common metric" for comparing the strength of dependence within and between all the nine families; ρ_p, ρ_s, τ, and η^2 can be zero when the variables are not independent (t and the last four families).

4 Location-Scale Family

Calsaverini and Vicente (2009) noted that M of the t family decomposes into two parts and Ebrahimi et al. (2010b) noted that the same type of decomposition holds for the location-scale family. The distribution of a random vector $\boldsymbol{X}$ is said to have d-dimensional pdf $f(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma)$ in the location-scale (L-S) family with location vector $\boldsymbol{\mu}$ and scale matrix Σ if

$$\boldsymbol{X} \overset{st}{=} \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{X}^o,$$

where $\overset{st}{=}$ denotes the stochastic equality, $\boldsymbol{\theta}$ is the model parameters other than $\boldsymbol{\mu}$ and Σ and the distribution of $\boldsymbol{X}^o$ is in the same family with $\boldsymbol{\mu} = \mathbf{0}$, the vector of zeros, and $\Sigma = I_d$, the identity matrix; for $d > 1$ we refer to $\boldsymbol{X}^o$ as the orthogonalized version of $\boldsymbol{X}$.

In the multivariate case the mutual information measures are defined for independence between two or more subvectors of a d-dimensional random vector $\boldsymbol{X}$. Examples include the independent model $F_1 \cdots F_d$ and independence of two disjoint subvectors, $F_j(\boldsymbol{x}_j)F_h(\boldsymbol{x}_h)$, $j + h = d$. The divergence measure and information index map the strength of the dependence on the scale between the respective independent structure and a complete dependence.

4.1 Minimum Dependent Model

It is well-known that the entropy of L-S family is given by

$$H(\boldsymbol{X}|\boldsymbol{\theta}, \Sigma) = H(\boldsymbol{X}^o|\boldsymbol{\theta}) + \frac{1}{2}\log|\Sigma|, \tag{25}$$

where $H(\boldsymbol{X}^o|\boldsymbol{\theta})$ is free from $\boldsymbol{\mu}$ and Σ and $|\Sigma|$ denotes the determinant of Σ (Zellner 19??).

Let $M(\boldsymbol{X})$ denote the mutual information for a dependence relationship between the components of a d-dimensional random vector $\boldsymbol{X}$. The entropy decomposition (25) provides the following decomposition for the L-S family:

$$M(\boldsymbol{X}|\boldsymbol{\theta}, \Sigma) = M(\boldsymbol{X}^o|\boldsymbol{\theta}) + M(\Omega) \geq M_{\mathcal{G}}(\Omega), \tag{26}$$

where $M(\boldsymbol{X}^o|\boldsymbol{\theta})$ measures the intrinsic dependence of the orthogonal (unrotated) random vector $\boldsymbol{X}^o$, $\Omega = D^{-1/2}\Sigma D^{-1/2}$, $D = \text{Diag}[\sigma_{11}, \cdots, \sigma_{dd}]$, and $M(\Omega)$ is the portion of dependence induced by the rotation; for M, the marginal scales σ_i's, *cancel out*.

In the L-S family, the Gaussian distribution has following unique information properties.

- The maximum entropy model in the L-S family having the same scale matrix Σ is the Gaussian distribution whose entropy is given by (25) with $H(\boldsymbol{X}^o|\boldsymbol{\theta}) = \frac{d}{2}\log(2\pi e)$.
- The minimum dependence model in the L-S family having the same scale matrix Σ is the Gaussian distribution. In (26), $M(\Omega) = M_{\mathcal{G}}(\Omega)$ is the mutual information of the Gaussian model with correlation matrix Ω. That is, the Gaussian copula represents the minimal dependence structure among the copulas of all L-S models.

Table 2 gives some examples for the multivariate Gaussian model with correlation matrix Ω. The multivariate information index δ^2 is defined by using $M(\boldsymbol{X})$ in (16). The measures in the first row are for the dependence between all components. The measures in the second row are for dependence between two disjoint subvectors, where λ_j, $j = 1, \cdots, \min\{d_k\}$ denote the canonical correlations of the two subvectors $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ given by the nonzero eigenvalues of $\Omega_{11}^{-1/2}\Omega_{12}\Omega_{22}^{-1/2}\Omega_{12}\Omega_{11}^{-1/2}$, and Ω_{ij} are the partitions of Ω for $(\boldsymbol{X}_1, \boldsymbol{X}_2)$. The measures in the last row is for dependence between X_d and other components, where $\rho^2_{x_d|x_1,\cdots,x_{d-1}}$ is the squared multiple correlation.

Decomposition (26) gives the following representation for dependence information index (16):

$$\delta^2_{LS}(\boldsymbol{\theta}, \Sigma) = \delta^2_{\mathcal{G}} + (1 - \delta^2_{\mathcal{G}})\delta^2(\boldsymbol{\theta}), \tag{27}$$

where $\delta^2_{\mathcal{G}}$ is the corresponding multivariate normal information index. Representation (27) is insightful. For example in the bivariate case, $\delta^2_{\mathcal{G}}(\Omega) = \rho^2$, so $\delta^2_{LS}(\boldsymbol{\theta}, \rho)$ is given in terms of the Gaussian (linear) information index of dependence adjusted upward by the information index of the intrinsic dependence of the family.

The decomposition of the L-S dependence in terms of the copulas is insightful,

$$I[C_{LS}(\boldsymbol{\theta}, \Sigma)] = I[C(\boldsymbol{\theta})] + I[C_{\mathcal{G}}(\Omega)], \tag{28}$$

where $I[C_{\mathcal{G}}(\Omega)]$ is an information measure for the multivariate Gaussian copula. This decomposition gives the following insights:

(a) $I[C_{\mathcal{G}}(\Omega)]$ provides a measure of linear dependence and $I[C(\boldsymbol{\theta})]$ provides a measure for other aspects of the dependence intrinsic to the family.

Table 2. Examples of mutual information measures and Indices for Gaussian Dependence

$M(\boldsymbol{X})$	$M_{\mathcal{G}}(\Omega)$	Index $(\delta^2_{\mathcal{G}})$
$M(X_1, \cdots, X_d)$	$-.5 \log \lvert\Omega\rvert$	$1 - \lvert\Omega\rvert$
$M(\boldsymbol{X}_1, \boldsymbol{X}_2)$	$-.5 \sum_{j=1}^{\min\{d_k\}} \log(1-\lambda_j)$	$1 - \prod_{j=1}^{\min\{d_k\}} (1-\lambda_j)$
$M[X_d, (X_1, \cdots, X_{d-1})]$	$-.5 \log\left(1 - \rho^2_{\boldsymbol{x}_d\mid \boldsymbol{x}_1,\cdots,\boldsymbol{x}_{d-1}}\right)$	$\rho^2_{\boldsymbol{x}_d\mid \boldsymbol{x}_1,\cdots,\boldsymbol{x}_{d-1}}$

(b) In terms of entropy, (28) is an analysis of variance-type decomposition, where the total spread of the L-S copula C_{LS} is partitioned into the spread due to the spread of scale-free copula $C(\boldsymbol{\theta})$ and the spread due to the spread of rotation matrix $C_{\mathcal{G}}(\Omega)$.

4.2 Elliptical Families

The most important and widely used L-S models in dependence studies belong to the family of the elliptically contoured distributions defined by the pdf's of the form

$$f(\boldsymbol{x}|\Sigma, \boldsymbol{\mu}, h) = k|\Sigma|^{-1/2} h\Big((\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\Big),$$

where $h = h(\cdot, \boldsymbol{\theta})$ is referred to as the scale function which may include parameters $\boldsymbol{\theta}$ in addition to the scale matrix Σ (Fang et al. 2002). Micheas and Zografos (2006) gave an integral expression for the ϕ-divergence dependence measure of this family. However, other than M given by the KL divergence, the other ϕ-divergence measures do not give decompositions like (26).

Danaher and Smith (2011) discuss some advantages of elliptical copulas for applications and note that "dependence in the data is parameterized by a correlation matrix". However, this advantage stands only after a specific model has been selected. Since the elliptical distributions are in the L-S family, (26) applies and consequent decomposition of the mutual information are applicable to these distributions. Thus, the correlation only can estimate the portion of dependence due to the scale, hence underestimates the dependence of an elliptical distribution. We have illustrate this for the t family where $\boldsymbol{\theta}$ is the degrees of freedom ν; see Demarta and McNeil (2005) for a comprehensive review. In fact, the association measures also underestimate the dependence. Fang et al. (2002) showed that the Kendall's τ for all elliptical distributions is given by is the same as for the bivariate normal distribution. That is, Kendall's τ cannot distinguish between the members of these families which have various levels of the dependence.

4.3 Multivariate Normal Mixture

Consider the following well-known representation of t as the normal scale mixture:

$$\boldsymbol{Y} \overset{st}{=} \boldsymbol{\mu} + \Phi^{1/2}\boldsymbol{Z}, \tag{29}$$

where $\boldsymbol{Z} \sim N(\mathbf{0}, \Omega)$, and $\Phi^{-1} \sim G(\nu/2, \nu/2)$, $\nu = 1, 2 \cdots$ see, e.g., Demarta and McNeil (2005). The equivalent Bayesian formulation for (29) is

$$f(\boldsymbol{x}|\phi) = N(\boldsymbol{\mu}, \phi\Omega), \quad \phi \sim \text{Ga}(\nu/2, \nu/2), \ \nu = 1, 2 \cdots, \tag{30}$$

where Ω is a known correlation matrix; this representation is used by Carlin et al. (1992) and others. The predictive (marginal) distribution is multivariate-t, $f(\boldsymbol{x}) = t(\nu, \boldsymbol{\mu}, \Omega)$. Suppose that for a sequence of random variables $X_1, X_2, \cdots$, the conditional distribution of any vector of length n, given ϕ and the distribution of ϕ are as in (30). The predictive information for the unconditional sequence $M(\boldsymbol{X}, X_{n+1})$ is given by (28) with the L-S family parameter $\boldsymbol{\theta} = \nu$, $I[C(\nu)]$ which can be easily computed by using the expression for the entropy of t distribution (Zografos and Nadarajah 2005) in (12) and $I[C_{\mathcal{G}}(\Omega)]$ is given in the last row of Table 2, which is free from the scale ϕ. That is, the normal scale mixing by $G(\nu/2, \nu/2)$ increases dependence, $M(\boldsymbol{X}, X_{n+1}) > I[C_{\mathcal{G}}(\Omega)]$.

5 Index for Singular Models

When one variable is not completely dependent on the other, but a functional dependence is probable, $0 < P[X_1 = g_1(X_2)] = \pi < 1$ or $0 < P[X_2 = g_2(X_1)] = \pi < 1$, the joint distribution is singular. The most well known singular model is the Marshall-Olkin Bivariate Exponential (MOBE) distribution whose survival function has the following representation:

$$\bar{F}(x_1, x_2) = (1 - \pi)\bar{F}_a(x_1, x_2) + \pi\bar{F}_s(x_1, x_2). \tag{31}$$

Here, $\bar{F}_a(x_1, x_2)$ is the survival function with an absolutely continuous bivariate pdf

$$f_a(x_1, x_2) = \begin{cases} \dfrac{\lambda_1}{\lambda_1 + \lambda_2}\lambda(\lambda_2 + \lambda_3)\, e^{-\lambda_1 x_1 - (\lambda_2 + \lambda_3)x_2}, & x_1 < x_2 \\[2ex] \dfrac{\lambda_2}{\lambda_1 + \lambda_2}\lambda(\lambda_1 + \lambda_3)\, e^{-(\lambda_1 + \lambda_3)x_1 - \lambda_2 x_2}, & x_1 > x_2, \end{cases} \tag{32}$$

where $\lambda = \lambda_1 + \lambda_2 + \lambda_3$. This pdf is also known as the absolutely continuous bivariate exponential (ACBED) (Block and Basu 1974); $\bar{F}_s(x_1, x_2)$ is the survival for a singular part with a univariate pdf $f_s(x) = \lambda e^{-\lambda x}$ for $x_i = x_j$, $j \neq i = 1, 2$;

and $\pi = P(X_1 = X_2) = \dfrac{\lambda_3}{\lambda} > 0$. The marginal distributions are exponential with parameters $\lambda_i + \lambda_3, i = 1, 2$ and $\lambda_3 = 0$ implies the independent bivariate exponential model.

For the MOBE distribution, $\rho_p = \tau = \pi$ and $\eta^2_{i|j}$ can be calculated. However, more generally as in the absolutely continuous case, these indices cannot distinguish between singular distributions with same probability for a functional relationship and for some models ρ_p and $\eta^2_{i|j}$ are not even defined, e.g, when the marginals are Pareto. Due to the lack of absolute continuity, M is not directly applicable to these models, though an attempt has been made to use it for the copula of a MOBE distribution (Mercier et al. 2006). For example, for $\lambda_0 = \lambda_1 = \lambda_2$, using the entropy of MOBE (Nadarajah and Zografos 2005a, and Ebrahimi et al. 2007) and the marginal entropies $H(X_i|\lambda_0, \lambda_3) = 1 - \log(\lambda_0 + \lambda_3)$, $i = 1, 2$ in (12) we get the following expression:

$$M(X_1, X_2|\lambda_0, \pi) = \pi\left(1 - \log\frac{\lambda_0}{2}\right) + \log\frac{(1-\pi)\pi^{\pi}}{(1+\pi)^{1-\pi}}; \qquad (33)$$

calculation is shown in the Supplementary file. The second term is non-positive and equal to zero by continuity $0 \log 0 = 0$. Then, clearly $M(X_1, X_2|\lambda_0, \pi) = 0$ when $\pi = 0$ (the independent model) and negative for $\lambda_0 > 2$. In fact $M(X_1, X_2|\lambda_0, \pi) < 0$ for λ_0 just above .2 and all $\pi > 0$, and for $\lambda_0 \leq .2$ it has multiple roots without the variables being independent. Thus, the important properties of M such as non-negativeness fails and the measure does not identify independence. We should mention that (33) is different than the expression given in Mercier et al. (2006) for the Marshall-Olkin copula (Cuadras-Augè distribution, Cuadras and Augè 1981), however their expression is also negative.

5.1 Generalized Dependence Information Index

As we illustrated, a direct application of the mutual information to singular distributions is not feasible. We propose an indirect application of M for singular distributions with representation (31) through a modification of the index (16). We apply the probabilistic argument of Marshall and Olkin (1967) to dependence between X_1 and X_2. Let $\mathcal{S}_s = \{(x_1, x_2) : x_i = g_i(x_j),\ j \neq i = 1, 2\}$ and $\mathcal{S}_a = \{(x_1, x_2)\}$ such that:

$$P(\mathcal{S}_s) = P[X_i = g_i(X_j)] = \pi > 0, \quad j \neq i = 1, 2 \qquad (34)$$

$$P(\mathcal{S}_a) = P[X_i \neq g_i(X_j)] = 1 - \pi > 0, \quad j \neq i = 1, 2. \qquad (35)$$

By (34) and (35), the local dependence measures are $M(X_1, X_2|\mathcal{S}_s) = \infty$ and $M(X_1, X_2|\mathcal{S}_a) < \infty$. That is, for a singular distribution, $M(X_1, X_2) = \infty$ with probability π and $M(X_1, X_2) < \infty$ with probability $1 - \pi$; details are given in the Appendix. Similarly, in terms of the information indices $\delta^2(X_1, X_2) = 1$ and $\delta^2(X_1, X_2) < 1$ with probabilities π and $1 - \pi$. This interpretation suggests the following modification of the index (16) for the singular distributions:

$$\delta^2_{\pi}(X_1, X_2) = \pi + (1 - \pi)\delta^2_a(X_1, X_2), \qquad (36)$$

where $\delta_a^2(X_1, X_2)$ is the dependence information index for the absolutely continuous distribution $\bar{F}_a$. This index is the average of indices for the singular and absolutely continuous parts of F, $\{1, \delta_a^2(X_1, X_2)\}$, with probabilities given by (34) and (35). Here, π adjusts $\delta_a^2(X_1, X_2)$ for the singularity. The adjustment factor π can be interpreted as a shrinkage factor for $\delta_s^2(X_1, X_2) = \delta(\infty) = 1$ toward $\delta_a^2(X_1, X_2)$. It can also be interpreted reversely, as the inflater for $\delta_a^2(X_1, X_2)$ toward $\delta^2(\infty) = 1$, as well. Note that $\delta_\pi^2(X_1, X_2) = \delta_\infty^2(X_1, X_2) = 1$ if $\pi = 1$ and $\delta_\pi^2(X_1, X_2) = \delta_a^2(X_1, X_2)$ if $\pi = 0$ which is the case when $F \ll F_1 F_2$ where $\delta_a^2(X_1, X_2) = 0$ if and only if the variables are independent. Also note that when the absolutely continuous part is an independent model, we have $\delta_a^2(X_1, X_2) = 0$ and $\delta_\pi^2(X_1, X_2) = \pi$.

The modified index (36) inherits the following important properties of (16):

- $\delta_\pi^2(X_1, X_2) \geq 0$, where the equality holds if and only if X_1 and X_2 are independent.
- If $Y_i = \phi_i(X_i)$, $i = 1, 2$ is a one-to-one transformation, then $\delta_\pi^2(Y_1, Y_2) = \delta_\pi^2(X_1, X_2)$.

The invariance under one-to-one transformations is noted from the invariance of the volume under the change of variables in the first integral in (A.2) in the Appendix.

The modified index δ_π^2 takes the departure of the absolutely continuous part from the independence into account through δ_a^2, thereby enabling us to rank the dependence among various singular models that have the same probability for the singular part. Figure 5 illustrates comparison of δ_π^2 for the MOBE model and another well known singular model having the same exponential pdf for the singular part. As a benchmark for the comparison, we have included the model with the independent bivariate distribution for the absolutely continuous part and a singular part with probability π. This model is from Christensen et al. (2010) where it is used as the prior for a Bayesian test of a sharp hypothesis which will be discussed in sequel. Since $f_a(x_1, x_2)$ is an independent model, $\delta_a^2(X_1, X_2) = 0$, hence $\delta_\pi^2 = \pi$, the straight line. All three models have the same Kendall's index, $\tau = \pi$, so this index cannot distinguish between them.

The nearly straight line in Figure 5 is $\delta_\pi^2(X_1, X_2)$ for the MOBE model. This index is computed using M for the ACBED distribution (32):

$$M_a(X_1, X_2) = \pi + \sum_{k=1}^{\infty} \frac{1}{k} \left([\Psi_1(-k\lambda_2) - w_1]\alpha_1^k + [\Psi_2(-k\lambda_1) - w_2]\alpha_2^k - \pi^k \right), \quad (37)$$

where

$$\alpha_i = \frac{\lambda_3}{\lambda_i + \lambda_3}, \quad w_i = \frac{\lambda_i}{\lambda_1 + \lambda_2}, \quad \Psi_i(-k\lambda_j) = E\left(e^{-k\lambda_j X_i}\right), \quad j \neq i = 1, 2,$$

is the moment generating function corresponding to $f_a(x_1, x_2)$ evaluated at $-k\lambda_j$ (Block and Basu 1974); details are given in the Supplementary file.

The curved that dominates the MOBE index in Figure 5 is $\delta_\pi^2(\boldsymbol{X}_n, X_{n+1})$ for the following exponential autoregressive (EAR) process

$$X_{n+1} = \rho X_n + \epsilon_{n+1}, \quad (38)$$

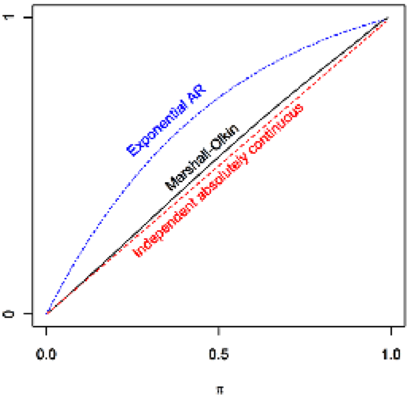


Fig. 5. Plots of the generalized information indices for the independent absolutely continuous, Marshall-Olkin ($\lambda_1 = \lambda_2$), and the exponential autoregressive distributions

where $\{X_n\}$ is a sequence of identically distributed exponential random variables $P(X_n > x) = \bar{F}(x) = e^{-\lambda x}$ and $\{\epsilon_n\}$ is an iid sequence, and ϵ_{n+1} and X_n are independent. Since the dependence structure of the EAR model is Markovian, the information provided by a sequence of n observations $\boldsymbol{X}_n$ for prediction of the next one is given by $M(\boldsymbol{X}_n, X_{n+1}) = M(X_n, X_{n+1})$. The series $\{X_n\}, n = 1, 2, \cdots$ is stationary, but the distribution of (X_n, X_{n+1}) which is singular (Gaver and Lewis 1980). The bivariate survival function is in the form of (31) with $\pi = \rho < 1$,

$$f_a(x_n, x_{n+1}) = \lambda^2 e^{-\lambda(1-\rho)x_n - \lambda x_{n+1}}, \quad x_{n+1} > \rho x_n$$
$$f_s(x_n) = \lambda e^{-\lambda x_n}, \quad x_{n+1} = \rho x_n.$$

For the absolutely continuous part,

$$M_a(X_n, X_{n+1}) = \rho + \log(1-\rho) + \sum_{k=1}^{\infty} \frac{\rho}{k[1 + k(1-\rho)][\rho + k(1-\rho)]}; \qquad (39)$$

details are given in the Supplementary file, where it also is shown that $\pi = \rho = \tau$.

By the invariance property of (36), the information dependence indices of the MOBE and EAR models are applicable to transformations of $X_i, i = 1, 2$. Some well-known transformations of the MOBE are $Y_i = X_i^{1/\alpha_i}$, $i = 1, 2$ which gives Weibull marginals (Marshall and Olkin 1967, Moeschberger 1974, Lee 1979), $Y_i = e^{X_i}$, $i = 1, 2$ (Muliere and Scarsini 1987), and the Cuadras-Augè copula (Cuadras and Augè, 1981) also known as the Marshall-Olkin copula. Further applications include a family of lifetime models, referred to as the time-transformed exponential (Barlow and Hsiung, 1983).

6 Summary and Conclusions

The information index of dependence δ^2 was originally motivated by Linfoot (1957) through a mathematical generalization of the bivariate normal $\rho_p^2 = \rho^2$. We summarized the information notion of dependence from its most basic level in terms of the interpretations of ρ_p^2 for dependence of the most widely used and easily understood model. In the Gaussian world, where the reference points for dependence are orthogonality and linear dependence, ρ_p^2 is both, an index of departure of the joint distribution from the independence, and the utility index in terms of variance reduction by using one variable to predict the other. In doing so, we synthesized seemingly unrelated classics, Lancaster (1963) and Kimeldorf and Sampson (1978) with Lindley (1956), DeGroot (1962), and Bernardo (1979) for Bayesian analysis. This excursion led us to the mutual information, the unique measure that is endowed by the interpretations of the bivariate normal ρ_p^2 as well as being the shared information between the variables. This measure is applicable to all absolutely continuous distributions, discrete distributions, and distributions of categorical variables. We have shown that its immediate generalizations by Rènyi's measures do not provide such a unique measure of dependence, even for the normal model.

In addition to its theoretical merits, we illustrated the efficacy of M for practice. Dependence between two variables is a more general notion than association between them and more complex than the variance reduction through conditioning. An index of departure from independence such as δ^2 is needed for a "common metric" in studies that involve comparison of dependence between and within different families of distributions. We illustrated rankings of dependence between and within the bivariate Gaussian, t, Cauchy, Pareto, and F-G-M families along with four families from the recent literature which have pdf's in the form of a generalized Sarmanov family. For all distributions in the Gaussian and t families, ρ_p, τ, and ρ_s remain the same function of ρ while the departures from the independence vary considerably, and $\eta^2_{i|j}$ is not defined for all members (Table 1). Through δ^2 we gained insights about the Student-t family: the range of dependence is narrow and gradation of dependence is coarse when the degrees of freedom is low and the variables are near orthogonal. Expansion to the entire Pearson VII family corrects both issues (Figure 2). For the same τ, the dependence of the Clayton copula is bounded below by the Gaussian copula and above by the Cauchy copula. Similar examples were shown with the F-G-M copula (Figure 3). The dependence of distributions in the location-scale family decomposes additively into two parts, one measuring the dependence with identity scale matrix and one measuring the Gaussian dependence. Among all distributions in the multivariate L-S family having the same scale matrix, the multivariate Gaussian copula represents the minimal dependence structure. Consequently, the association measures underestimate the dependence for the important class of elliptical families.

We close with some statements of limitations and proposals. Association between variables is an important notion with important consequences for practice, as well. As a measure of departure from the independence δ^2 cannot serve this purpose and indices such ρ_p, ρ_s, and τ are needed. Using δ^2 along with the sign of τ or ρ_s for the departure and its direction from the independence. The mutual information is not applicable to singular distributions. We noted an improper use, though for the purpose of motivating this measure. We illustrated the lack of applicability and proposed a modified information index for singular distributions. We showed an application of the modified index through ranking dependence of two singular distributions with the same probability for the singularity π and the same Kendall's $\tau = \pi$: the dependence of the most celebrated singular model, MOBE, is just a bit stronger than that for the independent model for the absolutely continuous part and the dependence of EAR model uniformly dominates the dependence of the MOBE model.

A. Appendix

Proof of Equation (14)

First note that

$$\begin{aligned}
K(F:G_1G_2) &= \int\int f(x_1,x_2)\log\frac{f(x_1,x_2)}{g_1(x_1)g_2(x_2)}dx_1dx_2\\
&= \int\int f(x_1,x_2)\log\frac{f(x_1,x_2)f_1(x_1)f_2(x_2)}{f_1(x_1)f_2(x_2)g_1(x_1)g_2(x_2)}dx_1dx_2\\
&= M(F)+\int\int f(x_1,x_2)\log\frac{f_1(x_1)f_2(x_2)}{g_1(x_1)g_2(x_2)}dx_1dx_2\\
&= M(F)+\int\int f(x_1,x_2)\log\frac{f_1(x_1)}{g_1(x_1)}dx_1dx_2\\
&\quad +\int\int f(x_1,x_2)\log\frac{f_2(x_2)}{g_2(x_2)}dx_1dx_2\\
&= M(F)+K(F_1:G_1)+K(F_2:G_2). \qquad (A.1)
\end{aligned}$$

The invariance property of (7) allows transforming $K(F:G_1G_2)$ by $U_i = G_i(X_i)$ and $M(F)$ by $V_i = F_i(X_i)$ and obtain (14). The inequality is from $K(F_i:G_i) \geq 0$, where the equality holds if and only if $g_i(x) = f_i(x)$ almost everywhere. Extension to multivariate is straightforward. Mercier and Inglada (2008) give (A.1) incorrectly in terms of $K(G_i:F_i)$.

Technical details for (36)

For the bivariate distribution with survival function (31),

$$\begin{aligned} P(\mathcal{S}_a) &= \int_{\Re}\int_{\Re} f_\mu(x_1,x_2)dx_1dx_2 = 1-\pi > 0, \\ P(\mathcal{S}_s) &= \int_{\Re} f_\mu(x)dx = \pi > 0, \end{aligned} \tag{A.2}$$

where f_μ is the density relative to a measure μ defined by Bemis et al. (1972) as follows. For any $\mathcal{A} \subseteq \mathcal{S}$, $\mu(\mathcal{A}) = \mu_1([\mathcal{A} \cap \mathcal{S}_s]_p) + \mu_2(\mathcal{A})$, where μ_1 and μ_2 are one and two dimensional Lebesgue measures and the subscript p denotes the projection of the set onto the x_1-axis. For any $\mathcal{A} \subseteq \mathcal{S}$ with $P(\mathcal{A}) = \pi \in (0,1)$, let $f(x_1,x_2|\mathcal{A}) = \frac{1}{\pi}f(x_1,x_2)$ be the conditional (truncated) joint pdf and denote its marginals by

$$f(x_j|\mathcal{A}) = \begin{cases} \frac{1}{\pi}\int_{\mathcal{A}_i} f(x_j,x_i)dx_i, & i \neq j, \quad (x_1,x_2) \in \mathcal{A} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{A}_i$, $i = 1,2$ denotes the projection of $\mathcal{A}$ on the x_i-axis. The mutual information of the truncated distribution, $M(X_1,X_2|\mathcal{A})$, is given by (10) with pdf's $f(x_1,x_2|\mathcal{A})$ and $f_i(x_i|\mathcal{A})$, $i = 1,2$. If $F \ll F_1F_2$, then the decomposition of the KL information gives:

$$M(X_1,X_2) = K(\boldsymbol{p},\boldsymbol{p}^*) + \pi M(X_1,X_2|\mathcal{A}) + (1-\pi)M(X_1,X_2|\mathcal{A}^c), \tag{A.3}$$

where $K(\boldsymbol{p},\boldsymbol{p}^*)$ is the information divergence between two Bernoulli distributions:

$$\boldsymbol{p} = (\pi, 1-\pi), \quad \boldsymbol{p}^* = (\pi^*, 1-\pi^*), \quad \pi^* = P_{f_1f_2}(\mathcal{A}) = \int\int_{\mathcal{A}} dF_1(x_1)dF_2(x_2).$$

$M(X_1,X_2|\mathcal{A})$ and $M(X_1,X_2|\mathcal{A}^c)$ are measures of *local* dependence, with the associated probabilities π and $1-\pi$. Now let $\mathcal{A} = \mathcal{S}_s$ in (34) and $\mathcal{A}^c = \mathcal{S}_a$ in (35).

References

Abe, S. and Rajagopal, A. K. (2001), "Information Theoretic Approach to Statistical Properties of Multivariate Cauchy-Lorentz Distributions," *J. Phys. A: Math. and General*, 34, 8727-8731.

Balakrishnan, N. and Lai, C-D. (2009). *Continuous Bivariate Distributions*, 2nd Ed. N. Y.: Springer.

Barlow, R.E. and Hsiung, J.H. (1983), "Expected Information from a Life Test Experiment," *The Statistician*, 48, 18-21.

Bedford, T. and Cooke, R.M. (2002), "A New Graphical Model for Dependent Random Variables," *Ann. Statist.*, 30, 1031-1068.

Bemis, B.M., Bain, L.J., and Higgins, J.J. (1972), "Estimation and Hypothesis Testing for the Parameters of Bivariate Exponential Distribution," *J. Amer. Statist. Assoc.*, 67, 927-929.

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *Ann. Statist.*, 7, 686-690.

Block, H. and Basu, A.P. (1974), "A Continuous Bivariate Exponential Extension," *J. Amer. Statist. Assoc.*, 69, 1031-1037.

Calsaverini, R.S. and Vicente, R. (2009), An Information-Theoretic Approach to Statistical Dependence: Copula Information," *Europhysics Letter*, 88 doi: 10.1209/0295-5075/88/68003.

Card, S. W. and Mohan, C. K. (2009), "An Application of Information Theoretic Selection to Evolution of Models with Continuous-valued Inputs," *Genetic Programming Theory and Practice VI: Genetic and Evolutionary Computation*, 1-14, DOI: 10.1007/978-0-387-87623-8_3.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling," *J. Amer. Statist. Assoc.*, 87, 493-500.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T.E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press.

Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, New York: Wiley.

Cuadras, C.M. and Augè, J. (1981), "A Continuous General Multivariate Distribution and its Properties," *Comm. Statist. – Theo. Meth.*, 10, 339-353.

Danaher, P. J. and Smith, M. S. (2011), "Modeling Multivariate Distributions Using Copulas: Applications in Marketing," *Marketing Science*, 30, 4-21.

Darbellay, G.A. and Vajda, I. (2000), "Entropy Expressions for Multivariate Continuous Distributions," *IEEE Trans. Info. Theory*, 46 , 709-712.

DeGroot, M. H. (1962), "Uncertainty, Information, and Sequential Experiments," *Ann. Math. Statist.*, **33** 404-419.

Demarta, S. and McNeil, A.J. (2005), "The *t* Copula and Related Copulas," *International Statistical Review*, 73, 111-129.

De Paula, A. (2008), "Conditional Moments and Independence," *Amer. Statist.*, 62, 219-221.

Drouet-Mari, D. and Kotz, S. (2001), *Correlation and Dependence.* London: Imperial College Press.

Ebrahimi, N., Kirmani, S.N.U.A. and Soofi, E.S. (2007), "Dynamic Multivariate Information," *Journal of Multivariate Analysis*, 98, 328-349.

Ebrahimi, N., Soofi, E.S., and Soyer, R. (2010a), "On the Sample Information about Parameter and Prediction," *Statistical Science*, 25, 348-367.

Ebrahimi, N., Soofi, E.S., Soyer, R. (2010b), "Information Measures in Perspective," *International Statistical Review*, 78, 383-412.

Ebrahimi, N., Soofi, E.S., Zhao, S. (2011), "Information Measures of Dirichlet Distribution with Applications," *Applied Stochastic Models in Business and Industry*, 27, 131-150.

Ebrahimi, N., Hamedani, G.G., Soofi, E. S., and Volkmer, H. (2010c), "A Class of Models for Uncorrelated Random Variables," *Journal of Multivariate Analysis*, 101, 1859-1871.

Fang, H-B., Fang, K-T., and Kotz. S. (2002), "The Meta-elliptical Distributions with Given Marginals," *Journal of Multivariate Analysis*, 82, 1-16.

Frees, E.W. and Valdez, E.A.(2008), "Hierarchical Insurance Claim Modeling," *J. Amer. Statist. Assoc.*, 103,1457-1469.

Gaver, D.P. and Lewis, P.A.W. (1980), "First-order Gamma Sequences and Point Processes," *Advances in Applied Probability*, 12, 727-745.

Golden, L. L., Brockett, P. L, and Zimmer, M. R. (1990), "An Information Theoretic Approach for Identifying Shared Information and Asymmetric Relationships Among Variables," *Multivariate Behavioral Research* 25, 479-502.

Granger, C. W., Maasoumi, E, and Racine J. (2004), "A Dependence Metric for Possibly Nonlinear Processes," *J. Time Ser. Anal.*, 25, 649-669.

Hirschberg, J., Maasoumi, E., and Slottje, D.J. (1991), "Cluster Analysis and the Quality of Life Across Countries," *J. Econometrics*, 50, 131-150.

Joe, H. (1989), "Relative Entropy Measures of Multivariate Dependence," *J. Amer. Statist. Assoc.*, 84, 157-164.

Kimeldorf, G. and Sampson, A.R. (1978), "Monotone Dependence," *Annals of Statist.*, 6, 895-903.

Kotz, S. and van Dorp, J. R. (2010), "Generalized Diagonal Band Copulas with Two-Sided Generating Densities," *Decision Analysis*, 7, 1-19.

Kullback, S. (1959). *Information Theory and Statistics*, N.Y.: Wiley (reprinted in 1968 by Dover).

Kullback, S. and Leibler, R. A. (1951), "On Information and Sufficiency", *Annals of Mathematical Statistics*, 22, 79-86.

Kundu, D. and Gupta, R. (2010), "Modified Sarhan-Balakrishnan, Singular Bivariate Distribution," *Journal of Statistical Planning and Inference*, 140, 526-538.

Lancaster, H. O. (1963), "Correlation and Complete Dependence of Random Variables," *Annals of Mathematical Statistics*, 34, 1315-1321.

Lee, L. (1979), "Multivariate Distributions Having Weibull Properties," *Journal of Multivariate Analysis*, 9, 267-277.

Lindley, D.V. (1956), "On a Measure of Information Provided by an Experiment," *The Annals of Mathematical Statistics*, 27, 986-1005.

Linfoot, E. (1957), "An Informational Measure of Correlation," *Information and Control*, 1, 85-89.

Marshall, A. W. and Olkin, I. (1967), "A Multivariate Exponential Distribution," *J. Amer. Statist. Assoc.*, 62, 30-44.

Mercier, G. and Inglada, J. (2008). "Change Detection with Misregistration Errors and Heterogeneous Data Through the Orfeo toolbox," Technical Report of TELECOM Bretagne # RR-2008003-ITI. (http://www.gregoire-mercier.fr/articles/2008/RR-2008003-ITI.pdf)

Mercier, G., Derrode, S., Pieczynski, W., Nicolas, J-M., Joannic-Chardin, A., and Inglada, J. (2006), "Copula-based Stochastic Kernels for Abrupt Change Detection," *IEEE International Conference on Geoscience and Remote Sensing Symposium*, 204-207.

Moeschberger, M.L. (1974), "Life Tests Under Dependent Competing Causes of Failure," *Technometrics*, 16, 39-47.

Muliere, P. and Scarsini, M. (1987), "Characterization of a Marshall-Olkin Type Class of Distributions," *Annals of Institute of Mathematical Statistics*, 39, 429-441.

Nadarajah, S. and Zografos, K. (2003), "Formulas for Rènyi Information and Related Measures for Univariate Distributions," *Inform. Sci.*, 155, 119138.

Nadarajah, S. and Zografos, K. (2005), "Expressions for Rènyi and Shannon Entropies for Bivariate Distributions," *Inform. Sci.*, 170, 173-189.

Nelson, R.B. (2006), *An Introduction to Copulas, 2nd ed.*, Springer.

Principe, J.C. (2010), *Information Theoretic Learning: Rènyi's Entropy and Kernel Perspective*, Springer, New York.

Rènyi, A. (1961), "On Measures of Entropy and Information," *Proc. Fourth Berkeley Symp.*, 1, pp.547-561. Berkeley: UC Press.

Retzer, J.J., Soofi, E.S., and Soyer R. (2009), "Information Importance of Predictors: Concepts, Measures, Bayesian Inference, and applications," *Comput. Statist. Data Anal.*, 53, 2363-2377.

Sarhan, A.M. and Balakrishnan, N. (2007), "A New Class of Bivariate Distributions and its Mixture," *Journal of Multivariate Analysis*, 98, 1508-1527.

Seth, S., Rao, M, Park, Il and Principe, J.C. (2011), ”A Unified Framework for Quadratic Measures of Independence," *IEEE Transactions on Signal Processing*, 59, 3624-3635.

Shannon, C. E. (1948), "A Mathematical Theory of Communication," *Bell Sys. Tech. J.*, 27, 379-423.

Sims, C.A. (2010), ”Rational Inattention and Monetary Economics," to appear in *Handbook of Monetary Policy*, Elsevier.

Sklar, A. (1959). Fonctions de rèpartition à n dimensions et leurs marges. *Publications de lÌnstitut Statistique se lÙniversitè Paris*, 8, 229-231.

Smith, M. S., Min, A., Almeida, C., and Czado, C. (2010), "Modeling Longitudinal Data using a Pair-Copula Decomposition of Serial Dependence," *J. Amer. Statist. Assoc.*, 105, 1467-1479.

Szèkely, G.J., Rizzo, M.L., and Bakirov, N.K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *Ann. Statist.*, 35, 2769-2794.

Theil, H. and Chung, C. (1988), "Information-theoretic Measures of Fit for Univariate and Multivariate Linear Regressions," *The American Statistician*, 42, 249-252.

Tourassia, G.D., Frederick, E.D., Markey, M.K., and Floyd, C. E. J. (2001), "Application of the Mutual Information Criterion for Feature Selection in Computer-aided Diagnosis," *Medical Physics*, 28, 2394-2402.

Tsallis, C. (1988), "Possible Generalization of Bolzmann-Gibbs Statistics," *J. Statist. Phys.*, 52, 479-487.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley: New York (reprinted in 1996 by Wiley)

Zografos, K. and Nadarajah, S. (2005), "Expressions for Rènyi and Shannon Entropies for Multivariate Distributions," *Statistics & Probability Letters*, 71, 71-84.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 329-344

MODELLING DIFFUSION OF INNOVATIONS WITH HOMOGENEOUS AND HETEROGENEOUS POPULATIONS

Md. Abud Darda, Renato Guseo and **Cinzia Mortarino**
Department of Statistical Sciences, Via C. Battisti 241, 35121 Padova, Italy
E-mail: darda@stat.unipd.it, guseo@stat.unipd.it, mortarino@stat.unipd.it

ABSTRACT

Diffusion of innovation can be defined as a system of adoption by which the innovation of a new product or service is spread among the members or agents in a social system. Diffusion models help to forecast the demand and work as a decision aid in making pre-launch, launch and help post-launch strategic choices. This complex system can be approached, in aggregate form, as a function of the number of adopting agents or the number of adoptions, as a function of time and controlling covariates or as a function of the time and spatial aspects simultaneously in the related structure. The process can also be described with simultaneous and dynamic models depending upon whether innovations are independent, complementary or substitutes. The present study offers a framework for systematizing the diffusion models in the perspective of existing homogeneity or heterogeneity among the agents. Different models are compared, new models are proposed and their advantages and disadvantages discussed with reference to the Algerian natural gas production time series. Some guidelines for further research extensions are also suggested.

1. INTRODUCTION

In recent decades, the pattern of diffusion of innovations for products or services has become an interesting matter of study for social scientists, mathematicians, marketing experts, statisticians, engineers and biologists. Researchers are drawn to the topic not only to examine trends and underlying factors in the diffusion process but also for forecasting purposes. In this context, researchers try to understand the behaviour of the existing individuals (agents) in society and their attitude towards newly introduced goods or services and explain them with special mathematical models. This attitude can be termed “adoption” the marketing language that divides the existing non-homogeneous agents into several mutually exclusive groups. Bass (1969) considers two sub-populations of adopters, innovators and imitators, and develops an aggregate model for the adoption assuming homogeneity in the sub-population units. This paradigm of innovation diffusion modelling proceeds with further diversification of understanding the process also followed by a number of studies and reviews (see Mahajan et al. (1990), Bass (2004), Meade and Islam (2006)). The important thing here is to model the social contagion and adoption of goods/services in a social system that is characterised by specific regime changes that cumulatively follow a sigmoid shape. This system of adoption also depends on the structure of the social system, on its internal rules or external influences (policies, marketing strategies etc.) that may be considered to vary with respect to time. Rogers and

Shoemaker (1971) define the diffusion of innovation as the process by which innovation spreads among the members of a social system. The innovation itself, adopters of the innovation, innovation channels, time and space, change agents and the social system dynamics seem to be associated with this process. Various modelling approaches were followed by the researchers to obtain time patterns of the diffusion process. The fundamental approach is to consider the diffusion process as a direct function of time. The alternative approach is to consider the process as a function of the number of previous adopters over time through special differential equations which may include theoretical assumptions and therefore, easy-to-interpret parameters. Others extended the fundamental diffusion model to study the time and spatial aspects of the diffusion process simultaneously. Some researchers also attempted to model a simultaneous and dynamic diffusion model, depending upon whether innovations are independent, complementary or substitutes.

The objective of the present study is to conduct an in-depth study of the existing diffusion of innovation models with both homogeneity and heterogeneity assumptions in the population and make a valid comparison of their parameter estimates. In Section 2, a short discussion on the standard Bass model and its extensions is presented. Section 3 contains the discussion of diffusion of innovation models appropriate for a heterogeneous population with special emphasis on Gamma-shifted Gompertz models and their extensions. Model parameter estimates and their comparisons are discussed in Section 4. Section 5 is devoted to the analysis of the Algerian natural gas production within the logic of a diffusion of innovation process. Some interesting results are provided with reference to the evolving dynamics, peak time and reserves estimation. Finally, an overall discussion and further extension guidelines are presented in Section 6.

2. THE BASS MODEL: A PARADIGM OF A HOMOGENOUS DIFFUSION MODEL

The fundamental diffusion model by Bass (1969) is based on the assumptions that the probability of adoption of a new product or innovation at time *t* given that it has not yet been adopted would depend on a convex combination of two factors: the number of independent initial adopters and the number of existing adopters (i.e., imitators). The innovation coefficient measures the propensity of potential adopters to become adopters, and the imitation coefficient measures the propensity of potential adopters to imitate previous adopters. The Bass model was built on Roger’s conceptual framework by developing a mathematical model that captures the non-linear structure of the S-shaped curve (Robertson et al. (2007)). Introducing the coefficient of innovation *p (p>0)* and the coefficient of imitation *q (q>0)*, the Bass model can be described by the following equation:

$$f(t) = [p + qF(t)][1 - F(t)] \tag{1}$$

where *F(t)* depicts the distribution over time of adoptions and *f(t)=F(t)* denotes the corresponding density of the adoption process over time. Under the initial condition *F(0)=0,* its solution (Bass (1969)) defines the following distribution function:

$$F(t) = 1 - e^{-(p+q)t} \Big/ \left(1 + \frac{q}{p} e^{-(p+q)t}\right), \qquad t \geq 0 \tag{2}$$

Let *m* be the number of potential adopters (or adoptions) in the market. Then the total number of adoptions until time *t* is therefore obtained as follows:

$$Y(t) = mF(t) = m\frac{1 - e^{-(p+q)t}}{\left(1 + \frac{q}{p} e^{-(p+q)t}\right)}; \qquad t \geq 0$$

Considering the time unit as unity (year, quarter, month, week, days etc.), the rate of diffusion, in other words, let's say, sales *S(t)* in the time interval $t-1, t$, is given by the following:

$$\begin{aligned} S(t) &= mf(t) + \varepsilon(t) \\ &\approx m[F(t) - F(t-1)] + \varepsilon(t) \\ &= m\left[\frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}} - \frac{1 - e^{-(p+q)(t-1)}}{1 + \frac{q}{p} e^{-(p+q)(t-1)}}\right] + \varepsilon(t) \end{aligned}$$

Considering $\varepsilon\,(t)$, the error term in the equation as distributed with variance σ^2, the parameters *p*, *q* and *m* can be estimated by the non-linear least squares (NLS) procedure (Srinivasan and Mason (1986)).

A better approximation for *f(t) = F´(t)* can also be obtained through the following:

$$f(t) \approx F(t+0.5) - F(t-0.5)$$

The Bass model is the first and foremost formal way to separate the innovators (leaders) and imitators (followers) in an innovation process that better explains Roger's perspective based on a normal distribution assumption. Innovators and imitators characterise a latent distinction, since the observed data just report on the adoption of a susceptible agent without any other specification.

A very important extension of the Bass model developed by Bass et al. (1994) is based on the Generalised Bass Model (GBM), which introduces a general time dependent intervention function *x(t)*, able to take into account the possible effect of the exogenous variables on the diffusion process. Thus, an extension of Equation (1), under the initial condition *F(0)=0*, is given by the following:

$$f(t) = [1 - F(t)][p + qF(t)]x(t) \tag{3}$$

and its solution gives an expression for the total number of adopters until time *t* as follows:

$$Y(t)=m\frac{\left\{1-e^{-(p+q)\int_0^t x(\tau)d\tau}\right\}}{\left\{1+\frac{q}{p}e^{-(p+q)\int_0^t x(\tau)d\tau}\right\}}=mF(t), \qquad 0\le t\le +\infty \tag{4}$$

Bass et al. (1994) called this function *x(t)* the "current marketing effort" that reflects the current effect of dynamic marketing variables on the conditional probability of adoption at time *t*. Notice that the closed-form solution (4) is extremely general, because the control function *x(t)* may assume, under local integrability, any shape without special limitations. For *x(t)=1*, the model reduces to the standard Bass model, and for *x(t)>1*, the adoption process is accelerated over time; otherwise, it is delayed (Guseo et al. (2007) and Dalla Valle & Furlan (2011)). Therefore, this intervention function may modify the time elapsing between adoption events within a general closed-form solution very powerfully in applied contexts.

The basic Bass model fits very well to real data, and many other versions of the model appeared later to explain different aspects of diffusion. A special application of the GBM has been made in the energy sector, crude oil in particular (Guseo & Dalla Valle (2005) and Guseo et al. (2007), Guseo (2011)) where the rationale for these applications is grounded on the related diffusion technologies that are directly or indirectly energy consuming.

Guseo et al. (2007) model the intervention function *x(t)* through some exponential shocks, under the assumption that the memory effect has a non-uniform distribution over time. Thus, the function *x(t)* can be defined by the following:

$$x(t)=1+c_1e^{b_1(t-a_1)}I_{t\ge a_1}+c_2e^{b_2(t-a_2)}I_{t\ge a_2}$$

where a_i *(i=1,2)* denotes the starting times of exponential shocks, b_i *(i=1,2)* describes the effect's persistence and c_i *(i=1,2)* controls the intensity of perturbations. For the values of parameter $b_i<0$, the process is mean reverting (i.e., the memory is decaying to the stationary position; in other words, *x(t)=1*). If $b_i>0$, the process introduces a permanent acceleration in the saturation of a life cycle.

Despite recent developments, the Bass model still suffers conceptual limitations in applications and forecasting. It assumes that the internal influence (word-of-mouth effect) remains uniform over time frame over the diffusion process period. Conversely, in practice, the later adopters are not as likely to discuss the product with non-adopters as are the early adopters, and they are also less likely to exhibit the same enthusiasm in discussing the new product. In other situations, the internal influence becomes increased due to the influence of the reluctance of later adopters to the word-of-mouth effects. In many occurrences, the late adopters have different characters than the early adopters and would respond differently (Rogers (2003)) and the diffusion model should allow for this. Therefore, researchers have suggested a number of alternative structures to the intervention function *x(t)* to comply with the existing population structures modelled with the shocks as exponential or rectangular or both types in the observed dataset.

Bass models, BM and GBM, have a fixed market potential over the assumed life cycle. An important extension, the dynamic market potential, *m(t)*, is introduced in Guseo (2004) and Guseo and Guidolin (2009, 2010, 2011). In particular, Guseo and Guidolin (2009) obtain a Riccati closed-form solution for general *m(t)* and *x(t)* functions that emphasizes the different role of policies over time *x(t)* and over scale *m(t)* in order to describe the time modulation of a non-constant carrying capacity (market potential).

3. A NON-HOMOGENOUS DIFFUSION MODEL: THE GAMMA SHIFTED GOMPERTZ MODEL

In recent days, considering the heterogeneity of the population, a limited number of diffusion models have been introduced that incorporate individual-level heterogeneity and/or heterogeneity in the diffusion penetration rate. Researchers try to develop a segmental diffusion model (Robertson et al. 2007), or models considering several distributional assumptions of market penetration rate and adoption at the individual level (Bemmaor (1994); Van den Bulte and Lilien (1997); Gutierrez-jaimez et al. (2007)). The principal matter of interest in this case is to obtain a parsimonious and flexible closed-form diffusion model that can accommodate both symmetric and non-symmetric diffusion patterns with a point of inflection that can occur at any stage of the diffusion process (Mahajan and Wind (1986)).

The research paradigm on diffusion of innovation in a social system by Bass (1969) and Mansfield (1961) and their generalisations addressed the market as an aggregate structure, with little attention to micro-level processes that characterise adoption decisions (Chatterjee and Eliashberg (1990); Mahajan et al. (1990)). The main issue in this line is to understand and explain the diffusion process across a population of adopting units. The existence of a heterogeneous population of adopters has been largely ignored in this perspective. In the individual-level perspective, the diffusion of innovation can either be modelled as individual adoption probability with the timing of adoption or derivation of adoption behaviour at the individual level in a decision-theoretic framework (Chatterjee and Eliashberg (1990); Rose and Joskow (1990); Sinha and Chandrashekaran (1992)).

The model by Chatterjee and Eliashberg (1990) considers the heterogeneity in initial perceptions of the innovation’s performance, consumers’ preference structure and the perceived reliability of information on which updating takes place. Their approach is an important step in modelling diffusion at the micro level, but its application is limited by its dependence on extensive perceptual data about adoption. Sinha and Chandrashekaran (1992) use a hazard model approach that explicitly incorporates covariates in the adoption time specification so that population is heterogeneous in timing of adoption. The considered split hazard model framework allows for modelling the adoption decision at the individual level as well as describing and forecasting new product acceptance at the aggregate market level.

Bemmaor (1994) suggests an alternative approach to explain the changes in the parameter estimates of the Bass model considering the underlying heterogeneity of the population. He considers that diffusion can equivalently be explained by the variation of individual propensities to buy across consumers. Therefore, a shifted Gompertz density can explain the timing of first purchase, and the individual propensity across consumers follows a gamma distribution. The aggregate diffusion process results in a mixture of these two densities. Bemmaor and Lee (2002) briefly analysed the consequence misspecification in the Bass model mentioned by Van den Bulte and Lilien (1997) and found the supremacy of the Bemmaor (1994) model with strong forecasting capacity. Their results also suggested that the Gamma-shifted Gompertz model is a flexible model to analyse the systematic changes in parameter estimates when specification error andill-conditioning occur.

The Gamma-shifted Gompertz model postulates that the ratio *q/p* of Bass model parameters varies with scale parameter *β* of the heterogeneity distribution of *η* and *η* has to be distributed according to a Gamma (*α, β*) law. Therefore, the individual level model (model of first adoption timing) can be identified with a shifted Gompertz density as follows:

$$F(t \mid \eta, b) = (1 - e^{-bt})e^{-\eta e^{-bt}}, \qquad t > 0, \quad \eta, b > 0$$

with the following density function:

$$f(t \mid \eta, b) = be^{(-bt - \eta e^{-bt})}[1 + \eta(1 - e^{-bt})], \qquad t > 0, \quad \eta, b > 0$$

For a fixed value of *b*, the small values *η* imply a low mean time of adoption (i.e., a strong individual propensity to buy).

If the heterogeneity parameter *η* varies according to a Gamma distribution with shape parameter *α* and scale parameter *β*, the aggregate-level diffusion model can be described by the following cumulative distribution function:

$$F(t) = (1 - e^{-bt}) / (1 + \beta e^{-bt})^{\alpha} \tag{5}$$

With the following density function:

$$f(t) = be^{-bt}(1 + \beta e^{-bt})^{-(\alpha+1)}[1 + \alpha\beta + \beta e^{-bt}(1 - \alpha)]$$

If we re-parameterise Equation (3) with the equivalent Bass model coefficient by letting *b*= *p*+*q* and *β*= *q/p*, we obtain the following aggregate-level diffusion model:

$$F(t) = 1 - e^{-(p+q)t} \Bigg/ \left(1 + \frac{q}{p}e^{-(p+q)t}\right)^{\alpha}, \quad t \geq 0 \tag{6}$$

For the value of shape parameter *α=1*, the model reduces to the standard Bass model. When *α=0*, the model reduces to an exponential model, and when *α*=∞, it converges to the shifted Gompertz model (Mahajan and Peterson (1985)). Therefore, as *α* approaches to zero, the shape of the diffusion curve resembles an exponential diffusion curve, and for larger values of *α*, it approaches a logistic curve. The role of *α* becomes the vital driver in the diffusion process that can be used to explain the impact of contagion or lack of contagion in the diffusion process. Thus, the Gamma-shifted Gompertz model comprehends several other models of diffusion, including the standard Bass model.

The Gamma-shifted Gompertz model by Bemmaor (1994) is an important contribution in diffusion modelling that provides grounds to investigate jointly “the speed takeoff” and “the diffusion speed after takeoff” observed in the process. As all other diffusion patterns are nested within the Bemmaor modelling approach, our idea is to extend it and make this model able to incorporate the related exogenous variables in the diffusion process considering the modifications with the GBM approach, incorporating the intervention function *x(t)*.

Considering an intervention to the Bemmaor model with one exponential shock that starts at time *a*, with intensity *c* and persistent effect *b*, a GBM Bemmaor model can be given by the following equation:

$$F(t)=\{\begin{matrix} 1-e^{-(p+q)t} \Big/ \left(1+\frac{q}{p}e^{-(p+q)t}\right)^{\alpha}; \quad t<a \\ \left(1-e^{-(p+q)(t+\frac{c}{b}e^{\{b(t-a)-1\}}}\right) \Big/ \left(1+\frac{q}{p}e^{-(p+q)(t+\frac{c}{b}e^{\{b(t-a)-1\}}}\right)^{\alpha}; \quad t\geq a \end{matrix} \tag{7}$$

It should be noted that the Bemmaor model has two main portions: the numerator that explains the influence of innovators and the denominator that explains the effect of imitators on the ultimate penetration. Therefore, it can be applied to model (6) with a simple modification after introducing a non-negative exponent to the innovators’ influence that gives the expression for the following model:

$$F(t)= 1-e^{-(p+q)t} \ {}^{\delta} \Big/ \left(1+\frac{q}{p}e^{-(p+q)t}\right)^{\alpha}, \qquad t\geq 0 \tag{8}$$

The new parameter *δ* will speed up/suppress the initial start to the curve and modify the curve peakedness. This intervention of the modified model is important to describe the quick/delayed entrance of the innovators, which could be mixed up with other contagion processes. With a fixed *α*, for *δ=1*, the modified model becomes the standard Bemmaor model, and *δ<1* will delay the innovators’ contagion process, whereas *δ>1* will speed up the diffusion at the very beginning. In other words, the parameter *δ* can be considered a measure of the propensity of the innovators to participate in the adoption process. For an ideal situation, *δ* should be greater than *α*, and for *δ=α=1*, the proposed modified model equals the standard Bass model.

The modified Bemmaor model can also be used with further modifications by adding some exponential or rectangular shocks. An equivalent expression as the model in Equation (7), the modified Bemmaor model with intervention function can be given by the following cumulative distribution function:

$$F(t)=\{\begin{matrix} 1-e^{-(p+q)t} \ {}^{\delta} \Big/ \left(1+\frac{q}{p}e^{-(p+q)t}\right)^{\alpha}; \quad t<a \\ \left(1-e^{-(p+q)(t+\frac{c}{b}e^{\{b(t-a)-1\}}}\right)^{\delta} \Big/ \left(1+\frac{q}{p}e^{-(p+q)(t+\frac{c}{b}e^{\{b(t-a)-1\}}}\right)^{\alpha}; \quad t\geq a \end{matrix} \tag{9}$$

These modified models in Equations (7-9) appearing for the first time, are very simple but important in terms of explaining the contagion effect with specific parameterisation of innovators' and imitators' penetration in the diffusion process. The analytical validation of the above postulates will be discussed in Section 5 with a real dataset.

4. MODEL PARAMETER ESTIMATES AND INFERENCE

The GBM in Equation (4) and the heterogeneity models in Equations (6-9) can be specified in a non-linear regressive framework as follows:

$$z(t) = f(\underline{\beta}, t) + \varepsilon(t) \tag{10}$$

where *z(t)* represents the cumulative observed data, $f(\underline{\beta}, t)$ is the deterministic component of the model specified through the cumulative mean function of *f(t)* of adoption over time, $\underline{\beta}$ is the vector of parameters and $\varepsilon(t)$ is a white noise process. The model parameters can be estimated using the NLS method following the Levenberg-Marquardt algorithm (Seber and Wild (1989)). At the second step, the estimated function $f(\hat{\underline{\beta}}, t)$ can be used in an ARMAX model in order to obtain a convenient expression of the residual structure in *ε(t)* that may be characterised by auto dependence effects very far from a standard white noise.

Following Guseo et al. (2007), the significance of the gain from the simpler model to the more complex model can be evaluated in two steps. As a first step, the squared multiple partial correlation coefficient is computed by the following:

$$\tilde{R}^2 = R^2_{M_1} - R^2_{M_2} \Big/ 1 - R^2_{M_2} \tag{11}$$

where $R^2_{M_2}$ denotes the determination index of the reduced model that has to be compared to model M_1. If *N* denotes the total number of observations used to fit the models, and *λ* is the number of parameters included in model M_1, the significance for the number of *κ* parameters of the model M_1 that are not included in model M_2 can be evaluated by a special form of F-statistics defined as follows:

$$F = \left[\tilde{R}^2 (N - \lambda) \right] \Big/ \left[1 - \tilde{R}^2 \ \ \kappa \right] \tag{12}$$

which is distributed as a Snedecor's-*F* with $\kappa, (N - \lambda)$ degrees of freedom under the assumption of equivalence of models M_1 and M_2 with *ε(t)* is normal i.i.d. Considering the common threshold 4 for the F-ratio in (12) as an approximate robust criterion to compare model M_2 nested in model M_1, the comparative performance can be evaluated (Guseo et al. (2007)).

5. AN APPLICATION TO THE ALGERIAN NATURAL GAS PRODUCTION DATA

Algeria has one of the biggest natural gas reserves in the world. Algeria is the owner of the eighth-largest natural gas reserve, having 159 trillion cubic feet (TCF) of proven natural gas, according to *Oil and Gas journal*. Results from the BP Statistical Review of World Energy 2010 indicate Algeria as the holder of 2.4% of the total world gas reserves. The reserve-to-production ratio is 55.3 years, but this type of index is often questioned, because it does not take into account the nonlinear extraction dynamics. The country is the third-largest exporter of natural gas to Europe. Algeria's natural gas sector has witnessed rapid expansion on the heels of increased production. Recent successes are aided by the international partnerships and technological advances, and the country is, at the same time, looking forward to solidifying its standing as a regional transit hub for natural gas, Global Arab Network reports according to Oxford Business Group (OBG). 'Sonatrach' dominates the country's natural gas production and wholesale distribution; however, foreign investments in the sector are continuously increasing. Foreign producers such as PCI, BP, Statoil, Total, BHP-Billiton, Eni and Repsol have entered into partnership agreements with 'Sonatrach' from the early 1970s.

The present study uses the Algerian natural gas production (in billion cubic meters, BCM) data obtained from British Petroleum (2011) for the period from 1970 to 2010. As seen in Figure-1, starting from the 1970's, the scenario of Algerian gas production follows an increasing trend until 2005 with some ups and downs at different sections but a slow decreasing trend afterwards. The increment trend is due to an increasing demand from the three top consumers (i.e. Italy, Spain and France), but the unexpected slow decrement trend after 2005 is a matter of discussion.

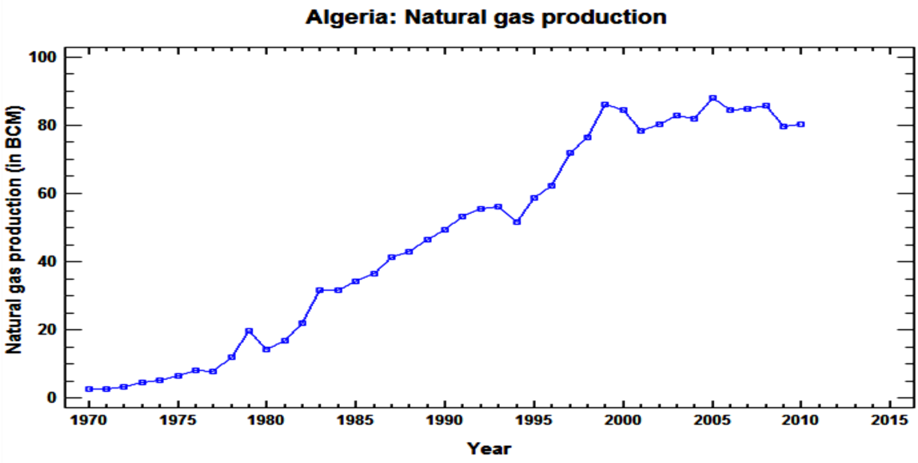


Figure-1: Natural Gas Production in Algeria (Instantaneous Data)
[Source: British Petroleum (2011)]

To describe the cumulative annual Algerian natural gas production, the study considers the model described in the previous section in Equations (2), (4), (6) and (7) with different specifications. Starting with the standard Bass (B1) model, it considers a GBM with an exponential shock for the intervention function *x(t)*. Afterwards, considering the heterogeneity, the standard Bemmaor model (BM) and Bemmaor model with GBM (GBMBM) are considered. A further generalisation as in Equations (8-9), modified Bemmaor model (BMM) and modified Bemmaor model with intervention function (GBMBMM), is also taken into consideration to identify the best fitted model and validate parameters. Obtained results are given in Table 1.

Table 1
Parameter Estimates and Asymptotic Standard Errors
(In Parentheses) for Different Models

Model parameters		**B1**	**BM**	**BMM**	**GBM**	**GBMBM**	**GBMBMM**
General penetration parameters	**m**	2687 (59.71)	4948 (437.989)	3029 (184.45)	2993 (64.129)	3057 (267.21)	2833 (189.288)
	p	0.0018 (0.00004)	0.0379 (0.00128)	0.0013* (0.00066)	0.0012 (0.00005)	0.00092* (0.0018)	0.00068* (0.00081)
	q	0.124 (0.002469	0.0096* (0.00528)	0.1155 (0.0151)	0.1118 (0.0025)	0.1099 (0.0269)	0.1298 (0.0241)
Exponential shock parameters	**a**	---	---	---	11.42 (0.4132)	9.44 (0.6602)	12.75 (0.5334)
	b	---	---	---	-0.264 (0.0483)	-0.178 (0.8083)	-0.255* (0.1472)
	c	---	---	---	1.158 (0.1485)	1.202 (0.4949)	0.6384 (0.1535)
Propensity parameters	**α**	---	22.17 (9.7556)	0.763 (0.1103)	---	0.971 (0.4865)	0.756 (0.2158)
	δ	---	---	3.233 (0.4525)	---	---	2.218 (0.4614)
R^2		0.999407	0.999877	0.999907	0.999924	0.999923	0.999946
Model SE		14.9322	6.81086	5.9028	5.33395	5.38814	4.50681

B1: Standard Bass model. BM: Bemmaor model.
BMM: Modified Bemmaor model.
GBM: Generalised Bass model with one exponential shock.
GBMBM: Bemmaor model with GBM.
GBMBMM: Modified Bemmaor model with intervention function.
*Parameter not significant.

As shown in Table- 1, the analysis results from the Algerian gas production dataset prove the efficacy of the newly introduced modification of the existing Bemmaor model with respect to the parameter estimates and respective fitness of the model. Compared to the standard Bass model (B1) or Bemmaor model (BM), the modified Bemmaor models (BMM, GBMBM and GBMBMM) reach better R^2 values. Parameter estimates for *m*, the carrying capacity, associated with the limiting behaviour of the cumulative production process and represent a current estimate of the Ultimate Recoverable Resources (URR). All the above models except BM suggest that the natural gas production crossed the middle of the life cycle and the maximum production level was already reached.

The standard Bass model (B1) predicts a moderate net reserve for net natural gas reserve and shows a very slow contribution to the innovators and comparatively large decrements of imitators' contribution to the process. The R^2 indicates the requirements for further modification of the fitted model. Since 2011, Algeria has produced 1921 BCM of natural gas, according to the Bass model, only 28% of the total reserve remains for the future. The GBM with one exponential shock shows a better fit to the data. It shows a mean-reverting positive shock around 1981/1982 when the Algerian state-dominated oil and gas company commissioned the Sonatrach Skikda LNG plant and refinery (GL-1K complex) and the government signed a 20-year agreement with France.

The Bemmaor model (BM) and the Bemmaor model with intervention function (GBMBM) improve the model fitness in terms of R^2 and estimated standard error. A large value of the additional heterogeneity parameter in BM indicates the existing heterogeneity in the annual gas production and therefore describes the possibility for explaining the observed process in a shifted Gompertz setup. The BM model indicates that 61% of total natural gas is still unused. The GBMBM, on the other hand, indicates the suitability of the usual GBM model with an estimate for the heterogeneity parameter of approximately 1 and an observed positive exponential shock in gas production around the 1980s that was absorbed in time.

The Modified Bemmaor model (BMM) and the modified Bemmaor model with intervention function (GBMBMM) proposed in this study are completely new in the literature and include one additional parameter for the innovators' heterogeneous propensity. Both models fit well with this additional parameter in terms of R^2 compared to all other considered models. The models have somewhat similar values for the imitators' propensity level. The large innovators' propensity coefficients for the BMM models indicate the existence of heterogeneity among initial productions, and a very accelerated trend with a late start is observed at the early stage of the diffusion process. The predicted reserve level is comparatively better for GBMBMM, indicating that the maximum level of production was already reached and 67.8% of the Algerian natural gas URR had been extracted by 2011. The process also shows a positive mean-reverting exponential shock around 1983, when Algeria signed another gas export agreement with Italy and the first BTUs of gas were delivered through the Transmed pipeline.

Finally, based upon the comparative parameter estimates, model standard errors and R^2 values help to select the best fitted model among the postulated models. In all respects, the GBMBMM model performs better for describing natural gas production. To obtain an improved short-term prediction for the regressive approach of the postulated models, an ARMAX model, based upon one regressor or more lagged regressors depending upon the predictive values of the first regressive step, was implemented. Obtained forecasts for the different models are given in Figures (2-7). Results obtained for model estimation and forecast performance are described in Table 2.

Figures (2–4) show the graphs for the observed Algerian natural gas production and the forecast with the standard Bass model (B1), Bemmaor model (BM) and modified Bass model (BMM) respectively with a convenient ARMAX sharpening for a better short-term prediction. The results show that the Algerian natural gas production already crossed the maximum production level in the year 2000 by B1 and BM predictions, whereas the modified Bemmaor model (BMM) indicated that maximum production was reached in 2006. The BM model shows a recovery trend and forecast another peak production level between 2012 and 2016. Therefore, the predicted life cycle becomes a little longer. Other models do not support this prediction, and a decrement trend is observed after maximum production in 2000 with some stationarity in the process.

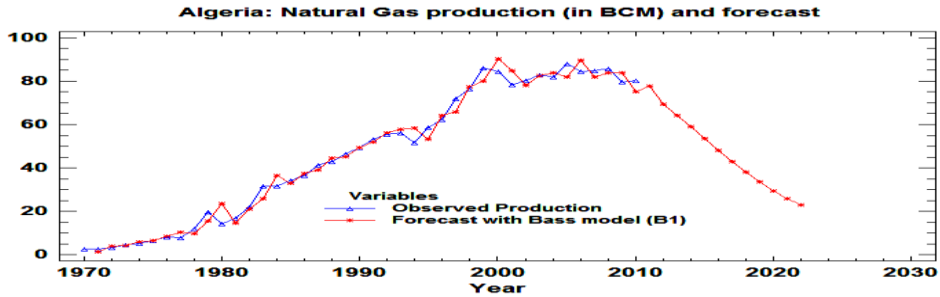


Figure- 2: Algerian Natural Gas: Forecast with Bass (B1) Model with ARIMA(2,0,2)

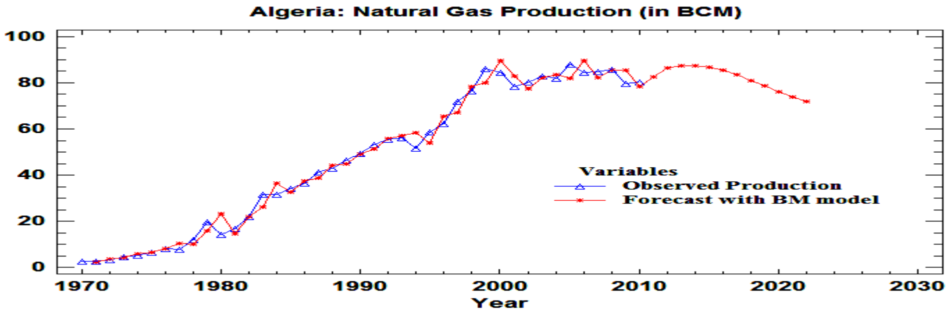


Figure-3: Algerian Natural Gas: Forecast with Bemmaor (BM) Model with ARIMA(2,0,2)

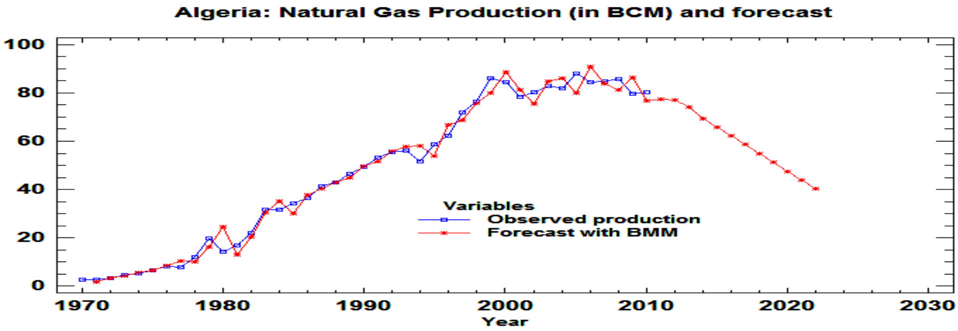


Figure-4: Algerian Natural Gas: Forecast with Modified Bemmaor (BMM) Model with ARIMA(1,0,4)

Model forecasts from the Algerian gas production data with the GBM with an intervention function of one exponential shock, described by GBM, GBMBM and GBMBMM are shown in Figures (5-7). Results show that the maximum production level had already been achieved in 2006, which is different from the year predicted by the standard Bass model (B1) or Bemmaor model (BM) due to the consideration of the existing exponential shock in the data and its consequence after the strategic and planning decisions taken by the respective authority. A rapidly decreasing production process was also predicted by GBMBMM followed by the respective prediction with GBM and GBMBM set up with ARMAX corrections for short-term prediction.

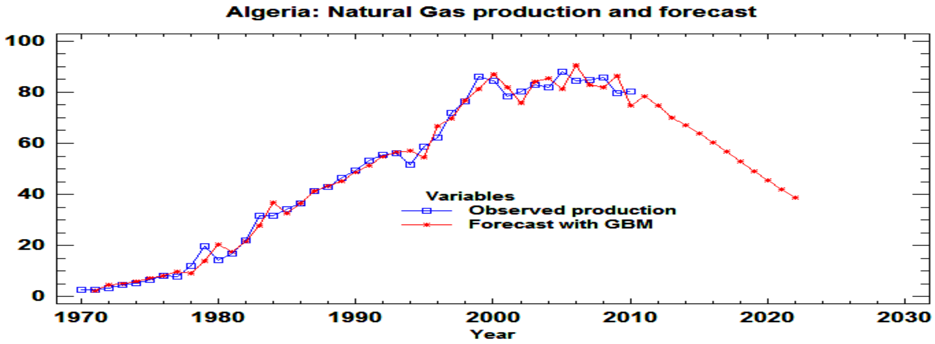


Figure-5: Algerian Natural Gas: Forecast with Generalised Bass (GBM) Model with ARIMA(2,0,3)

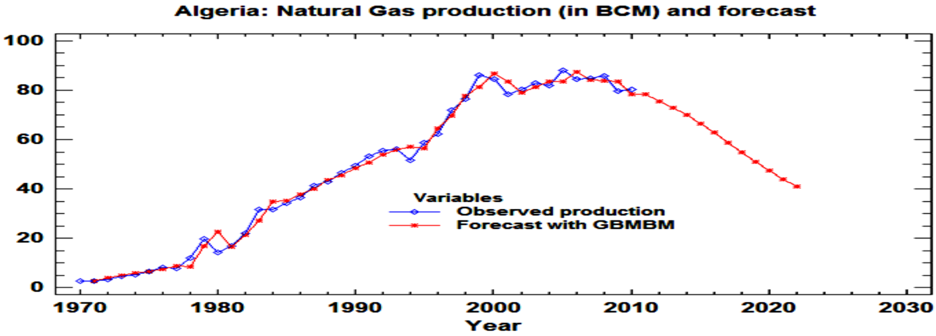


Figure-6: Algerian Natural Gas: Forecast with Generalised Bass & Bemmaor (GBMBM) Model with ARIMA(2,0,1)

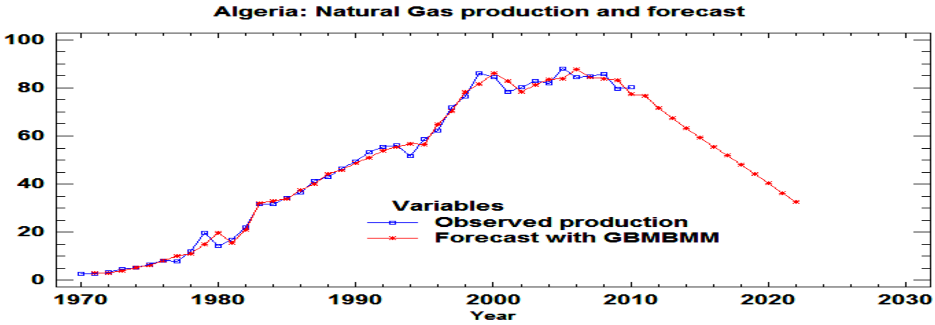


Figure-7: Algerian Natural Gas: Forecast with Generalised Bass & Modified Bemmaor (GBMBM) Model with ARIMA(2,0,1)

Results from Table 2 indicate that both the root-mean squared error (RMSE) and mean-absolute prediction error (MAPE) attain minimum values for the GBMBMM model after the forecast with ARIMA (2,1) set-up when regressed with the predicted estimates. When compared with the standard Bass model (B1), the significance for the inclusion of additional parameter/s passes the F-test for all other postulated models.

Similar results are also found for the parameter/s when compared with the Bemmaor model (BM). For the GBMBMM model, the squared partial correlation co-efficient $\tilde{R}^2$ = 0.9089 (F=65.85), $\tilde{R}^2$=0.5610 (F=14.483) and $\tilde{R}^2$=0.2895 (F=6.723) when compared with B1, BM and GBM, respectively. Therefore, strong evidence for the significance of an additional parameter in GBMBMM and its forecast capacity is established.

Table 2
Model Performance for Estimation and Forecast

Model	RMSE	MAPE	No. of parameters	R^2	$\tilde{R}^2$ w.r. to B1(F)	$\tilde{R}^2$ w.r. to BM (F)	$\tilde{R}^2$ w.r. to GBM (F)
B1+ARIMA (2,0,2)	2.9690	2.1389	3	0.999407	NA	NA	NA
BM+ARIMA (2,0,2)	2.63093	2.10323	4	0.999877	0.7926 (141.40)	NA	NA
BMM+ARIMA (1,0,4)	2.78532	2.77534	5	0.999907	0.8432 (96.80)	0.2439 (11.612)	NA
GBM+ARIMA (2,0,3)	2.37584	3.29014	6	0.999924	0.8718 (79.34)	0.3821 (10.822)	NA
GBMBM+ARIMA (2,0,1)	2.24863	2.1383	7	0.999923	0.8702 (56.99)	0.3740 (6.771)	NA
GBMBMM+ARIMA (2,0,1)	1.72866	1.77848	8	0.999946	0.9089 (65.85)	0.5610 (14.483)	0.2895 (6.723)

6. CONCLUSIONS

Modelling a complex system is always a difficult task. Diffusion of innovation modelling in this context is facing new challenges for incorporating the interventions of new ideas, technologies and other influencing variables with a parsimonious model that helps to explain and modify the evolutionary shape of the curve with respect to time. The main aim of this paper was to compare some existing diffusion models. Considering the Bemmaor model and then extending it with modifications of the existing models considering heterogeneity levels. Because the individual propensity should not be the same for the innovator and imitator groups, it is important to consider different parameterisations to identify and validate the heterogeneity level. Therefore, the complete life cycle of the process can be studied more efficiently with minimum prediction errors.

The application of the proposed modified model in parallel to other existing models of innovation diffusion gives a fruitful comparison of the efficacy of the proposed parameter and its estimates. Results obtained from the Algerian gas production data perfectly match those of recent studies, which support the decrement trends identified by the model forecasts. Overall production of natural gas decreased 3% in 2011 as compared to the previous year, as British Petroleum (2012) reports.

For the first time, this study used the concept of the existence of heterogeneity among the early adopters/innovators in the diffusion process that could be further considered for multiple products' diffusion models or diffusion models with seasonally varying time series.

REFERENCES

1. Bass, F.M. (1969). A new product growth model for consumer durables. *Management Science*, 15, 215-227.
2. Bass, F.M. (2004). Comments on a new product growth model for consumer durables: The Bass Model. *Management Science*, 50, 1833-1840.
3. Bass, F.M., Krishnan, T.V. and Jain, D.C. (1994). Why the Bass model fits without decision variables. *Marketing Science*, 13, 203-223.
4. Bemmaor, A.C. (1994). Modeling the diffusion of new durable goods: Word-of mouth effect versus consumer heterogeneity. *Research Traditions and Marketing*, 201-223, Boston, MA: Kluwer Academic.
5. Bemmaor, A.C. and Lee, J. (2002). The impact of heterogeneity and ill-conditioning on diffusion model parameter estimates. *Marketing Science*, 21, 209-220.
6. British Petroleum (2011). BP Statistical Review of World Energy (www.bp.com), London.
7. British Petroleum (2012). BP Statistical Review of World Energy (www.bp.com), London.
8. Chatterjee, R. and Eliashberg, J. (1990). The innovation diffusion process in a heterogeneous population: A micro-modeling approach. *Management Science*, 36(9), 1057-1079.
9. Dalla Valle, A. and Furlan, C. (2011). Forecasting accuracy of wind power technology diffusion models across countries. *International Journal of Forecasting,* 27(2), 592-601.
10. Guseo, R. (2004). Strategic interventions and competitive aspects in innovation life cycle. Tech. Rept. *Working paper Series* N. 11/2004.
11. Guseo, R. and Dalla Valle, A. (2005). Oil and gas depletion: Diffusion models and forecasting under strategic intervention, *Statistical Methods and Applications*, 148(3), 375-387.
12. Guseo, R., Dalla Valle, A. and Guidolin, M. (2007). World oil depletion models: Price effects compared with strategic or technological interventions. *Technological Forecasting and Social Change*, 74(4), 452-469.
13. Guseo R. and Guidolin, M (2009). Modelling a dynamic market potential: A Class of automata networks for diffusion of innovations. *Technological Forecasting and Social Change*, 76(6), 806-820.
14. Guseo, R. and Guidolin, M. (2010). Cellular automata with network incubation in information technology diffusion. *Physica A: Statistical Mechanics and its Applications*, 389(12), 2422-2433.
15. Guseo, R. and Guidolin, M. (2011). Market potential dynamics in innovation diffusion: Modelling the synergy between two driving forces. *Technological Forecasting and Social Change*, 78(1), 13-24.
16. Guiterrez-Jaimez, R., Roman, P., Romero, D., Serrano, J.J. and Torres, F. (2007). A new Gompertz-type diffusion process with application to random growth. *Mathematical Biosciences,* 208(1), 147–165.
17. Mahajan, V. and Wind, Y. (1986). *Innovation diffusion models of new product acceptance*. Series of Econometrics and Management Sciences, Ballinger, Cambridge, Massachusetts.
18. Mahajan, V., Muller, E. and Bass, F.M. (1990). New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing,* 54, 1-26.

19. Mahajan, V. and Peterson, R.A. (1985). *Models for Innovation Diffusion*, Sage, London.
20. Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica: Journal of Econometric Society,* 741-766.
21. Meade, N. and Islam, T. (2006). Modeling and forecasting the diffusion of innovation- A 25-year review. *International Journal of Forecasting,* 22(3), 519-545.
22. Robertson, A., Soopramanien, D. and Fildes, R. (2007). Segmental new-product diffusion of residential broadband services. *Telecommunication Policy,* 31(5), 265-275.
23. Rogers, E.M. (2003). *Diffusion of Innovations*. 5th Edition, Free Press, New York.
24. Rogers, E.M. and Shoemaker, F.F. (1971). *Communication of innovations: A Cross-Cultural Approach*, Free Press, New York.
25. Rose, N.L. and Joskov, P.L. (1990). The diffusion of new technologies: Evidence from the electric utility industry, *The RAND Journal of Economics*, 31(3), 354-374.
26. Sinha, R.K. and Chandrashekharan, M. (1992). A split hazard model for analyzing the diffusion of innovations. *Journal of Marketing Research,* 29(1), 116-127.
27. Srinivasan, V. and Mason, C.H. (1996). Technical note- Non-linear least squares estimation of new product diffusion models. *Marketing Science*, 5(2), 169-178.
28. Seber, G. and Wild, C. (1989). *Nonlinear Regression*, Wiley, New York.
29. Van Den Bulte, C. and Lilien, G.L. (1997). Bias and systematic change in the parameter estimates of macro-level diffusion models. *Marketing Science,* 338-353.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 345-358

ON TWO-STAGE LAO MULTI-HYPOTHESES TESTING FOR MANY DISTINCT FAMILIES OF PROBABILITY DISTRIBUTIONS

Farshin Hormozi nejad[1] and **Evgueni Haroutunian**[2]
[1] Islamic Azad University, Ahvaz Branch, Iran
Email: hormozi-nejad@iauahvaz.ac.ir
[2] Institute for Informatics and Automation Problems,
National Academy of Sciences, Republic of Armenia
Email: evhar@ipia.sci.am

ABSTRACT

Problem of multihypotheses two-stage testing for a model which includes many separated families of hypothetical probability distributions is considered. The studied object follows one of S hypothetical probability distributions. The problem of multiple hypotheses testing in a pair of stages is introduced such that in the first stage one family of distributions must be distinguished and then in the second stage, the object's distribution must be denoted between mentioned family. The reliabilities matrix of logarithmically asymptotically optimal (LAO) hypothesis testing by a pair of stages is considered and compared with the case of similar one-stage testing.

KEYWORDS

Logarithmically asymptotically optimal test, multihypotheses testing, two-stage test, reliabilities matrix.

1. INTRODUCTION

Hoeffding [12], Tusnady [15], Csiszár and Longo [6] and Blahut [3] studied asymptotically optimal tests. Ahlswede and Haroutunian [1] and in paper [8] Haroutunian formulated new problems on multiple hypotheses testing and identification. In paper [7] Haroutunian the solved problem of logarithmically asymptotically optimal (LAO) testing of multiple statistical hypotheses. In [9] Haroutunian et al investigated reliability criteria in information theory and in statistical hypothesis testing and Haroutunian and Hakobyan in [10] considered the problem of multiple hypotheses LAO testing for many independent objects. The two-stage LAO testing of multiple hypotheses for a pair of families of distributions is investigated in [11, 13, 14]. In this paper the two-stage LAO multihypotheses testing for a model consisting of many disjoint families of probability distributions (PDs) is studied.

Section 2 contains some notations and definitions and the problem statement. Section 3 introduces the multiple hypothesis testing and expose the one-stage LAO test from [7]. Section 4 shows the two-stage test by a sample. Section 5 discusses the procedure of making decision in two stages on the base of a pair of samples. Section 6 compares matrices of reliabilities of the one-stage and the two-stage LAO hypotheses testing.

2. PRELIMINARIES

Random variable (RV) X characterizing an object takes values in the finite set $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$ is the space of all distributions on $\mathcal{X}$. S possible probability distributions(PDs) $P_s,\ \ s=\overline{1,S}$ of $\mathcal{X}$ are given, they are grouped in K disjoint families of PDs. The first family includes R_1 hypotheses, the second family includes R_2 hypotheses and etc., the last family includes R_K hypotheses such that $\sum_{k=1}^{K} R_k = S$.

Let sample $\mathrm{x}=(x_1,x_2,\dots,x_N)$ be a vector of results of N independent observations of the RV X. Let $N(x|\mathrm{x})$ be the number of repetitions of the element $x\in\mathcal{X}$ in the vector $\mathrm{x}\in\mathcal{X}^N$, and

$$Q_{\mathrm{x}}(x) \overset{\Delta}{=} \frac{N(x|\mathrm{x})}{N}\ ,\quad x\in\mathcal{X},$$

be the PD, called in statistics *the empirical probability distribution* of the sample x, but we prefer shorter term from information theory, *the type* of x [4, 5].

Let $\mathcal{P}^N(\mathcal{X})$ be the set of all possible types on $\mathcal{X}^N$ for N-sample and let T_Q^N be the set of all vectors x of the type $Q\in\mathcal{P}^N(\mathcal{X})$. The Shannon entropy of RV X with PD Q and the divergence (Kullback-Leibler distance) of PDs P and Q, are defined [4, 5, 7] as follows:

$$\mathrm{H}_Q(X) \overset{\Delta}{=} -\sum_{x\in\mathcal{X}} Q(x)\log Q(x),$$

$$\mathrm{D}(Q\parallel P) \overset{\Delta}{=} \sum_{x\in\mathcal{X}} Q(x)\log\frac{Q(x)}{P(x)}.$$

We need the following useful properties of types [4, 5].

$$|\mathcal{P}^N(\mathcal{X})| \le (N+1)^{|\mathcal{X}|},$$

$$(N+1)^{-|\mathcal{X}|}.\exp\{N\mathrm{H}_Q(X)\} \le |T_Q^N| \le \exp\{N\mathrm{H}_Q(X)\},$$

$$\mathrm{P}^N(\mathrm{x}) = \exp\{-N(\mathrm{H}_Q(X)+\mathrm{D}(Q||P))\},\quad \mathrm{f}or\quad \mathrm{x}\in T_Q^N.$$

3. ONE-STAGE LAO TEST FOR MULTIHYPOTHESES TESTING

We call the procedure of making decision on the base of N-sample the test ϕ^N. For detecting actual PD between S PDs, the test ϕ^N, can be defined by division of the sample space $\mathcal{X}^N$ to S disjoint subsets $G_s^N,\ \ s=\overline{1,S}$. The set G_s^N consists of all samples x for which s-th PD is adopted,

$$G_s^N \overset{\Delta}{=} \{\mathrm{x}: \phi^N(\mathrm{x})=s\},\quad s=\overline{1,S},$$

Let $\alpha_{l|s}$ be the probability of the erroneous selection of PD P_l provided that P_s is true,

$$\alpha_{l|s}(\phi^N) \overset{\Delta}{=} P_s^N(G_l^N),\quad l,s=\overline{1,S},\quad l\ne s.$$

The probability to reject P_s, when it is true, is

$$\alpha_{s|s}(\phi^N) \overset{\Delta}{=} P_s^N\left(\overline{G}_s^N\right) = \sum_{l\neq s} \alpha_{l|s}(\phi^N), \quad l,s = \overline{1,S}.$$

We denote by ϕ the infinite sequences of tests, corresponding reliabilities (error probability exponents) are

$$E_{l|s}(\phi) \overset{\Delta}{=} \limsup_{N\to\infty}\{-\frac{1}{N}\log\alpha_{l|s}(\phi^N)\},\ l,s = \overline{1,S}.$$

The matrix $E(\phi) = \{E_{l|s},\ l,s = \overline{1,S}\}$ is called the *reliabilities matrix*.

The test ϕ^* is called LAO if for given by consumer positive values of corresponding $S-1$ diagonal elements of the reliabilities matrix $E(\phi^*)$, the procedure provides maximal values for other elements of it. The following theorem contains the solution of problem of LAO test ϕ^* construction and conditions of existence of the test of elements of matrix $E(\phi^*)$ of which are positive.

Theorem 1 [8, 9]. Consider an object with S hypotheses $P_s,\ s = \overline{1,S}$. For given positive numbers $E_{1|1}, E_{2|2}, \ldots, E_{S-1|S-1}$ let us introduce the regions:

$$R_s = \{Q\colon D(Q \parallel P_s) \leq E_{s|s}\}, \quad s = \overline{1,S-1},$$
$$R_S = \{Q\colon \mathrm{D}(Q \parallel P_s) > E_{s|s}, \quad s = \overline{1,S-1}\},$$

and the following values for elements of the future reliabilities matrix $E(\phi^*)$ of the LAO test sequence ϕ^*:

$$E_{s|s}^* = E_{s|s}, \quad s = \overline{1,S-1},$$
$$E_{l|s}^* = \inf_{Q\in R_l} \mathrm{D}(Q \parallel P_s), \quad l,s = \overline{1,S},\ l\neq s,$$
$$E_{S|S}^* = \min_{l\neq S} E_{l|S}^*.$$

If the following compatibility conditions take place

$$0 < E_{1|1} < \min_{s=\overline{2,S}} \mathrm{D}(P_s \parallel P_1),$$

$$0 < E_{s|s} < min[\min_{l=\overline{1,s-1}} E_{l|s}^*, \min_{l=\overline{s+1,S}} \mathrm{D}(P_l \parallel P_s)], \quad 2 \leq s \leq S-1,$$

then there exists a LAO sequence of tests ϕ^* with reliabilities matrix

$$E(\phi^*) = \{E_{l|s}^*,\ l,s = \overline{1,S}\}.$$

Even if one of the compatibility conditions is violated, then the reliabilities matrix of such test has at least one element equal to zero.

4. THE TWO-STAGE LAO TEST BY ONE SAMPLE

We denote the two-stage test on the base of N observations by Φ_1^N such that may be composed by a pair of tests φ_1^N and φ_2^N for two consecutive stages and we write $\Phi_1^N = (\varphi_1^N, \varphi_2^N)$. The first stage is for choice of a family of PDs, it is executed by a non-randomized test $\varphi_1^N(\mathrm{x})$ using the sample x. The next stage is for making decision in the determined family of PDs, it is shown by a non-randomized test $\varphi_2^N(\mathrm{x})$ based on the sample x and the result of the test φ_1^N.

4.1. First Stage of the Two-Stage Test by One Sample

Consider for convenience the cumulative numbers $M_k = \sum_{i=1}^k R_i$ and the sets of indexes

$$D_1 = \{\overline{1, M_1}\}, D_2 = \{\overline{M_1+1, M_2}\}, \dots, D_k = \{\overline{M_{k-1}+1, M_k}\}, \dots, D_K = \{\overline{M_{K-1}+1, S}\}.$$

Therefore suppose there are K disjoint families of PDs P_1, P_2,..., P_K such that

$$P_k = \{P_s, \quad s \in D_k\}, \quad k = \overline{1, K}.$$

The first stage of decision making consists in using sample x for selection of one family of PDs by a test $\varphi_1^N(\mathrm{x})$, which can be defined by division of the sample space X^N on K disjoint subsets

$$A_k^N \overset{\Delta}{=} \{\mathrm{x}_1 : \varphi_1^N(\mathrm{x}) = k\}, \quad k = \overline{1, K}.$$

The set A_k^N consists of all vectors x for which k-th family of PDs is adopted.

Let $\alpha'_{m|k}(\varphi_1^N)$ be the probability of the erroneous acceptance of m-th family of PDs provided that k-th family of PDs contains the correct PD:

$$\alpha'_{m|k}(\varphi_1^N) \overset{\Delta}{=} \max_{s: s \in D_k} P_s^N(A_m^N), \ m \neq k, \quad m, k = \overline{1, K}. \tag{1}$$

The probability to reject k-th family of PDs, when it is true, is

$$\alpha'_{k|k}(\varphi_1^N) \overset{\Delta}{=} \max_{s: s \in D_k} P_s^N(\overline{A}_k^N) = \sum_{m \neq k} \alpha'_{m|k}(\varphi_1^N), \quad k = \overline{1, K}. \tag{2}$$

We have to consider reliabilities of the sequence of tests φ_1

$$E'_{m|k}(\varphi_1) \overset{\Delta}{=} \limsup_{N \to \infty} \{-\frac{1}{N} \log \alpha'_{m|k}(\varphi_1^N)\}, \quad m, k = \overline{1, K}. \tag{3}$$

The *reliabilities matrix* for the first stage of the test is $\mathbf{E}'(\varphi_1) = \{E'_{m|k}, \ m, k = \overline{1, K}\}$ and it follows from (1), (2) and (3) that

$$E'_{k|k}(\varphi_1) = \min_{m \neq k} E'_{m|k}(\varphi_1).$$

The test φ_1^* is called LAO if for given by consumer positive values of corresponding $K-1$ diagonal elements of the reliabilities matrix $E'(\varphi_1^*)$, the procedure provides maximal values for other elements of it.

Theorem 2. Consider S different hypotheses P_s, $s = \overline{1, S}$ that take place in K disjoint family of PDs. For given positive numbers $E'_{1|1}, E'_{2|2}, \dots, E'_{K-1|K-1}$ let us introduce the regions:

$$R'_k = \{Q : \min_{s \in D_k} \mathrm{D}(Q \parallel P_s) \leq E'_{k|k}\}, \quad k = \overline{1, K-1},$$

$$R'_K = \{Q : \min_{s \in D_k} \mathrm{D}(Q \parallel P_s) > E'_{k|k}, \quad k = \overline{1, K-1}\},$$

and the following values for elements of the future reliabilities matrix $E'(\varphi_1^*)$ of the LAO test sequence φ_1^*:

$$E'^*_{k|k} = E'_{k|k}, \quad k = \overline{1, K-1},$$

$$E'^*_{m|k} = \min_{s \in D_k} \inf_{Q \in R'_m} \mathrm{D}(Q \parallel P_s), \quad m, k = \overline{1, K}, \ m \neq k,$$

$$E'^*_{K|K} = \min_{m \neq K} E'^*_{m|K}.$$

If the following compatibility conditions take place

$$0 < E'_{1|1} < \min_{s \in D_1, l \in D_k, k = \overline{2, K}} \mathrm{D}(P_l \parallel P_s),$$

$$0 < E'_{k|k} < min[\min_{m = \overline{1, k-1}} E'^*_{m|k}, \min_{m = \overline{k+1, K}, l \in D_m, s \in D_k} \mathrm{D}(P_l \parallel P_s)], \quad 2 \leq k \leq K-1,$$

then there exists a LAO sequence of tests φ_1^* with reliabilities matrix $E'(\varphi_1^*)$.

Even if one of the compatibility conditions is violated, then the reliabilities matrix of such test has at least one element equal to zero.

We omit the proof of Theorem 2 because it is analogous to Theorem 2 in [13].

Corollary 1. If $K = 2$, $S = 2$ we have hypotheses P_1 *and* P_2 *and* every family of PDs have only one PD and Theorem 2 is equivalent to Hoeffding's Theorem [12], where for $E'^*_{1|1} < D(P_2 || P_1)$,

$$E'^*_{2|2}(E'^*_{1|1}) = \inf_{Q: \mathrm{D}(Q || P_1) \leq E'^*_{1|1}} \mathrm{D}(Q || P_2).$$

4.2. Second Stage of the Two-Stage Test by One Sample

Since a family of PDs is selected, it is necessary to detect one PD in this family. So we consider test $\varphi_2^N(\mathrm{x}, \varphi_1)$ which can be defined by division of the sample space $\mathcal{X}^N$ to R_k disjoint subsets B_s^N, $s \in D_k$, if the k-th family of PDs is accepted, then

$$B_s^N \overset{\Delta}{=} \{\mathrm{x}: \varphi_2^N(\mathrm{x}, \varphi_1) = s\}, \quad s \in D_k, \ k = \overline{1, K},$$

Let $\alpha"_{l|s}(\varphi_2^N)$ be the probability of the erroneous acceptance at the second stage of test, in which PD P_l is accepted when P_s is true,

$$\alpha"_{l|s}(\varphi_2^N) \overset{\Delta}{=} P_s^N(B_l^N), \ l \neq s, \ l \in D_k.$$

The probability to reject P_s, when it is true, is

$$\alpha"_{s|s}(\varphi_2^N) \overset{\Delta}{=} P_s^N\left(\overline{B}_s^N\right) = \sum_{l \neq s} \alpha"_{l|s}(\varphi_2^N) + P_s^N(\overline{A}_k^N), \quad l, s \in D_k. \qquad (4)$$

Corresponding reliabilities for the second stage of test, are defined as

$$E"_{l|s}(\varphi_2) \overset{\Delta}{=} \limsup_{N \to \infty} \{-\frac{1}{N} \log \alpha"_{l|s}(\varphi_2^N)\}, \quad l, s = \overline{1, S}. \qquad (5)$$

Using lemma of types we introduce the following equalities and notations

$$\lim_{N\to\infty}\{-\frac{1}{N}\log P_s^N(\overline{A}_k^N)\} = \inf_{Q:\min\limits_{l\in D_k} D(Q||P_l)>E'^*_{k|k}} D(Q||P_s) \overset{\Delta}{=} E^I_{k|s}. \qquad (6)$$

From (6) we can see when $s \notin D_k$ then $E^I_{k|s} = 0$ and it follows from (4) and (5) that

$$E"_{s|s}(\varphi_2) = \min[\min_{l\neq s} E"_{l|s}(\varphi_2), E^I_{k|s}], \quad s \in D_k.$$

If at the first stage, the k-th family of PDs is accepted correctly, the part of reliabilities matrix for the second stage of test is $\mathbf{E}"^{(k|k)}(\varphi_2) = \{E"_{l|s}, \quad l, s \in D_k\}$ and if at the first stage, the m-th family of PDs is accepted erroneously, but the k-th family of PDs is correct, the part of reliabilities matrix is $\mathbf{E}"^{(m|k)}(\varphi_2) = \{E"_{l|s} \quad l \in D_m, \ s \in D_k\}$.

Theorem 3 (See [8, 9]). If at the first stage of test the k-*th* family of PDs is accepted, then for given $R_k - 1$ positive values $E"_{M_{k-1}+s|M_{k-1}+s}, \quad s = \overline{1, R_k - 1}$ of the reliabilities matrix $\boldsymbol{E}"^{(k|k)}(\varphi_2)$, let us consider the regions:

$$R"_s = \{Q: \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big) \leq E''_{M_{k-1}+s|M_{k-1}+s}\}, \quad s = \overline{1, R_k - 1},$$

$$R''_{R_k} = \{Q: \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big) > E''_{M_{k-1}+s|M_{k-1}+s}, s = \overline{1, R_k - 1}, \ \min_{s\in D_k} D(Q||P_s) \leq E'^*_{k|k}\},$$

and the following values of elements of the future reliabilities matrix $E''^{(k|k)}(\varphi_2^*)$ of the LAO test sequence:

$$E"^*_{M_{k-1}+s|M_{k-1}+s} = E"_{M_{k-1}+s|M_{k-1}+s}, \quad s = \overline{1, R_k - 1},$$

$$E"^*_{M_{k-1}+l|M_{k-1}+s} = \inf_{Q\in R"_l} \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big), \quad l, s = \overline{1, R_k}, \ l \neq s,$$

$$E"^*_{M_k|M_k} = \min[\min_{l\neq R} E"^*_{l|M_k}, E^I_{k|M_k}].$$

When the following compatibility conditions are valid

$$E"_{1|1} < min[\underline{\min}_{s=2,R_k} \mathrm{D}(P_{M_{k-1}+s} \parallel P_{M_{k-1}+1}), \ E^I_{k|1}],$$

$$E"_{s|s} < min\left[\underline{\min}_{l=1,s-1} E"^*_{M_{k-1}+l|M_{k-1}+s}, \underline{\min}_{l=s+1,R_k} \mathrm{D}\big(P_{M_{k-1}+l} \parallel P_{M_{k-1}+s}\big), \ E^I_{k|s}\right],$$
$$2 \leq s \leq R_k - 1,$$

then there exists a LAO sequence of tests φ_2^*, elements of reliabilities matrix $E''^{(k|k)}(\varphi_2^*)$ of which are defined above and are positive.

Even if one of the compatibility conditions is violated, then the reliabilities matrix of such test has at least one element equal to zero.

4.3. Reliabilities of the Two-Stage Test by One Sample

The test on the base of N-sample denoted $\Phi_1^{*N} = (\varphi_1^{*N}, \varphi_2^{*N})$, which is formed by a pair of LAO tests φ_1^{*N} and φ_2^{*N}. If in the first stage of LAO test the k-th family of PDs is accepted, then in the two-stage decision making, the test Φ_1^{*N} can be assigned by division of the sample space $\mathcal{X}^N$ to disjoint subsets as follows

$$C_s^N \overset{\Delta}{=} A_k^{*N} \cap B_s^N, \quad s \in D_k.$$

Let $\alpha'''_{l|s}$ be the probability of the erroneous acceptance by two-stage test by one sample of PD P_l when P_s is true:

$$\alpha'''_{l|s}(\Phi_1^{*N}) \overset{\Delta}{=} P_s^N(C_l^N), \quad l,s = \overline{1,S}, \quad l \neq s.$$

And the probability to reject P_s in two-stage test by one sample, when it is correct, is

$$\alpha'''_{s|s}(\Phi_1^{*N}) \overset{\Delta}{=} P_s^N\left(\overline{C}_s^N\right) = \sum_{l\neq s} \alpha'''_{l|s}(\Phi_1^{*N}), \quad s = \overline{1,S}.$$

We denote by $\Phi_1^* = (\varphi_1^*, \varphi_2^*)$ the infinite sequences of tests and define reliabilities:

$$E'''_{l|s}(\Phi_1^*) \overset{\Delta}{=} \limsup_{N\to\infty}\{-\frac{1}{N}\log\alpha'''_{l|s}(\Phi_1^{*N})\}, \quad l,s = \overline{1,S}.$$

hese are the relationship between error probabilities and reliabilities for the two-stage test by one sample and the first and the second stages of LAO tests:

a) if $l \in D_k,\ s = \overline{1,S}$ then

$$\alpha'''_{l|s}(\Phi_1^{*N}) = P_s^N(A_k^{*N} \cap B_l^N) = P_s^N(B_l^N) = \alpha''_{l|s}(\varphi_2^{*N}) \tag{7}$$

b) if $s \in D_k$ then

$$\begin{aligned}\alpha'''_{s|s}(\Phi_1^{*N}) &= P_s^N\left(\overline{C}_s^N\right) = P_s^N(A_k^{*N} \cap \overline{B}_s^N) + P_s^N(\overline{A}_k^{*N}) \\ &= \sum_{l\neq s, l\in D_k} \alpha''_{l|s}(\varphi_2^{*N}) + E_{k|s}^I = \alpha''_{s|s}(\varphi_2^{*N})\end{aligned} \tag{8}$$

According to (7)–(8) and definition of reliabilities we get

Theorem 4. If all distributions P_s, $s = \overline{1,S}$, are different and positive values $E_{k|k}'^{*}$, $k = \overline{1,K-1}$ and $E"_{M_{k-1}+r|M_{k-1}+r}$, $r = \overline{1,R_k-1}$, satisfy compatibility conditions of Theorems 2 and 3, then elements of matrix of reliabilities $E'''(\Phi_1^*)$ of the two-stage test by one sample Φ_1^* are

$$E'''_{l|s}(\Phi_1^*) = E"_{l|s}(\varphi_2^*), \quad l,s = \overline{1,S}.$$

When one of the compatibility conditions is violated, then the reliabilities matrix of such test has at least one element equal to zero.

5. THE TWO-STAGE LAO TEST BY A PAIR OF SAMPLES

We will discuss another version of the two-stage testing. Suppose $N = N_1 + N_2$ be such that:

$$N_1 = \lceil \psi N \rceil, \quad N_2 = \lceil (1-\psi)N \rceil, \quad 0 < \psi < 1,$$

$$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2),\ \mathrm{x} \in \mathcal{X}^N, \quad \mathrm{x}_1 \in \mathcal{X}^{N_1}, \quad \mathrm{x}_2 \in \mathcal{X}^{N_2}.$$

The two-stage test by a pair of samples on the base of N-sample we denote by $\Phi_2^N = (\varphi_1^{N_1}, \varphi_2^{N_2})$. The first stage is a non-randomized test $\varphi_1^{N_1}(\mathrm{x}_1)$ based on the sample x_1. The next stage is a non-randomized test $\varphi_2^{N_2}(\mathrm{x}_2, \mathrm{x}_1)$ based on the sample x_2 and the outcome of the test $\varphi_1^{N_1}(\mathrm{x}_1)$.

5.1. First Stage of Two-Stage Test by A Pair of Samples

The first stage of decision making for selection of a family of PDs denoted by a test $\varphi_1^{N_1}(\mathrm{x}_1)$, can be defined by division of the sample space $\mathcal{X}^{N_1}$ on K disjoint subsets

$$A_k^{N_1} \overset{\Delta}{=} \{\mathrm{x}_1 : \varphi_1^{N_1}(\mathrm{x}_1) = k\}, \quad k = \overline{1,K},$$

The set $A_k^{N_1}$ consists of all vectors x_1 for which k-th family of PDs is adopted.

Let $\alpha'_{m|k}(\varphi_1^{N_1})$ be the probability of the erroneous acceptance of m-th family of PDs provided that k-th family of PDs is true that consists the correct PD:

$$\alpha'_{m|k}(\varphi_1^{N_1}) \overset{\Delta}{=} \max_{s\in D_k} P_s^{N_1}(A_m^{N_1}),\ m \neq k, \quad m,k = \overline{1,K}. \tag{9}$$

The probability to reject k-th family of PDs, when it is true, is

$$\alpha'_{k|k}(\varphi_1^{N_1}) \overset{\Delta}{=} \max_{s\in D_k} P_s^{N_1}(\overline{A}_k^{N_1}) = \textstyle\sum_{m\neq k} \alpha'_{m|k}(\varphi_1^{N_1}), \quad k = \overline{1,K}. \tag{10}$$

We have to consider reliabilities of the sequence of tests φ_1:

$$E'_{m|k}(\varphi_1) \overset{\Delta}{=} \limsup_{N_1\to\infty} \left\{-\frac{1}{N_1}\log \alpha'_{m|k}(\varphi_1^{N_1})\right\}, \quad m,k = \overline{1,K}. \tag{11}$$

and it follows from (9), (10) and (11) that there are

$$E'_{k|k}(\varphi_1) = \min_{m\neq k} E'_{m|k}(\varphi_1).$$

We can get the LAO testing with the similar of Theorem 2 and use it for the first stage of two-stage test by a pair of samples.

5.2. Second Stage of Two-Stage Test by A Pair of Samples

After selecting a family of PDs, it is necessary to detect one PD in this family by test $\varphi_2^{N_2}(\mathrm{x}_2, \varphi_1^{N_1})$ which can be defined by division of the sample space $\mathcal{X}^{N_2}$ to R_k disjoint subsets $B_s^{N_2},\ s \in D_k$. So if the k-th family of PDs is accepted, then

$$B_s^{N_2} \overset{\Delta}{=} \{\mathrm{x}_2 : \varphi_2^{N_2}(\mathrm{x}_2, \varphi_1^{N_1}) = s\}, \quad s \in D_k,\ \ k = \overline{1,K},$$

The probability of the erroneous acceptance at the second stage of test, in which PD P_l is accepted when P_s is true, is

$$\alpha"_{l|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}(B_l^{N_2}),\ \ l \neq s,\ \ l \in D_k.$$

The probability to reject P_s, when it is true, is

$$\alpha"_{s|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}\left(\overline{B}_s^{N_2}\right) = \textstyle\sum_{l\neq s} \alpha"_{l|s}(\varphi_2^{N_2}), \quad s = \overline{1,S}. \tag{12}$$

Corresponding reliabilities for the second stage of test, are defined as

$$E"_{l|s}(\varphi_2) \overset{\Delta}{=} \limsup_{N_2\to\infty} \{-\frac{1}{N_2}\log \alpha"_{l|s}(\varphi_2^{N_2})\}, \quad l,s = \overline{1,S}. \tag{13}$$

It follows from (12) and (13) that

$$E"_{s|s}(\varphi_2) = \min_{l \neq s} E"_{l|s}(\varphi_2), \quad l, s = \overline{1, S}.$$

Theorem 5 (See [8, 9]). If at the first stage of test the k-th family of PDs is accepted, then for given $R_k - 1$ positive and finite values $E"_{M_{k-1}+s|M_{k-1}+s}$, $s = \overline{1, R_k - 1}$ of the reliabilities matrix $\boldsymbol{E}"^{(k|k)}(\varphi_2)$, let us consider the regions:

$$R''_s = \{Q\colon \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big) \leq E"_{M_{k-1}+s|M_{k-1}+s}\}, \quad s = \overline{1, R_k - 1},$$

$$R''_{R_k} = \{Q\colon \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big) > E''_{M_{k-1}+s|M_{k-1}+s}, s = \overline{1, R_k - 1}, \ \min_{s \in D_k} D(Q||P_s) \leq E'^{*}_{k|k}\},$$

and the following values of elements of the future reliabilities matrix $E"^{(k|k)}(\varphi_2^*)$ of the LAO test sequence:

$$E"^{*}_{M_{k-1}+s|M_{k-1}+s} = E"_{M_{k-1}+s|M_{k-1}+s}, \quad s = \overline{1, R_k - 1},$$

$$E"^{*}_{M_{k-1}+l|M_{k-1}+s} = \inf_{Q \in R''_l} \mathrm{D}\big(Q \parallel P_{M_{k-1}+s}\big), \quad l, s = \overline{1, R_k}, \ l \neq s,$$

$$E"^{*}_{M_k|M_k} = \min_{l \neq R} E"^{*}_{l|M_k}.$$

When the following compatibility conditions are valid

$$E"_{1|1} < \min_{s=\overline{2,R_k}} \mathrm{D}(P_{M_{k-1}+s} \parallel P_{M_{k-1}+1}),$$

$$E"_{s|s} < min[\min_{l=\overline{1,s-1}} E"^{*}_{M_{k-1}+l|M_{k-1}+s}, \min_{l=\overline{s+1,R_k}} \mathrm{D}(P_{M_{k-1}+l} \parallel P_{M_{k-1}+s})], \quad 2 \leq s \leq R_k - 1,$$

then there exists a LAO sequence of tests φ_2^*, elements of reliabilities matrix $E"^{(k|k)}(\varphi_2^*)$ of which are defined above and are positive.

Even if one of the compatibility conditions is violated, then the reliabilities matrix of such test has at least one element equal to zero.

5.3. Reliabilities of Two-Stage Test by A Pair of Samples

The tool of making decision according to N-sample denoted $\Phi_2^{*N} = (\varphi_1^{*N_1}, \varphi_2^{*N_2})$, which is organized by a pair of LAO tests $\varphi_1^{*N_1}$ and $\varphi_2^{*N_2}$. In the two-stage decision making, the test Φ_2^{*N} can be defined by division of the sample space $\mathcal{X}^N$ to S disjoint subsets as follows:

$$C_s^N \overset{\Delta}{=} A_k^{*N_1} \times B_s^{N_2}, \quad s = \overline{1, S},$$

such that at the first stage of test the k-th family of PDs is accepted and

$$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2) \in C_s^N\colon \quad \mathrm{x}_1 \in A_k^{*N_1}, \quad \mathrm{x}_2 \in B_s^{N_2}, \ \ s \in D_k.$$

Let $\alpha'''_{l|s}$ be the probability of the erroneous acceptance by two-stage test with a pair of samples of PD P_l when P_s is true:

$$\alpha'''_{l|s}(\Phi_2^{*N}) \overset{\Delta}{=} P_s^N(C_l^N), \quad l, s = \overline{1, S}, \quad l \neq s.$$

And the probability to reject P_s in two-stage test by a pair of samples, when it is correct, is

$$\alpha'''_{s|s}(\Phi_2^{*N}) \overset{\Delta}{=} P_s^N\left(\overline{C_s}^N\right) = \sum_{l\neq s} \alpha'''_{l|s}(\Phi_2^{*N}), \quad s = \overline{1,S}.$$

We score by $\Phi_2^* = (\varphi_1^*, \varphi_2^*)$ the infinite sequences of tests and define reliabilities:

$$E'''_{l|s}(\Phi_2^*) \overset{\Delta}{=} \limsup_{N\to\infty}\{-\frac{1}{N}\log\alpha'''_{l|s}(\Phi_2^{*N})\}, \quad l,s = \overline{1,S}.$$

We want to review the relationship between error probabilities and reliabilities for the two-stage test by a pair of samples and the first and the second stages of LAO tests. For that we consider error probabilities as follows

a) if $l,s \in D_k$ then

$$\alpha'''_{l|s}(\Phi_2^{*N}) = P_s^{N_1}(A_k^{*N_1}) \cdot P_s^{N_2}(B_l^{N_2}) \tag{14}$$

b) if $s \in D_k$ and $l \notin D_k$ then

$$\alpha'''_{l|s}(\Phi_2^{*N}) = P_s^{N_1}(\overline{A}_k^{*N_1}) \cdot P_s^{N_2}(B_l^{N_2}) \tag{15}$$

Using properties of types we create the following equalities:

$$\lim_{N_1\to\infty}\{-\frac{1}{N_1}\log P_s^{N_1}(\overline{A}_k^{*N_1})\} = \inf_{Q:\min\limits_{l:l\in D_1} D(Q||P_l)>E'^*_{k|k}} D(Q||P_s) \overset{\Delta}{=} E^I_{k|s}, \tag{16}$$

we can understand, when $s \notin D_k$, then $E^I_{k|s} = 0$ and if $s \in D_k$

$$\lim_{N_1\to\infty}\{-\frac{1}{N_1}\log P_s^{N_1}(A_k^{*N_1})\} = \inf_{Q:\min\limits_{l:l\in D_1} D(Q||P_l)\leq E'^*_{k|k}} D(Q||P_s) = 0. \tag{17}$$

According to (14)–(17) and definition of reliabilities we obtain

a) if $l,s \in D_k$ then

$$E'''_{l|s}(\Phi_2^*) = (1-\psi)E''^*_{l|s}, \tag{18}$$

b) if $s \in D_k$ and $l \notin D_k$ then

$$E'''_{l|s}(\Phi_2^*) = \psi E^I_{k|s} + (1-\psi)E''^*_{l|s}, \tag{19}$$

c) $E'''_{s|s}(\Phi_2^*) = \min\limits_{l\neq s} E'''_{l|s}(\Phi_2^*).$ (20)

Theorem 6. If all distributions P_s, $s = \overline{1,S}$, are different and positive values $E'^*_{k|k}$, $k = \overline{1,K-1}$ and $E''_{M_{k-1}+r|M_{k-1}+r}$, $r = \overline{1,R_k-1}$, satisfy compatibility conditions of Theorems 2 and 5, then elements of matrix of reliabilities $E'''(\Phi_2^*)$, of the two-stage test by a pair of samples Φ_2^* are defined in (18)–(20).

When one of the compatibility conditions is violated, then at least one element of the matrix $E'''(\Phi_2^*)$ is equal to zero.

6. COMPARISON OF RELIABILITIES MATRICES OF THREE METHODS

We compare the reliabilities matrix of the two-stage test and the reliabilities matrix of the one-stage test described in Theorem 1. For comparison we will give the same diagonal elements $E_{s|s} = E'''_{s|s}$, $s = \overline{1, S-1}$ of the reliabilities matrices. For $s = \overline{1, M_k - 1}$, $k = \overline{1,K}$ we have

$$R'''_s = \{Q\colon \mathrm{D}(Q \parallel P_s) \leq E'''_{s|s},\ \min_{s\in D_k} D(Q||P_s) \leq E'^{*}_{k|k}\}$$
$$= \{Q\colon \mathrm{D}(Q \parallel P_s) \leq E_{s|s},\ \min_{s\in D_k} D(Q||P_s) \leq E'^{*}_{k|k}\} = R_s,$$

and reliabilities are

$$E'''_{s|l} = \inf_{Q\in R'''_s} \mathrm{D}(Q \parallel P_l) = \inf_{Q\in R_s} \mathrm{D}(Q \parallel P_l) = E_{s|l}.$$

Theorem 7. If all distributions P_s, $s = \overline{1,S}$, are different and positive numbers of diagonal elements $E_{s|s} = E'''_{s|s}$, $s = \overline{1, S-1}$ of the reliability matrices of two-stage and one-stage cases, satisfy compatibility conditions shown in Theorem 2-6, then for columns $s = \overline{1, M_k - 1}$, $k = \overline{1,K}$, reliabilities of two matrices are equal, but for columns M_k, $k = \overline{1,K}$, reliabilities can be different.

The answer to the question: which is the best value of $E'^{*}_{k|k}$, $k = \overline{1, K-1}$ giving the best value to reliabilities $E''^{*}_{l|s}$, $l, s \in D_k$, is in the following .

Theorem 8. If distributions P_s, $s = \overline{1,S}$, are different and for given positive diagonal values $E_{s|s} = E''_{s|s}$, $s = \overline{1, M_k - 1}$, $k = \overline{1,K}$, of reliabilities matrix of the second stage of test, the bounds of reliability $E'^{*}_{k|k}$ of the first stage, satisfy the following conditions

$$\max_{s=\overline{1,M_k-1}} E''_{s|s} \leq E'^{*}_{k|k} \leq \min_{s\notin D_k, l\in D_k} \mathrm{D}(P_s \parallel P_l),$$

and the best value for it equals to the lower bound $\max_{s=\overline{1,M_k-1}} E''_{s|s}$.

Example. Suppose $X = \{a, b, c\}$, the first family of PDs P_1 includes three PDs $\{P_1, P_2, P_3\}$, the second family P_2 consists five PDs $\{P_4, P_5, P_6, P_7, P_8\}$ and the third family of PDs P_3 includes four PDs $\{P_9, P_{10}, P_{11}, P_{12}\}$ such that PDs are as follows:

$P_1 = (0.1,0.4,0.5)$, $P_2 = (0.2,0.5,0.3)$, $P_3 = (0.3,0.6,0.1)$, $P_4 = (0.4,0.3,0.3)$,
$P_5 = (0.5,0.4,0.1)$. $P_6 = (0.6,0.2,0.2)$, $P_7 = (0.7,0.2,0.1)$, $P_8 = (0.8,0.1,0.1)$,
$P_9 = (0.1,0.1,0.8)$, $P_{10} = (0.1,0.2,0.7)$. $P_{11} = (0.2,0.1,0.7)$, $P_{12} = (0.2,0.2,0.6)$.

The reliabilities matrix of the first stage of test for given value are the following:

$E''_{1\|1}$	$E''_{2\|2}$	$E''_{4\|4}$	$E''_{5\|5}$	$E''_{6\|6}$	$E''_{7\|7}$	$E''_{9\|9}$	$E''_{10\|10}$	$E''_{11\|11}$
0.005	0.0100	0.007	0.002	0.007	0.0004	0.009	0.001	0.001

$$\mathbf{E}'(\Phi_1^*) = \begin{bmatrix} 0.0100 & 0.0137 & 0.0100 \\ 0.0101 & 0.0070 & 0.0070 \\ 0.0127 & 0.1281 & 0.0127 \end{bmatrix}.$$

The reliabilities matrices of the one-stage test $\mathbf{E}(\phi^*)$ and the two-stage test $\mathbf{E}'''(\Phi_1^*)$ and the same diagonal elements are as follows

$$\mathbf{E}(\phi^*) = \begin{bmatrix}
0.0050 & 0.0112 & 0.0050 & 0.0762 & 0.2368 & 0.2341 & 0.4365 & 0.2020 & 0.0590 & 0.0310 & 0.0865 & 0.0100 \\
0.0183 & 0.0100 & 0.0100 & 0.0215 & 0.0851 & 0.1094 & 0.2353 & 0.0642 & 0.1682 & 0.1259 & 0.1662 & 0.0100 \\
0.1643 & 0.0223 & 0.0223 & 0.0556 & 0.0245 & 0.0947 & 0.1482 & 0.0137 & 0.4748 & 0.4084 & 0.4399 & 0.0100 \\
0.0674 & 0.0167 & 0.0460 & 0.0070 & 0.0354 & 0.0111 & 0.0769 & 0.0070 & 0.1593 & 0.1404 & 0.1276 & 0.0070 \\
0.2133 & 0.0560 & 0.0101 & 0.0273 & 0.0020 & 0.0189 & 0.0349 & 0.0020 & 0.4690 & 0.4218 & 0.4116 & 0.0070 \\
0.1914 & 0.0926 & 0.0933 & 0.0111 & 0.0322 & 0.0070 & 0.0124 & 0.0070 & 0.2824 & 0.2753 & 0.2261 & 0.0070 \\
0.3189 & 0.1557 & 0.1025 & 0.0517 & 0.0292 & 0.0026 & 0.0004 & 0.0004 & 0.4880 & 0.4700 & 0.4126 & 0.0070 \\
0.4294 & 0.2710 & 0.2293 & 0.1043 & 0.1043 & 0.0157 & 0.0141 & 0.0141 & 0.5165 & 0.5259 & 0.4317 & 0.0070 \\
0.0863 & 0.1855 & 0.0680 & 0.1808 & 0.4557 & 0.3143 & 0.5387 & 0.4168 & 0.0090 & 0.0112 & 0.0112 & 0.0090 \\
0.0206 & 0.0840 & 0.0127 & 0.1177 & 0.3429 & 0.2660 & 0.4838 & 0.3073 & 0.0017 & 0.0010 & 0.0202 & 0.0010 \\
0.0854 & 0.1465 & 0.0671 & 0.0963 & 0.3165 & 0.1708 & 0.3391 & 0.2835 & 0.0017 & 0.0202 & 0.0010 & 0.0010 \\
0.0234 & 0.0523 & 0.0151 & 0.0430 & 0.2097 & 0.1281 & 0.2854 & 0.1827 & 0.0125 & 0.0102 & 0.0102 & 0.0102
\end{bmatrix}.$$

$$\mathbf{E}'''(\Phi_1^*) = \begin{bmatrix}
0.0050 & 0.0112 & \mathbf{0.0890} & 0.0762 & 0.2368 & 0.2341 & 0.4365 & 0.4395 & 0.0590 & 0.0310 & 0.0865 & \mathbf{0.0050} \\
0.0183 & 0.0100 & 0.0100 & 0.0215 & 0.0851 & 0.1094 & 0.2353 & 0.2491 & 0.1682 & 0.1259 & 0.1662 & 0.0100 \\
0.1643 & 0.0223 & 0.0223 & 0.0556 & 0.0245 & 0.0947 & 0.1482 & 0.1751 & 0.4748 & 0.4084 & 0.4399 & \mathbf{0.0223} \\
0.0674 & 0.0167 & \mathbf{0.0280} & 0.0070 & 0.0354 & 0.0111 & 0.0769 & 0.0761 & 0.1593 & 0.1404 & 0.1276 & 0.0070 \\
0.2133 & 0.0560 & \mathbf{0.0027} & 0.0273 & 0.0020 & 0.0189 & 0.0349 & 0.0496 & 0.4690 & 0.4218 & 0.4116 & \mathbf{0.0020} \\
0.1914 & 0.0926 & \mathbf{0.0644} & 0.0111 & 0.0322 & 0.0070 & 0.0124 & 0.0081 & 0.2824 & 0.2753 & 0.2261 & 0.0070 \\
0.3189 & 0.1557 & \mathbf{0.0729} & 0.0517 & 0.0292 & 0.0026 & 0.0004 & 0.0004 & 0.4880 & 0.4700 & 0.4125 & \mathbf{0.0004} \\
0.4294 & 0.2710 & \mathbf{0.1818} & 0.1043 & 0.1043 & 0.0157 & 0.0141 & 0.0141 & 0.5165 & 0.5259 & 0.4317 & \mathbf{0.0141} \\
0.0863 & 0.1855 & \mathbf{0.3665} & 0.1808 & 0.4557 & 0.3143 & 0.5387 & 0.4882 & 0.0090 & 0.0112 & 0.0112 & 0.0090 \\
0.0206 & 0.0840 & \mathbf{0.2243} & 0.1177 & 0.3429 & 0.2660 & 0.4838 & 0.4589 & 0.0017 & 0.0010 & 0.0202 & 0.0010 \\
0.0854 & 0.1465 & \mathbf{0.2842} & 0.0963 & 0.3165 & 0.1708 & 0.3391 & 0.2871 & 0.0017 & 0.0202 & 0.0010 & 0.0010 \\
0.0234 & 0.0523 & \mathbf{0.1527} & 0.0430 & 0.2097 & 0.1281 & 0.2854 & 0.2609 & 0.0125 & 0.0102 & 0.0102 & 0.0102
\end{bmatrix},$$

The bold-face numbers are unequal reliabilities of the second matrix.

The reliabilities matrices of the one-stage test $\mathbf{E}(\phi^*)$ and the two-stage test with two samples $\mathbf{E}'''(\Phi_2^*)$ with $\psi = 0.1$ and the same diagonal elements, are as follows

$$\mathbf{E}(\phi^*) = \begin{bmatrix}
0.0050 & 0.0112 & 0.0958 & 0.0762 & 0.2368 & 0.2341 & 0.4365 & 0.4487 & 0.0590 & 0.0310 & 0.0865 & 0.0050 \\
0.0183 & 0.0100 & 0.0125 & 0.0215 & 0.0851 & 0.1094 & 0.2353 & 0.2559 & 0.1682 & 0.1259 & 0.1662 & 0.0100 \\
0.1643 & 0.0223 & 0.0185 & 0.0556 & 0.0245 & 0.0947 & 0.1482 & 0.1799 & 0.4748 & 0.4084 & 0.4399 & 0.0185 \\
0.0674 & 0.0167 & 0.0324 & 0.0070 & 0.0354 & 0.0111 & 0.0769 & 0.0798 & 0.1593 & 0.1404 & 0.1276 & 0.0070 \\
0.2133 & 0.0560 & 0.0042 & 0.0273 & 0.0020 & 0.0189 & 0.0349 & 0.0524 & 0.4690 & 0.4218 & 0.4116 & 0.0020 \\
0.1914 & 0.0926 & 0.0721 & 0.0111 & 0.0322 & 0.0070 & 0.0124 & 0.0094 & 0.2824 & 0.2753 & 0.2261 & 0.0070 \\
0.3189 & 0.1557 & 0.0805 & 0.0517 & 0.0292 & 0.0026 & 0.0004 & 0.0007 & 0.4880 & 0.4700 & 0.4125 & 0.0004 \\
0.4294 & 0.2710 & 0.1941 & 0.1043 & 0.1043 & 0.0157 & 0.0141 & 0.0124 & 0.5165 & 0.5259 & 0.4317 & 0.0124 \\
0.0863 & 0.1855 & 0.3797 & 0.1808 & 0.4557 & 0.3143 & 0.5387 & 0.4972 & 0.0090 & 0.0112 & 0.0112 & 0.0090 \\
0.0206 & 0.0840 & 0.2349 & 0.1177 & 0.3429 & 0.2660 & 0.4838 & 0.4678 & 0.0017 & 0.0010 & 0.0202 & 0.0010 \\
0.0854 & 0.1465 & 0.2966 & 0.0963 & 0.3165 & 0.1708 & 0.3391 & 0.2936 & 0.0017 & 0.0202 & 0.0010 & 0.0010 \\
0.0234 & 0.0523 & 0.1615 & 0.0430 & 0.2097 & 0.1281 & 0.2854 & 0.2674 & 0.0125 & 0.0102 & 0.0102 & 0.0102
\end{bmatrix}.$$

$$\mathbf{E}'''(\Phi_2^*) = \begin{bmatrix}
0.0050 & 0.0091 & 0.0050 & 0.0669 & 0.2122 & 0.2071 & 0.3923 & 0.1794 & 0.0581 & 0.0286 & 0.0838 & \mathbf{0.0115} \\
0.0155 & 0.0100 & 0.0100 & 0.0190 & 0.0761 & 0.0964 & 0.2117 & 0.0565 & 0.1550 & 0.1135 & 0.1550 & \mathbf{0.0115} \\
0.1447 & 0.0185 & 0.0185 & 0.0487 & 0.0221 & 0.0838 & 0.1334 & 0.0121 & 0.4284 & 0.3666 & 0.4006 & 0.0115 \\
0.0602 & 0.0150 & \mathbf{0.0406} & 0.0070 & 0.0311 & 0.0091 & 0.0686 & 0.0070 & 0.1470 & 0.1266 & 0.1207 & \mathbf{0.0085} \\
0.1904 & 0.0493 & \mathbf{0.0092} & 0.0229 & 0.0020 & 0.0160 & 0.0311 & 0.0020 & 0.4230 & 0.3792 & 0.3752 & \mathbf{0.0085} \\
0.1710 & 0.0818 & \mathbf{0.0819} & 0.0091 & 0.0283 & 0.0070 & 0.0110 & 0.0070 & 0.2569 & 0.2475 & 0.2088 & \mathbf{0.0085} \\
0.2851 & 0.1377 & \mathbf{0.0902} & 0.0446 & 0.0256 & 0.0019 & 0.0004 & 0.0004 & 0.4405 & 0.4221 & 0.3760 & \mathbf{0.0085} \\
0.3841 & 0.2404 & \mathbf{0.2027} & 0.0911 & 0.0924 & 0.0130 & 0.0124 & 0.0124 & 0.4661 & 0.4726 & 0.3934 & 0.0085 \\
0.0766 & 0.1639 & 0.0594 & 0.1601 & 0.4085 & 0.2793 & 0.4845 & 0.3717 & 0.0090 & 0.0097 & 0.0097 & 0.0090 \\
0.0187 & 0.0739 & 0.0115 & 0.1038 & 0.3075 & 0.2360 & 0.4347 & 0.2736 & 0.0013 & 0.0010 & 0.0177 & 0.0010 \\
0.0758 & 0.1288 & 0.0587 & 0.0851 & 0.2836 & 0.1514 & 0.3052 & 0.2525 & 0.0013 & 0.0177 & 0.0010 & 0.0010 \\
0.0211 & 0.0458 & 0.0134 & 0.0379 & 0.1878 & 0.1130 & 0.2566 & 0.1624 & 0.0103 & 0.0089 & 0.0089 & 0.0089
\end{bmatrix},$$

the bold-faced numbers are the reliabilities of two-stage test which are greater than the reliabilities of one-stage test.

7. CONCLUSION

In the two-stage test at the first stage we can give $K-1$ elements of reliabilities matrix $E'(\varphi_1^*)$ and at the second stage we can give $\sum_{i=1}^{K}(R_k-1)=S-K$ diagonal elements of reliabilities matrix $E"^{(k|k)}(\varphi_2^*),\ k=\overline{1,K},$ at result at the two-stage test we can use in calculations $S-1$ elements. We have shown that the number of the preliminarily given elements of the reliabilities matrix in one-stage and in two-stage tests by one sample and by two samples would be the same. Some element of the reliabilities matrix of the two-stage test by one sample and by two samples can be even greater than corresponding elements of the one-stage test. So the consumer has possibility to use the method which is preferable. We can show that the number of operations of the two-stage test by one sample is less than this of one-stage test and is more than quantity of operations of two-stage test by a pair of samples. This was observed also during experimental calculations of example.

REFERENCES

1. Ahlswede, R. and Haroutunian, E.A. (2006). On statistical hypotheses optimal testing and identification. *Lecture Notes in Computer Science*, vol. 4123, General Theory of Information Transfer and Combinatorics, Springer, 462-478.
2. Birge L. (1981). Ritessess maximates de decroissence des errors et tests optimaux associts, Wahr sehein lichkeitstheorie Verv. Geliete, vol. 55, 261-273.
3. Blahut, R.E. (1974). Hypotheses testing and information theory, *IEEE Trans. Inform Theory*, 20(4), 405-417.
4. Cover, T.M. and Tomas, J.A. (2006). Elements of Information Theory, Second edition, Wiley, New York.
5. Csiszár, I. and Körner, J. (1981). Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic Press, New York.
6. Csiszár, I. and Longo, G. (1971). On the error exponent for source coding and for testing simple statistical hypotheses. *Studia Sc. Math. Hungarica*, 6, 181-191.
7. Haroutunian, E.A. (1990). Logarithmically asymptotically optimal testing of multiple statistical hypotheses. *Problems of Control and Information Theory*, 19(5-6), 413-421.
8. Haroutunian, E.A. (2005). Reliability in multiple hypotheses testing and identification. *Proceedings of the NATO-ASI Conference*, vol. 198 of NATO Science Series III: Computer and Systems Sciences, Yerevan, Armenia, 189-201, IOS Press.
9. Haroutunian, E.A., Haroutunian, M.E. and Haroutunian, A.N. (2008). Reliability criteria in information theory and in statistical hypothesis testing. *Foundations and Trends in Communications and Information Theory*, 4(2-3).
10. Haroutunian, E.A. and Hakobyan, P.M. (2009). Multiple hypotheses LAO testing for many independent objects. Scholarly Research Exchange.
11. Haroutunian, E.A., Hakobyan, P.M. and Hormozi nejad, F. (2012). On two-stage logarithmically asymptotically optimal testing of multiple hypotheses concerning distributions from the pair of families". Transactions of IIAP of NAS of RA and of YSU, Mathematical Problems of Computer Science, 37, 34-42.
12. Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, 36, 369-401.

13. Hormozi nejad, F., Haroutunian E.A. and Hakobyan, P.M. (2012). On Two-stage LAO Testing of Multiple Hypotheses for the Pair of Families of Distributions. *Electronic Journal of Statistics*, 6, 1-25.
14. Hormozi nejad, F., Haroutunian E.A. and Hakobyan P.M. (2011). On LAO testing of multiple hypotheses for the pair of families of distributions. *Proceeding of the Conference Computer Science and Information Technologies*, Yerevan, Armenia, 135-138.
15. Tusnady, G. (1977). On asymptotically optimal tests. *Annals of Statistics*, 5(2), 385-393.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 359-366

GENERALIZED EXTREME VALUE DISTRIBUTIONS

M. Ahsanullah
Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08648, USA
Email: ahsan@rider.edu

ABSTRACT

Extreme Value Distributions (GEV) are the limiting distributions for the normalized maximum or minimum of a very large collection of independent random variables from the same distribution. Fisher and Tippett (1928) showed that maximum of a sample independent and identically distribution (iid) after proper normalization converges in distribution to one of the three possible distribution, the Gumbel distribution, the Frechet distribution and the Weibull distribution. The same is true for the minimum of a sample from an iid. These three distributions are special cases of generalized extreme value distribution. In this talk some new results dealing with records of the the GEV will be presented

1. INTRODUCTION

A random variable X is said to have the generalized extreme value distribution if its cumulative distribution function is of the following form:

$$F(x) = \exp[-\{ 1- \gamma \sigma^{-1} (x - \mu)\}^{1/\gamma}] \quad (1.0.1)$$

where $\sigma > 0$, $\gamma \neq 0$ and

$$x < \mu + \sigma \gamma^{-1}, \text{ for } \gamma > 0$$

$$x > \mu + \sigma \gamma^{-1}, \text{ for } \gamma < 0. \quad (1.0.2)$$

If $\gamma = 0$ then

$$F(x) = \exp[-\exp\{ - (x - \mu) / \sigma \}], \sigma > 0, - \infty < x < \infty. \quad (1.0.3)$$

We will write $X \in \text{GEV}(\mu,\sigma, \gamma)$ if X has the cdf as given in (1.0.1).

Since

$$\lim_{\gamma \to 0} \{1-\gamma \sigma^{-1} (x-\mu)\}^{1/\gamma} = \exp\{-\sigma^{-1} (x-\mu)\} \text{, we can take}$$

$$\lim_{\gamma \to 0} GEV(\mu , \sigma ,\gamma) = GEV(\mu ,\sigma ,0) .$$

The density function of $\text{GEV}(\mu,\sigma, \gamma)$ is

$$f(x) = \sigma^{-1}\{1-\gamma\sigma^{-1}(x-\mu)\}^{\frac{1-\gamma}{\gamma}} \exp[-\{1-\gamma\,\sigma^{-1}(x-\mu)\}^{1/\gamma}],\ \gamma \neq 0$$

$x < 1/\gamma$, for $\gamma > 0$,

$x > 1/\gamma$, for $\gamma < 0$,

and

$f(x) = e^{-x}\exp(-e^{-x})$, for $\gamma = 0$, for all x.

The extreme value distribution for $\gamma = 0$, is also known Gumbel distribution.

The largest order statistic $X_{n,n}$ when properly standardized tends to one of the following three types of limiting distributions as n→ ∞.

1) Type 1: (Gumbel) $F(x) = \exp(e^{-x})$, for all x,

2) Type 2: (Frechet) $F(x) = \exp(-x^{-\delta})$, $x > 0$, $\delta > 0$

3) Type 3: (Weibull) $F(x) = \exp(-(-x)^{\delta})$, $x<0$, $\delta >0$.

Since the smallest order statistic $X_{1,n} = Y_{n,n}$, where $Y = -X$, $X_{1,n}$ when properly standardized will also converge to one of the above three limiting distributions. Gumbel(1958) has given various applications of these distributions. The Type 1 (Gumbel distribution) is the limiting distribution of $X_{n,n}$ when F(x) is normal, log normal, logistic, gamma etc. The generalized extreme value distribution (1.0.1) has been discussed by Jenkinson (1955). It includes as special case the above three well known extreme value distributions.

The type 2 and type 3 distributions can be transformed to Type 1 distribution by the transformations $V_2 = \ln X$ and $V_3 = -\ln X$ respectively.

These distributions were originally introduced by Fisher and Tippet (1928). Extreme value distributions have been used in the analysis of data concerning floods, extreme sea levels and air pollution problem; for details see Gumbel (1958), Horwitz (1980), Jenkinson (1955) and Roberts (1979).

For a given set of n observations, let $X_{1,n} < \ldots. < X_{n,n}$ be the associated order statistics. Suppose that $P\{a_n(X_{n,n} - b_n) < x\} \to G(x)$ as $n \to \infty$ for some suitable constants a_n and b_n. Then it is known (see Leadbetter et al, 1983, p.33) that

$$P\{a_n\,(X_{n-m,n} - b_n) \le x\} \xrightarrow{d} G(x)\sum_{s=0}^{m-1}\frac{[-\ln G(x)]^s}{s!} \quad .$$

We have already seen that the right hand side of the above expression is the cdf of the m th lower record value from the distribution function G(x).

Thus the limiting distribution of the (n- m + 1)th order statistic (m finite) as n → ∞ from the generalized extreme value distribution is the same as the m th lower record value from the generalized extreme value distribution. In this chapter we will study the lower record values of GEV (μ,σ,γ).

2. DISTRIBUTIONAL PROPERTIES OF RECORD VALUES

If X $\in$ GEV (μ,σ,γ), then using (1.1.14) we can write for $\gamma \neq 0$, the pdf f(m) of the m th lower record value as

$$f_{(m)}(x) = \{1-\gamma\,\sigma^{-1}(x-\mu)\}^{(m-1)/\gamma}\, f_m^*(x) \qquad (2.1)$$

where

$$f_m^*(x) = \frac{\{1-\gamma\,\sigma^{-1}(x-\mu)\}^{(1-\gamma)/\gamma}}{\sigma\,(m-1)\,!}\exp\{-(1-\gamma\sigma^{-1}\,(x-\mu)\,)\}^{1/\gamma}$$

and for $\gamma = 0$,

$$f_{(m)}(x) = \frac{e^{-m\,\sigma^{-1}(x-\mu)}}{\sigma\,(m-1)\,!}\exp\{-e^{-\sigma^{-1}(x-\mu)}\}\ ,\ m = 1,2,. \qquad (2.2)$$

From (2.1) and (2.2) it can be shown that

$$X_{L(m)} \underset{=}{d} \quad \mu + \sigma\,\gamma^{-1}\,\{\,1 - (\,W_1 + \cdots\cdots + W_m)^{\gamma}\,\}\,,\ \gamma \neq 0, \qquad (2.3)$$

and

$$X_{L(m)} \underset{=}{d} \quad X - \sigma\,(\,W_1 + \frac{W_2}{2} + \ \ldots\ldots \ + \frac{W_{m-1}}{m-1})\,, \qquad \gamma = 0, \qquad (2.4)$$

where W_1, W_2, W_m are independently distributed as exponential random variables with mean unity.

$$E(X_{L(m)}) = \mu + \sigma\,\gamma^{-1}\,.\{\,1 - \Gamma\,(m+\gamma\,)\,/\,\Gamma(m)\}.$$

$$\begin{aligned}\mathrm{Var}\,(X_{L(m)}) &= \sigma^2\,\gamma^{-2}\,[\,E(\,W_1 + \cdots + W_m)^{2m} - \{\,E(W_1 + \cdots + W_m)\}^2]\\ &= \sigma^2\,\gamma^{-2}[\ \frac{\Gamma(m+2\,\gamma)}{\Gamma(m)} - \{\,\frac{\Gamma(m+\gamma)}{\Gamma(m)}\}^2\ \ ].\end{aligned}$$

For r < m

$$\begin{aligned}\gamma^2\sigma^{-2}\ \mathrm{Cov}\,(X_{L(r)}, X_{L(m)}\,) &= \{(\sum_{j=1}^{r} W_j)^{\gamma}\,(\sum_{j=1}^{m} W_j)^{\gamma} - E(\sum_{j=1}^{r} W_j\,)^{\gamma}\ E(\sum_{j=1}^{m} W_j\,)^{\gamma}\\ &= \int_0^{\infty}\int_0^{\infty} u^{\gamma}\,(u+v)^{\gamma}\ \frac{e^{-u}\,u^{r-1}}{\Gamma(r)}\ \frac{e^{-v}\ v^{m-r-1}}{\Gamma(m-r)}\,du\,dv\\ &= \frac{\Gamma(r+\gamma)\,\Gamma(\,r+2\,\gamma)}{\Gamma(r)\,\Gamma(r+\gamma)} - \frac{\Gamma(r+\gamma)\,\Gamma(m+\gamma)}{\Gamma(r)\,\Gamma(m)}\ ,\end{aligned}$$

since u and v are independent. Thus

$$\mathrm{Cov}\,\{X_{L(r)}, X_{L(m)}\} = \sigma_o^2\ a_r\ b_m, \qquad r < m$$

where

$$a_r = \frac{\Gamma(\,r+\gamma)}{\Gamma(r)}\,, \quad b_m = \frac{\Gamma(m+2\gamma)}{\Gamma(m+\gamma)} - \frac{\Gamma(m+\gamma)}{\Gamma(m)} \;\; and\; \sigma_o^2 = \frac{\sigma^2}{\gamma^2}\;.$$

For X ∈ GEV (μ,σ,0), the joint pdf of $Y = H(X_{L(m+1)}) / H(X_{L(m)})$ is

$$f_Y{}^{*}\,(y) = m\, y^{\,m-1,}\; 0 < y < \infty. \qquad (2.5)$$

Thus $(Y)^m$ is distributed as uniform over the interval (0,1). Consequently $m[-\ln H(X_{L(m)}) + \ln H(X_{L(m+1)})]$ is distributed as negative exponential distribution with mean unity. Since $-\ln H(x) = \dfrac{x-\mu}{\sigma}$, we have $m[X_{L(m)} - X_{L(m+1)}]$ is E(0,σ). Using this property or (2.4), we obtain for $\gamma = 0, E(X_{L(r)}) = \mu + \upsilon_r^* \sigma$

$$\mathrm{Var}\,(X_{L(m)}) = \sigma^2 V_{r,r}^*\,,\ r = 1,2, ...$$

$$\mathrm{Cov}\,(X_{L(r)}, X_{L(m)}) = \mathrm{Var}\,(X_{L(m)}),\ r < m,$$

with

$$\upsilon_1^* = \upsilon$$

$$\upsilon_j^* = \upsilon_{j-1}^* - (j-1)^{-1}\,,\ j \ge 2,$$

$$V_{1,1}^* = \frac{\pi^2}{6}\,,$$

....

$$V_{j,\,j}^* = V_{j-1,\,j-1}^* - (j-1)^{-2}\,, \quad j \ge 2$$

2. CHARACTERIZATION

We have $S_{(m)} = m(\,X_{L(m)} - X_{L(m+1)})$, m=1,2,…as identically distributed negative exponential. random variables. Arnold and Villasenor (1997) raised the question whether the identical distribution of S1 and 2 S_2 are i.i.d. negative exponential with unit mean can characterize the Gumbel distribution. Al-Zaid and Ahsanullah (2003) proved the following theorem.

Theorem 2.1.

Let $\{X_j,\ j = 1,….\}$ be a sequence of independent and identically distributed random variables with absolutely continuous (with respect to Lebesgue measure) distribution function F(x). Then the following two statements are identical.

(a) $F(x) = e^{-e^{-x}}\,,\ -\infty < x < \infty,$

(b) for a fixed m >1, the condition $X_{L(m)} \underset{=}{d} X_{L(m+1)} + \dfrac{W}{m}$, where W is independent of $X_{\,L(n)}$ and $X_{L(n+1)}$ and $X_{L(n+1)}$ and is distributed as exponential mean unity.

Proof.

It is easy to show that (a) ⇒ (b),

We will prove here that (b) ⇒ (a).

Suppose that for a fixed m > 1, $X_{L(m)} \underset{=}{d} X_{L(m+1)} + \frac{W}{m}$, then

$$F_{(m)}(x) = \int_{-\infty}^{x} P(W \le m(x-y)\, f_{(m+1)}(y)\,dy$$

$$= \int_{-\infty}^{x} [1 - e^{-m(x-y)}\, f_{(m+1)}(y)\,dy$$

$$= F_{(m+1)}(x) - \int_{-\infty}^{x} e^{-m(x-y)}\, f_{(m+1)}(y)\,dy. \qquad (2.1)$$

Thus

$$e^{mx}[F_{(m+1)}(x) - F_{(m)}(x)] = \int_{-\infty}^{x} e^{my}\, f_{(m+1)}(y)\,dy \qquad (2.2)$$

Using the relation (1.1.7), we obtain

$$e^{mx}\frac{F(x)\,H^m(x)}{\Gamma(m+1)} = \int_{-\infty}^{x} e^{my}\, f_{(m+1)}(y)\,dy \qquad (2.3)$$

Taking the derivatives of both sides of (2.3), we obtain

$$\frac{d}{dx}\left[e^{mx}\frac{H^m(x)}{\Gamma(m+1)}F(x)\right] = \quad e^{mx}\, f_{(n+1)}(x) \qquad (2.4)$$

This implies that

$$\frac{d}{dx}\left[e^{mx}\frac{H^m(x)}{\Gamma(m+1)}\right]F(x) = 0 \ . \qquad (2.5)$$

Thus

$$\frac{d}{dx}\left[e^{mx}\frac{H^m(x)}{\Gamma(m+1)}\right] = 0 \ . \qquad (2.6)$$

Hence

$$H(x) = c\, e^{-x}, \ -\infty < x < \infty \qquad (2.7)$$

Thus

$$F(x) = e^{-ce^{-x}} \ -\infty < x < \infty\ . \qquad (2.8)$$

Since F(x) is a distribution, assuming F(0) =e^{-1} , we obtain

$$F(x) = e^{-e^{-x}}, \ -\infty < x < \infty. \qquad (2.9)$$

Corollary 2.1.

If for some fixed m > 1, $X_{U(m+1)} \underset{=}{d} X_{U(m)} + \frac{W}{m}$, then we get a characterization of the Gumbel distribution with F(x) = $1 - e^{-e^{x}}$, $-\infty < x < \infty$.

Corollary 2.2.

If m =1, then relation $X_{U(2)} \underset{=}{d} X_{U(1)} + W,$ will give a characterization of the negative exponential distribution.

Remark 2.3.

The condition that any one of the statistics $m(X_{L(m)} - X_{L(m+1)})$, $m(X_{U(m+1)} - X_{U(m)})$, $(X_{L(m)} - X_{L(m+1)})$ or $(X_{U(m+1)} - X_{U(m)})$ is distributed as negative exponential do not characterize any distribution.

$\gamma = 0.60$,

$\hat{\mu} = 78.74, \hat{\sigma} = 3.97$, log likelihood for MVLUE's = 78.33

$\hat{Var}(\hat{\mu}) = 0.78\sigma^2, \hat{Var}(\hat{\sigma}) = 0.13\sigma^2$

$\tilde{\mu} = 78.33, \tilde{\sigma} = 3.51, \tilde{MSE}(\tilde{\mu}) = 0.77\sigma^2$ and $\tilde{MSE}(\tilde{\sigma}) = 0.12\sigma^2$.

The Log likelihood for BLIE's = -21.70.

Since Type 2 and Type 3 distributions are related by a change of sign, we give here characterizations of type 2 distribution.

Theorem 2.2.1.

Let$\{X_j$, j =1,….} be a sequence of independent and identically distributed random variables with absolutely continuous (with respect to Lebesgue measure) distribution function F(x). Then the following two statements are identical.

(a) F(x) = exp(- x $^{-\delta}$), x > 0, $\delta > 0$

(b) $X^{-\delta}_{L(n+`1)}) \underset{=}{d} X^{-\delta}_{L(n)} + W, n \geq o,$

where W is independent of $X_{L(2)}$ and X_i and W and is distributed as exponential with unit mean and F(0) = $e^{-1.}$.

Proof.

Let $Y_i = X_i^{-\delta}$ and then the pdf of Y_i is f(y) =exp(-y), $y \geq 0_=$ let $Y_{U(i)}$ be the upper record values of the sequence $\{Y_{i,\ i=1,2,\ldots\}.}$, Then the relation (b) is equivalent to

$$Y_{U(n+1)} \underset{=}{d} Y_{U(n)} + W \tag{2.10}$$

Using (3,1) it is easy to show that (a) $\Longrightarrow$(b).

Now we prove (b) $\Longrightarrow$(a).

Let F^*_n and f^*_n be the cdf and pdf of $Y_{U(n)}$, then, we have

$$F^*_{n+1}(x) = \int_0^x (1-e^{-(x-y)}) f^*_n(y)dy\,. \tag{2.11}$$

We assume F^*_k and f^*_k as the cdf and pdf of $Y_{U(k)}, k \geq 0,$ respectively.

From (3.2), we obtain

$$e^x[F^*_n(x) - F^*_{n+1}(x) = \int_0^x e^y f^*_n(y)dy \tag{2.11}$$

It is known (see Ahsanullah (2004) p.) that

$$F^*_n(x) - F^*_{n+1}(x) = (1-F^*(y))\frac{(-\ln(1-F^*(x))^n}{\Gamma(n+1)} \tag{2.12}$$

Thus

$$e^x[F^*_n(x) - F^*_{n+1}(x) = \int_0^x e^y f^*_n(y)dy \tag{2.13}$$

It is known (see Ahsanullah (2004) p.4) that

$$e^x(1-F^*(x))\frac{(-\ln(1-F^*(x))^n}{\Gamma(n+1)}$$
$$= \int_0^x e^y \frac{(-\ln(1-F^*(x))^{n-1}}{\Gamma(n)} f^*(y)dy, \tag{2.16}$$

where $F^*(y)$ and $f^*(y)$ are the cdf and pdf of Y_i respectively.

Differentiating both sides of (3.6) with respect to x , we obtain on simplification

$$1-F^*(x))-f^*(x)=0 \quad \text{for all } x \tag{2.17}$$

The solution of (2.177) with the boundary condition $F^*(0)=0$ and $F^*(\infty)=1$ is

$$F^*(x)=0=1-e^{-x}\ ,x \geq 0. \tag{2.18}$$

Hence the distribution of X is type 2 extreme value.

REFERENCES

1. Ahsanullah, M. (2004). *Record Values-Theory and Applications*. University Press of America, Lanham, MD.
2. Ahsanullah, M. and Nevzorov, V.B. (2001a). *Ordered Random Variables*. Nova Science Publishers Inc, New York, NY.
3. Al Zaid, A.A. and Ahsanullah, M.A. (2003). Characterization of the Gumbel distribution based on Record Values. *Communications in Statistics-Theory and Methods*, 32(12), 2101-2103.
4. Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N. (1998). *Records*. John Wiley& Sons Inc., New York. NY.
5. Fisher, R.A. and Tipette, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Procs Cambridge Philos Soc*., 24, 180-190.
6. Galambos, J. (1987). *The Asymptotic Theory of Extreme Order Statistics*. Robert E. Krieger Publishing Co. Malabar, FL.
7. Gnedenko, B. (1943). Sur la Distribution Limite du Terme Maximum d'une Serie Aletoise. *Ann. Math*., 44, 423-453.
8. Kotz, S. and Nadaraja, S. (2000). *Extreme Value Distributions- Theory and Applications*. Imperial College Press.
9. Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York, NY.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 367-374

COMPARISON OF METHODS FOR FITTING COX MODELS WITH RANDOM HERD AND TREATMENT-BY-HERD EFFECTS

Adel Elghafghuf[1] and **Henrik Stryhn**[2]

[1] Centre for Veterinary Epidemiological Research, University of Prince Edward Island, Charlottetown, PEI C1A4P3, Canada. Email: aelghafghuf@upei.ca
Department of Statistics, University of Misurata, P.O. Box 2478, Misurata, Libya
[2] Centre for Veterinary Epidemiological Research, University of Prince Edward Island, Charlottetown, PEI C1A4P3, Canada. Email: hstryhn@upei.ca

ABSTRACT

In many veterinary studies, including clinical trials and epidemiological investigations, data are clustered into groups such as herds. When such data are analyzed, it has become more and more popular to look for possible heterogeneity in outcome between herds. However, beyond the investigation of such heterogeneity, it is also interesting to consider heterogeneity in treatment effect over herds. For time-to-event outcomes, this may be investigated by including a random herd effect and random treatment-by-herd interaction in a Cox model.

Various estimation methods have been proposed to deal with such models. Four estimation methods used to fit these models are compared through a simulation study with settings resembling a real dataset. The performance of the four methods is investigated in terms of the bias of point estimates and their empirical standard errors in both fixed and random effect parts. The simulation study showed some differences between the four estimation methods.

KEY WORDS

Survival analysis, Cox model, hazard, random effects, estimation methods.

1. INTRODUCTION

Survival data from epidemiological veterinary studies involving animals from multiple herds is a typical example of multilevel survival data, also referred to as correlated or clustered survival data. For simplicity here, we will refer throughout this paper to animals within herds as subjects nested in clusters. Proportional hazards models with random effects acting multiplicatively on an unspecified baseline hazard, also referred to as random effects Cox models, are commonly used for multilevel survival data. The Cox model with univariate random effect, also called a shared frailty model, in which subjects within a same cluster share the same random cluster effect (frailty), provides a simple way to account for within cluster correlations. A more complex survival model of potential interest for epidemiological studies is a Cox model with random cluster effects and random treatment-by-cluster effects where treatment effects vary between clusters. Several estimation methods have been proposed to fit this complex survival model.

In this paper, we compare through a simulation study four commonly used estimation methods in epidemiology, all accessible in standard statistical software, for estimating random effects Cox models. These methods are: the maximum penalized partial likelihood (MPPL) method proposed by Ripatti et al. (2000), the Poisson maximum likelihood (PML) method used by Rabe-Hesketh et al. (2004), the maximum penalized likelihood approach (MPL) proposed by Rondeau et al. (2008), and a Bayesian approach applied by Yamaguchi et al. (2002).

The rest of this paper is organized as follows. In Section 2, we introduce a Cox model with random effects. In Section 3, the four estimation approaches are reviewed. Section 4 contains a simulation study and its results. The paper ends with a discussion of our findings in Section 5.

2. COX MODEL WITH RANDOM EFFECTS

In the following, we consider clustered survival data from a total of N subjects that come from m different clusters. The subject j in the cluster i is observed until either an event time T_{ij} or a right censoring time C_{ij} independent of T_{ij}. Let $Y_{ij} = \min\,(T_{ij}, C_{ij})$ be the observed time and $\delta_{ij} = I_{\{T_{ij} \leq C_{ij}\}}$ be the event indicator. For each subject, we also observe a binary predictor x_{ij} representing a treatment at the subject level ($x_{ij} = 0$: no treatment, $x_{ij} = 1$: treatment). The simplest model for such data that takes into account the correlation occurring in the data is the univariate random effects Cox model. This model is given by

$$h(t_{ij}|b_{i0}) = h_0(t_{ij})\,exp(b_{i0} + \beta x_{ij}) \qquad (1)$$

where $h(t_{ij}|.)$ is the conditional hazard function for the j^{th} individual from the i^{th} cluster at time t_{ij}, $h_0(t_{ij})$ is the baseline hazard at time t_{ij}, β is a fixed overall treatment coefficient, and b_{i0} is the random effect for the i^{th} cluster. In model (1), the baseline hazard $h_0(t_{ij})$ is left completely arbitrary. The random effects $b_{i0}; i = 1, \dots, m$, are $i.i.d$ from a density function $f(.)$. Model (1) can be rewritten as follows:

$$h(t_{ij}|u_{i0}) = h_0(t_{ij})u_{i0}\,exp(\beta x_{ij}) \qquad (2)$$

where $u_{i0} = e^{b_{i0}}$ is the frailty term acting multiplicatively on the baseline, and the $i.i.d$ random effects $u_{i0}\,; i = 1, \dots, m$, which are assumed to have a specific density function $f_U(.)$. The most common choices for frailty density functions are from the one-parameter gamma and log-normal distributions. Model (1) can be extended by adding an extra random effect within the same cluster to account for possible variation occurring in the treatment effect, this random effect represents the possible treatment-by-cluster interaction. The proportional hazard Cox model with random cluster and treatment-by-cluster effects takes the form,

$$h(t_{ij}|\beta, b_i) = h_0(t_{ij})exp\left(b_{i0} + x_{ij}(\beta + b_{i1})\right) \equiv h_0(t_{ij})\,exp(x_{ij}^{'}\beta + z_{ij}^{'}b_i) \quad (3)$$

where b_{i1} is the random treatment effect. The variables b_{i0} and b_{i1} are assumed to be normally distributed with mean zero and variance-covariance matrix $\Sigma(\theta)$. Given the random effects b_i, observations within cluster i are assumed to be independent. Therefore, the conditional likelihood function for cluster i is:

$$L_i^c(h_0,\beta,b_i) = \prod_{j=1}^{n_i}\left[h_0(t_{ij})\, exp\left(x_{ij}'\beta + z_{ij}'b_i\right)\right]^{\delta_{ij}} \\ exp\left\{-H_0(t_{ij})\, exp\left(x_{ij}'\beta + z_{ij}'b_i\right)\right\} f(b_i) \qquad (4)$$

where $H_0(t_{ij}) = \int_0^{t_{ij}} h_0(v)\, dv$ is the cumulative baseline hazard.

Assuming conditional independence for observations within a cluster and independence between clusters, the conditional likelihood function takes the form

$$L^c(h_0,\beta,b) = \prod_{i=1}^{m} L_i^c(h_0,\beta,b_i) = \prod_{i=1}^{m} exp\{l_i^c(h_0,\beta,b_i)\} \qquad (5)$$

where $l_i^c = \ln L_i^c$. Thus, the conditional log-likelihood for all the clusters can be expressed as

$$l^c(h_0,\beta,b) = \sum_{i=1}^{m} l_i(h_0,\beta,b_i) \qquad (6)$$

3. ESTIMATION METHODS

3.1. Maximum Penalized Partial Likelihood (MPPL)

This approach was proposed by Ripatti et al. (2000) and became in the recent years one of the most commonly used estimation approaches for semiparametric survival models with random effects. Ripatti et al. (2000) approximate the marginal log-likelihood using the Laplace method for integral approximation. Assuming normally distributed random effects with variance-covariance matrix $\Sigma(\theta)$, the marginal likelihood can be defined as

$$L_M(h_0,\beta,\theta) = c|\Sigma(\theta)|^{-\frac{m}{2}} \int e^{-k(b)} db \qquad (7)$$

where

$$k(b) = l^c(h_0,\beta,b) - \frac{1}{2} b'\Sigma(\theta)^{-1} b \qquad (8)$$

Ripatti et al. (2000) used the Laplace method for integral approximation to approximate the logarithm of (7). Ignoring the constant in (7), the approximated marginal log-likelihood takes the form,

$$l_M(h_0,\beta,\theta) \approx -\frac{m}{2} ln|\Sigma(\theta)| - \frac{1}{2} ln\left|k''(\tilde{b})\right| - k(\tilde{b}) \qquad (9)$$

The $k'(b)$ and $k''(b)$ are the first two order partial derivatives of $k(b)$ with respect to b, and $\tilde{b} = \tilde{b}(\beta,\theta)$ is the solution of $k'(\tilde{b}) = 0$.

For fixed θ, Ripatti et al. (2000) show that the values $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$, that maximize the penalized log-likelihood in (8), also maximize the penalized partial log-likelihood

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \delta_{ij} \left[\left(x_{ij}'\beta + z_{ij}'b_i\right) - ln \sum_{t_{kl} \geq t_{ij}} exp\left(x_{kl}'\beta + z_{kl}'b_k\right)\right] - \frac{1}{2} b'\Sigma(\theta)^{-1} b \qquad (10)$$

Based on the penalized partial log-likelihood (10), the estimating equations for $\beta(\theta)$ and $b(\theta)$ given θ can be derived. By alternating between solving the estimating equations, the $\hat{\beta}(\theta)$, $\hat{b}(\theta)$ can be obtained. Once $\hat{\beta}(\theta)$ and $\hat{b}(\theta)$ are computed, θ can be updated by maximizing the approximate profile likelihood derived from (7):

$$l_{PPL}\left(\hat{\beta}(\theta), \hat{b}(\theta), \theta\right) \approx -\frac{m}{2} ln|\Sigma(\theta)| - \frac{1}{2} ln\left|k''(\hat{b})\right| - \frac{1}{2} \hat{b}'\Sigma(\theta)^{-1}\hat{b} \qquad (11)$$

Ripatti et al. (2000) suggested using $k''_{PPL}(\hat{b}) = \frac{\partial^2 l_{PPL}}{\partial b \partial b'}$ instead of $k''(\hat{b})$ and they motivated that by its better empirical performance in the simulations. The estimates of the variance-covariance matrix for $\hat{\beta}$, can be obtained as in standard Cox regression using the random effects estimates as an offset. On the other hand, for estimating the variance-covariance matrix for $\hat{\theta}$, we need to differentiate the approximate profile likelihood in (11) twice with respect to θ and take the expectation with respect b.

3.2. Maximum Penalized Likelihood (MPL)

This approach is proposed by Rondeau et al. (2008); they use the full likelihood instead of a partial likelihood and penalize the hazard function instead of penalizing the frailties. Rondeau et al. (2008) introduce a smooth baseline hazard estimator by maximizing the penalized log-likelihood and propose to add a penalty term for the roughness of the baseline hazard to the marginal log-likelihood. This roughness penalty term is product of a smoothing parameter ν and the integral of the squared second derivative of the baseline hazard. The penalized log-likelihood is thus defined as:

$$l_{PL}(h_0, \beta, \theta) = l_M(h_0, \beta, \theta) - \nu \int_0^{\infty} [h_0''(t)]^2 dt \qquad (12)$$

where $l_M(h_0, \beta, \theta)$ is the logarithm of marginal likelihood defined in (7), and ν is a positive smoothing parameter that controls the trade-off between the data fit and the smoothness of the function $h_0(.)$.

For modeling the hazard, Rondeau et al. (2008) suggest to approximate $h_0(.)$ on a basis of splines. As they use cubic M-splines that are easy to integrate or differentiate, the second derivative of h is approximated by a linear combination of piecewise polynomials. Such approximation reduces the number of parameters and allows for flexible shapes of hazard functions. The approximation error can be reduced by increasing the number of knots. In other words, the more knots are used; the closer was the estimate to the true hazard. Rondeau et al. (2003), Rondeau et al. (2006), and Rondeau et al. (2008) mentioned that the smoothing parameter can be chosen either by maximizing a likelihood cross-validation criterion as in Joly et al. (1998) or by fixing the number of degrees of freedom to estimate the hazard function as described in Gary (1992) and Rondeau et al. (2003). The estimated of variance-covariance matrix for model parameters is obtained directly from the converged Hessian matrix in the maximization process.

3.3. Poisson Maximum Likelihood (PML)

One approach to fitting model (3) is to translate the Cox model with random effects into a random-effects Poisson model framework (Rabe-Hesketh et al., 2004) because the likelihood function of a proportional hazards mixed model with unspecified baseline hazards can be shown to be proportional to the likelihood function of a suitably defined mixed Poisson model (Feng et al., 2005). The follow-up period is divided into as many intervals (say l intervals) as there are unique failure times. Each interval starts at a unique failure time and ends at the next unique failure time. So the interval length is sufficiently short to assume a constant baseline hazard $h_0(t)$ for each of these intervals:

$$h_0(t) = h_{0k}, t \in \Omega_k = (t_{k-1}, t_k],\ k = 1, \dots, l.$$

Let t_{ijk} be the total time of subject j within cluster i in Ω_k, and δ_{ijk} is the event indicator for subject j within cluster i in Ω_k, and $\delta_{ijk} = 1$ if the event happens in Ω_k and 0 otherwise. As shown in Feng et al. (2005), under an independent and non-informative censoring assumption for the interval Ω_k and assuming normal random effects, the likelihood function of the observed data in cluster i from a random-effects Cox model is proportional to the observed likelihood function from a random-effects Poisson model with log interval lengths between unique failure times as an offset. Thus the likelihood is defined as:

$$L_M(h_{01},\dots,h_{0l},\beta,\theta) \propto c|\Sigma(\theta)|^{-1/2}\int\prod_{k=1}^{l}\prod_{j=1}^{n_i}\left[h_{0k}t_{ijk}\exp\left(x_{ij}'\beta+z_{ij}'b_i\right)\right]^{\delta_{ijk}}$$
$$\times\exp\left[-h_{0k}t_{ijk}\exp\left(x_{ij}'\beta+z_{ij}'b_i\right)\right]\exp\left[-\tfrac{1}{2}b_i'\Sigma(\theta)^{-1}b_i\right]db_i \qquad (13)$$

Using a Laplace approximation to approximate the high-dimensional integral in (13), the likelihood for the entire data set is then

$$L(h_{01},\dots,h_{0l},\beta,\theta)=\prod_{i=1}^{m}L_i^{Lap}(h_{01},\dots,h_{0l},\beta,\theta) \qquad (14)$$

3.4. Bayesian

Bayesian techniques can be used to fit random-effects survival models with unspecified baseline hazard, where the cumulative baseline hazard is specified in terms of increments over particular intervals without knowing any information about the hazard function itself. These increments are assumed to be independent and to follow a gamma process. Similar to the Poisson modeling approach in Section 3.3, the follow-up time is divided into l intervals with the boundaries corresponding to the observed event time. The increase of the cumulative baseline hazard in the interval Ω_k is $h_{0k}=H_0(t_k)-H_0(t_{k-1})$. Assuming a normal distribution for random effects b_i, the likelihood function for all clusters is

$$c|\Sigma(\theta)|^{-1/2}\prod_{i=1}^{m}L_i(h_{01},\dots,h_{0l},\beta,b_i)\exp\left(-\frac{1}{2}b_i'\Sigma(\theta)^{-1}b_i\right) \qquad (15)$$

where

$$L_i(h_{01},\dots,h_{0l},\beta,b_i)$$
$$=\prod_{k=1}^{l}\prod_{j=1}^{n_i}\left[h_{0k}t_{ijk}\exp\left(x_{ij}'\beta+z_{ij}'b_i\right)\right]^{\delta_{ijk}}\exp\left[-h_{0k}t_{ijk}\exp\left(x_{ij}'\beta+z_{ij}'b_i\right)\right] \qquad (16)$$

is the conditional likelihood for the observed data in the i^{th} cluster. The δ_{ijk} takes value 1 if failure occurs and 0 if otherwise within the k^{th} time interval and t_{ijk} is the follow-up time in that interval. The joint posterior distribution is

$$c|\Sigma(\theta)|^{-1/2}\prod_{i=1}^{m}L_i(h_{01},\dots,h_{0l},\beta,b_i)\exp\left(-\frac{1}{2}b_i'\Sigma(\theta)^{-1}b_i\right)\times\pi(\beta)\times\pi(h_{0k})\times\pi(\Sigma) \qquad (17)$$

where $\pi(.)$ indicates prior distribution.

Prior distributions used are generally non-informative: $N(0,\sigma^2)$ for β; the multivariate normal distribution $N(0,\Sigma)$ for random effects $b_i|\Sigma$ with aWishart distribution $W(\alpha,V)$ for the inverse covariance matrix Σ^{-1} with α degrees of freedom and a diagonal variance-covariance matrix V; the increments of the baseline hazard are assumed gamma distributed with parameters $r[H_0^*(t_k)-H_0^*(t_{k-1})]$ and r. The $H_0^*(.)$ is specified based on a constant hazard h_0^* and thus $H_0^*(t_k) = h_0^* t_k$ (Duchateau et al., 2008), and r representing the degree of confidence in this prior.

4. SIMULATION STUDY

We study the performance of the four estimation methods described in Section 3 for a Cox model with a random cluster effect and a random treatment-by-cluster interaction by using a simulation study. The simulation structure was built based on a real data set from a veterinary science field. The data set has a two-level of hierarchy; the upper level is a herd and the lower level is an animal. Only one independent variable was considered: a treatment procedure applied at the lower level. A previous analysis of the data set gave the following estimates: $-0.72, 2.26, 0.49$ and 0.8 for β, σ_0^2, σ_1^2 and ρ, respectively.

4.1. Simulation of Data

Model parameters were chosen to mimic the real data set. A total of 3556 subjects nested in 22 clusters were considered. The true values of the parameters β, σ_0^2, σ_1^2 and ρ were set to be -0.8, $2.0, 0.5$ and 0.8, respectively. An exponential baseline hazard was used and assumed to be equal to 0.002. Using the technique of Bender et al. (2005) for generating survival times from a Cox model, 300 data sets were generated in R version 2.12.1 from model (3). The two random effects b_{i0} and b_{i1} were generated from a zero-mean multivariate normal distribution. The fixed effect variable was generated from a Bernoulli distribution with probability 0.26. The event time for each individual was randomly generated from an exponential distribution with a parameter equal to the hazard function defined in (3). The censoring time was generated from a uniform distribution $U(0, \gamma)$, where γ was determined based on the amount of censoring. The survival time is equal to the minimum of event time and censoring time. These choices of simulation parameters resulting in approximately 89% censored observations. The point estimates and their standard errors were extracted for each simulated data set. The average of the parameter estimates and the average of their standard errors and absolute bias across the 300 data sets, probability coverage, and the empirical standard deviation were computed. The significance of bias was assessed by a t-test at the 5% significance level.

4.2. Software implementation

The simulated data sets were analyzed using the four estimation methods described in Section 3. The analyses based on MPPL, MPL and PML were done by the R-packages ‘coxme’, ‘frailtypack’ and ‘lme4’ implemented in R v2.12.1, respectively. In the MPL analysis, the smoothing parameter and number of knots were set to 10,000 and 8, respectively. For PML analysis, each simulated data sets is split at failure times using the R-package ‘Epi’, and then a mixed-effects Poisson model with a 4-order smooth function of time is applied to the split data set. Finally, the Bayesian analysis used MCMC estimation. The model was implemented in WinBUGS v1.4 and called from within R with non-informative prior distributions. Markov chains were run for the simulated data sets 1000 burn-in samples and estimates as posterior medians based on 1000 samples. Markov chain diagnostics were carried out for a sample of simulated data sets and found to be satisfactory.

4.3. Simulation results

The simulation results are presented in Table 1. The average of estimated values was reported based only on the datasets when convergence was reached. In other words, the analyses for certain data sets or approaches were excluded if non-convergence or non-sensible estimates occurred.

Table 1: Simulation study results. Mean of the estimate, empirical standard deviation, mean of the model-based standard error and the average of the absolute value of bias over 300 simulated data sets

Method		β	σ_0^2	σ_1^2	ρ	95% CI	
	True value	-0.8	2.0	0.5	0.8	Coverage	Convergence
MPPL	**Estimate**	-0.671	1.758	0.419	0.744	72%	100%
	Emp. Std	0.347	0.632	0.333	0.402		
	Model Se	0.216	--[a]	--[a]	--[a]		
	Abs. Bias	0.129*	0.242*	0.081*	0.056*		
MPL	**Estimate**	-0.918	1.334	0.530	0.589	93%	19%
	Emp. Std	0.256	0.554	0.452	0.466		
	Model Se	0.406	0.515	0.472	0.387		
	Abs. Bias	0.118*	0.666*	0.030	0.211*		
PML	**Estimate**	-0.801	1.780	0.512	0.740	79%	100%
	Emp. Std	0.363	0.689	0.403	0.496		
	Model Se	0.228	--[a]	--[a]	--[a]		
	Abs. Bias	0.001	0.220*	0.012	0.060*		
Bayesian	**Estimate**	-0.823	2.006	0.746	0.709	95%	100%
	Emp. Std	0.352	0.757	0.335	0.450		
	Model Se	0.378	0.904	0.691	0.619		
	Abs. Bias	0.023	0.006	0.246*	0.091*		

[a] No available standard error.

* Significant bias ($P < 0.05$).

Only the MPL procedure experienced convergence problems and failed to reach convergence for most of the simulated data sets. The fixed effect parameter β was estimated well by PML and Bayesian models with relative absolute biases of 0.13% for PML and 2.9% for Bayesian analysis. On the other hand, MPPL and MPL methods produced estimates with a higher relative absolute bias; 16.1% for MPPL method and 14.8% for MPL method (however, MPL estimates were computed based on a small number of simulated data sets). MPPL and PML methods produced on average underestimated standard errors for β (confidence interval converges of 72% and 79%, respectively) and the MPL method overestimated it (with confidence interval converge of 93%), while the Bayesian approach gave on average a standard error close to the empirical standard deviation (with confidence interval converge of 95%). The Bayesian model produced a good estimate for random cluster effects variance σ_0^2, and overestimated the random treatment variance σ_1^2 with a significant bias. Other methods gave estimates for random cluster effects variance σ_0^2 with strongly significant downward bias. The MPPL method underestimated the variance σ_1^2, and PML and MPL yielded estimates with non-significant bias. The correlation parameter ρ was somewhat underestimated by all methods. Overall, and despite the biased estimates of σ_1^2, the performance of the Bayesian model seemed to be the best among the four methods.

5. CONCLUSIONS

In this paper we reviewed four common estimation methods to fit a Cox model with random cluster and random treatment effects. We have compared these methods through a simulation study based on a Cox model with treatment fixed effects, random cluster effects and random treatment effects assuming correlation between the two random effects. The structure of the simulations was designed to mirror a real data set structure. The simulation study showed significant biases of all model parameters by the MPPL method. The MPL method was computationally problematic and behaved poorly in estimating the standard errors. The PML method is more flexible than the others; however this flexibility comes at the price of having to split the data set to a huge number of records resulting in potentially very large data sets and thus compromising computing time. The method reported significant bias in estimating the parameters of random cluster effects. In contrast, Bayesian model is somewhat slow and gave computationally intensive, and it produced estimates with a significant bias for random treatment effects but it gave good probability coverage for fixed effect estimates and a good estimate for random herd effects.

Overall it is hard to formulate a clear recommendation, because the performance of the estimation methods might be affected by many factors, such as sample size, amount of censoring, variance of random effects and these factors have to be taken into account. This study can give some guidance within its boundaries in this respect.

REFERENCES

1. Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*. 24, 1712-1723.
2. Duchateau, L. and Janssen, P. (2008). *The frailty model*. New York: Springer.
3. Feng, S., Wolfe, R. and Port, F. (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using Poisson variance structures. *J. Amer. Statist. Assoc.*, 100, 718-735.
4. Gray, R.J. (1992). Flexible methods for analyzing survival data using splines with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.*, 87, 942-951.
5. Joly, P., Commenges, D. and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*. 54, 185-194.
6. Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004). GLLAMM Manual. U.C. Berkeley Division of Biostatistics working paper series, working paper 160.
7. Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*. 56, 1016-1022.
8. Rondeau, V. and Gonzalez, J. (2003). Maximum penalized likelihood estimation in a Gamma-frailty model. *Lifetime Data Analysis.* 9, 139-153.
9. Rondeau, V., Filleul, L. and Joly, P. (2006). Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*, 25, 4036-4052.
10. Rondeau, V., Michiels, S., Liquet, B. and Pignon, J. (2008). Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Statistics in Medicine.* 27, 1894-1910.
11. Yamaguchi, T., Ohashi, Y. and Matsuyama, Y. (2002). Proportional hazards models with random effects to examine centre effects in multicentre cancer clinical trials. *Statistical Methods in Medical Research.* 11, 221-236.

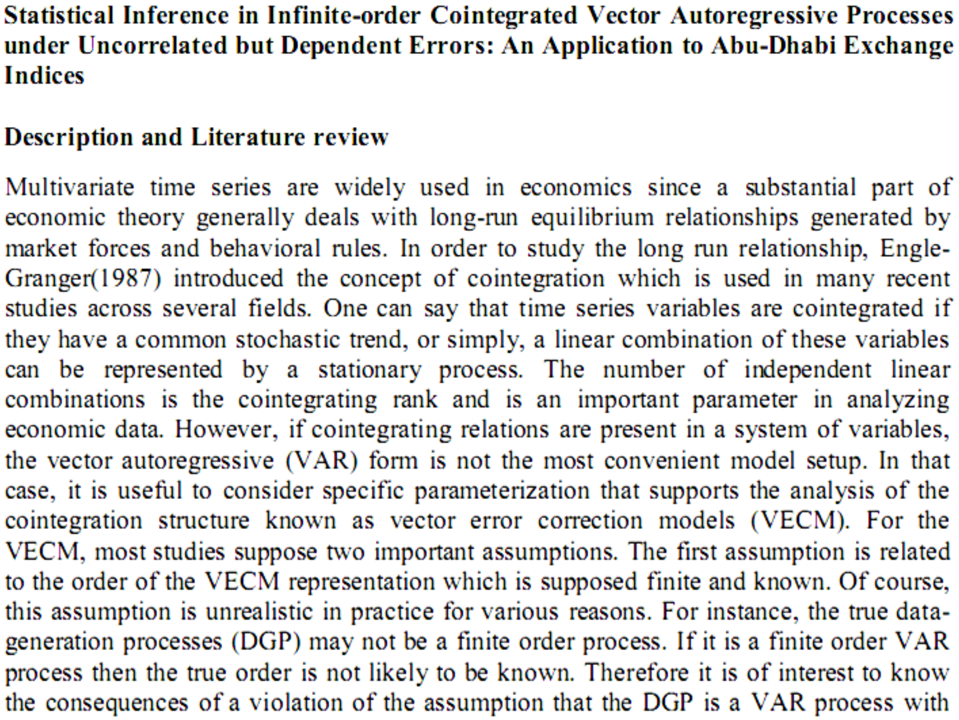
Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 375-380

STATISTICAL INFERENCE IN INFINITE-ORDER COINTEGRATED VECTOR AUTOREGRESSIVE PROCESSES UNDER UNCORRELATED BUT DEPENDENT ERRORS

Chafik Bouhaddioui
United Arab Emirates University, Al-Ain, UAE
Email: chafikb@uaeu.ac.ae

ABSTRACT

The concept of cointegration processes is one of the most used concepts in economics and finance. Mainly, researchers are interested in behavior of the estimators of the model parameters. In this project, we will investigate the asymptotic behavior of the estimators of an infinite-order cointegrated vector autoregressive series under non-independent errors by showing its asymptotic distribution. Using this result, we will construct a Likelihood Ratio (LR) test of the cointegration rank. One can also develop a method under unrestrictive assumptions to select the autoregressive order. Monte Carlo experiments illustrate the finite sample performance of the LR test.



Statistical Inference in Infinite-order Cointegrated Vector Autoregressive Processes under Uncorrelated but Dependent Errors: An Application to Abu-Dhabi Exchange Indices

Description and Literature review

Multivariate time series are widely used in economics since a substantial part of economic theory generally deals with long-run equilibrium relationships generated by market forces and behavioral rules. In order to study the long run relationship, Engle-Granger(1987) introduced the concept of cointegration which is used in many recent studies across several fields. One can say that time series variables are cointegrated if they have a common stochastic trend, or simply, a linear combination of these variables can be represented by a stationary process. The number of independent linear combinations is the cointegrating rank and is an important parameter in analyzing economic data. However, if cointegrating relations are present in a system of variables, the vector autoregressive (VAR) form is not the most convenient model setup. In that case, it is useful to consider specific parameterization that supports the analysis of the cointegration structure known as vector error correction models (VECM). For the VECM, most studies suppose two important assumptions. The first assumption is related to the order of the VECM representation which is supposed finite and known. Of course, this assumption is unrealistic in practice for various reasons. For instance, the true data-generation processes (DGP) may not be a finite order process. If it is a finite order VAR process then the true order is not likely to be known. Therefore it is of interest to know the consequences of a violation of the assumption that the DGP is a VAR process with known finite order. The second condition is related to the innovations process which are supposed to be independent and identically distributed (*i.i.d*). This assumption is too restrictive when economic or financial data is to study. Most of the macroeconomic time series exhibit a conditional heteroscedasticity or any nonlinear form.

In this project, under the more general model which is represented by an infinite-order cointegrated process with uncorrelated but dependent errors, denoted by WIVAR(∞), we will address three main aims. The First aim is to study the behavior of the estimators of the cointegration. The difficulty of this problem comes from the fact that we have to consider the approximation of the infinite-order cointegrated process by a finite-order autoregressive process where the order of the fitted autoregression is a function of the sample size and the fact that the errors are uncorrelated but dependent which needs the use of a more general result than the central limit theorem which assume that the errors are independent. In the literature, studying the behavior of the parameter estimators of an IVAR(∞) with *i.i.d* errors was done by Saikkonen (1992) and Saikkonen and Lutkepohl (1996). They showed the asymptotic properties of the estimated coefficients of the autoregressive error correction model (VECM) and the pure vector autoregressive (VAR) representations derived under the assumption that the autoregressive order goes to infinity with the sample size. Under the same strong assumptions on the innovations process, Saikkonen and Lutkepohl (1996) constructed a test for linear (zero) restrictions which arise in exogeneity or Granger causality analyses.

The second aim is to construct a likelihood ratio (LR) tests of the cointegration rank where the process is WIVAR(∞) . The problem is more complicate than the regular LR tests proposed by Johanson (1988, 1991). Lutkepohl and Saikkonen (1999) extended the LR tests proposed by Ng and Perron (1995) to the case of the IVAR(∞) while Raissi (2009) proposed an extension of the LR tests where the order of the VECM is finite and errors are uncorrelated but dependent.

The third aim is to study the order selection model. Lutkepohl and Saikkonen (1999) proposed two ways for specifying the VAR order. The first possibility is to use a deterministic rule for choosing the order and the second one is to assume that the order is chosen by some data-dependent rule. The main problem with these two approaches is that they assume that the errors are (*i.i.d*).

Challenges and outcomes

In the three parts of this project the challenge comes from the study of more general cointegrated process under weak assumption. This task is complicated because the order of the VECM is infinite and needs to be set as a function of the sample size and the errors are not supposed to be *i.i.d*. This introduces at least two complications; the first comes from the fact that the order has to be supposed that it goes to infinity when the sample size goes to infinity too. The asymptotic distribution of the estimators will be independent of the autoregressive order. The asymptotic properties of the estimators will be derived and discussed. The second complication will come from the fact that the errors will be assumed uncorrelated but dependent. Under some mixing condition on the error process, one can propose a valid LR test statistic and its asymptotic distribution. This will produce excellent quality papers but more importantly it can be used by practitioner in checking their model adequacy. Applications to Abu-Dhabi exchange indices will be considered. This will make the research paper complete and makes the research more attractable to practitioner.

Research significance

In fact any result which gives practitioner a valid model under unrestrictive assumptions is quite welcomed. And any solid research in this direction will receive good review and will be a welcomed publication. Also, it will open the door for more sophisticated model in analyzing economic and financial data.

Researchers' prior experience in the field

Bouhaddioui have been working the multivariate time series and specifically the infinite-order autoregressive processes under different conditions. Jointly with Roy from University of Montreal we developed the asymptotic properties of the cross-correlations vectors between two infinite-order autoregressive stationary processes with *i.i.d* errors, see Bouhaddioui and Roy (2006a). On the same way, we developed a generalized Portmanteau test of independence between two infinite-order autoregressive processes in Bouhaddioui and Roy (2006b). In the context of infinite-order cointegrated processes, I develop a test of non-correlation using the cross-correlation and partial cross-correlations matrices, see Bouhaddioui and Dufour (2008). A test of independence and causality was proposed for the same processes in Bouhaddioui and Dufour (2010). Bouhaddioui and Dufour (2009) proposed a generalized causality test in multiple horizons for infinite-order autoregressive processes. In all these papers, reviewers suggest that if we consider the case of dependent errors, this will be an important contribution to the econometric/time series literature and to practitioners. Prof Dufour has a substantial work in econometrics/time series and finance. We collaborated in different articles related to the cointegration analysis. Also, his expertise in these domains will be an asset when dealing with the heteroscedasticity of the errors (ARCH/GARCH models). He published tens of papers in high qualified journals in econometrics and statistics (Econometrica, JASA,..). Prof Kilani, his expertise in time series analysis and nonparametric methods will be an asset when dealing with the bootstrap part and simulations, see Ghoudi and al.(2007) and Ghoudi and Remillard (2009).

Methodology

This project uses a central limit theorem under weak assumptions and properties of the quasi-likelihood estimation. To establish the result one has first to write the VECM process in a way to obtain the quasi-likelihood estimators of the cointegration parameters. By showing the consistency of these estimators, one can show the asymptotic distribution of these estimators by using an appropriate central limit theorem. This requires lots of technicalities and heavy use of the model structure. Though the proofs are similar in structure of the IVAR(∞) with *i.i.d* errors the details when the errors have a dependent structure are completely different. Also, we will show that the test statistic of the LR test has a different distribution than the case of *i.i.d* Gaussian case. By proposing a consistent estimator to the variance of the errors, we can propose two ways to identify the autoregressive order. A bootstrap method for choosing the order will be also proposed and compared to the other two methods.

Likelihood of success

Given our prior experiences in this field of the research, the project is quite likely to be successful. In fact we will start by consider the more general IVAR(∞) model and its VECM representation. By using the maximum likelihood estimators of the cointegration parameters, we already show the consistency of these estimators. Then we will tackle the asymptotic distribution of these estimators or function of these estimators. Professor Dufour will visit our college and department during this semester and we will be mainly working on showing the asymptotic distribution of the QL estimators.

References:

- Bouhaddioui, C. & Dufour, J. (2010). Semiparametric innovation-based tests of orthogonality and causality between two infinite-order cointegrated ceries with application to Canada/US monetary interactions, Journal of Business and Economics Statistics, Under revision.
- Bouhaddioui, C. & Dufour, J. (2009). Tests of causality between two infinite-order vector autoregressive series. In JSM Proceedings, Business and Economic Statistics Section. Alexandria, VA: American Statistical Association. 1661-1669
- Bouhaddioui, C. & Dufour, J. (2008). Tests for non-correlation of two infinite order cointegrated vector autoregressive series. Journal of Applied Probability and Statistics, 3 (1), 77-94.
- Bouhaddioui, C. & Roy, R. (2006b). A generalized portmanteau test for independence of two infinite order vector autoregressive series. Journal of Time Series Analysis, **27** (4), 505-544.
- Bouhaddioui, C. & Roy, R. (2006a). On the distribution of residual cross-correlations of infinite order vector autoregressive series and applications. Statistics and Probability Letters, **76** (1), 58-68.
- Engle, R.F. & Granger, C.W.J. 1987. Co-integration and error correction: representation, estimation, and testing. *Econometrica,* 55:251–276.
- Johansen, S. (1988) Statistical analysis of cointegration vectors. Journal of Economic Dynamic and Control, 12, 231–254.
- Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica , 59, 1551–1581.
- Lutkepohl, H. & Saikkonen, P. (1999) Order selection in testing for the cointegrating rank of a VAR process. In Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive W.J. Granger, Engle, R.F., White, H., Eds.; Oxford University Press: Oxford, 168–199.
- Ng, S. & Perron, P. (1995) Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. Journal of the American Statistical Association, 90, 268–281.
- Raïssi, H. (2009) Testing the Cointegrating Rank with Uncorrelated but Dependent Errors,Stochastic Analysis and Applications, 27 (1), 24–50

- Saikkonen, P. (1992) Estimation and testing of cointegrated systems by an autoregressive approximation. Econometric Theory, 8, 1–27.
- Saikkonen, P. & L¨utkepohl, H. (1996), 'Infinite-order cointegrated vector autoregressive processes: Estimation and inference', *Econometric Theory* **12**, 814–844.
- Genest, C., Ghoudi, K., Remillard B. (2007). Rank based extensions of the Brock, Dechert, and Scheinkman test. *Journal of the American Statistical Association,* **102**, 1363-1376.
- Ghoudi, K. and Remillard, B. (2009). Empirical distribution functions and copulas based on residuals of ARMA models (Submitted).

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 381-394

MANY HYPOTHESES TESTING WITH POSSIBILITY OF REJECTION OF DECISION FOR THE PAIR OF FAMILIES OF PROBABILITY DISTRIBUTIONS

Farshin Hormozi nejad
Islamic Azad University, Ahvaz Branch, Iran
Email: hormozi-nejad@iauahvaz.ac.ir

ABSTRACT

In this paper, it is considered multiple statistical hypotheses testing with possibility of rejecting to make choice between hypotheses concerning the pair of families of probability distributions in a pair of stages, such that in the first stage one family of distributions must be distinguished and then in the second stage, the object's distribution must be denoted between mentioned family of probability distributions. It is investigated description of characteristics of logarithmically asymptotically optimal (LAO) hypotheses testing with permission of decision rejection. The matrix of optimal interdependencies of all pairs of the error probability exponents (reliabilities) is studied. The goal of research is to express the functional relation between the pair of error probabilities exponents of LAO hypothesis testing by a pair of stages and to compare with the case of similar one-stage testing.

KEYWORDS

Logarithmically asymptotically optimal test, multiple hypotheses testing, testing with rejection of decision, two-stage test, reliabilities matrix.

1. INTRODUCTION

In some works results of probability theory and statistics were obtained with application of information-theoretical methods and there are studies where statistical results provide ground for new findings in information theory [3], [5], [6], [8], [9], [13], [20]. The classical problem of statistical hypothesis testing refers to two hypotheses. Hoeffding [17], Tusnady [20], Csiszár and Longo [7] dealt with the error exponent for testing simple statistical hypotheses. Blahut [5] reviewed relationship between hypotheses testing and information theory. We call the exponent of error probability the reliability. In case of two hypotheses both reliabilities corresponding to two possible error probabilities could not be increased simultaneously, it is an accepted way to fix the value of one of the reliabilities and try to make the tests sequence get the greatest value of the remaining reliability. Such a test is called logarithmically asymptotically optimal (LAO). The term LAO for testing of two hypotheses was proposed by Birge' [4]. The need of testing of more than two hypotheses in many scientific and applied fields has essentially increased recently. Ahlswede et al [1, 2] and Haroutunian [11] formulated some problems of multiple hypotheses testing and identification. Haroutunian [10] and Haroutunian et al [12] investigated the problem of LAO testing of multiple statistical hypotheses. Haroutunian and Hakobyan [13] solved the problem of multiple hypotheses LAO testing for many independent objects. The model of the two-stage LAO testing in multiple hypotheses for a pair of families of distributions is investigated in [14, 18, 19]. The problem of multiple hypotheses LAO testing with rejection of decision is studied in [15, 16].

In this paper the problem of two-stage LAO multihypotheses testing for a model with possibility of rejecting to make choice between hypotheses consisting of many disjoint families of probability distributions (PDs) is studied and a solution of that problem is exposed. The matrices of optimal asymptotic interdependencies of all pairs of the error probability exponents are studied.

2. PRELIMINARIES

Random variable (RV) X characterizing the studied object takes values in the finite set $\mathcal{X}$ and $\mathcal{P}(\mathcal{X})$ is the space of all distributions on X. S hypothetical probability distributions (PDs) of X are given that divided in two disjoint families of distributions. The first family includes R hypotheses $P_1, P_2, \ldots, P_R$ and the second family consists of $S-R$ hypotheses $P_{R+1}, P_{R+2}, \ldots, P_S$.

Let N-sample $\mathrm{x} = (x_1, x_2, \ldots, x_N)$, be a vector of results of N independent observations of the RV X. The purpose of the test is using sample $\mathbf{x}$ to detect the actual distribution from given list.

The entropy of RV X with PD Q and the divergence (Kullback-Leibler distance) of PDs Q and P, are defined [6, 8, 10] as follows:

$$\mathrm{H}_Q(X) \overset{\Delta}{=} -\sum_{x\in\mathcal{X}} Q(x)\log Q(x),$$

$$\mathrm{D}(Q \parallel P) \overset{\Delta}{=} \sum_{x\in\mathcal{X}} Q(x)\log\frac{Q(x)}{P(x)}.$$

The method of types is a base of our proofs, so we here remind some definitions and estimates [6]. Let $N(x|\mathrm{x})$ be the number of repetitions of the element $x \in \mathcal{X}$ in the vector $\mathrm{x} \in \mathcal{X}^N$, and

$$Q_{\mathrm{x}}(x) \overset{\Delta}{=} \{N(x|\mathrm{x})/N \;,\quad x \in \mathcal{X}\},$$

is the PD, called in information theory *the type* of $\mathbf{x}$ [6]. Let $\mathcal{P}^N(\mathcal{X})$ be the set of all possible types on $\mathcal{X}^N$ for N observations and suppose T_Q^N be the set of all vectors $\mathbf{x}$ of the type $\mathcal{P}^N(\mathcal{X})$.

3. THE TWO-STAGE LAO TESTING WITH REJECTION OF DECISION BY ONE SAMPLE

We denote the two-stage test on the base of N-sample by Φ_1^N. Such test may be realized by a pair of tests φ_1^N and φ_2^N for two consecutive stages and we write $\Phi_1^N = (\varphi_1^N, \varphi_2^N)$. The first stage is for choice of a family of PDs, it is executed by a non-randomized test $\varphi_1^N(\mathrm{x})$ using sample $\mathbf{x}$. The next stage is for making decision in the determined family of PDs, it is shown by a non-randomized test $\varphi_2^N(\mathrm{x})$ based on the sample $\mathbf{x}$ and on the result of the test φ_1^N.

First stage of two-stage test with rejection of decision by one sample is as follows.

Let us introduce two sets of indices $D_1 = \{\overline{1, R}\}$ and $D_2 = \{\overline{R+1, S}\}$ and a pair of disjoint families of PDs $\mathcal{P}_1$ and $\mathcal{P}_2$

$$\mathcal{P}_1 = \{P_s, \quad s \in D_1\}, \qquad \mathcal{P}_2 = \{P_s, \quad s \in D_2\}.$$

The first stage of decision making consists in using sample $\mathbf{x}$ for selection of one family or rejecting the decision is denoted by a test $\varphi_1^N(\mathrm{x})$. It can be defined by division of the sample space $\mathcal{X}^N$ on three disjoint subsets

$$A_i^N \overset{\Delta}{=} \{\mathrm{x}: \varphi_1^N(\mathrm{x}) = i\}, \quad i = \overline{1,3}, \qquad \bigcup_{i=1}^{3} A_i^N = \mathcal{X}^N.$$

The set $A_i^N,\ i = 1,2$, consists all vectors $\mathbf{x}$ for which i-th family of PDs is adopted and A_3^N is all vectors $\mathbf{x}$ for rejecting of the pair of family of PDs.

Let $\alpha'_{i|j}(\varphi_1^N),\ i \neq j,\ i,j = 1,2$, be the probability of the erroneous acceptance of the i-th family of PDs provided that the j-th family of PDs is true (that is the correct PD is in the j-th family). we define

$$\alpha'_{i|j}(\varphi_1^N) \overset{\Delta}{=} \max_{s \in D_j} P_s^N(A_i^N), \quad i \neq j,\ i,j = 1,2. \tag{1}$$

And let $\alpha'_{3|j}(\varphi_1^N),\ j = 1,2$, be the probability of the erroneous rejection of the pair of families of PDs provided that the j-th family of PDs is true

$$\alpha'_{3|j}(\varphi_1^N) \overset{\Delta}{=} \max_{s \in D_j} P_s^N(A_3^N), \quad j = 1,2. \tag{2}$$

Let $\alpha'_{j|j}(\varphi_1^N)$ be the probability to reject the j-th family of PDs when it is right,

$$\alpha'_{j|j}(\varphi_1^N) \overset{\Delta}{=} \max_{s \in D_j} P_s^N(\overline{A}_j^N), \quad j = 1,2. \tag{3}$$

We consider reliabilities of the infinite sequence of tests φ_1:

$$E'_{i|j}(\varphi_1) \overset{\Delta}{=} \limsup_{N \to \infty} \{-\frac{1}{N} \log \alpha'_{i|j}(\varphi_1^N)\}, \quad i = \overline{1,3},\ j = \overline{1,2}. \tag{4}$$

The reliabilities matrix for the first stage of the test is $\mathbf{E}'(\varphi_1)$ and one can see from (1)-(4) that

$$E'_{j|j} = \min_{i \neq j} E'_{i|j}, \quad i = \overline{1,3},\ j = \overline{1,2}.$$

For construction of the necessary LAO test for preliminarily given positive values $E'_{j|j},\ j = \overline{1,2}$, we define the following subsets of distributions:

$$R'_j \overset{\Delta}{=} \{Q: \min_{s \in D_j} \mathrm{D}(Q||P_s) \leq E'^{*}_{j|j}\}, \quad j = \overline{1,2}, \tag{5}$$

$$R'_3 \overset{\Delta}{=} \{Q: \min_{s\in D_j} \mathrm{D}(Q||P_s) > E'^{*}_{j|j}, \quad j=\overline{1,2}\}, \tag{6}$$

$$E'^{*}_{j|j} \overset{\Delta}{=} E'_{j|j}, \quad j=\overline{1,2}, \tag{7}$$

$$E'^{*}_{i|j} \overset{\Delta}{=} \min_{s\in D_j} \inf_{Q\in R'_i} D(Q||P_s), \;\; i=\overline{1,3}, \;\; j=\overline{1,2}, \;\; i \not\asymp j, \tag{8}$$

Theorem 1. If all distributions P_s, $s=\overline{1,S}$, are different and the positive values $E'^{*}_{j|j}$, $j=\overline{1,2}$, are such that the following inequalities hold

$$E'^{*}_{1|1} < \min_{l\in D_2, s\in D_1} D(P_l||P_s), \qquad E'^{*}_{2|2} < E'^{*}_{1|2}, \tag{9}$$

then there exists a LAO sequence of tests, all elements of the reliabilities matrix $E'(\varphi_1^*) = \{E'^{*}_{i|j}\}$ of which are positive and are defined in (7) – (8) .

When one of the inequalities (9) is violated, then at least one element of the matrix $E'(\varphi_1^*)$ is equal to 0.

The second stage of the two-stage test with rejection of decision by one sample is in coming:

If the first family of PDs is accepted, then we consider test $\varphi_2^N(\mathrm{x})$ which can be defined by division of the sample space A_1^{*N} to $R+1$ distinct subsets

$$B_s^N \overset{\Delta}{=} \{\mathrm{x}: \varphi_2^N(\mathrm{x}) = s\}, \quad s=\overline{1,R+1}.$$

The set B_s^N, $s=\overline{1,R}$, consists all vectors x for which s-th PD is adopted and B_{R+1}^N is all vectors x for rejecting of the first family of PDs.

Let $\alpha''_{l|s}(\varphi_2^N)$ be the probability of the erroneous acceptance at the second stage of test, in which PD P_l is admitted when P_s is true

$$\alpha''_{l|s}(\varphi_2^N) \overset{\Delta}{=} P_s^N(B_l^N), \qquad l\in D_1, s=\overline{1,S}, l\neq s.$$

When decision is rejected, but PD P_s is true, we consider the following probability of error:

$$\alpha''_{R+1|s}(\varphi_2^N) \overset{\Delta}{=} P_s^N(B_{R+1}^N), \qquad s=\overline{1,S}.$$

The probability to reject P_s, when it is true, is

$$\alpha''_{s|s}(\varphi_2^N) \overset{\Delta}{=} P_s^N\left(\overline{B}_s^N\right) = \sum_{l\neq s, l=1}^{R+1} \alpha''_{l|s}(\varphi_2^N) + P_s(\overline{A}_1^{*N}), \quad s\in D_1, \tag{10}$$

Corresponding reliabilities for the second stage of test, are defined as

$$E''_{l|s}(\varphi_2) \overset{\Delta}{=} \limsup_{N\to\infty}\{-\frac{1}{N}\log\alpha''_{l|s}(\varphi_2^N)\}, \quad l=\overline{1,R+1}, \; s=\overline{1,S}. \tag{11}$$

Using properties of types we introduce the following definition

$$\lim_{N\to\infty}\{-\frac{1}{N}\log P_s^N(A_i^{*N})\} = \inf_{Q:\min_{l\in D_i} D(Q||P_l)\leq E'^{*}_{i|i}} D(Q||P_s) \overset{\Delta}{=} E^{I}_{i|s}, \quad s\notin D_i. \tag{12}$$

From (10)–(12) it follows that

$$E''_{s|s}(\varphi_2) = \min[\min_{l\neq s} E''_{l|s}(\varphi_2), \min_{i=2,3} E^I_{i|s}], \quad s \in D_1.$$

Theorem 2. If at the first stage of test the first family of PDs is accepted, then for given positive values $E''_{s|s}$, $s=\overline{1,R}$ of the reliabilities matrix $\mathrm{E}''(\varphi_2)$ let us consider the regions:

$$R''_s = \{Q: \min_{l\in D_1} \mathrm{D}(Q||P_l) \le E'^*_{1|1}, \quad \mathrm{D}(Q \parallel P_s) \le E''_{s|s}\}, \quad s=\overline{1,R},$$

$$R''_{R+1} = \{Q: \min_{l\in D_1} \mathrm{D}(Q||P_l) \le E'^*_{1|1}, \quad \mathrm{D}(Q \parallel P_s) > E''_{s|s}, \quad s=\overline{1,R}\},$$

and the following values of elements of the future reliabilities matrix $\mathrm{E}''(\varphi_2^*)$ of the LAO test sequence:

$$E''^*_{s|s} = E''_{s|s}, \quad s=\overline{1,R},$$

$$E''^*_{l|s} = \inf_{Q\in R''_l} \mathrm{D}(Q \parallel P_s), \quad l=\overline{1,R+1}, s=\overline{1,S}, l\neq s,$$

If the following compatibility conditions are valid

$$E''_{1|1} < min[\min_{s=\overline{2,R}} \mathrm{D}(P_s \parallel P_1), \qquad \min_{i=2,3} E^I_{i|1}],$$

$$E''_{s|s} < min[\min_{l=\overline{1,s-1}} E''^*_{l|s}, \min_{l=\overline{s+1,R}} \mathrm{D}(P_l \parallel P_s), \min_{i=2,3} E^I_{i|s}], \quad 2\le s\le R-1,$$

$$E''_{R|R} < min[\min_{l=\overline{1,R-1}} E''^*_{l|R}, \min_{i=2,3} E^I_{i|R}],$$

then there exists a LAO sequence of tests φ_2^*, elements $E''^*_{l|s}$ of reliabilities matrix $E''(\varphi_2^*)$ of which are defined above and are positive.

If one of the compatibility conditions is violated, then at least one element of the matrix $E''(\varphi_2^*)$ is equal to 0.

When the second family of PDs is accepted, then the test $\varphi_2^N(\mathrm{x})$ is realized by division of the sample space A_2^{*N} to $S-R+1$ distinct subsets

$$B_s^N \overset{\Delta}{=} \{\mathrm{x}: \varphi_2^N(\mathrm{x}) = s\}, \quad s=\overline{R+1,S+1}.$$

In this case the definitions of error probabilities and reliabilities for the second stage of test is similar to mentioned definitions of Section 3. So if in the first stage of test, the second family of PDs is accepted then Theorem 2 with replacing $s=\overline{R+1,S}$ will be used.

Which is the best value of $E'^*_{j|j}$, $j=1{,}2$, giving the best value to reliabilities $E''^*_{l|s}$? The answer to the question is in the following

Theorem 3. If distributions P_s, $s=\overline{1,S}$, are different then for given positive diagonal values $E''_{s|s}$, $s=\overline{1,S}$, of reliabilities matrix of the second stage, the bounds of reliabilities $E'^*_{j|j}$, $j=1{,}2$, of the first stage, satisfy the following conditions

$$\max_{s=\overline{1,R}} E"_{s|s} \leq E'^{*}_{1|1} \leq \min_{s\in D_2, l\in D_1} \mathrm{D}(P_s \parallel P_l), \tag{13}$$

$$\max_{s=\overline{R+1,S}} E"_{s|s} \leq E'^{*}_{2|2} \leq E'^{*}_{1|2}, \tag{14}$$

and the best value for them are equal to the lower bounds $E'^{*}_{1|1} = \max_{s=\overline{1,R}} E"_{s|s}$ and $E'^{*}_{2|2} = \max_{s=\overline{R+1,S}} E"_{s|s}$.

The reliabilities are investigated for the two-stage testing by one sample with rejection of decision as follows:

In two-stage decision making, if at the first stage of LAO test the i-th family of PDs is accepted, then the test Φ^{*N}_1 can be assigned by division of the sample space X^N to $S+1$ disjoint subsets as follows

$$C^N_s \overset{\Delta}{=} A^{*N}_i \cap B^N_s, \quad s \in D_i, i = 1,2, \qquad C^N_{S+1} \overset{\Delta}{=} (A^{*N}_1 \cap B^N_{R+1}) \cup (A^{*N}_2 \cap B^N_{S+1}) \cup A^{*N}_3.$$

The set C^N_s, $s = \overline{1,S}$ consists of all vectors $\mathbf{x}$ for which in the two-stage test s-th PD is adopted and C^N_{S+1} consists of all vectors $\mathbf{x}$ such that the two-stage test is rejected.

Let $\alpha'''_{l|s}$ be the probability of the false acceptance by two-stage test of PD P_l when P_s is true:

$$\alpha'''_{l|s}(\Phi^{*N}_1) \overset{\Delta}{=} P^N_s(C^N_l), \quad l = \overline{1,S}, \quad s = \overline{1,S}, \quad l \neq s.$$

When decision is rejected, but PD P_s is true, we consider the following probability of error:

$$\alpha'''_{S+1|s}(\Phi^{*N}_1) \overset{\Delta}{=} P^N_s(C^N_{S+1}), \qquad s = \overline{1,S}.$$

And the probability to reject P_s, when it is right, is

$$\alpha'''_{s|s}(\Phi^{*N}_1) \overset{\Delta}{=} P^N_s\left(\overline{C}^N_s\right), \quad s = \overline{1,S}.$$

We denote by $\Phi^*_1 = (\varphi^*_1, \varphi^*_2)$ the infinite sequences of tests and define reliabilities:

$$E'''_{l|s}(\Phi^*_1) \overset{\Delta}{=} \limsup_{N\to\infty}\{-\frac{1}{N}\log\alpha'''_{l|s}(\Phi^{*N}_1)\}, \quad l = \overline{1,S+1}, \quad s = \overline{1,S}. \tag{15}$$

These are the relationships between error probabilities and reliabilities for the two-stage test by one sample and the first and the second stages of LAO tests:

a) if $l \in D_i$, $s = \overline{1,S}$, then

$$\alpha'''_{l|s}(\Phi^{*N}_1) = P^N_s(A^{*N}_i \cap B^N_l) = P^N_s(B^N_l) = \alpha''_{l|s}(\varphi^{*N}_2) \tag{16}$$

b) for $s = \overline{1,S}$,

$$\begin{aligned}\alpha'''_{S+1|s}(\Phi^{*N}_1) &= P^N_s(A^{*N}_1 \cap B^N_{R+1}) + P^N_s(A^{*N}_2 \cap B^N_{S+1}) + P^N_s(A^{*N}_3)\\ &= P^N_s(B^N_{R+1}) + P^N_s(B^N_{S+1}) + P^N_s(A^{*N}_3)\\ &= \alpha''_{R+1|s}(\varphi^{*N}_2) + \alpha''_{S+1|s}(\varphi^{*N}_2) + P^N_s(A^{*N}_3)\end{aligned} \tag{17}$$

According to (16)–(17) and definition (15) of reliabilities we get

$$E'''_{l|s}(\Phi_1^*) = E''_{l|s}(\varphi_2^*), \quad l,s = \overline{1,S}. \tag{18}$$

$$E'''_{S+1|s}(\Phi_1^*) = \min[E''_{R+1|s}(\varphi_2^*),\ E''_{S+1|s}(\varphi_2^*),\ E_{3|s}^{I}], \quad s = \overline{1,S}. \tag{19}$$

Theorem 4. If all distributions P_s, $s = \overline{1,S}$, are different and positive values $E_{j|j}^{\prime *}$, $j = 1,2$ and $E''_{s|s}$, $s = \overline{1,S}$, satisfy compatibility conditions of correspondingly, Theorems 2 and 3, then elements of matrix of reliabilities $E'''(\Phi_1^*)$ of the two-stage test by one sample Φ_1^* can be found by (18)–(19).

When one of the compatibility conditions is violated, then at least one element of $E'''(\Phi_1^*)$ is equal to zero.

4. THE TWO-STAGE LAO TESTING WITH REJECTION OF DECISION BY A PAIR OF SAMPLES

Now we will discuss another version of testing. Suppose $N = N_1 + N_2$ be such that:

$$N_1 = \lceil \psi N \rceil, \quad N_2 = [(1-\psi)N], \quad 0 < \psi < 1,$$

$$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2), \qquad \mathrm{x} \in \mathcal{X}^N, \quad \mathcal{X}^N = \mathcal{X}^{N_1} \times \mathcal{X}^{N_2}.$$

The two-stage test by a pair of samples on the base of N-sample is denoted by $\Phi_2^N = (\varphi_1^{N_1}, \varphi_2^{N_2})$. The first stage is a non-randomized test $\varphi_1^{N_1}(\mathrm{x}_1)$ based on the sample x_1. The next stage is a non-randomized test $\varphi_2^{N_2}(\mathrm{x}_2, \mathrm{x}_1)$ based on sample x_2 and the outcome of test $\varphi_1^{N_1}(\mathrm{x}_1)$.

The first stage of two-stage test with rejection of decision by a pair of samples is in coming:

The first stage of decision making for choice of a family of PDs by a test $\varphi_1^{N_1}(\mathrm{x}_1)$ can be defined by division of the sample space $\mathcal{X}^{N_1}$ on three distinct subsets

$$A_i^{N_1} \overset{\Delta}{=} \{\mathrm{x}_1 : \varphi_1^{N_1}(\mathrm{x}_1) = i\}, \quad i = \overline{1,3}.$$

The set $A_i^{N_1}$, $i = 1,2$, consists all vectors x_1 for which i-th family of PDs is adopted and $A_3^{N_1}$ is all vectors x_1 for rejecting of the pair of family of PDs.

We define error probabilities $\alpha'_{i|j}(\varphi_1^{N_1})$, $i = \overline{1,3}$, $j = 1,2$ and reliabilities $E'_{i|j}(\varphi_1)$, $i = \overline{1,3}$, $j = 1,2$ analogous by Section 3. For construction of the LAO test for preliminarily given values $E'_{j|j}$, $j = \overline{1,2}$, we can define the subsets of distributions and conditions (5)–(8) and Theorem 1 similarly is used.

The second stage of two-stage test with rejection of decision by a pair of samples is as follows:

The test $\varphi_2^{N_2}(\mathrm{x}_2, \mathrm{x}_1)$ can be defined by division of the sample space $\mathcal{X}^{N_2}$ to $R+1$ (or $S-R+1$) distinct subsets. If the first family of PDs is accepted, then

$$B_s^{N_2} \overset{\Delta}{=} \{x_2: \varphi_2^{N_2}(x_2, x_1) = s\}, \quad s = \overline{1, R+1},$$

And if the second family of PDs is accepted, then $s = \overline{R+1, S+1}$.

The probability of the fallacious acceptance at the second stage of test of PD P_l, when P_s is correct, is

$$\alpha"_{l|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}(B_l^{N_2}), \qquad l \neq s, \qquad l, s = \overline{1, S}.$$

When at the first stage, the first family of PDs is accepted and at the second stage of test decision is rejected, but PD P_s is true, we have the following probability of error:

$$\alpha"_{R+1|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}\left(B_{R+1}^{N_2}\right), \qquad s = \overline{1, S},$$

and in this case if the second family of PDs is accepted, then error probability is

$$\alpha"_{S+1|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}\left(B_{S+1}^{N_2}\right), \qquad s = \overline{1, S},$$

The probability to reject P_s, when it is true and the first family of PDs is accepted, is

$$\alpha"_{s|s}(\varphi_2^{N_2}) \overset{\Delta}{=} P_s^{N_2}(\overline{B}_s^{N_2}) = \sum_{l \neq s, l = \overline{1, R+1}} \alpha"_{l|s}(\varphi_2^{N_2}), \quad s \in D_1. \tag{20}$$

Corresponding reliabilities for the second stage of test, are

$$E"_{l|s}(\varphi_2) \overset{\Delta}{=} \limsup_{N_2 \to \infty}\{-\frac{1}{N_2} \log \alpha"_{l|s}(\varphi_2^{N_2})\}, \quad l = \overline{1, R+1}, \quad s = \overline{1, S}. \tag{21}$$

It follows from (20) and (21)

$$E"_{s|s}(\varphi_2) = \min_{l \neq s} E"_{l|s}(\varphi_2), \quad l = \overline{1, R+1}, \quad s = \overline{1, S}.$$

Theorem 5. If at the first stage of test the first family of PDs is accepted, then for given positive and finite values $E''_{s|s}$, $s = \overline{1, R}$ of the reliabilities matrix $\boldsymbol{E}''(\varphi_2)$, *let us investigate the regions:*

$$R''_s = \{Q: \min_{l \in D_1} \mathrm{D}(Q || P_l) \leq E'^{*}_{1|1}, \quad \mathrm{D}(Q \parallel P_s) \leq E"_{s|s}\}, \quad s = \overline{1, R},$$

$$R''_{R+1} = \{Q: \min_{l \in D_1} \mathrm{D}(Q || P_l) \leq E'^{*}_{1|1}, \quad \mathrm{D}(Q \parallel P_s) > E"_{s|s}, \quad s = \overline{1, R}\},$$

and the following values of elements of the future reliabilities matrix $E''(\varphi_2^*)$ of the LAO test sequence:

$$E"^{*}_{s|s} = E"_{s|s}, \quad s = \overline{1, R},$$

$$E"^{*}_{l|s} = \inf_{Q \in R"_l} \mathrm{D}(Q \parallel P_s), \quad l = \overline{1, R+1}, \quad s = \overline{1, R}, l \neq s,$$

When the following compatibility conditions are valid

$$E"_{1|1} < \min_{s = \overline{2, R}} \mathrm{D}(P_s \parallel P_1),$$

$$E"_{s|s} < min[\underline{\min}_{l=\overline{1,s-1}} E"^{*}_{l|s}, \underline{\min}_{l=\overline{s+1,R}} \mathrm{D}(P_l \parallel P_s)], \quad 2 \le s \le R-1,$$

$$E"_{R|R} < \underline{\min}_{l=\overline{1,R-1}} E"^{*}_{l|R},$$

then there exists a LAO sequence of test φ_2^*, elements of reliabilities matrix $E''(\varphi_2^*)$ of which are defined above and are positive.

Even if one of the compatibility conditions is violated, then $E''(\varphi_2^*)$ has at least one element equal to zero.

If in the first stage of test, the second family of PDs is accepted, then for $S-R$ given positive values $E"_{s|s}$, $s=\overline{R+1,S}$ of reliabilities matrix $E''(\varphi_2^*)$, the procedure is analogous.

The reliabilities are surveyed for the two-stage test with rejection of decision by a pair of samples as follows:

The tool of making decision according to N-sample denoted $\Phi_2^{*N} = (\varphi_1^{*N_1}, \varphi_2^{*N_2})$ is organized by a pair of LAO tests $\varphi_1^{*N_1}$ and $\varphi_2^{*N_2}$. In the two-stage decision making, the test Φ_2^{*N} can be defined by partition of the sample space $\mathcal{X}^N$ to $S+1$ separate subsets as follows

$$C_s^N \overset{\Delta}{=} A_i^{*N_1} \times B_s^{N_2}, \quad s \in D_i, \qquad i = 1,2,$$

$$C_{S+1}^N \overset{\Delta}{=} (A_1^{*N_1} \times B_{R+1}^{N_2}) \cup (A_2^{*N_1} \times B_{S+1}^{N_2}) \cup A_3^{*N_1},$$

such that we can see

$$\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2) \in C_s^N: \quad \mathrm{x}_1 \in A_i^{*N_1}, \quad \mathrm{x}_2 \in B_s^{N_2}.$$

We can use definition of error probabilities $\alpha'''_{l|s}(\Phi_2^{*N})$, $l=\overline{1,S+1}$, $s=\overline{1,S}$ and reliabilities $E'''_{l|s}(\Phi_2^*)$, $l=\overline{1,S+1}$, $s=\overline{1,S}$ similar to Section 3. So we can consider error probabilities as follows

$$\alpha'''_{l|s}(\Phi_2^{*N}) = P_s^{N_1}\big(A_i^{*N_1}\big) \cdot P_s^{N_2}\big(B_l^{N_2}\big), \quad l,s \in D_i, \quad i = 1,2 \tag{22}$$

$$\alpha'''_{l|s}(\Phi_2^{*N}) = P_s^{N_1}(A_j^{*N_1}) \cdot P_s^{N_2}(B_l^{N_2}), \quad s \in D_i,\ l \in D_j, \quad i,j = 1,2,\ i \ne j \tag{23}$$

$$\alpha'''_{S+1|s}(\Phi_2^{*N}) = P_s^{N_1}\big(A_1^{*N_1}\big) \cdot P_s^{N_2}\big(B_{R+1}^{N_2}\big) + P_s^{N_1}\big(A_2^{*N_1}\big) \cdot P_s^{N_2}\big(B_{S+1}^{N_2}\big) + P_s^{N_1}\big(A_3^{*N_1}\big),\ s = \overline{1,S}, \tag{24}$$

Using properties of types we create the following equalities:

$$\lim_{N_1\to\infty}\{-\frac{1}{N_1}\log P_s^{N_1}(A_i^{*N_1})\} = \inf_{Q:\min_{l\in D_i} D(Q||P_l)\le E'^{*}_{i|i}} D(Q||P_s) \overset{\Delta}{=} E^{I}_{i|s}, \quad s \notin D_i, i = 1,2. \tag{25}$$

According to (22)–(25) and definition of reliabilities we obtain

$$E'''_{l|s}(\Phi_2^*) = (1-\psi)E"^{*}_{l|s}, \qquad l,s \in D_i, \quad i = 1,2, \tag{26}$$

$$E'''_{l|s}(\Phi_2^*) = \psi E_{j|s}^I + (1-\psi)E''^*_{l|s}, \quad s \in D_i, \quad l \in D_j, i,j = 1,2, \quad i \neq j, \qquad (27)$$

$$E'''_{s|s}(\Phi_2^*) = \min_{l \neq s} E'''_{l|s}(\Phi_2^*), \qquad s \in D_i, \; i = 1,2, \qquad (28)$$

$$E'''_{S+1|s}(\Phi_2^*) = E'''_{s|s}(\Phi_2^*), \qquad s = \overline{1,S}. \qquad (29)$$

Theorem 6. If all distributions P_s, $s = \overline{1,S}$, are different and positive values $E'^*_{j|j}$, $j = 1,2$ and $E''_{s|s}$, $s = \overline{1,S}$, satisfy compatibility conditions of Theorems 1 and 5, then elements of matrix of reliabilities $E'''(\Phi_2^*)$, of the two-stage test by a pair of samples Φ_2^* are defined in (26)–(29).

When one of the compatibility conditions is violated, then at least one element of $E'''(\Phi_2^*)$ is equal to zero.

5. COMPARISON OF RELIABILITIES MATRICES OF THREE METHODS

We compare the reliabilities matrices for the two-stage test by one sample and for the one-stage test. For comparison we will give some diagonal elements $E_{s|s} = E'''_{s|s}$, $s = \overline{1,S}$ of the reliabilities matrices. Taking in consideration that for $s \in D_i$, $i = 1,2$, by Theorem 3 we have

$$R'''_s = \{Q: \min_{l \in D_i} \mathrm{D}(Q||P_l) \leq E'^*_{i|i}, \quad \mathrm{D}(Q \parallel P_s) \leq E'''_{s|s}\}$$
$$= \{Q: \mathrm{D}(Q \parallel P_s) \leq E'''_{s|s}\} = \{Q: \mathrm{D}(Q \parallel P_s) \leq E_{s|s}\} = R_s,$$

and reliabilities are

$$E'''_{s|l} = \inf_{Q \in R''_s} \mathrm{D}(Q \parallel P_l) = \inf_{Q \in R_s} \mathrm{D}(Q \parallel P_l) = E_{s|l}.$$

And at result we have $R'''_{S+1} = R_{S+1}$ and for the $(S+1)$-th column reliabilities are equal too. So we will receive to following result.

If all distributions P_s, $s = \overline{1,S}$, are different and positive values of diagonal elements $E_{s|s} = E'''_{s|s}$, $s = \overline{1,S}$ of the reliabilities matrices of one-stage test and two-stage test by one sample, satisfy compatibility conditions shown in Theorems 1-6, then reliabilities of two matrices are equal but they are greater than the tow-stage test by a pair of samples.

Example. Suppose $\mathcal{X} = \{a,b,c\}$, the first family of PDs contains two PDs $P_1 = (0.1,0.1,0.8)$ and $P_2 = (0.2,0.2,0.6)$ and the second family contains three PDs $P_3 = (0.4,0.4,0.2)$, $P_4 = (0.5,0.4,0.1)$ and $P_5 = (0.6,0.2,0.2)$.
We present values of divergences of all pairs of PDs.

Values of $\mathrm{D}(P_s \parallel P_l)$

No.	$l = 1$	$l = 2$
$s = 3$	0.3612	**0.1454**
$s = 4$	0.5	0.2415
$s = 5$	0.4067	0.1908

Applying Theorem 1 we see that

$$0 < E'^{*}_{1|1} < \min_{s=\overline{3,5},l=\overline{1,2}} \mathrm{D}(P_s||P_l) = 0.1454,$$

and consequently for example if $E'^{*}_{1|1} = 0.01$ then we have

$$E'^{*}_{2|2} < E'^{*}_{1|2} = \min_{s=\overline{3,5}} \inf_{\min\limits_{l=1,2} Q:\mathrm{D}(Q||P_l)\leq 0.01} \mathrm{D}(Q||P_s) = 0.0933.$$

Let the following preliminarily values are given

$E_{1\|1}$	$E_{2\|2}$	$E_{3\|3}$	$E_{4\|4}$	$E_{5\|5}$
0.01	0.009	0.004	0.005	0.02

The reliabilities matrix of the first stage of test is in coming:

$$\boldsymbol{E}'(\Phi_1^*) = \begin{bmatrix} 0.0100 & 0.0649 & 0.0100 \\ 0.0933 & 0.0200 & 0.0200 \end{bmatrix}.$$

The reliabilities matrices of the one-stage test $\boldsymbol{E}(\phi^*)$ and the two-stage test by one sample $\boldsymbol{E}'''(\Phi_1^*)$ with possibility of rejection of decision and the same values of diagonal elements, is as follows

$$\boldsymbol{E}(\phi^*) = \boldsymbol{E}'''(\Phi_1^*) = \begin{bmatrix} 0.0100 & 0.0129 & 0.2978 & 0.4294 & 0.2558 & 0.0100 \\ 0.0114 & 0.0090 & 0.1058 & 0.1914 & 0.0910 & 0.0090 \\ 0.2629 & 0.0969 & 0.0040 & 0.0050 & 0.0063 & 0.0040 \\ 0.4624 & 0.2225 & 0.0062 & 0.0050 & 0.0067 & 0.0050 \\ 0.2779 & 0.1202 & 0.0250 & 0.0234 & 0.0200 & 0.0200 \end{bmatrix}.$$

The reliabilities matrix of the two-stage test by a pair of samples $\boldsymbol{E}'''(\Phi_2^*)$ with $\psi = 0.05$ and possibility of rejection of decision and the same diagonal elements, is as follows

$$\boldsymbol{E}'''(\Phi_2^*) = \begin{bmatrix} 0.0100 & 0.0116 & 0.2829 & 0.4074 & 0.2405 & 0.0100 \\ 0.0105 & 0.0090 & 0.1010 & 0.1817 & 0.0854 & 0.0090 \\ 0.2531 & 0.0956 & 0.0040 & 0.0045 & 0.0054 & 0.0040 \\ 0.4418 & 0.2140 & 0.0056 & 0.0050 & 0.0058 & 0.0050 \\ 0.2671 & 0.1180 & 0.0233 & 0.0216 & 0.0200 & 0.0200 \end{bmatrix},$$

This matrix show that the reliabilities of two-stage test by a pair of samples are near to the reliabilities of the one-stage test.

7. CONCLUSION

We have shown that the number of the preliminarily given values of elements of the reliabilities matrices of the one-stage test and of the two-stage tests by one sample and by two samples would be the same but the procedure of calculations for the two-stage tests would be shorter. So the consumer has possibility to use the method which is preferable. We can show that the number of operations of the two-stage test by one sample is less than this of one-stage test and is more than quantity of operations of two-stage test by a pair of samples. This was observed also during experimental calculations of example.

ACKNOWLEDGMENT

The author is indebted to his Professor Evgueni Haroutunian for many helpful suggestions and criticism during the course of this work.

REFERENCES

1. Ahlswede R. and Haroutunian E.A. (2006). On statistical hypotheses optimal testing and identification." Lecture Notes in Computer Science, vol. 4123, General Theory of Information Transfer and Combinatorics, Springer, pp. 462-478.
2. Ahlswede R., Aloyan E. and Haroutunian E. (2006). On logarithmically asymptotically optimal hypothesis testing for arbitrarily varying source with side information. *Lecture Notes in Computer Science*, vol. 4123, General Theory of Information Transfer and Combinatorics, Springer Verlage, pp. 457-461.
3. Ahlswede R. and Csisz´ar I. (1997). Hypothesis testing with communication constraints. *IEEE Transactions on Information Theory*, 32(4), 533-542.
4. Birge L. (1981). Vitesses maximales de decroissence des errors et tests optimaux associees. *Wahrsch. Verw. Gebiete*, 55, 261-273.
5. Blahut R.E. (1974). Hypotheses testing and information theory. *IEEE Trans. Information Theory*, 20(4), 405-417.
6. Cover T.M. and Tomas J.A. (2006). *Elements of Information Theory*. Second edition, Wiley, New York.
7. Csiszár I. and Longo G. (1971). On the error exponent for source coding and for testing simple statistical hypotheses. *Studia Sc. Math. Hungarica*, 6, 181-191.
8. Csisz´ar I. and Shields P. (2004). Information theory and statistics: *A tutorial. Foundations and Trends in Communications and Information Theory*, 1(4).
9. Fu F.W. and Shen S.Y. (1998). Hypothesis testing for arbitrarily varying source with exponential-type constraint. *IEEE Transactions on Information Theory*, 44(2), 892-895.
10. Haroutunian E.A. (1990). Logarithmically asymptotically optimal testing of multiple statistical hypotheses. *Problems of Control and Information Theory*, 19(5-6), 413-421.
11. Haroutunian E.A. (2005). Reliability in multiple hypotheses testing and identification. *Proceedings of the NATO-ASI Conference, vol. 198 of NATO Science Series III: Computer and Systems Sciences*, Yerevan, Armenia, pp. 189-201, IOS Press.
12. Haroutunian E.A., Haroutunian M.E. and Harutyunyan A.N. (2008). Reliability criteria in information theory and in statistical hypothesis testing. Foundations and Trends in Communications and Information Theory, vol. 4, nos. 2-3.
13. Haroutunian E.A. and Hakobyan P.M. (2009). Multiple hypotheses LAO testing for many independent objects." Scholarly Research Exchange.
14. Haroutunian E.A., Hakobyan P.M. and Hormozi nejad F. (2012). On two-stage logarithmically asymptotically optimal testing of multiple hypotheses concerning distributions from the pair of families. Transactions of IIAP of NAS of RA and of YSU. *Mathematical Problems of Computer Science*, 37, 34-42.
15. Haroutunian E.A., Hakobyan P.M. and Yessayan A.O. (2011). Many hypotheses LAO testing with rejection of decision for arbitrarily varying object. Transactions of IIAP of NAS of RA. *Mathematical Problems of Computer Science*, 35, 77-85.

16. Haroutunian E.A, Hakobyan P.M. and Yessayan A.O. (2011). On multiple hypotheses LAO testing with rejection of decision for many independent objects. *Proceedings of International CSIT Conference*. pp. 141-144.
17. Hoeffding W. (1965). “Asymptotically optimal tests for multinomial distributions. *Annals of Mathematical Statistics*, 36, 369-401.
18. Hormozi nejad F., Haroutunian E.A. and Hakobyan P.M. (2011). On LAO testing of multiple hypotheses for the pair of families of distributions. *Proceeding of the Conference “Computer Science and Information Technologies"*, Yerevan, Armenia, pp. 135-138.
19. Hormozi nejad F., Haroutunian E.A. and Hakobyan, P.M. (2012). On Two-stage LAO Testing of Multiple Hypotheses for the Pair of Families of Distributions. *Electronic Journal of Statistics*, 6, 1-25.
20. Tusnady G. (1977). On asymptotically optimal tests. *Annals of Statistics*, 5(2), 385-393.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 395-408

COPYRIGHT PROTECTING USING VISIBLE REMOVABLE WATERMARKING

Ahmed Mohmed Abushaala and **Zaineb Ateia Moammer Elghoul**
Faculty of Information Technology, Misurata University, Misurata-Libya
Email: am_rata@yahoo.co.uk; zoba.moammer@yahoo.com

ABSTRACT

A recent proliferation and success of the Internet, together with availability of relatively inexpensive digital recording and storage devices has created an environment in which it became very easy to obtain, replicate and distribute digital content without any loss in quality. This has become a great concern to the multimedia content (music, video, and image) publishing industries, because technologies or techniques that could be used to protect intellectual property rights for digital media and prevent unauthorized copying did not exist.

This need attracted attention from the research community and industry leading to a creation of a new information hiding form, called *Digital Watermarking*. Basic idea is to create a metadata containing information about a digital content to be protected, and hide it within that content. The information to hide, the metadata, can be invisible or visible watermarking.

In our work, we use the frequency domain transform applying Discrete Cosine Transform (DCT), which has high efficiency in terms of performance in two phases embedding the watermark and extracting the original image without impacting significantly on the images. A visible watermark (logo) is embedded to standard images (like Lena and Pepper images). The visible watermark can take many scales of transparency depending on random embedding factor.

we obtained good results through subject testing of the embedded watermark images, and by the retrieval process we obtained a high measure of PSNR.

KEYWORDS

Watermarked image, visible Watermarking, host image, digital watermark, Copyright Protection, Discrete Cosine Transform, Peak Signal-to-Noise Ratio.

1. INTRODUCTION

1.1 Steganography and Watermarking-History and Terminology:

The idea to communicate secretly is as old as communication itself. First stories, which can be interpreted as early records of covert communication, appear in the old Greek literature, for example, in Homer's *Iliad*, or in tales by Herodotus. The word "*steganography* ", which is still in use today, derives from the Greek language and means covert communication. have investigated the history of covert communication in great detail, including the broad use of techniques for secret and covert communication before and during the two World Wars, and steganographic methods for analog signals. Although the historical background is very interesting, we do not cover it here in detail.

Paper watermarks appeared in the art of handmade paper making nearly 700 years ago. The oldest watermarked paper found in archives dates back to 1292 and has its origin in *Fabriano*, Italy, which is considered the birthplace of watermarks. At the end of the thirteenth century, about 40 paper mills were sharing the paper marked in *Fabriano* and producing paper with different format, quality, and price. They produced raw, coarse paper which was smoothed and post processed by artisans and sold by merchants. Competition not only among the paper mills but also among the artisans and merchants was very high, and it was difficult to keep track of paper provenance and thus format and quality identification. The introduction of watermarks helped avoiding any possibility of confusion. After their invention, watermarks quickly spread over Italy and then over Europe, and although originally used to indicate the paper brand or paper mill, they later served as indication for paper format, quality, and strength and were also used to date and authenticate paper. A nice example illustrating the legal power of watermarks is a case in 1887 in France called “Des Decorations”. The watermarks of two letters, presented as pieces of evidence, proved that the letters had been predated and resulted in considerable sensation and, in the end, in the resignation of President Gr´evy.

The analogy between *paper watermarks*, *steganography*, and *digital watermarking* is obvious, and in fact, paper watermarks in money bills or stamps, see figure(1), actually inspired the first use of the term watermarking in the context of digital data.

Figure (1): stamp is an example of the watermark.

The idea of digital image watermarking arose independently in 1990, and around 1993. coined the word “water mark” which became “watermark” later on. It took a few more years until 1995/1996 before watermarking received remarkable attention. Since then, digital watermarking has gained a lot of attention and has evolved very quickly. Frank Hartung (1999).

There are many applications that use the digital watermark, including: Ownership Assertion, Fingerprinting, Authentication and integrity verification, Content labeling, Usage control, Content protection. Memon & Wong (1998).

What is the different between "steganography" and "watermarking"?

Steganography is the act of adding a hidden message to an image or other media file. It is similar to encrypting a document, but instead of running it through a cypher, the document is broken up and stored in unused, or unnoticeable, bits within the overall image.

Watermarking is similar, but has a completely different purpose. Placing a watermark in an image or other media file serves to identify the artist or author of the work. It isn't so much an attempt to hide a message as it is to tag a document for later identification. Watermarking is used to protect the copyright of the original owner, to make their claim stronger, should the image be used without permission by someone else.

2. DIGITAL WATERMARKING

Digital watermarking is the process by which identifying data is woven into media content such as images, movies, music or programming, giving those objects a unique, digital identity. It is the method of embedding data into digital multimedia content. This is used to verify the credibility of the content or to recognize the identity of the digital content's owner. Saraju (1999). Digital watermarking can be classified into different category according to the host signal:

(1) *Digital Image Watermarking*: Most of the research about digital watermarking is on image watermarking. This might be due to that there are so many images available on World Wide Web free of charge and without any copyright protection.

(2) *Digital Video Watermarking*: A video sequence consists of still images , therefore all the watermarking methods applied on image can be applied on video. However, video watermarking has other problems. For example, it is dangerous to use the same watermarking key for a whole video. If the same key is used for all the frames or shots in a video sequence, it would make the watermarking algorithm vulnerable to the collusion attack.

(3) *Digital Audio Watermarking:* In case of audio signals, the term “watermarking” can be defined as “robust and inaudible transmission of additional data along with audio signals”. Audio watermarking is based on the perceptual audio coding techniques.

(4) *Others:* Such as holograph, text, software and database watermarking.

2.1 Types of Digital Image Watermarking:

There are several types of watermark, including:

- Visible watermarking: This type that is clear or transparent and can be seen by the eye, but it is significant, and often seal, logo or name of the owner and that when you use digital material on the internet, and is placed in this way so that people view it can't be content distribution without presents it and can't be edited easily, see figure (2).

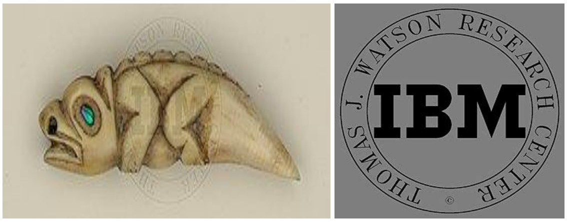


Figure (2): images for the visible watermark

- Invisible watermarking: This type that is not clear which does not appear on the digital content. Where encrypted within the same content, and used to track whether the content is the real owner, see figure(3)

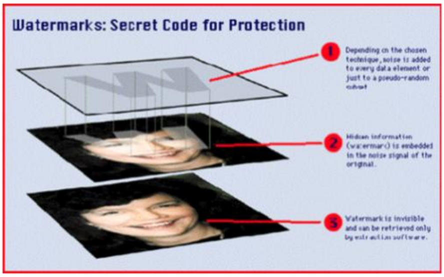

Watermarks: Secret Code for Protection
1
Depending on the chosen technique, noise is added to every data element or just to a pseudo-random subset
2
Hidden information (watermark) is embedded in the noise signal of the original.
3
Watermark is invisible and can be retrieved only by extraction software.

Figure (3): image for the invisible watermark

2.2 Digital Watermark Requirements:

There are three key requirements for digital watermark are: transparency, Robustness and capacity, Vidyasagar et al. (2005).

- ***Transparency:*** Digital watermark should not affect the quality of the original image after you add them. Cox et al (2002) identified transparency or accuracy as "perceptual similarity between versions or originals and copies added to the watermark." The water should not mark clear distortions in the picture, because if you find these distortions they reduce the commercial value of the image.
- ***Robustness:*** Cox et al (2002) defined toughness as "the ability to detect the watermark after common signal processing operations." And durability also means the ability to resist changes and amendments to the original file, such as resizing, cropping, rotation, and attacks such as adding noise. Accordingly, we rated the watermark to:

  Strong watermark: designed for combat against manipulation, all applications that require security watermark systems require this type of watermark.

  Fragile watermark: included with very low durability. This type can be destroyed less manipulation, in this sense be similar to send hidden messages in ways Steganographic, and can be used to verify the integrity of the goals.

- ***Capacity:*** Cox et al (2002) identified the ability or capacity as "the number of bits encrypted watermark within the unit of time or work." And also suggest that the watermark capacity does not depend on the algorithm used, but is linked to the characteristics of the host signal. After the top 3 listing requirements for the watermark will complete the rest of the other requirements:
- ***Complexity:*** Complexity describes spending to detect and encrypt the watermark information, it is recommended that the watermark intricately designed, for example, can integrate different watermarks with each other.
- ***Security***: There are two levels of security. In the first level the user can not unauthorized reading and decoding the watermark and can't be disclosed if given data contain a watermark. The second level allows the user to unauthorized disclosure if the data contains the watermark, and with that information can't be read without the knowledge of the secret key.

2.3 Digital Watermark Techniques:

Many different techniques to embed watermark proposed during the past few years. Generally all watermarks can be divided into the spatial domain techniques and frequency domain techniques.

- **Spatial Domain Techniques:** Of these techniques recall the following technique:
- ***Least Significant Bits:*** This method is the simplest technique in the spatial domain techniques, and is dependent on finding bits important Bits of others in the digital file as these bits can be replaced with information that will hide, and from abroad are not edited the file significantly. LSB working on files that have a high degree of clarity and audio files which have many different sounds, and this method usually does not increase the size of the file, but depending on the amount of information that will hide inside the file, it can become distorted significantly.
- **Frequency Domain Techniques:** It is include Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT). In this paper we dealt with (DCT).
- ***Discrete Cosine Transform:*** Is one of the most important in the process of transfers include a watermark. And the main goal for DCT transform representation's spatial representation areas bandwidth which is suitable for the disposal of items with high frequency in the matrix images, it represents is like a group of Sinusoids to modify frequencies, according to the following equations:

$$F(u,v) = \alpha_u \alpha_v \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} f(x,y) \cos\left(\frac{\pi u(2x+1)}{2N}\right) \cos\left(\frac{\pi v(2y+1)}{2M}\right) \ldots\ldots\ldots (1)$$

$$F(u,v) = \alpha_u \alpha_v \sum_{y=0}^{M-1} \sum_{x=0}^{N-1} f(x,y) \cos\left(\frac{\pi u(2x+1)}{2N}\right) \cos\left(\frac{\pi v(2y+1)}{2M}\right) \ldots\ldots\ldots (2)$$

$$\alpha_v = \begin{cases} \sqrt{\frac{1}{M}}, & u = 0 \\ \sqrt{\frac{2}{M}}, & otherwise \end{cases} \qquad \alpha_u = \begin{cases} \sqrt{\frac{1}{N}}, & u = 0 \\ \sqrt{\frac{2}{N}}, & otherwise \end{cases}$$

Ahmed, Natarajan, and Rao (1974) first came DCT discrete cosine in early Seventies, and since that time has increased his popularity, In particular class by Wang (1984) to four different transfers named DCT-I, DCT-II, DCT-III, DCT-IV. Ziad & Martin (2001).

In discrete cosine transform DCT split the image to the non-overlapping blocks and apply the DCT on each block. This produces us three areas for different frequencies: Low Frequency Sub-band, Mid Frequency Sub-band, and High Frequency Sub-band. Embedded watermark using DCT based on two facts:

The first truth is that most of the energy signals emerge in the low frequency region that contains the most important visual parts in the picture, and second fact is that the high-frequency components of the image still usually through pressure and noise attacks. So that the watermark included modifying frequency coefficients medium in order not to be affected by the clarity of the image, and the watermark will still process pressure. Saeed &Ahmad (2009).

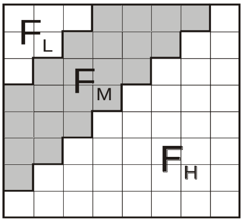


Figure (4): areas for inclusion in discrete cosine transform DCT

Visual watermark technology useful and correct must meet the following requirements Yeung et al. (1997):

- must be clear in both color images and monochrome.
- spread in a large and important area of the image in order to prevent deleted by cutting them.
- should be visible but does not obscure the original image underneath.
- must be difficult removal, where the removal of the image through the purchase by the owner.
- must be applied automatically with some intervention by the person.

3. ARCHITECTURAL SCHEME WATERMARK

Most algorithms include visual watermark does not take into account the process of removing the watermark and recover the original image. But in some applications be required of users remove this tag of the image that ensured within, for example, sends a contagious programs television program with the slogan television users, authorized users can get on the way still from which the logo of the television program, while maintaining the quality high visual of the original television program. Hongyuan et al (2010).

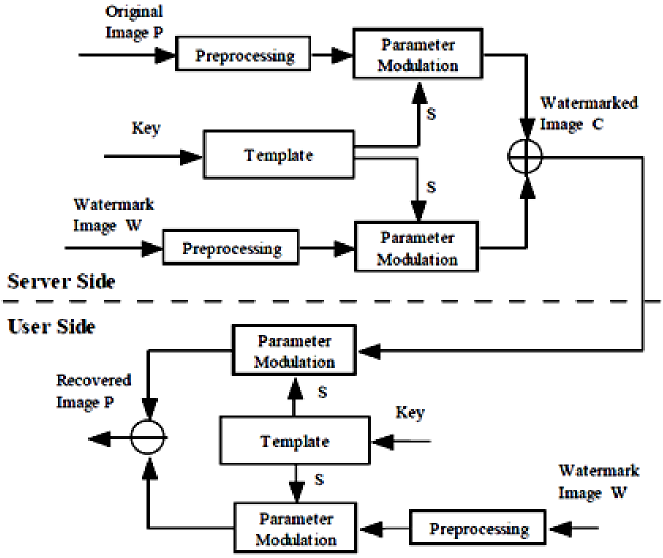


Figure (5): include architectural and remove the visible watermark

The architectural scheme shown in figure (5) is composed of three parts: a coefficient modulated generator and is a pseudorandom sequence, Watermarking Embedding, and Watermarking Removable. At the server side, the original image is embedded with visible watermark into a watermarked image. Then, the image is transmitted to the users by networks. At the user side, the users remove the visible watermark and recover the original image.

Before the process of embedding or removing the watermark pseudorandom sequence (coefficient modulated) have been generated and the process of generating the pseudorandom sequence **S** known as follows:

$$S=T(a,b)$$

$$S=s0,\ s1,.......,s(n-1)\ ,\ a=<s(i)=<b\ ,\ i=1,2,.......n-1 \quad ....(3)$$

Where T (.) operation to generate modulated coefficient, **n** is the number of elements in each block, and a, b is the period that limit coefficient modulated and be (0,1) .

4. STEPS EMBEDDING VISIBLE WATERMARKING

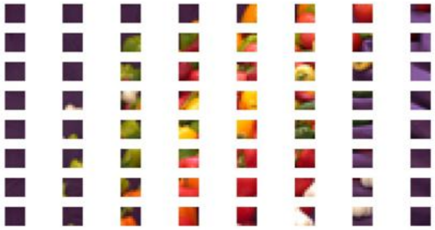
- ***Image Preparation Phase:*** Is pre-basic steps to embed the watermark, where is inserted and read color image using directive reading (imread), and sometimes we need to change logo image (watermark) type of JPG to PNG ,because it has colorful background may affect the output of the embedding process and PNG format provides us with the high quality of the image, as well as the transparency we need.
- ***The First Step:*** In the process of embedding, embedding the watermark image directly into the RGB components are not suitable for color space be greatly complicated. In this scheme adopt YUV color space to include a watermark. To convert the watermark image and the host image from RGB to YUV:

$$\begin{cases} Y = 0.299R + 0.587G + 0.114B \\ U = -0.148R - 0.289G + 0.437B \\ V = 0.615R - 0.515G - 0.100B \end{cases} \quad \text{--------}\ (4)$$

where Y is Luminance component in the images.

- ***The Second Step:*** In this step we have more of the process:

1- Dividing images into non-overlapping blocks the size of 8x8, elements of these blocks range (0-255).



$$Original = \begin{bmatrix} 154 & 123 & 123 & 123 & 123 & 123 & 123 & 136 \\ 192 & 180 & 136 & 154 & 154 & 154 & 136 & 110 \\ 254 & 198 & 154 & 154 & 180 & 154 & 123 & 123 \\ 239 & 180 & 136 & 180 & 180 & 166 & 123 & 123 \\ 180 & 154 & 136 & 167 & 166 & 149 & 136 & 136 \\ 128 & 136 & 123 & 136 & 154 & 180 & 198 & 154 \\ 123 & 105 & 110 & 149 & 136 & 136 & 180 & 166 \\ 110 & 136 & 123 & 123 & 123 & 136 & 154 & 136 \end{bmatrix}$$

Figure (6): dividing the image into blocks of equal size 8X8
Figure (7): block show is divided into 8X8

2- Because DCT is designed to work on the pixel values that range from -128 to 127, blocks and subtractions from the original value 128 of each block.

3- converts the blocks to a frequency domain by applying the 2D-DCT on each block.

$$D = \begin{bmatrix} 162.3 & 40.6 & 20.0 & 72.3 & 30.3 & 12.5 & -19.7 & -11.5 \\ 30.5 & 108.4 & 10.5 & 32.3 & 27.7 & -15.5 & 18.4 & -2.0 \\ -94.1 & -60.1 & 12.3 & -43.4 & -31.3 & 6.1 & -3.3 & 7.1 \\ -38.6 & -83.4 & -5.4 & -22.2 & -13.5 & 15.5 & -1.3 & 3.5 \\ -31.3 & 17.9 & -5.5 & -12.4 & 14.3 & -6.0 & 11.5 & -6.0 \\ -0.9 & -11.8 & 12.8 & 0.2 & 28.1 & 12.6 & 8.4 & 2.9 \\ 4.6 & -2.4 & 12.2 & 6.6 & -18.7 & -12.8 & 7.7 & 12.0 \\ -10.0 & 11.2 & 7.8 & -16.3 & 21.5 & 0.0 & 5.9 & 10.7 \end{bmatrix}$$

Figure (8): the block Coefficients after applying DCT

This block consists of 64 coefficient from DCT coefficients, the coefficient that located at the top of the left D (0,0) represents the low frequency of the block of the original image, and coefficient D (7,7) represents the highest frequency of the image block. It is important to know that the human eye is very sensitive to low frequencies.

- ***The Third Step :*** Host image coefficients Pi and logo image Wi adjusted by using modulated coefficient that is randomly generated Si through:

$$\begin{cases} p_i' = (1-s(i))p_i \\ w_i' = s(i)w_i \end{cases} \quad i = 0, 1, \ldots, n-1 \quad \text{-----} \ (5)$$

- ***The Fourth Step :*** The modulated watermark image is embedded in the host image by :

$$c(i) = p_i^{'} + w_i^{'} \quad i = 0,1,\dots,n-1 \quad \text{------}(6)$$

- ***The Fifth Step:*** Applying inverse discrete cosine transform IDCT on C, then convert the watermarked image from color space YUV to RGB.

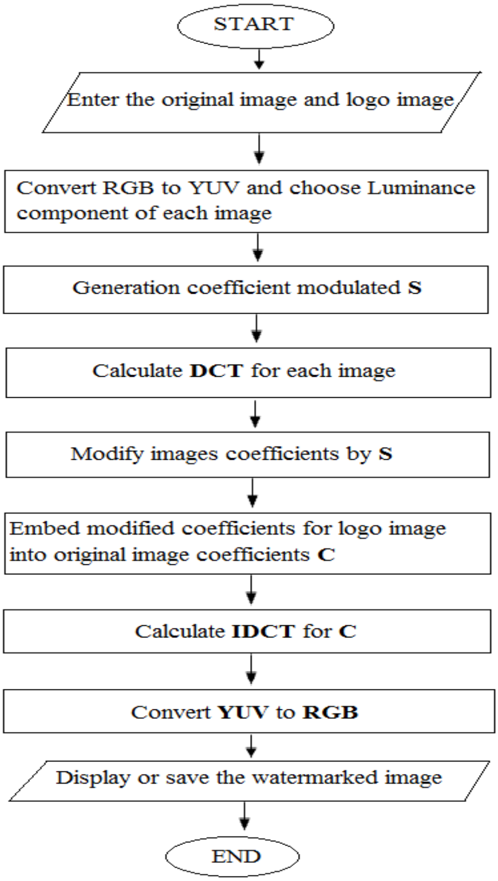

START
Enter the original image and logo image
Convert RGB to YUV and choose Luminance component of each image
Generation coefficient modulated S
Calculate DCT for each image
Modify images coefficients by S
Embed modified coefficients for logo image into original image coefficients C
Calculate IDCT for C
Convert YUV to RGB
Display or save the watermarked image
END

Figure (9): flowchart of the embedding process

5. STEPS OF RETRIVAL THE ORIGINAL IMAGE

1- on the side of the authorized user, modulated coefficient have been born and is the same coefficient that we used in the embedding process.

2- modifying of the DCT coefficients for the watermark image by the equation:

$$w_i^{'} = s(i)\ w_i \quad i = 0,1,\dots,n-1 \quad \text{------}\ (7)$$

where S (i) is the coefficient modulated, Wi is the DCT coefficients of the original watermark image.

3- after modifying coefficients for the watermark image, the retrieval process is given by:

$$p(i) = \frac{c(i) - w_i^{*}}{1 - s(i)} \quad i = 0, 1, \ldots, n-1 \quad \text{-------} \ (8)$$

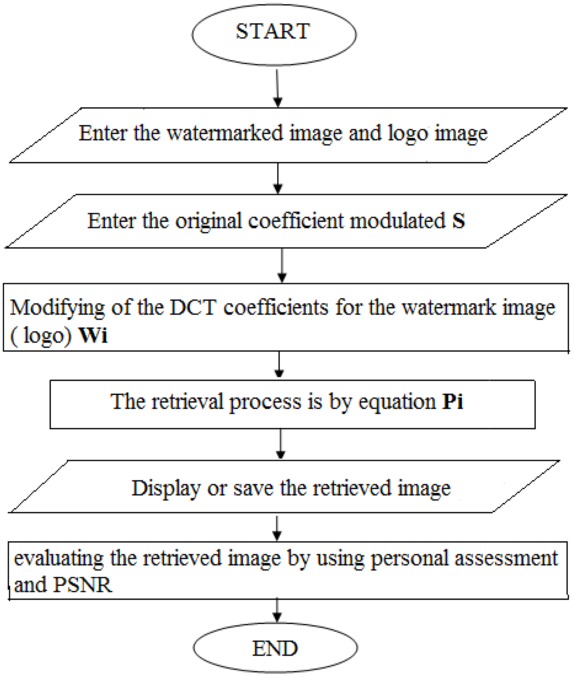

START
Enter the watermarked image and logo image
Enter the original coefficient modulated **S**
Modifying of the DCT coefficients for the watermark image (logo) **Wi**
The retrieval process is by equation **Pi**
Display or save the retrieved image
evaluating the retrieved image by using personal assessment and PSNR
END

Figure (10): flowchart of the retrieval process

6. EVALUATION AND RESULTS AND CONCLUSIONS

After doing the process Embed visual watermark and the original host image retrieval, we evaluated the results that we have acquired from the two processes through the adoption of the two methods of evaluation are:

- Evaluation Personal : Is by looking at the naked eye to the resulting images and evaluated, and given estimate, either excellent or very good or good or poor, and different appreciation given from one person to another by his vision and accepted the resulting images.
- Image Quality Scale : Is one of the historical algorithms used in image processing, and is an a shortcut for ***Peak Signal-to-Noise Ratio*** , it represents the great ability of the signal to noise, and is used to measure the improved image quality recovered, and calculated by the following equation:

$$MSE = \frac{1}{NxM} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} [X(i,j) - Y(i,j)]^2$$

$$PSNR = 10\log_{10}\frac{L^2}{MSE} \quad \text{------ (9)}$$

where L represents the range of values that can be taken pixel, For example, Y channel encoded depth 8 bits, L = 2 ^ 8-1 = 255.

The more the value of *PSNR* increased strength of the similarities between the original image and the image recovered, and vice versa. Top similarity between the images when *MES = 0* shall be *PSNR* tends to infinity. Excellent values range from *30 dB* to *50 dB*, while the acceptable range settles about *25 dB*. د. عصـام عبود (2006).

Note: To use this scale photographs must be of the same type and the same size.

To get the results we used a set of color images with changing the size of the area include:

- We used RGB images of type JPG and the dimensions of 128 * 128 to represent the host image, but for the watermark image were used color images of type PNG dimensions of 128 * 128 in order to achieve transparency.
- We used RGB images of type JPG and dimensions 256 * 256 to represent the host image, but for the watermark image were used color images of type PNG dimensions of 128 * 128, we will add the watermark image to the host image in a specific area and not in every image.

We noticed that when we used images of the same size, or different images in size added in specific area, the process of embedding done correctly and in the selected area, and the value of PSNR of pictures is infinite or is very high, and this means that the image recovered mismatched with the host image before the embedding process.

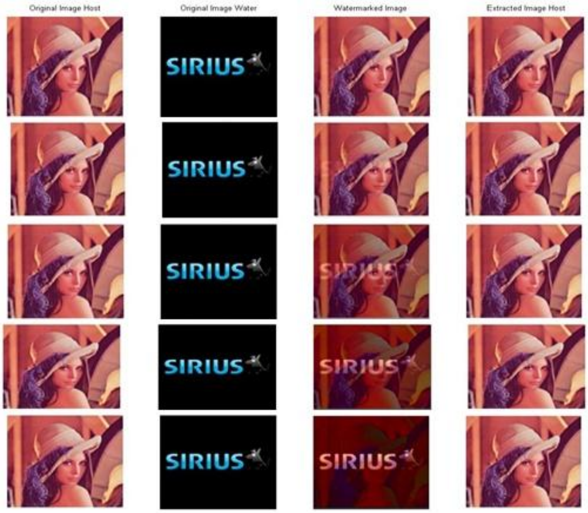


Figure (11): images illustrating the process of embedding and retrieval of images equal size (lenaa)

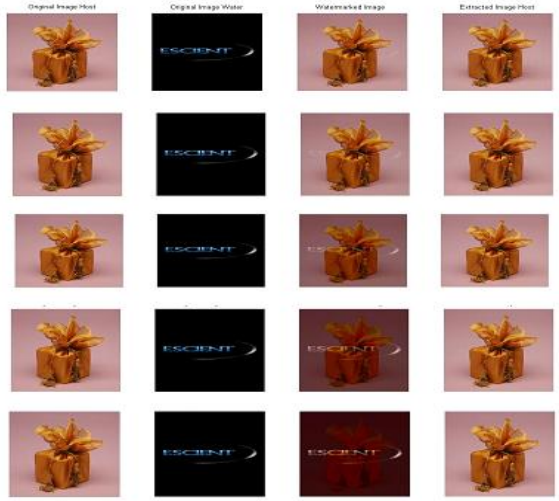


Figure (12): images illustrating the process of embedding and retrieval of images equal size (pic5)

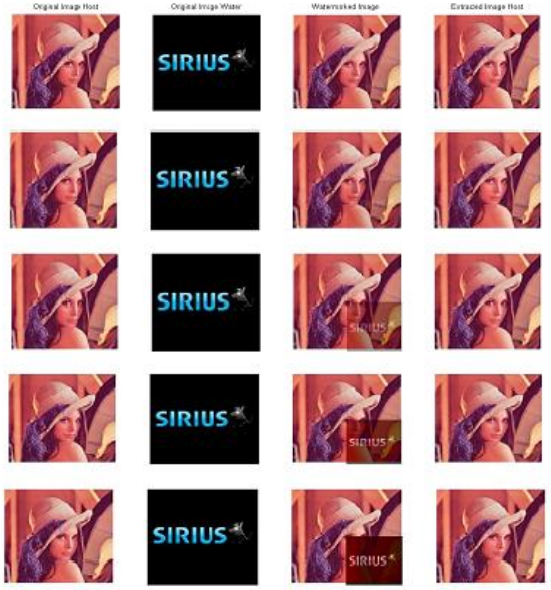


Figure (13): Photos illustrating the process of embedding and retrieval of different sized images (lenaa)

Figure (14): Photos illustrating the process of embedding and retrieval of different sized images (pic5)

Table (15)
Results of a scale equal PSNR for images and others of equal size

PSNR	Watermark image Png	Host image Jpg	S (I)
Inf	Logo 4	Pic 5	0.04
Inf	Logo 4	Pic 5	0.1
Inf	Logo 4	Pic 5	0.36
Inf	Logo 4	Pic 5	0.6
Inf	Logo 4	Pic 5	0.85
Inf	Logo 3	Lenaa	0.04
Inf	Logo 3	Lenaa	0.1
Inf	Logo 3	Lenaa	0.36
Inf	Logo 3	Lenaa	0.6
Inf	Logo 3	Lenaa	0.85

Here we have a simple test and is a different value for the coefficient of modulation, ie S used in the modulated by the addressee value different from the S in the recovery process in the future. We note the following results:

- Whenever the value of the coefficient modulated incorrect smaller than the correct coefficient modulated values to be part of the watermark image still exists in the recovered image, and the value of PSNR rise the closer we get to the correct coefficient modulated value.
- Whenever the value of the coefficient modulated incorrect larger than the correct coefficient modulated values disappear watermark image from the recovered image and shows the effect on recovered image itself, and the value of PSNR rise the closer we get to the correct coefficient modulated value.
- Value RSNR be high as long as the effect is not affects on the recovered image itself.

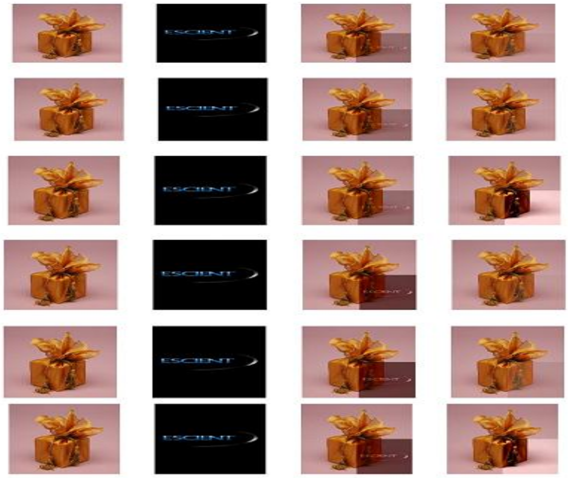


Figure (16): using incorrect coefficient modulated

Table (17)
Results PSNR scale images coefficient modulated wrong

PSNR	S (I) Receiver	Watermark image Png	Host image Jpg	S (I) Sender
43.2734	0.05	Logo 4	Pic 5	0.3
49.8354	0.2	Logo 4	Pic 5	0.3
29.1418	0.7	Logo 4	Pic 5	0.3
45.6107	0.5	Logo 4	Pic 5	0.6
39.0361	0.3	Logo 4	Pic 5	0.6
30.8356	0.8	Logo 4	Pic 5	0.6

REFERENCES

1. Dr.عصام عبود, (2006). انقاص التأثير الكلي الناتج عن ضغط الصور المعتمد على تحويل التجيب المتقطع".
2. Frank Hartung, Student Member, IEEE, and Martin Kutter, (JULY 1999). Multimedia Watermarking Techniques.
3. Hongyuan Li, Guangjie Liu, Yuewei Dai, Zhiquan Wang, (November 2010). Copyright Protecting Using the Secure Visible Removable Watermarking In JPEG Compression.
4. Memon, N. and Wong, P.W. (1998). Protecting digital media content.
5. Saeed K. Amirgholipour, Ahmad R. Naghsh-Nilchi, (June 2009). Robust Digital Image Watermarking Based on Joint DWT-DCT.
6. Saraju Prasad Mohanty, (JANUARY 1999), Watermarking of Digital Images.
7. Vidyasagar M. Potdar, Song Han, Elizabeth Chang, (INDIN 2005), A Survey of Digital Image Watermarking Techniques, © 2005 IEEE.
8. Yeung et al. (1997) IEEE. Digital Watermarking for High-Quality Imaging.
9. Ziad M. Hafed and Martin D. Levine, (2001). Face Recognition Using the Discrete Cosine Transform.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 409-426

RESAMPLING METHOD FOR THE ADAPTIVE CHOICE OF TUNING CONSTANT AND VARIABLE SELECTION IN ROBUST REGRESSION

Zafar Mahmood[1] and **Salahuddin**[2]
[1] Department of Mathematics, Statistics and Computer Sciences,
Khyber Pakhtunkhwa Agricultural University Peshawar, Pakistan
[2] Department of Statistics University of Peshawar, Peshawar, Pakistan

ABSTRACT

Robust regression estimators are designed to easily fit contaminated data sets. A common problem associated with these estimators to be completely specified is the proper choice of tuning constant *c*. This choice is somewhat arbitrary and is largely the matter of the personal preference. Several authors suggested different value of '*c'* for various M estimators. We propose K-fold cross-validation for choosing optimal choice of '*c'* to minimize cross-validated absolute Median Predicted Residual. We also propose a resampling variable selection method in robust regression with unusual estimates of prediction error. The study found that the proposed technique is working well.

1. INTRODUCTION

The ordinary least square method is the optimal procedure for fitting linear regression model when the necessary assumptions are fulfilled, see, Draper and smith (1998). But when the regression model does not meet the fundamental assumptions or if the data contain missing observations or outliers, the sample estimates can be misleading and the predictions of the model may become biased, see, Rousseeuw and Leroy (1987), Ho & Naugher, (2000). If the basic assumption of the normality of the residuals is violated that is, the distribution of errors is heavy tailed; the least squares method may not be appropriate, see, Andrews et al., (1972). To deal with long-tail error distribution, one approach is to construct outlier diagnostics, remove the largest residuals as unusual points and still use least square method, see, Zafar and Salahuddin (2011). However, least squares may not be effective if many outliers are present in the data set because of the leave-one-out nature of the outlier test. Accordingly, robust regression procedures provide an alternative and have been introduced to adjust the least squares methods for coping with the outliers to have least influence on the final estimates.

There are several methods of robust regression, such as, M-estimators developed by Huber (1981), least trimmed squares (LTS) developed by Rousseeuw (1984), S-estimators developed by Rousseeuw & Yohai (1984), least median of squares regression (LMS) developed by Rousseeuw (1984) and MM-estimators developed by Yohai(1987). We discuss only M-estimators for regression as these estimators are relatively simple, perform well and are easy to compute. The choice of tuning constant is somewhat arbitrary for these M-estimators and we applied K-fold cross-validation procedure to choose best value of tuning constant for popular and some newly established M-estimators.

Furthermore, an attempt has been made to apply resampling variable selection method in robust regression with alternative estimates of prediction error based on Winsor's principle.

1.1. M-Estimation:

Robust regression procedures are concerned with modification of linear least squares while the distribution of error is not normal, mainly when the errors are heavy-tailed. The term robust regression is to utilize a fitting criterion that is not as susceptible as least squares to unusual observations.

The most generally used method of robust regression is M-estimation. This category of estimators can be considered as a generalization of maximum likelihood estimation, hence the name 'M' estimation.

Considering the linear model,

$$Y = X\beta + \varepsilon$$

And suppose the errors ε_i are independent random variables and follow a double exponential distribution:

$$f(\varepsilon_i) = \frac{1}{2}\sigma\left[\exp\left(-|\varepsilon_i|\right)/\sigma\right]$$

Here the estimates of β are obtained by maximum likelihood method using the likelihood function:

$$L(\beta) = \prod_{i=1}^{n}\frac{1}{2}\sigma\left[\exp\left(-|\varepsilon_i|\right)/\sigma\right]$$

$$= \left(\frac{1}{2}\sigma\right)^{n}\exp\left(-\sum|\varepsilon_i|/\sigma\right)$$

Maximizing the likelihood function would involve minimizing the sum of absolute error that is $\sum|\varepsilon_i|$. Minimizing the sum of absolute error is often called L1-norm regression problem, while least square is called the L2-norm regression problem that minimizes the sum of the square of errors. Thus for heavy-tailed distribution, least square method is no longer the optimum choice and L1-norm regression is quite preferable. For more details on L1-norm regression, see Book et al. (1980), Narula and Wellington (1982), and Dodge (1987).

Similar to L1-norm regression, another approach for heavy-tailed distribution is the use of M-estimation. The basic idea of M-estimations is to minimize some function of residual rather then the sum of the square residuals. The common M-estimator minimizes the objective function:

$$\sum\rho\left(u_i\right) = \sum\rho\left(y_i - x_i^{T}\beta\right)$$

where, ρ is a symmetric function (that is $\rho(u) = \rho(-u)$ for all u) with a sole minimum at zero and provides the contribution of each residual to the objective function. The function

ρ is linked to the likelihood function for a suitable option of error distribution. For example, the least square estimation, $\rho\left(u_i\right)=u_i^2$.

Differentiating the objective function with respect to the coefficients, β and setting the partial derivative equals zero provides k+1 estimating equations for coefficients. Let $\psi=\rho'$ (the derivative of ρ), then

$$\sum\psi\left(y_i-x_i^T\beta\right)x_i^T=0$$

The solution of this equation that minimizes the objective function is called the M-estimator of β. These M-estimators are not necessarily scale invariant that is if the error were multiplied by a constant, the new result may not be the same as the previous one. The scale invariant version of these M-estimates equations:

$$\sum\psi\left(y_i-x_i^T\beta/s\right)x_i^T=0 \qquad (1)$$

where s is a robust estimate of scale and the commonly used scale estimate is the median absolute deviation:

$$s=Med\left|u_i-Med\left(u_i\right)\right|/0.6745$$

where u_i is the residuals of initial fit and the value 0.6745 is an approximately unbiased estimator for large sample from normal distribution.

The weight function is defined $w(u)=\psi(u)/u$, $u=y_i-x_i^T\beta/s$ and let the weights $w_i=w(u_i)$, and then the estimating equations may be written as:

$$\sum w_i\left(y_i-x_i^T\beta\right)x_i^T=0$$

Now solving these estimating equations for M-estimators is a weighted least-square problem that is minimizing $\sum w_i^2u_i^2$. However, the weights depend upon the residuals, the residual depend upon the estimated coefficients, and the estimated coefficients depend upon the weight. Therefore, a convenient computational scheme is the iteratively reweighted least-squares, IRLS method proposed by Holland and Welsch (1977). The scheme consists of the following steps:

1. Choose initial estimate $\beta^{(0)}$, such is least square or regression by medians and calculate residuals and the scale estimate s.
2. Compute residuals $u_i^{(t-1)}$ at each iteration and associated weights $w_i^{(t-1)}=w(u_i^{(t-1)})$ from the prior iteration.
3. Compute the new weighted least square estimates,

$$\beta^{(t)}=\left(X^TW^{(t-1)}X\right)^{-1}X^TW^{(t-1)}Y$$

This iteration procedure is continued until the estimated coefficients converge.

Several M-estimators have been proposed: the familiar least-square estimator (OLS); least absolute residuals (LAR); the Huber estimator; the Tukey biweight (or bisquare) estimator; the Andrews' wave (or sine) estimator and some other recently developed re-descending M-estimators. The objective functions, and the corresponding ψ and weight functions for these M-estimators are given in Table1. Both the Huber objective and the least square functions increase without bound as residual u departs from 0, but the LS objective function increases more quickly. Least –squares assign equal weight to each observation while Huber estimator has a monotone ψ-function, and does not weight large residuals as heavily as least square. Huber estimator is the compromise between least absolute residuals estimators and least-squares estimators.

The Andrews' wave estimator, Tukey's biweight estimator, Qadir objective function, Asad function and Insha's function belongs to the class of re-descending M-estimators because their ψ-functions equal zero for sufficiently large $\vert u\vert$ that is observations having large residuals will receive zero weights.

Table 1: Objective functions, and the consequent ψ and weight functions for various M-estimators

Method	Objective- function $\rho(u)$	ψ –function $\rho^{/}(u)$	Weight- function w(u)	Range of u
LS	$½u^2$	u	1	$\vert u\vert < \infty$
LAR	$\vert u\vert$	sign(u)	sign(u)/u	$\vert u\vert < \infty$
Huber	$½u^2$ $c\vert u\vert-½c^2$	u c sign(u)	1 $c/\vert u\vert$	$\vert u\vert \le c$ $\vert u\vert > c$
Bisquare	$c^2/6[1-\{1-(u/c)^2\}^3]$ $c^2/6$	$u[1-(u/c)^2]^2$ 0	$[1-(u/c)^2]^2$ 0	$\vert u\vert \le c$ $\vert u\vert > c$
Andrews	$c^2[1-\cos(u/c)]$ $2\ c^2$	c sin(u/c) 0	sin(u/c)/(u/c) 0	$\vert u\vert \le c\pi$ $\vert u\vert > c\pi$
Qadir	$u^2/96c^4(3c^4-3c^2u^2+u^4)$ $c^2/96$	$u/16c^4(c+u)^2(c-u)^2$ 0	$1/16c^4(c+u)^2(c-u)^2$ 0	$\vert u\vert \le c$ $\vert u\vert > c$
Asad	$u^2/45c^8(3u^8-10c^4u^4+15c^{8)}$ $8c^2/45$	$2u/3[1-(u/c)^4]^2$ 0	$2/3[1-(u/c)^4]^2$ 0	$\vert u\vert \le c$ $\vert u\vert > c$
Insha	$c^2/4[\text{Arc tan}(u/c)^2 + c^2u^2/c^4+u^4]$	$u[1+(u/c)^4]^{-2}$	$[1+(u/c)^4]^{-2}$	$\vert u\vert \ge 0$

The constant '*c*' in these re-descending M-estimators is usually called tuning constant and determine the properties of the associated estimators (such as efficiency, influence function, and gross-error sensitivity). Smaller values of '*c*' generate more resistance to outlier, but at the cost of lower efficiency when the errors are distributed normally.

The choice of the tuning constant is usually selected arbitrarily and largely is the matter of personal preference. Several authors suggested different value of tuning constant *c* for various M estimators. For Wave function Andrew (1974) uses *c*=1.5, Gross (1976) suggest *c*=1.8 & 2.4, Hogg (1979) uses values of *c*=1.5, 2.0 and Rey (1983) suggests *c*=1.3387.

We used K-fold cross-validation procedure for choosing best tuning constant in these re-descending M-estimators to give reasonably high efficiency.

2. RESAMPLING CHOICE OF TUNING CONSTANT IN ROBUST REGRESSION

The robust regression estimators of re-descending M-estimations involve the use of ψ –function which replaces the derivative of the square function of the LS estimator. This ψ –function is not completely specified and needs the choice of tuning constant. Kelly (1996) reports the results of simulation study to investigate tuning constant c which minimize the jackknife asymptotic mean-squared error of the estimators. Yohai (1974) considered a class of error distribution in the linear regression model and showed how to choose *c* for Huber's robust regression estimator so that the resulting estimator was minimax over the class of error distribution. Salahuddin (1990) used leave-one-out cross-validation (LOOCV) to choose an optimal value of tuning constant for Andrews' wave estimator.

We propose a resampling strategy of applying K-fold cross-validation method to choose that value of tuning constant that minimizes the Median Predicted Residual (MedPR).

Let $\hat{\beta}_{(i)}$ be a robust estimate of β obtained from the data with the nth group of data eliminated. Then the best value of tuning constant $\{c = 1.1\ (0.1)3.0(0.2)5.0\}$ is when;

$$\text{Minimum (MedPR)} = Med\left|Y_i - \hat{Y}_{(i)}\right|$$
$$= Med\left|Y_i - X_i^T\hat{\beta}_{(i)}\right|$$

where Med. stands for median, $X_i^T\hat{\beta}_{(i)}$ is the predicted value of observation(s) Y_i deleted from the data set. The strategy is to choose that value of '*c*' which minimizes the median predicted residual instead of PRESS in linear and generalized regression.

Solving the estimating equations for M-estimators, we start using iteratively re-weighted least-squares (IRLS) to obtain a robust fit. To initiate IRLS, we need a resistant fit (that is least-squares residuals, LSR or least absolute residuals, LAR) to compute the preliminary fit. However, these two methods do not protect against high-leverage observations and a robust regression might have difficulty in recovering from poor preliminary fit. Therefore, we need a suitable resistant fit to arrive at a good robust fit. Andrews (1974) developed a robust method to provide preliminary fit called as regression by medians. One will prefer this method as a source of initial estimates and residuals as it converge in fever iteration and suffer less computational cost.

For simple regression the review of Andrews' procedure can be described as follows:

1. Arrange the X-values of the data in ascending order.
2. Remove a certain number say np_1 of the smallest and largest X- values.
3. Remove further a certain number say np_2 of X-values immediately above and below the median.
4. Compute the medians (say $Med.X_L$ and $Med.X_H$) of the two remaining subsets corresponding to the lower and higher values of X.

5. Also compute the medians (say $Med.Y_L$ and $Med.Y_H$) for the corresponding Y-values.
6. The slope of the fit is than computed as:

$$\hat{\beta} = Med.Y_H - Med.Y_{L\,/}\;\; Med.X_H - Med.X_L$$

This estimator has a high breakdown point because half of the data on either subset can determine the fitted line. We use np1+np2 ≈ 25% of the total number of observations.

Andrew generalized this procedure to a multiple regression case by applying a sweep operator to predictors successively and then to the outcome variable. Suppose we consider a multiple regression model with three predictor variables, X1, X2 and X3, then the procedure can be summarized as follows:

The predictor variable X1 is used to modify the variables X2, X3 and Y by sweeping X1 out of these variables respectively:

$$X2.1 = X2 - \hat{\beta}_1\ X1$$

$$X3.1 = X3 - \hat{\beta}_2\ X1$$

$$Y.1 \ = Y \ - \hat{\beta}_3\ X1$$

Then the predictor variable X2 is used to modify the variables X3 and Y by sweeping X2 out of these variables respectively:

$$X3.12 = X3.1 - \hat{\beta}_4\ X2.1$$

$$Y.12 \ = Y.1 \ - \hat{\beta}_5\ X2.1$$

And the predictor variable X3 is used to modify the variable Y by sweeping X3 out of Y:

$$Y.123 \ = Y.12 \ - \hat{\beta}_6\ X3.12$$

The subscript after a dot shows list of integers indicating that these predictor variables have been swept out and β 's the estimates of slope. The procedure proceeds for more predictor variables in the regression models in similar fashion. The procedure may be iterated and the total number of iteration is (p/2+2), where p is the number of predictor variables. At each iteration, the procedure is applied to a set of adjusted variables obtained at a previous iteration. The procedure is used to compute the initial residuals for robust fit. The above procedure is used to compute a set of residuals and initiate the iterative procedure.

K-fold cross-validation is used to choose the tuning constant *c* which minimizes the median of the absolute prediction. The selected value of *c* is the best choice to compute robust estimates from the full data-set.

The Algorithm:

Our proposed algorithm consists of the following steps.

1. Generate or read input data for m-rows (observations) and n = p+1 columns (predictors and a response variable).
2. Initiate tuning constant $c = 1.1$ and its increment $\Delta c = 0.01$

3. Run cross-validation procedure by removing a group of observations.
4. To initiate IRLS, run least-squares method or least absolute residuals or Andrews' method of regression by medians to compute the preliminary fit. Use the residuals from this preliminary fit to compute scale estimate, s. Our algorithm runs Andrews' method of regression by medians for initial fit.
5. Run any weight function of re-descending M-estimators and compute weights W_i for residuals. Also calculate the weighted data of variables, their sum, means, and the weighted sum of squares and cross-products by matrix X^TWX, W is the diagonal matrix of order m-1 * m -1 and X is a matrix of order m-1 * n. Then compute the weighted residual sum-of-squares and cross-products from the above quantities and normalize to simple correlation matrix.
6. Compute standard regression coefficients, transfer back into original units and calculate residuals. Then compute new weights while using the calculated residuals.
7. Return to step 5 and repeat the whole procedure. This iteration procedure is continued until the estimated coefficients converge. The process terminates once the maximum change in the coefficients from one step to another is less than 0.1% or the number of iteration exceed 20. Fit the resultant robust estimate and compute the predicted residuals.
8. Return to step 4 and repeat the whole procedure by deleting instead group 2, group 3, and so on until all groups have been deleted once. Compute the median of the absolute values of the predicted residual.
9. Return to step 3 and increment the tuning constant c accordingly $\{c = 1.1 (0.1)3.0(0.2)5.0\}$. Repeat the whole procedure and choose the best value of c, the one that gives the smallest median predication error.
10. Use the selected value of tuning constant for the respective weight function and fit the robust regression for the complete data set.

(A computer program in R for this algorithm can be provided on demand from principle author).

Model Selection in Robust Regression:

Many robust regression procedures have been proposed in the last 3 decades but little guidance is available for model selection in robust regression. Usually, the variable selection methods in robust regression are based on robust version of the general linear test that utilizes the asymptotic covariance matrix; see Hampel et al. (1986). Hertier and Ronehetti (1994) and Markatou and He (1994) suggested the Wald test (analogous to t-test) and drop-in-dispersion tests (analogous to F-test) to generalized M and compound estimators. Field and Welsh (1998) and Field (1997) proposed saddle point estimatation of tail area probabilities for testing robust regression hypothesis as improvement to the asymptotic approach. Ronchetti and Staudte (1994) proposed a robust edition of Mallow's Cp. There technique multiplies the squared residuals by the final weights from a robust regression and two supplementary measures are also added to the residual sum of squares that are the selected robust estimator and a function of the number of parameters. The robust Cp seems working well for their examples, but they have not confirmed the results by any simulation study.

Davison and Hinkley (1997) discussed the applications of resampling techniques in robust regression. They suggested removing gross outliers from the analysis since excessive outliers could appear in the resample data leading to inefficiency and breakdown and then applying any of the least squares prediction error method to robust regression. Wilcox (1998) proposed a bootstrap resampling scheme for the selection of predictor variables in robust regression. He used a bootstrap approach based on percentile to find critical values for the joint confidence region on the Mahalanobis distance for the model parameters. Wilcox stated that there is a room for improvement with this technique since the probability of Type 1st error can be considerably smaller than the nominal levels in many situations. He reported that this method does not work well with least squares and hence correction factors through simulation are essential to get the right coverage probabilities.

Wisnowski et al. (2003) provide a new resampling variable selection method by calculating alternative estimates of prediction error. They proposed, relaxing the absolute minimum prediction error condition and choosing a regression model with the smallest number of predictor variables and a smaller (not essentially minimum) prediction error. They are not advocating using a specific percentage for the suggested change in prediction errors to be used for variable selection in robust regression.

It is clear from the above review that little guidance is available for variable selection in robust regression especially applying cross validation or bootstrap estimates of prediction error in these regression models.

A Proposed Variable Selection Criterion in Robust Regression

There are many variable selection methods in regression analysis to select best possible subset regression model. The most common are the automated model selection methods (forward, backwad, stepwise), subset regression and all possible regression. These variable selection methods are based on R^2, Adjusted R^2, F test statistics (F-to-enter and F-to-remove), AIC, Akaike (1973), C_p, Mallows (1973), and BIC, Schwarz (1978)**.** But these methods based on least square regression parameters lose power in the presence of unusual points. Breiman (1995) preferred some measure of prediction error for variable selection in regression based on resampling methods of cross validation and bootstrapping. Cross validation and bootstrap methods are well reputed in least-square variables selection but cannot be directly applied to contaminated data sets using robust regression models. Wisnowski et al. (2003) proposed, relaxing the requirement for the absolute minimum prediction error and selecting a model with the fewest number of predictor variables and a low (not necessarily minimum) prediction error.

We applied a new resampling variable selection method by introducing alternative estimates of prediction error based on Winsor's principle.

The cross-validatory scheme for choosing subset predictor variables in robust regression is as follows:

Let $\hat{\beta}_{(i)}$ be a robust estimate of parameters obtained from data with the ith observation deleted, then a subset predictor variable(s) in robust regression is a good choice that corresponds to minimum occurrence of Winsorized Prediction Error (WPE) computed as the average of Winsorized Prediction Sum of Squares ($\text{PRESS}_{\text{Winsor}}$).

$$\text{PRESS}_{\text{Winsor}} = \sum\left[\left(Y_i - \hat{Y}_{(i)}\right)^2_{Winsor}\right]$$

$$\text{WPE} = \text{PRESS}_{\text{Winsor}}/\text{n}$$

The idea is to choose the subset model that produces the near minimum Winsorized PRESS and then a final robust model with the selected predictors variables are fitted for full data set. The suggested algorithm is given below,

1. Read input data consisting of n-rows (data points) and m-columns (predictors and a response variable).
2. Apply the Cross-validation to choose appropriate tuning constant value for any M-estimation robust regression (see section 2).
3. Initialize cross-validation (LOOCV) by deleting first case or K-fold cross validation (KFCV) by deleting first group of observation.
4. Run a robust regression for a subset of predictor variables using the selected tuning constant value and compute the predicted residual for the deleted case.
5. Return to step 3 and repeat the whole procedure by deleting instead case 2, case 3 and so on or group 2, group 3 and so on until all the observations has been deleted once and calculate Winsorized predicted residuals sum of squares ($\text{PRESS}_{\text{Winsor}}$).
6. Compute the Winsorized Predicted Error (WPE).
7. Return to step 3 and repeat the procedure for all possible subset of predictor variables.
8. Select the subset of predictor variables that produces the minimum WPE. For superior results we further suggest relaxing the requirement for strict minimum WPE and selecting a model with a fewest number of predictor variables corresponding to near minimum WPE. A reasonable strategy involves observing a line connected scatter plot (screeplot), the subset model with the fewest number of predictor variables where the screeplot levels off is selected.
9. Use the selected subset of predictor variables and compute the robust estimates for the complete data set.

(A computer program in R for this algorithm can be provided on demand from principle author).

3. BROWNLEE'S STACK LOSS PLANT DATA

This is an experimental data of a plant for the oxidation of ammonia to nitric acid. The data frame consisting of 21 observations on 4 variables. The Stack Loss data set has the initial three predictor variables, X1 (Flow of cooling air), X2 (Cooling water inlet temperature), X3 (Concentration of acid) and a response variable Y (Stack loss). The air flow represents the rate of operation of the plant, water temperature represents the cooling temperature of water circulated through coils in the absorption tower and the concentration of acid represents the acid concentration circulating, minus 50, times 10: that is, 89 correspond to 58.9 percent acid. The dependent variable Stack loss is 10 times the percentage of the incoming ammonia to the plant that flees from the absorption column unabsorbed; i.e., an inverse measure of the over-all efficiency of the plant, see, Brownlee, K. A. (1960)

We analyzed this data by applying K-fold cross-validation resampling technique in robust regression for choosing the appropriate tuning constant for various M-estimators. Table 2 represents the median prediction residuals for various tuning constant values resulting from K-fold cross-validation technique for Andrew's and Tukey's type M-estimators. Table 2 shows the summary of the resultant robust regression models and Table 3 represents the residuals for various fit.

Table 2: Cross validated tuning constants and their corresponding Median Predicted Residuals (Med. PR) for Andrews and Bisquare estimators resulting from K-fold CV

Tune. C	**Andrews**		**Bisquare**	
	Resistant fit for initial residuals		**Resistant fit for initial residuals**	
	Med	**LS**	**Med**	**LS**
1.10	.9263	1.9026	2.7654	2.0593
1.20	.9235	3.0018	2.8323	2.0438
1.30	1.2452	3.0577	2.9254	2.1454
1.40	1.3267	3.1049	2.9624	2.1047
1.50	1.3711	3.2163	2.5921	1.9431
1.60	1.4013	3.3016	2.6741	2.8433
1.70	1.4233	3.3849	2.3576	2.8016
1.80	1.4402	3.4854	2.1028	2.7649
1.90	1.4537	2.7804	2.0323	2.7367
2.00	1.4646	2.8314	1.9924	1.9582
2.10	1.4736	2.8557	1.9666	1.9635
2.20	1.4812	2.8696	1.9486	1.9428
2.30	1.4876	2.8809	1.9352	1.5745
2.40	1.4932	2.8905	1.9251	1.5696
2.50	1.4980	2.8985	1.9170	1.5650
2.60	1.5021	2.9053	1.9106	1.5608
2.70	1.5058	2.9112	1.9053	1.7241
2.80	1.5090	2.9162	1.9010	1.7475
2.90	1.5119	2.9207	1.8976	1.8614
3.00	1.5145	2.9246	1.8950	1.8569
3.20	1.5189	2.9389	.8704	1.8497
3.40	1.5224	2.9511	.8870	2.0459
3.60	1.5254	2.9613	.9003	2.0673
3.80	1.5279	2.9698	.9111	2.0897
4.00	1.5300	2.9770	.9200	2.8165
4.20	1.5318	2.9832	.9260	2.8919
4.40	1.5334	2.9885	.9245	2.9546
4.60	1.5347	2.9931	1.1909	3.5717
4.80	1.5359	2.9971	1.1820	3.6878
5.00	1.5369	3.0007	1.2594	3.7053

The procedure selects c = 1.20, 1.10 corresponding to the minimum Med. PR = 0.9235, 1.9026 respectively when regression by median and least square method is used for initial fit in Andrews estimator. For Tukey's estimator the procedure selects c = 2.60, 3.120 corresponding to the minimum Med. PR = 0.8704, 1.5608 when regression by median and least square method is used for initial fit. The resultant robust regression model for Andrews procedure on the basis of selected c = 1.20 is exactly the same equation found by Andrews (1974) using c = 1.5 and different value of np_1 and np_2 for obtaining the initial fit. We used np_1 = 3 and np_2 = 3 (total of which is approximately 30% of the full data set) to obtain the initial fit using method of regression by medians. Our program than uses the selected tuning constant values on the whole data set and fit the robust regression model. It is clear from the residuals in Table 4 that four observations 1, 3, 4 and 21 have large residuals receiving zero weights by each M-estimator we described.

Table 3:
Robust regression models for various M-estimators using the selected tuning constant by our procedure to the stack-loss data.

M-estimator	Initial fit	Model Coefficients				Tuning Constant
		Intercept	X1	X2	X3	
Andrew's	Med	-37.1590	0.8172	0.5225	-0.0723	1.20
	LS	-37.4178	0.7548	0.7654	-0.0849	1.10
Bisquare	Med	-37.0661	0.8208	0.5127	-0. 0733	3.00
	LS	-36.7170	0.8490	0.4336	-0.0764	2.60
Qadir	Med	-37.2153	0.8148	0.5291	-0.0717	3.40
	LS	-36.7170	0.8491	0.4336	-0.0764	2.60
Asad	Med	-37.4913	0.8023	0.5645	-0.0689	3.20
	LS	-35.6166	0.8464	0.4434	-0.0897	1.80
LS(full data)	-	-39.9197	0.7156	1.2953	-0.1521	-
LS(reduced data)	-	-37.6525	0.7977	0.5773	-0.0671	-

For the purpose of comparison, we have computed the least square estimator for both full and reduced data sets as shown in Table 3. For reduced data set that is without observations 1, 3, 4 and 21 a least square fit is,

$$\hat{Y} = -37.6525 + 0.7977X_1 + 0.5773X_2 - 0.0671X_3$$

Comparing model coefficients in Table 3, we see that all the robust regression model coefficients has close agreement with the coefficients of least square fit without unusual observations.

Table 4:
Residuals for various M-estimators using the selected tuning constant by our procedure to the stack-loss data.

Cases	Residuals							
	M-Estimators (Initial fit by Med)				M-Estimators (Initial fit by LS)			
	Andrew	Tukey	Qadir	Asad	Andrew	Tukey	Qadir	Asad
1	6.1062	6.0831	6.1216	6.1912	5.9258	5.8841	5.8841	5.9187
2	1.0339	1.0098	1.0499	1.1223	0.8409	0.8077	0.8077	0.8289
3	6.3095	6.2857	6.3255	6.4008	6.3158	6.0729	6.0729	6.1272
4	8.2388	8.2487	8.2321	8.1892	7.6391	8.3152	8.3152	8.3048
5	-0.7161	-0.7259	-0.7096	-0.6818	-0.8300	-0.8177	-0.8177	-0.8085
6	-1.2386	-1.2386	-1.2388	-1.2463	-1.5954	-1.2512	-1.2512	-1.2519
7	-0.3276	-0.3116	-0.3380	-0.3976	-0.8511	-0.2264	-0.2264	-0.1569
8	0.6724	0.6884	0.6620	0.6024	0.1488	0.7735	0.7735	0.8431
9	-0.9698	-0.9555	-0.9796	-1.0369	-1.5761	-0.8549	-0.8549	-0.8662
10	0.1370	0.0950	0.1646	0.3035	0.6564	-0.2218	-0.2218	-0.2775
11	0 .7874	0.7545	0.8095	0.9233	1.4209	0.4657	0.4657	0.5300
12	0 .2377	0.1939	0.2670	0.4190	1.1014	-0.1771	-0.1771	-0.1164
13	-2.7184	-2.7584	-2.6921	-2.5587	-2.1737	-3.0691	-3.0691	-3.0981
14	-1.4461	-1.4650	-1.4331	-1.3657	-1.0047	-1.6623	-1.6623	-1.5544
15	1.3250	1.3208	1.3278	1.3421	1.4596	1.2582	1.2582	1.3013
16	0.1082	0.1009	0.1129	0.1355	0.2047	0.0290	0.0290	0.0321
17	-0.4261	-0.4377	-.4194	-0.3932	-0.7501	-0.4740	-0.4740	-0.6674
18	0.0798	0.0753	0.0822	0.0889	-0.1554	0.0607	0.0607	-0.0393
19	0.6295	0.6359	0.6247	0.5933	0.1641	0.7035	0.7035	0.6071
20	1.8709	1.8577	1.8792	1.9169	1.8050	1.7619	1.7619	1.7080
21	-8.9195	-8.9737	-8.8831	-8.6961	-7.9979	-9.4375	-9.4375	-9.3342

It is clear from the residuals of these robust regression models in Table 4 when regression by median is used as initial fit that cases 1, 3, 4 and 21 have large residuals (see columns 2, 3, 4 & 5 of Table 4). Thus the robust regression procedure with the K-fold cross-validated choice of tuning constant successfully identify these four unusual points by leaving their residuals much larger and hence receiving zero weights. Starting with the ordinary least square as an initial fit, Andrews estimator shows large residuals for the four unusual points (see column 6 of Table 4) while Tukey's, Qadir and Asad estimators have large residuals (>3) for cases 1, 3, 4, 13 and 21 see columns 7, 8 & 9 of Table 4) but we find cases 1, 3, 4 and 21 are receiving zero weights while case 13 does not (that is wt > 0.0 for case 13).

Table 5:
Cross-validated Wensorized Prediction Error (WPE) as a function of the number of variable(s) in the model using Tukey's bi-square estimator for Stack loss data

Model Order	Model Parameters	WPE	
		LOOCV	KFCV
1	β_0	140.670	149.916
2	β_1	3.316	4.332
3	β_2	71.686	80.471
4	β_3	8.692	16.846
5	$\beta_1\beta_2$	1.014	3.193
6	$\beta_1\beta_3$	5.333	5.988
7	$\beta_2\beta_3$	37.251	80.755
8	$\beta_1\beta_2\beta_3$	5.545	4.962

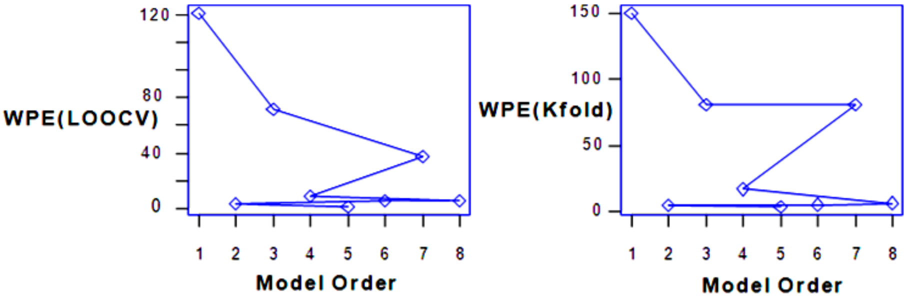


Figure 1: Screeplot of Winsorized Prediction Error

The procedure selects the subset model having predictor variables X_1 and X_2 that correspond to minimum Winsorized Prediction Error, WPE (=1.041 using LOOCV and =3.193 using KFVV). We believe that superior results for robust model selection are obtainable by relaxing the requisition for the absolute smallest Winsorised prediction error. The strategy is to observe a screeplot (line connected scatter plot of subset model order versus Winsorized prediction error as shown in Figure 1. The subset model with fewest number of predictor variable(s) that is X_1 corresponding to near minimum WPE, where the screeplot levels of, is selected. The final robust regression for full data is,

$$\hat{Y} = -39.88 + 0.95X_1 \qquad (8.1)$$

For the reduced data set that is without observations 1, 3, 4 and 21 a least square fit for predictor variable X_1 in the model yields,

$$\hat{Y} = -40.03 + 0.95X_1 \qquad (8.2)$$

Comparing equations 8.1 and 8.2, we see that these two models are quite close.

The Simulation Experiment

For the better understanding of the proposed resampling method performance, we run a design experiment approach by means of Monte Carlo simulation. In this study, the data sets consist of n = 40 data points and p = 5 parameters as extensively used in Shao (1993, 1996) and Wisnowski et al. (2003). The response variable is generated as $Y = Z\beta + \varepsilon$, where $Z \overset{iid}{\sim} N(0,1)_{40}$, β is the vector of the recognized parameters [2, 3, 6, 0, 0] and $\varepsilon \overset{iid}{\sim} N(0,1)$. A value of 10 and/or 15 is added to create outliers for the last 4 or 8 observations. The data sets thus contain 10% and 20% residuals outliers at a distance of 10σ and /or 15σ. From the previous knowledge and pilot studies, the subsequent factors are included.

- Percentage of Outliers. This important factor reveals the number of outliers in the data set. In our generated data sets, the outliers' density levels are 10% and 20%.
- Outlying Distance. This factor shows how many standard deviation (SD) from the means make the outlier in residuals. Our generated data sets contain residuals outliers at distance of 10 standard deviations and 15 standard deviations.
- Cross-validation Assessment Size. This factor is most important factor to correctly identify important predictor variables. The levels are 1 (leave-one-out cross-validation), 5 (K-fold) cross-validation.

To illustrate the methodology, the winsorized predicted error for models with increasing number of predictor variables are observed. The cross-validation design is a full factorial 2^3. The results in Table 6 are the proportions of times that each model is selected out of 350 replicates. It is important to mention that incredibly little supplementary information is obtained if the number of replicates is increased. The shaded column in Table 6 reports the proportion of times that each cross-validation method using various criterions is able to find the correct model. The most striking results are the common failure of the minimum prediction error criterion in the majority of cases. It wrongly selects the largest roust regression model (p=5) in most of the cases instead of selecting the correct model. The proposed minimum Winsorized prediction error criterion and that of near minimum Winsorized prediction error criterion significantly outperforms the minimum prediction error criterion in robust regression. obviously, the best selection method is the near minimum Winsorized prediction error applied from the K-fold cross-validation.

Table 6:

Design experiment results for cross-validation methods using Winsorized residuals to compute prediction error in robust regression models. The top values in each cell of last four columns are the proportion of time that a model is chosen out of 350 replications by means of minimum prediction error applying least square method. The middle values are the proportion chosen using minimum Winsorized prediction error in robust regression models and the bottom values are the proportion chosen using near minimum Winsorized prediction error in robust regression models

% Outliers	**Outlier Distance**	**CV Size**	$\beta_0 - \beta_1$	$\beta_0 - \beta_2$	$\beta_0 - \beta_3$	$\beta_0 - \beta_4$
10	10	1	0.000 0.000 0.000	0.291 0.623 0.674	0.120 0.106 0.294	0.589 0.271 0.026
20	10	1	0.000 0.000 0.000	0.406 0.529 0.649	0.100 0.274 0.177	0.494 0.194 0.174
10	15	1	0.000 0.000 0.000	0.306 0.497 0.720	0.097 0.229 0.180	0.597 0.274 0.100
20	15	1	0.000 0.000 0.000	0.205 0.520 0.723	0.389 0.089 0.034	0.406 0.391 0.243
10	10	5	0.000 0.000 0.000	0.377 0.723 0.891	0.126 0.100 0.083	0.497 0.177 0.026
20	10	5	0.000 0.000 0.000	0.591 0.631 0.783	0.194 0.363 0.211	0.215 0.005 0.005
10	15	5	0.000 0.000 0.000	0.217 0.637 0.780	0.294 0.094 0.103	0.489 0.269 0.117
20	15	5	0.000 0.000 0.000	0.294 0.623 0.729	0.277 0.106 0.069	0.429 0.271 0.203

CONCLUSION

Linear least square estimates can behave badly for contaminated data sets. One remedy is to detect and remove unusual observations from the least square fit. Another approach, termed robust regression, designed to reduce the impact of unusual observations by reducing the weights given to large residuals. This can be done by the most common general method of robust egression called M-estimations. Several M-estimators have been proposed (see Table 1), each require the analyst to specify some tuning constant. In practice, the choice of the tuning constant for these robust regressions is somewhat arbitrary and do not adapt to each particular data set. We have applied the efficient K-fold cross-validation in a robust way by computing the Median Predicted Residual (MedPR) to choose tuning constant for each data set.

We have analyzed the well known Brownlee's Stack Loss Plant Data by applying K-fold resampling technique to choose the tuning constant and select subset model in robust regressions. We find that K-fold cross-validation in robust regression has led to an appropriate value of tuning constant which deletes the unusual points by assigning them zero weights

Comparing model coefficients in Table 3, we find that all the robust regression estimates has close agreement with the coefficients of least square fit when unusual observations are deleted from the data set.

The study also proposes a resampling variable selection strategy by establishing alternative estimates of prediction error based on Winsor’s principle for contaminated data sets. We suggest that, even though robust estimation and resampling techniques are computationally complex procedures, we can combine these procedures for better results. We find that a subset predictor variable(s) in robust regression is a good choice that corresponds to minimum occurrence of Winsorized Prediction Error (WPE) computed as the average of Winsorized Prediction Sum of Squares ($\text{PRESS}_{\text{Winsor}}$). We further suggest that better results are obtainable by relaxing the condition for strict minimum WPE and selecting a model with a smallest number of predictor variables corresponding to near minimum WPE. A reasonable plan involves examining a line connected scatter plot (screeplot), the subset model with the smallest number of predictor variables where the screeplot levels off is chosen.

The simulation experiment clearly reveals that decent results are possible with the resampling method using alternative estimate of prediction error in robust regression. We see that the proposed resampling criterion often outperform the minimum prediction error criterion in contaminated data sets.

REFERENCES

1. Andrews, D.F. (1974). A Robust Method for Multiple Linear Regressions. *Tachometrics*, 16, 523-531.
2. Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimation of Location: Survey and Advances. Princeton*, New Jersey: Princeton University Press.

3. Book, D., Booker, J., Hartley, H.O., and Sielken, R.I. (1980). Unbiased L1 Estimators and Their Covariance. *ONR THEMIS Technical Report* No. 64, Institute of Statistics, Texas A & M University.
4. Breiman, L., (1995). Better Subset Regression Using the Nonnegative Garrote. *Tachometrics*, 37, 373-384
5. Brownlee, K.A. (1960). *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley. (2nd ed. 1965) 491-500.
6. Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Method and Their Application*. Cambridge University Press, Cambridge, UK.
7. Draper, N.R. and Smith, H. (1998). Applied Regression Analysis (3rd ed.), *John Wiley,* New York.
8. Dodge, Y. (1987). Statistical Data Analysis Based on the L_1-Norm and Related Methods. North-Holland, Amsterdam.
9. Field, C.A. (1997). Robust Regression and Small Samples Confidence Intervals. J. Statist. Plann. Inference, 57, 39-48.
10. Field, C.A. and Welsh, A.H. (1998). Robust Regression Confidence Intervals for Regression Parameters. *Austral. N.Z.J. Statist.*, 40, 55-64.
11. Gross. A.M. (1976). Confidence Interval Robustness with Long-Tailed Symmetric Distributions. *J. Amer. Statist. Assoc.,* 71, 409-416.
12. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.
13. Hertier, S., and Ronchetti, E.M. (1994). Robust Bounded-Influence Tests in General Parametric Models. *J. Amer. Statist. Assoc.,* 89, 897-904.
14. Hogg, R.V. (1979). Statistical Robustness: One View of Its Use in Application Today. *The American Statisticians*, 33, 108-115.
15. Huber, P.J. (1981). *Robust Statistics*. New York, Wiley.
16. Holland, P.J. and Welsch, R.E., (1977). Robust Regression Using Iteratively Reweighted Least Square. *Communications in Statistics*, A6, 813-827.
17. Ho, K. and Naugher, J. (2000). Outliers Lie: An illustrative Example of Identifying Outliers and Applying Robust Models. *Multiple Linear Regression Viewpoints*, 26(2), 2-6.
18. Kelly, G.E. (1992). Robust Regression Estimators-The Choice of Tuning Constants. *The Statistician*, 41, 303-314.
19. Kelly, G. (1996). Adaptive Choice of Tuning Constant for Robust Regression Estimators. *The Statisticians*, 45, 35-40.
20. Markatou, M., He, X., (1994). Bounded Influence and High Breakdown Point Testing Procedures in Linear Models. *J. Amer. Statist. Assoc.,* 89, 543-549.
21. Narula, S.C. and Wellington, J.F. (1982). The Minimum Sum of Absolute Error Regression: A State of the Art Survey. *International Statistical Review*, 50, 317-326.
22. Salahuddin (1990). Cross-Validation in Regression Analysis. Ph.D. Thesis, University of Wales, Swansea.
23. Rey, W.J.J. (1993). *Introduction to Robust and Quasi-Robust Statistical Methods*. Belin, Springer-Verlag.
24. Ronchetti, E. and Staudte, R.G. (1994). A Robust Version of Mallows's C_p. *J. Amer. Statist. Assoc.,* 89, 550-559.

25. Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York, John Wiley.
26. Rousseeuw, P.J. (1984). Least Median of Square Regression. *J. Amer. Statist. Assoc.,* 79, 871-880.
27. Rousseeuw, P.J. and Yohai, V.J. (1984). Robust Regression by Means of S-Estimators: in Franke, J., Hardle, W. and Martin, R.D. (Eds.), *Robust and Nonlinear Time Series Analysis: Lecture Notes in Statistics*. Springe, Berlin, 26, 256-272.
28. Wisnoski, J.W., Simpson, J.R., Montgomery, D.C., and Runger, G.C., (2003). Resampling Method for Variable Selection in Robust Regression. Compt. Statist. *Data Analysis*, 43, 341-355.
29. Yohai, V.J. (1974). Robust Estimation in the Linear Model. *Ann. Statist.*, 2, 562-567.
30. Yohai, V.J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. *Ann. Statist*., 15(2), 642-656.
31. Zafar M. and Salahuddin, (2011). A Cross-Validation Approach to Optimize Unequal Cutoff Values in Stepwise Regression. *Pak. J. Satatis*., 27(2), 197-211.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 427-436

WHICH SECTOR IS EASILY ACCESSIBLE FOR FEMALE AS A SOURCE OF INCOME?

Mariam Abbas Soharwardi
Department of Economics, The Islamia University of Bahawalpur,
Bahawalpur, Pakistan. Email: ma_eco@hotmail.com

ABSTRACT

This research paper is investigating the determinants of the female earnings in the formal and informal sector. It also highlights which sector is easily accessible for female as a source of income. A comparative analysis of the female earnings between formal and informal sector have been done which shows that informal sector is more appropriate as a source of income for females. The data in this research paper has been collected from Ali Pur through questionnaires. One thousand women have been interviewed 500 form formal sector and five hundred from informal sector and then a comparison has been made considering the factors which affects the female earnings in formal and informal sectors. This comparison has been done the help of econometric models consisting of all the variables that affect the female earnings in both sectors whether they affect positively or negatively.

INTRODUCTION

Significance for the development of women, the labors force participation rate (LFPR) of women has remained substantially lower than that of man's in the world such that only 60 women per 100 men participating in the labor force in 100. The LFPR of women varies widely many countries as well. In 2001, in LFPR of women while staying below 30% in countries like Oman, Malta, Belize was above 60% in countries like Iceland, Sweden and Canada, north Cyprus is one of the countries having a relatively lower rate of labors force participation for women. Social policies can help to reduce the incompatibility between labors market participation and child care therefore induces higher female labors participation rates. In a cross-country perspective, this type of studies implies that women in different European countries have the same preference, but face different possibilities due to different social policies in the countries, leading to different choices. The first important social dimension is that women are becoming more integrated into formal production. Although women lag behind men considerably in terms of employment and wages, like in many countries, at least they are catching up. Second an important economic dimension is that an increase in the female contribution to formal production leads to higher economic growth. The high growth rates of the Dutch economy at the end of 1990s can be attributed to the substantial increase in female participation. Third, the development has a fiscal and demographic dimension as well. It is widely believed that an increase in the participation rate contributes to the fiscal sustainability of the welfare state which is under pressure due to the again of society. Women labor force participation rate in Pakistan, according to old data collection technique was exceptionally low at just 14.4%, as compared to 70.3% for men, while unemployment rate was 16.5% for women and 6.7% for men (FBS 2003, PP.15, 30) the share of women's earnings in earned income of household was 26% of that of men earnings while their economic activity rate as percentage of that of men was 40% (MHDC, 2000) According to revised data collection technique of federal Bureau of statistics, women's participation rate has been increased to 50% instead of 14.4% in 2003. According to revised data collection, if a women is involved doing work such as harvesting, sowing seeds, cotton-picking, maize and rice husking, livestock and poultry breeding, agricultural forming activities, construction work, collection of fire-wood and cotton-sticks, fetching water, making clothes, sewing, knitting, marketing and preparation of goods and materials, then she will be included in labor force. It explained that informally employed women have increased the labor force participation rate of women up to 50%. Informal sector employment is generally a larger source of employment for women than for men in the developing world. In the developing countries, 60% or more women workers are informally employed (outside agriculture), through in Asia the proportion of women and men is roughly equivalent.

THE COST OF BEING FORMAL AND THE COST OF NOT BEING FORMAL

In order for an informal sector to all exist, there need to be benefits greater than disadvantages when operating informally rather than conduct a proper official registration. In countries where female firms are of minor importance to the economy, the cost of becoming formal is obviously too high. Registration processes and the like are too complicated and, on many occasion the demanded amount of financial capital is unreasonably large for a small scale entrepreneur. Once the process of becoming a formal firm is passed, it must be profitable to stay so. Taxation on production, payroll tax, social insurance etc, must be set on a reasonable level further it must not be too complicated to follow laws and regulation also infrastructure must be at a satisfactory level or production and transportation costs will probably be unreasonably high. Even though there might be a great effort to start up and run an official firm, there are of course great benefits as well.

FORMAL SECTOR

All those types of employment which offer regular wages and hours, which carry with them employments rights and on which income tax is paid is called the formal sector. It is a sector which encompasses all jobs with normal hours and regular wages and is recognized as income sources on which income taxes must be paid which is opposite of informal sector. The employment sector, comprising 'proper' jobs that are usually permanent, with set hours of work, agreed levels of pay and sometimes pensions and social security rights.

INFORMAL SECTOR

It is a sector which encompasses all jobs which are not recognized as normal income sources, and on which taxes are not paid. The term is sometimes used to refer to only illegal activity, such as an individual who earns wages but does not claim them on his or her income taxes, or a cruel situation where people are forced to work without pay. However, informal sector could also be interpreted to include legal activities, such as jobs that are performed in exchange for something other than money which is opposite of formal sector.

FEMALE PARTICIPATION IN FORMAL SECTOR

Data from the Labor Force Survey indicate that females comprise a significant and rising portion of the occupational category of professionals and related workers. Between 1984-85 to 1987-88 the female share in the occupational group of professionals and related workers has risen from 15.5 to 18.3 percent of the total. Although there have been some inroads into non-traditional areas like engineering, banking, and law the numbers in these fields remain very limited, the major increase under this occupational group has been confined to the professions of teaching and medicine.

FEMALE PARTICIPATION IN INFORMAL SECTOR

The informal sector is steadily growing in almost all developing countries, for example in Latin America, 8.4 of every ten new jobs created between 1990 and 1994 were in the informal sector; in Asia, the informal sector absorbs between 40 and 50 percent of the urban labor force, and in Africa, the urban informal sector currently employs some 60 percent of the urban labor force and will create more than 90 percent of all additional jobs in this region.

OBJECTIVES

1. To find out determinants of female earnings in formal and informal sectors.
2. To find out which sector is easily accessible for females as a source of income.

METHOD AND METHODOLOGY FOR CONSTRUCTION THE ECONOMETRIC MODEL FOR COMPARISON

This research is based on the primary data and the field research is conducted in Alipur. We selected this area for the collection of our empirical data for several reasons. Firstly Pakistan is a developing country in which female earning in formal and informal sector plays an important role and Alipur constitutes a good location for our research related with formal and informal sector. Secondly, we live there and we have valuable contacts that helped us to touch with females. Total Population is 40000 and 1000 females for Formal Sector and 1000 females for Informal Sector are selected as a sample size for this study. And the data were collected within 2 weeks.

DEFINING THE VARIABLES

In our research the variables are

- Type of institution
- Age
- Education
- Response of the family
- Opinion about women’s job
- Working hours
- Family income
- Family size

MODEL CONSTRUCTION

Equation of the Female Earning in Formal Sector

$$FEF = \beta0 + \beta1TI + \beta2A + \beta3E + \beta4ROF + \beta5OPJ + Ui$$

where

FEF	=Female Earning in Formal sector
TI	= Type of Institution
A	= Age
E	= Education
ROF	=Response of Family
OPJ	=Opinion about Job
Ui	= Random error

β0, β1, β2, β3, β4, β5

Where female earning in formal sector is our dependent variable and the independent variables is type of institution, age, education, response of the family, opinion about women's job, and Ui is random error independently and identically distributed with zero mean and constant variance.

Equation of Female Earning in Informal Sector

$$FEF = \beta0 + \beta1FI + \beta2AFS + \beta3WH$$

where

FEI	=Female Earning in informal sector
FI	=Family Income
FS	=Family Size
WH	=Working Hours
Ui	= Random error

β0, β1, β2, β3

where female earning in informal sector is our dependent variables and the independent variables are working hours, family income, family size and Ui is random error independently and identically distributed with zero mean and constant variance.

For this we use the simple regression model. Data has analyzed using statistical package for social scientists (SPSS). The regression equations have also estimated with OLS regression model for comparison of female earnings in formal and informal sector.

REGRESSION RESULTS

Regression results are shown in table 1. All coefficients are the positive sign. All coefficients are also statistically significant.

Quantitative Analyses Of Informal Sector

In quantitative analysis we show the results of Female earnings in Informal Sector through this table no.

Table 1: Regression Results of female earnings in informal sector

Variables	Standardized β	t-values	Significance
Constant		.048	.962
Family Income	.730	11.742	.000
Family size	-.229	-3.600	.001
Hours	.205	3.224	.002

Source: Survey

R Square	Adjusted R Square	F Test
.630	.619	54.546

EXPLANATION

Female earning in informal sector is our dependent variable in this study and significantly dependent upon the family Income, Family size, working Hours**.** When there is increase in family income the female earnings increases. When the family size will increase the female earning will decrease. Working hours are also positively related with the dependent variable. When there is increase in working hours the total earning will also increase. R square shows the goodness of fit and its value is 0.630 and the value of adjusted R square is 0.619. T test is used to check the significance of the β's. Where β's are significant when the value is greater than 2. Here our result shows that all β's are significant. F test also check the significance of the overall model.

QUANTITATIVE ANALYSES OF FORMAL SECTOR

In quantitative analysis we show the results of Female earnings in Formal Sector through this Table No. 2.

Table 2: Regression results of female earning in formal sector

Variables	Standardized β	t-values	Significance
Constant		-6.152	.000
Type of Inst	.297	3.795	.000
Age	.338	4.335	.000
Education	.169	2.061	.042
Response of family	.254	3.274	.001
Opinion	.206	2.596	.011

Source: Survey

Square	Adjusted R Square	F Test
.467	.433	13.599

EXPLANATION

Female earning in formal sector is our dependent variable in this study and significantly dependent upon the type of institution, age, education, role of women and opinion. Type of institution is significant and positively related with the female earning. When there is increase in types of institution female learn more skills and increases the earnings. Age is also significant and positively related with the female earning. Education is significant and also positively related with the dependent variable. When the level of education is high the female earning will also be high. Family response is significant and also positively related with the female earning. Better family response will causes the higher level of female earning. Opinion about business is also significant and positively related with female earning in formal sector. If the opinion about job will good then chances of earnings will be high.R square shows the goodness of fit and its value is .467 and the value of adjusted R square is 0.433. T test is used to check the significance of the β’s. Where β’s are significant when the value is greater than 2. Here our result shows that all β’s are significant. F test also check the significance of the overall model.

5.3 Qualitative analysis of formal and informal sector Comparison

Table 3: Age

Age of respondent in formal sector			Age of respondent in Informal sector		
Age	**Frequency**	**Percent**	**Age**	**Frequency**	**Percent**
15-20	10	10.0	15-20	22	22.0
20-39	56	56.0	20-39	50	50.0
40-49	29	29.0	40-49	24	24.0
50-59	5	5.0	50-59	4	4.0
Total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This table shows the distribution of the respondent’s monthly earnings in formal and informal sector. Ten percent female earn in the age of 15-20 in formal sector and 22.0% in informal sector. Fifty six percent female earn in the age of 20-39 in formal sector and 50% in informal sector. Twenty percent female earn in the age of 40-49 in formal sector and 24% in informal sector. Five percent female earn in the age of 50-59 in formal sector and 4% in informal sector.

Table 4: Education

Education of respondent in formal sector			Education of respondent in Informal sector		
Education	**Frequency**	**Percent**	**Education**	**Frequency**	**Percent**
Primary	3	3.0	Primary	16	16.0
Metric	8	8.0	Metric	11	11.0
Intermediate	19	19.0	Intermediate	11	11.0
Graduation	11	11.0	Graduation	13	13.0
Master	39	39.0	Master	23	23.0
Above master	19	19.0	Above master	2	2.0
Uneducated	1	1.0	Uneducated	24	24.0
total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This shows the respondents education which is 3% in the Primary education in formal sector and 16% in informal sector. Eight percent are getting education of Metric in formal sector and 11% are in informal sector. In formal sector 19% are getting education of Intermediate and 11% in formal sector. Eleven percent are getting education of Graduation in formal sector and 13% in informal sector. Thirty nine percent have Master education in formal sector and 23% have in informal sector. Nineteen percent have education of above master in formal sector and 2% in informal sector and 1% female is uneducated in formal sector and 24% in informal sector.

Table 5: Opinion about Job

Opinion about women job in formal sector			Opinion about women job in Informal sector		
OJ	**Frequency**	**Percent**	**OJ**	**Frequency**	**Percent**
Favorable	87	87	Favorable	91	91
Unfavorable	13	13	Unfavorable	9	9
Total	100	100.0	Total	100	100.0

Source: Survey

Explanation:

This table shows that 87% female are in the favor of opinion about women's job in formal sector and 91% are in the favor of informal sector and 13% are unfavorable for women's job in formal sector and 9% are in informal sector.

Table 6: Type of Institution

Type of institution from where they got skill in formal sector			Type of institution from where they got skill in informal sector		
TOI	**Frequency**	**Percent**	**TOI**	**Frequency**	**Percent**
Govt	91	91.0	Govt	20	20.0
Private	9	9.0	Private	80	80.0
total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This table shows that 91% female get their skills from Government Institute in formal sector and 20% in informal sector and 9% female get their skill from Private Institution and 80% in informal sector.

Table 7: Response of Family

Response of your family in formal sector			Response of your family in Informal sector		
POF	**Frequency**	**Percent**	**ROF**	**Frequency**	**Percent**
V. Supportive	40	40.0	V. Supportive	68	68.0
Supportive	44	44.0	Supportive	22	22.0.
Indifferent	12	12.0	Indifferent	3	3.0
Non supportive	3	3.0	Non supportive	5	5.0
React badly	1	1.0	React badly	4	4.0
total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This table shows that 40% response of the family is very supportive for formal sector and 68% in formal sector. Forty four percent response of the family is supportive for formal sector and 22% in informal sector. Twelve percent response of the family is indifferent in formal sector and 3% in informal sector. Three percent are non supportive in formal sector and 5% in informal sector. One percent response of the family is badly reacted in formal sector and 4% are in informal sector.

Table 8: Family Size

Family size in formal sector			Family size in Informal sector		
FS	**Frequency**	**Percent**	**FS**	**Frequency**	**Percent**
1	0	0.0	1	1	1.0
2	3	3.0	2	4	4.0
3	3	3.0	3	2	2.0
4	11	11.0	4	10	10.0
5	15	15.0	5	15	15.0
6	17	17.0	6	13	13.0
7	16	16.0	7	17	17.0
8	21	21.0	8	7	7.0
9	7	7.0	9	8	8.0
10	7	7.0	10	9	9.0
11	0	0.0	11	3	3.0
12	0	0.0	12	10	10.0
13	0	0.0	13	1	1.0
Total	100	100.0	total	100	100.0

Source: Survey

Explanation:

According to the survey results table 8 shows that there are 2 members in the family that are 3% in formal sector and 4% in informal sector. There are 3% families in which there are 3 members in formal sector and 2% families in informal sector. There are 11% families in which there are 4 members in formal sector and 10% families in informal sector. In formal sector the minimum family size is 2 and in informal sector the minimum family size is 1 and the maximum family size in formal sector is 10 and in informal sector the maximum family size is 13.

Table 9: Family Income

Family income in formal sector			Family income in Informal sector		
FI	**Frequency**	**Percent**	**FI**	**Frequency**	**Percent**
1000-10000	12	12.0	1000-10000	39	39.0
10000-20000	12	12.0	10000-20000	20	20.0
20000-40000	42	42.0	20000-40000	23	23.0
40000-80000	28	28.0	40000-80000	20	20.0
80000-120000	5	5.0	80000-120000	1	1.0
Total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This table shows that family income from 1000-10000 have 12% in formal sector and 39% in informal sector. From 10000-20000 the family income is again 12% in formal sector and 20% in informal sector. From 20000-40000 the family income increases at 42% in formal sector and 23% in informal sector. From 40000-80000 the family income is 28% in formal sector and 20% in formal sector. From 80000-120000 the family income is 5% in formal and 1% in informal sector.

Table 10: Working Hours

Working hours in formal sector			Working hours in Informal sector		
H	**Frequency**	**Percent**	**H**	**Frequency**	**Percent**
2	0	0.0	2	4	4.0
3	0	0.0	3	3	3.0
4	1	1.0	4	12	12.0
5	12	12.0	5	21	21.0
6	17	17.0	6	11	11.0
7	24	24.0	7	10	10.0
8	40	40.0	8	17	17.0
9	2	2.0	9	12	12.0
10	4	4.0	10	5	5.0
11	0	0.0	11	0	0.0
12	0	0.0	12	5	5.0
Total	100	100.0	total	100	100.0

Source: Survey

Explanation:

This table shows the working hours of female earnings. One percent female work 4 hours in formal sector and 12% female work in informal sector. Twelve percent female work 5 hours in formal sector and 21% in informal sector. The minimum working hours of female earning in formal sector are 4 and in informal sector the minimum working hours are 2 and the maximum working hours in formal sector are 10 and in informal sector maximum working hours are 12.

CONCLUSION

The study has find out the comparison of female earnings in formal and informal sector. With the help of OLS regression model, results have been obtained. Percentage has also been calculated. The dependent variables are female earnings in formal sector and female earnings in informal sector. Results show that informal sector is easily accessible for females as a source of income. Results show that 40% response of the family is very supportive for formal sector and 68% in formal sector. Results also shows that 87% female are in the favor of opinion about women's job in formal sector and 91% are in the favor of informal sector and 13% are unfavorable for women's job in formal sector and 9% are in informal sector. Determinants of female earnings are different in formal sectors and informal sectors. Female earnings in formal sector is determined by type of institution from where you got skill, age of respondents, education and response of the family, opinion about job and Female earning in informal sector is determined by family income, family size, and working hours. Type of institution from where you got skill, age, education, response of the family, opinion about job, family income and working hours are positively affect the dependent variables and only family size is negatively affect the dependent variable.

POLICY IMPLICATION

The following policies may be adopted for the female earnings in formal and informal sector:

1. The government should establish special training institutes for household workers that provide skill for female workers.
2. The government can intervene for the establishment of educational institutions for female workers so that they can rationally manage their household budget and decrease expenditures.
3. The government should give the free education to the females so that they can easily get job.
4. The government should make the best policies for the working time for females.

REFERENCES

1. Jamali, K. (2009). The Role of Rural Women in Agriculture and its Allied Fields: A Case Study of Pakistan. *European Journal of Social Sciences*, 7(3).
2. Sikod, F. (2007). Gender Division of Labor and Women's Decision-Making Power in Rural Households in Cameroon. *Africa Development*, XXXII(3), 58-71. © Council for the Development of Social Science Research in Africa, 2007 (ISSN 0850-3907).
3. Rodgers, Y.D.M. (1999). Protecting Women and Promoting Equality in the Labor Market: Theory and Evidence November 1999. The World Bank, Development Research Group/Poverty Reduction and Economic Management Network.
4. Paula, A.D. and Scheinkman, J.A. (2006). The Informal Sector, First version: January 11, 2006, This Version: July 12, 2006.
5. Akintoye, I. R. (2008). Reducing Unemployment through the Informal Sector: A Case Study of Nigeria. *European Journal of Economics Finance and Administrative Sciences*, 11, ISSN 1450-2275.
6. Bhattacharya, P.C. (2007). Informal Sector, Income Inequality and Economic Development. September 2007, Discussion Paper 2007/09.
7. Granström, S.C. (2009). The informal sector and formal competitiveness in Senegal. Department of Economics at the University of Lund 2009: 8, Minor Field Study Series.
8. Bello, O.D., Venezuela, B.C.D. (1998). The informal sector, productivity shocks, and distortions in the labor market. Escuela de Economía Universidad Católica Andrés Bello.
9. Huda, S. S. M. S, Alam, M. S., and Khan, M. Y. (2009). A Comparative Study of Women Entrepreneurs in Formal and Informal Economy: A Study of Dhaka City. *Asian Journal of Business Management*, 1(1): 19-23, ISSN: 2041-8752© Maxwell Scientific Organization, 2000.
10. Blaauw, M.P.F. (2005). The dynamics of the informal sector in South Africa– a case study of day labourers in pretoria1" 1 Paper presented at the *biennial conference of the Economic Society of South Africa*, 7-9 September 2005 in Durban, South Africa, 2 Lecturer, Department of Economics, University of Johannesburg, South Africa.
11. Javed, A, Sadaf, S, and Luqman, M. (2006). Rural Women's Participation in Crop and Livestock Production Activities in Faisalabad. *Pakistan Journal of Agriculture & Social Sciences*, 1813–2235/2006/02–3–150–154.
12. http://www.google.com.articles
13. http//:www.pakistan.com
14. Contreras, D., and Gonzalo, P. (2008). *Female Labor Force Participation in Chile: How Important Are Cultural Factors*.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 437-450

DISTRIBUTION OF THE NUMBER OF OBSERVATIONS NEAR THE i-th DEPENDENT PROGRESSIVELY TYPE-II CENSORED ORDER STATISTIC

M. Rezapour, M.H. Alamatsaz and **N. Balakrishnan**
[1] Department of Statistics, Shahid Bahonar University, Kerman, Iran.
Email: mohsenrzp@gmail.com
[2] Department of Statistics, University of Isfahan, Isfahan, Iran.
Email: alamatho@sci.ui.ac.ir
[3] Department of Mathematics and Statistics, McMaster University,
Ontario L8S 4K1, Canada. Email: bala@univmail.cis.mcmaster.ca

ABSTRACT

In a life-testing problem, it may be of interest to investigate the number of observations near the median, maximum or more generally, the i-th progressively Type-II censored order statistic (PCOS-II). In this paper, we shall derive the probability mass and distribution functions of the number of observations near the i-th PCOS-II for a system with identical components for the cases with independent as well as dependent components. The type of dependence considered among the component lifetimes is through an Archimedean copula. We also describe a goodness-of-_t method for determining the best copula model for a given PCOS-II. Finally, an example is provided to illustrate all the results established here.

1. Introduction

A prominent method for lifetime investigation that saves time and cost is progressively Type-II censored sampling. When the system has dependent components, we should use a scheme which takes into account dependency. Let us now describe a dependent progressively Type-II censored sample. Under this sampling scheme, N identical but dependent units are placed on a life-test; after the i-th failure, R_i $(i = 1, \ldots, m \leq N)$ surviving units are removed at random from the surviving components. Thus, under the dependent progressively Type-II censored sampling scheme, we observe in all m failures, so that $N = m + R_1 + R_2 + \cdots + R_m$. Many authors have studied progressive Type-II censoring and properties of order statistics arising from such a progressively censored life-test. In this paper, the known results on the distribution of the number of observations near the minimum of PCOS-II [see [5]] are extended in two directions, namely, for the i-th order statistic as well as for the dependent case. The dependence considered is of Archimedean copula type. Suppose $X^{\tilde{R}}_{1:m:N}, X^{\tilde{R}}_{2:m:N}, \ldots, X^{\tilde{R}}_{m:m:N}$ are the order statistics obtained by the scheme mentioned above, where $\tilde{R} = (R_1, \ldots, R_m)$. Define the variable $\mu(i, m, N, a)$ as the number of observations near the i-th PCOS-II, i.e.,

$$(1) \qquad \mu(i, m, N, a) = \#\{X^{\tilde{R}}_{j:m:N} | X^{\tilde{R}}_{j:m:N} \in (X^{\tilde{R}}_{i:m:N}, X^{\tilde{R}}_{i:m:N} + a)\},$$

1991 *Mathematics Subject Classification.* 60E15 and 60K10.

Key words and phrases. Archimedean copula, order statistics, progressive censoring, progressively Type-II censored order statistics, reliability systems, goodness-of-fit test.

where a is a positive constant. If n identical dependent units are tested for reliability, the weakest and the most reliable items will fail first and last, respectively. Thus, the smallest and the largest order statistics in the sample of dependent units that are being tested will give their corresponding lifetimes. Here, we are primarily concerned with the number of items with their lifetimes being close to the i-th order statistic in a dependent progressively Type-II censored sample.

The joint density function of such order statistics has also been presented in [1]. If the marginal distribution function of the components is F, one can write the joint density function of $X_{1:m:N}^{\bar{R}}, X_{2:m:N}^{\bar{R}}, \ldots, X_{m:m:N}^{\bar{R}}$ as

$$\begin{aligned} &P(x_1 < X_{1:m:N}^{\bar{R}} \leq x_1 + dx_1, \ldots, x_m < X_{m:m:N}^{\bar{R}} \leq x_m + dx_m) \\ &\quad = \quad \vartheta_1 \cdots \vartheta_{m-1}(1-F(x_1))^{R_1} dF(x_1) \cdots (1-F(x_m))^{R_m} dF(x_m) + o(1), \end{aligned}$$

where $\vartheta_l = N - l + 1 - R_1 - \cdots - R_{l-1}$ $(2 \leq l \leq m-1)$, $\vartheta_m = R_m + 1$, $\vartheta_1 = N$ and $\vartheta_{m+1} = 0$. In practice, however, there may be dependence among components of the system, because they may be sharing a common source or common random environment. Such a dependence will also arise if the components are affected by a common shock, for example. Copulas are convenient tools for modeling the dependence structure in such situations. Modeling a system with dependent components using a copula framework is a very practical method, and in fact [11] and [10] recently applied Archimedean copulas to model reliability systems with dependent components. Marginal density and distribution functions of PCOS-II arising from a dependent and non-identical sample were derived in [8] and [9]. A detailed description of copulas and their subclass of Archimedean copulas can be found in [6] and [7]. In this paper, we derive the density function of μ defined in (1) for a system whose dependent components are jointly distributed according to an Archimedean copula.

A function $\psi : \Re_+ \mapsto [0,1]$ is said to be *d-alternating* if $(-1)^k \psi^{(k)} \geq 0$ for $k \in \{1, \ldots, d\}$. It is said to be *completely monotone* if it is d-alternating for all $d \in \aleph$. A copula C_ψ with the following form is called an Archimedean copula :

$$C_\psi(u_1, \ldots, u_N) = \psi\left(\sum_{i=1}^{N} \psi^{-1}(u_i)\right), \tag{2}$$

where $\psi : \Re_+ \mapsto [0,1]$ is an N-alternating $(N \geq 2)$ function such that $\psi(0) = 1$ and $\lim_{x \to \infty} \psi(x) = 0$. Here, ψ is said to be the generator function of the copula. In this paper, we assume that the generator of the Archimedean copula is completely monotone. In this case, we can rewrite (2) as

$$C_\psi(u_1, \ldots, u_N) = \int_0^\infty \prod_{j=1}^{N} G^\alpha(u_i) dM_\psi(\alpha), \tag{3}$$

where $G(u) = \exp\{-\psi^{-1}(u)\}$ and M_ψ is a distribution function with Laplace transform ψ.

Now, consider the random vector $\mathbf{X} = (X_1, \ldots, X_N)$ with joint survival function

$$\begin{aligned} (4) \qquad & P(X_1 > x_1, \ldots, X_N > x_N) \\ & = \psi\left(\sum_{i=1}^{N} \psi^{-1}(\bar{F}(x_i))\right) = \int_0^\infty \prod_{i=1}^{N} G^\alpha(\bar{F}(x_i)) dM_\psi(\alpha), \end{aligned}$$

where $\bar{F} = 1 - F$ and F is the marginal distribution function. Let us further assume that the distribution function F has density function f. Then, the joint density function of $\mathbf{X}$ equals

$$(5) \qquad \int_0^\infty \prod_{i=1}^{N} \alpha g(\bar{F}(x_i)) f(x_i) G^{\alpha-1}(\bar{F}(x_i)) dM_\psi(\alpha),$$

where g is the derivative of G. But, (5) yields

$$\begin{aligned} (6) \quad & P\Big(x_1 < X_1 \le y_1, \ldots, x_n < X_n \le y_n | X_{n+1} = x_{n+1}, \ldots, X_N = x_N\Big) \\ & \qquad \cdot f_{X_{n+1}, \ldots, X_N}(x_{n+1}, \ldots, x_N) \\ & = \int_0^\infty \int_{x_1}^{y_1} \ldots \int_{x_n}^{y_n} \prod_{s=1}^{n} \alpha g(\bar{F}(w_s)) G^{\alpha-1}(\bar{F}(w_s)) f(w_s) dw_s \\ & \quad \cdot \prod_{s=n+1}^{N} \alpha g(\bar{F}(v_s)) G^{\alpha-1}(\bar{F}(v_s)) f(v_s) dM_\psi(\alpha) \\ & = \int_0^\infty \prod_{s=1}^{n} \Big\{ G^\alpha(\bar{F}(x_s)) - G^\alpha(\bar{F}(y_s)) \Big\} \prod_{s=n+1}^{N} \alpha g(\bar{F}(x_s)) G^{\alpha-1}(\bar{F}(x_s)) f(x_s) dM_\psi(\alpha), \end{aligned}$$

where $x_i < y_i$ for $i = 1, 2, \ldots, n$. For more details, one may refer to [6] or [7].

The rest of this paper is organized as follows. In the next section, we derive some distributional results for the variable μ in (1) when the components are independent. We then investigate the distribution of μ when the components are dependent in Section 3. A goodness-of-fit method using the probability mass function of μ is also introduced in this section. Finally, an illustrative example that displays the behavior of the variable μ is provided in Section 4 and the proposed goodness-of-fit method is then applied to determine the best copula model.

2. Probability mass function of μ

In [5], the asymptotic properties of $\mu(1, m, N, a)$ with independent components have been investigated. In this section, we consider, more generally, the distribution function of $\mu(i, , m, N, a)$ for any $i = 1, \ldots, m$ when the components are independent and then for the dependent case. Let us consider the following progressive Type-II censoring scheme: Suppose N randomly selected dependent units from a population with joint survival function $\psi\left(\sum_{i=1}^{N} \psi^{-1}(\bar{F}(x_i))\right)$ are placed on a life-test, where $\bar{F}$ is the common marginal survival function. Further, suppose that at the time of the

i-th failure, R_i, $i = 1, 2, \ldots, m$, number of surviving units are randomly withdrawn from the test. Let $X^{\tilde{R}}_{j:m:N}$ denote the j-th dependent PCOS-II from a sample of size m from N items put on test with censoring scheme $\tilde{R} = (R_1, R_2, \ldots, R_m)$. It is evident that $N = m + R_1 + R_2 + \cdots + R_m$. Then, the joint probability density function of $X^{\tilde{R}}_{1:m:N}, \ldots, X^{\tilde{R}}_{m:m:N}$ can be expressed as

$$\begin{aligned}(7)\qquad \sum_{D_N} P\Big\{ & x_1 < X_{i_1} \le x_1 + dx, X_{i_2} > x_1, \ldots, X_{i_{R_1+1}} > x_1, \\ & x_2 < X_{i_{R_1+2}} \le x_2 + dx, X_{i_{R_1+3}} > x_2, \ldots, X_{i_{R_1+R_2+2}} > x_2, \\ & x_3 < X_{i_{R_1+R_2+3}} \le x_3 + dx, \ldots, X_{i_{R_1+\cdots+R_m+m}} > x_m\Big\},\end{aligned}$$

where the summation D_N, over all permutations $(i_1, \ldots, i_N)$ of $\{1, \ldots, N\}$. Then, by using (6), we can express the joint density function as

$$(8)\qquad \int_0^\infty c_{m-1} \prod_{i=1}^m \alpha G^{\alpha-1}(\bar{F}(x_i)) g(\bar{F}(x_i)) f(x_i) \Big\{G^\alpha(\bar{F}(x_i))\Big\}^{R_i} dM_\psi(\alpha)\,,$$

where $c_{m-1} = \prod_{i=1}^m \vartheta_i$, and $\vartheta_i = \sum_{v=i}^m (R_v + 1)$ for $i = 2, \ldots, m-1$, $\vartheta_1 = N$ and $\vartheta_m = 1$. Now, for convenience, we introduce the following notation:

$$\bar{G}(x, \alpha) = G^\alpha(\bar{F}(x)) \quad \textit{and} \quad g(x, \alpha) = -\frac{d\bar{G}(x, \alpha)}{dx} = \alpha G^{\alpha-1}(\bar{F}(x)) g(\bar{F}(x)) f(x)\,.$$

Then, (8) reduces to

$$(9)\qquad \int_0^\infty c_{m-1} \prod_{i=1}^m g(x_i, \alpha) \bar{G}^{R_i}(\alpha, x_i) dM_\psi(\alpha)\,.$$

Now, we present a lemma which is a key result for finding the distribution of μ.

Lemma 2.1. *Suppose the function $-f$ is the derivative of the absolutely continuous function $\bar{F}$. Then, we have*

$$\begin{aligned}(10)\qquad & \int_{x_s}^x \int_{x_{s+1}}^x \cdots \int_{x_{r-2}}^x \prod_{j=s+1}^{r-1} f(x_j) \{\bar{F}(x_j)\}^{R_j} dx_{r-1} \cdots dx_{s+1} \\ & = \sum_{i=s+1}^r a_i^{(s)}(r) \{\bar{F}(x)\}^{Q(i,r)} \{\bar{F}(x_s)\}^{Q(s+1,i)}\,,\end{aligned}$$

where $a_i^{(s)}(r) = \prod_{j=s+1, j \ne i}^r \frac{1}{Q(j,i)}$, $r+1 \le i \le s$, and $Q(j, i)$ equals $R_j + \cdots + R_{i-1} + i - j$ if $i > j$ and equals $-(R_i + \cdots + R_{j-1} + j - i)$, otherwise, and $Q(r, r) = 0$. In addition, if $\lim_{x \to -\infty} F(x) = 0$, we have

$$\text{(11)} \quad \int_{-\infty}^{x}\int_{x_1}^{x}\cdots\int_{x_{r-2}}^{x}\prod_{j=1}^{r-1} f(x_j)\{\bar{F}(x_j)\}^{R_j}\,dx_{r-1}\cdots dx_1 = \sum_{i=1}^{r} a_i(r)\{\bar{F}(x)\}^{Q(i,r)},$$

where $a_i(r) = \prod_{j=1, j\neq i}^{r} \frac{1}{Q(j,i)}$, $1 \leq i \leq r$.

Proof. If we assume $s = 0$ and let $x_0 \to -\infty$ in (10), the equality in (11) is obtained. Therefore, it is sufficient to prove (10). This can be proved by induction. Let $s = 0$ and $r = 3$. In this case, the left hand side of (10) equals

$$\int_{x_0}^{x} f(x_1)\{\bar{F}(x_1)\}^{R_1}\Big(-\frac{\{\bar{F}(x_2)\}^{R_2+1}}{R_2+1}\Big|_{x_1}^{x}\Big)dx_1$$
$$= \int_{x_0}^{x}\Big(\frac{1}{R_2+1}f(x_1)\{\bar{F}(x_1)\}^{R_2+R_1+1} - \frac{\{\bar{F}(x)\}^{R_2+1}}{R_2+1}f(x_1)\{\bar{F}(x_1)\}^{R_1}\Big)dx_1$$
$$= \frac{\{\bar{F}(x_0)\}^{R_2+R_1+2}}{(R_2+1)(R_2+R_1+2)} - \frac{\{\bar{F}(x_0)\}^{R_1+1}\{\bar{F}(x)\}^{R_2+1}}{(R_2+1)(R_1+1)} + \frac{\{\bar{F}(x)\}^{R_2+R_1+2}}{(R_1+1)(R_2+R_1+2)}$$
$$= a_3(3)\{\bar{F}(x_0)\}^{Q(1,3)} + a_2(3)\{\bar{F}(x_0)\}^{Q(1,2)}\{\bar{F}(x)\}^{Q(2,3)} + a_1^3(3)\{\bar{F}(x)\}^{Q(1,3)},$$

which yields the right hand side of (10). Now, we consider the general case. The left hand side of (10) equals

$$\int_{x_s}^{x}\int_{x_{s+1}}^{x}\cdots\int_{x_{r-3}}^{x}\prod_{j=s+1}^{r-2} f(x_j)\{\bar{F}(x_j)\}^{R_j}\Big(-\frac{\{\bar{F}(x_{r-1})\}^{R_{r-1}+1}}{R_{r-1}+1}\Big|_{x_{r-2}}^{x}\Big)dx_{r-2}\cdots dx_{s+1}$$
$$\text{(12)} \quad = \int_{x_s}^{x}\int_{x_{s+1}}^{x}\cdots\int_{x_{r-4}}^{x}\int_{x_{r-3}}^{x}\frac{1}{R_{r-1}+1}\Big(\prod_{j=s+1}^{r-2} f(x_j)\{\bar{F}(x_j)\}^{R_j^*}$$
$$-\frac{\{\bar{F}(x)\}^{R_{r-1}+1}}{R_{r-1}+1}\prod_{j=s+1}^{r-2} f(x_j)\{\bar{F}(x_j)\}^{R_j}\Big)dx_{r-2}\cdots dx_{s+1},$$

where $R_j^* = R_j$ for $j = s+1, \ldots, r-3$ and $R_{r-2}^* = R_{r-1} + R_{r-2} + 1$. But, by our induction assumption, we have

$$\int_{x_s}^{x}\int_{x_{s+1}}^{x}\cdots\int_{x_{r-3}}^{x}\prod_{j=s+1}^{r-2} f(x_j)\{\bar{F}(x_j)\}^{R_j}dx_{r-2}\cdots dx_{s+1}$$
$$= \sum_{i=s+1}^{r-2} a_i^{(s)}(r-1)\{\bar{F}(x)\}^{Q(i,r-1)}\{\bar{F}(x_s)\}^{Q(s+1,i)} + \{\bar{F}(x_s)\}^{Q(s+1,r-1)}a_{r-1}^{(s)}(r-1),$$

and

$$\int_{x_s}^{x}\int_{x_{s+1}}^{x}\cdots\int_{x_{r-3}}^{x}\prod_{j=s+1}^{r-2} f(x_j)\{\bar{F}(x_j)\}^{R_j^*}dx_{r-2}\cdots dx_{s+1}$$
$$= \sum_{i=s+1}^{r-2} a_i^{*(s)}(r-1)\{\bar{F}(x)\}^{Q^*(i,r-1)}\{\bar{F}(x_s)\}^{Q^*(s+1,i)} + \{\bar{F}(x_s)\}^{Q^*(s+1,r-1)}a_{r-1}^{*(s)}(r-1),$$

where $Q^*(i,r-1)=\sum_{v=i}^{r-2}(R_v^*+1)=\sum_{v=i}^{r-3}(R_v+1)+(R_{r-2}+R_{r-1}+2)=Q(i,r)$ for $i=s+1,\ldots,r-2$ and $Q^*(r-1,r-1)=Q(r-1,r-1)$, and

$$a_i^{*(s)}(r-1)=\prod_{j=s+1,j\neq i}^{r-1}\frac{1}{Q^*(j,i)}=\frac{1}{Q^*(r-1,i)}\prod_{j=s+1,j\neq i}^{r-2}\frac{1}{Q^*(j,i)}$$

for $i=s+1,\ldots,r-2$, where $Q^*(j,i)$ equals $R_j^*+\cdots+R_{i-1}^*+i-j=R_j+\cdots+R_{i-1}+i-j=Q(j,i)$ if $i>j$ and equals $-(R_i^*+\cdots+R_{j-1}^*+i-j)=-(R_i+\cdots+R_{j-1}+j-i)=Q(j,i)$ if $i<j$, for $j=s+1,\cdots,r-2$, and $Q^*(r-1,i)=-(R_i^*+\cdots+R_{r-2}^*+r-2-i)=-(R_i+\cdots+R_{r-2}+R_{r-1}+1+r-2-i)=Q(r,i)$. Therefore, $a_i^{*(s)}(r-1)$, for $i=s+1,\cdots,r-2$, equals

$$\frac{1}{Q(r,i)}\prod_{j=s+1,j\neq i}^{r-2}\frac{1}{Q(j,i)}. \tag{13}$$

But, $Q^*(i,r-1)=R_j^*+\cdots+R_{r-2}^*+r-2-j=R_j+\cdots+R_{r-2}+R_{r-1}+1+r-2-j=Q(j,r)$ and so we have

$$a_{r-1}^{*(s)}(r-1)=\prod_{j=s+1}^{r-2}\frac{1}{Q^*(j,r-1)}=\prod_{j=s+1}^{r-2}\frac{1}{Q(j,r)}, \tag{14}$$

and $Q^*(s+1,i)=\sum_{v=s+1}^{i-1}(R_v+1)=Q(s+1,i)$ for $i=1,\ldots,r-2$. Therefore, (12) equals

$$\sum_{i=s+1}^{r-2}\frac{a_i^{*(s)}(r-1)}{R_{r-1}+1}\{\bar F(x)\}^{Q(i,r)}\{\bar F(x_s)\}^{Q(s+1,i)}+\{\bar F(x_s)\}^{Q(s+1,r)}\frac{a_{r-1}^{*(s)}(r-1)}{R_{r-1}+1}$$
$$-\sum_{i=s+1}^{r-2}\frac{a_i^{(s)}(r-1)\{\bar F(x)\}^{R_{r-1}+1}}{R_{r-1}+1}\{\bar F(x)\}^{Q(i,r-1)}\{\bar F(x_s)\}^{Q(s+1,i)}$$
$$-\{\bar F(x_s)\}^{Q(s+1,r-1)}\frac{a_{r-1}^{(s)}(r-1)\{\bar F(x)\}^{R_{r-1}+1}}{R_{r-1}+1}. \tag{15}$$

Now, since $\frac{1}{R_{r-1}+1}=\frac{1}{Q(r-1,r)}$, by (14), we have $\frac{a_{r-1}^{*(s)}(r-1)}{Q(r-1,r)}=a_r^{(s)}(r)$ and $Q(i,r-1)+R_{r-1}+1=\sum_{v=i}^{r-1}(R_v+1)=Q(i,r)$ for $i=s+1,\ldots,r-1$ and $Q^*(s+1,r-1)=\sum_{v=s+1}^{r-2}(R_v+1)=Q(s+1,r)$, and $\frac{a_{r-1}^{(s)}(r-1)}{Q(r-1,r)}=-\frac{a_{r-1}^{(s)}(r-1)}{-Q(r,r-1)}=-a_{r-1}^{(s)}(r)$. Thus, (15) equals

$$\sum_{i=s+1}^{r-2}\frac{a_i^{*(s)}(r-1)}{R_{r-1}+1}\{\bar F(x)\}^{Q(i,r)}\{\bar F(x_s)\}^{Q(s+1,i)}+\{\bar F(x_s)\}^{Q(s+1,r)}a_r^{(s)}(r) \tag{16}$$
$$-\sum_{i=s+1}^{r-2}\frac{a_i^{(s)}(r-1)}{Q(r-1,r)}\{\bar F(x)\}^{Q(i,r)}\{\bar F(x_s)\}^{Q(s+1,i)}$$
$$+\{\bar F(x_s)\}^{Q(s+1,r-1)}a_{r-1}^{(s)}(r)\{\bar F(x)\}^{R_{r-1}+1}.$$

But, by (13), $\frac{1}{Q(r-1,r)}(a_i^{*(s)}(r-1)-a_i^{(s)}(r-1))$ equals

$$\frac{1}{Q(r-1,r)}\Big(\frac{1}{Q(r,i)}\prod_{j=s+1,j\neq i}^{r-2}\frac{1}{Q(j,i)}-\frac{1}{Q(r-1,i)}\prod_{j=s+11,j\neq i}^{r-2}\frac{1}{Q(j,i)}\Big)$$
$$=\frac{1}{Q(r-1,r)}\Big(\frac{Q(r-1,i)-Q(r,i)}{Q(r,i)Q(r-1,i)}\Big)\prod_{j=s+1,j\neq i}^{r-2}\frac{1}{Q(j,i)}=\frac{1}{Q(r,i)Q(r-1,i)}\prod_{j=s+1,j\neq i}^{r-2}\frac{1}{Q(j,i)},$$

which equals $a_i^{(s)}(r)$. Thus, (16) becomes

$$\sum_{i=s+1}^{r-2}a_i^{(s)}(r)\{\bar F(x)\}^{Q(i,r)}\{\bar F(x_s)\}^{Q(s+1,i)}+\{\bar F(x_s)\}^{Q(s+1,r)}a_r^{(s)}(r)$$
$$+\{\bar F(x_s)\}^{Q(s+1,r-1)}a_{r-1}^{(s)}(r)\{\bar F(x)\}^{Q(r-1,r)}\,.$$

This completes the proof of the lemma. □

In the following theorem, we present the probability mass function of μ defined in (1).

Theorem 2.2. *The probability that $\mu(i,m,N,a)=k$, $0\le k\le m-i$, is given by*

$$c_{i+k-1}\sum_{s=1}^{i}a_s(i)\sum_{j=i+1}^{i+k+1}a_j^{(i)}(i+k+1)\int_{-\infty}^{\infty}\frac{f(x)}{\psi'(\psi^{-1}(\bar F(x)))}$$
$$\cdot\psi'\Big((\vartheta_s-\vartheta_j)\psi^{-1}(\bar F(x))+\vartheta_j\psi^{-1}(\bar F(x+a))\Big)dx\,. \tag{17}$$

Proof. Let $0\le k\le m-i$. Since

$$P\Big(\mu(i,m,N,a)=k\Big)=P\Big(X_{k+i:m:N}^{\bar R}-X_{i:m:N}^{\bar R}<a,X_{k+i+1:m:N}^{\bar R}-X_{i:m:N}^{\bar R}>a\Big)$$
$$=\int_{-\infty}^{\infty}P\Big(X_{k+i:m:N}^{\bar R}<a+x_i,X_{k+i+1:m:N}^{\bar R}>a+x_i|X_{i:m:N}^{\bar R}=x_i\Big)f_{X_{i:m:N}^{\bar R}}(x_i)dx_i$$
$$=\int_{-\infty}^{\infty}\int_{-\infty}^{x_i}\int_{x_1}^{x_i}\cdots\int_{x_{i-2}}^{x_i}\int_{x_i}^{a+x_i}\int_{x_{i+1}}^{a+x_i}\cdots\int_{x_{k+i-1}}^{a+x_i}\int_{a+x_i}^{\infty}\int_{x_{i+k+1}}^{\infty}$$
$$\cdots\int_{x_{m-1}}^{\infty}f_{X_{1:m:N}^{\bar R},\ldots,X_{m:m:N}^{\bar R}}(x_1,\ldots,x_m)dx_m\cdots dx_{i+1}dx_{i-1}\cdots dx_1dx_i\,. \tag{18}$$

Upon using (9), we can rewrite (18) as

$$c_{m-1}\int_{-\infty}^{\infty}\int_{-\infty}^{x_i}\int_{x_1}^{x_i}\cdots\int_{x_{i-2}}^{x_i}\int_{0}^{\infty}\prod_{s=1}^{i-1} g(x_s,\alpha)\Big\{\bar{G}(x_s,\alpha)\Big\}^{R_s}$$

$$\cdot\int_{x}^{a+x_i}\int_{x_{i+1}}^{a+x_i}\cdots\int_{x_{i+k-1}}^{a+x_i}\prod_{s=i+1}^{i+k+1} g(x_s,\alpha)\Big\{\bar{G}(x_s,\alpha)\Big\}^{R_s}$$

$$\cdot\int_{a+x_i}^{\infty}\int_{x_{i+k+1}}^{\infty}\cdots\int_{x_{m-1}}^{\infty}\prod_{s=i+k+1}^{m} g(x_s,\alpha)\Big\{\bar{G}(x_s,\alpha)\Big\}^{R_s}$$

$$\cdot g(x_i,\alpha)\Big\{\bar{G}(x_i,\alpha)\Big\}^{R_i} dM_{\psi}(\alpha)dx_m\cdots dx_{i+1}dx_{i-1}\cdots dx_1 dx_i\,.$$

But, we can clearly see that

$$c_{m-1}\int_{a+x_i}^{\infty}\int_{x_{i+k+1}}^{\infty}\cdots\int_{x_{m-1}}^{\infty}\prod_{s=i+k+1}^{m} g(x_s,\alpha)\Big\{\bar{G}(x_s,\alpha)\Big\}^{R_s} dx_m\cdots dx_{i+k+1}$$

$$= c_{m-1}\frac{\Big\{\bar{G}(a+x_i,\alpha)\Big\}^{R_m+\cdots+R_{i+k+1}+m-(i+k)}}{\prod_{s=i+k+1}^{m}(\sum_{v=s}^{m}(R_v+1))} = c_{i+k-1}\Big\{\bar{G}(a+x_i,\alpha)\Big\}^{\vartheta_{i+k+1}}\,.$$

By Lemma 2.1, we have

$$\int_{-\infty}^{x_i}\int_{x_1}^{x_i}\cdots\int_{x_{i-2}}^{x_i}\prod_{s=1}^{i-1} g(x_s,\alpha)\Big\{\bar{G}(x_s,\alpha)\Big\}^{R_s} dx_{i-1}\cdots dx_1 = \sum_{s=1}^{i} a_s(i)\{\bar{G}(x_i,\alpha)\}^{Q(s,i)}$$

and

$$\int_{x_i}^{a+x_i}\int_{x_{i+1}}^{a+x_i}\cdots\int_{x_{i+k-2}}^{a+x_i}\prod_{j=i+1}^{i+k+1} g(x_j,\alpha)\{\bar{G}(x_j,\alpha)\}^{R_j} dx_{i+k+1}\cdots dx_{i+1}$$

$$= \sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\{\bar{G}(a+x_i,\alpha)\}^{Q(j,i+k+1)}\{\bar{G}(x_i,\alpha)\}^{Q(i+1,j)}\,.$$

Carrying out the required integration in (18), we obtain the probability mass function as

$$
\begin{aligned}
& c_{i+k-1}\int_{-\infty}^{\infty}\int_{0}^{\infty} g(x_i,\alpha)\Big\{\bar{G}(x_i,\alpha)\Big\}^{R_i}\Big\{\bar{G}(a+x_i,\alpha)\Big\}^{\vartheta_{i+k+1}}\sum_{s=1}^{i} a_s(i)\{\bar{G}(x_i,\alpha)\}^{Q(s,i)} \\
& \cdot\sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\{\bar{G}(a+x_i,\alpha)\}^{Q(j,i+k+1)}\{\bar{G}(x_i,\alpha)\}^{Q(i+1,j)}dM_\psi(\alpha)dx_i \\
&= c_{i+k-1}\sum_{s=1}^{i} a_s(i)\sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\int_{-\infty}^{\infty}\int_{0}^{\infty} g(x_i,\alpha)\{\bar{G}(x_i,\alpha)\}^{R_i+Q(s,i)+Q(i+1,j)} \\
&\qquad\qquad \cdot\{\bar{G}(a+x_i,\alpha)\}^{\vartheta_{i+k+1}+Q(j,i+k+1)}dM_\psi(\alpha)dx_i \\
&= c_{i+k-1}\sum_{s=1}^{i} a_s(i)\sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\int_{-\infty}^{\infty}\int_{0}^{\infty} g(x_i,\alpha)\{\bar{G}(x,\alpha)\}^{Q(s,j)-1} \\
&\qquad\qquad \cdot\{\bar{G}(a+x_i,\alpha)\}^{\vartheta_j}dM_\psi(\alpha)dx_i\,.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
& P\Big(\mu(i,m,N,a)=k\Big) \\
&\quad = c_{i+k-1}\sum_{s=1}^{i} a_s(i)\sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\int_{-\infty}^{\infty}\frac{f(x_i)}{\psi'(\psi^{-1}(\bar{F}(x_i)))} \\
&\quad \cdot\int_{0}^{\infty}\alpha\exp\bigg(-\alpha\Big(Q(s,j)\psi^{-1}(\bar{F}(x_i))+\vartheta_j\psi^{-1}(\bar{F}(x_i+a))\Big)\bigg)dM_\psi(\alpha)dx_i\,.
\end{aligned}
$$

Since $\psi'(x)=-\int_0^\infty \alpha e^{-\alpha x}dM_\psi(\alpha)$ and $Q(s,j)=\vartheta_s-\vartheta_j$, (17) follows which completes the proof of the theorem. □

Now, we consider the case when the components are independent. In this case, we obtain the following result.

Theorem 2.3. *Under the notation of Theorem 2.2, when the components are iid, the probability of* $\mu(i,m,N,a)=k$, $0\le k\le m-i$, *is given by*

$$
(19)\qquad c_{i+k-1}\sum_{s=1}^{i} a_s(i)\sum_{j=i+1}^{i+k+1} a_j^{(i)}(i+k+1)\int_{-\infty}^{\infty} f(x)\bar{F}^{\vartheta_s-\vartheta_j-1}(x)\bar{F}^{\vartheta_j}(x+a)dx\,.
$$

Proof. Consider $\psi(x)=e^{-x}$. In this case, the Archimedean copula reduces to the independent copula. Thus, substituting this completely monotone generator function into (17), we arrive at (19), as required. □

Remark 2.4. *If we let* $i=1$ *in Theorem 2.3, and replace* ϑ_{i+1} *by* γ_i, $i=0\ldots,m$, *the probability of* $\mu(1,m,N,a)=k$, $0\le k\le m-1$, *reduces to the expression given in* [5].

Goodness-of-fit test for PCOS-II. Using the probability mass function obtained in (17), we can propose a goodness-of-fit test for finding the best Archimedean copula that fits an observed PCOS-II sample. For this purpose, suppose the observed PCOS-II sample is from a specific Archimedean copula distribution with vector parameter $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1$ is the parameter of the marginal distribution and $\boldsymbol{\theta}_2$ is the parameter of the Archimedean copula. Then, from (17) for $i = 1, \dots, m$ and $a > 0$, let us define

$$\xi(i, a, \boldsymbol{\theta}) = \max_{0 \le k \le m-i} \left\{ P\Big(\mu(i, m, N, a) = k\Big) \right\}, \tag{20}$$

with its estimator being

$$\hat{\xi}(i, a, \boldsymbol{\theta}) = \#\{X_{j:m:N} | X_{j:m:N} \in (X_{i:m:N}, X_{i:m:N} + a)\}\,. \tag{21}$$

Now, if we consider the statistic

$$\zeta = \sum_{i=1}^{m} \Big(\xi(i, a, \boldsymbol{\theta}) - \hat{\xi}(i, a, \boldsymbol{\theta}) \Big)^2, \tag{22}$$

we can say that the Archimedean copula that best fits the observed PCOS-II sample is the one that minimizes the value of ζ in (22).

We may adapt the bootstrap methodology to compute the corresponding P-value and it proceeds as follows:

- Generate M random samples of size n from PCOS-II and, for each of these samples, find $\zeta_i, i = 1, \dots, M$, by the above described method;
- If $\zeta_{1:M} \le \dots \le \zeta_{M:M}$ denote the ordered values of the test statistics calculated from the first step, an estimate of the critical value of the test statistic at level α is then given by

$$\zeta_{[(1-\alpha)M]:M} \tag{23}$$

  and

$$\frac{1}{M} \#\{j : \zeta_{j:M} \ge \zeta\} \tag{24}$$

  yields an estimate of the P-value associated with the observed value ζ of the statistic in (22). Here, $[x]$ refers to the integer part of $x \in \mathbb{R}$.

In our computations in the following section, we used $M = 10,000$ bootstrap runs for determining the P-value of the test.

3. Illustrative Example

In this section, we provide an illustrative example in which the mass function of the number of observations near the i-th PCOS-II, in a system with dependent components distributed according to the generalized Clayton family, is obtained.

Example 3.1. Let $\psi(s) = (1+s)^{-\theta}$ for $\theta \geq 0$. Then, we have $C_\psi(u_1, \ldots, u_N) = (u_1^{-1/\theta} + \cdots + u_N^{-1/\theta} - N + 1)^{-\theta}$ which is known as the Clayton family. In this case, by Theorem 2.2, the probability mass function of $\mu(i, m, N, a)$ is given by

$$c_{i+k-1} \sum_{s=1}^{i} \sum_{j=i+1}^{i+k+1} a_s(i) a_j^{(i)}(i+k+1) \int_{-\infty}^{\infty} \frac{f(x)}{\theta \bar{F}^{1+\frac{1}{\theta}}(x)}$$
$$\cdot \theta \Big((\vartheta_s - \vartheta_j)(\bar{F}^{-\frac{1}{\theta}}(x) - 1) + \vartheta_j (\bar{F}^{-\frac{1}{\theta}}(x+a) - 1) + 1 \Big)^{-\theta-1} dx .$$

Thus, if $\bar{F}(x) = e^{-\lambda x}$, $x \geq 0$ and $\lambda > 0$, the probability mass function of $\mu(i, m, N, a)$ becomes

$$c_{i+k-1} \sum_{s=1}^{i} \sum_{j=i+1}^{i+k+1} a_s(i) a_j^{(i)}(i+k+1) \int_{0}^{\infty} \lambda e^{\frac{\lambda}{\theta} x}$$
$$\cdot \Big((\vartheta_s - \vartheta_j) e^{\frac{\lambda}{\theta} x} + \vartheta_j e^{\frac{\lambda}{\theta}(x+a)} - \vartheta_s + 1 \Big)^{-\theta-1} dx ,$$

which simplifies to

$$c_{i+k-1} \sum_{s=1}^{i} \sum_{j=i+1}^{i+k+1} \frac{a_s(i) a_j^{(i)}(i+k+1) \Big(\vartheta_j (e^{\frac{\lambda}{\theta} a} - 1) + 1 \Big)^{-\theta}}{\vartheta_s - \vartheta_j + \vartheta_j e^{\frac{\lambda}{\theta} a}} . \tag{25}$$

In Table 3.2 below, we have presented the values of the probability mass function of $\mu(i, m, N, a)$ in (25) for this example when $m = 13, N = 191, \lambda = 0.2, \theta = 2$, $\tilde{R} = (5, 7, 9, 5, 4, 3, 8, 15, 19, 18, 25, 36, 24)$, for $i = 1, 5, 7$ and $a = 0.06, 2.5, 10$.

Table 3.2. *Value of the probability mass function of $\mu(i, m, N, a)$ in Example 3.1 when $m = 13, N = 191, \lambda = 0.2, \theta = 2$, $\tilde{R} = (5, 7, 9, 5, 4, 3, 8, 15, 19, 18, 25, 36, 24)$, for $i = 1, 5, 7$ and $a = 0.01, 2.5, 10$.*

	$i = 1, a = 0.01$	$i = 5, a = 0.01$	$i = 7, a = 2.5$	$i = 7, a = 10$
1	0.121137658	0.145062958	0.000127719	3.07×10^{-8}
2	0.165682477	0.188021684	0.0005356	2.87×10^{-7}
3	0.168371223	0.181771414	0.001413472	1.67×10^{-6}
4	0.146194063	0.157969966	0.002973835	7.33×10^{-6}
5	0.116700081	0.126301229	0.006613652	3.47×10^{-5}
6	0.087873969	0.090394982	0.027645753	0.000406413
7	0.065147568	0.060844227	0.960689971	0.999549549
8	0.047935925	0.037786079		
9	0.034071796	0.011847486		
10	0.022328785			
11	0.013991936			
12	0.008146899			
13	0.002417621			

Now, we generate a sample of PCOS-II according to these assumptions. We produced random samples of size $N = 191$ from the Clayton family with parameter $\theta = 2$ and exponential marginal distribution with mean $\lambda = 0.2$. Then, with $m = 13$, we observed $X^{\tilde{R}}_{1:13:191}$ as the minimum value of a the generated sample. We then omitted 5 surviving components from the generated sample at random and observed $X^{\tilde{R}}_{2:13:191}$ as the minimum value of the remaining components. We followed this manner until all PCOS-II are recorded, which are as follows:

$$0.1882, 0.1920, 0.2473, 0.2695, 0.3114, 0.4020, 0.4372,$$
$$0.4885, 0.5391, 0.6387, 0.7519, 1.0394, 1.1958\,.$$

Now, for finding the best Clayton family that fits the observed PCOS-II, we determined ζ and the P-value of the hypothesis test in (22) for various values of θ and λ. The values so obtained are presented in Table 3.3.

Table 3.3. *Goodness-of-fit statistic ζ in (22) for various values of θ, λ and $a = 0.01, 2.5, 10$ for the simulated data.*

		$\theta = 0.5$		$\theta = 1$		$\theta = 1.5$	
λ	a	ζ	*P-value*	ζ	*P-value*	ζ	*P-value*
0.05	0.01	10.72220	0.000376	10.73005	0.013008	10.73195	0.012564
0.05	2.5	585.3782	0.009640	577.0205	0.022404	572.3473	0.021640
0.05	10	532.9293	0.013836	524.1712	0.026660	518.1261	0.025750
0.1	0.01	9.728561	0.002609	9.698003	0.015273	9.683874	0.014751
0.1	2.5	558.1850	0.011816	545.5925	0.024610	537.7770	0.023771
0.1	10	513.7623	0.015370	512.3113	0.028215	509.8802	0.027252
0.15	0.01	8.931059	0.004401	8.846248	0.017090	8.808590	0.016507
0.15	2.5	543.0629	0.013026	531.7852	0.025837	524.5497	0.024956
0.15	10	508.6211	0.015781	508.9034	0.028632	508.0094	0.027655
0.2	0.01	8.275219	0.005875	8.133902	0.018585	8.070786	0.017951
0.2	2.5	532.9293	0.013836	524.1712	0.026660	518.1261	0.025750
0.2	10	507.3716	0.015881	507.7029	0.028733	507.3948	0.027753
0.25	0.01	7.725357	0.007110	7.531243	0.019838	7.443612	0.019161
0.25	2.5	525.6786	0.014416	519.4169	0.027248	514.5267	0.026318
0.25	10	507.0837	0.015904	507.2608	0.028757	507.1622	0.027775
0.3	0.01	7.257041	0.008163	7.016199	0.020905	6.906389	0.020192
0.3	2.5	520.3956	0.014839	516.2174	0.027677	512.3160	0.026732
0.3	10	507.0187	0.015909	507.0966	0.028762	507.0682	0.027780
0.4	0.01	6.500390	0.009863	6.186055	0.022630	6.040701	0.021858
0.4	2.5	513.7623	0.015370	512.3113	0.028215	509.8802	0.027252
0.4	10	507.0009	0.015911	507.0131	0.028763	507.0125	0.027782

		$\theta = 2$		$\theta = 2.5$		$\theta = 3$	
λ	a	ζ	*P-value*	ζ	*P-value*	ζ	*P-value*
0.05	0.01	10.73276	0.001008	10.73320	0.000974	10.73347	0.001115
0.05	2.5	569.2302	0.025837	566.9447	0.024956	565.1711	0.028580
0.05	10	514.5218	0.037082	512.2801	0.035817	510.8252	0.041019
0.1	0.01	9.675965	0.006992	9.670934	0.006754	9.667458	0.007735
0.1	2.5	532.5411	0.031667	528.7867	0.030587	525.9680	0.035030
0.1	10	508.5486	0.041192	507.8567	0.039786	507.4898	0.045565
0.15	0.01	8.787671	0.011795	8.774398	0.011393	8.765237	0.013048
0.15	2.5	519.9972	0.034910	516.9665	0.033718	514.8566	0.038616
0.15	10	507.4856	0.042294	507.2341	0.040851	507.1156	0.046784
0.2	0.01	8.035559	0.052365	8.013142	0.050578	7.997635	0.057924
0.2	2.5	514.5218	0.048956	512.2801	0.047285	510.8252	0.054153
0.2	10	507.1810	0.042562	507.0802	0.041109	507.0359	0.047081
0.25	0.01	7.394444	0.019057	7.363056	0.018406	7.341300	0.021080
0.25	2.5	511.7247	0.038637	510.0800	0.037318	509.0762	0.042739
0.25	10	507.0737	0.042624	507.0311	0.041169	507.0129	0.047149
0.3	0.01	6.844511	0.021877	6.804915	0.021130	6.777428	0.024200
0.3	2.5	510.1389	0.039769	508.9214	0.038412	508.2170	0.043992
0.3	10	507.0316	0.042637	507.0129	0.041182	507.0051	0.047164
0.4	0.01	5.958411	0.026434	5.905640	0.025532	5.868962	0.029240
0.4	2.5	508.5486	0.041192	507.8567	0.039786	507.4898	0.045565
0.4	10	507.0062	0.042641	507.0025	0.041186	507.0009	0.047168

We observe that the value of ζ is the smallest for $\theta = 2$ and $\lambda = 0.2$ and, in this case, the P-value is greater than 0.05 which implies that at the usual 5% significance level, the null hypothesis that the observed PCOS-II order statistics are from a random vector $\mathbf{X} = (X_1, \ldots, X_N)$ with joint survival function (4) with Clayton generator function with parameter $\theta = 2$ and $\bar{F}(x) = e^{-0.2x}$ can not be rejected. This illustrates that the proposed method is useful for such a testing purpose.

References

[1] Aggarwala, R., and Balakrishnan, N. (1998). Some properties of progressive censored order statistics from arbitrary and uniform distributions with applications to inference and simulation. *J. Statist. Plann. Inference* **70**, 35-49.

[2] Balakrishnan, N. and Aggarwala, R. (2000) *Progressive Censoring: Theory, Methods, and Applications*. Birkhäuser, Boston.

[3] Balakrishnan, N. (2007). Progressive censoring methodology: an appraisal (with Discussions). *Test* **16**, 211-296.

[4] Balakrishnan, N. and Cramer, E. (2008). Progressive censoring from heterogeneous distributions with applications to robustness. *Ann. Inst. Statist. Math.* **60**, 151-171.

[5] Balakrishnan, N., and Stepanov, A. (2008). Asymptotic properties of numbers of near minimum observations under progressive Type-II censoring. *J. Statist. Plann. Inference*, **38**, *1010-1020.*

[6] Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.

[7] Nelsen, R.B. (2006). *An Introduction to Copulas*. Springer, New York.

[8] Rezapour, M., Alamatsaz, M. H., Balakrishnan, N. and Cramer, E. (2012). On properties of progressively Type-II censored order statistics arising from dependent and non-identical random variables. *Statistical Methodology* **10**, *1, 5871.*

[9] Rezapour, M., Alamatsaz, M. H. and Balakrishnan, N. (2010). On properties of dependent general progressively Type-II censored order statistics. *To appear in Metrika.*

[10] Rezapour, M., Alamatsaz, M.H. and Pellerey, F. (2011). Multivariate aging with Archimedean dependence structures in high dimensions. *Commun. Statist.-Theor. Meth. (to appear).*

[11] Rezapour, M., Salehi, E. and Alamatsaz, M.H. (2011). Stochastic comparison of residual and past lifetimes of $(n-k+1)$-out-of-n systems with dependent components. *Commun. Statist.-Theor. Meth. (to appear).*

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 451-456

ANALYSIS OF A SINGLE SERVER WAITING LINE QUEUING SYSTEM USING A NEW SERVICE TIME DISTRIBUTION

Masood Alam[1] and **Sabir Ali Siddiqui**[2]

[1] Department of Mathematics and Statistics, Sultan Qaboos University,
Muscat, Sultanate of Oman. Email: syedmasoodalam@gmail.com

[2] Department of Mathematics and Sciences, Dhofar University,
Salalah, Sultanate of Oman. Email: siddiqui.sabir@gmail.com

ABSTRACT

In our present article we have analyzed a system with single server where the arrivals follow Poisson distribution and the service time distribution is Inverse Gaussian distribution.

The same system has been analyzed by using Mukherji-Islam (1983) model as the service time model.

Inverse Gaussian distribution has been used in queuing theory but the Mukherjee-Islam distribution which is basically a failure model has been tested first time as the service time model in our study.

In this article parameters of the system like average number of customers in the system, variability and coefficient of variation of system size have been evaluated.

1. INTRODUCTION

The Inverse Gaussian family of distributions are often used in analyzing many of the realistic situations arising at life testing, economical analysis, Insurance studies etc. The major advantage of this distribution is the interpretation of the Inverse Gaussian random variable as the first passage time distribution of Brownian motion with positive drift. In textile industries the printing or bleaching processed are distributed approximately as Inverse Gaussian distribution. Here unit of cloth is to be taken as customer, the printing or bleaching is viewed as service. In spite of wide applicability of the Inverse Gaussian distribution as approximate model for skewed data and having simple exact sampling theory, it has not been used much in analyzing waiting line systems.

2. QUEUING MODEL WITH INVERSE GAUSSIAN SERVICE TIME DISTRIBUTION

Consider a single server queuing with infinite capacity having FCFS (First Come First Serve) queue discipline and the arrivals are Poisson with arrival rateλ. The service time distribution of the process is a Inverse Gaussian of the form;

$$f(t;\mu,\theta)=\left(\frac{\theta}{2\pi t^3}\right)^{1/2}\exp-\left\{\frac{\theta(\mu t-1)^2}{2.t}\right\} \quad ;\mu,\theta\geq 0$$

$$;0\leq t<\infty \qquad (1)$$

With Mean $=\dfrac{1}{\mu}$ and Variance $=\dfrac{1}{\theta\mu^3}$.

3. MAXIMUM LIKELIHOOD ESTIMATES

Parameters μ and θ are involved in the service time distribution given in equation (1). Consider a random sample $t_1, t_2, \ldots\ldots t_n$ from the population with p.d.f. (1). The likelihood function is given as;

$$L(t\ ;\mu,\theta) = \prod_{i=1}^{n} f_i(t_i;\mu,\theta)$$

$$= \prod_{i=1}^{n}\left(\frac{\theta}{2\pi t_i^{3}}\right)^{1/2} \exp-\left\{\frac{\theta(\mu t_i - 1)^2}{2.t_i}\right\}$$

$$= \left(\frac{\theta}{2\pi}\right)^{n/2} \prod_{i=1}^{n}\left(\frac{1}{t_i}\right)^{3/2} \exp-\left\{\frac{\theta}{2}\sum_{i=1}^{n}\frac{(\mu t_i - 1)^2}{t_i}\right\} \tag{2}$$

Which give the mle for μ and θ as;

$$\hat{\mu} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\overline{T}} \tag{3}$$

and $$\hat{\theta} = \frac{n}{\sum_{i=1}^{n}\frac{(\mu t_i - 1)^2}{t_i}}$$

$$= \frac{n.\overline{T}^2}{\sum_{i=1}^{n}\frac{(t_i - \overline{T})^2}{t_i}} \tag{4}$$

4. ANALYSIS OF THE MODEL

Let H_n be the probability that there are n arrivals during the service time of a customer. Let $H(z)$ denotes the probability generating function (p.g.f.) of H_n given as

$$H(z) = \sum_{n=1}^{\infty} H_n z^n\ ;\left| z \right| \le 1 \tag{5}$$

Following heuristic argument of Kendall (1953) and Gross and Haris (1974), the probability H_n that there are *n* arrivals during the service time is given by,

$$= H_n = \int_0^{\infty} \frac{e^{-\lambda t}\left(\lambda t\right)^n}{n\,!}\left(\frac{\theta}{2\pi t^3}\right)^{1/2} \exp-\left\{\frac{\theta(\mu t - 1)^2}{2.t}\right\} d\,t \tag{6}$$

Then the probability generating function of H_n is

$$H(z) = \sum_{n=0}^{\infty} z^n \int_0^{\infty} \frac{e^{-\lambda t}\left(\lambda t\right)^n}{n\,!}\left(\frac{\theta}{2\pi t^3}\right)^{1/2} \exp-\left\{\frac{\theta(\mu t - 1)^2}{2.t}\right\} d\,t$$

Which, finally comes out to be,

$$H(z) = \exp\left\{\theta\mu\left[1-\left(1+\frac{2(\lambda-\lambda z)}{\theta\mu^2}\right)^{1/2}\right]\right\} \tag{7}$$

The average number of arrivals during the service time is

$$H'(z)\Big|_{z=1} \;=\; \frac{\lambda}{\mu} \tag{8}$$

Let P_n be the probability that there are n customers in the system at the steady state. Let $P(z)$ be the probability generating function of P_n.
Therefore,

$$P(z) = \frac{\left(1-\frac{\lambda}{\mu}\right)(1-z)\exp\left\{\theta\mu\left[1-\left(1+\frac{2(\lambda-\lambda z)}{\theta\mu^2}\right)^{1/2}\right]\right\}}{\exp\left\{\theta\mu\left[1-\left(1+\frac{2(\lambda-\lambda z)}{\theta\mu^2}\right)^{1/2}\right]\right\}-z} \tag{9}$$

By expanding P (z) and collecting the coefficients of z^n, we get P_n the probability that there are n customers in the system.

The probability that the system is empty is

$$P_0 = 1 - \frac{\lambda}{\mu} \tag{10}$$

After substituting estimated value of μ in equation (10), P_0 can be evaluated for different values of λ. It is also observed that P_0 is independent of θ i.e. P_0 is not influenced by the variability of the service time.

The average number of customers in the system is obtained as ,

$$L = P'(z)\Big|_{z=1}$$

$$= \frac{\lambda\left[\,\lambda+\theta\mu\left(\,2\mu-\lambda\right)\right]}{2\left(\mu-\lambda\right)\theta\mu^2} \tag{11}$$

From the equation (11) it can be observed that the average number of customers in the system influenced by θ. The value of L can be computed by using estimated values of μ and θ and different values of λ.

The variability of the system size can be obtained by using the formula.

$$\text{V} = [\text{P''(z)} + \text{P'(z)} - (\text{p'(z)})^2]\Big|_{z=1}$$

$$= \frac{3A^2 + (\rho-1)\{2\rho-3).A - 4.B - 6\rho(2\rho-1)(\rho-1)}{12\left(\rho-1\right)^2} \tag{12}$$

where

$$A = \frac{\lambda^2}{\mu^2}\left(\frac{1}{\theta\mu}+1\right)$$

$$B = \frac{\lambda^3}{\mu^3}\left(\frac{3}{\theta\mu^2}+\frac{3}{\theta\mu}+1\right)$$

and $\rho = \frac{\lambda}{\mu}$`

The coefficient of variation (C.V.) of the system size will be

$$C.V. = \frac{\sqrt{V}}{L} x100 \qquad (13)$$

For the different values of λ and estimated values of μ and θ, the values of variability of the system size and the coefficient of variation can be computed.

It has been studied by Murty (1993) that the variability of the system size decreases as μ increases for fixed values of λ and θ. As λ increases the variability of the system size increases for fixed values of θ and μ, Further it has been observed that the coefficient of variation increases as μ increases for fixed values of 'λ' and 'θ'. As λ increases the coefficient of variation decreases for fixed values of θ and λ. As θ increases the coefficient of variation decreases for fixed values of 'λ' and 'θ'. For given arrival and service rates, the mean queue length of $M/M/1$ and $M/IG/1$ model are compared and it has been observed that when $\frac{1}{\mu} \leq \theta,$ the mean queue length of $M/IG/1$ is less than that of the $M/M/1$ model.

It concludes that by controlling 'θ', the mean queue length of M/IG/1 model can be controlled, which has influence on the optimal operating policies of the system. Further the model can be analyzed in a better by using estimated values of θ and μ in place of hypothetical value of θ and μ. In this model we have used only hypothetical value to λ.

5. QUEUEING MODEL WITH MUKHERJI-ISLAM SERVICE TIME DISTRIBUTION

Again consider a single server queuing with infinite capacity having FCFS (First Come First Serve) queue discipline. The arrivals are assumed to be Poisson with arrival rate λ. But the service time distribution of the process is a new finite range probability distribution which is originally introduced by Mukherjee-Islam (1983) as a life testing model.

$$f(t,\theta,p) = \frac{p}{\theta^p} t^{p-1} \frac{p}{p+1} .\theta \, ; \qquad p,.\theta > 0 \, ,$$

$$t \geq 0 \qquad (14)$$

The above model is monotonic decreasing and highly skewed to the right. The graph is J-shaped thereby showing the uni-model feature. The distribution function of above model is;

$$F(t)=\left[\frac{t}{\theta}\right]^{p}H_n \tag{15}$$

With Mean = $\dfrac{p}{p+1}.\theta$

and Variance = $\dfrac{p}{(p+1)^2(p+2)}.\theta^{\,2}$

6. MAXIMUM LIKELIHOOD ESTIMATES

For the sample of size n, the likelihood function for the model will by

$$L(t:\theta,p)=p^{n}\theta^{-np}\prod_{i=1}^{n}t_i^{p\ -1} \tag{16}$$

Finally the m.l.e. of p is obtained as,

$$\hat{p}=\frac{n}{n\log\theta-\sum\log t_i} \tag{17}$$

And the mle of θ is'

$$\hat{\theta}=t_{(n)}=\max(t_1,t_2,........,t_n) \tag{18}$$

7. ANALYSIS OF THE MODEL

Let H_n be the probability that there are n arrivals during the service time of a customer. Let $H(z)$ denotes the probability generating function (p.g.f.) of H_n given as

$$H(z)=\sum_{n=1}^{\infty}H_n z^n\ ;\left|\,z\,\right|\le 1$$

Following heuristic argument of Kendall (1953) and Gross and Hariss (1974), the probability H_n that there are n arrivals during the service time is given by

$$H_n=\int_0^{\theta}\frac{e^{-\lambda t}\left(\lambda t\right)^n}{n\,!}\left(\frac{p}{\theta^p}\right)t^{p-1}\,d\,t \tag{19}$$

Then the probability generating function of H_n is

$$H(z)=\sum_{n=0}^{\infty}z^n\int_0^{\theta}\frac{e^{-\lambda t}\left(\lambda t\right)^n}{n\,!}\left(\frac{p}{\theta^p}\right)t^{p-1}\,d\,t$$

Which, becomes as;

$$H(z)=p.\sum_{j=0}^{\infty}\frac{\left(-(\lambda-\lambda z)\right)^j}{j\ !}\frac{\theta^j}{p+j} \tag{20}$$

The average number of arrivals during the service time is

$$H'(z)=\Big|_{z=1}\ =\ \frac{p}{p+1}.\theta\lambda \tag{21}$$

Let $\mu=\dfrac{p+1}{p}.\theta$ (the reciprocal of the mean) then

$$H'(z)\Big|_{z=1} = \frac{\lambda}{\mu} \qquad (22)$$

Now, let P_n be the probability that there are 'n' customers in the system at the steady state and $P(z)$ be the probability generating function of P_n. Then by expanding $P(z)$ and collecting the coefficients of Z^n, we get P_n.

Furthermore, the analysis can be carried out in the same manner as in previous section for Inverse Gaussian service time distribution system. The successful application of these models as service time models in queuing theory leads the enlargement of the family of such distributions.

8. COMMENTS AND CONCLUSION

Important parameters of queuing system have been obtained here using two different service time distributions, Inverse Gaussian and Mukherji – Islam distributions. Inverse Gaussian distribution has been used by many workers before in the analysis of a queuing system.

In this article Mukherji-Islam (1983) model has been used first time as the service time distribution in analyzing a queuing model.

The successful analysis of the present queuing system using this distribution, that is a failure time distribution, leads the path to use other failure models also as the service time distribution in the analysis of queuing systems.

In our present article we have analyzed a system with single server where the arrivals follow Poisson distribution and the service time distribution is Inverse Gaussian distribution.

The same system has been analyzed by using Mukherji-Islam (1983) model as the service time model.

Inverse Gaussian distribution has been used in queuing theory but the Mukherjee-Islam distribution which is basically a failure model has been tested first time as the service time model in our study.

In this article parameters of the system like average number of customers in the system, variability and coefficient of variation of system size have been evaluated.

REFERENCES

1. Gross, D. and Harris, C.M. (1974). *Fundamentals of Queuing Theory*, 2nd Edition, John Wiley and Sons, New York.
2. Kendall, D.G. (1953). Stochastic process occurring in the theory of queues and their analysis by the method of imbedded Markov chains. *Ann. Math. Stat.*, 24.
3. Mukherji, S.P. and Islam, R. (1983). A finite range distribution of failures times. *Naval Research Logistics Quarterly,* 30, 487-491.
4. Murty, T.S. (1993). *Some waiting time models with Bulk service*, Ph.D. Thesis, Andhra University, India.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 457-474

SPLIT SAMPLE BOOTSTRAP METHOD

Alamgir Khalil
Department of Statistics University of Peshawar, KPK, Pakistan
Email: profalamgir@yahoo.com

ABSTRACT

The bootstrap technology introduced by Efron (1979) has wide applications, particularly, in regression analysis. It has been used by researchers to construct empirical distributions for estimates of the regression coefficients. In case of outliers in the data, the classical bootstrap procedure fails to give us fine results even if robust regression estimates are used. In this paper we introduced a new bootstrap procedure, called “split sample bootstrap” to handle outliers. The proposed bootstrap procedure gives bootstrap estimates having smaller bootstrap estimates of the standard errors and as a result, we get narrow confidence intervals of the estimates.

1. INTRODUCTION

In regression analysis we come across a situation where we need to construct empirical distributions for estimates of the regression coefficients when analytical distributions cannot be derived, particularly, in case of small sample situation when asymptotic results do not hold. For this purpose, Efron (1979) introduced a very important computational technique called “bootstrap procedure”. In regression scenario, most commonly used bootstrap methods are residual bootstrap and pair bootstrap (also known as classical bootstrap methods). These bootstrap techniques work well when there are no outliers in regression data. The presence of outliers may cause violation of normality assumption. While drawing “B” bootstrap samples with replacement from the original sample, some of the bootstrap samples may consist of larger number of outliers as compared to the number of outliers in the original sample (Willems and Aelst, 2005). The excessive number of outliers in the bootstrap samples definitely poses serious threats to the classical bootstrap estimates of the standard errors of the regression coefficients and hence the bootstrap confidence intervals of the regression estimates are affected, leading to breakdown of the bootstrap methodology. Thus, the problem leads us to draw inaccurate conclusions. In some cases, the percentage of outliers in some of the bootstrap samples may exceed the break down value of the estimation procedure. Hence, even using robust regression estimation procedure may fail to give us satisfactory results in such a situation as almost all robust procedures show resistance to a limited number of outliers present in the data, called breakdown value of the estimation procedure (Willems and Aelst, 2005, Norazan et al., 2009). Willems and Aelst (2005) argued that the classical bootstrap estimates are less robust as compared to even the robust estimators which are bootstrapped.

Several researchers have introduced various remedial bootstrap methods to deal with the presence of outliers. Barrera and Zamar (2002) proposed a “robust and fast bootstrap method” for the class of robust regression estimation procedure. Aelst and Willems

(2002) also proposed similar bootstrap method for S- estimators of multivariate regression. Willems and Aelst (2005) proposed a "short- cut bootstrap procedure" for LTS. This procedure is simple, fast and robust (also known as fast and robust bootstrap method). Willems and Aelst (2005) also suggested another bootstrap procedure, that is, deleting the outliers identified by some initial robust estimation procedure and then to bootstrap the remaining clean data points. This idea led Imon and Ali (2005) to propose another bootstrap method, called, "diagnostic- before- bootstrap method" for regression. According to this procedure, the high leverage points are detected first and are removed before drawing the bootstrap samples. The method ignores the outliers in the bootstrap samples, therefore, by doing so and then to apply the classical bootstrap is likely to under estimate the standard errors or may result in invalid inferential statements and predictions intervals. Recently, Midi et al. (2009) suggested a modified bootstrap method which they called "Dynamic Robust Bootstrap for LTS (DRBLTS)". This new bootstrap method is based on LTS estimators and is used to control the percentage of outliers in bootstrap samples. This method again does not take into account the under estimation of the standard errors of the estimates. Most recently, Norazan et al. (2009) presented a new bootstrap method which they called "weighted bootstrap with probability (WBP)" in regression. According to WBP method, the outlying observations are less likely to be selected in the bootstrap samples as small probabilities are associated with outlying observations and hence the method is expected to minimize the effect of outlying observations on the proposed bootstrap method.

In this paper, we first propose a new bootstrap method that we called "Split- Sample Bootstrap (SSB)" method which ensures that the number of outliers in each of the bootstrap samples equals the number of outliers in the original sample. The new method will then be used to construct a bootstrap distribution of Aalmgir redescending M- estimator (ALARM). This ALARM estimator has the weight function given below:

$$\mathrm{w(r)} = \begin{cases} \dfrac{16e^{-2(r/c)^2}}{(1+e^{-(r/c)^2})^4} & ,|r| \le c \\ 0 & ,|r| > c \end{cases} \tag{1}$$

where r is the residual of the fitted model and the parameter C is the tuning content and is set equal to 3 to have 95% efficiency at normal case. This estimator has slow decaying ψ -function given by

$$\psi(\mathrm{r}) = \begin{cases} \dfrac{16re^{-2(r/c)^2}}{(1+e^{-(r/c)^2})^4} & ,|r| \le c \\ 0 & ,|r| > c \end{cases} \tag{2}$$

The estimator based on this ψ - function is very robust and efficient.

The SSB method is based on the bootstrap algorithm using redescending ALARM estimator to estimate the model parameters. In section 2, we discuss some bootstrap procedures existing in the literature along with the proposed bootstrap method.

2. SOME BOOTSTRAP PROCEDURES

Let us consider the general regression model

$$Y = X\beta + \varepsilon \tag{3}$$

where Y is a $(n \times 1)$ vector of responses, X is a $(n \times k)$ data matrix containing values of the predictors ($k = p+1$), the parameter vector β is $(k \times 1)$ (including the intercept) which is to be estimated from the observed data and ε is a $(n \times 1)$ vector of unobservable random error terms.

We discuss various bootstrap procedures in brief in the following section.

ROBUST RESIDUAL BOOTSTRAP (RRB1)

We define both fixed-x and random- x bootstrap procedures here with a slight modification, that is, instead of OLS estimator $\hat{\beta}$ is chosen as the resdecending ALARM estimator as the OLS estimation procedure is greatly affected by the presence of outliers and consequently the performance of bootstrap procedure is badly affected.

We summarize the residuals bootstrap scheme in steps as follows:

Step-1: By assuming an original sample, fit the robust regression model (using the ALARM estimate) to the sample to estimate the regression coefficients $\hat{\beta}_{(A)}$ and the fitted values $\hat{y}$.

Step-2: Use the observed y_i and $\hat{y}$ to compute the residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Step-3: Draw $\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_n^*$ randomly by resampling from $\hat{\varepsilon}_i$ to compute $\hat{y}_i^*$ using the relation

$$y_i^* = f(x_i, \hat{\beta}_{(A)}) + \varepsilon_i^*$$

Step-4: Fit the regression model to $(X,\ y^*)$ to get $\hat{\beta}_{(A)}^*$.

Step-5: Replicate step-3 and step-4 B times to obtain the bootstrap replicates

$$\hat{\beta}_{(A)}^{*1}, \hat{\beta}_{(A)}^{*2}, \ldots, \hat{\beta}_{(A)}^{*B}.$$

ROBUST RANDOM- X (PAIR) BOOTSTRAP (RRB2)

For the model given in (3), we consider a sample of n observations $z_i' = (y_i,\ x_{i1},\ x_{i2}, \ldots,\ x_{ip}),\ i = 1,\ 2, \ldots,\ n.$ The following steps describe the procedure for random-*x* bootstrap samples based on robust estimation procedure.

Step-1: Select a bootstrap sample from z_i'. The bootstrap sample observations are drawn independently from the original sample with equal probabilities $1/n$ and are denoted by $z_1^{*\prime},\ z_2^{*\prime}, \ldots,\ z_n^{*\prime}$.

Step-2: Fit the regression model to $z_1^{*\prime}, z_2^{*\prime}, \ldots, z_n^{*\prime}$ and compute the regression coefficients $\hat{\beta}^*_{(A)}$.

Step-3: Replicate the above steps B times in order to compute the bootstrap replicates $\hat{\beta}^{*1}_{(A)}, \hat{\beta}^{*2}_{(A)}, \ldots, \hat{\beta}^{*B}_{(A)}$.

The above two bootstrap procedures are commonly known as classical bootstrap procedures.

WEIGHTED BOOTSTRAP WITH PROBABILITY (WBP)

In this method, Hampel weight function is used for assigning weights to data points and to identify the outlying observations. According to WBP, the selection probability of the *i*th observation is given by

$$p_i = \frac{w_i}{\sum_{j=1}^{n} w_j}, \; 0 \le p_i \le 1 \text{ and} i = 1, 2, \ldots, n. \tag{5}$$

Thus, through this weight mechanism, outliers were controlled in the bootstrap procedure by assigning probabilities $p_1, p_2, \ldots, p_n$ to $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$. Prior to describing the WBP procedure, we follow some notations used in WBP. Let D denotes the set of deleted data points and R denotes the set of remaining observations where R contains only those data points for which $p_i > 0$. Let $\hat{\beta}^{(-D)}$ be the estimate of the parameter obtained by fitting the regression model from the observations after deleting *d* cases, that is, (X_R, Y_R). The WBP procedure is described as follows:

Step-1: Use LMS to fit the original data. Use the LMS residuals and apply Hampel weight function to identify outlying observations. Obtain $\hat{\beta}^{(-D)}$ from the remaining data points (with $w_i > 0$) and the fitted values

$$\hat{y}^{(-D)} = f(x_i, \hat{\beta}^{(-D)}).$$

Step-2: Based on the fitted values, obtain the residuals $\hat{\varepsilon}_i^{(-D)} = y_i - f(x_i, \hat{\beta}^{(-D)})$.

Step-3: Draw $\hat{\varepsilon}_1^{*(-D)}, \hat{\varepsilon}_2^{*(-D)}, \ldots, \hat{\varepsilon}_n^{*(-D)}$ randomly by resampling from $\hat{\varepsilon}_i$ to compute y_i^* using the relation $y_i^* = f(x_i, \hat{\beta}_{(A)}) + \hat{\varepsilon}_i^{*(-D)}$.

Step-4: Obtain $\hat{\beta}^{*(-D)}$ by regressing the bootstrapped values $y_i^{*(-D)}$ on the fixed X_R.

Step-5: Repeat Step-3 and Step-4 B times to obtain $\hat{\beta}^{*1(-D)}, \hat{\beta}^{*2(-D)}, \ldots, \hat{\beta}^{*B(-D)}$.

THE PROPOSED BOOTSTRAP PROCEDURE: SPLIT SAMPLE BOOTSTRAP (SSB)

Several researchers have considered the theory for bootstrapping the distributions of robust regression estimates. For more details, see Shorack (1982), Parr (1985), Yang (1985), Shao (1990), Liu and Singh (1992), Barrera and Zamar (2002), among others.

Barrera and Zamar (2002) argued that the asymptotic variances of could be used to estimate the standard error of robust estimates only when the central normal model holds which in the presence of outliers does not hold. Numerical instability of the robust regression estimates may be one of the problems if the percentage of outliers in the bootstrap sample is higher than in the original sample irrespective of the robustness of the estimation procedure being used and irrespective of the robustness of the estimate being bootstrapped (Barrera and Zamar, 2002). Several researchers proposed various strategies so that the bootstrap procedure becomes less sensitive to the presence of outliers in the sample. See, for example, Singh (1998), Stromberg (1997), Amado and Pires (2004), Imon and Ali (2005) and Norazan et al. (2009) for more details.

In the current study we propose a new and simple bootstrap procedure called "Split Sample Bootstrap Procedure (SSB)" to protect the bootstrap procedure against undue effect of the increased number of outliers in some bootstrap samples in order to avoid numerically instability of the robust regression estimates. The basic idea of SSB is to first identify the exact number of outliers, say n_1, present in the original sample using LTS procedure and then to form two groups of observations; one group (G_1) consisting of n_1 outliers and the second group (G_2) consisting of n_2 (= n- n_1) remaining observations (clean data points) such that the number of observations in both group equals n (sample size). The bootstrap samples are drawn from these two groups. That is, n_1 observations in each bootstrap are drawn from G_1 and n_2 clean data points from G_2 such that the size of each bootstrap sample equals n. ALARM estimator is used to estimate the model parameters for each bootstrap sample. Our proposed bootstrap procedure not only protects the bootstrap procedure against excessive number of outliers but also ensures efficiency as it does not loose information (all observations are retained) collected from all data points. The following steps describe our proposed bootstrap procedure:

Step-1: Given n data points in the sample. Use LMS to fit the original data. Use the standardized LMS residuals to identify the outlying observations in the sample data. Obtain $\hat{\beta}_{(A)}$ from the data points and the fitted values

$$\hat{y}_i = f(x_i,\ \hat{\beta}_{(A)}).$$

Step-2: Based on the fitted values, obtain the residuals $\hat{\varepsilon}_i = y_i - f(x_i,\ \hat{\beta}_{(A)})$. Split all the residuals and group them in to two groups G_1 (consisting of n_1 residuals associated with outliers) and G_2 (consisting of n_2 remaining residuals associated with clean data points).

Step-3: Draw $\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, \ldots, \hat{\varepsilon}_n^*$ randomly by resampling from $\hat{\varepsilon}_i$ (of which n_1 are drawn from G_1 and n_2 from G_2) to compute y_i^* using the relation $y_i^* = f(x_i, \hat{\beta}_{(A)}) + \hat{\varepsilon}_i^*$.

Step-4: Obtain $\hat{\beta}_{(A)}^*$ by regressing the bootstrapped values y_i^* (using the ALARM estimator) on the fixed X_R.

Step-5: Repeat Step-3 and Step-4 B times to obtain $\hat{\beta}_{(A)}^{*1}, \hat{\beta}_{(A)}^{*2}, \ldots, \hat{\beta}_{(A)}^{*B}$.

Using the bootstrap replications obtained in Step-5 one can construct the bootstrap distribution of the ALARM estimator and hence the distribution can be used to compute the bootstrap standard error and bootstrap confidence intervals for the parameters estimates.

3. ASSESSMENTS OF VARIOUS BOOTSTRAP PROCEDURES

In order to evaluate the performance of various bootstrap procedures, we compute bias, standard error and length of the confidence interval for the parameters estimates. The estimation procedure is our proposed redescending M- estimator, that is, ALARM estimator. Let $\hat{\beta}_{(A)}$ be the ALARM estimator for the parameter β computed from a given sample. The RRB1 estimate of β is computed using the expression

$$\hat{\beta}^{*}_{(A)} = \sum_{b=1}^{B} \hat{\beta}^{*b}_{(A)} / B$$

and its bias is given by

$$Bias(\hat{\beta}^{*}_{(A)}) = \hat{\beta}^{*}_{(A)} - \hat{\beta}_{(A)}$$

The bootstrap standard error of the estimate is given by

$$SE(\hat{\beta}^{*}_{(A)}) = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\beta}^{*b}_{(A)} - \hat{\beta}^{*}_{(A)})^2}{B}}$$

Let $\left[\hat{\beta}_{(A)}{}^{*(\alpha/2)},\ \hat{\beta}_{(A)}{}^{*(1-\alpha/2)}\right]$ be the $100(1-\alpha)\%$ bootstrap percentile interval for β. Then the length (L) of the interval is defined as

$$L = \hat{\beta}_{(A)}{}^{*(1-\alpha/2)} - \hat{\beta}_{(A)}{}^{*(\alpha/2)}$$

The calculations of estimates for other bootstrap procedures are the same, the difference being only in the notations.

4. NUMERICAL RESULTS AND DISCUSSION

We consider real data examples and simulation studies to illustrate and compare the performance of our proposed bootstrap procedure with WBP and RRB1.

REAL DATA EXAMPLES

In this section we consider two examples based on real data sets available in the literature which have been extensively analyzed by researchers for identification of outliers and for measuring performance of various robust methods developed by researchers. We also compare the performance of our newly proposed bootstrap procedure with RRB1 and WBP using the data sets in the two examples given below.

Example 1(Hawkins, Bradu, and Kass Data)

In this example we consider a data set artificially generated by Hawkins, Bradu, and Kass (1984). This data consist of 75 observations having one response and three predictors. The first 10 observations in the data set were generated as outliers. Our proposed ALARM estimator detects all 10 observations as outliers.

Figures 1 and 2 show how well ALARM estimator identifies all 10 outliers. Here we apply our newly proposed bootstrap procedure as well as RRB1 and WBP to this data set be performing B= 2000 bootstrap replications in all cases. We present bootstrap SE’s, confidence intervals (CI) and lengths (L) for all three bootstrap procedures in Tables 1- 3. As expected, our proposed bootstrap procedure (SSB) performs better than the rest of the two procedures. The bootstrap SE’s of the estimates are the smallest among all. Our proposed procedure also provides shortest bootstrap confidence intervals for all parameters. The WBP is a good competitor of SSB as compared to RRB1. WBP provides more efficient results than RRB1 but our proposed SSB outperforms other two bootstrap procedures.

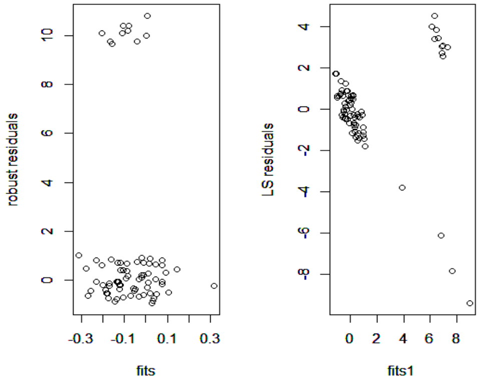


Fig 1: Robust and OLS residuals versus fitted values

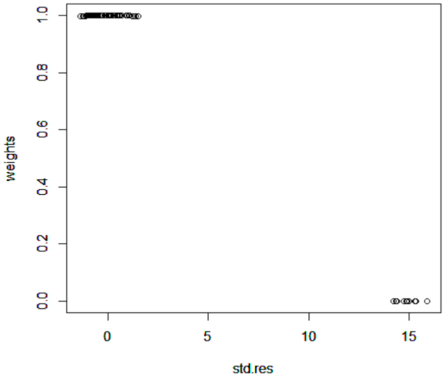


Fig 2: Robust weights versus Std. Residuals

Table 1:
Robust Parameter Estimates obtained from Hawkins-Bradu-Kass data

Parameter	OLS	ALARM
β_0	-0.388	-0.181
β_1	0.239	0.081
β_2	-0.335	0.039
β_3	0.383	-0.051

Table 2:
Bootstrap standard Errors of Parameter Estimates obtained from Hawkins-Bradu-Kass data

Parameter	RRB1	WBP	SSB
β_0	4.7265	1.6121	0.3614
β_1	0.9203	0.6974	0.0323
β_2	1.8763	0.7103	0.0745
β_3	0.8133	0.8137	0.1168

Table 3:
95% Bootstrap Confidence Intervals of the Parameter Estimates Obtained from Hawkins-Bradu-Kass data

Parameter	RRB1	WBP	SSB
β_0	(-3.7833, 0.2343) **(4.0176)**	(-0.2388, 1.2369) **(1.4757)**	(-0.9844, -0.1410) **(0.8436)**
β_1	(-2.4692, 1.0672) **(3.5364)**	(-0.2892, 0.5179) **(0.8071)**	(0.0533, 0.1691) **(0.1158)**
β_2	(-1.8128, 2.2731) **(4.0858)**	(-0.5598, 0.4353) **(0.9951)**	(0.0273, 0.2179) **(0.1906)**
β_3	(-0.2333, 2.2802) **(2.5135)**	(-0.1003, 0.7563) **(0.8567)**	(-0.0720, 0.2055) **(0.2776)**

Example 2 (Condroz data)

The Condroz sample data provide information about the pH- value and the Calcium (Ca) contents in soil samples (Goegebeur, Planchon, Beirlant and Oger, 2005). The same data have been reported and analyzed under the context of robust estimation for identifying outliers by Vandewalle et al. (2004). We have considered a subset of the entire data containing 428 observations where the pH- value lies between 7.0 and 7.5. It can be seen from the plot of Calcium content that the distribution of the content is skewed. Waive

We initially used our robust estimator, ALARM to estimate model parameters and the results are presented in Table 4. We compute the bootstrap estimates of standard errors (SE's) based on all three procedures and are presented in Table 5. We also have computed the bootstrap confidence intervals and their lengths for the parameters from the data and are presented in Table 6. The results in Table 5 and Table 6 clearly show that due the presence of outliers in the data, RRB1 produce very inefficient results giving largest SE's and widest confidence intervals.

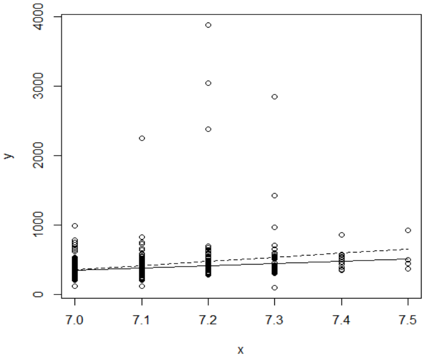


Fig 3: Plot of the Condroz data

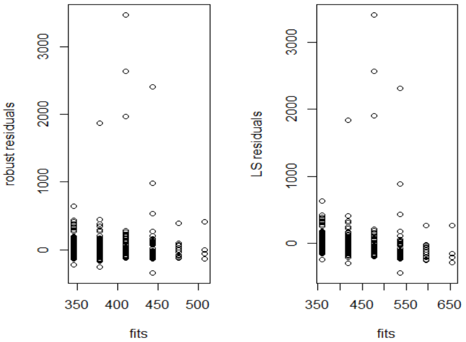


Fig 4: Robust and OLS residuals versus fitted values

Table 4:
Robust Parameter Estimates obtained from Condroz data

Parameter	OLS	ALARM
β_0	-3783.04	-1920.23
β_1	591.68	323.65

Table 5:
Bootstrap standard Errors of Parameter Estimates for Condroz data

Parameter	RRB1	WBP	SSB
β_0	48.62	56.30	49.78
β_1	6.60	6.03	4.87

Table 6:
95% Bootstrap Confidence Intervals of the Parameter Estimates for Condroz data

Parameter	RRB1	WBP	SSB
β_0	(-2051.25, -1831.09) **(220.16)**	(-2040.08, -1836.52) **(203.56)**	(-2032.79, -1852.29) **(180.50)**
β_1	(313.15, 370.89) **(57.74)**	(319.97, 342.64) **(22.67)**	(314.18, 331.12) **(16.94)**

The proposed procedure SSB produce numerically stable results and outperforms the other two bootstrap procedures.

5. SIMULATION STUDIES

We have considered real data examples for investigating the performance of our newly proposed split sample bootstrap procedure. We also carryout simulation studies

SIMULATION DESIGN

We carry out an extensive simulation study to assess the robustness and efficiency properties of our proposed bootstrap procedure, SSB and to compare its performance with various bootstrap procedures existing in the literature. We have considered four (4) different simulation designs in this study based on changes in the error and leverage structure. In all these designs, the regressors are generated from N (0, 1) distribution and contamination is done only in Y- space.

(1) Single outlier ($\lambda = 1$) scenario: Errors are generated from the N (0, 1) distribution and the last observation is generated from N (100, 1) as an outlier.

(2) Multiple outliers ($\lambda = 10\%$) scenario: Errors are generated from N (0, 1) distribution and the last observations 10% observations are generated from N (100, 1) as outliers.

(3) Multiple outliers ($\lambda = 20\%$) scenario: Errors are generated from N (0, 1) distribution and the last observations 20% observations are generated from N (100, 1) as outliers.

(4) Multiple outliers ($\lambda = 40\%$) scenario: Errors are generated from N (0, 1) distribution and the last observations 10% observations are generated from N (100, 1) as outliers.

For all designs stated above, Y is computed using the expression (1) by taking $\beta_0 = \beta_1 = ... = \beta_p = 2,$ for $p = 1, 2, 3.$ We consider n = 50, 100 and 200. In all the competing bootstrap procedures considered in this study, the redescending ALARM estimator is used to estimate the parameters of the regression model. The tuning constant parameter for the ALARM estimator is chosen so that 95% efficiency is obtained at normal distribution. MAD is used as a scale estimate for population variance. For comparison purpose, the bias, standard error (SE) and length (L) of the confidence are computed for the estimates based on 1000 bootstrap samples and are averaged over 1000 simulations. The results are given in Table 7-12.

Table 7:
Bias and Standard Error for the bootstrap estimates ($\lambda = 1$)

n	*P*	Parameter	RRB1		WBP		SSB	
			Bias	SE	Bias	SE	Bias	SE
50	1	β_0	0.030	1.897	0.015	0.441	0.009	0.039
		β_1	0.076	1.859	-0.019	0.788	0.015	0.039
	2	β_0	-0.018	1.168	0.016	0.113	0.010	0.035
		β_1	0.014	1.046	-0.017	0.929	0.011	0.039
		β_2	0.012	1.020	0.009	0.745	0.007	0.033
	3	β_0	-0.025	1.631	-0.101	0.618	0.008	0.039
		β_1	-0.018	1.129	0.016	0.650	0.013	0.040
		β_2	0.016	1.066	0.011	0.820	0.011	0.040
		β_3	-0.025	1.091	0.016	0.773	0.011	0.040
100	1	β_0	-0.025	1.059	0.006	0.182	0.011	0.018
		β_1	-0.015	1.160	-0.003	0.117	0.014	0.019
	2	β_0	0.019	1.117	0.012	0.125	0.012	0.018
		β_1	0.014	1.022	0.009	0.303	0.006	0.028
		β_2	-0.010	1.018	-0.060	0.200	0.008	0.029
	3	β_0	0.017	1.026	-0.062	0.206	0.006	0.039
		β_1	-0.011	1.023	0.010	0.150	0.007	0.029
		β_2	0.012	1.035	-0.007	0.256	0.013	0.016
		β_3	-0.015	1.021	0.010	0.303	0.002	0.028
200	1	β_0	-0.013	0.238	0.010	0.107	0.003	0.010
		β_1	0.009	0.260	0.005	0.117	0.001	0.010
	2	β_0	-0.014	0.294	0.009	0.029	0.005	0.019
		β_1	-0.008	0.128	0.006	0.038	0.004	0.019
		β_2	-0.008	0.284	-0.008	0.038	0.003	0.011
	3	β_0	0.010	0.169	-0.002	0.033	0.004	0.009
		β_1	0.005	0.174	0.001	0.035	0.001	0.010
		β_2	0.008	0.192	0.002	0.049	0.002	0.020
		β_3	-0.016	0.103	-0.005	0.034	0.002	0.010

Table 8:
Bias and Standard Error for the bootstrap estimates ($\lambda = 10\%$)

n	*P*	Parameter	RRB1		WBP		SSB	
			Bias	SE	Bias	SE	Bias	SE
50	1	β_0	0.430	2.097	-0.021	0.541	0.014	0.044
		β_1	-0.376	1.955	-0.019	0.888	0.003	0.044
	2	β_0	0.416	1.669	-0.026	0.312	-0.021	0.042
		β_1	-0.313	1.646	-0.017	0.529	-0.007	0.047
		β_2	0.312	1.220	0.109	0.847	0.006	0.039
	3	β_0	0.425	1.832	-0.101	0.728	-0.011	0.044
		β_1	-0.219	1.159	0.026	0.950	0.013	0.044
		β_2	-0.417	1.664	0.111	0.820	0.010	0.045
		β_3	-0.325	1.593	0.026	0.773	-0.005	0.046
100	1	β_0	-0.325	1.153	-0.016	0.282	0.007	0.022
		β_1	-0.214	1.262	0.013	0.216	0.006	0.022
	2	β_0	0.219	1.217	0.012	0.225	0.002	0.022
		β_1	0.214	1.121	0.010	0.303	-0.007	0.022
		β_2	0.213	1.114	-0.110	0.210	-0.015	0.022
	3	β_0	-0.217	1.129	-0.016	0.217	0.007	0.022
		β_1	0.311	1.124	0.012	0.215	0.005	0.022
		β_2	0.111	1.136	-0.017	0.256	0.008	0.022
		β_3	-0.115	1.221	0.010	0.403	-0.003	0.022
200	1	β_0	-0.113	0.738	0.010	0.211	-0.009	0.011
		β_1	0.109	0.860	0.011	0.127	0.006	0.011
	2	β_0	-0.114	0.794	0.009	0.129	0.006	0.011
		β_1	-0.109	0.729	0.006	0.137	0.003	0.012
		β_2	-0.208	0.888	-0.008	0.038	0.001	0.012
	3	β_0	0.110	0.693	-0.006	0.033	0.002	0.010
		β_1	0.103	0.740	0.006	0.065	0.004	0.013
		β_2	0.108	0.920	0.008	0.057	0.002	0.011
		β_3	-0.116	0.903	-0.005	0.065	0.004	0.013

Table 9:
Bias and Standard Error for the bootstrap estimates (λ = 20 %)

n	*P*	Parameter	RRB1		WBP		SSB	
			Bias	SE	Bias	SE	Bias	SE
50	1	β_0	-0.540	3.198	0.023	0.561	0.018	0.054
		β_1	0.474	2.956	0.029	0.866	0.012	0.048
	2	β_0	1.516	2.369	-0.026	0.342	-0.011	0.052
		β_1	0.630	3.546	0.027	0.429	-0.012	0.059
		β_2	0.312	2.820	0.102	0.657	0.013	0.053
	3	β_0	-0.445	2.842	0.141	0.821	-0.011	0.044
		β_1	0.439	2.489	0.029	0.971	0.016	0.058
		β_2	-0.817	2.667	-0.110	0.820	0.013	0.055
		β_3	-0.627	2.993	0.029	0.873	-0.015	0.058
100	1	β_0	0.425	1.537	0.010	0.380	0.009	0.044
		β_1	0.514	1.628	-0.010	0.226	0.006	0.042
	2	β_0	-0.329	1.754	0.016	0.238	0.008	0.032
		β_1	0.321	1.670	0.015	0.312	-0.008	0.043
		β_2	0.243	2.14	0.110	0.211	-0.015	0.031
	3	β_0	0.217	2.109	0.022	0.224	0.008	0.043
		β_1	-0.311	2.224	-0.021	0.206	0.008	0.042
		β_2	-0.211	1.956	-0.019	0.354	0.008	0.050
		β_3	0.215	2.221	0.013	0.408	-0.007	0.042
200	1	β_0	0.203	1.433	0.011	0.111	-0.009	0.021
		β_1	-0.199	1.686	0.010	0.144	0.007	0.021
	2	β_0	0.194	1.423	-0.011	0.136	0.007	0.022
		β_1	-0.179	1.290	-0.011	0.157	0.008	0.012
		β_2	-0.208	0.888	0.008	0.088	0.005	0.024
	3	β_0	0.122	1.293	-0.016	0.033	0.004	0.021
		β_1	0.138	1.170	-0.009	0.094	0.004	0.029
		β_2	0.187	1.205	0.011	0.069	0.005	0.025
		β_3	-0.163	1.203	0.008	0.095	0.007	0.026

Table 10:
Bias and Standard Error for the bootstrap estimates ($\lambda = 40\%$)

n	*P*	Parameter	RRB1		WBP		SSB	
			Bias	SE	Bias	SE	Bias	SE
50	1	β_0	17.194	11.18	-0.436	1.393	0.019	0.074
		β_1	-1.915	8.867	0.052	0.871	-0.018	0.077
	2	β_0	20.91	13.25	-0.448	1.614	0.019	0.077
		β_1	-1.939	6.34	-0.75	0.834	0.023	0.081
		β_2	-0.948	5.40	0.011	0.834	-0.015	0.083
	3	β_0	21.285	13.52	-0.371	1.926	-0.016	0.078
		β_1	2.122	7.12	0.027	1.839	-0.018	0.079
		β_2	-1.066	6.54	0.078	1.887	0.019	0.081
		β_3	0.987	7.35	-0.068	1.820	0.019	0.082
100	1	β_0	10.840	10.67	0.113	0.746	-0.011	0.055
		β_1	1.898	3.228	-0.034	0.561	-0.016	0.066
	2	β_0	11.650	12.6	0.093	0.708	-0.009	0.055
		β_1	-0.881	5.104	-0.107	0.586	-0.012	0.056
		β_2	-0.513	5.232	0.005	0.573	-0.012	0.057
	3	β_0	12.42	8.901	0.071	0.680	-0.011	0.056
		β_1	0.842	5.137	-0.045	0.584	-0.009	0.057
		β_2	-0.869	5.199	-0.028	0.575	-0.010	0.058
		β_3	-0.901	5.234	0.043	0.560	0.010	0.047
200	1	β_0	8.520	5.03	0.079	0.442	-0.006	0.037
		β_1	-1.078	2.736	0.063	0.405	-0.008	0.037
	2	β_0	9.850	5.321	0.045	0.463	0.007	0.037
		β_1	0.141	3.744	0.007	0.401	0.006	0.036
		β_2	-0.655	3.806	-0.031	0.392	0.006	0.038
	3	β_0	10.023	8.095	-0.139	0.439	0.007	0.038
		β_1	-0.697	4.834	0.048	0.389	-0.007	0.037
		β_2	-0.709	4.854	0.060	0.383	-0.005	0.0139
		β_3	-0.751	3.806	0.008	0.411	0.006	0.038

Table 11:
Average Length of the bootstrap Confidence Interval ($\lambda = 1\&\ 10\%$)

n	*P*	Parameter	$\lambda = 1$			$\lambda = 10\%$		
			RRB1	WBP	SSB	RRB1	WBP	SSB
			L	*L*	*L*	*L*	*L*	*L*
50	1	β_0	2.531	1.383	0.153	2.531	1.582	0.173
		β_1	2.528	1.588	0.161	2.528	1.787	0.179
	2	β_0	2.666	1.842	0.134	2.666	2.842	0.170
		β_1	3.172	1.176	0.169	3.172	1.376	0.194
		β_2	3.059	1.036	0.135	3.059	1.237	0.162
	3	β_0	3.971	1.830	0.153	3.971	1.930	0.173
		β_1	2.468	1.863	0.163	2.468	1.993	0.179
		β_2	1.168	1.719	0.163	1.168	1.989	0.184
		β_3	2.290	1.950	0.162	2.290	2.950	0.184
100	1	β_0	1.380	0.449	0.073	1.380	0.849	0.086
		β_1	1.084	0.072	0.078	1.084	0.672	0.093
	2	β_0	1.578	0.902	0.073	1.578	0.902	0.084
		β_1	1.336	0.658	0.077	1.336	0.998	0.090
		β_2	1.572	0.360	0.078	1.572	0.866	0.090
	3	β_0	1.487	0.788	0.075	1.487	1.117	0.085
		β_1	1.520	0.485	0.078	1.520	0.885	0.091
		β_2	1.719	0.417	0.079	1.719	0.917	0.089
		β_3	1.313	0.542	0.079	1.313	0.742	0.089
200	1	β_0	0.925	0.280	0.037	0.925	0.440	0.043
		β_1	1.015	0.331	0.038	1.015	0.431	0.044
	2	β_0	0.807	0.126	0.037	0.807	0.326	0.042
		β_1	0.892	0.878	0.039	0.892	0.878	0.045
		β_2	0.046	0.016	0.039	0.046	0.416	0.045
	3	β_0	0.120	0.087	0.037	0.120	0.287	0.043
		β_1	0.126	0.154	0.039	0.126	0.154	0.045
		β_2	0.283	0.080	0.034	0.283	0.089	0.046
		β_3	0.124	0.055	0.037	0.124	0.087	0.046

Table 12:
Average Length of the bootstrap Confidence Interval ($\lambda = 20\% \,\&\, 40\%$)

n	*P*	Parameter	$\lambda = 20\ \%$			$\lambda = 40\%$		
			RRB1	WBP	SSB	RRB1	WBP	SSB
			L	*L*	*L*	*L*	*L*	*L*
50	1	β_0	3.541	1.812	0.208	9.635	4.626	0.288
		β_1	3.228	1.837	0.225	8.960	3.039	0.314
	2	β_0	3.656	1.841	0.226	10.62	4.377	0.301
		β_1	3.982	1.760	0.222	8.917	2.902	0.328
		β_2	3.259	1.371	0.220	8.633	2.917	0.333
	3	β_0	3.871	1.935	0.226	10.93	4.434	0.309
		β_1	4.168	2.293	0.228	8.982	2.877	0.318
		β_2	3.016	2.089	0.227	8.543	3.100	0.326
		β_3	3.290	2.150	0.221	8.149	2.896	0.326
100	1	β_0	2.530	0.942	0.099	6.685	2.522	0.137
		β_1	2.684	0.722	0.107	5.623	1.960	0.148
	2	β_0	2.257	0.913	0.100	7.931	2.439	0.139
		β_1	2.235	0.995	0.106	7.362	2.045	0.146
		β_2	2.677	0.966	0.107	7.685	2.017	0.151
	3	β_0	1.987	1.217	0.102	5.310	2.345	0.142
		β_1	2.522	0.984	0.107	7.185	2.032	0.154
		β_2	2.708	0.938	0.108	7.534	1.993	0.158
		β_3	2.316	0.942	0.108	7.688	1.953	0.158
200	1	β_0	1.925	0.540	0.049	3.402	1.536	0.066
		β_1	1.815	0.531	0.053	3.419	1.408	0.071
	2	β_0	1.917	0.425	0.049	4.435	1.594	0.068
		β_1	1.929	0.670	0.053	4.508	1.401	0.074
		β_2	1.146	0.326	0.054	3.662	1.360	0.072
	3	β_0	1.712	0.387	0.051	5.360	1.506	0.069
		β_1	1.162	0.254	0.054	4.789	1.358	0.073
		β_2	1.816	0.189	0.054	3.771	1.338	0.075
		β_3	1.912	0.286	0.055	3.698	1.436	0.074

The results presented in Table 7 show that, for small sample, the classical bootstrap procedure RRB1 does not perform well as compared to WBP and SSB when the data is contaminated by even single outlier in Y-space as it biases as well as standard errors are largest among the three procedures. But its performance gets improved with increasing sample size and hence all the three methods are close enough as far as their biases and standard errors are concerned for larger sample sizes. WBP performs better than RRB1 but its performance is poorer than our proposed procedure for all sample sizes in case of single outlier in the data. The SE’s and biases are the smallest for SSB among all and decrease further with increasing sample sizes. The RRB1 also produce widest confidence intervals as compared to WBP and SSB, particularly, for smaller sample sizes. SSB produces shortest length confidence interval among the three procedures. From the results presented in Table 8 one can clearly see that the RRB1 is more sensitive to increasing percentage of outliers ($\lambda = 10\%$) as compared to WBP and SSB as the increase in the biases and SE's for RRB1 is much greater than the increase in biases and SE's for WBP and SSB. The proposed procedure seems to be the least sensitive among the three with respect to its biases, SE's and length (L). However, in case of $\lambda = 10\%$, WBP and SSB give almost identical results for larger sample sizes.

The biases and SE's for WBP and SSB are consistently the smallest for all *P*, *n*, and percentage of outliers λ. As the percentage of outliers increases from $\lambda = 10\%$, not only the RRB1 but also WBP performs poorly as their biases, SE's and *L* tend to increase significantly. Only our proposed procedure shows resistance against increasing number of outliers for all *n*, *P* and λ. However, WBP produces better results as compared to RRB1 but not as promising as our proposed method does. WBP protects against outliers up to some extent when $\lambda = 1$, 10%, but things become worse for larger $\lambda = 20\%$ and 40%, as it yields poor results in terms of biases, SE's and lengths of the bootstrap confidence intervals. From all these results it is very clear that RRB1 is very sensitive to outliers present in the data, the WBP is the second most sensitive procedure showing less resistance to higher percentage of outliers. The biases, SE's are the largest for RRB1, second largest for WBP and the smallest for SSB. Our proposed method is very robust showing high protection against outliers even up to $\lambda = 40\%$. It is hardly affected by the presence of even higher percentages of outliers as it produces minimum bias and gives smallest SE in all scenarios. Shortest bootstrap confidence intervals are obtained even for highly contaminated data. The results are true for 1000 bootstraps and for every *n* = 50, 100 and 200.

6. COMMENTS AND CONCLUSION

The presence of outliers in the data requires a very comprehensive and details examination not only for usual regression analysis but also for bootstrap procedures. In the present study, we have introduced a new bootstrap procedure, called “split sample bootstrap (SSB)” for regression analysis to get enhanced protection against outliers and to get numerically stable results. We presented two numerical examples and conducted simulation studies to evaluate the performance of our proposed procedure. The results obtained from both sample data and simulated data clearly show that SSB is a better choice as compared to RRB1 and WBP, particularly, when the data contain higher percentage of outliers. The proposed bootstrap procedure is a very good robust alternative to other bootstrap procedures.

REFERENCES

1. Amado, C. and Pires, A.M. (2004). Robust bootstrap with nonrandom weights based on the influence function. *Commun. in Statist., Simula. and Comput.*, 33, 377-396.
2. Efron, B. (1979). Bootstrap Methods. Another Look at the Jackknife. *Ann. Statist.*, 7, 1-26.
3. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics,* 26, 197-208.

4. Goegebeur, Y., Planchon, V., Beirlant, J. and Oger, R. (2005). Quality Assessment of Pedochemical Data Using Extreme Value Methodology. *Journal of Applied Science,* 5, 1092-1102.
5. Imon, A.H.M.R. and Ali, M.M. (2005). Bootstrapping regression residuals. *J. Korean Data Inform. Sci. Soc.*, 16, 665-682.
6. Liu, R.Y. and Singh, K. (1992). Efficiency and robustness in resampling. *Ann. Statist*., 20, 370-384.
7. Midi, H., Uraibi, H.S. and Talib, B.A. (2009). Dynamic Robust Bootstrap Method Based on LTS Estimators. *Euro. J. Scientific Res.*, 32(3), 277-287.
8. Norazan, M.R., Habsha, M. and Imon, A.H.M.R. (2009). Weighted Bootstrap with Probability in Regression. *Proceeding of the 8th WSEAS International conference on Applied Computer and Applied Computational Science*, 2009. ISSN: 1790- 5117.
9. Barrera, M. and Zamar, R. (2002). Bootstrapping robust estimates of regression. *J. Ann. Statist.*, 30(2), 556-582.
10. Parr, W.C. (1985). The bootstrap: Some large sample theory and connections with robustness. *Statist. Probab. Lett.,* 3, 97-100.
11. Shao, J. (1990). Bootstrap estimation of the asymptotic variances of statistical functionals. *Ann. Inst. Statist. Math.*, 42, 737-752.
12. Shorack, G.R. (1982). Bootstrapping robust regression. *Comm. Statist. Theory Methods*, 11, 961-972.
13. Singh, K. (1998). Breakdown theory for bootstrap quantiles. *The Ann. Statist.*, 26, 1719-1732.
14. Stromberg, A.J. (1997). Robust covariance estimates based on resampling. *J. Statist. Plann. Inference*, 57, 321-334.
15. Van Aelst, S. and Willems, G. (2002). Robust bootstrap for S-estimators of multivariate regression. Statistics in Industry and Technology: *Statistical Data Analysis*, 201-212.
16. Vandewalle, B., Beirlant, J. and Hubert, M. (2004). A robust estimator of the tail index based on an exponential regression model, Theory and Applications of Recent Robust Methods, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, *Series: Statistics for Industry and Technology, Birkhauser, Basel*, 367-376.
17. Willems, G. and Aelst, S.V. (2005). Fast and Robust Bootstrap for LTS. *Computational Stat. and data Analysis*, 48, 703-715.
18. Yang, S.-S. (1985). On bootstrapping a class of differentiable statistical functionals with applications to L- and M-estimates. *Statist. Neerlandica*, 39, 375-385.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 475-486

CAUSES AND MEASURES TO CONTROL INFANT MORTALITY IN RURAL AREAS (A CASE STUDY OF UCH SHARIF)

Mariam Abbas Soharwardi[1] and **Umar Farooq**
Department of Economics, The Islamia University of Bahawalpur, Pakistan
Email: ma_eco@hotmail.com

ABSTRACT

Infant mortality in Uch sharif is increasing day by day. Purpose of this study is to illustrate the factors determining infant mortality in Uch sharif and measures to control infant mortality in rural areas. Primary data has been used for this study by using interview technique and 120 mothers were interviewed. We used binary logistic model in our study. The result showed the significance results of birth gap, doctor availability and mother's employment with infant mortality. These results are explained by the odd ratio which showed that we can control by the birth gap, provision of doctors and promote the concept of carry out jobs after marriages.

INTRODUCTION

Health is an important an important aspect of life .In general terms better health status of individuals reflects reduced illness, low level of morbidity and less burden of disease in a given pollution .It is a widely recognized that improved health not only lowers mortality, morbidity and level of fertility, but also contributes to increased productivity and regular school attendance of children as a result of fewer work days lost due to illness , which in turn have implications for economic and social well- being of the population at large. Hence investing in Health is vital for promoting human resource development and economic growth in a country. [World Bank (1993)]

A view of Pakistan s health profile indicates that the sector has expanded considerably in terms of Physical infrastructure and its man power in both the public and private sector. This has contributed to some in selected Health status indicator over the years. However the public health care delivery system has been inadequate in meeting the needs of the fast growing population and in filtering down its benefits to the gross-root level. As such, Pakistan still has one of the Highest rates of infants and child mortality, total mortality and maternal mortality when compared with many other countries in Asia region [UN(2000). Due to low priority given to social sector development in the past and the low budgetary allocations made to the health desired level and large gaps remain in the quality of care indicators, especially in rural areas like Uch Sharif, high level of infant and child mortality and fertility in Pakistan point toward the fact that health and illness problems are sever for young children and Mothers (Mahmud 1993). Infant and child death rates in Pakistan are high even in the context of the Asian region and progress in health and survival of children has been much been less then desired level [world Bank(1993)]. Although estimates of infant and child mortality rates as derived from various data sources in Pakistan show great variation, the available evidence indicates that nearly 58% of all deaths occur among children under five years of age. 36% die during infancy and more than half of all infant deaths occur within the four weeks of their birth. Health care facilities in Pakistan are concentrated mostly in urban areas contributing to lower risks of death among children from infections and diarrhea disease. (Mahmud 1994). Health is main indicator of development. In grappling problem of what forces may be useful for promoting human development resource, a considerable degree of attention has focused upon general health sector especially infant health. What is infancy? Infancy is generally the period from birth until age one year. It is a time of a lot of growth and change for children and families. Health for all by the year 2000 has became the slogan of primary health care since the declaration of Alama Ata in 1978 (World Health Organization (WHO) 1978) which recommended that, a man social target of government international organization and the whole world community in the coming decades should be the attainment by all people of the world by the year 2000 of a level of health that will permit them to lead a socially and economically productive life' the task seem as formidable especially when it is considered that of the 122 million infant born each year in the world. More than 12 millions die before reaching their first birthday, more than 10 million of these deaths occurring in the developing world (WHO, 1980).

Infant mortality is defined as the number of death of infants per 100 live birth. The most common cause of infant mortality worldwide has traditionally been dehydration from diarrhea. Because of the success of spreading information about oral dehydrations solutions (a mixture of salts sugar and water) to mother around the world, the rate of children dying from dehydration has been decreasing and has became second most common cause in late 1990s. Currently the most common cause is pneumonia.

Uthman (2008) has investigated the effect of low birth weight on infant mortality. To examine the relationship between high-risk of infant born with low birth weight and infant mortality in Nigeria. Birth weight is a strong indicator not only of a birth mother's health nutritional status but also newborn's chances for survival, growth, long term health and psychosocial development. Chaudhury, et al. (2006) have investigated the district level variations in infant mortality in Sri Lanka. The purpose of this paper was to study the inter-district variation in infant mortality in relation to certain aspects of economic (access to safe drinking water) and nutrition (birth weight) status; access to health (public health and mid-wife population ratio) and use of health care. This paper is an attempt to identify some proximate determinant of IMR. Data was taken from the Registrar's general development, Family health bureau and the medical statistics unit of ministry of health. The data was secondary. Multiple regression model used for analysis. They conclude that the focused attention to neonatal survival will be an important policy imperative. Three of the factors (birth weight, access to safe drinking water registration status of pregnant women) examined here directly impact neonatal mortality, which accounts groups for almost three fourths of deaths during infancy in Sri Lanka. Souza, et al. have analyzed the determinants of child mortality in slums of Karachi, Pakistan. This study was undertaken to identify risk factors for under-five child mortality. This paper attempts to uncover the role of behavior issues like restricted maternal autonomy and patterns of health seeking behavior alongside the more conventional, socio demographic predictors for under-five child mortality. This study was initiated in January 1993 in Six Slums of Karachi where the community health sciences (CHS) department of the Aga khan University (AKU) has operated primary health care (PHC) programs since 1985. For the theoretical frame work Mosley and Chen's frame work was used. Logistic regression model was used for analysis. And the data was secondary. They concluded that the primary determinant of a programs success will be largely is effectiveness in introducing relevant social and behavioral change in these village-like settings, rather than the biological effectiveness of the technologies themselves exclusively.

MAJOR OBJECTIVE OF THE STUDY

The objective of the study is examining the levels and determinant of Infant mortality in uch sharif. The other purpose is to see the effect of disease, infections that cause infant death.

Data and Methodology

Primary data is collected with the help of questionnaire through field survey by interviewing selected sample. In this regard random sample survey was conducted and 120 mother were interviewed. The sample design was prepared aiming at to produce accurate data within the permitted time and expenditure. Furthermore variables of Infant health Health were also computed from census data from Sub Tehsil Uch sharief. These variables include Health status. Major disease at the time of birth major disease after birth, birth gap, breastfeeding, vaccination, mother employment status, mother qualification, gender of child mother major disease, hospital distance, Doctor availability, visit of vaccination team, and LHW visit in home

Construction of Model

The Economic model provides a frame work to identify the investigated relationship and use of the resulting information. The aim of this is to explain the Impacts of Health status of Infants on Child Mortality .So "the next step is to specify the statistical model that consisting with the sampling process by which underlying data is generated. A model is formulated to suggest a conceptual framework for Infants mortality. In our analysis, we will use binary logistic model. The regress and can take only two values, say 1 if the Infant is died, and 0 if the Infant is not died.

Function

We estimate non linear maximum likelihood function for the binary logistic model. We start with the general function's

$$Yi=f(X1, X2, X3, X4\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.)$$

where Yi denote infant mortality, Y is equal to 1, if infant is died, and Y is equal to 0 if infant is not died.

X1, X2, X3 and X4 are various factors that affect the children health positively or negatively.

Equation

To find out the determinants of rural children health, we have formed one equation. This is as under

$$INM= \beta_0+ \beta_1\ BG+ \beta_2\ DCA+ \beta_3\ ME+ U_i$$

where

INF= Infant Mortality
BG= Birth Gap
DCA= Doctor Availability
ME= Mother Employment

β_0, β_1 , $\beta 2$ and β_3 are regression coefficient to be estimated by using binary logistic model. U= Random error term independently and identically distributed with zero mean and constant variance. The direction and strength of INM and Explanatory variables are determined from sign coefficient and significance of t ratios.

Binary Logistic model.

Logit analysis is in many ways the natural complement of ordinary linear regression where the regress was and is not a continuous variable but a state may or may not hold or a category in a given classification. When such discrete variables occur among the independent variables or repressors or a regression equation, they are deal with the introduction of one or several dummy variables. But when the dependent variables belong to this type, the regression model brakes down. Legit analysis and logistic analysis provides a ready alternative. At first sight it is quite different from the familiar linear regression models, and slightly fright men by its apparent complexity, they logit model belongs to the class of probability models. Major disease at the time of birth. Major disease at the time of birth. Major disease at the time of birth mine discrete probabilities over a limited number of possible outcomes.

RESULTS AND DISCUSSION

Qualitative Analysis

In qualitative analysis, averages and percentages has been calculated. Descriptive analysis of child Health with the independent variables is as follows

Table 1:
How Many of Infant Are Died

Infants	Frequency	Percentage.
Nil	33	27.5
One Child	53	44.2
Two Child	34	28.3
Total	120	100.0

Explanation

According to survey of 120 houses. Of which, child mortality is nil in 27.5 % , only 1 child in 44.2% and two child in 28.3%.

Table 2:
Distribution of respondent by Health Status of Infants

Health Status	Frequency	Percentage
Weightless	16	13.3
Healthy	53	44.2
Normal	51	42.5
Total	120	100.0

Explanation

According to the results 13.3% children are Weightless 44.2 % children’s are Healthy, and 42.5% are normal Health.

Table 3
Distribution of Gender of Child

Gender	Frequency	Percentage
Female	51	42.5
Male	69	57.5
Total	120	100.0

Explanation

The Table shows that 57.5% male infants, and 42.5% female infants. Male infants are preferred to female infants in this survey because male infants are less healthy than that of female infants.

Table 4
Distribution of Respondents Disease after Birth.

Disease	Frequency	Percentage
Fever	17	14.2
Diarrhea	25	20.8
Respiratory	10	8.3
Pneumonia	20	16.7
Malnutrition	17	14.2
Malaria	15	12.5
Hemoglobin	8	6.7
Hepatitis	8	6.7
Total	120	100.0

Source: Survey

Explanation

Nine major diseases are found affecting the infants during the survey. According to our results 14.2% infants are suffering in fever, 20.8% are suffering in Diarrhea, 8.3% infants are suffering in Respiratory disease, 16.7% infants are suffering in pneuomia, 14.2% infants are suffering in malnutrition and 12.5% infants are suffering in malaria, 6.7 % infants are victim the lack of hemoglobin, 6.7% infants are suffering in hepatitis .The most horrible disease to infants found during the survey is Diarrhea.

Table 5
Baby vaccinated completely

Vaccinated	Frequency.	Percentage.
No	55	45.8
Yes	65	54.2
Total	120	100%

Source: Survey

Explanation

Vaccination is necessary to every infant as it prevents from many diseases. According to our result 45.8% infants are not vaccinated, and 54.2% is vaccinated completely.

Table 6
Infant Feeding

Infant Feeding	Frequency	Percentage
Breast feeding	59	49.2
Powder milk	40	33.3
Cow milk	21	17.5
Total	120	100.0

Source: Survey

Explanation

According to our results 49.2% infants are breast fed, 33. % infants are fed powdered milk and 17.5% infants are fed cow milk. In Pakistan, most infants are breast fed.

Table 7
Supplement Food

Supplement Food	Frequency	percentage
No	76	63.3
Yes	44	36.7
Total	120	100.0

Source: Survey

Explanation

According to our result 63.3% are not fed by the supplement food where as 36.7% infants are fed by supplement food.

Table 8
Mother Age at Delivery

Age	Frequency	Percentage
16	72	60
19	34	28.4
21	10	8.3
26	4	3.3
Total	120	100.0

Source: Survey

Explanation

According to results 60% of Women give birth to their first child at age of 16 sixteen, 28.4 % mothers are age of nineteen at the time of their first delivery. Percentage of women of age of 21 at delivery time is 8.3 % while 3.3 % of women are of the age of 26 years at the time of their delivery .Most of the girls are married at the age of fifteen according to survey.

Table 9
Mother Qualification

Qualification	Frequency	Percentage
Uneducated	50	41.5
Primary	3	2.5
Middle	32	26.8
Metric	16	13.3
Up to Metric	19	15.9
Total.	120	100.0

Source: Survey

Explanation

According to the result 41.5 % mothers are uneducated .Whereas 2.5% mothers are educated at primary level and 26.8 % are educated at elementary level. 13.3% mothers are educated at secondary level. 15.9% are educated above the metric.

Table 10
Mother Vaccination during Pregnancy.

Vaccinated	Frequency	Percentage
No	55	45.8
Yes	65	54.2
Total	120	100.0

Source: Survey

Explanation

Vaccination is not only necessary for infants but as well as for mothers. 45.8 % of mothers are not vaccinated according to survey. 54.2% mothers are vaccinated.

Table 11
Mother Major Disease

Disease	Frequency	Percentage
T.B	12	10
Hepatitis	21	17.5
Asthma	161	13.3
Heart Problem	14	11.7
Gastric Problem	15	12.5
Nil	42	35
Total	120	100.0

Explanation

According to survey five diseases were found in mothers. 10 % of mothers are suffering from 17.5% mothers are patient of Hepatitis. Asthma is found in 13.3% of mothers. 11.7% mothers are suffering from heart problems and Gastric problem is common in 12.5% whereas 35% of mothers are considered well.

Table 12
Source of Income

Source	Frequency	Percentage
Father	101	84.2
Mother	19	15.8
Total	120	100.0

Source: Survey

Explanation

Source of income is necessary for food, health or treatment of infants and mothers. 84.2% families depend upon father. Whereas 15.8% depend on mothers.

Table 13
Head of House Hold

Head	Frequency	Percentage
Father	107	89.2
Mother	13	10.8
Total	120	100.0

Source: Survey

Explanation

According to results 89.2% families have father as head of the house hold but 10.8% have mothers as the head of family.

Table 14
Medical center Distance

Distance	Frequency	Percentage
1-5	86	71.7
2-8	27	22.5
8-10	7	5.8
Total	120	100.0

Source: Survey

Explanation

According to survey some of the infants and mothers are not getting the cure on time due to the distance of nearest medical center situated. 71.7 % of medical centers are situated at 1 to 5 km away, 22.5 % of which are situated at 2 to 8 km away. Centers at 8 to 10 km are of 5.8%.

Table 15
Public or Private Medical Center

Medical center	Frequency	Percentage
Public	81	67.5
Private	17	14.2
Both	22	18.3
Total	120	100.0

Source: Survey

Explanation

According to result 67.5 % of families are near to public medical centers 14.2% are near to private medical centers while 18.3% have both public and private medical centers near to them.

Table 16
visit vaccination team.

Visit	Frequency	Percentage
No	55	45.8
Yes	65	54.2
Total	120	100.0

Source: Survey

Explanation

According to our survey 45.8% houses are not visited by vaccination teams. While 54.2% houses are visited by the vaccinated teams as being near to the cities.

Table 17
LHW visits

Visit	Frequency	Percentage
No	50	41.7
Yes	70	58.3
Total	120	100.0

Source: Survey

Explanation

According to our survey 41.7 % houses are not visited by vaccination teams. While 58.3% houses are visited by the vaccinated teams as being near to the cities.

QUANTITATIVE ANALYSIS

Infant Mortality is used as dependent variable in the Logistic Regression model and the r results of model show that child health is effective by three independent variables

Birth gap
Doctor availability
Mother employment

There are many variable used in the Logestic Regression model. But stepwise regression showed these three variables are closely related to the dependent variable.

Binary Logistic Model

Dependent Variable	Independent Variables	Coefficient	P-value	Odd Ratio
Infant Mortality	Bg	-.388	.185	.678
	Dca	-.576	.081	1.780
	Memp	-.964	.030	.381
	Constant	1.562	.071	4.769

Source: Survey
P.V = Level of significance
O.R = Odd Ratio

Explanation

The results showed that the infants where birth gap is low their health is .678 times affected than those who have more birth gap. Birth gap positively associated with infant health. Doctor’s availability is deeply effect on infant mortality. The areas in which doctors are not available the infant health is 1.786 times affecting than those areas where doctors available. Mothers who had employed, who have highly expenditure on health of their children. The Health of their children is improved .381 times than the mothers who had not employed Mother Employment have positive impact on infant health. In rural areas, average expenditure on children is very low because there are not more facilities of health and parents have low level of income. As well as our average expenditure is more the health of the children is better

CONCLUSION

After Analysis the impacts of Health status of infants on child mortality in Sub Tehsail, uch Sharif Following conclusion are made. Ensuring the survival and well being of children is a concern of families, communities and nations throughout world. In quantitative analysis, Birth gap, Doctor Availability and Mother Employment are very closely related with children health. The Nine major Diseases Fever Diarrhea, Respiratory, Pneumonia, malnutrition malaria, Hemoglobin, hepatitis Which are badly effected on the Infant Health. In depth study of rural area has shown that the main Reason for non acceptance of immunization is the obstruction created by Husband and Mothers in law because of the children’s crying at the time of immunization. In rural areas, a strong issue is a preference for sons which is reflected in discrimination against girls in decision about health care, Schooling and feeding. In the Rural areas, economic activity and poverty in Pakistan have adversely affected the child health particularly their nutritional status. Poor income, education of mother, household income ,occupation of father, standard of living ,birth interval are variable which affecting of children Health. child survival programmer are inexpensive, basic interventions that save the lives of children under five from the leading causes of child death and promote healthy and productive families and communities.

REFERENCES

1. Aga, S. (2000). The determinants of infant mortality in Pakistan. *Social Science and medicine*, 51, 199-208.
2. Arshad, Mahmood (2002). Determinants of Neonatal and Post neonatal Mortality in Pakistan. *The Pakistan Development Review*, 41(4), 723-744.
3. Akhter, N. and Zafar, I. (2005). Factor affecting child health. *Jou. Agri. Soci. Sci,* 50-53.
4. Brandt, W. (1980). A program for Survival Report of the independent commission on international development issue. *The journal of Human Resources*, 81(5), 52-120.
5. Castro, F. (1983). The world Economics and social crisis. *Child and Mother Development*, 47(4), 95-138.
6. Caldwell, J.C.P.H. Reddy and P. Caldwell (1983). The social Component of Mortality Decline in South India Employing Alternatives Methodologies. *The Population Studies*, 37(2), 185-205.
7. Chaudhury, R.H, Gunasekera, P. (2006). District level variations in infant mortality in Sri Lanka A challenge to achieving the millennium development goal on child survival. *Regional health Forum*, 10(1).
8. Gokhale, R.S.S. and Gorale, V. R (2002). Infant mortality in India; Use of maternal and child health services in relation to literacy status. *J Health Popul Nutr,* 20(2), 138-149.
9. Hendrickse, R.G. (1971). The Causes Effects of infant health Status in Third World. *The Journal of Child Health*, 40(3), 156-290.
10. Huq, M.N. and Tasnim, T. (2008). Maternal education and child health care in Bangladesh. *Maternal Child Health J.*, 12, 43-51.
11. Majumder, A.K, May, M. and Pant, P.D. (1997). Infant and child mortality determinants in Bangladesh: Are they changing. *J. Biosoc. Sci*, 29, 385-399.
12. Mahmood, N. (1994). The disease Pattern and utilization of health Care Service in Pakistan. *The Pakistan Development Review*, 41(4), 723-744.
13. Macfarlane, S.B. (1984). Some opportunity for biometry in promoting child health the third World. *The Journal of Human Resources*, 40,525-513.
14. Mahmood, A.M. (2002). Determinants of neonatal and post-neonatal mortality in Pakistan. *The Pakistan Development Review*, 41(4), 723-744.
15. Mahfooz and Surur., (2009). Level and determinants of infant and child mortality in malakal Town. *Southern Sudden,* 4(2).
16. Omotesho. O.A, M.O, Adewumi and K.S (2007). Food security and poverty of rural households In Kwara state, Nigeria. *AAAE Conference Proceedings*, 571-575.
17. Stock well, E.G., Goza, W.F. and Balistreri, K.S. (2005). Infant mortality and socioeconomic status. *Family and Demographic Research*, 24(4), 387-399.
18. Sugathan, K.S, Mishra, V. and Retherfor, R.D. (2001). Promoting institutional deliveries in rural India: The role of Antenatal care services. *National Family Health Survey*, Report No. 20.
19. Souza, R.M.D. and Bryant, J.H. (1999). Determinants of childhood mortality in Slums of Karachi, Pakistan. *Journal of health and population in developing countries*, 2(1), 33-44.

20. Tezcan, A.G. (1992). Infant mortality: a Turkish Puzzle. *Health Transition Review*, 2(2).
21. Uthman, O.A. (2008). Effect of low birth weight on infant mortality: Analysis using weilbull Hazard model. *The internet journal of Epidemiology*, 6(1).
22. World Bank (1993). World Development Report 1993). Health Development Issues. Health for All Series, 40(3), 156-290.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 487-494

SOME RECENT WORKS ON TWO PHASE SAMPLING FOR RATIO AND REGRESSION ESTIMATION

M.S. Ahmed
Department of Mathematics and Statistics, Sultan Qaboos University,
Muscat, Sultanate of Oman. Email: msahmed@squ.edu.om

ABSTRACT

Tripathi & Ahmed (1993, 1995) suggested a class of estimators for two-phase sampling by using multiple auxiliary variables. It noticed that Chand (1975), Kiregyera (1980, 1984) and Mukerjee *et al*. (1987) were the particular cases of that class of estimators. They provided the optimum estimator, which have minimum variance. Further, Ahmed and Triphati (1993) suggested other class of estimators which was better than Srivastava *et al*. (1990). Recently, Roy (2003), Singh *et al*. (2004), Upadhyay and Singh (2001), Singh (2001), Samiuddin and Hanif (2007), Hanif*et al*. (2009), Senapati and Sahoo (2006), Pradhan (2005), and Dash and Mishra (2011) suggested some estimators with same setup of Chand (1995). This paper will present the comparisons of these recent estimators with Ahmed and Tripathi (1993) and Tripathi& Ahmed (1993, 1995). Finally, we will show that most of these recent estimators have no credibility.

1. INTRODUCTION

Two-phase sampling scheme is widely used for estimating population parameters when the auxiliary variables are readily available cheaply related to survey variable. Chand (1975), Kiregyera (1980, 1984), Mukerjee *et al*. (1987) and Srivastava *et al*. (1990) suggested several estimators for finite population mean using two auxiliary variables (the mean of one is known while other is unknown. Tripathi and Ahmed (1993, 1995) extended these results for more than two auxiliary variables for each phase in two-phase sampling. Further, Ahmed *et al*. (1995&96) and Ahmed (2003) extended their result for multiphase sampling. Suppose a finite population composes with N observation units; each of the units can be identified by a level of which the set is denoted $U = \{1,2,\dots,N\}$. A simple random sample without replacement S is a subset of U.

Suppose $\boldsymbol{X}_i = \left(\boldsymbol{X}_i^{(1)'}, \boldsymbol{X}_i^{(2)'}\right)'$ are q-auxiliary variables, where $\boldsymbol{X}_i^{(1)} = (X_{1i}, X_{2i}, \dots, X_{ki})'$ and $\boldsymbol{X}_i^{(2)} = (X_{k+1i}, X_{k+2i}, \dots, X_{qi})'$ and they are available with moderate cost to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i\in U} Y_i$ of the study variable Y.

Suppose that $\boldsymbol{X}_i^{(1)}$ is available for all $i \in U$ and $\boldsymbol{X}_i^{(2)}$ is available for all $i \in S_1$, where S_1 is a sub-sample drawn from U under the sampling design D_1. The study variable Y_i is observed for all $i \in S_2$ a sub-sample drawn from S_1 under the sampling design D_2 comparatively small sample with moderate cost.

Suppose $\overline{\boldsymbol{X}}_{(1)} = \left(\bar{X}_{h(1)}\right)_{q\times 1} \forall h = 1,2,..,q$ is an unbiased estimate of the population mean $\overline{\boldsymbol{X}} = \left(\bar{X}_h = \frac{1}{N}\sum_{i\epsilon U} X_{hi}\right)_{q\times 1}$ $\forall h = 1,2,..,q$ under the sampling design D_1; and $\bar{Y}_{(2)}$is an unbiased estimate of the population mean $\bar{Y}$ under the sampling design D_2 at the 2nd phase (Also, $\overline{\boldsymbol{X}}_{(2)}$ is an unbiased estimate of the $\overline{\boldsymbol{X}}_{(1)}$ under the sampling design D_2 at the 2nd phase).

The layout of auxiliary variables and study variable for generalized multiphase sampling are given as follows:

Source		**Size**	**Auxiliary Variables**	**No. of variables**
Population	U	N	$X_1, X_2, \dots, X_k$	k
1st phase	S_1	n_1	$X_{k+1}, X_{k+2}, \dots, X_q$	$q - k\ (k < q)$
2nd phase	S_2	n_2	$(X_1, X_2, \dots, X_k, X_{k+1}, X_{k+2}, \dots, X_q), Y$	$q + 1$

2. DIFFERENT ESTIMATOR AND THEIR PROPERTIES

Tripathi and Ahmed (1993, 1995) suggested the following estimator

$$\begin{aligned} d_T &= \bar{Y}_{(2)} + \left(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(1)}_{(1)}\right)'\boldsymbol{T} + \left(\overline{\boldsymbol{X}}_{(1)} - \overline{\boldsymbol{X}}_{(2)}\right)'\boldsymbol{\Gamma} \\ &= \bar{Y}_{(2)} + \textstyle\sum_{h=1}^{k} T_h\left(\bar{X}_h - \bar{X}_{h(1)}\right) + \sum_{h=1}^{q} \Gamma_h\left(\bar{X}_{h(1)} - \bar{X}_{h(2)}\right) \end{aligned} \qquad (2.1)$$

where $\boldsymbol{T}$ and $\boldsymbol{\Gamma}$ are suitable chosen constant vectors.

Then d_T is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$V(d_T) = V_0 + \boldsymbol{T}'\boldsymbol{A}_1\boldsymbol{T} + \boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} - \boldsymbol{2T}'\boldsymbol{G}_1 - \boldsymbol{2\Gamma}'\boldsymbol{G} \qquad (2.2)$$

where $\boldsymbol{A}_1 = C_1\left(\overline{\boldsymbol{X}}^{(1)}_{(1)}, \overline{\boldsymbol{X}}^{(1)}_{(1)}\right), \boldsymbol{A} = E_1C_2\left(\overline{\boldsymbol{X}}_{(1)}, \overline{\boldsymbol{X}}_{(1)}\right),$

$\boldsymbol{G_1} = C_1\left(\overline{\boldsymbol{X}}^{(1)}_{(1)}, E_2(\bar{Y}_{(2)})\right), \boldsymbol{G} = E_1C_2\left(\overline{\boldsymbol{X}}_{(2)}, \bar{Y}_{(2)}\right),$

Optimum values of constants $\boldsymbol{T} = (T_h); T_{h0} = B_{yh.1,2,\dots,h-1,h+1,..k}$ and $\boldsymbol{\Gamma} = (\Gamma_h); \Gamma_{h0} =$ $\mathcal{B}yh.1,2,\dots,h-1,h+1,..q$ are partial regression coefficients $\mathcal{B}yh.1,2,\dots,h-1,h+1,..k$means regression coefficient of the linear plane $y\ on\ x_1, x_2, \dots, x_k$ after eliminating the linear effect of $x_1, x_2, \dots, x_{h-1}, x_{h+1}, \dots, x_k$)

Under SRSWOR, the minimum variance of d_T is

$$V(d_T)_{min} = \frac{N\sigma_y^2}{N-1}\left[\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(1 - \rho^2_{y.1,2,\dots,k}\right) + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\left(1 - \rho^2_{y.1,2,\dots,k,\dots.q}\right)\right] \quad (2.3)$$

where $\rho_{y.1,2,\dots,k}$ and $\rho_{y.1,2,\dots,k,\dots.q}$ multiple correlation coefficients.

Ahmed *et al*. (1994) suggested the following class of estimators

$$d_A = \bar{Y}_{(2)} \prod_{h=1}^{k}\left(\frac{\bar{X}_h}{\bar{X}_{h(1)}}\right)^{\alpha_h} \prod_{j=1}^{q}\left(\frac{\bar{X}_{j(1)}}{\bar{X}_{j(2)}}\right)^{\gamma_j}$$

The optimum values of α_h and γ_j are given below

$$\alpha_{h0} = \frac{B_{yh.1,2,\dots,h-1,h+1,..k}}{R_h} \text{ and } \gamma_{j0} = \frac{B_{yj.1,2,\dots,j-1,j+1,..,q}}{R_j}; \text{ where } R_j = \frac{\bar{X}_j}{\bar{Y}}$$

$$\hat{\alpha}_{h0} = \frac{b_{yh.1,2,\dots,h-1,h+1,..k}}{r_h} \text{ and } \hat{\gamma}_{j0} = \frac{b_{yj.1,2,\dots,j-1,j+1,..,q}}{r_j}; \text{ where } r_j = \frac{\bar{X}_{j(2)}}{\overline{Y_{(2)}}}$$
$$V(g_a)_{min} = V(d_T)_{min}$$

3. EACH PHASE ONE AUXILIARY VARIABLE CASES AND COMPARISONS

Most of another used Z and X be two auxiliary variables available for estimating the population mean $\bar{Y}$ of the study variable Y, where the population mean of Z $(X_1 = Z)$ is known while the population mean of X $(X_2 = X)$ is not known but available for large sample.

(It means that $\bar{Y}_{(2)} = \bar{y}, \bar{X}_1 = \bar{Z}, \bar{X}_{1(1)} = \bar{z}', \bar{X}_{1(2)} = \bar{z}, \bar{X}_{2(1)} = \bar{x}', \bar{X}_{2(2)} = \bar{x}$)

If $k = 1$ and $q - k = 1$, say, $X_1 = Z$and $X_2 = X$, then Tripathi and Ahmed (1995) showed the estimator as

$$d = \bar{y} + t_1(\bar{Z} - \bar{z}') + \tau_1(\bar{z}' - \bar{z}) + \tau_2(\bar{x}' - \bar{x}) \qquad (3.1)$$
$$\mathrm{E}(t_1) \approx T_1,\ \mathrm{E}(\tau_1) \approx \mathit{\Gamma}_1 \text{and } \mathrm{E}(\tau_2) \approx \mathit{\Gamma}_2$$
$$d_0 = \bar{y} + T_1(\bar{Z} - \bar{z}') + \mathit{\Gamma}_1(\bar{z}' - \bar{z}) + \mathit{\Gamma}_2(\bar{x}' - \bar{x}) \qquad (3.2)$$
$$V(d_0) = \frac{N}{N-1}[\left(\frac{1}{n_2} - \frac{1}{N}\right)\sigma_y^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)\left(T_1^2\sigma_1^2 - 2T\sigma_{y1}\right) + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\left(\mathit{\Gamma}_1{}^2\sigma_1^2 + \mathit{\Gamma}_2{}^2\sigma_2^2 + \right.$$
$$2\mathit{\Gamma}1\mathit{\Gamma}2\sigma 12{-}2\mathit{\Gamma}1\sigma y1{-}2\mathit{\Gamma}2\sigma y2] \qquad (3.3)$$

Optimum values of $T_{10} = B_{yz}, \Gamma_{10} = B_{yz.x}$ and$\Gamma_{20} = B_{yx.z}$ and the minimum variance is

$$V(d_T)_{min} {=} \frac{N\sigma_y^2}{N-1}\left[\left(\frac{1}{n_1} - \frac{1}{N}\right)\left(1 - \rho_{yz}^2\right) + \left(\frac{1}{n_2} - \frac{1}{n_1}\right)\left(1 - \rho_{y.zx}^2\right)\right] \qquad (3.4)$$

Best choice of the estimate of $T_{10} = B_{yz}, \Gamma_{10} = B_{yz.x}$ and$\Gamma_{20} = B_{yx.z}$ of are respectively,

$$t_{10} = b_{yz}, \tau_{10} = b_{yz.x} \text{ and} \tau_{20} = b_{yx.z}$$

Most of the available works has been done for two auxiliary variables, where the population mean of one is known while that of other is not known. Thus we will make a comparison with these available estimators. Let X_1 and X_2 be two auxiliary variables available for estimating the population mean $\bar{Y}$ of the study variable Y, where the population mean of $X_1$is known while the population mean of X_2 is not known but available for large sample. In this situation, number estimators are available. Chand (1975) considered the chain based ratio and product estimators as

$$d_1 = \left(\frac{\bar{Y}_{(2)}}{\bar{X}_{2(2)}}\right)\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{2(2)}}\right)\bar{X}_1 \text{and} d_2 = \left(\frac{\bar{Y}_{(2)}}{\bar{X}_{2(1)}}\right)\left(\frac{\bar{X}_{2(2)}}{\bar{X}_1}\right)\bar{X}_{1(1)} \text{ respectively.}$$

Kiregyera (1980,1984) considered chain based ratio-to-regression, ratio-in-regression and regression-in-regression estimators as

$$d_3 = \left(\frac{\bar{Y}_{(2)}}{\bar{X}_{2(2)}}\right)\bar{X}_{2(2)} + b_{21}\left(\bar{X}_1 - \bar{X}_{1(1)}\right),\ d_4 = \bar{Y}_{(2)} + b_{y2}\left[\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{1(1)}}\right)\bar{X}_1 - \bar{X}_{2(2)}\right]$$

and $d_5 = \bar{Y}_{(2)} + b_{y2}\left[\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{21}(\bar{X}_1 - \bar{X}_{1(1)})\right]$ respectively.

Mukerjee *et al*. (1987) revised the estimators of Kiregyera (1984) and suggested two improved estimators as

$$d_6 = \bar{Y}_{(2)} + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{y1}(\bar{X}_1 - \bar{X}_{1(2)})$$

and

$$d_7 = \bar{Y}_{(2)} + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{y2}b_{21}\left(\bar{X}_1 - \bar{X}_{1(1)}\right) + b_{y1}\left(\bar{X}_1 - \bar{X}_{1(2)}\right)$$

Srivastava *et al*. (1990) gave a general expression for chain based ratio and product estimators and one may obtain the following two estimators from their estimator as

$$d_8 = \left(\frac{\bar{Y}_{(2)}}{\bar{X}_{2(1)}}\right)\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{1(1)}}\right)\bar{X}_1 \text{ and } d_9 = \left(\frac{\bar{Y}_{(2)}}{\bar{X}_{1(1)}}\right)\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{2(2)}}\right)\bar{X}_1$$

Besides these, we may consider the following estimators

$$d_{10} = \bar{Y}_{(2)}\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{2(2)}}\right) + b_{y1}\left(\bar{X}_1 - \bar{X}_{1(1)}\right)$$

$$d_{11} = \bar{Y}_{(2)}\left(\frac{\bar{X}_1}{\bar{X}_{1(1)}}\right) + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right)$$

$$d_{12} = \left[\bar{Y}_{(2)} + b_{y1}\left(\bar{X}_1 - \bar{X}_{1(1)}\right)\right]\left(\frac{\bar{X}_{2(1)}}{\bar{X}_{2(2)}}\right)$$

$$d_{13} = \left[\bar{Y}_{(2)} + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right)\right]\left(\frac{\bar{X}_1}{\bar{X}_{1(1)}}\right)$$

$$d_{14} = \bar{Y}_{(2)} + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{y1}\left(\bar{X}_{1(1)} - \bar{X}_{1(2)}\right)$$

$$d_{15} = \bar{Y}_{(2)} + b_{y2.1}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{y1.2}\left(\bar{X}_{1(1)} - \bar{X}_{1(2)}\right)$$

$$d_{16} = \bar{Y}_{(2)} + b_{y2}\left(\bar{X}_{2(1)} - \bar{X}_{2(2)}\right) + b_{y1}\left(\bar{X}_{1(1)} - \bar{X}_{1(2)}\right)$$

where $\bar{X}_{j(h)}$ is the sample mean of j-th auxiliary variable from the sample S_h (j,h=1,2) and b_{y1}, b_{y2} and b_{21} are the sample regression co-efficient of Y on X_1, Y on X_2 and of X_2 on X_1 respectively, obtained from the second phase sample.

Table 1

d	t_1	τ_1	τ_2	$E(t_1) \approx T_1$	$E(\tau_1) \approx \Gamma_1$	$E(\tau_2) \approx \Gamma_2$
d_1	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{1(1)}}$	0	$\frac{\bar{Y}_{(2)}\bar{X}_1}{\bar{\bar{X}}_{1(1)}\bar{\bar{X}}_{2(2)}}$	R_1	0	R
d_2	$-\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_1}=-\frac{\bar{y}}{\bar{\bar{z}}}$	0	$\frac{\bar{Y}_{(2)}\bar{X}_{1(1)}}{\bar{\bar{X}}_1\bar{\bar{X}}_{2(1)}}$	$-R_1$	0	$-R_2$
d_3	b_{21}	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(2)}}$	B_{21}	0	R
d_4	$\frac{\bar{X}_{2(1)}}{\bar{\bar{X}}_{1(1)}} b_{y2}$	0	b_{y2}	$\frac{R_1}{R_2} B_{y2}$	0	B
d_5	$b_{y2}b_{21}$	0	b_{y2}	$B_{y2}B_{21}$	0	B
d_6	b_{y1}	b_{y1}	b_{y2}	B_{y1}	B_{y1}	B
d_7	$b_{y1}+b_{y2}b_{21}$	b_{y1}	b_{y2}	$B_{y1} + B_{y2}B_{21}$	B_{y1}	B
d_8	$\frac{\bar{Y}_{(2)}\bar{X}_{2(2)}}{\bar{\bar{X}}_{2(1)}\bar{\bar{X}}_{1(1)}}$	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(1)}}$	R_1	0	$-R_2$
d_8	$\frac{\bar{Y}_{(2)}\bar{X}_{2(2)}}{\bar{\bar{X}}_{2(1)}\bar{\bar{X}}_{1(1)}}$	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(1)}}$	R_1	0	$-R_2$
d_9	$-\frac{\bar{Y}_{(2)}\bar{X}_{2(1)}}{\bar{\bar{X}}_{2(2)}\bar{\bar{X}}_1}$	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(2)}}$	$-R_1$	0	R_2
d_{10}	b_{y1}	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(2)}}$	B_{y1}	0	R
d_{11}	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{1(1)}}$	0	b_{y2}	R_1	0	B
d_{12}	$-\frac{\bar{X}_{2(1)}}{\bar{\bar{X}}_{2(2)}} b_{y1}$	0	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{2(2)}}$	B_{y1}	0	R
d_{13}	$\frac{\bar{Y}_{(2)}}{\bar{\bar{X}}_{1(1)}}$	0	$\frac{\bar{X}_1}{\bar{\bar{X}}_{1(1)}} b_{y2}$	R_1	0	B_{y2}
d_{14}	0	b_{y1}	b_{y2}	0	B_{y1}	B
d_{15}	0	$b_{y1.2}$	$b_{y2.1}$	0	$B_{y1.2}$	B
d_{15}	0	$b_{y1.2}$	$b_{y2.1}$	0	$B_{y1.2}$	B
d_{16}	$b_{y1.2}$	0	$b_{y2.1}$	$B_{y1.2}$	0	B
d_0	b_{y1}	$b_{y1.2}$	$b_{y2.1}$	B_{y1}	$B_{y1.2}$	B

The estimator d_1 to d_{16} are generated from the class of estimators (d) for different choices of the statistics t, τ_1 and τ_2 and their respective approximate variance will be obtained for respective values of T_1, Γ_1 and Γ_1 (see, Table 1).It is showed that d_0 attends minimum variance.

Ahmed *et al*. (1994) suggest the following general estimators

$$\tilde{d}_A = \bar{y}\left(\frac{\bar{Z}}{\bar{z}\prime}\right)^{\alpha_1}\left(\frac{\bar{z}\prime}{\bar{z}}\right)^{\gamma_1}\left(\frac{\bar{x}\prime}{\bar{x}}\right)^{\gamma_2} \quad (3.5)$$

The optimum choices of α_1, γ_1 and γ_2

$$\alpha_{10} = \frac{B_{yz}}{r_z}, \gamma_1 = \frac{b_{yz.x}}{r_z} \text{ and} \hat{\gamma}_2 = \frac{b_{yx.z}}{r_x}; r_z = \bar{y}/\bar{z}\text{and} r_x = \bar{y}/\bar{x}$$

$$\hat{\alpha}_{10} = \frac{b_{yz}}{r_z}, \hat{\gamma}_1 = \frac{b_{yz.x}}{r_z} \text{ and} \hat{\gamma}_2 = \frac{b_{yx.z}}{r_x}; r_z = \bar{y}/\bar{z}\text{and} r_x = \bar{y}/\bar{x}$$

$$V(g_a)_{min} = V(d_T)_{min} = \frac{N\sigma_y^2}{N-1}\left[\left(\frac{1}{n_1}-\frac{1}{N}\right)\left(1-\rho_{yz}^2\right)+\left(\frac{1}{n_2}-\frac{1}{n_1}\right)\left(1-\rho_{y.zx}^2\right)\right] \quad (3.6)$$

Samiuddin and Hanif (2007)'s full, partial and no information cases are particular cases Tripathi and Ahmed (1995), and Ahmed *et al*. (1994).

Tripathi and Ahmed (1995) showed that Chand (1975), Kiregyera (1980, 1984), Mukerjee *et al*. (1987) and Sahoo *et al*. (1993, 1994) all were the particular cases of their class of estimators.

Ahmed *et al*. (1994) showed that Chand (1975) and Srivastava *et al*. (1990) were the particular cases of their class of estimators.

It seems that Samiuddin and Hanif (2007) did not aware of Tripathi and Ahmed (1995) and Ahmed *et al*. (1994) papers.

Further, recently proposed Mishra, G., Rout, K. (1997), Senapati and Sahoo (2006), Pradhan (2005), and Dash and Mishra (2011) all these estimator are particular case of Triphati and Ahmed (1995).

REFERENCES

1. Ahmad, Z. (2007). *Generalized multivariate ratio and regression estimators for multi-phase sampling*. Ph.D. thesis, School of Business Administration, National College of Business, Administration & Economics, Lahore, Pakistan.
2. Ahmed, M.S. (1995). Some estimation procedure using multivariate auxiliary information in sample surveys. Unpublished Ph.D. thesis, Department of Statistics & Operations Research, Aligarh Muslim University, Aligarh-202002, India.
3. Ahmed, M. S. and Ali, M.A. (1996).The general class of chain estimators for the product of two means using double sampling. *Journal of Statistical Studies*, 16, 65-68.
4. Ahmed, M.S. (1997). The general class of chain estimators for the ratio of two means using double sampling. *Communication in Statistics-Theory and Methods*, 26(9), 2247-2254.
5. Ahmed, M.S. (1998). A note on regression type estimators using multiple auxiliary information. *Australian & New Zealand Journal Statistics*, 40(3), 373-376.
6. Ahmed, M.S. (2003). General Chain estimators under multi-phase sampling. *J. Applied Statist. Sci*., 12(4), 243-250.
7. Ahmed, M. S. and Atsu S.S. (2009). A General Class of Estimators Under Multi-phase Sampling. *Statistics in Transition*, 10(2), 183-192.
8. Ahmed, M.S., Khan, S.U. and Tripathi, T.P. (1995&96). Model based regression estimators using multiphase sampling. *Aligarh Journal of Statistics*, 15&16, 69-74.

9. Ahmed, M.S., Khan, S.U. and Tripathi, T.P. (1994). Two general class of chain ratio and product estimators for a finite population mean based on two-phase sampling and multivariate information. *Journal of Statistical Studies*, 14, 86-99.
10. Ahmed, M.S. and Atsu S.S. Dorvlo (2006). A class of model based estimators under multiphase sampling. Presented on *International Conference on Mathematical Modelling and Computation, University of Brunei Darussalam*, Brunei, June 5-8, 2006.
11. Chand, L. (1975). Some ratio-type estimators based on two or more auxiliary variables. Ph.D. thesis submitted to Iowa State University, Ames, Iowa.
12. Hanif, M., Ahmed, Z. and Ahmad, M. (2009). Generalized multivariate ratio estimators using multi-auxiliary variables for multi-phase sampling. *Pak. J. Statist*. 25(4), 615-629.
13. Kiregyera, B. (1980). A chain ratio-type estimator in finite population two phase sampling using two-auxiliary variables. *Metrika*, 27, 217-223.
14. Kiregyera, B. (1984). Regression type estimators using two-auxiliary variables and the model of double sampling from finite populations. *Metrika*, 31, 215-226.
15. Mishra, G. and Rout, K. (1997). A regression estimator in two phase sampling in presence of two auxiliary variables. *Metron*, 12,177-186.
16. Mukerjee, R., Rao, T.J. and Vijayan, K. (1987). Regression-type estimators using multiple auxiliary information. *Aust. Jour. Stat*., 29(3), 244-254.
17. P.R. Dash and G. Mishra (2011). An Improved Class of Estimators in Two-Phase Sampling Using Two Auxiliary Variables. *Communications in Statistics - Theory and Methods*, 40(24), 4347-4352.
18. Samiuddin, M. and Hanif, M. (2007). Estimation of population mean in single and two phase sampling with or without additional information. *Pak. J. Statist*., 23(2), 99-118.
19. Singh,V.K., Singh, Hari P., Singh, Housila P., and Shukla, D. (1994). A general class of chain estimators for ratio and product of two means of a finite population. *Comm.Stat.-Theo. Math*., 23(5), 1341-1355.
20. Srivastava, S. Rani., Khare, B.B., and Srivastava, S.R. (1990).A generalized chain ratio estimator for mean of a finite population. *Jour. Ind. Sco.Ag. Stat*., 42, 108-117.
21. Tripathi, T.P. and Ahmed, M.S. (1993). A class of estimators for a finite population mean based on multivariate information and general two-phase sampling". Tech. Report No. 17/93 (August 23, 1993), Stat-Math. Unit, Indian Statistical Institute, Calcutta, India.
22. Tripathi, T.P. and Ahmed, M. S. (1995). A class of estimators for a finite population mean based on multivariate information and two-phase sampling. *Cal. Stat. Asso. Bull*., 45, 179-180, 203-218.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 495-506

THE GENDER DISPARITY IN EDUCATION (A CASE STUDY OF REGIONAL PUNJAB)

Shahzad Mushtaq and **Mariam Abbas Soharwardi**
Department of Economics, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
Email: mianb19@yahoo.com; ma_eco@hotmail.com

ABSTRACT

This study examined Gender disparity in net primary school enrolment among the districts of Punjab (Pakistan). The secondary data of 34 districts of Punjab is used for this study. The study found that there exist wide disparities in primary school enrolment among the districts of Punjab. There is a high disparity in primary school enrolment in districts of lower Punjab as compared to the districts of upper and central Punjab. The disparity in net primary enrolment among the districts of Punjab have a negative relationship with number of schools, number of teachers, male adult literacy rate and female adult literacy rate and positive relationship with per capita income and poverty status.

INTRODUCTION

Education is very important to enhance human capabilities and to achieve the desired objectives of socio-economic development. Education enables individuals to broaden their visions and opportunities and to have a voice in public decision making. At the macro level, education means human capital and sustainable economic development due to productive and skilled labour force. At the micro level, education is strongly correlated to higher income generating opportunities and a more informed and aware existence.

In Pakistan, education has suffered from many issues including underinvestment, failure to implement five-year plans, and lack of purpose and policy direction. Since independence, Pakistan has increased the number of primary schools eighteen fold and multiplied enrolment sixteen times. But these gains have been defeated by rising population and lack of quality education [HDR (1998)].

Achieving economic growth is an important goal of any economy. However in recent years, it has been realized that economic growth is a necessary but not a sufficient condition for human development. Pakistan provides a good example of a country which has historically enjoyed a respectable GDP growth rate and yet failed to translate this positive development into satisfactory level of human development. Since its independence in 1947, Pakistan's development policies have focused primarily on realizing high economic growth and only incidentally on the task of providing social necessities. Such a process has given rise to a structure of production and distribution which has been only indirectly responsive to social goals.

WORLD LEVEL ORGANIZATIONS

Education for All (EFA)

The Education For All (EFA) movement, started more than a decade ago in 1990, accelerated the process of human resource development in many developing countries. The EFA refers to the global commitment to ensure that all children would complete Primary Education of good quality. A decade after, the Millennium Declaration resolved to ensure, by 2015, that all children would be able to complete a course of primary education.

At the World Conference on Education for All (Jomtien, Thailand 1990), 1500 participants comprising delegates from 155 governments, policy makers and specialists in education and health, social and economic development from around the World, met to discuss major aspects of EFA. The World Declaration on Education For All and the Framework for Action to meet Basic Learning Needs, adopted at Jometien, foresaw the need for an end of decade assessment of progress as a basis for a comprehensive review of policies concerning basic education. A number of meetings, conferences and forums were held in 1990's to assess the achievement and revise the targets, goals and policies in EFA. A brief overview of these meetings/conferences is as follows:

Jomtien Conference 1990

The Jomtien Conference clearly defined the basic learning needs of the child i.e. learning tools (such as literacy, oral expression, numeric, and problem solving) as well as basic learning contents (such as knowledge, skills, values and attitudes). The framework for action to meet basic learning needs identified six main areas of action:

i) Expansion of early childhood care and development activities;
ii) Universal access to and completion of primary education;
iii) Improvement in learning achievements;
iv) Reduction of adult illiteracy;
v) Expansion of basic education and skills training for youth and adults.

Goals and targets agreed upon in the Jomtien conference were:

1. Universal access to and 80% completion of Primary education by the year 2000.
2. Reduction of adult illiteracy rate to one half of its 1990 level by the year 2000, with sufficient emphasis on female literacy.
3. Improvement in lemming achievement so that an agreed percentage of an appropriate age cohort (e.g. 80 percent of 14 years-old) attains or surpasses a defined level of necessary learning achievements.
4. Expansion early childhood care and developmental activities, including family and community interventions, especially for poor, disadvantaged and disabled children.

The World Education Forum in Dakar (2000)

Ten years after Jomtien, delegates of several countries and funding agencies gathered in Dakar and reaffirmed their commitment in providing Education For All (EFA). The World Education Forum, convened by UNESCO, UNDP, and UNFPA brought together 1500 participants from 182 countries, as well as major funding agencies. It ended with the adoption of the Dakar Framework for Action, wherein ministers of education and other government representatives, heads of United Nation agencies, the donor community and representative of NGOs, indeed all participants, committed themselves to achieve the goals and targets in EFA by the year 2015.

Education for All and Human Development

Pakistan, like other developing countries, responded positively to the declaration. Measures like the Education Sector Reforms (ESR) Action Plan for 2001-04 and National Plan of Action (NPA) for education, a long term framework (2001-15) indicate its commitment with EFA goals. However the facts contained in the recent Human Development Reports reveal an alarming situation regarding current human resource status in Pakistan. According to the Human Development Index (HDI) ranking, Pakistan is at the 142th place among 175 countries, lying in the Low Human Development class. According to Education Development Index (EDI) Pakistan rank number is 118 in the world.

Punjab Education Sector Reform Program (PESRP)

The Punjab Education Sector Reform Program (PESRP), which stated in 2003, has three strategic pillars: (1) public finance reforms to ensure increased public spending for education; (2) devolution of public sector management reform; and, (3) improvement in access, quality and governance of education.

At the time of it launching, the Punjab had been witnessed insignificant improvements in the education sector with net primary enrollments rates of only 45 percent. The education reforms focus on increasing enrolments and retention especially for girls and in improving sector governance and monitoring.

The program has been supported by three IDA development sector policy credits, with the third credit, PEDPC III, approved in June 2006. In a period of three years, enrollment increases have been registered for both boys and girls although at a higher rate for girls. Consequently Punjab is seeing a narrowing of the gender gap. Sector governance has improved through robust monitoring, independent validations, and improvements in financial management.

Disparity

The word disparity refers to lack of equality or parity among different groups, regions, individuals (males and females), countries and etc, for some comparable conditions. Disparity is also defined as difference between two or more things.

Educational Disparity

The disparity in education means that the difference in the educational achievements such as enrolment for different individuals (gender disparity), groups, regions etc.

Educational Disparities in Pakistan

There are great disparities in access among the four provinces; plus there are high variations in rural urban education indicators. A large proportion of the literate population is concentrated in the national and provincial capitals. The areas with low literacy rate are also backward in terms of economic development (Husain and Qasim 2005). Punjab being the most populated provinces hosted the largest number of state schools, while Balochistan hosts the smallest number.

However, the status of education across the provinces is not equal. Literacy rate is highest in Sindh at 56 percent and lowest in Balochistan at 37 percent. This inter-provincial difference is most pronounced in literacy rates among females: as opposed to a female literacy rate of 44 percent in Punjab, in Balochistan the rate is only 19 percent.

Further there is great variation in performance across the rural and urban areas within each province and across males and females. The Gross Enrolment Rate (GER) is high as 111 percent in urban areas of Punjab while it is as low as 41 percent in the rural areas of Balochistan.

The access to education is also marked by income difference: the overall literacy rate among the poor is 28 percent, while that for the non-poor is 49 percent; the net enrolment rate is 37 percent for the poor as opposed to 59 percent for the non-poor (World Bank 2002). The enrolments remain the lowest among the poorest quintile and dropouts highest among this group. This pattern persists across rural and urban regions of all provinces (World Bank 2002).

Against these challenges the government has failed to increase education facilities at the national level to meet the needs of all. It has also failed to develop strategies to bridge the gap of disparities on basis of income, region, and urban/rural divide. The annual increase in the number of public primary schools is below the need: during 2005-6, only 1221 primary state schools were established (MoF 2006). Emphasis is also being placed on opening state financed non-formal schools through NGOs. The Ministry of Education claims to have already established 10374 Non Formal Basic Education (NFBE) schools across the country and aims to take the number up to 82000 (GoP & UNESCO 2005). There are no independent assessments of the performance of children in these schools but according to government's claims they have a 75 percent pass rate in the government administrated fifth grade examinations (GoP & UNESCO 2005). A National and Four Provincial Education Foundations, which are government established NGOs, have also been setup to promote community schools. Even if these schools are providing acceptable education, they confront the problem of mainstreaming. There are not enough state middle schools to absorb children completing primary in these schools. The NGOs are also unable to upgrade their own schools to middle or secondary due to lack of availability of qualified teachers in remote areas to teach at middle and secondary-levels.

Gender Disparities in Education in Pakistan

The disparities in access on basis of gender also continue after the adopting the strategies for EFA. The female enrollment rates are lower than males and drop out rates among girls are very high, (World Bank 2002).

These gender disparities are compounded not only due to poor supply of educational facilities but also due to cultural values and norms which makes it difficult to access education for girls: for example, religious and cultural emphasis on 'purdah' makes parents reluctant to send girls to schools at a distance. However, the high turn out of girls in NGO run non-formal schools and the recent World Bank sponsored stipend scheme suggests that the cultural values are not against female education, rather parents require institutional arrangements responding to their cultural requirements: for example, establishing schools close to home to ensure female security, providing female teachers to respect purdah (World Bank 2002; Sarwar 2006).

Intra-Provincial Educational Disparity in Punjab

The analysis based on PSLM (2004-05) data suggest that Punjab Education Sector Reform Program has contributed significantly in improving the gross and net enrolment rates in the province, which was a major and immediate focus of these reforms. The lessons learnt from this program indicate that if the financial constraints can be eased, appropriate physical infrastructure facilities can be provided, and committed quality teachers are employed then progress can be accelerated towards substantially improving the educational access and outcomes in Pakistan.

The Punjab Education Sector Reform Program me (PESRP) no doubt has increased the net primary enrolment as well as gross primary enrolment in Punjab but the disparity in net primary enrollment and gross primary enrollment still persists among the districts of Punjab. These disparities in primary enrolment among districts of Punjab causing severely effect to education in Punjab. The disparities are high in lower districts of Punjab as compared to upper and central districts of Punjab.

LITERATURE REVIEW

Sathar and Lloyd (1994) have discussed the determinants of primary school enrolment and completion among children in Pakistan as well as the level of parental expenditures on children enrolled at primary level and giving particular attention to factors at household and community levels. The secondary data is used based on the 1991 Pakistan Integrated Household Survey (PIHS), which includes a national sample of 4711 households. To analyze the data the Multivariate Analysis technique is used. The dependent variables are primary school enrolment, completion of primary school and ever attended school; and independent variables are education expenditures in form of tuition, uniforms, books, transportation, private tutor, examination fees and others; distance of school, child characteristics in form of child’s age, number of total children in a household and birth order of a child; parent’s characteristics in form of their literacy; household characteristics in form of mother headship, neither parents headship, number of male and female adults, household income, household expenditures, household cultivated land and household business; and community characteristics in form of availability of public and private schools within one kilometer. This study shows that inequalities across households provide a major explanation for variation among children in primary school levels. The basic decision relating to children’s entry into school and completion of the primary level are largely determined by parent’s education, particularly that of mothers and household income. Only a small percentage of school-age children in Pakistan having mothers with any education or parents with sufficient income, the cycle of poverty and unequal opportunity is perpetuated. The accessibility of “appropriate” single-sex schools and the availability of quality schools are important additional factors in children’s schooling outcomes, particularly for girls in the rural areas. The study also shows that larger numbers of children in a household reduces the probability of primary school completion for children in the urban areas and significantly reduce average educational expenditures. This study concludes by recommending a substantially increased government commitment to primary education, with particular emphasis on the needs of girls. Expected gains would include greater gender equality, a substantial improvement in human development and possibly decline in fertility.

Sabir (2002) has discussed that which income group actually benefits from the government's subsidized education services and how are these benefits distributed between males and females in Pakistan. The "Benefit Incidence Analysis" technique is used to asses gender differentials in public service provision. The secondary data is used and estimates are based on Pakistan Integrated Household Survey (PIHS) 1998-99 and Provincial Demand for Grants 1999-2000. The study shows that government subsidies directed towards primary education are pro poor in all four provinces of Pakistan and females has disadvantage in access to primary education. The government subsidies directed towards higher education poorly targeted and poorest income group receives less than the riches income group and indeed favor those who are better off. Similarly, the gender disparity in access to public subsidy is higher at tertiary level and lowest at primary level, which also reflects poor targeting.

Khan (1997) shows that investment in primary education has higher return for the economy than investment in any physical capital be it agriculture, industry or infrastructure. Hence, if it investments were strictly made on economic criteria of rates of return then primary education should have received the highest priority in the development plans of Pakistan.

Arif et al. (1999) have analyzed the effects of poverty and gender on primary school enrolment in Pakistan and they also discussed the determinants of primary school enrolment. The secondary data is used in this study and data source is the Pakistan Socio-economic Survey (PSES) carried by the PIDE (Pakistan Institute of Development Economics). The Logistic Regression and Multivariate analysis technique is used to analyze the data. The study shows that the percentage of enrolled children who belong to poor households is less than that for the children who belong to non-poor households. The primary school enrolment is very low in rural areas which are 49% as compared to 72% enrolment in urban areas. The negative effect of poverty on primary school enrolment is more pronounce in the rural areas and for girls. The study shows that poverty, gender and place of residence have significant effects on primary school enrolment. The poverty exerts a significant negative influence on a child's probability to enroll in a primary school and this effect cannot be entirely explained by the household income. The study shows that poverty affects male and female enrolment rates alike, but this is not the case with the income. The parents' decision to enroll boys in school is not significantly influenced by household income; girl's chances of attending school depend on the availability of additional financial resource.

Hazarika (2001) has analyzed gender differences in the sensitivity of primary school enrollment to the costs of post-primary schooling in rural Pakistan. The study shows that all measures of the costs of schooling, only distance from primary school is found to be a statistically significant determinant of female primary school enrollment.

Sawada and Locksbin (2001) have analyzed the household schooling decisions in rural Pakistan. This study is based on the field survey to investigate household decisions about schooling in rural Pakistan. This study shows that hiring more female teachers and providing more primary schools for girls closer to villages will improve the chances of rural Pakistani girls entering school and staying enrolled.

Although the Punjab Education Sector Reform Program causes to increase the net and primary enrolment in the Punjab but the intra-provincial disparity of net and gross primary enrolment still persists. The Punjab province can be divided into three main regions upper, central and lower based on their socioeconomic characteristics. The people of upper Punjab are mainly working in service sector. The people of central Punjab are mainly related to industries. The people of lower Punjab are mainly related to agriculture.

Conflict theory suggests that property ownership (economic structure) determines social and political structure. Thus, inequality in the economic system permeates the social and political, and by extension, educational systems.

According to conflict theory, inequality issues in Punjab education may simply be a symptom of the social, economic and political disparities in which education is found deeply rooted. The Upper and Central Punjab are more modernized, urbanized and advanced industrially and technologically than the Lower Punjab. Conflict theory therefore would point to the education system as a reflection of these inequalities, and that educational development, either progressing or lagging behind, should be explained by the larger social, economic and political context.

Critical social theory shares with conflict theory the point of view that inequality is inherent in education. Nonetheless, whereas conflict theory attributes educational inequality to the larger social, economic and political structures, critical social theory seeks the source of inequality in the education system itself; according to critical social theory, in any system there is co-existence of two opposite groups of objects, entities or people, namely the oppressors and the oppressed.

DATA METHODOLOGY

In this study, the secondary data is used for 34 districts of Punjab as follows;

1. Net Primary School Enrolment for age 5-9 years
2. Adult literacy Rate for Males (15 years & older)
3. Adult literacy Rate for Females (15 years & older)

Source: Pakistan Social and Living Standards Measurement Survey (PSLM) 2004-05

4. No. of Schools at Primary level
5. No. of Teachers at Primary level

Source: Academy of Educational Planning & Management (AEPAM) 2003-04

6. Per Capita Income Per Month in Pakistani rupees
7. Poverty Status (Percentage of people living below Rs.1000 per month)

Source: Planning and Development Department, Govt. of Punjab.

Econometric Analysis

I have used some variables such as number of primary schools, number of teachers in primary schools, average per capita income per month, poverty status (percentage of population living below Rs.1000 per month), male adult literacy rate (age 15 years and older) and female adult literacy rate (age 15 years and older) as determinants of disparity in Net Primary Enrolment in the districts of Punjab. For econometric analysis, I used the OLS (Ordinary Least Squares) method by the help of computer software SPSS (Statistical Package for Social Science) to find out the results.

Hypothesis

The disparity in Net Primary School Enrolment has negative relationship with number of schools, number of teachers, male adult literacy rate and female adult literacy rate; and positive relationship with the per capita income and poverty status.

Model

The model used for the econometric analysis is given in the following,

$$\text{DNPE} = f\ (\text{NOS, NOT, PCI, MALR, FALR})$$

Table 1:
The definitions of variables and their expected sign

Variables	**Description**	**Expected sign**
Dependent Variable		
DNPE	Disparity in Net Primary School Enrolment	
Independent Variables		
NOS	Number of Schools	Negative
NOT	Number of Teachers	Negative
PCI	Per Capita Income	Positive
PS	Poverty Status	Positive
MALR	Male Adult Literacy	Negative
FALR	Female Adult Literacy	Negative

The Model is given below

$$\text{DNPE}=\beta_0+\beta_1\text{NOS}+\beta_2\text{NOT}+\beta_3\text{PCI}+\beta_4\text{PS}+\beta_5\text{MALR}+\beta_6\text{FALR}+\mu$$

RESULTS AND DISCUSSION

In the following table results of regression are given,

Table 2:
Econometric results of the model.

Variables	Coefficients	Significant level
Constant	-20.934 (-1.16)	.256
NOS	-.008 (-1.147)	.261
NOT	-.00255 (.987)	.333
PCI	.02172 (3.705)	.001
PS	.736 (4.222)	.0
MALR	-.455 (-2.241)	.022
FALR	-.253 (-1.553)	.132
R2	.792	
F	17.110	

t values are given in parenthesis.

Discussion

The discussions of the econometric results are given as follows:

(a) Number of Schools

Theoretically disparity in net primary school enrolment and number of school are negatively related. As number of school increases then disparity in net primary school enrollment decreases. In Table 2 the coefficient of NOS is negative, which confirm theoretical expectation. The magnitude of coefficient of NOS is -.008 which indicates that one unit increase in NOS would leads to -.008 unit decrease in DNPE. The variable NOS is not significant. Khan and Ali (2005) shows that if certain facilities and institutions such as schools are not locally available and there are social taboos or difficulties about girls' use of non-local facilities, or if there are affirmative action policies in place for girl's participation in certain levels of education, household's behavior towards girls may be negatively biased not due to parental discrimination per se but rather due to these supply side conditions.

(b) Number of Teachers

Theoretically disparity in net primary school enrolment and number of teachers are negatively related. As number of teachers increases then disparity in net primary school enrollment decreases. In Table 2 the coefficient of NOT is negative, which confirm theoretical expectation. The magnitude of coefficient of NOT is -.00255 which indicates that one unit increase in NOT would leads to -.00255 unit decreases in DNPE. The variable NOT is highly insignificant. But literature shows that the biggest challenges for the government in improving quality is to ensure provision of required number of teachers within state schools and to improve pre and in-service teachers training. There are insufficient numbers of teachers in state schools: during 2004-2005, the average teacher school ratio in primary schools in Punjab and NWFP was 3 and 2 (Shami et al. 2006). Given that many primary schools in the city areas have more than five teachers as they are popular postings and are used as political bribes, the average of 2 to 3 teachers per school means that some schools in the rural areas end up being multi-grade one-teacher schools (Aly 2007). This means that in some government schools one teacher ends up teaching children from first to five grades and unlike the teachers in the non-formal schools does not even get any specialized training in multi-grade teaching.

(c) Per Capita Income

Regression result posits that there exists a positive relationship between disparity in net primary school enrolment and per capita income, which approved theoretical expectations. It shows that increase in per capita income alternatively the growth is like that it supports the disparity, higher growth will result into disparity. There is a need to redirect the growth so that disparity may be declined. The magnitude of coefficient PCI is .021 indicate the one unit increase in PCI leads to .021 unit increase in DNPE. The variable PCI is highly significant. Deon (2000) suggested that gaps in educational enrollment and attainment across different wealth groups are large in almost all developing countries.

(d) Poverty Status

Variable Poverty status depicts that there exist a positive relation between disparity in net primary school enrolment and poverty status. It shows that if PS of individuals increases then the disparity in net primary school enrolment also increases. The poverty status of people also means that they have poor conditions of living, less access to health facilities and other opportunities of daily life. The poverty status of parents no doubt affects the children’s health and their nutrition level which decreases their enrolment in the primary schools. From Table 2 the value of coefficient of PS is .736 and highly significant. It shows that one unit change in PS would leads to .736 unit change in DNPE. Alderman, Behrman, Lavy and Menon (1997) have showed that poor health and nutrition conditions negatively affect the children primary schooling in rural Pakistan. Maluccio, Hoddinott, Behrman, Martorell, Quisumbing and Stein (2006) have showed that childhood nutrition have significantly effected the educational achievement for both males and females among Guatemalan adults.

(e) Male Adult Literacy Rate

The coefficient of MALR shows that there is a negative relationship between disparity in net primary enrolment and male adult literacy rate, which is theoretically consistent. It shows that if the male adult literacy rate increases in the districts of Punjab then the disparity in net primary school enrolment decreases. The magnitude of coefficient of MALR is -.455 and is highly significant. Generally, in Punjab households are headed by the males, hence if they are educated then there should be an increase in the primary school enrolment. Khan et al. (2005) shows that parents education (separately of fathers and mothers) has positive impact (as a continuous variable---number of years of education) on the sons and daughter’s schooling but the impact on son’s schooling is stronger than daughters.

(f) Female Adult Literacy Rate

The coefficient FALR posits that there exists a negative relationship between disparity in net primary school enrolment and female adult literacy rate. It shows that if the female adult literacy rate increases in the districts of Punjab then disparity in net primary school enrolment decreases. In table 2 the coefficient of FALR is -.253 and significant at 10 percent level. This value shows that one unit change in FALR would leads to change .253 units in DNPE.

CONCLUSION

The primary education has a very important role in the life of individuals because it has higher returns. In Punjab province the enrolment in the primary schools has improved during the last few years but there is still present a wide disparity in net primary enrolment among the upper, central and lower districts of Punjab. Our analysis shows that there exists a lower net primary enrolment in the districts of lower Punjab causing a high disparity in net primary enrolment as compared to the upper and central districts of Punjab. Our econometric analysis shows that disparity in net primary enrolment among the districts of Punjab has a negative relationship with number of schools, number of teachers, male adult literacy rate and female adult literacy rate; and positive relationship with the per capita income and poverty status.

POLICY RECOMMENDATIONS

The policy recommendations are given below.

1. The number of primary schools should be increased in the districts of Punjab where the disparity in net primary enrolment is high.
2. To increase the quality of education measures should be taken.
3. To decrease the gap in the per capita income among the districts of Punjab measures should be taken which will increase the children schooling.
4. Measures should be taken to increase the adult literacy rate in high primary enrolment disparity districts.
5. In Pakistan there exists lot of ghost school or zero efficiency. The govt. should take instant action for this.

REFERENCES

1. Akram and Khan (2007). Public Provisions of Education and Government Spending in Pakistan, *Pakistan Institute of Development Economics*, 40, 402-415.
2. Alderman, Behrman, Lavy and Menon (1997). *Child Nutrition, Child Health, and School Enrollment*. The World Bank, Policy Research Department, Poverty and Human Resources Division.
3. Arif, Saqib and Zahid (1999). Poverty, Gender, and Primary School Enrolment in Pakistan. *The Pakistan Development Review,* 38, 979-992.
4. Asadullah and Niaz (2006). Educational Disparity in East and West Pakistan, 1947-71: Was Pakistan Discriminated Against? *Discussion Papers in Economic and Social History*, University of Oxford.
5. Aslam, Monazza (2005). *Rates of Return to Education by Gender in Pakistan*, University of Oxford.
6. Bano, Masooda (2007). *Education for all by 2015: Will we make it*? Case study of Pakistan, UNESCO.

7. Filmer, Deon (2000). *The Structure of Social Disparities in Education: Gender and Wealth*. Development Research Group, Poverty and Human Resources, The World Bank.
8. Hazarika, Gautam (2001). The Sensitivity of Primary School Enrollment to the Costs of Post Primary Schooling in Rural Pakistan: a Gender Perspective. *The Pakistan Development Review*, 32, 215-225.
9. Husain, Qasim and Sheikh (2003). An Analysis of Public Expenditure on Education in Pakistan. *The Pakistan Development Review*, 42, 771-780.
10. *Impact Analysis of Punjab Education Sector Reforms*. First Quarterly Report for FY06.
11. Ismail and Zafar (1996). Gender Differentials in the Cost of Primary Education: A Study of Pakistan. *The Pakistan Development Review*, 35, 835-849.
12. Jones and Gavin (2000). Human Resources, Poverty, and Regional Development. *The Pakistan Development Review*, 39, 389-413.
13. Khan and Kantar (1997). Education in Pakistan: Fifty Years of Neglect. *The Pakistan Development Review*, 36, 647-667.
14. Khan and K. Ali (2005). Bargaining Over Sons and Daughters Schooling: Probit Analysis of Household Behavior in Pakistan. *The Pakistan Development Review*, 5, 123-142.
15. Khan (2003). Children in Different Activities: Child Schooling and Child Labor. *The Pakistan Development Review*, 42, 137-160.
16. Kotani and Keiko (2004). *Growing Together: Regional Disparities in Primary Education Achievement in Brazil*. School of Education, Stanford University.
17. Maluccio, John, Hoddinott, Behrman, Martorell, Quinsumbing and Stein (2006). The Impact of an Experimental Nutritional Intervention in Childhood on Education among Guatemalan Adults. International Food Policy Research Institute, Food Consumption and Nutrition Division, *Discussion Paper* No. 207.
18. Pal and Ghosh (2007). *Elite Dominance and Under-Investment in Mass Education: Disparity in Social Development of Indian States*, 1960-92.
19. Rodriguez, Pose and Andres (2003). Human Capital and Regional Disparities in the EU, Department of Geography and Environment, London School of Economics.
20. SAARC Regional Education Strategy Update, January 2007.
21. Sabir (2002). Gender and Public Spending on Education in Pakistan: A Case Study of Disaggregated Benefit Incidence. *The Pakistan Development Review*, 41, 477-493.
22. Saqib (2004). Willingness to Pay for Primary Education in Rural Pakistan. *The Pakistan Development Review*, 43, 27-51.
23. Sather and Lloyd (1994). Who Gets Primary Schooling in Pakistan: Inequalities among and within Families. *The Pakistan Development Review*, 33, 103-134.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 507-518

LINEAR MODEL INFERENCE WITH NON-SAMPLE PRIOR INFORMATION

Shahjahan Khan[1], **Budi Pratikno**[2] and **Rossita M. Yunus**[3]

[1] University of Southern Queensland, Toowoomba,
Queensland, Australia. Email: khans@usq.edu.au
[2] General Soedirman University, Purwokerto, Indonesia
[3] University of Malaya, Kuala Lumpur, Malaysia

ABSTRACT

Arguably the most widely used statistical technique is the linear model. Traditionally all classical inferences on the parameters of linear model are based exclusively on the available sample data. Often valuable non-sample prior information on the value of the parameter of interest is available from the expert knowledge or previously conducted studies. Inclusion of such information, in addition to the sample data, is likely to improve the quality of the inference. This paper uses both sample and non-sample information to define estimators of linear model and investigate their statistical properties. It also incorporates the non-sample prior information in defining tests for a subset of parameters when information on the other subset is available. The comparisons of power of the tests are also explored under different conditions.

Keywords: Pretest and shrinkage estimators,bias and mean squared error, pretest test, power and size of test, and non-central bivariate t distribution.

1. Introduction

Classical or frequentist statistics exclusively uses sample information to make inference on population parameters. Incorporation of non-sample prior information with the sample data is likely to improve the quality of inference. However, any such improvement depends on the accuracy of the non-sample prior information. Bayesian approach includes the prior distribution of the parameters of the model along with the sample data to draw inference. The prior distribution is not unique, indeed often subjective, and the posterior distribution depends on the choice of the prior distribution, and that affects the ultimate inference. Nevertheless, non-sample prior information (NSPI) on the value of any parameter from reliable sources can be accurate and lead to correct inference.

As a common practice, classical inferences about population parameters are always drawn from the sample data alone. This applies to methods used in parameter estimation and hypothesis testing. Inferences about population parameters could be improved using non-sample prior information (NSPI) from trusted sources (cf Bancroft, 1944). Such information, which is usually available from previous studies or expert knowledge or experience of the researchers, is un-related to the sample data. It is expected that the inclusion of NSPI in addition to the sample data improves the quality of the estimator and the performance of the test. However, any NSPI on the value of any parameter is likely to be uncertain (or unsure). In this case, the information can be articulated in the form of a null hypothesis. An appropriate statistical test on this null hypothesis is useful to eliminate the uncertainty on the NSPI. Then the outcome of the preliminary test is used in the hypothesis testing or estimation. This approach is likely to improve the quality of the estimator and the performance of the statistical test (see Khan and Saleh 2001; Saleh 2006, p. 1; Yunus 2010; Yunus and Khan, 2011a; and Pratikno 2012).

The NSPI can be classified as (i) unknown (unspecified) if NSPI on the value of the parameter(s) is unavailable, (ii) known (certain or specified) if the exact value of the parameter(s) is available, and (iii) uncertain if the suspected value is unsure (that is, suspected to be a fixed quantity, but not sure). For the three different scenarios, three different estimators, namely the (i) unrestricted estimator (UE), (ii) restricted estimator (RE) and (iii) preliminary test estimator (PTE) are defined in the literature (see,e.g., Judge and Bock, 1978; Saleh, 2006, p. 58). Khan (2003), and Khan and Hoque (2003) provide the UE, RE, and PTE for different linear models. Many authors have contributed to this area to the estimation of parameter(s) in the presence of uncertain NSPI. Bancroft (1944, 1964, 1965) and Han and Bancroft (1968) introduced a preliminary test estimation of parameters. Later, Sclove et al. (1972), Stein (1981), Bhoj and Ahsanullah (1994), Khan (1998, 2003, 2005, 2006a, 2006b, 2008), Khan and Saleh (1995, 1997, 2001, 2005, 2008), Khan et al. (2002a, 2002b, 2005), Khan and Hoque (2003), and Saleh (2006, p. 55) covered various work in the area of improved estimation using NSPI.

For the testing purpose, three different statistical tests, namely the (i) unrestricted test (UT), (ii) restricted test (RT) and (iii) pre-test test (PTT) are defined along the same line as the three different estimators. The UE and UT use the sample data alone but the RE and RT do not use the sample data alone. The PTE and PTT use both the NSPI and the sample data. The PTE is a choice between the UE and RE, whereas the PTT is a choice between the UT and RT. The choice depends on the outcome of the pre-testing on the uncertain NSPI value. Note that by definition the test statistics of the PT and UT are correlated but that of the PT and RT are uncorrelated, indeed independent.

There are a very limited number of studies on the testing of parameters in the presence of uncertain NSPI. Tamura (1965), Saleh and Sen (1978, 1982), Yunus and Khan (2008, 2011a, 2011b), and Yunus (2010) used the NSPI for testing hypothesis using nonparametric methods. Some authors have studied the UE, RE and PTE for parametric cases (for instance Bechhofer (1951), Bozivich et al. (1956), Bancroft (1964), Saleh (2006)), and Hoque et al. (2009) but not the tests. Pratikno (2012) covered the testing after pretest under the parametric framework for a number of linear regression models. The non-parametric approach, namely the M-test method, is used by Yunus (2010) and Yunus and Khan (2011b).

2. The Simple Regression Model

The simple regression model for $[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$ can be represented by

$$\boldsymbol{y} = \theta \boldsymbol{1}_n + \beta \boldsymbol{x} + \boldsymbol{e}, \tag{2.1}$$

where θ and β are the intercept and slope parameters, $\boldsymbol{x}$ is the vector of explanatory variable, $\boldsymbol{y}$ is the vector of the response variable, and the error vector $\boldsymbol{e} \sim N(\mu,\ \sigma^2 I_n)$ in which I_n is the identity matrix of order n and σ^2 is the spread parameter. Let uncertain NSPI on the value of β be available, and the degree of distrust in the NSPI be $0 \leq d \leq 1$.

In addition to the simple regression model above, the following linear models may be considered to estimate parameters and perform tests: estimate/test (1) the intercept vector of the *multivariate simple regression model* (MSRM) when there is NSPI on the slope vector, (2) a subset of regression parameters of the *multiple regression model* (MRM) when NSPI is available on another subset of the regression parameters, and (3) the equality of the intercepts for $p(\geq 2)$ lines of the *parallel regression model* (PRM) when there is NSPI on the slopes. In this paper we do not include these model. It may be noted that to study the properties (power and size) of the ptetest test of any multivariate model the bivariate noncentral chi-square (cf Yunus and Khan, 2011c) and F distributions are essential. For details on testing after pretest under parametric model, see Pratikno (2012), and Khan and Pratikno (2012, 2013).

3. The Estimation Problem

From the sample data alone the unrestricted estimator (UE) of the parameters are

$$\tilde{\beta} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y},\ \tilde{\theta} = \bar{y} - \tilde{\beta}\bar{x},\ \text{and}\ S_n^2 = \frac{1}{n-2}(\boldsymbol{y} - \tilde{\boldsymbol{y}})'(\boldsymbol{y} - \tilde{\boldsymbol{y}})\ \text{where}\ \tilde{\boldsymbol{y}} = \tilde{\theta}\mathbf{1}_n + \tilde{\beta}\boldsymbol{x}. \quad (3.2)$$

Note S_n^2 has a scaled χ^2 distribution with d.f. $\nu = (n-2)$. Consider β_0 to be the value of the slope from a credible source. Then this NSPI can be expressed as a null hypothesis $H_0 : \beta = \beta_0$. When the NSPI is correct, the restricted estimator (RE) of β is $\hat{\beta} = \beta_0$, so the RE of θ becomes $\hat{\theta} = \bar{Y} - \beta_0\bar{X}$. If the NSPI is under suspicion, its uncertainty is removed by testing $H_0 : \beta = \beta_0$ against $H_0 : \beta \neq \beta_0$ using the test statistic $\mathcal{L}_\nu = S_n^{-1} S_{xx}^{\frac{1}{2}}(\tilde{\beta} - \hat{\beta}) \sim t_\nu$ with $\nu = (n-2)$ df. This test statistic is used to define the preliminary test estimator (PTE). Note, in general, $t_\nu^2 = F_{1,\nu}$.

Let $0 \leq d \leq 1$ be the coefficient of distrust on the NSPI. The value of d is 0 if there is no distrust in the NSPI. Now the RE, PTE & shrinkage estimator (SE) of the intercept parameter are defined as follows:

$$\begin{aligned}
\text{Restricted estimator (RE)} : \hat{\theta}^{\text{RE}}(d) &= d\tilde{\theta} + (1-d)\hat{\theta}, \quad 0 \leq d \leq 1, \\
\text{Pretest estimator (PTE)} : \hat{\theta}^{\text{PTE}}(d) &= \hat{\theta}^{\text{RE}}(d)I(F < F_\alpha) + \tilde{\theta}I(F \geq F_\alpha) \\
&= \tilde{\theta} + (\hat{\theta} - \tilde{\theta})(1-d)I(F < F_\alpha), \\
\hat{\theta}^{\text{PTE}}(d=0) &= \tilde{\theta} + (\hat{\theta} - \tilde{\theta}I(F < F_\alpha)\ \text{when}\ d = 0, \\
\text{Shrinkage estimator (SE)} : \hat{\theta}^{\text{SE}}(d) &= \tilde{\theta} + (1-d)(\hat{\theta} - \tilde{\theta})cS_n[\sqrt{S_{xx}}|\tilde{\beta}||]^{-1}, \quad (3.3)
\end{aligned}$$

where c is the shrinkage constant, and $I(\cdot)$ is a binary indicator function.

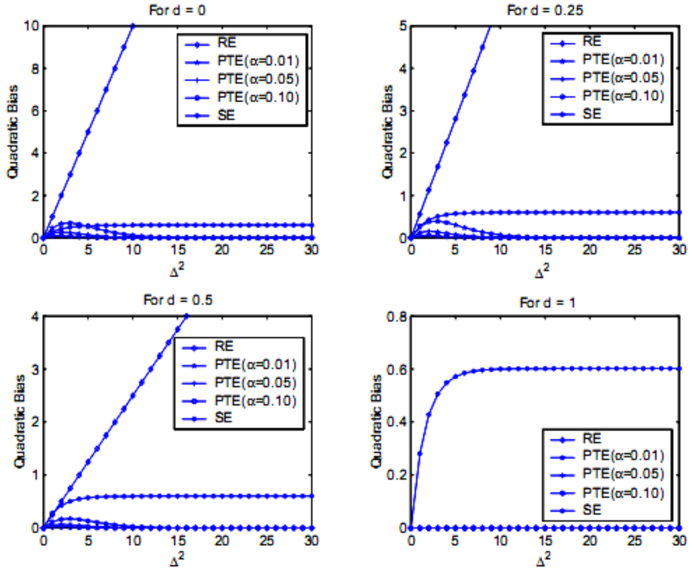


Figure 1: Graph of the quadratic bias of the RE, PTE and SE against Δ^2.

3.1. The Bias of the RE, PTE and SE

The UE is unbiased. The expression for bias of the other estimators are given below.

$$\begin{aligned}
B_2[\hat{\theta}^{\mathrm{RE}}(d)] &= S_{xx}^{-1/2}\bar{x}\sigma(1-d)\Delta, \quad \text{where } \Delta^2 = \sigma^{-2}S_{xx}(\beta-\beta_0)^2. \\
B_3[\hat{\theta}^{\mathrm{PTE}}(d)] &= (1-d)\bar{x}(\beta-\beta_0)G_{3,\nu}\left(3^{-1}F_\alpha;\Delta^2\right), \\
B_4[\hat{\theta}^{\mathrm{SE}}(d)] &= (1-d)S_{xx}^{-1/2}c\bar{x}(\beta-\beta_0)E[S_n]E\left[Z|Z|^{-1}\right],
\end{aligned}$$

where $Z = \sigma^{-1}\sqrt{S_{xx}}(\tilde{\beta}-\beta_0) \sim \mathcal{N}(\Delta,1)$ and $G_{n_1,n_2}(\cdot;\Delta^2)$ is the c.d.f. of a non-central F-distribution with (n_1,n_2) degrees of freedom and non-centrality parameter Δ^2.

The quadratic bias of the RE, PTE and SE are

$$\begin{aligned}
QB_2[\hat{\theta}^{\mathrm{RE}}(d)] &= S_{xx}^{-1}\bar{x}^2\sigma^2(1-d)^2\Delta^2, \\
QB_3[\hat{\theta}^{\mathrm{PTE}}(d)] &= S_{xx}^{-1}\bar{x}^2\sigma^2(1-d)^2\Delta^2\left\{G_{3,\nu}\left(3^{-1}F_\alpha;\Delta^2\right)\right\}^2, \\
QB_4[\hat{\theta}^{\mathrm{SE}}(d)] &= S_{xx}^{-1}\sigma^2\bar{x}^2K_\nu^2\{2\Phi(\Delta)-1\}^2,
\end{aligned}$$

where $K_\nu = \sqrt{\frac{2}{n-2}}\,\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})}$.

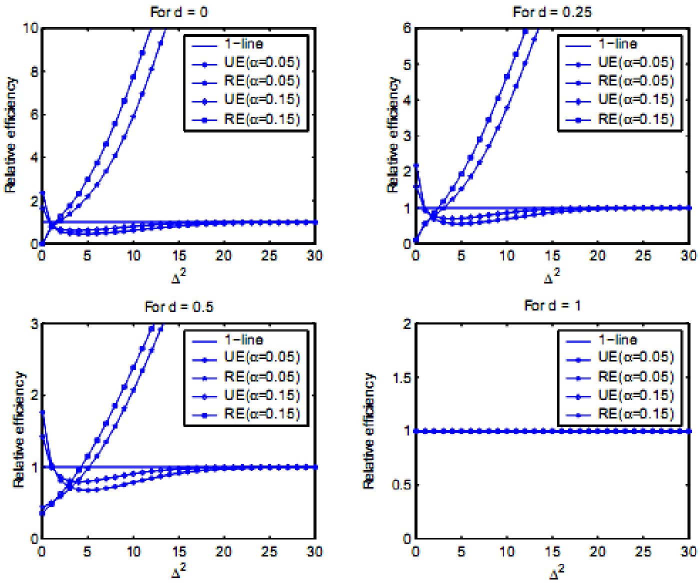


Figure 2: Graph of the relative efficiency of PTE relative to UE and RE against Δ^2.

3.2 The MSE of the RE, PTE and SE

The mean squared error of the estimators are

$$\begin{aligned}
M_2[\hat{\theta}^{\mathrm{RE}}(d)] &= \sigma^2\left[d^2H+(1-d)^2S_{xx}^{-1}\bar{x}^2\Delta^2\right], \\
M_3[\hat{\beta}^{\mathrm{PTE}}(d)] &= \sigma^2H+S_{xx}^{-1}\sigma^2\bar{x}^2\left[\Delta^2\left\{2(1-d)G_{3,v}\left(3^{-1}F_\alpha;\Delta^2\right)\right.\right. \\
&\qquad \left.\left.-(1-d^2)G_{5,v}\left(5^{-1}F_\alpha;\Delta^2\right)\right\}-(1-d^2)G_{3,v}\left(3^{-1}F_\alpha;\Delta^2\right)\right], \\
M_4[\hat{\theta}^{\mathrm{SE}}(d)] &= \sigma^2\left[n^{-1}+S_{xx}^{-1}\bar{x}^2\left\{1+2\pi^{-1}K_\nu^2\left(1-2e^{-\frac{\Delta^2}{2}}\right)\right\}\right].
\end{aligned}$$

4. The Testing Problem

For testing the base load (see Kent 2009) of the energy consumption in a production plant test of intercept is appropriate. The three tests for testing $H_0^*:\theta=0$ are (1) the unrestricted test (UT) if β is unspecified; (2) the restricted test (RT) if β is specified ($\beta=\beta_0$); and (3) the pretest test (PTT), if there is uncertainty on the NSPI, after a preliminary test (PT) on the slope, that is, after testing $H_0^{(1)}:\beta=\beta_0$ to remove any uncertainty.

From now on we only consider tests for unknown σ^2.

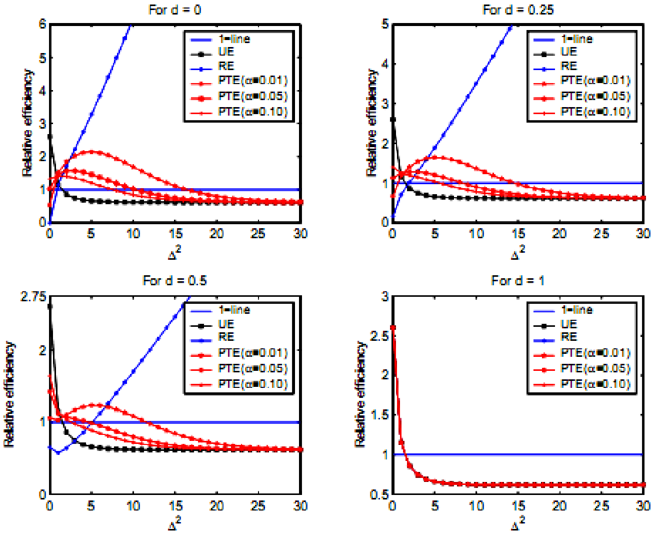


Figure 3: Graph of relative efficiency of SE relative to UE, RE and PTE against Δ^2.

1. $T_1^{(1)} = \sqrt{n}(\tilde{\theta} - \theta_0)[s_*(1 + \frac{n\bar{x}^2}{S_{xx}})^{\frac{1}{2}}]^{-1} = \sqrt{n}(\bar{y} - \tilde{\beta}\bar{x})[s_*(1 + n\bar{x}^2 S_{xx})^{\frac{1}{2}}]^{-1} \sim t_{(n-2)}$.

2. $T_2^{(1)} = \frac{\sqrt{n}(\hat{\theta}-\theta_0)}{s} = \frac{\sqrt{n}\bar{y}}{s} \sim t_{(n-1)}$under H_0^*, where $s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$ and $s_*^2 = \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2}{n-2}$.

3. $T_3^{(1)} = \frac{\tilde{\beta}-\beta_0}{\mathrm{SE}(\tilde{\beta})} = \frac{\sqrt{S_{xx}}\beta\tilde{-}\beta_0}{s_*} \sim t_{n-2}$ under $\mathrm{H}_0^{(1)}$

The associated test functions are defined as (1) $\Omega_1^{(1)} = I\left[T_1^{(1)} > t_{n-2,\alpha_1}\right]$; (2) $\Omega_2^{(1)} = I\left[T_2^{(1)} > t_{n-1,\alpha_2}\right]$; and (3) $\Omega^* = \begin{cases} 1, & \text{if } \{\Psi_1 \text{ or } \Psi_2\} \\ 0, & \text{otherwise,} \end{cases}$
where $\Psi_1 = \left(T_3^{(1)} \leq t_{n-2,\alpha_3}, T_2^{(1)} > t_{n-1,\alpha_2}\right)$ and $\Psi_2 = \left(T_3^{(1)} > t_{n-2,\alpha_3}, T_1^{(1)} > t_{n-2,\alpha_1}\right)$.

Consider the local alternative hypothesis

$$K : (\theta, \beta) = (\delta_1/\sqrt{n}, \delta_2/\sqrt{n}),$$

where $\delta_1 = \sqrt{n}\theta, \delta_2 = \sqrt{n}\beta$ are (fixed) real values. For unknown σ^2 (1) $T_1^{(t)} = T_1^{(1)} - \frac{\delta_1}{s_*\sqrt{1+n\bar{x}^2/S_{xx}}} \sim t_{n-2}$; (2) $T_2^{(t)} = T_2^{(1)} - \frac{\delta_1 + \delta_2\bar{x}}{s} \sim t_{n-1}$; and (3) $T_3^{(t)} = T_3^{(1)} - \frac{\delta_2\sqrt{S_{xx}}}{\sqrt{n}s_*} \sim t_{n-2}$.
Then

$$\boldsymbol{t} = \begin{pmatrix} T_1^{(t)} \\ T_3^{(t)} \end{pmatrix} = \frac{\boldsymbol{Z}}{\sqrt{\frac{(n-1)s_*^2}{\sigma^2}/(n-2)}}$$

is bivariate t with d.f.$=(n-2)$, location vector $(0,0)'$ and scale matrix $\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$. Here,

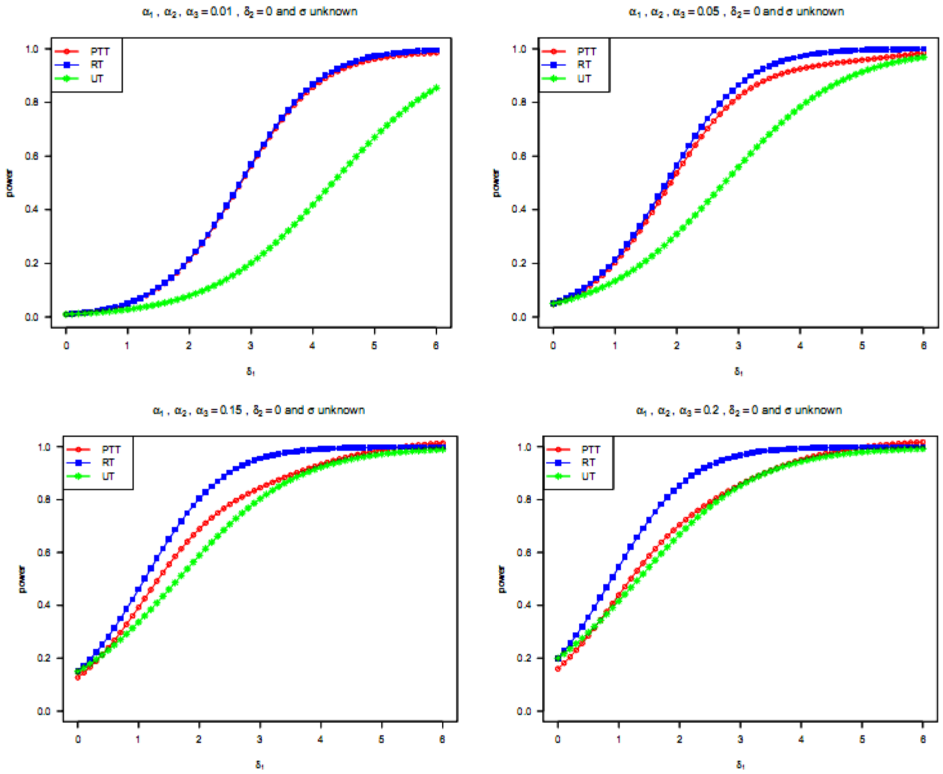


Figure 4: Power function of different tests for different values of δ_1 when $\delta_2 = 0$. $T_2^{(t)}$ and $T_3^{(t)}$ are independent.

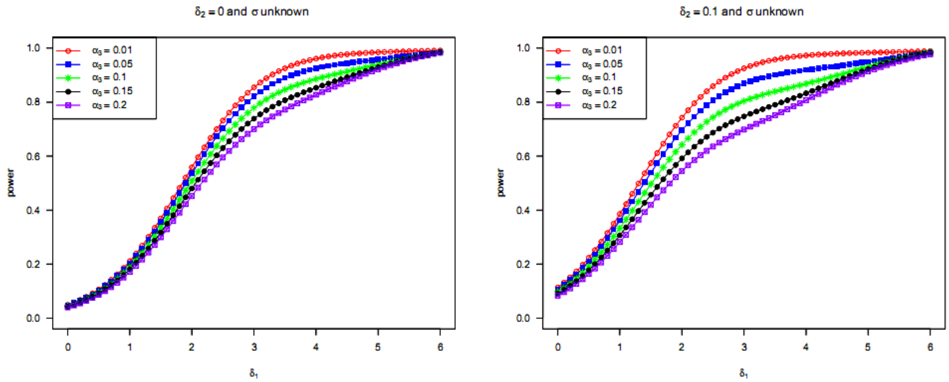
4.1. Properties of the tests - unknown σ^2

The power functions of the three test are given by

- $\Pi_1^{(1)}(\boldsymbol{\delta}) = P\left(T_1^{(t)} > t_{n-2,\alpha_1} - \delta_1[1 + \frac{n\bar{x}^2}{S_{xx}}]^{-1/2}\right)$.
- $\Pi_2^{(1)}(\boldsymbol{\delta}) = P\left(T_2^{(t)} > t_{n-1,\alpha_2} - \delta_1 - \bar{x}\delta_2\right)$, where $\delta_1 = \frac{\lambda_1}{s}(\approx \frac{\lambda_1}{s_*})$ and $\delta_2 = \frac{\lambda_2}{s}(\approx \frac{\lambda_2}{s_*})$.
- $\Pi_3^*(\boldsymbol{\delta}) = P\left(T_3^{(t)} \leq t_{n-2,\alpha_3} - \frac{\delta_2\sqrt{S_{xx}}}{\sqrt{n}}\right) \times P\left(T_2^{(t)} > t_{n-1,\alpha_2} - \delta_1 - \delta_2\bar{x}\right)$ $+ d_{1\rho(a,b,\rho)}\left\{t_{n-2,\alpha_3} - \frac{\delta_2\sqrt{S_{xx}}}{\sqrt{n}}, t_{n-2,\alpha_1} - \delta_1[1 + \frac{n\bar{x}^2}{S_{xx}}]^{-1/2}, -\rho\right\}$

where d_1 is a bivariate Student's t probability integrals defined as

$$d_{1\rho}(a,b,\rho) = \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) n\pi\sqrt{1-\rho^2}} \int_a^{\infty}\int_b^{\infty}\left[1 + \frac{1}{\nu(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right]^{-\frac{\nu+2}{2}} dxdy, \quad (4.4)$$



in which $-1 < \rho < 1$ is the correlation coefficient between the T^{UT} and T^{PT}, $a = t_{n-2,\alpha_3} - \frac{\delta_2\sqrt{S_{xx}}}{\sqrt{n}}$ and $b = t_{n-2,\alpha_1} - \delta_1[1 + \frac{n\bar{x}^2}{S_{xx}}]^{-1}$. For large sample, $\Omega_1 = \Omega_1^{(1)}$, $\Omega_2 = \Omega_2^{(1)}$ and $\Omega^* = \Omega_1^*$. It follows that for any reasonable small n and for some moderate values of δ_2 and α_3, the size of the UT, RT and PTT will be relatively smaller when σ^2 is unknown than when σ^2 is known because $t_{cric} \geq z_{citc}$.

4.2. Illustrations

To compare the power and size of the tests graphically consider the values of the independent variable to be $1, 2, 3, \ldots\ldots, n$. Then $\bar{x} = \frac{n+1}{2}$, and $S_{xx} = \frac{n(n^2-1)}{12}$. Hence

- $\Pi_1^{(1)}(\boldsymbol{\delta}) = P\left(T_1^{(t)} > t_{n-2,\alpha} - \delta_1\sqrt{\frac{n-1}{2n}}\right)$.
- $\Pi_2^{(1)}(\boldsymbol{\delta}) = P\left(T_2^{(t)} > t_{n-1,\alpha} - \delta_1 - (\frac{n+1}{2})\delta_2\right)$.
- $\Pi_3^*(\boldsymbol{\delta}) = P\left(T_3^{(t)} \leq t_{n-2,\alpha_3} - \delta_2\sqrt{\frac{n^2-1}{12}}\right) P\left(T_2^{(t)} > t_{n-1,\alpha} - \delta_1 - \delta_2(\frac{n+1}{2})\right)$
  $+ d_1\left\{t_{n-2,\alpha_3} - \delta_2\sqrt{\frac{n^2-1}{12}}, t_{n-2,\alpha} - \delta_1\sqrt{\frac{n-1}{2n}}, -\frac{\sqrt{3n+3}}{\sqrt{4n+2}}\right\}$.

The power of the tests, for different values of its arguments, are computed from the generated data and plotted in the graphs for comparison.

The size of the PTT is is given in the following table.

4.3. Comparing power and size

For $\bar{x} > 0$: (1) The RT is the best choice for its largest power but the worst choice for its largest size (2) The UT is the best choice for its smallest size but the worst choice for its smallest power (3) The size of the PTT is smaller than that of the RT regardless the value of the slope and the power of the PTT is larger than that of the UT for small and moderate values of the slope. For $\bar{x} = 0$: (1) The size and power of RT, UT and PTT are the same. For $\bar{x} < 0$: (1) The RT is the best choice for its smallest size but the worst choice for its smallest power. (2) The size and power of the UT and PTT are not much different

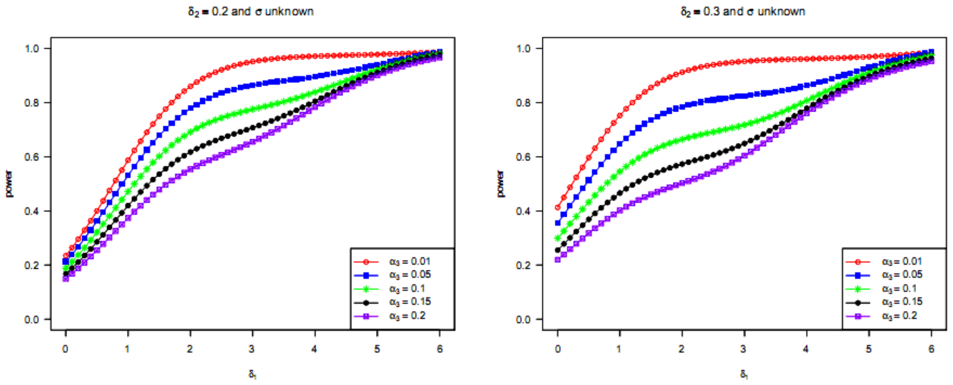
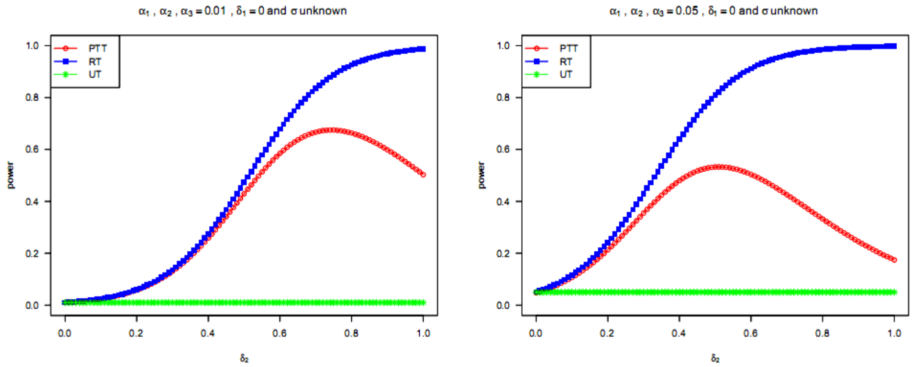


Figure 5: Power function of the PTT for different values of α_3 and varying δ_2.

Table 1: Size of the PTT following PT on slope.

Source of σ	$\delta_2 \backslash \alpha_3$	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90
σ known	0	.04	.04	.04	.03	.03	.02	.02	.01	.01	.01
σ unknown		.04	.04	.04	.03	.03	.02	.02	.01	.01	.01
σ known	.10	.12	.11	.09	.08	.06	.05	.04	.02	.02	.02
σ unknown		.10	.09	.08	.06	.05	.04	.03	.02	.02	.02
σ known	.20	.25	.22	.17	.14	.10	.08	.06	.04	.03	.03
σ unknown		.21	.18	.14	.11	.09	.07	.05	.04	.03	.03
σ known	.30	.39	.33	.24	.18	.13	.09	.07	.05	.04	.04
σ unknown		.35	.29	.21	.16	.12	.09	.06	.05	.04	.04
σ known	.40	.49	.39	.26	.18	.13	.09	.06	.05	.04	.04
σ unknown		.48	.38	.25	.18	.13	.09	.07	.06	.05	.05
σ known	1	.11	.07	.05	.05	.05	.05	.05	.05	.05	.05
σ unknown		.17	.10	.06	.06	.06	.06	.06	.06	.07	.07
σ known	2	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
σ unknown		.07	.07	.07	.07	.07	.07	.07	.07	.07	.07



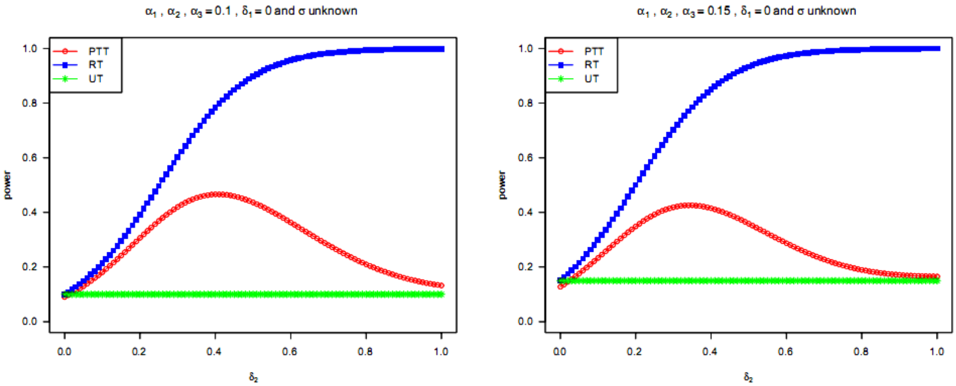


Figure 6: Sizes of UT, RT and PTT for different values of $\alpha_1 = \alpha_2 = \alpha_3$.

for moderate and large values of the slope. The power of the PTT are larger than that of the UT and RT for small values of the slope.

The size and power of the PTT is large when the nominal size of pre-test is very close to 0 especially when the slope (δ_2) is large (& other arguments fixed). This is because it approaches to the size of the RT (which is large when the slope (δ_2) is large).

5. Concluding remarks

Under the unbiasedness criterion, the UE is the best, and RE is the worst if d away from 1. But the PTE is a compromise between the two. The SE is also biased, but it is better than the RT and worse than PTE. For $d = 1$, all estimators are unbiased except the RE.

The UT has the lowest size and lowest power. The RT has the highest power and highest size. The PTT protects against the lowest power of UT and highest size of RT.

REFERENCES

Bancroft, T. A. (1944). On biases in estimation due to the use of the preliminary tests of singnificance. *Annals of Mathematical Statistics*, **15**, 190-204.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of the preliminary test(s) of singnificance. *Biometrics*, **20**(3), 427-442.

Bancroft, T. A. (1965). Inference for incompletely specified models in the physical sciences (with discussion). Bull ISI, Proc. 35th Section, Beograd, **41**, 497-515.

Bechhofer, R. E. (1951). The effect of preliminary test of significance on the size and power of certain tests of univariate linear hypotheses, PhD thesis, Columbia Univ.

Bhoj, B. S. and Ahsanullah, M. (1994). Estimation of a conditional mean in a linear regression model after a preliminary test on regression coefficient. *Biometrical Journal*, **36**(2), 153-163.

Bozivich, H., Bancroft, T. A. and Hartley, H. O. (1956). Power of analysis of variance test procedures for certain incompletely specified models. *Ann. of Math. Statis.*, **27**, 1017-1043.

Han, C.P. and Bancroft, T. A. (1968). On pooling means when variance is unknown. *Journal of American Statistical Association*, **63**, 1333-1342.

Hoque, Z., Khan, S. and Wesolowiski, J. (2009). Performance of preliminary test estimator under linex loss function. *Journal of Communications in Statistics - Theory and Methods*, **38**(2), 252-261.

Judge, G. G. and Bock, M. E. (1978). The Statistical implications of pre-test and stein-rule estimators in econoetrics. North-Holland, New York.

Kent, R. (2009). Energy miser-know your plants energy fingerprint. (accesed 23 May 2011). URL: *http://www.ptonline.com/articles/know-your-plants-energy-fingerprint*.

Khan, S. (1998). On the estimation of the mean vector of Student-t population with uncertain prior information. *Pakistan Journal of Statistics*, **14**, 161-175.

Khan, S. (2003). Estimation of the parameters of two parallel regression lines under uncertain prior information. *Biometrical Journal*, **44**, 73-90.

Khan, S. (2005). Estimation of parameters of the multivariate regression model with uncertain prior information and Student-t errors. *Journal of Statistical Research*, **39**(2), 79-94.

Khan, S. (2006a). Prediction distribution of future regression and residual of squares matrices for multivariate simple regression model with correlated normal responses. *Journal of Applied Probability and Statistics*, **1**, 15-30.

Khan, S. (2006b). Shrinkage estimation of the slope parameters of two parallel lines under uncertain prior information. *Journal of Model Assisted Statistics and Applications*, **1**, 195-207.

Khan, S. (2008). Shrinkage estimators of intercept parameters of two simple regression models with suspected equal slopes. *Communications in Statistics - Theory and Methods*, **37**, 247-260.

Khan, S. and Pratikno, B. (2012). Testing Base Load with Non-Sample Prior Information on Process Load, *Statistical Papers*, In Press.

Khan, S. and Pratikno, B. (2013). Testing Base Load with Non-Sample Prior Information on Process Load (Multivariate Model), Book Chapter in Statistics in Scientific Investigations, edited by Maman Djauhari. In press

Khan, S. and Saleh, A. K. Md. E. (1995). Preliminary test estimators of the mean based on p-samples from multivariate Student-t populations. *Bulletin of the International Statistical Institute*. 50th Session of ISI, Beijing, 599-600.

Khan, S. and Saleh, A. K. Md. E. (1997). Shrinkage pre-test estimator of the intercept parameter for a regression model with multivariate Student-t errors. *Biometrical Journal*, **39**, 1-17.

Khan, S. and Saleh, A. K. Md. E. (2001). On the comparison of the pre-test and shrinkage estimators for the univariate normal mean. *Statistical Papers*, **42**(4), 451-473.

Khan, S., Hoque, Z. and Saleh, A. K. Md. E. (2002a). Improved estimation of the slopeparameter for linear regression model with normal errors and uncertain prior information. *Journal of Statistical Research*, **31**(1), 51-72.

Khan, S., Hoque, Z. and Saleh, A. K. Md. E. (2002b). Estimation of the slope parameter for linear regression model with uncertain prior information. *Journal of Statistical Research*, **36**, 55-73.

Khan, S. and Hoque, Z. (2003). Preliminary test estimators for the multivariate normal mean based on the modified W, LR and LM tests. *Journal of Statistical Research*, **37**, 43-55.

Khan, S., Hoque, Z. and Saleh, A. K. Md. E. (2005). Estimation of intercept parameter for linear regression with uncertain non-sample prior information. *Statistical Papers*, **46**, 379-394.

Khan, S. and Saleh, A. K. Md. E. (2005). Estimation of intercept parameter for linear regression with uncertain non-sample prior information. *Statistical Papers*, **46**, 379-394.

Khan, S. and Saleh, A. K. Md. E. (2008). Estimation of slope for linear regression model with uncertain prior information and Student-t error. *Communications in Statistics-Theory and Methods*, **37**(16), 2564-2581.

Pratikno, B. (2012). Test of Hypotheses for Linear Regression Models with Non-Sample Prior Information, Unpublished PhD thesis, University of Southern Queensland, Australia.

Saleh, A. K. Md. E. (2006). Theory of preliminary test and Stein-type estimation with applications. Wiley, New Jersey.

Saleh, A. K. Md. E. and Sen, P. K. (1978). Nonparametric estimation of location parameter after a preliminary test on regression. *Annals of Statistics*, **6**, 154-168.00.

Saleh, A. K. Md. E. and Sen, P. K. (1982). Nonparametric tests for location after a preliminary test on regression. *Communications in Statistics-Theory and Methods*, **12**(16), 1855-1872.

Sclove, S. L., Morris, C. and Rao, C. R. (1972). Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Ann. Math. Statist.*, **43**, 1481-1490.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**, 1135-1151.

Tamura, R. (1965). Nonparametric inferences with a preliminary test. *Bull. Math. Stat.* **11**, 38-61.

Yunus, R. M. (2010). Increasing power of M-test through pre-testing. Unpublished PhD Thesis, University of Southern Queensland, Australia.

Yunus, R. M. and Khan, S. (2008). Test for intercept after pre-testing on slope - a robust method. *In: 9th Islamic Countries Conference on Statistical Sciences (ICCS-IX): Statistics in the Contemporary World - Theories, Methods and Applications*, 81-90.

Yunus, R. M. and Khan, S. (2011a). Increasing power of the test through pre-test - a robust method. *Communications in Statistics-Theory and Methods*, **40**, 581-597.

Yunus, R. M. and Khan, S. (2011b). M-tests for multivariate regression model. *Journal of Nonparamatric Statistics*, **23**, 201-218.

Yunus, R. M. and Khan, S. (2011c). The bivariate noncentral chi-square distribution - A compound distribution approach. *Applied Mathematics and Computation*, **217**, 6237-6247.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 519-528

SHRINKAGE ESTIMATION OF ELLIPTICAL REGRESSION MODEL UNDER BALANCED LOSS FUNCTION

M. Arashi[1] and Shahjahan Khan[2]

[1] Faculty of Mathematics, Shahrood University of Technology,
Shahrood, Iran. Email:m_arashi_stat@yahoo.com
[2] University of Southern Queensland, Toowoomba,
Queensland, Australia. Email: khans@usq.edu.au

ABSTRACT

For the multiple regression model with error vector following elliptically contoured distribution we propose Bayesian shrinkage estimators under balanced loss function. Comparing a set of competing estimators for the regression vector, it is shown that the shrinkage factor of the Stein estimator is robust with respect to the regression parameters and unknown density generator of elliptical models. The dominance relation of the estimators is also provided.

Key Words: Bayes estimator; Elliptically contoured distribution; Preliminary test estimator; Stein-type shrinkage estimator; Positive-rule shrinkage estimator.
2010 Mathematics Subject Classification: 62J05; 62J07; 62F10.

1. Introduction

Consider the following multiple regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is an n-vector of responses, $\boldsymbol{X}$ is an $n \times p$ non-stochastic design matrix with full rank p, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is p-vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$ is the n-vector of random noises distributed as any member of the elliptically contoured distributions (ECDs), $\mathcal{E}_n(\boldsymbol{0}, \sigma^2\mathbf{V}, g_n)$ for some un-structured known matrix $\mathbf{V} \in S(n)$, where $S(n)$ denotes the set of all positive definite matrices of order $(n \times n)$. The density of $\boldsymbol{\epsilon}$ is given by

$$f(\boldsymbol{\epsilon}) = d_n|\sigma^2\mathbf{V}|^{-\frac{1}{2}} g_n[2\sigma^2]^{-1}\boldsymbol{\epsilon}'\mathbf{V}^{-1}\boldsymbol{\epsilon}, \tag{1.2}$$

where $d_n^{-1} = \pi^{\frac{p}{2}}[\Gamma(\frac{p}{2})]^{-1} \int_{\mathbb{R}^+} y^{\frac{p}{2}-1} g_n(y)\mathrm{d}y$ and for some density generator function $g_n(.)$. The existence of the density generator $g_n(x)$ is dependent on the condition (Fang et al., 1990) $\int_0^\infty x^{\frac{n}{2}-1} g_n(x)\mathrm{d}x < \infty$. If $g_n(.)$ does not depend on n, we use the notation g instead. In this paper, we consider the estimation problem under the following loss function

$$\begin{aligned} L_{\omega,\boldsymbol{\delta}_0}^{\boldsymbol{W}}(\boldsymbol{\delta};\boldsymbol{\beta}) &= \omega r\left(\|\boldsymbol{\beta}\|^2\right)(\boldsymbol{\delta}-\boldsymbol{\delta}_0)'\boldsymbol{W}(\boldsymbol{\delta}-\boldsymbol{\delta}_0) \\ &\quad +(1-\omega)r\left(\|\boldsymbol{\beta}\|^2\right)(\boldsymbol{\delta}-\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{\delta}-\boldsymbol{\beta}), \end{aligned} \tag{1.3}$$

where $\omega \in [0, 1]$, $r(.)$ is a positive weight function, $\boldsymbol{W}$ is a weight matrix, and $\boldsymbol{\delta}_0$ is a target estimator. This loss is pioneered by Jozani et al. (2006) inspiring by Zellner's (1994) balanced loss function. This loss function takes both goodness of fit and error of estimation into account. The $\omega r\left(\|\boldsymbol{\beta}\|^2\right)(\boldsymbol{\delta}-\boldsymbol{\delta}_0)'(\boldsymbol{\delta}-\boldsymbol{\delta}_0)$ part of the loss is analogous to a penalty term for lack of smoothness in nonparametric regression. The weight ω in (1.3) calibrates the relative importance of these two criteria. Dey et al. (1999) also considered issues of admissibility and dominance, under the loss (1.3) ignoring the term $r(.)$ when $\boldsymbol{W} = \boldsymbol{I}_p$. For the case $\omega = 0$, we will simply write $L_0^{\boldsymbol{W}}(\boldsymbol{\delta}; \boldsymbol{\beta})$ as the quadratic loss function. Of course, duty of the weight function $r(.)$ is clearly apparent in the Bayesian viewpoint. In this paper, we take it into consideration for the sake of generality. As it can be seen later, the structure of $r(.)$ does not alter the whole superiority conclusions.

This paper aims at the estimation of the regression parameter vector, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ when it is suspected that $\boldsymbol{\beta}$ may belong to any sub-space defined by $\boldsymbol{H\beta} = \boldsymbol{h}$ where $\boldsymbol{H}$ is a $q \times p$ matrix of constants and $\boldsymbol{h}$ is a q-vector of known constants with focus on the Stein-type shrinkage estimator of $\boldsymbol{\beta}$ in addition to preliminary test estimator (PTE).

Saleh (2006) presents an overview on the topic under normal and nonparametric theory covering many standard models. Other relevant works in the area include Arashi (2012), Khan (2008, 2000), Arashi et al. (2008), Hoque et al. (2009), and Khan and Saleh (1997).

2. Preliminaries for Bayesian estimation

It is easy to show that the unrestricted estimator (UE) of $\boldsymbol{\beta}$ and σ^2 are

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y} = \boldsymbol{C}^{-1}\mathbf{X}'\boldsymbol{V}^{-1}\boldsymbol{y}, \qquad \boldsymbol{C} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}. \tag{2.4}$$

$$\tilde{\sigma}^2 = n^{-1}(\boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \text{ and} \tag{2.5}$$

$$S^2 = (\boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})/(n-p) \tag{2.6}$$

is an unbiased estimator of $\sigma_{\boldsymbol{\epsilon}}^2 = -2\psi'(0)\sigma^2$, where $\psi'(0)$ is the first derivative of characteristic generator of the elliptical model at point zero. Using the invariant theory due to Jeffreys (1961), we define the following prior of ignorance

$$\pi(\boldsymbol{\beta}, \sigma^2) \propto \text{constant}, \quad \pi(\sigma^2) \propto \sigma^{-2}. \tag{2.7}$$

Assume in the multiple regression model (1.1), $\boldsymbol{\epsilon} \sim \mathcal{E}_n(\mathbf{0}, \sigma^2\boldsymbol{V}, \boldsymbol{g})$, where $\boldsymbol{V} \in \mathcal{S}_n$. Then w.r.t. the prior distribution given by (2.7), the posterior distribution of $\boldsymbol{\beta}$ is multivariate Student's t distribution, denoted by $\boldsymbol{\beta}|(\mathbf{X}, \boldsymbol{y}) \sim t_p(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}, m)$, where $\boldsymbol{\Sigma} = S^2\boldsymbol{C}^{-1}$, with *pdf*

$$f(\boldsymbol{\beta}|\mathbf{X}, \boldsymbol{y}) = |\boldsymbol{\Sigma}|^{-\frac{1}{2}}[c(m,p)\pi^{\frac{p}{2}}]^{-1}\left[1 + \frac{1}{m}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right]^{-\frac{n}{2}},$$

where $c(m,p) = m^{\frac{p}{2}}\Gamma\left(\frac{m}{2}\right)[\Gamma\left(\frac{n}{2}\right)]^{-1}$, and $m = n - p$. Thus the Bayes estimator is the posterior mean given by

$$\hat{\boldsymbol{\beta}}_B = \tilde{\boldsymbol{\beta}}. \tag{2.8}$$

Under (1.1), the distribution of Bayes estimator is $\mathcal{E}_p(\boldsymbol{\beta}, \sigma^2\boldsymbol{C}^{-1}, \boldsymbol{g})$. Thus the 1st moment of $\hat{\boldsymbol{\beta}}_B$ is the zero vector and the 2nd central moment is $E(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta})'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma_\epsilon^2 \boldsymbol{tr}(\boldsymbol{C}^{-1})$.

Since the Bayes estimator is nothing more than that of the classical least square estimator of β, one may ask what would be the benefit of putting prior on the model? The answer is that the role of the prior distribution is obvious in the loss function dealing with the function $r(\|\boldsymbol{\beta}\|^2)$.

For the elliptically contoured family distributions the function $r(.)$ is given by

$$r(\|\boldsymbol{\beta}\|^2) = g(\|\boldsymbol{\beta}\|^2). \tag{2.9}$$

Actually by taking this assumption, the loss function relates to the density generator of the base model and therefore the prior information has direct impact on the model understudy. We note that $r(.)$ can be independent of $g(.)$.

Overall, what we need is to compute $E\left[r(\|\boldsymbol{\beta}\|^2)\right]$ which by making use of (2.7), and taking the constant to be 1, is given by

$$E\left[r(\|\boldsymbol{\beta}\|^2)\right] = \int_{\mathbb{R}^p} g(\|\boldsymbol{\beta}\|^2)\mathrm{d}\boldsymbol{\beta} = \frac{\pi^{\frac{p}{2}}}{\Gamma\left(\frac{p}{2}\right)} \int_{\mathbb{R}+} y^{\frac{p}{2}-1} g(y)\mathrm{d}y = d_n^{-1}, \tag{2.10}$$

where d_n is the normalizing constant in (1.2).

3. Shrinkage Estimators

The restricted estimator (RE) is given by

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} - \boldsymbol{C}^{-1}\boldsymbol{H}'\boldsymbol{V}_1(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{h}), \quad \boldsymbol{V}_1 = [\boldsymbol{H}\boldsymbol{C}^{-1}\boldsymbol{H}']^{-1}. \tag{3.11}$$

By making use of (1.1) one can easily see that $\hat{\boldsymbol{\beta}} \sim \mathcal{E}_p\left(\boldsymbol{\beta} - \boldsymbol{\delta}, \sigma^2\boldsymbol{V}_2, \boldsymbol{g}\right)$ for $\boldsymbol{\delta} = \boldsymbol{C}^{-1}\boldsymbol{H}'\boldsymbol{V}_1(\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{h})$ and $\boldsymbol{V}_2 = \boldsymbol{C}^{-1}(\boldsymbol{I}_p - \boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{H}\boldsymbol{C}^{-1})$. Similarly, under $H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$, the following estimator is unbiased for σ_ϵ^2.

$$S^{*2} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/(n - p + q), \tag{3.12}$$

from least squares theory.

Let $w = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}, \sigma > 0, \boldsymbol{V} \in S(n)\}$ and $\Omega = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p, \sigma > 0, \boldsymbol{V} \in S(n)\}$. Then to remove the uncertainly in the suspected value of $\boldsymbol{h}$, we test $H_0 : \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{h}$ (where $q < p$) against $H_a : \boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{h}$, using Corollary 1 from Anderson et al. (1986), which gives the likelihood ratio test statistic

$$\mathcal{L}_n = (\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{h})'\boldsymbol{V}_1(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{h})/(qS^2). \tag{3.13}$$

Under H_0, the pdf of $\mathcal{L}_n$ is given by

$$\boldsymbol{g}^*_{q,m}(\mathcal{L}_n) = \left(\frac{q}{m}\right)^{\frac{q}{2}} \mathcal{L}_n^{\frac{q}{2}-1}\left[B\left(\frac{q}{2}, \frac{m}{2}\right)\left(1 + \frac{q}{m}\mathcal{L}_n\right)^{\frac{1}{2}(q+m)}\right]^{-1} \tag{3.14}$$

which is the central F-distribution with (q, m) degrees of freedom.

In many practical situations, along with the model one may suspect that $\boldsymbol{\beta}$ belongs to the sub-space defined by $\boldsymbol{H\beta} = \boldsymbol{h}$. In such situation one combines the estimate of $\boldsymbol{\beta}$ and the test-statistic to obtain shrinkage estimators as in Saleh (2006). The preliminary test estimator (PTE) of $\boldsymbol{\beta}$ which is a convex combination of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}^{PT} = \tilde{\boldsymbol{\beta}} I(\mathcal{L}_n \geq F_\alpha) + \hat{\boldsymbol{\beta}} I(\mathcal{L}_n < F_\alpha), \tag{3.15}$$

where $I(A)$ is the indicator function of the set A and F_α is the upper α^{th} percentile of the central F-distribution with (q, m) d.f. The PTE depends on α $(0 < \alpha < 1)$, the level of significance and also it yields the extreme results, namely $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ depending on the outcome of the test. Therefore we define Stein-type shrinkage estimator (SE) of $\boldsymbol{\beta}$, as

$$\hat{\boldsymbol{\beta}}^{S} = \hat{\boldsymbol{\beta}} + (1 - d\mathcal{L}_n^{-1})(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{\beta}} - d\mathcal{L}_n^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}), \tag{3.16}$$

where

$$d = (q-2)m/[q(m+2)] \text{ and } q \geq 3. \tag{3.17}$$

The SE has the disadvantage that it has strange behavior for small values of $\mathcal{L}_n$. Also, the shrinkage factor $(1 - d\mathcal{L}_n^{-1})$ becomes negative for $\mathcal{L}_n < d$. Hence we define a better estimator namely the positive-rule shrinkage estimator (PRSE) of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}^{S+} \;=\; \hat{\boldsymbol{\beta}} + (1 - d\mathcal{L}_n^{-1}) I[\mathcal{L}_n > d](\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}). \tag{3.18}$$

4. Properties of the estimators

The bias of the unrestricted (LSE) and restricted estimators are given by

$$\boldsymbol{B}_1(UE) = E[\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}] = \boldsymbol{0}, \text{ and } \boldsymbol{B}_2(RE) = E[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}] = -\boldsymbol{\delta}, \text{ respectively.} \tag{4.19}$$

Following Arashi et al. (2012) the bias of the PTE becomes

$$\begin{aligned} \boldsymbol{B}_3(PT) &= E(\hat{\boldsymbol{\beta}}^{PT} - \boldsymbol{\beta}) = E[\tilde{\boldsymbol{\beta}} - I(\mathcal{L}_n \leq F_\alpha)(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}] \\ &= -\boldsymbol{CH}'\mathbf{V}_1^{1/2}\, E[I(\mathcal{L}_n \leq F_\alpha)\mathbf{V}_1^{1/2}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{h})] = -\boldsymbol{\delta} G_{q+2,m}^{(2)}\left(F_\alpha; \Delta_*^2\right). \end{aligned} \tag{4.20}$$

where

$$G_{q+2i,m}^{(2-h)}(l_\alpha, \Delta_*^2) = \sum_{r=0}^{\infty} K_r^{(h)}(\Delta_*^2) I_{l_\alpha}\left[\frac{q+2i}{2} + r, \frac{m}{2}\right],$$

$l_\alpha = \frac{qF_{q,m}(\alpha)}{m+qF_{q,m}(\alpha)}$, $I_x[a, b] = \int_0^x u^{a-1}(1-u)^{b-1}du$ is the incomplete beta function and

$$K_r^{(h)}(\Delta_*^2) = [-2\psi'(0)]^r \left(\frac{\Delta_*^2}{2}\right)^r \int_0^\infty \frac{(t^{-1})^{-r+h}}{r!}\, e^{\frac{-t\Delta_*^2[-2\psi'(0)]}{2}} W(t)dt.$$

The bias of the SE is

$$\begin{aligned} \boldsymbol{B}_4(S) &= E(\hat{\boldsymbol{\beta}}^{S} - \boldsymbol{\beta}) = E[\tilde{\boldsymbol{\beta}} - d\mathcal{L}_n^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}] \\ &= -d\boldsymbol{C}^{-1}\boldsymbol{H}'\boldsymbol{V}_1^{1/2}E[\mathcal{L}_n^{-1}\boldsymbol{V}_1^{1/2}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{h})] = -dq\boldsymbol{\delta}E^{(2)}[\chi_{q+2}^{*-2}(\Delta_*^2)], \end{aligned} \tag{4.21}$$

where

$$E^{(2-h)}[\chi_{q+s}^{*^{-2}}(\Delta_*^2)] = \sum_{r\geq 0}\frac{1}{r!}K_r^{(h)}(\Delta_*^2)(q+s-2+2r)^{-1},$$

and that of the PRSE is

$$\begin{aligned}\boldsymbol{B}_5(S+) &= E(\hat{\boldsymbol{\beta}}^S-\boldsymbol{\beta}) - E[I(\mathcal{L}_n \leq d)(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})] + dE[\mathcal{L}_n^{-1}I(\mathcal{L}_n \leq d)(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})] \\ &= -dq\boldsymbol{\delta}E_N^{(2)}[\chi_{q+2}^{*^{-4}}(\Delta_*^2)] + \boldsymbol{\delta}G_{q+2,m}^{(2)}\left(d; \Delta_*^2\right) \\ &\quad + \frac{qd}{q+2}\,\boldsymbol{\delta}E^{(2)}\left[F_{q+2,m}^{-1}(\Delta_*^2)I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right], \end{aligned} \tag{4.22}$$

where

$$\begin{aligned}&E^{(2-h)}[F_{q+s,n-p}^{-j}(\Delta_*^2)I(F_{q+s,n-p}(\Delta_*^2) < d_1)] \\ &= \sum_{r=0}^{\infty}K_r^{(h)}(\Delta_*^2)\left(\frac{q+s}{n-p}\right)^j\frac{B(\frac{q+s+2r-2j}{2},\frac{m+2j}{2})}{B(\frac{q+s+2r}{2},\frac{m}{2})}I_{x'}\left[\frac{q+s+2r-2j}{2},\frac{m+2j}{2}\right],\end{aligned}$$

in which $d_1 = \frac{dq}{q+2}$, and $x' = \frac{dq}{m+dq}$. Note that as the non-centrality parameter $\Delta_*^2 \to \infty$, $\boldsymbol{B}_1 = \boldsymbol{B}_3 = \boldsymbol{B}_4 = \boldsymbol{B}_5 = \boldsymbol{0}$ while $\boldsymbol{B}_2$ becomes unbounded. However, under $H_0 : \boldsymbol{H\beta} = \boldsymbol{h}$, because $\boldsymbol{\delta} = \boldsymbol{0}$, $\boldsymbol{B}_1 = \boldsymbol{B}_2 = \boldsymbol{B}_3 = \boldsymbol{B}_4 = \boldsymbol{B}_5 = \boldsymbol{0}$.

The risk function for any estimator $\boldsymbol{\beta}^*$ of $\boldsymbol{\beta}$ under balanced loss function is

$$\boldsymbol{R}_{\omega,\boldsymbol{\delta}_0}^{\boldsymbol{W}}(\boldsymbol{\beta}^*;\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}\left\{E_{\chi}[L_{\omega,\boldsymbol{\delta}_0}^{\boldsymbol{W}}(\boldsymbol{\beta}^*;\boldsymbol{\beta})]|\boldsymbol{\beta}\right\}. \tag{4.23}$$

Using the above definition we find the risk function (4.23) when $\boldsymbol{\delta}_0 = \tilde{\boldsymbol{\beta}}$, as the target estimator, and $\boldsymbol{W} = \boldsymbol{C}$, given by (2.1), evaluate the risks of the five different estimators. For the case $\omega = 0$, we will simply write $\boldsymbol{R}_0^{\boldsymbol{W}}(\boldsymbol{\beta}^*;\boldsymbol{\beta})$.

For the risk of the Bayes estimator, from $\boldsymbol{R}_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(.;\boldsymbol{\beta})$ given in (4.23), we have

$$\begin{aligned}\boldsymbol{R}_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}) &= (1-\omega)E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right)E[(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{C}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})]|\boldsymbol{\beta}\right\} \\ &= p\,\sigma_\epsilon^2(1-\omega)E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right)\right\} = p\,d_n^{-1}\sigma_\epsilon^2(1-\omega).\end{aligned} \tag{4.24}$$

Noting $\boldsymbol{V}_1^{\frac{1}{2}}(\boldsymbol{H}\tilde{\boldsymbol{\beta}}-\boldsymbol{h}) \sim \mathcal{E}_q(\boldsymbol{V}_1^{\frac{1}{2}}(\boldsymbol{H\beta}-\boldsymbol{h}), \sigma^2\boldsymbol{I}_q, \boldsymbol{g})$, the risk of the RE is

$$\begin{aligned}\boldsymbol{R}_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) &= \omega E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right)E[(\boldsymbol{H}\tilde{\boldsymbol{\beta}}-\boldsymbol{h})'\boldsymbol{V}_1(\boldsymbol{H}\tilde{\boldsymbol{\beta}}-\boldsymbol{h})]|\boldsymbol{\beta}\right\} \\ &\quad +(1-\omega)E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right)E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{C}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]|\boldsymbol{\beta}\right\} \\ &= \boldsymbol{R}_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}) - q\,d_n^{-1}\sigma_\epsilon^2 + (1-\omega)d_n^{-1}\theta,\end{aligned} \tag{4.25}$$

where $\theta = \boldsymbol{\delta}'\boldsymbol{C}\boldsymbol{\delta} = (\boldsymbol{H\beta}-\boldsymbol{h})'\boldsymbol{V}_1(\boldsymbol{H\beta}-\boldsymbol{h})$. Note that $\boldsymbol{R} = \boldsymbol{C}_1^{-1/2}\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{H}\boldsymbol{C}_1^{-1/2}$ is a symmetric idempotent matrix of rank $q \leq p$. Thus, there exists an orthogonal matrix $\boldsymbol{Q}$ $(\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_p)$ such that $\boldsymbol{QRQ}' = \begin{bmatrix}\boldsymbol{I}_q & 0 \\ 0 & 0\end{bmatrix}$. Now we define random vector $\boldsymbol{w} =$

$\boldsymbol{Q}\boldsymbol{C}_1^{1/2}\tilde{\boldsymbol{\beta}} - \boldsymbol{Q}\boldsymbol{C}_1^{-1/2}\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{h}$, then $\boldsymbol{w} \sim \mathcal{E}_p(\boldsymbol{\eta}, \sigma^2\boldsymbol{I}_p, \boldsymbol{g})$, where $\boldsymbol{\eta} = \boldsymbol{Q}\boldsymbol{C}_1^{1/2}\boldsymbol{\beta} - \boldsymbol{Q}\boldsymbol{C}_1^{-1/2}\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{h}$. Partitioning the vector $\boldsymbol{w} = (\boldsymbol{w}_1', \boldsymbol{w}_2')'$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2')'$ where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are sub-vectors of order q and $p-q$ respectively, we can represent the test statistic $\mathcal{L}_n$ given by (2.6) as

$$\mathcal{L}_n = \boldsymbol{w}_1'\boldsymbol{w}_1/(qS^2), \text{ and } \theta = \boldsymbol{\eta}_1'\boldsymbol{\eta}_1. \tag{4.26}$$

For the risk of the PTE, note $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} = \boldsymbol{C}^{-1}\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{H}\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{w}$, then simplifications yield

$$\begin{aligned}
\boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{PT}; \boldsymbol{\beta}) &= \omega E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right) E[I(\mathcal{L}_n < F_\alpha)(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\boldsymbol{C}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})]|\boldsymbol{\beta}\right\} \\
&\quad +(1-\omega)E_{\boldsymbol{\beta}}\left\{r\left(\|\boldsymbol{\beta}\|^2\right) E[(\hat{\boldsymbol{\beta}}^{PT} - \boldsymbol{\beta})'\boldsymbol{C}(\hat{\boldsymbol{\beta}}^{PT} - \boldsymbol{\beta})]|\boldsymbol{\beta}\right\} \\
&= \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}) - (1-2\omega)q\sigma_\epsilon^2 d_n^{-1} G^{(1)}_{q+2,m}\left(F_\alpha; \Delta_*^2\right) \\
&\quad +2\theta(1-\omega)d_n^{-1}\left[2G^{(2)}_{q+2,m}\left(F_\alpha; \Delta_*^2\right) - G^{(2)}_{q+4,m}\left(F_\alpha; \Delta_*^2\right)\right].
\end{aligned} \tag{4.27}$$

On simplifications, the risk of the SE becomes

$$\begin{aligned}
\boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{S}; \boldsymbol{\beta}) &= \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}; \boldsymbol{\beta}) + q\, d_n^{-1}\Big\{ \left[d^2\omega - 2d(1-\omega)\right] E^{(1)}[\chi^{*^{-2}}_{q+2}(\Delta_*^2)] \\
&\quad +d^2(1-\omega)E^{(1)}[\chi^{*^{-4}}_{q+2}(\Delta_*^2)]\Big\} + \theta\, d_n^{-1} \\
&\quad \times\Big\{ \left[d^2\omega - 2d(1-\omega)\right] E^{(2)}[\chi^{*^{-2}}_{q+4}(\Delta_*^2)] - 2d(1-\omega) \\
&\quad \times E^{(2)}[\chi^{*^{-2}}_{q+2}(\Delta_*^2)] + d^2(1-\omega)E^{(2)}[\chi^{*^{-4}}_{q+4}(\Delta_*^2)]\Big\},
\end{aligned} \tag{4.28}$$

where

$$E^{(2-h)}[\chi^{*^{-4}}_{q+s}(\Delta_*^2)] = \sum_{r\geq 0}\frac{1}{r!}K_r^{(h)}(\Delta_*^2)(q+s-2+2r)^{-1}(q+s-4+2r)^{-1}.$$

Finally, for the risk of PRSE, after some simplifications, we obtain

$$\begin{aligned}
\boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{S+}; \boldsymbol{\beta}) &= \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{S}; \boldsymbol{\beta}) \\
&\quad -d_n^{-1}\sigma_\epsilon^2\left\{qE^{(1)}\left[\left(1 - \frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right)^2 I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right]\right. \\
&\quad \left. +\frac{\theta}{\sigma_\epsilon^2}E^{(2)}\left[\left(1 - \frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right)^2 I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right]\right\} \\
&\quad -2d_n^{-1}\theta E^{(2)}\left[\left(1 - \frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right) I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right].
\end{aligned} \tag{4.29}$$

5. Performance comparison

This section provides risk analysis of the above estimators with the weight matrix $\boldsymbol{C}$. From equations (4.24) and (4.25) the risk difference of the UE and RE is given by

$$\mathcal{D}_{21} = \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) - \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}) = d_n^{-1}\left[(1-\omega)\theta - q\sigma_\epsilon^2\right]. \tag{5.30}$$

Then it can be directly concluded that $\hat{\boldsymbol{\beta}}$ performs better than $\tilde{\boldsymbol{\beta}}$ that is, $\hat{\boldsymbol{\beta}}$ dominates $\tilde{\boldsymbol{\beta}}$ ($\hat{\boldsymbol{\beta}} \succeq \tilde{\boldsymbol{\beta}}$) provided $0 \le \theta \le \frac{q\sigma_\epsilon^2}{1-\omega}$, for $\omega \neq 1$ since $d_n > 0$.

For comparing the $\hat{\boldsymbol{\beta}}^{PT}$ and $\tilde{\boldsymbol{\beta}}$, the risk difference is

$$\begin{aligned} \mathcal{D}_{13} &= \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}) - \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}) = (1-2\omega)q\, d_n^{-1}\sigma_\epsilon^2 G^{(1)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right) \\ &\quad -2\theta\, d_n^{-1}(1-\omega)[2G^{(2)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right) - G^{(2)}_{q+4,m}\left(F_\alpha;\Delta_*^2\right)]. \end{aligned} \tag{5.31}$$

The right hand side of (5.31) is nonnegative i.e. $\hat{\boldsymbol{\beta}}^{PT} \succeq \tilde{\boldsymbol{\beta}}$ for $\omega \neq 1$ whenever

$$\theta \;\le\; \frac{(1-2\omega)q\sigma_\epsilon^2 G^{(1)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right)}{2(1-\omega)\left[2G^{(2)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right) - G^{(2)}_{q+4,m}\left(F_\alpha;\Delta_*^2\right)\right]}. \tag{5.32}$$

Moreover, under $H_0 : \boldsymbol{H\beta} = \boldsymbol{h}$, because of $\theta = 0$, $\hat{\boldsymbol{\beta}}^{PT} \succeq \tilde{\boldsymbol{\beta}}$ for values ω such that $\omega \le \frac{1}{2}$. Now we compare $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$ by the risk difference

$$\begin{aligned} \mathcal{D}_{23} &= \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) - \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}) \\ &= -q\, d_n^{-1}\sigma_\epsilon^2[1-(1-2\omega)G^{(1)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right)] + \theta\, d_n^{-1}(1-\omega)[1-2G^{(2)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right) \\ &\quad +G^{(2)}_{q+4,m}\left(F_\alpha;\Delta_*^2\right)]. \end{aligned} \tag{5.33}$$

Thus $\hat{\boldsymbol{\beta}}^{PT} \succeq \hat{\boldsymbol{\beta}}$ whenever

$$\theta \;\ge\; \frac{q\sigma_\epsilon^2\left[1-(1-2\omega)G^{(1)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right)\right]}{(1-\omega)\left[1-2G^{(2)}_{q+2,m}\left(F_\alpha;\Delta_*^2\right)+G^{(2)}_{q+4,m}\left(F_\alpha;\Delta_*^2\right)\right]}, \tag{5.34}$$

and vice versa. However, under H_0, the dominance order of $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{PT}$ is as follows

$$\hat{\boldsymbol{\beta}} \succeq \hat{\boldsymbol{\beta}}^{PT} \succeq \tilde{\boldsymbol{\beta}}, \quad \text{or} \quad \hat{\boldsymbol{\beta}}^{PT} \succeq \hat{\boldsymbol{\beta}} \succeq \tilde{\boldsymbol{\beta}}, \tag{5.35}$$

depending on the value α satisfying (5.5).

In order to determine the superiority of $\hat{\boldsymbol{\beta}}^{S}$ to $\tilde{\boldsymbol{\beta}}$, we give the following results.

Theorem 5.1. *Consider the model* (1.1) *where the error-vector belongs to the ECD,* $\mathcal{E}_n(\mathbf{0},\sigma^2\mathbf{V},\boldsymbol{g})$. *Then the Stein-type shrinkage estimator,* $\hat{\boldsymbol{\beta}}^{S}$ *of* $\boldsymbol{\beta}$ *given by*

$$\hat{\boldsymbol{\beta}}^{S} = \tilde{\boldsymbol{\beta}} - d^*\mathcal{L}_n^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})$$

uniformly dominates the Bayes estimator $\tilde{\beta}$ with respect to the balanced loss function $L_0^{\boldsymbol{C}}(\boldsymbol{\delta};\boldsymbol{\beta})$ and is minimax if and only if $0 < d^* \leq \frac{2m}{m+2}$. The largest reduction of the risk is attained when $d^* = \frac{m}{m+2}$.

Following Srivastava and Bilodeau (1989), the risk difference of the SE and Bayes estimator under balanced loss function, is given by

$$\begin{aligned}\mathcal{D}_{41} &= E_{\beta}\left\{E(\hat{\boldsymbol{\beta}}^S - \boldsymbol{\beta})'\boldsymbol{C}(\hat{\boldsymbol{\beta}}^S - \boldsymbol{\beta}) - E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{C}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})|\boldsymbol{\beta}\right\} \\ &= d_n^{-1}\left\{\frac{q^2(m+2)}{m}(d^*)^2 E_{\tau}\left(\frac{\tau^{-2}}{\dot{\boldsymbol{z}}'\boldsymbol{C}^{-1}\dot{\boldsymbol{z}}}\right) - 2q^2 d^* E_{\tau}\left(\frac{\tau^{-2}}{\dot{\boldsymbol{z}}'\boldsymbol{C}^{-1}\dot{\boldsymbol{z}}}\right)\right\},\end{aligned}$$

since $\left(\frac{mS^2}{\sigma^2}\right)\Big|\tau \sim \tau^{-1}\chi_m^2$ and $\tilde{\boldsymbol{\beta}}'\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{H}\tilde{\boldsymbol{\beta}} \mid \tau \sim \tau^{-2}\sigma^4\chi_q^2(\dot{\boldsymbol{\delta}})$, where $\dot{\boldsymbol{\delta}} = \boldsymbol{\beta}'\boldsymbol{H}'\boldsymbol{V}_1\boldsymbol{H}\boldsymbol{\beta}$, where E_N means getting expectation with respect to multivariate normal with covariance $\tau^{-1}\sigma^2\boldsymbol{V}$ and E_{τ} means getting expectation with respect to measure $dW(.)$.
Therefore, $\mathcal{D}_{41} \leq 0$ if and only if $0 < d^* \leq \frac{2m}{m+2}$ since $\int_0^{\infty}\frac{\tau^{-2}}{\dot{\boldsymbol{z}}'\boldsymbol{C}^{-1}\dot{\boldsymbol{z}}}\,dW(\tau) > 0$.

Theorem 5.2. *Suppose in the model* (1.1), $\boldsymbol{\epsilon} \sim \mathcal{E}_n(\boldsymbol{0}, \sigma^2\boldsymbol{V}, \boldsymbol{g})$. *Then the Stein-type shrinkage estimator*

$$\hat{\boldsymbol{\beta}}_*^S = \tilde{\boldsymbol{\beta}} - d(1-\omega)\mathcal{L}_n^{-1}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \tag{5.36}$$

uniformly dominates $\tilde{\boldsymbol{\beta}}$ under the balanced loss function $L_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta})$.

Corollary 5.1. *Suppose in the model* (1.1), $\boldsymbol{\epsilon} \sim \mathcal{E}_n(\boldsymbol{0}, \sigma^2\boldsymbol{V}, \boldsymbol{g})$. *Then $\hat{\boldsymbol{\beta}}^S \succeq \tilde{\boldsymbol{\beta}}$ under the balanced loss function $L_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta})$.*

The proof directly follows from Theorem 5.2 for the special case $\omega = 0$. To compare $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^S$, it is easy to show that

$$\begin{aligned}\boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^S;\boldsymbol{\beta}) &= \boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) + d_n^{-1}\bigg(q\sigma_{\boldsymbol{\epsilon}}^2 - \theta - dq^2\sigma_{\boldsymbol{\epsilon}}^2\Big\{(q-2)E[\chi_{q+2}^{*\,-4}(\Delta_*^2)] \\ &\quad + \left[1 - \frac{(q+2)\theta}{2q\sigma_{\boldsymbol{\epsilon}}^2\Delta_*^2}\right](2\Delta_*^2)E[\chi_{q+4}^{*\,-4}(\Delta_*^2)]\Big\}\bigg).\end{aligned} \tag{5.37}$$

Under H_0, this becomes

$$\boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}}^S;\boldsymbol{\beta}) = \boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) + qd_n^{-1}\sigma_{\boldsymbol{\epsilon}}^2(1-d) \geq \boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}), \text{ with} \tag{5.38}$$

$$\boldsymbol{R}_0^{\boldsymbol{C}}(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = \boldsymbol{R}_0^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}) - qd_n^{-1}\sigma_{\boldsymbol{\epsilon}}^2 \leq \boldsymbol{R}_0^{\boldsymbol{C}}(\tilde{\boldsymbol{\beta}};\boldsymbol{\beta}). \tag{5.39}$$

Therefore, $\hat{\boldsymbol{\beta}} \succeq \hat{\boldsymbol{\beta}}^S$ under H_0 with the balanced loss $L_0^{\boldsymbol{C}}(\boldsymbol{\beta}^*,\boldsymbol{\beta})$. Therefore under H_0, $\hat{\boldsymbol{\beta}} \succeq \hat{\boldsymbol{\beta}}^S$ with the balanced loss $L_{\omega,\tilde{\boldsymbol{\beta}}}^{\boldsymbol{C}}(\boldsymbol{\beta}^*;\boldsymbol{\beta})$. However, as $\boldsymbol{\eta}_1$ moves away from 0, θ increases and the risk of $\hat{\boldsymbol{\beta}}$ becomes unbounded while the risk of $\tilde{\boldsymbol{\beta}}^S$ remains below the risk of $\tilde{\boldsymbol{\beta}}$; thus for similar reasons, $\tilde{\boldsymbol{\beta}}^S$ dominates $\hat{\boldsymbol{\beta}}$ outside an interval around the origin under the balanced loss $L^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}^*;\boldsymbol{\beta})$. This scenario repeats when we compare $\hat{\boldsymbol{\beta}}^S$ and $\hat{\boldsymbol{\beta}}^{PT}$. Under H_0

$$\boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^S;\boldsymbol{\beta}) \quad = \quad \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}) + qd_n^{-1}\sigma_\epsilon^2[1-\alpha-d] \geq \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}),$$

for all α such that $F^{-1}_{q+2,m}(d,0) \leq \frac{qF_\alpha}{q+2}$. This means, $\hat{\boldsymbol{\beta}}^S$ does not always dominate $\hat{\boldsymbol{\beta}}^{PT}$ under H_0. So under H_0, with α satisfying $F^{-1}_{q+2,m}(d,0) \leq \frac{qF_\alpha}{q+2}$, under the balanced loss function we have $\hat{\boldsymbol{\beta}} \succeq \hat{\boldsymbol{\beta}}^{PT} \succeq \hat{\boldsymbol{\beta}}^S \succeq \tilde{\boldsymbol{\beta}}$.. The risk difference of $\hat{\boldsymbol{\beta}}^{S+}$ and $\hat{\boldsymbol{\beta}}^S$ is given by

$$\begin{aligned}
\mathcal{D}_{54} \quad = \quad & \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^{S+};\boldsymbol{\beta}) - \boldsymbol{R}^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^S;\boldsymbol{\beta}) = \\
& -d_n^{-1}\sigma_\epsilon^2\left\{qE^{(1)}\left[\left(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right)^2 I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right]\right. \\
& \left.+\frac{\theta}{\sigma_\epsilon^2}E^{(2)}\left[\left(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right)^2 I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right]\right\} \\
& -2d_n^{-1}\theta E^{(2)}\left[\left(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(\Delta_*^2)\right) I\left(F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}\right)\right].
\end{aligned}$$

The r.h.s. of the above equality is $-$ve since for $F_{q+2,m}(\Delta_*^2) \leq \frac{qd}{q+2}$, $(\frac{qd}{q+2}F_{q+2,m}(\Delta_*^2)-1) \geq 0$ and also the expectation of a positive random variable is positive. Thus $\hat{\boldsymbol{\beta}}^{S+} \succeq \hat{\boldsymbol{\beta}}^S$.

Remark 5.1. *The positive-rule shrinkage estimator* $\hat{\boldsymbol{\beta}}^{S+}$ *of* $\boldsymbol{\beta}$ *is minimax.*

Continue the comparisons under $L^{\boldsymbol{C}}_0(\boldsymbol{\beta}^*;\boldsymbol{\beta})$. The results are the same for the balanced loss $L^{\boldsymbol{C}}_{\omega,\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}^*;\boldsymbol{\beta})$. To compare $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{S+}$, note under H_0, i.e., $\boldsymbol{\eta}_1 = 0$,

$$\begin{aligned}
\boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{S+};\boldsymbol{\beta}) \quad = \quad & \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) + qd_n^{-1}\sigma_\epsilon^2\left\{(1-d) - E\left[(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(0))^2\right.\right. \\
& \left.\left.\times I(F_{q+2,m}(0) \leq \frac{qd}{q+2})\right]\right\} \geq \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}),
\end{aligned}$$

since $E\left[(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(0))^2 I(F_{q+2,m}(0) \leq \frac{qd}{q+2})\right] \leq E\left[(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(0))^2\right] = 1-d$. Thus under H_0, $\hat{\boldsymbol{\beta}} \succeq \hat{\boldsymbol{\beta}}^{S+}$. But, as $\boldsymbol{\eta}_1$ moves away from 0, θ increases and the risk of $\hat{\boldsymbol{\beta}}$ becomes unbounded while the risk of $\tilde{\boldsymbol{\beta}}^{S+}$ remains below the risk of $\tilde{\boldsymbol{\beta}}$; thus $\tilde{\boldsymbol{\beta}}^{S+}$ dominates $\hat{\boldsymbol{\beta}}$ outside an interval around the origin. When H_0 holds, $G^*_{q+2,m}(F_\alpha,0) = 1-\alpha$,

$$\begin{aligned}
\boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{S+};\boldsymbol{\beta}) \quad = \quad & \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}) + qd_n^{-1}\sigma_\epsilon^2\left\{1-\alpha-d-E\left[(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(0))^2\right.\right. \\
& \left.\left.\times I(F_{q+2,m}(0) \leq \frac{qd}{q+2})\right]\right\} \geq \boldsymbol{R}^{\boldsymbol{C}}_0(\hat{\boldsymbol{\beta}}^{PT};\boldsymbol{\beta}), \text{ for all } \alpha \text{ satisfying} \\
& E\left[(1-\frac{qd}{q+2}F^{-1}_{q+2,m}(0))^2 I(F_{q+2,m}(0) \leq \frac{qd}{q+2})\right] \leq 1-\alpha-d.
\end{aligned}$$

Thus, $\hat{\beta}^{S+}$ does not always dominates $\hat{\beta}^{PT}$ when the null-hypothesis H_0 holds.

Therefore the dominance order of five estimators under the balanced loss function $L^{C}_{\omega,\tilde{\beta}}(\beta^{*};\beta)$ can be determine under following two categories

$$1. \quad \hat{\beta} \succeq \hat{\beta}^{PT} \succeq \hat{\beta}^{S+} \succeq \hat{\beta}^{S} \succeq \tilde{\beta}, \text{ and } 2. \quad \hat{\beta} \succeq \hat{\beta}^{S+} \succeq \hat{\beta}^{S} \succeq \hat{\beta}^{PT} \succeq \tilde{\beta}.$$

Acknowledgments

First author's research is supported by Shahrood University of Technology, Iran.

References

1. Anderson, T. W., Fang, K. T. and Hsu, H. (1986). Maximum-likelihood estimates and likelihood-ratio criteria for multivariate elliptically contoured distributions, *The Canadian J. Statist.*, **14**, 55–59.
2. Arashi, M., (2012). Preliminary test and Stein estimators in simultaneous linear equations, *Lin. Alg. Appl.*, **436**(5), 1195-1211.
3. Arashi, M., Tabatabaey, S M M. and Khan, S. (2008). Estimation in multiple regression model with elliptically contoured errors under MLINEX loss, Journal of Applied Probability and Statistics, **3**(1), 23-36.
4. Dey, D., Ghosh, M. and Strawderman, W. E., (1999). On estimation with balanced loss function. *Statist. Prob. Lett.*, **45**, 97-101.
5. Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London, New York.
6. Hoque, Z., Khan, S., and Wesolowiski, J. (2009). Performance of Preliminary Test Estimator under Linex Loss Function, Communications in Statistics-Theory and Methods, **38**(2), 252-261.
7. Jeffreys, H., (1961). *Theory of Probability*, Oxford: Clarendon.
8. Jozani, M. J., Marchand, E. and Parsian, A. (2006). On estimation with weighted balanced-type loss function, *Statist. Prob. Lett.*, **76**, 773-780.
9. Khan, S. (2008). Shrinkage Estimators of Intercept Parameters of Two Simple Regression Models with Suspected Equal Slopes. Communications in Statistics - Theory and Methods, **37**, 247-260.
10. Khan, S (2000). Improved estimation of the mean vector for student-t model, *Communications in Statistics - Theory and Methods*, **29**(**3**), 507-527.
11. Khan, S. and Saleh, A. K. Md. E. (1997). Shrinkage Pre-Test Estimator of the Intercept Parameter for a Regression Model with Multivariate Student-t Errors, *Biometrical Journal*, **39**(**2**), 131147.
12. Saleh, A. K. Md. E. (2006), *Theory of Preliminary Test and Stein-type Estimation with Applications*, John Wiley, New York.
13. Srivastava, M. and Bilodeau, M., (1989). Stein estimation under elliptical distribution, *J. Mult. Annal.*, **28**, 247-259.
14. Zellner, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In: Berger, J. O. and Gupta, S. S. (Eds.), *Statistical Decision Theory and Methods V*, Springer, New York, pp. 337-390.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 529-534

DISTINGUISHING ASTROPHYSICAL SIGNALS USING WAVELET TECHNIQUE

M Ayub Khan Yousuf Zai[1] and **Khusro Mian**[2]

[1] Department of Applied Physics, University of Karachi,
Karachi, Pakistan. Email: ayubzai@yahoo.com

[2] Institute of Space and Planetary Astrophysics,
University of Karachi, Karachi, Pakistan

ABSTRACT

This concept has been studied using wavelet analysis that includes spectral analysis, FFT and 1-D Haar wavelets analysis in different Levels for approximation and detail reconstruction. Most of the small wavelength fractions of the solar radiation that penetrates the upper atmosphere for example energetic UV and EUV radiation that cause turbulence Wavelet analysis is a technique used to extract successfully the limited period that has an average value. Distinguished astrophysical signals using this technique can be detected. We presented estimation with respect to varying ionospheric conditions.

KEYWORDS

Ionosphere, Turbulence, Reconstruction, Spectral analysis, Wavelet, Haar, FFT, UV, EUV.

INTRODUCTION

Ultraviolet radiation from the Sun ionizes molecules of air in the thin upper atmosphere of the Earth. The ions accumulate in several layers, forming the ionosphere at altitudes between about 80 Km and 1500 Km above the Earth's surface. It is formed of five layers D, E, F_1 and F_2 layers from the base to the top. Each layer can reflect radio waves. The thicknesses and ionization of the layers change during the course of a day, all but one or two layers on the night side of the earth disappear while they thicken and strengthen on the day side. A radio transmitter on the day side can bounce signals off the ionosphere that then travels around the world by multiple reflections between the ground surface and ionosphere. The E layer is used by short wave amateur radio enthusiasts. The F layers are the most intensely ionized. Electric currents in the ionosphere arise from systematic motions of the ions, which are a variations in insulation and the periodic fluctuation in ionization related to the eleven years sunspot cycle.

PLASMA WAVE TURBULENCE

Plasma is considered as a quasi neutral gas consisting of ions and electrons upper atmosphere and exhibits collective behavior. It remains in the plasma state as ionized. The ionosphere lies in the region of upper atmosphere of the earth. The ionosphere changes greatly from hour to hour, day to day, day to night and so on. The structure of the ionosphere differs from altitude to different layers. Primarily, this difference is in the density of ions and electrons.

It has been observed that if the electrons in the ionospheric plasma are displaced from a uniform background of ions then electric field will be setup in such a direction as to restore the neutrality of the plasma by pulling the electrons back to their original positions. Because of their inertia, the electrons will overshoot and oscillate around their equilibrium positions with a characteristic frequency known as plasma frequency. This oscillation is so fast that the massive ions do not have time to respond to the oscillating field and may be considered fixed.

It is known that any motion of a fluid like plasma can be represented by Fourier series and decomposed by Fourier analysis into a superposition of sinusoidal oscillations with different frequencies ω and wavelength λ

When the oscillation amplitude is small, the waveform is generally sinusoidal and there is only one component.

The density of ionospheric plasma is a sinusoidal oscillating quantity and thus can be represented as follows:

$$n = \bar{n} \exp[i\,(k\,.\,r\ \_\ \omega\, t)]$$

$K.r = k_x\, x + k_y\, y + k_z\, z$, in Cartesian coordinates

Here $\bar{n}$ is a constant defining the amplitude of the plasma wave and k is the propagation constant. The linearization process needs low power dependent variables that leads to one component of Fourier transform. When one observes that waves have grown to larger steady amplitudes then linear theory does not exit and some nonlinear effect is limiting the amplitude. It has been explained that the amplitude at saturation is rather small and thus a wave experiences a number of changes when its amplitude becomes large. Taking the case of nonlinear Landau damping, if a plasma is so strongly excited that a continuous spectrum of frequencies is present, it is in the state of turbulence. This state must be described statistically as in the case of ordinary fluid hydrodynamics. Plasma wave turbulence leads to anomalous resistivity in which the electrons are slowed down by collision with random electric field fluctuations. The turbulence means crowd of vortices in nonlinear interaction. Turbulence is characterized by the Raynolds number that is the ratio of the nonlinear inertial forces responsible for the flow instability to the linear dissipation damping that converts kinetic energy into thermal energy. Plasma turbulence is the event that plasma and electromagnetic field variation randomly exhibit.

COMPUTATION OF PLASMA WAVE TURBULENCE

We have computed plasma Turbulence flux of Ionospheric layer at Pakistan region, by using perturbed electron concentration (N‘) and perturbed parcel velocity (V'). Turbulent flow varies randomly with time at a location as shown in Figure 1, thus turbulent flow is unsteady.

The product $\overline{U^{/} N^{/}}$ represents the transport of Kinematics Turbulent Flux:

$$\overline{U^{/} N^{/}} = \frac{U_E^{/} N_E^{/} + U_{F2}^{/} N_{F2}^{/}}{2}$$

$U_{F2}^{/}$ is an Instantaneous and Local Perturbation Scalar Velocity content

$U_E^{/}$ is an Instantaneous and Local Perturbation Scalar Velocity content

$N_{F2}^{/}$ is an Instantaneous and Local Perturbation Plasma Concentration content

$N_E^{/}$ is an Instantaneous and Local Perturbation Plasma Concentration content [11]

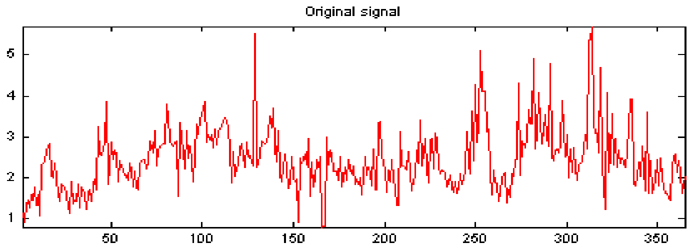


Fig 1: Shows original astrophysical signal generated by plasma turbulence data.

WAVELET ANALYSIS AND ITS 1-D ASPECT

It has been observed that the astrophysical signals exist in the universe with noise. Model conditions are such that this noise may reduce to certain levels it damages the signal and it must be removed in order to recover the desired signal and proceed with further data analysis. This noise removal takes place in the original signal or in a transform domain.

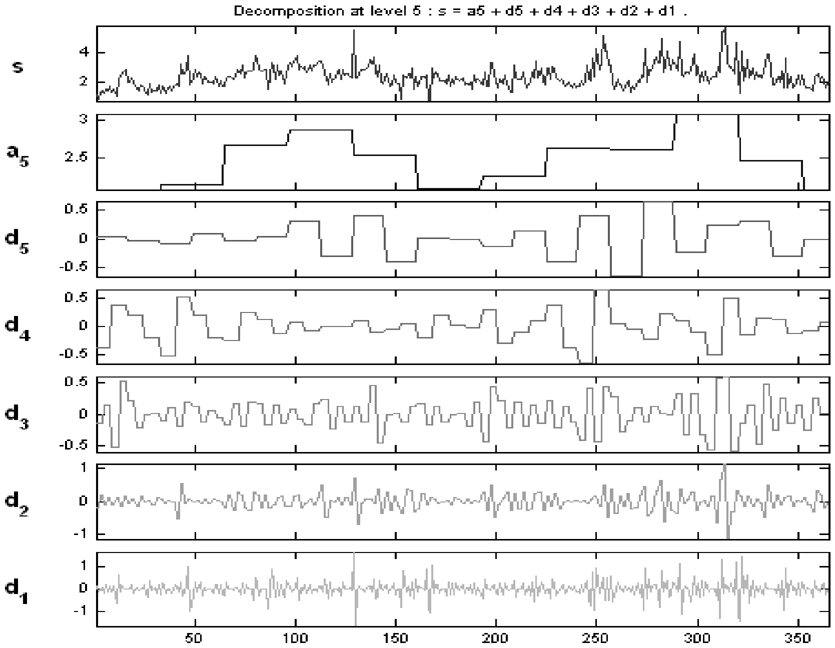


Fig. 2: Representation of the decomposition of the signal s lives in V_0 of order 5 and its approximation $a_5 \in V_{-5}$ and increasingly finer details $d_j \in W_{-j}$, j = 5, 4, 3, 2, 1∈.

In practical concepts 1-D represents the three steps of wavelet with de-noising conditions with plots of a known test signal with added noise will comprise the following:

1-the wavelet transform,

2-the de-noised wavelet transform,

3-and the de-noised signal estimate.

The signal is taken in some V$_{j,}$ state and then the decomposition may be illustrated as

$$L^2(R) = V_j \oplus (\underset{j \geq j_0}{\oplus W_j}),$$

Multi resolution emphasizing the normalization factor $a^{-1/2}$ and a^{-1} L^1 normalization for smll scales or high frequency containing the singularities of the signal and the choice $a^{-1/2}$ makes the transform unitary. L^2 norm is interpreted as total energy of the signal then replaced by the finite representation

$$V_j = V_{j_0} \oplus (\bigoplus_{j=j_0}^{j=1} W_j),$$

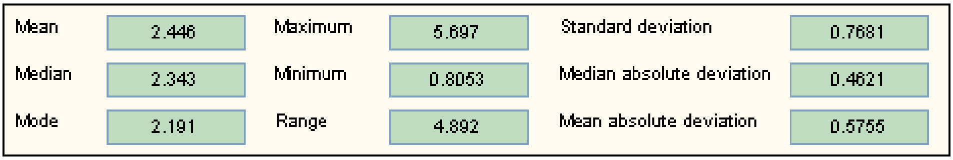
Fig.2. shows decomposition of signal, approximation and details order 5,

$$V_0 = V_{-5} \oplus W_{-5} \oplus W_{-4} \oplus W_{-3} \oplus W_{-2} \oplus W_{-1}$$

As we just saw, appropriate filters generate orthogonal wavelets bases. The signal s lives in V$_0$ and it is decomposed into its approximation $a_5 \in V_{-5}$ and the increasingly finer details $d_j \in W_{-j}, j = 1,2,3,4,5.$ in figure 2 to 7 we find, from top to bottom, the original signal s, the approximation at level 5, a$_5$ and details from the coarsest level d$_5$ to the finest level d$_1$. All the signals are expressed in the same of time unit, which allows a synchronous reading of all the graphs.

We can state that d$_1$ contains the components of the signals of periods between d$_1$ to d$_5$. The analysis makes it possible to track possible outliers, which are detected to the very large values of d$_1$ around the position 125 and 325. We can distinguish the details d$_1$ and d$_2$ measurement and state noises which give in details oscillating to zero. In details d$_3$ also have noise at the position 300 to 325. We cannot distinguish periods in the details d$_4$ and d$_{5.}$

Table 1 shows descriptive statistics astrophysical of plasma turbulence data, the St. Dev. is 0.7681, Mean is 2.44 and MAD is 0.5755.



Mean	2.446	Maximum	5.697	Standard deviation	0.7681
Median	2.343	Minimum	0.8053	Median absolute deviation	0.4621
Mode	2.191	Range	4.892	Mean absolute deviation	0.5755

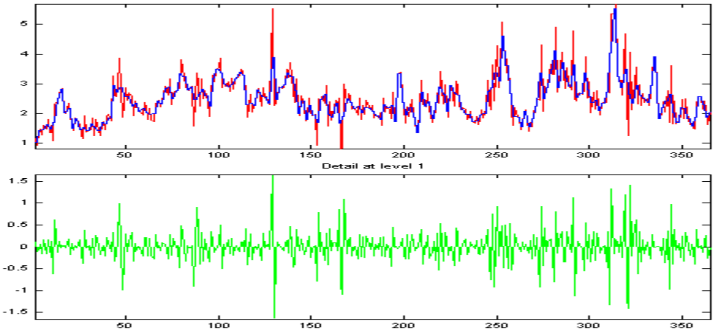


Fig 3: shows upper panel signal, approximation level 1 and lower panel details level 1,
Fig 4: shows upper panel signal, approximation level 2 and lower panel details level 2,

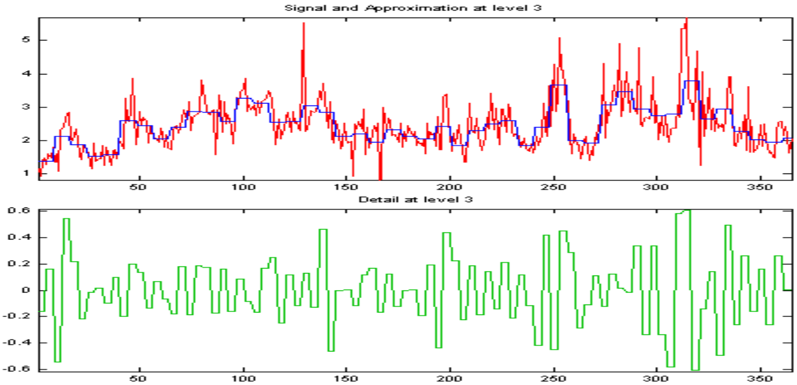


Fig 5: shows upper panel signal, approximation level 3 and lower panel details level 3,

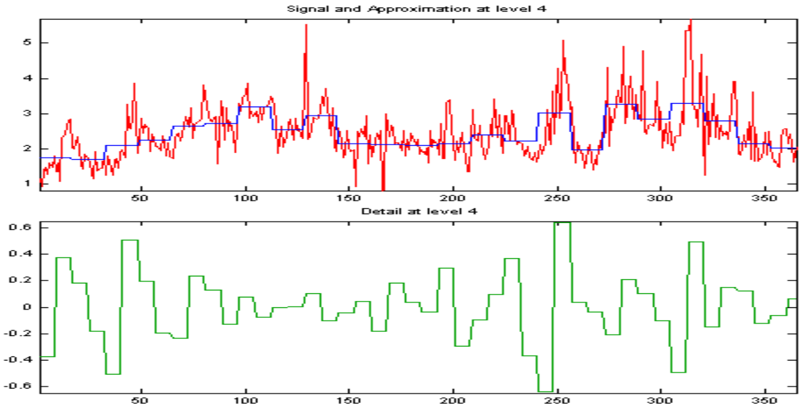


Fig 6: shows upper panel signal, approximation level 4 and lower panel details level 4,

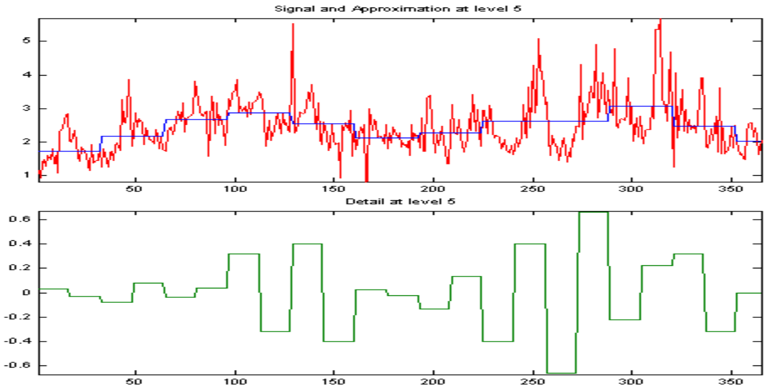


Fig 7: shows upper panel signal, approximation level 5 and lower panel details level 5,

Notice that the red colour indicates original signal, blue signal depicts approximation level. And Green signal illustrates the detailed

CONCLUSION

One of basic results in the theory of weakly plasma turbulence is the reconstruction of 1-D Haar wavelets in different Levels for approximation and detail filters. Wavelet approach is a currently developed Mathematical Statistical tool that illustrates to bring new insights to evaluate recent estimates it has been inspected that they will lead to a better understanding of plasma turbulence in upper atmosphere. Distinguished astrophysical signals with decomposition of signal, approximation and details order 5 have been discussed the problem of the determination of the influence of the plasma turbulence.

ACKNOWLEDGEMENT

The author (s) would like to thank the organizers of this conference for providing local hospitality in the form of free accommodation and registration fees waiver. This work was completed under the auspices of Institute of Space and Planetary Astrophysics (ISPA) University of Karachi, Pakistan. The services rendered by the technical staff of SUPARCO for delivery of ionospheric plasma layers data were also appreciated.

REFERENCES

1. Mark Z. Jacobson, (1999). *Fundamentals of Atmospheric Modeling*. IST Published Cambridge University Press USA p-60-65.
2. SM Pandit, (1983). *Time Series and System Analysis with Application*. John Wiley & Sons, Inc., USA p 24, 29, 119.
3. Chatfield C., (1989). *The Analysis of Time Series: An Introduction.* 4/e, Chapman & Hall, London p105-35.
4. Michel, Misiti, (2007). *Wavelets and Their Applications*. ISTE, Ltd. USA.p-13-23.
5. J C Van Den Berg. (2004). *Wavelets in Physics.* Cambridge University Press UK. p 8-15.
6. Don Hong, (2005). *Real Analysis with an Introduction to Wavelets and Applications*. Elsevier Inc India. p-123-213.
7. W Lowrie, (1997). *Fundamentals of Geophysics*. Cambridge university press, UK.
8. H Rishbeth, (1969). *Introduction to Ionospheric Physics.* Academic Press New York and London. p 4-31.
9. May–Britt Kallenrode (2004). *Space Physics (An Introduction `to Plasma and Particles in the Heliosphere and Magnetospheres.* 3rd Ed Springer –Verlag Berlin Heidelberg New York p 301-302, 166-167.
10. Chen M.C. (1929). *Introduction to Plasma Physics.* Plenum Publishing Corporation New York p.243.
11. Barclay L, (2003). *Propagation of Radio waves.* The Institution of Electrical Engineers, London.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 535-542

DISTANCE-BASED PARTITION OF MULTIVARIATE DATA

Maryam Tayefi and **Sharad Gore**
Department of Statistics, University of Pune, Pune, India
Email: maryam_tayefi@yahoo.com; sdgore@stats.unipune.ac.in

ABSTRACT

Multivariate data has what is popularly called curse of dimensionality. High dimensionality of data makes it hard to visualize, analyze or even handle data for different inferential methods. One way of reducing the effect of the curse of dimensionality is to reduce the dimension of data by carrying out procedures like the principal component analysis or factor analysis. This method reduces the dimension of the data by reducing the number of variables but the data size remains unchanged. The other method is to reduce the data size by identifying observations similar to each other so that they can be represented by only one. Thus reducing the data-size rather than the data dimension.

Grouping of data is simple in case of univariate data due to complete ordering. It is not so simple in case of multivariate data due to the data complexity. Several researchers have suggested different clustering algorithms for this purpose and most of these algorithms work well under suitable conditions. These clustering algorithms have the main problem of time- and space-complexity of the algorithm. The standard clustering algorithms are either iterative or recursive, thus increasing the time-complexity. A new algorithm is suggested in this paper that obtains a partition of the given data set using the distance matrix.

One a partition of the given data set is obtained, it can be used as the initialization for a partitioning clustering algorithm or as an intermediate stage in hierarchical clustering. Either way, this approach will reduce the time and space complexity of a clustering problem for a high dimensional large data set.

KEYWORDS

Clustering; Partition; Hierarchical; K-means; Time and Space Complexity.

1. Introduction

Cluster analysis is one of the core activities in data mining. Data mining aims at processing huge amounts of data for finding useful patterns in data. Patterns in data can be explored and exploited for developing prediction models. Data mining begins with exploratory data analysis and proceeds with supervised or unsupervised statistical learning methodologies with a view to identify useful features in the data. Features are easier to identify in homogeneous data. It is therefore common to first carry out a classification or clustering method in order to obtain homogeneous segments of the given data. Data segments can then be used to develop descriptive or predictive models for later use. For more discussion on application of cluster analysis, see Anderberg, 1973.

There are two standard methods of forming clusters. These are hierarchical and partitioning. Information on different methods of clustering can be found in [7]. Details about the hierarchical method can be found in [8], [9], and [10], while more information on k-means clustering can be found in [11]. The hierarchical methods begin with all individual observations as clusters of size one each and then proceeds with merging the nearest clusters at successive steps, ultimately merging all observations into a single clusters that contains all the observations. For more details, see Everitt, 1974 and Rui Xu *et al.*, 2009. The partitioning method, by contrast, begins with the specification of the number K of clusters and K cluster centroids. Every observation is assigned to the cluster having the centroid nearest to the observation. After assigning all observations to the nearest clusters, cluster centroids are computed using the observations assigned to those clusters. Individual observations are again assigned to the clusters according to the distances between observations and a cluster centroids. For more details on K-means clustering method, refer to Hartigan, 1979 and Rui Xu *et al.*, 2009. In this way, both hierarchical and partition-based methods are repetitive and hence have a high time complexity. The use of distances is common among all these methods, but the hierarchical methods also have to incorporate a linkage rule to generalize the distance between two individual observations to the distance between two clusters. According to the nature of repetitive computations in the two algorithms, hierarchical algorithms are recursive while partition-based algorithms are iterative.

An obvious question then is whether it is possible to segment multivariate data into homogeneous sub-samples without a repetitive algorithm. This paper describes an attempt to do so by partitioning the given data set with help of the distance matrix. For more discussion on partitioning, see Philip, 1966, and Kokolakis *et al.*, 2009. An additional feature of the method proposed in this paper is that the partition obtained through this procedure is purely guided by data through the distance matrix and there is no subjectivity. Hierarchical clustering has no facility to identify the optimal number of clusters, and is subjective in this aspect. K-means clustering has no unique choice of initial cluster centroids and thus has subjectivity in this regard. The proposed algorithm avoids either type of subjectivity and requires only the distance matrix for implementation.

The following section (Section 2) describes the concept of partitioning of a given data set. It also contains some discussion of limitations or drawbacks of the hierarchical and K-means clustering algorithms. It finally establishes the need for a non-iterative and non-recursive partitioning algorithm. The next section (Section 3) describes the proposed partitioning algorithm in detail. The algorithm is simple and straightforward and can be programmed easily. As a matter of fact, the R code for this algorithm is included in the paper as an appendix. The section (Section 3) detailing the proposed algorithm is followed by an illustrative example (Section 4). Here, we consider the `iris` data and form clusters using different clustering algorithms. It is interesting to note that the partitioning algorithm results in a large number of partition sets. It may then be appropriate to merge some of these partition sets using either K-means clustering or hierarchical clustering. These options are also discussed and results are obtained using alternative methods available. The paper ends with Appendix A giving the R script for implementing the proposed algorithm and a list of references (including some websites) used in preparation of this paper.

2. Partitioning of Multivariate Data

Classification algorithms aim at partitioning of multivariate data into distinct and identifiable classes. If the classes are defined through parameters like the mean vectors, then the classification will be supervised, classically called the discriminant analysis (2 classes) or classification problem (more than two classes). If the classes are not pre-defined, then the classification is unsupervised and is often called clustering. Clustering of the given data set is classically done using one of the two types of algorithms: partitioning or hierarchical. Both the types ultimately provide a partition of the specified data, but the procedures differ. All clustering algorithms require an appropriate distance function, while the hierarchical clustering algorithm also requires a linkage rule.

Since the purpose of partitioning the given data is to form partition sets in such a way that the point-to-point similarity within a set is maximal and set-to-set similarity (i.e, between sets) is minimal, it is important to define a measure of similarity between two observations. It is often more convenient to measure distance between two observations than similarity between them. Some of the commonly used distance functions are introduced in the following subsection.

2.1. Distance Function

The purpose of partitioning a given data set is usually to identify observations that are similar to one another and allocate them to a partition set. It implies that observations allocated to different partition sets are not as similar (to one another) as observations allocated to the same partition set. This would require that the distance function permits a comparison among observations and allows us to decide which observations are closest to each other.

Researches have introduced and defined several distance functions. The five commonly used distance functions are *Euclidean, Manhattan, Maximum, Mahalanobis,* and *Canberra.* The choice of distance function depends on the nature of the variables in the data. For instance, the Euclidean distance is available only if all variables are quantitative or measurement type, and not qualitative by nature. Mahalanobis distance may be appropriate when variables have different units of measurement and hence are not comparable in magnitude. Let us now consider the standard clustering algorithms, namely, the K-means clustering and the hierarchical clustering.

2.2. K-Means Clustering

The $K-$means clustering algorithm is implemented to form K clusters from a multivariate data in such a way that the within cluster variation is minimized while maximizing the between clusters variation. This is achieved by assigning an individual observation to the cluster whose mean vector is closest to the observation. More precisely, let $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ be the n observations in the sample. Also, let $\underline{\mu}_1, \underline{\mu}_2, \cdots, \underline{\mu}_K$ be means of the K clusters. For $i = 1, 2, \cdots, n$, the observation $\underline{X}_{i^\star}$ is classified in to the cluster number $j^\star$ if

$$\left|\underline{X}_{i^\star} - \underline{\mu}_{j^\star}\right| = \min_{i,j} \left|\underline{X}_i - \underline{\mu}_j\right|.$$

Note that the K-means clustering algorithm defines a partition of the data set of size n into K mutually exclusive and exhaustive clusters $C_1, C_2, \cdots, C_K$.

Also note that the K-means clustering algorithm may not produce a unique partition. The choice of initial cluster means and the number of iterations can influence the final cluster formation. There is no remedy to this subjectivity of the K-means clustering algorithm. Another drawback of the K-means clustering is its iterative nature. Large data sets with a significant amount of heterogeneity may take a huge amount of time to converge or may produce incorrect results. Finally, the K-means clustering algorithm requires specification of the number of clusters, K, and a wrong specification may result in a meaningless partition. It is therefore possible to use a hierarchical clustering algoeithm that does not require specification of the number of clusters.

2.3. Hierarchical Clustering

The hierarchical clustering algorithm is a recursive algorithm. The agglomerative clustering algorithm merges the closest two clusters at every repetition, thus going through a total of $n-1$ repetitions until all n data points form a single cluster. The divisive clustering algorithm splits a single cluster into two at every repetition, thus having a total $n-1$ repetitions until every data point forms a separate (singleton) cluster. The agglomerative clustering algorithm uses the distance matrix as the clustering criterion. The raw data is used to compute distances between data points, but the agglomerative clustering algorithm also requires a measure for distance between clusters. This is specified by a linkage rule. Four commonly used linkage rules are single linkage, complete linkage, average linkage and centroid linkage. It is important to note that the original distance function measures the distance between two observations, while the linkage rule defines the distance between two clusters. While the first two linkage rules can work only with the distance matrix, the last two require the original data points in addition to the distance matrix. In this way, if one selects one of the five distance functions listed above and one of the four linkage rules, there are 20 different versions of the hierarchical agglomerative clustering algorithm. Note that the hierarchical clustering algorithm is often preferred to the K-means clustering algorithm because the former does not require specification of the number of clusters, though researchers have worked on variations of the algorithm to determine the "optimal" number of clusters in one way or the other.

Note a feature of the hierarchical clustering algorithm. It forms several small clusters and can merge any of two of them at the next step. This makes it difficult to keep track of any single large cluster as it gets formed. It is natural for a researcher to trace the formation of one cluster at a time. The K-means clustering modifies all the K clusters at every iteration and therefore has the same disadvantage. A new partitioning algorithm is suggested in this paper that avoids this situation. The new algorithm allows one cluster to grow at a time up to its natural potential. Once a cluster can grow no more, another cluster is allowed to grow. This is hoped to help the researcher understand and interpret the clusters formed by the new algorithm. Another feature of the new algorithm is that it is neither recursive nor iterative. As a result, it does not require a linkage rule even though one data point is incrementally added to a cluster at a time. The new method does not remove a data point from a cluster after adding it to the cluster. The number of clusters is neither specified nor arbitrary. The number of clusters is determined by the distance matrix and hence is unique for a given data set. The hierarchical algorithms are recursive while the partitional algorithms are iterative. Due to their time complexity, both of these types of algorithms can be time taking and hence slow. An alternative partitioning method is proposed in this paper that avoids the repetitive nature of both the hierarchical and partitional algorithms. The proposed method is neither recursive nor iterative. It processes the distance matrix only once and produces a partition of the given data.

3. The Partitioning Algorithm

The data $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ is arranged in the form of a matrix

$$\mathcal{X} = \begin{bmatrix} \underline{X}_1' \\ \underline{X}_2' \\ \vdots \\ \underline{X}_N' \end{bmatrix}.$$

Rows of $\mathcal{X}$ represent cases or observations and its columns represent variables or atributes. The data matrix has n rows and P columns. The distance between two observations $\underline{X}_i$ and $\underline{X}_j$ is denoted by $d_{i,j} = d(\underline{X}_i, \underline{X}_j)$ for $i, j = 1, 2, \cdots, n$. All the distances are arranged in the form of the distance matrix

$$D = ((d_{ij})) = ((d(\underline{X}_i, \underline{X}_j))).$$

Note that D is an $n \times n$ symmetric matrix having zeros along the diagonal.

We use the distance matrix D to obtain a partition of the data set $\underline{X}_1, \underline{X}_2, \cdots, \underline{X}_n$ into mutually exclusive and exhaustive subsets. The number or size of partition sets is not pre-determined and it emerges as the partition sets are formed sequentially. Nevertheless, these are unique for any given data set in the sense that rearranging the data by permuting observations or variables will not change the final result. In this sense, the partitioning algorithm proposed in this paper produces a unique partition of the given data set.

The partitioning algorithm can be described as follows.

Step 1. Initialization.

In every column of the distance matrix, identify and mark the smallest (non-zero) entry, so that every column has at least one marked entry. Let R_j denote the row that has a marked entry in column number j for $j = 1, 2, \cdots, N$.

Step 2. Begin a new partition set.

Suppose $d_{i^*,j^*} = \min_{i,j}\{d_{i,j}\}$ so that observations $\underline{X}_{i^*}$ and $\underline{X}_{j^*}$ are closest to each other. Then we begin a new partition set $P = i^*, j^*$ by including these two observations in the set.

Step 3. Check completion of partition set.

In the distance matrix, check if there are any marked entries in rows numbered i^* and j^*. If there is no marked entry in either of these rows, then follow step 4. Otherwise follow step 6.

Step 4. The partition set is completed.

In this case, modify the distance matrix D by removing rows and columns numbered i^* and j^*.

Step 5. Check for termination of procedure.

Check if the distance matrix has at least two rows and at least two columns. If so, follow step 2 to begin a new partition set.

Otherwise, there are no more observations in data set to partition. Hence the procedure terminates.

Step 6. Add observations to the partition set P.

Let $C_{i\cdot}$ denote the column number having a marked entry in row number i^*. Add observation numbered $C_{i\cdot}$ to the partition set P. This procedure is followed for every marked entry in row numbered i^*.

Step 7. Add more observations to the partition set P.

Let $C_{j\cdot}$ denote the column number having a marked entry in row numbered j^*. Add observation numbered $C_{j\cdot}$ to the partiton set P. This procedure is followed for every marked entry in row numbered j^*.

Step 8. For every observation numbered K that is added to the partition set P in step 6 and step 7, check if there is any marked entry in row numbered K and add the corresponding observation to the partition set P.

Step 9. Repeat Step 8 until there is no marked entry in any row corresponding to the observations in the partition set P. Declare completion of the partition set P.

Step 10. Update the distance matrix D by removing rows and columns corresponding to observations in the partition set P.

Follow Step 2 to begin a new partition set.

Step 11. The algorithm is complete and has partitioned the given data set.

4. Illustrative Example

We use Fisher's `iris` data for the purpose of illustrating the partitioning algorithm. The `iris` data has 150 observations on four variables, namely sepal length, sepal width, petal length and petal width. There are three species, namely Setosa, Versicolor and Virginica, with 50 observations on every species. The partitioning algorithm produces 42 partition sets as listed below.

$P_1 = \{1, 18, 41\}$, $P_2 = \{8, 36, 40, 50\}$, $P_3 = \{2, 10, 13, 26, 35, 46\}$, $P_4 = \{11, 17, 37, 49\}$,
$P_5 = \{105, 109, 129, 133\}$, $P_6 = \{3, 48\}$, $P_7 = \{5, 38\}$, $P_8 = \{9, 14, 39, 42, 43\}$,
$P_9 = \{20, 22, 45, 47\}$, $P_{10} = \{28, 29\}$, $P_{11} = \{4, 7, 12, 25, 30, 31\}$, $P_{12} = \{58, 61, 94, 99\}$,
$P_{13} = \{64, 74, 79, 92\}$, $P_{14} = \{55, 59, 66, 76, 77\}$, $P_{15} = \{70, 80, 81, 82\}$, $P_{16} = \{63, 65, 68, 83, 93\}$,
$P_{17} = \{62, 89, 95, 96, 97, 100\}$, $P_{18} = \{104, 117, 135, 138\}$, $P_{19} = \{71, 128, 139, 150\}$,
$P_{20} = \{113, 140\}$, $P_{21} = \{124, 127, 147\}$, $P_{22} = \{23, 24, 27, 44\}$ $P_{23} = \{54, 60, 90\}$,
$P_{24} = \{56, 67, 85, 91, 107\}$ $P_{25} = \{72, 75, 98\}$, $P_{26} = \{111, 112, 148\}$, $P_{27} = \{110, 121, 125, 144\}$,
$P_{28} = \{101, 116, 137, 149\}$ $P_{29} = \{141, 145\}$ $P_{30} = \{142, 146\}$ $P_{31} = \{51, 53, 78, 87\}$,
$P_{32} = \{52, 57, 86\}$, $P_{33} = \{69, 88\}$, $P_{34} = \{102, 114, 115, 122, 143\}$, $P_{35} = \{106, 119, 123\}$,
$P_{36} = \{108, 131, 136\}$, $P_{37} = \{21, 32\}$, $P38 = \{6, 19\}$, $P_{39} = \{73, 84, 120, 134\}$,
$P_{40} = \{15, 16, 33, 34\}$, $P_{41} = \{103, 126, 130\}$, $P_{42} = \{118, 132\}$.

In order to compare the partitioning algorithm with the traditional clustering algorithms, both the K-means and hierarchical algorithms are applied to the same data with the following results.

The K-means clustering algorithm with $K = 42$ produces the following clusters.

Finally, we implemented the hierarchical agglomerative clustering algorithm with Euclidean distance function and complete linkage rule to obtain 42 clusters. The result was the following clusters.

It is interesting to note that eight clusters are identical in partitioning and K-means clustering. These are listed in the following table.

Cluster Number	K_{10}	K_{16}	K_{20}	K_{25}	K_{26}	K_{29}	K_{30}	K_{38}
Cluster Number	P_{34}	P_{12}	P_{33}	P_{19}	P_{32}	P_{35}	P_{42}	P_{21}
Cluster Size	5	4	2	4	3	3	2	3

Similarly, it was found that eleven clusters are identical in partitioning and hierarchical clustering. These are listed in the following table.

Cluster Number	H_4	H_{10}	H_{13}	H_{14}	H_{16}	H_{17}	H_{25}	H_{26}	H_{29}	H_{32}	H_{37}
Cluster Number	P_{38}	P_9	P_{31}	P_{32}	P_{14}	P_{24}	P_{33}	P_{19}	P_{34}	P_{35}	P_{42}
Cluster Size	2	4	4	3	5	4	2	4	5	3	2

The four variables, namely sepal length, sepal width, petal length and petal width are supposed to vary between the three species, namely Setosa, Versicolor and Virginica. It would therefore be interesting to note the clusters that contain observations from more than one species, indicating smaller differences between species in comparison to the differences within species.

5. Program Listing

```
rm(list = ls());
data(iris);
mydata=iris[,1:4];
M=as.matrix(dist(mydata,method="euclidean",diag=TRUE,upper=TRUE));
M[M==0]=Inf; nM=nrow(M); index=c(); C=c(); Call=c(); rowmin=c(); clusmat=c();
sink("partitioneuclidean.txt",append=FALSE,split=FALSE)
case=1; while (length(Call)<nM)
{ for (i in 1:ncol(M)){index[i]=min(which(M[,i]==min(M[,i])))};
index[Call]=NA; matmin=min(M);
for (i in 1:nrow(M)){rowmin[i]=min(M[i,])};
ind2=which(rowmin==min(rowmin));
r=ind2[1];c=index[r]; C=c(r,c); c=which(index==r);C=append(C,c);
for (r in c) {new=which(index==r);C=append(C,new);}
if(length(new)>1) {for (r in new){ c=which(index==r); C=append(C,c);}}
C=sort(unique(C)); Call=append(Call,C); cnum=rep(case,length(C));
mydatac=cbind(C,cnum); clusmat=rbind(clusmat,mydatac);
cat("Cluster Number",case,":  {");cat(C,"}","\n");
for (r in C) {M[r,]=Inf; }; for (c in C) {M[,c]=Inf; };
case=case+1; }
sink();
```

REFERENCES

1. Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Academic Press: New York.

2. Everitt, B. (1974). *Cluster Analysis* , London: Heinemann Educ.

3. Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics*, **28**, 100-108.

4. Kokolakis, G. and Fouskakis, D., (2009). Importance partitioning in micro-aggregation, *Computational statistics and Data Analysis*,**53**, 2439-2445.

5. Philip, A.E. and Mcculloch, J. W., (1966). Use of social indices in psychiatric epidemiology, *British Journal of Preventive and Social Medicine*, **20**, 122-126.

6. Rui Xu and Donald C. Wunsch (2009). *Clustering*, John Wiley and Sons, Inc.

7. http://sites.google.com/site/dataclusteringalgorithms

8. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

9. http://cgm.cs.mcgill.ca/∼soss/cs644/projects/siourbas/sect5.html

10. http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-clustering-1.html

11. aut http://www.autonlab.org/tutorials/kmeans.html

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 543-552

ESTIMATION FOR STOCHASTIC VOLATILITY MODEL: QUASI-LIKELIHOOD APPROACH

Raed Alzghool
Department of Mathematics, Faculty of Science
Al-Balqa' Applied University, Al-Salt, Jordan.
Email: raedalzghool@bau.edu.jo

ABSTRACT

The quasi-likelihood approach for estimation of Stochastic Volatility Model (SVM) is proposed in this paper. In the proposed method both state variables and unknown parameters are estimated by quasi-likelihood approach. The quasi likelihood approach is quite simple and standard and can also be carried out without full knowledge on the probability structure of relevant Stochastic Volatility Model.

Keywords: Stochastic Volatility Model(SVM), Quasi-likelihood(QL), Quasi-Score Estimating Function, Quasi-likelihood Estimate(QLE)

1. Introduction

The Stochastic Volatility Model (SVM) is a frequently used model for returns of financial assets. the stochastic volatility process is defined by

$$y_t = \sigma_t \xi_t = e^{\alpha_t/2}\xi_t, t = 1, 2, \cdots, T, \tag{1}$$

and

$$\alpha_t = \gamma + \phi\alpha_{t-1} + \eta_t, t = 1, 2, \cdots, T, \tag{2}$$

where both ξ_t and η_t i.i.d respectively; η_t has mean 0 and variance σ_η^2 and ξ_t has mean 0 and variance σ_ξ^2. Applications, together with estimation for (SVM), can be found in Jacquier, et al (1994); Briedt and Carriquiry (1996); Harvey and Streible (1998); Sandmann and Koopman (1998); Pitt and Shepard (1999); Alzghool and Lin (2007, 2010); Alzghool (2008). Sandmann and Koopman (1998) introduced the Monte Carlo maximum likelihood method of estimating stochastic volatility models (SVM). Davis and Rodriguez-Yam (2005) proposed an alternative estimation procedure which is based on an approximation to the likelihood function.

In this paper we will follow a different approach, the quasi-likelihood (QL) approach, to estimate the parameters and predictors of Stochastic Volatility Model(SVM). In literature, the QL approach has been applied to stochastic volatility models (SVM) by Papanastassiou and Ioannides (2004). They use and extend set of Kalman filter equations in their estimation procedure, and have restricted themselves to a linear state space model. The Kalman filter and the smoother are methods used to estimate predictors of state-variables and one-step-ahead predictors of observations. Usually the Kalman filter is derived through the maximum likelihood method. This means that the probability structure of the underlying model needs to be known. However, in practice, knowing system probability structure usually is not realistic. Furthermore, the likelihood function is often difficult to calculate. For these reasons, the maximum likelihood method is often difficult to apply in practice. Furthermore, Kalman filter involves many complex matrices calculation, which sometime makes the estimation procedure complex in practice. Not like the QL approach given by Papanastassiou and Ioannides (2004), To avoid complexity expression of Kalman filter matrix, We propose to apply the QL method only to the whole estimation procedure of SVM. We will demonstrate and show the proposed estimating procedure is less complex and easily implemented in estimating the state and parameters in stochastic volatility models (SVM).

The QL method was first introduced by Wedderburn (1974). Wedderburn's work was mainly based on the generalized linear model. At the same time, a similar technique was also independently developed by Godambe and Heyde. This technique later was called "quasi-likelihood" (see, Godambe and Heyde, (1987)). The later technique is more focused on the applications to the inference of stochastic processes. These two independently developed quasi-likelihood methods are defined in different ways because the original approaches are different. The definition given by Godambe and Heyde (1987) is more general than that given by Wedderburn (1974). For this aspect of discussion see Lin and Heyde (1993). In this paper, we will adopt the definition of the quasi-likelihood given by Godambe and Heyde (1987). For detail knowledge on the quasi-likelihood method see Heyde (1997).

Consider a stochastic process $y_t \in R^r$,

$$y_t = \mu_t(\theta) + m_t, 0 \le t \le T \tag{3}$$

where $\theta \in \Theta \in R^p$ is the parameter needed to be estimated; μ_t is a function vector of $\{y_s\}_{s<t}$; (in the other words, μ_t is $\mathcal{F}_{t-1}$-measurable); m_t is an error process with $E(m_t|\mathcal{F}_{t-1}) = E_{t-1}(m_t) = 0$. When the following estimating function space

$$\mathcal{G}_T = \{\sum_{t=1}^{T} A_t(y_t - \mu_t)|A_t \text{ is a } \mathcal{F}_{t-1}\text{-measurable } p \times r \text{ matrix}\}$$

is considered, the standard quasi-score estimating function in the space has form

$$G_T^*(\theta) = \sum_{i=1}^{T} E_{t-1}(\dot{m}_t)(E_{t-1}(m_t m_t'))^{-1} m_t, \tag{4}$$

where $\dot{m}_t = \frac{\partial m_t}{\partial \theta}$ and "$'$" denotes transpose. The solution of $G_T^*(\theta) = 0$ is the quasi-likelihood estimator of θ. For a special scenario, if we only consider sub estimating function spaces of $\mathcal{G}_T$, for example,

$$\mathcal{G}_T^{(t)} = \{A_t(y_t - \mu_t)|A_t \text{ is a } \mathcal{F}_{t-1}\text{-measurable } p \times r \text{ matrix}\} \subset \mathcal{G}_T, t < T,$$

then, the standard quasi-score estimating function in this space is

$$G^*_{(t)}(\theta) = E_{t-1}(\dot{m}_t)(E_{t-1}(m_t m_t'))^{-1} m_t \tag{5}$$

and $G^*_{(t)}(\theta) = 0$ will give the quasi-likelihood estimator based on the information provided by $\mathcal{G}_T^{(t)}$.

In next section, we use these results to develop an inference approach for estimating parameters in SVM. This approach can be carried out without full knowledge of the probability structure of underlying system y_t.

This paper is organized as follows. In Section 2.1, we apply the QL approach to SVM. In Sections 2.2, we demonstrate the QL approach via simulation study. In Section 2.3 we apply the QL approach to real data. A summary is given in Section 3.

2. Parameter estimation

2.1. The quasi-likelihood approach

In this subsection we introduce how to use the QL approach to estimate parameters in SVM without borrowing the transition matrix introduced in the standard Kalman filter method. Consider the following stochastic volatility model

$$y_t = f(\alpha_t, \theta) + \epsilon_t, t = 1, 2 \cdots, T, \tag{6}$$

$$\alpha_t = h(\alpha_{t-1}, \theta) + \eta_t, t = 1, 2 \cdots, T, \tag{7}$$

where $\{y_t\}$ represents the time series of observations, $\{\alpha_t\}$ the state variables, θ unknown parameter taking value in an open subset Θ of p-dimensional Euclidean space. Both f and h are functions satisfying certain regularity conditions, and the error terms ϵ_t and η_t are independent. Denote $\delta_t = (\epsilon_t, \eta_t)'$. Then δ_t is a martingale difference with

$$E_{t-1}(\delta_t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$Var_{t-1}(\delta_t) = \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}.$$

Traditionally, normality or conditional normality condition is assumed and the estimation of parameters are obtained by the ML approach. However, in many applications the normality assumption is not realistic. Further more, the probability structure of the model may not be known. Thus the maximum likelihood method is not applicable or it is too complex to estimate parameters through the ML method as the calculation involved is complex sometimes. In the following the QL approach for estimating the parameters in SVM is introduced. This approach can be carried out without full knowledge of the system probability structure. It involves in making decision about the initial values of θ and iterative procedure. Each iterative procedure consists of two steps. The first step is to use the QL method to obtain the optimal estimation for each α_t, say $\hat{\alpha}_t$. The second step is to combine the information of $\{y_t\}$ and $\{\hat{\alpha_t}\}$ to adjust the estimate of θ through the QL method.

In Step 1, assign an initial value to θ and consider the following martingale difference

$$\delta_t = \begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} = \begin{pmatrix} y_t - E(y_t|\mathcal{F}_{t-1}) \\ \alpha_t - E(\alpha_t|\mathcal{F}_{t-1}) \end{pmatrix}$$

and estimating function space

$$\mathcal{G}_T^{(t)} = \{A_t\delta_t | A_t \text{ is } \mathcal{F}_{t-1} \text{ measurable}\},$$

where α_t is considered as an unknown parameter. As mentioned in (5), a standardized optimal estimating function in this estimating function space is

$$G^*_{(t)}(\alpha_t) = E_{t-1}(\frac{\partial \delta_t}{\partial \alpha_t})[Var_{t-1}(\delta_t)]^{-1}\delta_t.$$

To obtain the QL estimate $\hat{\alpha_t}$ of α_t, we let $G^*_{(t)}(\alpha_t) = 0$ and solve the equation for α_t. This estimation is as same as the estimation given by Kalman filter approach when the underlying system has a normal probability structure. (For detailed discussion see Lin, (2007)).

In Step 2, θ is considered as an unknown parameter and the estimating function space

$$\mathcal{G}_T = \{\sum_{t=1}^{T} A_t\delta_t | A_t \text{ is } \mathcal{F}_{t-1} \text{ measurable}\}$$

is considered. Then the standardized optimal estimating function in this estimating function space is

$$G^*_T(\theta) = \sum_{t=1}^{T} E_{t-1}(\frac{\partial \delta_t}{\partial \theta})[Var_{t-1}(\delta_t)]^{-1}\delta_t.$$

To obtain the QL estimate $\hat{\theta}$ for θ we let $G^*_T(\theta) = 0$ and solve the equation while replacing α_t by $\hat{\alpha_t}$ obtained from Step 1. The $\hat{\theta}$ obtained from Step 2 will be used as a new initial value for the θ in Step 1 in the next iterative procedure. These two steps will be alternatively repeated till certain criterion is met.

When σ^2_ϵ and σ^2_η are unknown, a procedure for estimating σ^2_ϵ and σ^2_η will be involved. In Step 1, initial value for σ^2_ϵ and σ^2_η need to be provided. By the end of Step 2, the estimations of σ^2_ϵ and σ^2_η will be made and will be the new initial value for σ^2_ϵ and σ^2_η respectively in the next step. For details, see the simulation studies in next sections.

Simulation studies on this approach is presented below based on the basic Stochastic Volatility Model (SVM).

2.2. Stochastic volatility model (SVM)

For the simulation example, we consider the stochastic volatility process, which is often used for modelling log-returns of financial assets, defined by

$$y_t = \sigma_t \xi_t = e^{\alpha_t/2}\xi_t, t = 1, 2, \cdots, T, \tag{8}$$

and

$$\alpha_t = \gamma + \phi\alpha_{t-1} + \eta_t, t = 1, 2, \cdots, T, \tag{9}$$

where both ξ_t and η_t i.i.d respectively; η_t has mean 0 and variance σ_η^2. A key feature of the SVM in (8) is that it can be transformed into a linear model by taking the logarithm of the square of observations

$$\ln(y_t^2) = \alpha_t + \ln {\xi_t}^2, t = 1, 2, \cdots, T. \tag{10}$$

If ξ_t were standard normal, then $E(\ln \xi_t^2) = -1.2704$ and $Var(\ln \xi_t^2) = \pi^2/2$ (see Abramowitz and Stegun (1970), p943). Let $\boldsymbol{\varepsilon}_t = \ln \xi^2 + 1.2704$. The disturbance $\boldsymbol{\varepsilon}_t$ is defined so as to have zero mean. But If ξ_t were not standard normal, then $E(\ln \xi_t^2) = \mu$ and $Var(\ln \xi_t^2) = \sigma_\varepsilon^2$, So let $\boldsymbol{\varepsilon}_t = \ln \xi^2 - \mu$. Based on this situation, we consider the following martingale difference

$$\begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} = \begin{pmatrix} \ln(y_t^2) - \alpha_t - \mu \\ \alpha_t - \gamma - \phi\alpha_{t-1} \end{pmatrix}.$$

In Step 1, let α_t act as an unknown parameter. The standard quasi-score estimating function determined by the estimating function space

$$\mathcal{G} = \{A_t \begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \mid A_t \text{ is } \mathcal{F}_{t-1} \text{ measurable } \}$$

is

$$G_{(t)}(\alpha_t) = (-1, 1) \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix}^{-1} \begin{pmatrix} \ln(y_t^2) - \alpha_t - \mu \\ \alpha_t - \gamma - \phi\alpha_{t-1} \end{pmatrix}$$

$$= \sigma_\varepsilon^{-2}(\ln(y_t^2) - \alpha_t - \mu) + \sigma_\eta^{-2}(\alpha_t - \gamma - \phi\alpha_{t-1}). \tag{11}$$

Let $\hat{\alpha_0} = 0$ and initial values $\psi_0 = (\gamma_0, \phi_0, \sigma_{\eta_0}^2, \mu_0, \sigma_{\varepsilon_0}^2)$. Given $\hat{\alpha}_{t-1}$ the optimal estimation of α_{t-1}, the quasi-likelihood estimation of α_t, i.e. the optimal estimation of α_t, will be given by solving $G_{(t)}(\alpha_t) = 0$, i.e.

$$\hat{\alpha}_t = \frac{\sigma_{\eta_0}^2(\ln(y_t^2) - \mu) + \sigma_{\varepsilon_0}^2(\phi\hat{\alpha}_{t-1} + \gamma)}{\sigma_{\eta_0}^2 + \sigma_{\varepsilon_0}^2}, t = 1, 2, \cdots, T. \tag{12}$$

In Step 2, based on $\{\hat{\alpha_t}\}$ and $\{y_t\}$, let γ, μ and ϕ act as unknown parameters, and use the QL approach to estimate them. The standard quasi-score estimating function related to the estimating function space

$$\mathcal{G} = \{\sum_{t=1}^{T} A_t \begin{pmatrix} \epsilon_t \\ \eta_t \end{pmatrix} \mid A_t \text{ is } \mathcal{F}_{t-1} \text{ measurable } \}$$

is

$$G_T(\mu, \gamma, \phi) = \sum_{t=1}^{T} \begin{pmatrix} -1 & 0 \\ 0 & -1 \\ 0 & -\alpha_{t-1} \end{pmatrix} \begin{pmatrix} \sigma_{\varepsilon_0}^2 & 0 \\ 0 & \sigma_{\eta_0}^2 \end{pmatrix}^{-1} \begin{pmatrix} \ln(y_t^2) - \alpha_t - \mu \\ \alpha_t - \gamma - \phi\alpha_{t-1} \end{pmatrix}.$$

Replace α_t by $\hat{\alpha}_t$, $t = 1, 2, \cdots, T$, the QL estimate of μ, γ and ϕ will be given by solving $G_T(\mu, \gamma, \phi) = 0$. Therefore

$$\hat{\mu} = \frac{\sum_{t=1}^{T} \ln(y_t^2) - \sum_{t=1}^{T} \hat{\alpha}_t}{T}, t = 1, 2, \cdots, T. \qquad (13)$$

$$\hat{\phi} = \frac{\sum_{t=1}^{T} \hat{\alpha}_t \sum_{t=1}^{T} \hat{\alpha}_{t-1} - T \sum_{t=1}^{T} \hat{\alpha}_{t-1}\hat{\alpha}_t}{(\sum_{t=1}^{T} \hat{\alpha}_{t-1})^2 - T \sum_{t=1}^{T} \hat{\alpha}_{t-1}^2}, t = 1, 2, \cdots, T, \qquad (14)$$

$$\hat{\gamma} = \frac{\sum_{t=1}^{T} \hat{\alpha}_t - \hat{\phi} \sum_{t=1}^{T} \hat{\alpha}_{t-1}}{T}, t = 1, 2, \cdots, T. \qquad (15)$$

and let

$$\hat{\sigma_\eta^2} = \frac{\sum_{t=1}^{T} (\hat{\eta}_t - \bar{\hat{\eta}})^2}{T - 1} \qquad (16)$$

$$\hat{\sigma_\epsilon^2} = \frac{\sum_{t=1}^{T} (\hat{\epsilon}_t - \bar{\hat{\epsilon}})^2}{T - 1} \qquad (17)$$

where $\hat{\epsilon}_t = \ln(y_t^2) - \hat{\alpha}_t - \hat{\mu}$, and $\hat{\eta}_t = \hat{\alpha}_t - \hat{\gamma} - \hat{\phi}\hat{\alpha}_{t-1}, t = 1, 2, \cdots, T$. The above two steps will be iteratively repeated till certain criterion is meet. As mentioned before, $\hat{\psi} = (\hat{\mu}, \hat{\gamma}, \hat{\phi}, \hat{\sigma_\eta^2}, \hat{\sigma_\epsilon^2})$ will be used as an initial value for next step in the iterative procedure.

The final estimation results for SVM might be jointly affected by the initial values α_0 and ψ_0 which initially assigned to the underlying model during the inference procedure. For extensive discussion on a standard approach for assigning initial values in the Quasi-Likelihood (QL) estimation procedures see Alzghool and Lin (2011).

Table 1: QL estimates based on 1000 replication. Root mean square error of estimates are reported below each estimate.

	γ	ϕ	σ_η	μ	σ_ϵ	γ	ϕ	σ_η	μ	σ_ϵ
true	-0.821	0.90	0.675	-1.271	2.22	-0.411	0.95	0.484	-1.271	2.22
QL	-0.809	0.901	0.344	-1.366	2.15	-0.450	0.947	0.334	-1.048	2.094
	0.108	0.013	0.331	0.157	0.123	0.089	0.010	0.151	0.239	0.164
true	-0.736	0.90	0.363	-1.271	2.22	-0.368	0.95	0.260	-1.271	2.22
QL	-0.889	0.881	0.321	-1.199	2.02	-0.511	0.931	0.318	-1.185	2.01
	0.176	0.022	0.046	0.099	0.23	0.159	0.021	0.061	0.098	0.23
true	-0.706	0.90	0.135	-1.271	2.22	-0.353	0.95	0.096	-1.271	2.22
QL	-0.695	0.905	0.040	-1.043	2.21	-0.364	0.946	0.070	-1.660	2.17
	0.017	0.006	0.095	0.247	0.12	0.019	0.006	0.026	0.404	0.13
true	-0.147	0.98	0.166	-1.271	2.22	-0.141	0.98	0.061	-1.271	2.22
QL	-0.169	0.977	0.072	-1.327	2.23	-0.140	0.979	0.018	-1.705	2.22
	0.027	0.004	0.094	0.155	0.12	0.003	0.001	0.043	0.450	0.12

The format for this simulation study is the same as the layout considered by Rodriguez-Yam(2003). From empirical studies (e.g Harvey and Shepard, 1993; Jacquier et, al., (1994)) the values of ϕ between 0.9 and 0.98 are of primary interest. For this simulation study, we consider samples of size T=1000 and compute mean and root mean squared errors for $\hat{\phi}$, $\hat{\gamma}$, $\hat{\sigma_{\eta}^2}$, $\hat{\mu}$ and $\hat{\sigma_{\epsilon}^2}$ based on N=1000 independent samples. The results are shown in Table 1. QL denotes the quasi-likelihood estimate.

Table 2: QL estimates based on 1000 replication. Root mean square error of estimates are reported below each estimate.

		γ	ϕ	σ_η	μ	σ_ϵ
T=20	true	-0.141	0.98	0.061	-1.271	2.22
	QL	-0.1405	0.978	0.017	-2.428	2.16
		0.0041	0.005	0.044	1.258	0.567
T=50	QL	-0.1405	0.978	0.017	-2.179	2.19
		0.0036	0.002	0.044	0.961	0.367
T=100	QL	-0.141	0.979	0.018	-1.98	2.22
		0.0035	0.0015	0.043	0.750	0.251
T=200	QL	-0.1402	0.979	0.018	-1.809	2.22
		0.0034	0.0012	0.043	0.567	0.182
T=500	QL	-0.140	0.979	0.018	-1.705	2.22
		0.003	0.001	0.043	0.450	0.12

The effect of the sample size on the estimation of parameters is considered. Samples of sizes n = 20, 50, 100, 200, and 500 were generated. In Table 2, the simulation results also indicate that, the larger the sample size is, the smaller the root mean squared error will be.

2.3. Application to SVM

We apply the estimation procedure described in previous section to a real case where the observations are assumed to satisfy SVM (8) and (9)(see, Davis and Rodriguez-Yam (2005); Rodriguez-Yam (2003); Durbin and Koopman (2001)). The data are the pound/dollar of the daily observations of weekdays closing pound to dollar exchange rates $x_t, t = 1, \cdots, 945$ from 1/10/81 to 28/6/85 (http://staff.feweb.vu.nl/ koopman/sv/).

In literature, SVM (8) and (9) are used to model $y_t = log(x_t) - log(x_{t-1})$ where in the model, ξ has standard normal distribution. Set parameter $\psi = (\mu, \gamma, \phi, \sigma_\eta^2, \sigma_\epsilon^2)$.

Table 3: Estimation of γ, ϕ, σ_η, μ and σ_ϵ^2 for Pound/Dollar exchange rate data.

	γ	ϕ	σ_η	μ	σ_ϵ
QL	-0.0250	0.974	0.0210	-1.27	2.140
AL	-0.0227	0.957	0.0267		
MCL	-0.0227	0.975	0.0273		

Table 3 shows the estimates of ψ obtains by various method. QL denotes the estimate obtained by quasi-likelihood approach, AL the estimate obtained by maximizing the approximate likelihood proposed by Davis and Rodriguez-Yam (2003) and MCL estimate obtained by maximizing the estimate of the likelihood proposed by Durbin and Koopman

(1997). Note that AL and MCL outputs are taken from Rodriguez-Yam (2003). The QL estimations are slightly different from the estimation of AL and MCL.

3. Conclusion

This paper shows an alternative approach to estimate the parameters in SVMs. Instead of using traditional kalman filter formulae to estimate state variables, this approach use the QL method to estimate the state variables. It turns out the whole estimation processes looks very straightforward and is easily implemental. When the probability structure of underlying systems is complex or unknown, when maximum likelihood or mixture of maximum likelihood is not easily to implemented, the approach proposed in this paper is considerable approach for estimating parameters in SVMs.

REFERENCES

1. Abramovitz, M., Stegun, N. (1970). *Handbook of Mathematical Functions*, Dover Publication, New York.

2. Alzghool, R. and Lin, Y.-X. (2008). Parameters Estimation for SSMs: QL and AQL Approaches, *IAENG International Journal of Applied Mathematics*, **38**, pp. 34-43.

3. Alzghool, R. (2008). *Estimation for state space models: quasi-likelihood and asymptotic quasi-likelihood approaches*, PhD thesis, School of Mathematics and Applied Statistics, University of Wollongong, Australia.

4. Alzghool, R. and Lin, Y.-X. (2010). Estimation for State-Space Models: Quasi-likelihood, *Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X)*, **Volume I**, The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, pp. 409-423.

5. Alzghool and Lin (2011).Initial Values in Estimation Procedures for State Space Models (SSMs), *Proceedings of World Congress on Engineering 2011*, **Volume I**, WCE 2011, July 6-8, 2011, London, UK.

6. Breidt, F.J. and Carriquiry, A.L. (1996). Improved quasi-maximum likelihood estimation for stochastic volatility models. In: Zellner, A. and Lee, J.S. (Eds.), *Modelling and Prediction: Honouring Seymour Geisser*. Springer, New York, 228-247.

7. Davis, R. A. and Rodriguez-Yam, G. (2005). Estimation for State-Space Models: an approximate likelihood approach, *Statistica Sinica*, **15**, 381-406.

8. Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods.* Oxford, New York.

9. Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation, *Inter. Statist. Rev.*, **55**, 231-244.

10. Harvey, A. C. and Streible, M. (1998). Testing for a slowly changing level with special reference to stochastic volatility, *J. Econometrics*, **87**, 167-189.

11. Harvey, A. C. and Shepard, N. (1993). Estimation and testing of stochastic variance models. Unpublished manuscript, The London School of Economics.

12. Hedye, C. C. (1997). *Quasi-likelihood and its Application: a General Approach to Optimal Parameter Estimation*, Springer, New York.

13. Jacquire, E., Polson, N. G. and Rossi, P. E. (1994). Bayesian analysis of stochastic volatility models (with discussion), *J. Bus. Econom. Statist.*, **12**, 371-417.

14. Lin, Y.-X. (2007). An alternative derivation of the Kalman filter using the quasi-likelihood method, *J. of Statistical Planning and Inference*, **137**, 1627-1633.

15. Lin, Y.-X. and Heyde, C. C. (1993). Optimal estimating functions and Wedderburn's quasi-likelihood, *Comm. Statist.: Theory and Methods*, **22**, 2341-2350.

16. Papanastastiou, D. and Ioannides, D. (2004). The estimation of a state space model by estimating functions with an application, *Statistica Neerlandica* , **58**, No. 4, 407-427.

17. Pitt, M. K. and Shepard, N. (1999). Filtering via simulation: auxiliary particle filters, *J.Amer. Statist. Assoc.*, **94**, 590-599.

18. Rodriguez-Yam, G. (2003). Estimation for State-Space Models and Baysian regression analysis with parameter constraints, Ph.D. Thesis. Colorado State University.

19. Sandmann, G. and Koopman, S. J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood, *J. Econometrics*, **87**, 271-301.

20. Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika*, **61**, 439-447.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 553-558

MAPPING OF STOMACH CANCER INCIDENCE RATE IN IRAN FROM 2003 TO 2007 USING AREA-TO-AREA POISSON KRIGING

Naeimeh Sadat Asmarian[1], **Amir Kavousi**[2], **Masoud Salehi**[3]
and **Behzad Mahaki**[4*]

[1] Department of Biostatistics, School of Paramedicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

[2] Department of Sciences, School of Health, Safety and Environment, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

[3] Department of Statistics and Mathematics, School of Health Management and Information Sciences, Member of Health Management and Economics Research Center, Tehran University of Medical Sciences, Tehran, Iran.

[4] Department of Biostatistics, School of Public Health, Isfahan University of Medical Sciences, Isfahan, Iran. Email: behzad.mahaki@gmail.com

[*] Corresponding Author

ABSTRACT

Background and Goal: Stomach cancer has the highest prevalence of disease Gastrointestinal Cancer in Iran. Therefore, the aim of this study is mapping of county-level Stomach cancer incidence rate in Iran using Area-to-Area Poisson Kriging method and identify the high-risk areas.

Methods: This study is application and ecology. The methodology is illustrated using stomach cancer data recorded in Ministry of Health and Medical Education (in the non-infectious diseases management center) of Iran on years 2003-2007 to the 336 counties. Area-to-Area Poisson kriging method has been used to estimate the parameters of the map. The softwares SpaceStat and ArcGIS9.3 have been used for analysing the data and drawing maps.

Results: Incidence rate mean according to Area-to-Area Poisson Kriging method (5.19) has been estimated. Incidence rate variance mean using the Area-to-Area Poisson Kriging method (0.72) has been estimated. Maximum incidence rate using the Area-to-Area poisson kriging method (16.36) with variance (1.01) related to Divandareh county in north west of Iran and minimum incidence rate (0.12) with variance (0.67) related to Sarbaz county in south east of Iran have been estimated. Minimum variance incidence rate (0.01) related to Tehran county and Maximum variance incidence rate (2.51) related to Koohbanan county have been estimated.

Conclusion: The Area-to-Area poisson kriging method is recommended for estimation of disease mapping parameters since this method accounts spatial support and pattern in irregular spatial area. The results demonstrates that the counties in provinces Ardebil, Mazandaran and Kordestan have higher risk than other counties.

KEYWORDS

Disease Mapping, Area-to-Area Poisson kriging, Stomach cancer.

BACKGROUND

Cancer is the third cause of mortality after car accident and cardiovascular disease in Iran, so cancer is an important problem in public health in Iran. Stomach cancer has high incidence in north counties of Iran. The main aim of this study is the Stomach Cancer mapping for describe geographic of disease risk and identifying unusual high risk areas. The counties of Iran vary in size, shape and population size.

The area data used in this study is count data based on the Poisson distribution, So Area-to-Area Poisson kriging approach has been used for estimating the parameters of the map. Area-to-Area Poisson kriging approach is available to account for spatially varying population sizes and spatial patterns in the mapping of disease rates. This approach estimates disease risk more accurately and precisely. Area-to-Area Poisson kriging is a geostatistical techniques for the discrete distribution. Kaiser et al. (1997) introduced a spatial Poisson distribution. Oliver et al. (1998) employed binomial cokriging to produce a map of childhood cancer risk in the West Midlands of England. Monestiez et al. (2004; 2006) developed Poisson kriging to model spatially heterogeneous observation in the field of marine ecology. Goovaerts (2005) generalized Poisson kriging to analyse cancer data under the assumption that all geographic units are the same size, then Goovaert (2006) used Area-to-Area Poisson kriging technique for corporate the geometry of administrative units and the spatial repartition of the population at risk. This approach applied areal supports to predict area values by taking into account the spatial support of data as well as the varying population size, leading to more precise and accurate estimates of the risk.

Cancer data

The case of interest was Stomach cancer patients registered between the years 2003-2007 and adjusted using the 2006 population pyramid. Data recorded on incident cases of cancer were obtained from ministry of health and medical education (in the non-infectious diseases management center) of Iran. The major sources of data collection related to cancer were reports from pathology laboratories, hospitals and radiology clinics. The geographical units are 336 counties of display a wide range of size and shapes, which should favour Area-to-Area Poisson Kriging that implicitly accounts for the spatial support of the data in the analysis. The population-weighted average of Stomach cancer rate is 5.80 per 100,000 person- years.

METHOD

The Area-to-Area Poisson kriging technique for the estimation of risk values is described in detailed in Goovaerts (2006). This section provides a brief recall of the approach.

Let $u_\alpha = (x_\alpha, y_\alpha)$ represent area supports and $z(v_\alpha) = d(v_\alpha)/n(v_\alpha)$ denote the observed incidence rates where $d(v_\alpha)$ is the number of recorded incidence case and $n(v_\alpha)$ is the size of population at risk. Area-to-Area Poisson kriging spatial interpolation is predict any area value $z(v)$ using K area data neighboring units v_i :

$$z(v_\alpha) = \sum_{i=1}^{K} \lambda_i (v_\alpha) z(v_i) \tag{1}$$

where $D(v_\alpha)$ are weights assigned to rates are computed by solving the following equations:

$$\begin{gathered}\sum_{j=1}^{K} \lambda_j (v_\alpha) \left[\bar{C}_R (v_i, v_j) + \delta_{ij} \frac{m^*}{n(v_i)} \right] + \mu(v_\alpha) = \bar{C}_R (v_i, v_\alpha) \quad i = 1, \ldots, K \\ \sum_{j=1}^{K} \lambda_j (v_\alpha) = 1\end{gathered} \tag{2}$$

where $\mu(v_\alpha)$ is the Lagrange parameter, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. m^* Is the population-weighted mean, $\bar{C}(v_i, v_j)$ is the among-areas covariance and, $n(v_i)$ is the size of the population at risk in area v_i . The term $m^* / n(v_i)$ accounts for the variability resulting from the population size. The variance is calculated as:

$$\sigma^2 (v_\alpha) = \bar{C}_R (v_\alpha, v_\alpha) - \sum_{i=1}^{K} \lambda_i \bar{C}_R (v_i, v_\alpha) - \mu(v_\alpha) \tag{3}$$

where $\bar{C}(v_\alpha, v_\alpha)$ is the with-area covariance. The Area-to-Area Poisson kriging technique accomplished using the public-domain executable poisson-kriging.exe described in Goovaerts (2005)

RESULTS

According to Figure 1, crude rate and population at risk mapped in top graph, risk estimated using Area-to-Area Poisson kriging approach and corresponding prediction variance mapped in bottom graph. Crude rate mean is 4.88 and estimated risk mean is 5.19 and prediction variance mean is 0.82. North, north west and north east counties have higher risk than other counties.

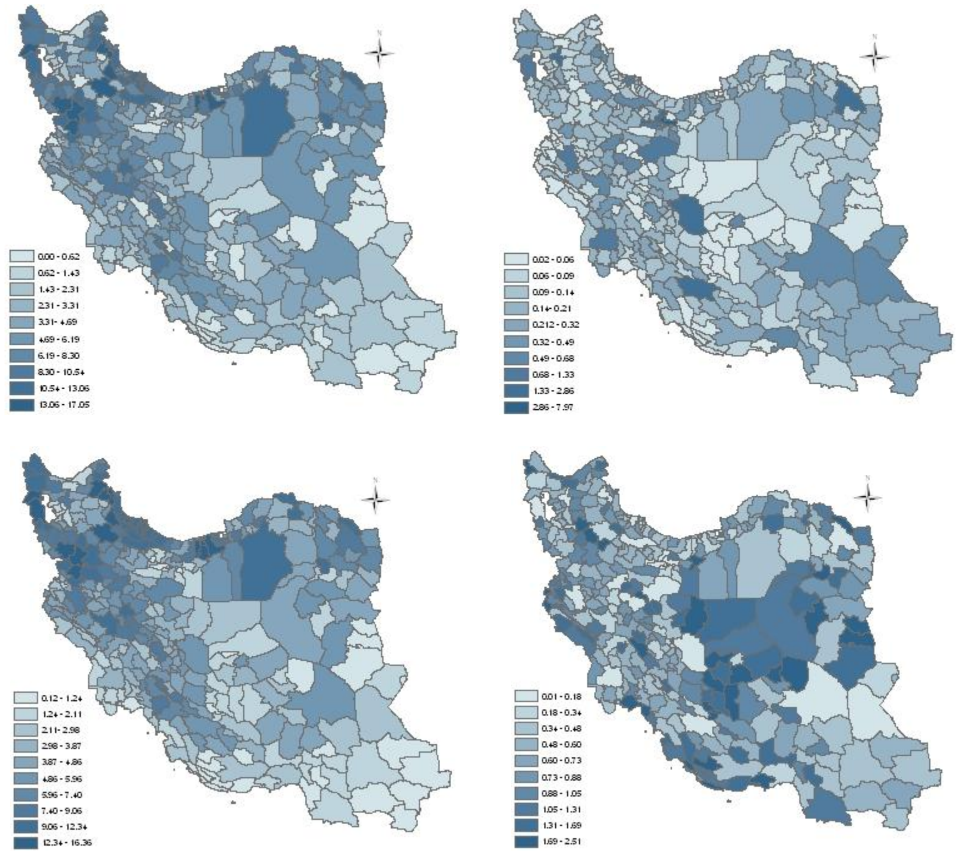


Fig. 1: Stomach Cancer incidence rate per 100,000 person years during the period 2003-2007 in Iran: crude rates (top, left), Population at risk (top, right), risk estimated (by ATA Poisson kriging) (botten, left), prediction variance (by ATA poisson kriging) (bottom, right).

CONCLUSIONS

Ignoring spatial support in parameter estimation may lead to more smoothing and less precise. The objective of this paper was to illustrate Stomach cancer incidence rate mapping by a geostatistic technique for health data to the popular Area-to-Area Poisson kriging. This approach account spatial support so generate less smoothing and more precise than other approaches such as BYM model and point Poisson kriging that ignore spatial support.

The Area-to-Area Poisson kriging approach is recommended for estimation of disease mapping parameters, since this method accounts spatial support and pattern in irregular spatial area. The results demonstrate that the counties in provinces Ardebil, Mazandaran and Gilan have higher risk than other counties.

In short, the risk of people developing Stomach cancer in Iran is heterogeneous during the period 2003-2007. People living north of Iran had a higher chance of cancer risk than people living in other areas. Epidemiologists and investigators believe that the cause of high incidence in north is nitrate including soil and particular nourishing in those areas.

REFERENCES

1. Goovaerts, P. (2005). Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics*.
2. Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *International Journal of Health Geographics*, 5, 52.
3. Kaiser, M.S. and N. Cressie (1997). Modeling Poisson variables with positive spatial dependence. *Statistical & Probability Letters*, 35(4), 423-432.
4. Kyriakids, P. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3).
5. Monestiez, P., L. Dubroca and E. Bonin (2004). Comparison of model based geostatistical methods in ecology: application to fin whales spatial distribution in northwestern Mediterranean Sea. *Geostatistics Banff*, 2, 777-786.
6. Monestiez, P., L. Dubroca and E. Bonin (2006). Geostatistical modelling of spatial distribution of Balaenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling*, 193, 615-628.
7. Iran Cancer Registry Report (2009). Provincial report, Islamic Republic of Iran-2007. Ministry of Health and Medical Education, Deputy for Health. Center for Diseases Control & Management. Tehran.
8. Iran Non-communicable Diseases Risk Factors Surveillance (2010). Provincial report, Islamic Republic of Iran (2007). Ministry of Health and Medical Education, Deputy for Health. Center for Diseases Control & Management, Tehran.
9. Oliver, M.A., R. Webster, C. Lajaunie and K.R. Muir (1998). Binomial cokriging for estimating and mapping the risk of childhood cancer, IMA. *Journal of Mathematics Applied in Medicine and Biology*, 15, 279-297.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 559-564

A BETTER APPROACH IN HYPOTHESIS TESTING

M. Shafiqur Rahman
Department of Operations Management and Business Statistics
College of Economics and Political Science, Sultan Qaboos University
Muscat, Sultanate of Oman. Email: srahman@squ.edu.om

ABSTRACT

This paper introduced a better approach in hypothesis testing. Should a research hypothesis or a claim be the null or alternative hypothesis? This paper proposed a simple way to select null and alternative hypotheses that can be applied to any practical situation. Considering two sets of hypotheses (a) H_0: $\theta \leq \theta_0$ against H_1: $\theta > \theta_0$ and (b) H_0: $\theta \geq \theta_0$ against H_1: $\theta < \theta_0$ for testing the significance of a population parameter θ if the absolute value of the test statistic obtained from a random sample is less than the absolute critical value then the two decisions corresponding to two sets of hypotheses (a) and (b) are contradictory. This paper proposed three regions hypothesis testing concept to overcome these issues and making consistent decisions.

KEY WORDS

Hypothesis testing, P-values, rejection region.

1. INTRODUCTION

Social science, Commerce and Business students usually have limited knowledge in Mathematics, especially in Algebra and Calculus. Mathematical Statistics basically deals with probability distributions of random variables, their properties, estimation and test of hypothesis regarding the parameters of probability distributions. In hypothesis testing the test statistic is selected in such a way that it follows some standard probability or sampling distribution when the null hypothesis is true. How can we select null and alternative hypotheses? Why a selected test statistic follows a certain probability or sampling distribution? How can we find the P-value? How can we make the decision? The derivation of the distribution of the test statistic requires extensive knowledge on Algebra and Calculus. It is difficult to explain these to students with limited knowledge in Mathematics or probability distribution theory. In order to give understanding of selecting null and alternative hypotheses, finding P-value and making decision in testing hypothesis to such type of students needs some simple way.

Selection of null and alternative hypotheses were discussed by many authors such as Anderson et al. (2011), Lehmann and Romano (2010), Shi and Tao (2008), Rao (1973), Bickel and Docksum (1977), Bain and Engelhardt (1992), McClave et al. (2005), etc. Anderson et al. (2011) proposed a general guideline for selecting null and alternative hypotheses. They considered a particular automobile model that currently attains an average fuel efficiency of 24 miles per gallon. They have mentioned that a product research group has developed a new fuel injection system specifically designed to increase the miles per gallon rating. In this case, the research hypothesis is that the new system will provide a mean miles-per-gallon rating exceeding 24 that is, $\mu > 24$. They have proposed as a general guideline, a research hypothesis should be stated as the alternative hypothesis. Hence, the appropriate null and alternative hypotheses for this study are H_0: $\mu \leq 24$ and H_a: $\mu > 24$.

As an illustration of testing the validity of a claim, Anderson et al. (2011) considered a manufacturer of soft drinks who states that two-liter containers of his product have an average of at least 67.6 fluid ounces. A sample of two-liter containers was selected, and the contents were measured to test the manufacturer's claim. In this type of situation, it is assumed that the manufacturer's claim is true unless the sample evidence proves otherwise. Using this approach for the soft-drink example, they stated the null and alternative hypotheses as H_0: $\mu \geq 67.6$ and H_a: $\mu < 67.6$. They said a manufacturer's claim is usually given the benefit of the doubt and stated as the null hypothesis. In any situation that involves testing the validity of a claim, the null hypothesis is generally based on the assumption that the claim is true.

Therefore, according to Anderson et al. (2011), a research hypothesis should be stated as the alternative hypothesis and in testing the validity of a claim, the null hypothesis is based on the claim. These two statements are contradictory. The hypotheses were selected correctly in the above two examples but generalization are **confusing** and **unacceptable**. This paper proposed some simple methods to select null and alternative hypotheses.

It is believed that people who live in northern Illinois are less likely to return sweepstakes entries than people who live in southern Illinois. Three hundred sweepstakes entries were sent to people in northern Illinois, and 300 entries were sent to people in southern Illinois. In response, 54% of the people in northern Illinois returned the entries, and 50% of the people in southern Illinois returned the entries. Let P_1 and P_2 be the proportions of peoples returning the sweepstakes entries for northern and southern Illinois respectively. Here we state the null and alternative hypotheses as H_0: $P_1 - P_2 \geq 0$ and H_a: $P_1 - P_2 < 0$. Then value of the test statistic is $Z = \dfrac{(\bar{P}_1 - \bar{P}_2) - (P_1 - P_2)}{\hat{\sigma}_{\bar{P}_1 - \bar{P}_2}}$, where

$$\hat{\sigma}_{\bar{P}_1 - \bar{P}_2} = \sqrt{\bar{P}(1 - \bar{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \text{ and } \bar{P} = \frac{n_1\bar{P}_1 + n_2\bar{P}_2}{n_1 + n_2},$$

= 0.98 and P-value = P(Z < 0.98) = 0.8365. For $\alpha = 0.05$, reject H_0 if P-value < 0.05 or, Z < -1.645. Therefore, do not reject H_0. That is people who live in Northern Illinois are not less likely to return sweepstakes entries than people who live in Southern Illinois. For the same problem if we are interested in testing the hypothesis that people who live in northern Illinois are more likely to return sweepstakes entries than people who live in southern Illinois. Then we would state the null and alternative hypotheses as H_0: $P_1 - P_2 \leq 0$ and H_a: $P_1 - P_2 > 0$. The value of the test statistic is Z = 0.98 and P-value = P(Z > 0.98) = 0.1635. For $\alpha = 0.05$, reject H_0 if P-value < 0.05 or, Z > 1.645. Therefore, do not reject H_0. That is people who live in Northern Illinois are not more likely to return sweepstakes entries than people who live in Southern Illinois. Therefore, using the same information we come up with two contradictory decisions. This paper proposed three regions hypothesis testing procedure to avoid these contradictory decisions.

2. PROPOSED METHODS

(a) Selecting null and alternative hypotheses:

In any research study, we usually interested in testing the validity of a claim or statement. In order to setup the null and alternative hypotheses, first we should identify the claim or statement that we are going to test in language or in Mathematical form. Then write down the alternative statement or claim in such a way that both statements/claims accommodate all the possibilities. Then the statement/claim that indicates or includes the **equality sign** will be the null hypothesis and other statement/ claim is the alternative hypothesis**.**

(b) Finding P-values:

An easy and straight forward way to find P-value is stated below:

(i) For one-tailed test P-value = $P\{T >$ (or < depending on H_a) T_{obs} }.

(ii) For two-tailed test P-value = $2P\{T > |T_{obs}|\}$.

(c) Three regions hypothesis testing procedures:

Let (x_1, x_2, x_3, …., x_n) be a random sample of size n drawn from a population the distribution of which has a known mathematical form say $f(x/\theta)$ but involves some unknown parameter θ. Let T be a function of x_1, x_2, x_3, …., x_n and is used as a test statistic to test (i) H_0: $\theta \le \theta_0$ and H_a: $\theta > \theta_0$ or, (ii) H_0: $\theta \ge \theta_0$ and H_a: $\theta < \theta_0$, or, (iii) H_0: $\theta = \theta_0$ and H_a: $\theta \ne \theta_0$. Here the parameter space can be divided into three regions $\Omega_L = \{\theta < \theta_0\}$, $\Omega_0 = \{\theta = \theta_0\}$, and $\Omega_U = \{\theta > \theta_0\}$. That is in (i) H_0 is the union of Ω_L and Ω_0, in (ii) H_0 is the union of Ω_0 and Ω_U, and in (iii) H_a is the union of Ω_L and Ω_U. These hypotheses can be tested in two ways: comparing the observed value of the test statistic T with the critical value or comparing P-value with the significance level α.

For testing (i), if the observed value of T (T_{obs}) obtained from the sample is less than $T_{1-\alpha}$ then conclude that $\theta < \theta_0$ and if $T_{1-\alpha} < T_{obs} < T_\alpha$ then conclude that $\theta = \theta_0$ and if $T_{obs} > T_\alpha$ then reject H_0 and conclude in favour of H_a that is $\theta > \theta_0$. Alternative way for testing (i): if P-value > 1-α then conclude that $\theta < \theta_0$ and if α < P-value < 1-α then conclude that $\theta = \theta_0$ and if P-value ≤ α then reject H_0 and conclude in favour of H_a that is $\theta > \theta_0$.

For testing (ii), if $T_{obs} < T_{1-\alpha}$ then reject H_0 and conclude in favour of H_a that is $\theta < \theta_0$, if $T_{1-\alpha} < T_{obs} < T_\alpha$ then conclude that $\theta = \theta_0$ and if $T_{obs} > T_\alpha$ then conclude that $\theta > \theta_0$. Alternative way for testing (ii): if P-value ≤ α then reject H_0 and conclude in favour of H_a that is $\theta < \theta_0$ and if α < P-value < 1-α then conclude that $\theta = \theta_0$ and if P-value > 1- α then conclude that $\theta > \theta_0$.

For testing (iii), if $T_{obs} < T_{1-\alpha/2}$ then conclude that $\theta < \theta_0$ and if $T_{1-\alpha/2} < T_{obs} < T_{\alpha/2}$ then conclude in favour of H_0 that is $\theta = \theta_0$ and if $T_{obs} > T_{\alpha/2}$ then conclude that $\theta > \theta_0$. Alternative way for testing (iii): if P-value ≤ α and the sign of T_{obs} is negative then conclude that $\theta < \theta_0$ and if P-value > α then conclude in favour of H_0 that is $\theta = \theta_0$ and if P-value $\leq\alpha$ and the sign of T_{obs} is positive then conclude $\theta > \theta_0$.

3. APPLICATION OF THE PROPOSED METHODS

(a) Selecting null and alternative hypotheses:

In Automobile model problem, the research group has developed a new fuel injection system specifically designed to increase the miles per gallon rating which is currently 24 miles per gallon. Here we are interested in testing the hypothesis that the new system will provide a mean miles-per-gallon rating exceeding 24 that is, $\mu > 24$. Other possibility is $\mu \leq 24$. Out of these two possibilities the one with equality sign is the null that is H_0: $\mu \leq 24$ and H_a: $\mu > 24$.

In soft drink containers problem, the manufacturer claimed that two-liter containers of his product has an average of at least 67.6 fluid ounces that is $\mu \geq 67.6$. Other possibility is $\mu < 67.6$. Out of these two possibilities the one with equality sign is the null that is H_0: $\mu \geq 67.6$ and H_a: $\mu < 67.6$.

(b) Finding P-values:

In the above example of returning sweepstakes entries for testing H_0: $P_1 - P_2 \geq 0$ against H_a: $P_1 - P_2 < 0$, the value of the test statistic $Z_{obs} = 0.98$.

P-value = P(Z < 0.98) = 0.8365.

In the same example of returning sweepstakes entries for testing H_0: $P_1 - P_2 \leq 0$ against H_a: $P_1 - P_2 > 0$, the value of the test statistic $Z_{obs} = 0.98$.

P-value = P(Z > 0.98) = 0.1635.

(c) Three regions hypothesis testing procedures:

In the above example of returning sweepstakes entries for testing H_0: $P_1 - P_2 \geq 0$ against H_a: $P_1 - P_2 < 0$, the value of the test statistic $Z_{obs} = 0.98$. At $\alpha = 0.05$ the critical value is -1.645 and the rejection rule is (a) reject H_0 if $Z_{obs} < -1.645$, (b) conclude $P_1 = P_2$ if $-1.645 < Z_{obs} < 1.645$ and (c) conclude $P_1 > P_2$ if $Z_{obs} > 1.645$. As $Z_{obs} = 0.98$, conclude $P_1 = P_2$. Alternative way is, as 0.05 < P-value = 0.8365 < 0.95, conclude $P_1 = P_2$.

On the other hand for testing H_0: $P_1 - P_2 \leq 0$ against H_a: $P_1 - P_2 > 0$, the value of the test statistic is $Z_{obs} = 0.98$. At $\alpha = 0.05$, the critical value is 1.645 and the rejection rule is (a) reject H_0 if $Z_{obs} > 1.645$, (b) conclude $P_1 = P_2$ if $-1.645 < Z_{obs} < 1.645$ and (c) conclude $P_1 < P_2$ if $Z_{obs} < -1.645$. As $Z_{obs} = 0.98$, conclude $P_1 = P_2$. Alternative way is, as 0.05 < P-value = 0.1635 < 0.95, conclude $P_1 = P_2$. Therefore, the decisions in both cases are same.

4. CONCLUSION

The proposed methods are simple, easy and straightforward.

REFERENCES

1. Anderson, D.R. Sweeny, D.J. and Williams, T.A. (2011). *Statistics for Business and Economics*, 11th Edition, Australia: South Western, Cengage Learning.
2. Bain, L.J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*, 2nd Edition, California: Duxbury Press.
3. Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics*, California: Holden-Day, Inc.
4. McClave, J.T. Benson, P.G. and Sincich, T. (2005). *Statistics for Business and Economics*, 9th Edition, New Jersey, Pearson Prentice Hall.
5. Lehmann, E.L. and Romano, J.P. (2010). *Testing Statistical Hypotheses*, Springer Texts in Statistics.
6. Rao, C.R. (1973). *Linear Statistical Inference and its application*, 2nd Edition, New York: John Wiley & Sons.
7. Shi, N. and Tao, J. (2008). *Statistical Hypothesis Testing: Theory and Methods*, World Scientific Publishing Company.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 565-572

ORGANIZATION LEARNING AS A MEDIATING MECHANISM BETWEEN TQM AND ORGANIZATIONAL PERFORMANCE: A REVIEW AND DIRECTIONS

Shahid Mehmood[1,2] and **Faisal Qadeer**[1]
[1] National College of Business Administration and Economics,
Lahore, Pakistan. Email: mfaisalqr@hotmail.com
[2] Department of Commerce, The Islamia University of Bahawalpur,
Bahawalpur, Pakistan. Email: shahidiub@hotmail.com

ABSTRACT

Despite wide acceptability of TQM as a process intervention to maintain organizational performance, the research on specifics mediating mechanism that translate TQM practices into performance still need to be progressed for coherence. This paper reviews the scientific literature on organization's learning as a mediator between TQM and organizational performance. The papers appearing after mid 1990s are selected through key word search, future citation of the most important papers and searching all important journals. Large number of research papers that investigates the impact of TQM on performance has identified organization learning as one of the most important mediating variables. This paper synthesis all such studies to examine why organizational learning is the most important mediator in this debate. Implications are discussed and suggestions for future research are provided.

KEY WORDS

Total quality management; organization learning and organizational performance.

INTRODUCTION

Total quality management (TQM) is an important philosophy and has gained a high degree of attention in improving organizational performance. For survival in a competitive environment TQM implementation is a goal of almost all organizations. To achieve this goal organizations are spending much more to implement TQM for continuous improvement and enhancing their performance but the empirical evidences on the relationship between TQM and organizational performance have been mixed. Some studies highlight the failure of TQM in enhancing performance. Dooyoung et al. (1998) reports an estimate of 60-67% failure rate of quality management. Fredrickson (1984) finds that in highly unsound product market broad decision making in total quality management negatively affects performance.

These mix findings and the need for an in-depth investigation of the relationship between TQM and organizational performance give us motivation to investigate learning as a mediating mechanism among TQM and organizational performance. In the section below we review the related literature and provide the research evidence on the importance of organizational learning as an important factor in association between TQM and organizational performance.

Most of the Scholars argue that adopting learning organization strategies should promote individual, team and organizational learning and that such improved learning should yield performance gains (Baker & Sinkula, 1999; Hunt & Morgan, 1994; Slater & Narver, 1995). To promote the learning culture, organizations should provide the ways that enable employees to contribute towards decision making and a change in implementation. This culture can be cultivated through the execution of TQM (Love et al. 2000). The major benefit to those organizations that have a capability to learn can enhance the performance of an organization, which predictably creates a sustainable competitive advantage for the firms (Brockmand & Morgan 2003; Fiol & Lyles 1985; Gnyawali et al. 1997).

Studies regularly find that the cultures that possess learning capability can improve individual, team, and organizational learning, and organizational performance (Egan et al., 2004; Kropp et al., 2006; Martinez-Costa & Jimenez Jimenez, 2008). Many scholars also found that the successful implementation of TQM produces effective learning that ensures a company's success (Barrow 1993; Denton 1998; Poole 2000). Irani et al. (2004) and McAdam and Harrison (1998) concluded that TQM stimulate the learning in an organization and when both are incorporated the organization can enhance the performance.

The existing research spread light on the importance of organizational learning in TQM performance relationship. Therefore, it may be suggested that TQM, organizational learning and performance are complementary and mutually dependent concepts.

METHODOLOGY

A Keyword search of published papers using terms '*Total quality management'* and organizational *performance* covering both theoretical and empirical studies excluded all papers published before 1990. The short listing of the papers is based on the relevancy to our aims of research and quality of journals. On the basis of above criteria we select 25 papers from different journals. The distribution of papers by journal and period is given in Table 1.

Table 1:
Distribution of Journal by Period

Name of Journal	**Years**		**Total**
	1990-00	**2001-2012**	
International journal of quality & reliability management	1	3	**04**
Journal of Operations Management		3	**03**
Academy of management Journal and proceedings	1	1	**02**
Management Sciences	2		**02**
Total quality management		2	**02**
International business review		1	**01**
Int. Journal of Production Research	1		**01**
Others (1 Per Journal)	2	8	**10**
Total	**7**	**18**	**25**

ANALYSIS AND FINDINGS

Findings from both theoretical and empirical papers are incorporated into this article. After comprehensive review of selected papers the summary of TQM and performance relationship is given in table 2.

Table 2:
TQM and Performance relationship

Paper	**Relationship**	**TQM and Performance type**
Lam et al. (2011) Terzivoski and Samson (1999) Fotopolus and Posmas (2010) Prajogo and Sohal (2003) Zehir et al. (2012) Douglas and Judge, (2001) Kanak, (2003) Corredor and Goni, (2011)	Positive & Significant	***Market*** performance ***Organizational*** performance, ***Innovation*** performance, ***Financial*** performance. TQM *Implementation* & performance Early implementation & performance
Macinati, M.S. (2008) Lemark et al. (1997) Hendricks & Singhal, (1996) Hendricks & Singhal, (1997)	Positive & Non-significant	***Financial*** performance TQM implementation & performance Quality award , financial performance TQM implementation & ***operational*** performance
Powell, (1995) Westphal et al. (1996)	Non-significant	TQM and performance

The summary of these findings describe that the relationship between TQM and performance is yet not clear and no linear relationship exists. On the basis of these findings we propose the mediating mechanism that may strengthen the relationship between TQM and performance.

After complete assessment of existing literature we have identified different factors as mediators among TQM and organizational performance as given in Table 3.

Table 3:
Mediators among TQM and Performance

Paper	**Mediator**	**Performance type**
Hung et al. (2011)	*Organizational learning*	Innovation performance
Sadikoglu and Zehir, (2010)	*Employee performance* *Innovation performance*	Innovation performance & Firm performance
Wang et al. (2012)	*Market orientation*	Organizational performance

On the basis of the findings of Table 3 while making comparison of organizational learning with other mediators it is identified that the learning is strongest mediator in TQM performance relationship because both TQM and learning focus on continuous improvement and competitive advantage. TQM is all about business management values consisting of different principles that help in continuous improvement. (Lin & Ogunyemi, 1996). TQM when successfully implemented helps in gaining sustainable competitive advantage (Prajogo & Sohal, 2004). In order to achieve continuous improvement firms must encourage learning to enhance knowledge that can be utilized in the future (Baker & Sinkula, 1999). Chang and Sun (2007) explores the correspondence between TQM and organization learning and identifies with the help of correspondence and cluster analysis that there exists close and significant correspondence among TQM and organization learning. Barrow (1993) argues that TQM closely relate to organizational learning as an anticipated product of TQM. Hill (1996) observes that scholars consider it the first step to continuous improvement. Popper and Lipshitz (2000) propose that productive learning can occurs in an organization where TQM culture is prevalent. Martinez-Costa and Jimenez-Jimenez (2008) finds that in Spanish firms the structure of TQM positively relate with the firms' organizational learning development. Lam et al., (2011) find that the Malaysian service firms practicing TQM also have learning orientation.

Very few studies attempt to explain organizational performance through a joint mechanism of TQM and organizational learning capability. One of such rare investigations, Martinez-Costa and Jimenez-Jimenez (2009) finds that in Spanish SMEs TQM, organizational learning and performance are connected. Similarly, Hung et al. (2011) recently conclude that in the high-tech Taiwanese firms TQM has positive associations with organizational learning. They also find that TQM as well as organizational learning have positive influences on the innovation performance.

Therefore, organizational learning not only promoting innovation performance of a firm, but it also acts as a mediating factor between TQM and innovation performance.

Therefore, it is believed that Organization learning is a strongest mediator between TQM and organizational performance. Some scholars put an emphasis on organization learning as a mediating mechanism between TQM and organizational performance. Sinkula (1994) and Slater and Narver (1995) ---"have proposed learning orientation as a mediator in the TQM–performance linkage". Lam et al (2011) in their future directions proposed to investigate the mediating effect of learning orientation in TQM-performance relationship.

CONCLUSION

Organizations need learning along with TQM to improve their performance in competitive environment. This study contributes to TQM literature by identifying different factors as mediating variables and the significance of organizational learning to strengthen the insignificant relationship between TQM and organizational performance. The study proposes that the influence of TQM on organizational performance is contingent with organizational learning.

FUTURE RESEARCH AND IMPLICATIONS

This is an initial investigation; further empirical study should be conducted to test the organizational learning as a mediating mechanism among TQM and organizational performance in a specific sector like manufacturing sector. This study also provides implications for practitioners and academicians. The managers of service and manufacturing sectors who intend to achieve higher organizational performance through the implementation of TQM must focus on organizational learning as a supporting factor to achieve the desired results. The academicians should form the strategies to maximize the influence of learning on implementation of TQM for enhancing performance of organizations.

REFERENCES

1. Baker, W.E. and Sinkula, J.M. (1999). The Synergistic Effect of Market Orientation and Learning Orientation on Organizational Performance. *Journal of the Academy of Marketing Science*, 27(4), 411-27.
2. Barrow, J.W. (1993). Does total quality management equal organizational learning? *Quality Progress*, 26(7), 39-43.
3. Brockmand, B. and Morgan, F. (2003). The role of existing knowledge in new product innovativeness and performance. *Decision Sciences*, 32(2), 385-419.
4. Chang, D.S. and Sun, K.L. (2007). Exploring the correspondence between total quality management and Peter Senge's disciplines of a learning organization: A Taiwan perspective. *Total Quality Management & Business Excellence*, 18(7), 807-822.
5. Corredor, P. and Goni, S. (2011). TQM and performance: Is the relationship so obvious? *Journal of Business Research*, 64(8), 830-838.
6. Denton, J. (1998). *Organizational learning and effectiveness*. London: Routledge.

7. Dooyoung, S., Kalinowski, J.G. and El-Enein, G. (1998). Critical implementation issues in total quality management. *SAM Advanced Management Journal,* 63(1), 10-14.
8. Douglas, T.J. and Judge, Jr., W.Q. (2001). Total quality management implementation and competitive advantage: the role of structural control and exploration. *Academy of Management Journal*, 44(1), 158-169.
9. Egan, T.M., Yang, B. and Bartlett, K. (2004). The effects of learning culture and job satisfaction on motivation to transfer learning and intention to turnover. *Human Resource Development Quarterly,* 15(3), 279-301.
10. Fiol, C.M. and Lyles, M.A. (1985). Organizational learning. *Academy of Management Review*, 10(4), 803-813.
11. Fotopoulos, C.V. and Psomas, E.L. (2010). The structural relationships between TQM factors and organizational performance. *The TQM Journal,* 22(5), 539-552.
12. Fredrickson, J.W. (1984). The comprehensiveness of strategic decision processes: extension, observation, future directions. *Academy of Management Journal,* 27(3), 445-466.
13. Gnyawali, D.R., Steward, A.C. and Grant, J.H. (1997). Creation and utilization of organizational knowledge: an empirical study of the roles of organizational learning on strategic decision making. *Academy of Management Proceedings*, 16-20.
14. Hendricks, K.B. and Singhal, V.R. (1996). Quality awards and the market value of the firm: an empirical investigation. *Management Sciences*, 42(3), 415-36.
15. Hendricks, K.B. and Singhal, V.R. (1997). Does implementing an effective TQM program actually improve operating performance? Empirical evidence from firms that have won quality awards. *Management Science,* 43(9), 1258-1274.
16. Hill, F.M. (1996).Organizational learning for total quality management through quality circles. *TQM Magazine*, 8(6), 53-57.
17. Hung, R.Y.Y., Lien, B.Y.H., Yang, B., Wu, C.H. and Kuo, Y.M. (2011). Impact of TQM and organizational learning on innovation performance in the high-tech industry. *International Business Review,* 20(2), 213-225.
18. Hunt, S.D. and Morgan, R.M. (1994). Organizational Commitment: One of Many Commitments or Key Mediating Construct? *Academy of Management Journal*, 37(6), 1568- 1587.
19. Irani, Z., Beskese, A. and Love, P. (2004). Total quality management and corporate culture: Constructs of organizational excellence. *Technovation*, 24(8), 643-650.
20. Kaynak, H. (2003). The relationship between total quality management practices and their effects on firm performance. *Journal of Operations Management*, 21(4), 405-435.
21. Kropp, F., Lindsay, N.J. and Shoham, A. (2006). Entrepreneurial, market and learning orientations and international entrepreneurial business venture performance in South African firms. *International Marketing Review*, 23(5), 504-523.
22. Lam, S.Y., Lee, V.H., Ooi, K.B. and Lin, B. (2011). The relationship between TQM, learning orientation and market performance in service organizations: an empirical analysis. *Total Quality Management*, 22(12), 1277-1297.
23. Lemak, D.J., Reed, R. and Satish, P.K. (1997). Commitment to total quality management: Is there a relationship with firm performance? *Journal of Quality Management,* 2(1), 67-86.

24. Lin, B. and Ogunyemi, F. (1996). Implications of total quality management in Federal services: the US experience. *International Journal of Public Sector Management*, 9(4), 4-11.
25. Love, P.E.D., Li, H., Irani, Z. and Faniran, O. (2000). Total quality management and the learning organization: a dialogue for change in construction. *Construction Management and Economics*, 18(3), 321-331.
26. Macinati, M.S. (2008). The relationship between quality management systems and organizational performance in the Italian National Health Service. *Health Policy*, 85(2), 228-241.
27. Martinez-Costa, M. and Jimenez-Jimenez, D., (2008). Are companies that implement TQM better learning organization? An empirical study. *Total Quality Management and Business Excellence,* 19(11), 1101-1115.
28. Martinez-Costa, M. and Jimenez-Jimenez, D. (2009). The effectiveness of TQM: the key role of organizational learning in small businesses. *International Small Business Journal*, 27(1), 98-125.
29. McAdam, R., Leitch, C. and Harrison, R. (1998). The links between organizational learning and total quality: A critical review. *Journal of European Industrial Training*, 22(2), 47-56.
30. Poole, S.W. (2000). The learning organization: motivating employees by integrating TQM philosophy in a supportive organizational culture. *Leadership and Organization Development Journal*, 21(8), 373-378.
31. Popper, M. and Lipshitz, R. (2000). Organizational learning: Mechanisms, culture and feasibility. *Management Learning*, 31(2), 181-196.
32. Prajogo, D.I. and Sohal, A.S. (2003). The relationship between TQM practices, quality performance and innovation performance. *The International Journal of Quality & Reliability Management,* 20(8), 901-918.
33. Prajogo, D.I., Power, D.J. and Sohal, A.S. (2004). The role of trading partner relationships in determining innovation performance: An empirical examination. *European Journal of Innovation Management*, 7(3), 178-186.
34. Powell, T. (1995). Total quality management as competitive advantage: A review and empirical study. *Strategic Management Journal*, 16 (1), 15 -37.
35. Sadıkoglu, E. Zehir, C. (2010). Investigating the effects of innovation and employee performance on the relationship between total quality management practices and firm performance: An empirical study of Turkish firms. *International Journal of Production Economics*, 127(1), 13-26.
36. Sinkula, J.M. (1994). Market information processing and organizational learning. *Journal of Marketing*, 58(1), 35-45.
37. Slater, S.F. and Narver, J.C. (1995). Market orientation and the learning organization. *Journal of Marketing,* 59(3), 63-74.
38. Terziovski, M. and Samson, D. (1999). The link between total quality management practice and organisational performance. *International Journal of Quality & Reliability Management*, 16(3), 226-237.
39. Wang, C.H., Chen, K.Y. and Chen, S.C. (2012). Total quality management, market orientation and hotel performance: The moderating effects of external environmental factors. *International Journal of Hospitality Management*, 31(1), 119-129.

40. Westphal, J.D., Gulati, R. and Shortell, S.M. (1996). The institutionalization of total quality management: the emergence of normative TQM adoption and the consequences for organizational legitimacy and performance. *Academy of Management Proceedings*, 249-253.
41. Zehir, C., Ertosun, O.G., Zehir, S. and Muceldilli, B. (2012). Total Quality Management Practices’ Effects on Quality Performance and Innovative Performance. *Procedia-Social and Behavioral Sciences*, 41, 273-280.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 573-584

COOPERATION BETWEEN NSOs AND ACADEMIC INSTITUTES TO BUILD CAPACITY IN OFFICIAL STATISTICS-QATAR PERSPECTIVE

Wadha Al-Jabor and **Pinar Ucar**
Qatar Statistics Authority, Economic Statistics and National Accounts Department
Doha Towers, P.O. Box 7283, Doha, Qatar
Email: waljabor@qsa.gov.qa; pucar@qsa.gov.qa

ABSTRACT

Official statistics is a cornerstone of good government and public confidence in government. Objective, reliable and accessible official statistics provide citizens and organizations, nationally and internationally, with confidence in the integrity of government and public decision-making on economic, social and environmental matters in each country.

National Statistics Offices (NSOs) are mandated to collect, analyze and disseminate data-official statistics to keep the public well-informed and meet policy makers requirements.

Academic institutes are responsible for supporting human capital by deepening expertise and knowledge, strengthening skills and contributing to the development of society through research activities.

Although statistics program graduates are the main recruitment source for NSOs, interaction between National Statistics Offices and academic institutes (particularly statistics programs) is not strong. Only few academic institutes have training programs which are directly applicable to official statistics environment. NSOs can cooperate with academic institutes in developing programs to enhance the skills for its required human capital/workforce.

The purpose of this paper is to review cooperation between NSOs and academic institutes both internationally and in Qatar.

KEYWORDS

Official Statistics, National Statistics Offices, Capacity Building, Statistical Literacy, Qatar Statistics Authority

1. INTRODUCTION

Human capital plays a crucial role in any organization’s growth. Valuing human capital development particularly in a National Statistics Office, will help directly increase the quality of collecting, processing, producing and disseminating official statistics. High quality products serve for a well-informed society and can better meet the expectations of policy makers.

Hence, the National Statistics Office should deploy efforts in developing programs to enhance technical, administrative, IT and soft skills for its required human capital/workforce.

In this section, focus will be given to three categories of workforce: Future Workforce, Fresh Graduates-New Recruits and Current Workforce (UNECE, 2011).

1.1. Future Workforce: It is very important to constantly focus on youth training in order to create a statistically literate generation. A generation familiar with official statistical concepts would be able to understand and appreciate the complexity of producing official statistics. This will also assist in motivating and preparing them for a career in government statistics.

1.2. Fresh Graduates–New recruits: Statistical Offices recruit staff generally from other government departments, research organizations and mostly from academic institutions. It is almost impossible for an academic institution to develop the appropriate skills in a graduate that a statistical organization desires, as there is a difference in preparing a student for a role in government or 'official' statistics, as opposed to 'general' statistics. The below table compares official statistics and "other" statistics that are often learned through the education system.

Table 1
Differences between official and other types of statistics

OFFICIAL STATISTICS	"OTHER" STATISTICS AND RESEARCH
Multi-purpose (collect once-use often)	Single Focus (on research or policy question)
Participation often mandatory (high response rates)	Voluntary participation (lower response rates-potential bias)
Often based on complex sample designs	Often designed experiments
Broad coverage (many variables-often high-level measures)	In-depth studies
Large- scale (provide comparisons between groups)	Usually relatively small scale (experiments or surveys)
Usually repeated regularly (provide long time series)	Mainly cross-sectional (single point of time)
Internationally comparable (agreed standards and classifications)	Relevant to population studied (focused on research or policy question)
Analysis provided by collectors usually simple (single variable or between two variables)	Sophisticated analysis (multivariate analysis methods used)
Provide primary data source	Can involve secondary analysis (of other data sources)
High cost	Generally involve lower costs

Source: Forbes, 2008

Statistics offices should develop programs that make the transition easier for fresh graduates from an academic statistics environment (new recruits from general statistics environment) to the official statistics environment. It can be carried out in three steps:

Step 1: Assign a mentor to provide guidance, share their statistical knowledge and offer advice and assistance when needed.

Step 2: Provide structured capacity development programs that helps them acquire the skills and experience within the statistics office. The best programs are based on action learning.

Step 3: Offer a career path and learning plan: Every new recruit has their own unique mix of professional, career and personal goals and priorities. Recruits should be aware of the career path and be motivated to work towards higher position within this career path.

1.3. Current Workforce: Strengthen the current employee’s skills through continuous learning programs. National Statistics Offices should provide training programs for current employees to upgrade their skills and to implement new practices, standards and processes and to gain a deeper understanding of the core business of statistics. Therefore, a blend of formal and informal learning activities should be offered to enhance employee competencies. They may include:

Formal training such as structured in-class courses either provided by trainers at the statistics office or in partnership with academic institutions, computer assisted courses or e-learning, external learning through university diploma or certificate programs some of which can be financially sponsored or provided as education leave.

Informal training such as networking opportunities, conferences in specific subject areas, information sessions, presentations, seminars and workshops.

2. LEARNING AND DEVELOPMENT STRATEGY IN QATAR STATISTICS AUTHORITY

Established in 2007, Qatar Statistics Authority (QSA) is the official source of statistical data and information replacing the Planning Council’s Statistics Department. QSA is responsible for the production, analysis and dissemination of official, demographic, social, economic, environmental, and other statistics.

Qatar’s economy, driven by huge revenues from natural gas and oil resources, has experienced a tremendous growth during 2004-2011 with GDP growing in real terms from USD 31.7 billion in 2004 to USD 88.2 billion in 2011 (preliminary estimates).

Qatar has formulated “Qatar National Vision 2030 (QNV 2030)” to manage its wealth in a sustainable manner. The 2030 vision rests on four pillars: Human development, Social development, Economic development and Environmental development. The human development pillar aims at developing a well-educated, motivated and capable workforce by 2030 (GSDP, 2008).

QSA prepared National Strategy for the Development of Statistics (NSDS) in line with Qatar’s national development policy and goals, to strengthen statistical capacity. Enhancing the development and potential of individual employees through human resource management strategies including training and career development is one NSDS objectives (QSA, 2008).

In line with Qatar National Vision, QSA developed Learning and Development Strategy to support business objectives of the organization. QSA seeks to create a learning environment to:

- Optimize the uniqueness of people, through learning and to achieve the desired business outcomes;
- Maintain and improve customer service;
- Ensure that all individuals have the opportunity for self-development;
- Ensure that employees are equipped to meet current and future competency requirements as specified by their job description;
- Achieve the succession planning goals of QSA;
- Support the achievement of QSA's employment and Legislative Qatarization goals;
- Develop Qatar workforce skills; and
- Contribute to Qatar's economic development

QSA delivers various types of programs: On-Line training, particularly E-Learning, Facilitator/Instructor-led learning, including facilitation by an external or internal provider, On-the-job-training OJT, conferences, seminars, and workshops (QSA, 2010).

3. COOPERATION BETWEEN QSA AND ACADEMIC INSTITUTES IN QATAR

To fulfill its objectives QSA has taken some initiatives to develop relations and boost cooperation with academic institutions in Qatar.

3.1. Cooperation with Qatar University (QU) Statistics program

Qatar University-established in 1973- is the only government university in the country, comprises of seven colleges which offer over 60 specializations: Arts and Sciences; Business and Economics; Education; Engineering; Law; Pharmacy; and Sharia and Islamic Studies. QU is also a leading center for research in the country. Statistics is a program under Stat, Math, Physics department at the College of Arts and Sciences. The program offers a major in statistics which concentrates on applied statistics along with some introductory probability and mathematical statistics courses. The program also offers a minor in statistics that is open for all students from other departments within the faculty or even from other faculties.

Supporting BSc. program Accreditation:

QSA has been supporting the program through the phase of improvement that the statistics program undertook while seeking accreditation from the Royal Statistical Society (RSS). The process started in March 2008 when a team from the Royal Statistical Society visited the Department and met with the program members. The program has been accredited in 2011 (QU, n.d.).

Supporting MSc. Program:

QSA researchers and interested individuals submitted request for a post graduate degree in applied statistics to be available for more than 200 statistics degree holders at QU-expected future statisticians in QSA and other agencies.

Supporting the Re-opening of Male Section:

QU approved the re-opening of the statistics program for male students which was halted for almost 3 years starting in 2009 due to reluctance of males to enroll in that program.

Statistics Poster Contest:

In 2011, the contest was launched jointly between QSA and the statistics program in QU. The aim was to increase statistical awareness of students, its relevance to scientific research, and to contribute in the development of statistical reasoning among future statisticians (QSA, 2011).

One -month Training for Students:

QSA offers one-month training course for students from Mathematics, Statistics and Physics Departments of Qatar University's College of Arts and Sciences. This course is open for any students during the year in conjunction with their academic breaks and can be as long or as short as they wish. During summer of 2012, four students trained by employees from Censuses and Household Surveys, Population and Social Statistics, and Economic Statistics and National Accounts Departments. The course's main aim was to inform students of QSA's methods in providing statistical data and information.

Students had the opportunity to gain better understanding of how Censuses and Household Surveys Department's prepare and implement Expenditure Household Survey, design the questionnaire, use PDA and census data, and census, statistical manuals and coding processes. Moreover, Population and Social Statistics Department introduced to the students a number of tasks carried out on regular basis, such as defining and calculating labor force and demographic indicators, analyzing population table, drawing population pyramid, data dissemination method, and the way of extracting the table from QALM website as well as saving it. Furthermore, students got familiar with the methods developed by the Economic Statistics and National Accounts Department to prepare the Gross Domestic Product (GDP); Consumer Prices Index (CPI); imports, exports and re-exports statistics; short-term economic indicators.

At the end of the training, students gave a presentation describing their experience during the period as well as recommendations to improve the training process for students. They were granted practical training certificates during a ceremony held by QSA (QSA, 2012).

Undergraduate Research Experience Program (UREP)-2012:

UREP is one the four funding programs operated by Qatar National Research Fund. It aims to promote “Learning by Doing” and "Hands-On" mentorship activities as effective methods for undergraduate education. In addition to a research-based education, students will gain experience with team-based research collaboration with faculty, postdoctoral fellows, graduate students, and other undergraduates or research staff in Qatar (QNRF, n.d.)

One of QSA staff from Economic Statistics Department is participating in a three-year UREP research project led by Qatar University Statistics Program.

3.2 Cooperation with Qatar Foundation

Qatar Foundation for Education, Science and Community Development (QF), established in 1995, aims to develop people's abilities through investments in human capital, innovative technology, state-of-the-art facilities and partnerships with elite organizations, thus raising youth competencies and the quality of life.

QF Education City offers branch campuses of eight international universities: Texas A&M University at Qatar, Weill Cornell Medical College in Qatar, Georgetown University, School of Foreign Service in Qatar, Virginia Commonwealth University School of the Arts in Qatar, Carnegie Mellon University in Qatar, Northwestern University in Qatar, HEC Paris in Qatar, University College London in Qatar, Qatar Faculty of Islamic Studies.

QF is also building a research base from both academic and applied research so that universities and businesses can collaborate on translating ideas into commercial products and services (QF, n.d.).

QSA signed Memorandum of Understanding (MoU) with Carnegie Mellon University in Qatar and Qatar Computing Research Institute.

MoU with Carnegie Mellon University in Qatar:

Carnegie Mellon University in Qatar (CMU-Q), established in 2004, offers undergraduate programs in Business Administration, Computer Science and Information Systems. CMU-Q also has a research agenda covering areas as diverse as social sciences, computer architecture, robotics and computer security.

QSA and CMU-Q signed in 2012 a MoU pertaining to executive education and training, and cooperation in scientific research between both parties. The MoU encourages students to take on volunteer opportunities in QSA’s activities and programs related to community service. As per the agreement both organizations will collaborate to hold joint conferences, seminars, meetings, workshops to share expertise and best practices in official statistics, business administration, computer science, and Information Technology.

The MoU reflects strong commitment of both parties to promoting knowledge in education, scientific research, executive education and community development.

QSA is willing to cooperate with CMU-Q in the field of scholarships and vocational training for Qatari students encouraging them to volunteer in QSA’s activities. The MoU is part of QSA’s efforts aim to develop methods for collecting and analyzing statistical data and information, and benefit from educational programs and strategic research, and reflects CMU-Q endeavors to enhance scientific research and strategic studies to meet the local market needs (QSA, 2012).

MoU with Qatar Computing Research Institute (QCRI):

QCRI is one of the three research institutes under Qatar Research Institute (QRI). QCRI aims at developing knowledge locally and support innovation aligned with national priorities by conducting cutting-edge, multidisciplinary applied computing research in coordination with Qatari institutions, corporations, and government. The Data Analytics group at QCRI tackles diverse data management problems that address Qatar’s growing needs and investments in a knowledge-based economy. Data analytics use highly sophisticated mathematical and statistical tools to examine raw data and consequently draw conclusions about the information buried in this data. Data analytics is used by government and industry to shape policies and improve decision making.

QSA and QCRI have signed a MoU to collaborate on research activities in the area of data analytics to enhance the efficiency of QSA operations and improve its data quality (QSA, 2011).

4. COOPERATION BETWEEN NSOs AND ACADEMIC INSTITUTES-EXAMPLES FROM WORLD-

Five national statistics offices are being used as an example of cooperation between NSOs and academic institutes: United Kingdom, New Zealand, Ireland, Slovenia, and Africa.

4.1. United Kingdom, Office for National Statistics (ONS)

MSc and Certificate in Official Statistics:

The University of Southampton and the Office for National Statistics have jointly structured a professional development training program targeting statisticians in the Government Statistical Service or equivalent organizations conducting large-scale statistical work. Through part-time study, the program enables statisticians and researchers to strengthen and update their professional skills and knowledge. The program is unique in the sense that it focuses on courses in both survey methodology and data analysis, addressed by leading experts in these fields.

The program currently comprises of 23 short course units in Survey Methods (including questionnaire design, survey quality, evaluation and monitoring), Sampling and Estimation (including weighting methods, variance estimation, non-response analysis and adjustments), Data Analysis (covering regression modeling, multilevel modeling, longitudinal data analysis, general linear models) and other courses related to Official Statistics (Time Series, Index Numbers, National Accounts, Small Area Estimation, Statistical Disclosure Control, Statistical Computing). Upon successful completion of 16 instructional units, students earn a diploma in Official Statistics. Starting 2012, an award of a Certificate in Official Statistics is available for those who cannot commit to the full training program but have successfully completed 8 instructional units. Diploma holders can undertake a supervised dissertation to earn a degree of MSc in Official Statistics. In addition, the short course program is ideal for professional development ‘one-off’ courses, with or without assessment, for refreshing statistical skills in a particular area (ONS, n.d.).

4.2 Statistics New Zealand

Statistics New Zealand has developed a strategy for raising statistical capacity. The strategy has three distinct parts: Enhancing staff skills within Statistics New Zealand (initially called The Power of Numbers); improving other agencies' capacities, particularly users and producers of official statistics, in the public sector (Beyond the Numbers) and up-skilling the public via communities of interest such as small businesses and schools (Understanding the Numbers). In this paper we will focus on Beyond the Numbers part.

Beyond the Numbers aims to raise capability in the government agencies (including Statistics New Zealand) and to make policy advisors aware of the power of statistics to provide an evidence base for decision making. There are currently three components (Campos et al., 2008):

Certificate in Official Statistics:

The Certificate of Official Statistics is a vocational (pre-university) level certificate that provides a recognized qualification in official statistics. The certificate comprises of four compulsory statistics units and some optional general units (project management, further statistics, research report, etc.). A group of academics from statistics departments in different universities agreed to work collectively to deliver the four below compulsory units:

a. Interpret statistical information to form conclusions for projects in a public sector context
b. Evaluate and use statistical information to make policy recommendations in a public sector context
c. Assess a sample survey and evaluate inferences in a public sector context
d. Resolve ethical and legal issues in the collection and use of data in a public sector context

This qualification is assessed on a competency basis with students able to take parts or whole units until they meet required standards. The Auckland University of Technology was contracted to formally assess the units. It is expected that students will take a year to complete the qualification. Candidates are mainly from agencies other than Statistics New Zealand, such as NZ Police and the Department of Labour.

Official Statistics System Seminar and Training Series:

The Official Statistics System Seminar Series (OSS Seminar Series) was formalized in 2006 with the creation of the Statistical Education and Research team. The Series is a monthly forum open to producers and users of official statistics with, in general, one speaker presenting on a particular topic. The training series is a one-day workshop with, to date, instruction from university statistics lecturers.

Adjunct Professor of Official Statistics:

The establishment of a half-time Adjunct Professor of Official Statistics (mainly funded by Statistics New Zealand was negotiated with Victoria University of Wellington). The professor has a co-ordination and leadership role providing tertiary training and research in official statistics, aimed at state sector employees and users of official statistics. That is to:

- Promote official statistics as a career choice
- Increase use of official statistics in academic and student research
- Improve use of official statistics in the state sector
- Enhance statistical capability of state sector.

4.3 Ireland the Central Statistics Office (CSO), University Collage Dublin (UCD)

Course on Official Statistics:

The Department of Statistics in University College Dublin offered in its curriculum a course on official statistics. It was first delivered in the academic year 1999-2000 to a class of final year undergraduate; graduate diploma and masters level students. The goals of this course were to teach students how official statistics pervade citizens' daily life; to describe the processes of data collection, analysis and dissemination in a National Statistics Office; and to introduce some techniques used extensively in NSOs that are not taught in other courses in the Statistics Department. The course consisted of six sections (Murphy, 2002):

Section 1-History: This section describes the history of official statistics with particular emphasis on Ireland. Some topics covered include the first censuses, and the history of international organisations such as the IMF and Eurostat. For international comparison purposes this reading concentrates especially on the development of official statistics in the UK.

Section 2-Legal and Institutional: This section begins by looking at the regulatory framework governing the collection of official statistics. In Ireland statistics are collected under the Statistics Act. The impact of International Organisations on the collection of official statistics is presented. The course then describes the operation of the Central Statistics Office in Ireland as an example of a National Statistics Office.

Section 3- The Statistics: This section examines different areas of official statistics such as: National Accounts, Balance of Payments, External Trade, Demography, Agriculture, Building, Business Register and Data Bank, Industry, Labour Market, Prices, Retail Sales, Services, Transport, Tourism and Vital Statistics. The course has been kept updated and relevant.

Section 4- Index Numbers: Some history of cost of living indices is presented. The Laspeyres, Paasche and Fisher index numbers are introduced. The calculation of a Consumer Price Index is discussed including how a fixed basket of goods is determined using a household budget survey. Purchasing Power Parities are mentioned.

Section 5-Databases: Over four lectures and six hours in the computer laboratory students are introduced to the basic ideas of a Relational Database Management System. They are shown how to design a database including several linked tables, how to design forms for data input and how to create SQL queries to extract information from the database.

Section 6- Additional Topics: Guest lecturers who are involved in the collection of official statistics or whose work makes significant use of official statistics are invited to speak to the students.

The Professional Diploma in Official Statistics for Policy Evaluation:

The Institute of Public Administration (IPA) is a recognized college of University College Dublin (UCD). The program has been developed in conjunction with the CSO. It introduces students to important Irish and international official statistics that would help them better understand the key statistics in relation to current economic and social developments in Ireland.

Course material is being prepared and taught by professional statisticians from CSO. This is a practical “hands-on” course, and emphasis is being placed on the visualization and presentation of statistics so that useful policy relevant information or knowledge can be derived. Students are also being introduced to data management and metadata best practice and to the broader principles of evidence-informed policy formulation and evaluation.

This program takes one academic year and is delivered via distance learning. Students receive comprehensive course material prepared by the CSO and the IPA. The distance learning is supported by attendance at scheduled seminars and workshops held at the IPA over the course of the academic year, where, in addition to receiving lectures and demonstrations (IPA, n.d.).

4.4 Slovenia

Online course:

A Course on European Economic Statistics (CEES) was developed at the Faculty of Economics in Ljubljana/Slovenia (FE) with the help of the consortium of partners: Faculty of Electrical Engineering from Ljubljana/Slovenia, Faculty of Economics and Business Administration from Sofia/Bulgaria (FEBA) and the Training of European Statisticians (TES) Institute from Luxembourg. The main objective of the CEES project was the development of an original course module on official statistics covering the field of economics for non-statisticians at the higher education level, taking into account recent developments of Eurostat statistics and deploying ICT (Information and Telecommunication Technology) in order to improve the quality of the learning process and increase users’ access and understanding of official statistics.

The CEES course aimed to provide learners with appropriately structured information on availability and quality of the Eurostat’s and national official statistical data. It also aimed to provide students with insight into harmonisation of national statistical practices with EU regulations. Finally, learners should be taught how to explore and appropriately use databases available on Internet (Bregar&Ograjensek, 1999).

4.5 Africa Central Statistics Office

It was a proposal to establish a partnership called SPAPGA (Statistics Partnership among Academia, Private Sector and Government in Africa) in order to advance collaboration between the private sector, government, and academia to improve statistical training and capacity building in Africa. It aimed to foster collaborations between partners through facilitating regular annual meetings and seminars; and conducting discussions on curricula development.

SPAPGA proposed two seperate programs to train statisticians (Thabane et al., 2008):

1. Training of future statisticians at the pre-employment stage (university or collage training)
   - inclass instruction: statistical concepts
   - experiential learning, coop, internship, apprenticeship
   - non-statisical skills: time/stress management, team dynamics
   - interdepartmental seminars
   - student-centered research meetings, seminars
   - student-centered pre-reviewed publications
   - mentorship
2. Training of statisticians for career development (on the job training)
   - mentorship
   - study groups
   - reflective logs
   - action research
   - peer coaching
   - in-service programs (e.g. discipline programs)
   - affiliation with statistical societies
   - inter-departmental/agency seminars
   - inter-country exchange programs

5. COMMENTS AND CONCLUSION

In a fast changing social and economic environment, demands from citizens and policy makers for timely, accurate and realiable statistics are increasing constantly. Highly qualified, well-trained workforce is essential for NSOs to meet this demand and produce better qualified statistics. NSOs should have sufficient capacity in place to respond to new and unexpected challenges and demands.

Today, a major challange for NSOs is recruiting and retaining the staff. It has become imperative to strengthen cooperation between NSOs and academic institutions in order to attract the youth to its labor force, as well as motivate and train the current workforce.

This paper highlights the cooperation between NSOs and academic institutions to build capacity in official statistics. Initiatives such as postgraduate programs, professional diploma programs, training courses, online courses, seminar series developed by two parties are introduced. Examples from UK, New Zealand, Slovenia, Ireland and Africa are given.

Current collaboration between QSA and Qatari academic institutions is MoUs. QSA may review the examples of other countries and in collobaration with academic institutes, may develop its own program that is tailor to specific needs of Qatar.

REFERENCES

1. Bregar, L. and Ograjenšek, I. (1999). Impact of Internet on Official Statistics: Users' Opportunities. Paper presented at *52nd session of International Statistical Institute*, Helsinki.
2. Campos, P., Forbes, S., Giacche, P., Helenius, R., Sanchez, J., Taylor, P. and Townsend, M. (2008). Government Statistical Offices and Statistical Literacy, Chapter 3: Raising Statistical Capacity. Publication of International Statistical Literacy Project of the International Statistical Institute.
3. GSDP (General Secretariat of Development Planning). (2008). Qatar National Vision. Doha.
4. GSDP (General Secretariat of Development Planning). (2012). Expanding the Capacities of Qatari Youth. Doha.
5. IPA (Institute of public Administration), Ireland (n.d.) Professional Diploma in Official Statistics for Policy Evaluation. Retrieved from http://www.ipa.ie/index.php?lang=en&p=page&id=363. Accessed November 2012.
6. Murphy, P. (2002). Teaching Official Statistics in an Irish University Statistics Department. Paper presented at ICOTS-6, The Sixth International Conference on Teaching Statistics, Cape Town.
7. Office for National Statistics, UK. (n.d.) MSc in Official Statistics. Retrieved from http://www.ons.gov.uk/ons/about-ons/what-we-do/training/courses/msc-in-official-statistics/index.html.Accessed November 2012.
8. QF (Qatar Foundation). (n.d.). Discover Qatar Foundation. Retrieved from http://www.qf.org.qa/discover-qf. Accessed November 2012.
9. QFRF (Qatar Foundation Research Fund). (n.d.) Undergraduate Research Experience Program (UREP). Retrieved from http://www.qnrf.org/funding_programs/urep/index.php. Accessed November 2012.
10. QSA (Qatar Statistics Authority). (2008). National Strategy for Development of Statistics. Doha.
11. QSA (Qatar Statistics Authority). (2010). Learning and Development Strategy (draft document).
12. QSA (Qatar Statistics Authority). (2012, June 19). MoU Between QSA and CMU in Qatar Signed. Retrieved from http://www.qsa.gov.qa/eng/News/2012/Article/31.htm
13. QSA (Qatar Statistics Authority). (2012, July 23). QSA Trains QU Students. Retrieved from http://www.qsa.gov.qa/eng/News/2012/Article/40.htm
14. QSA (Qatar Statistics Authority). (2011, October 5). QCRI and Qatar Statistics Authority to Collaborate on Data Analytics Research. Retrieved from http://www.qsa.gov.qa/eng/News/2011/artical/26.htm
15. QSA (Qatar Statistics Authority). (2011, December 15). QSA Encourages Future Statisticians. Retrieved from http://www.qsa.gov.qa/eng/News/2011/artical/43.htm
16. QU (Qatar University). (n.d.). Our History. Accessed November 2012. Retrieved from http://www.qu.edu.qa./theuniversity/history.php
17. QU (Qatar University) Statistics Program. (n.d.) About the Program. Retrieved from http://www.qu.edu.qa/artssciences/mathphysta/stats/index.php. Accessed November 2012.
18. Thabane, L., Chinganya, O. and Ye, C. (2008). Training Young Statisticians for the Development of Statistics in Africa. The African Statistical Journal, Volume 7.
19. UNECE (United Nations Economic Commission for Europe). (2011). Making Data Meaningful Part 4, Chapter 8: Improving Statistical Literacy within Statistical Organisations - Training the Workforce, Geneva.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 585-592

AGE ESTIMATION OF SCHOOL GOING CHILDREN OF KARACHI, LARKANA AND QUETTA BY NUMBER OF ERUPTED TEETH USING MEDIAN REGRESSION

Nazeer Khan and **Sundus Iftikhar**
[1] Dow University of Health Sciences, Karachi, Pakistan
Email: n.khan@duhs.edu.pk
[2] Karachi University, Karachi, Pakistan
Email: sundusiftikhar@hotmail.com

ABSTRACT

Introduction: Dental age is very much important in the areas where birth records are not properly maintained as it helps presuming age of not only alive but also children at death (Khan N; 2010). There have been several methods available to estimate dental age 1) Number of teeth at the time of emergence of permanent teeth 2) regression analysis 3) probit analysis 4) Demirjian methods. Median regression is not being used yet anywhere in the world to estimate age using number of erupted teeth. In this paper, age will be estimated through regression analysis and median regression. **Objective:** To estimate age of school going children of Pakistan by number of erupted teeth using median regression and to compare its results with simple linear regression analysis. **Methodology:** This study is a part of a larger national clinical survey for “Time of eruption of permanent teeth”. A sample of 4400 students from 102 schools of Karachi, 1200 students from 28 schools of Larkana and 1267 from 25 schools of Quetta were collected using systematic random sampling. Schools were randomly selected from the list of schools, using systematic random sampling procedure. A team of dentists (1 male and 1 female) and assistants (1 male and 1 female) visited each school to collect the data. Among those students if a child has have just erupted tooth, the child was taken away from the class and the questionnaire was completed for the selected child. The criterion of just erupted teeth was defined as: a tooth deemed to have emerged if any part of it was visible in the mouth. **Quantile Regression:** Quantile regression is an extension of ordinary least squares Ordinary least square regression examines location of response variable only; on the other hand quantile regression gives a complete picture of location, shape and scale of a distribution of response variable. **Results:** Three methods were used to estimate the age: method 1: Total teeth erupted with square and cube terms; Method 2: Total teeth with height and weight; Method 3: Quantile regression of total teeth erupted with square and cube terms. Method 1 was found to be best to estimate the calendar age of Quetta children, also independent variables height and weight in Method 2 was statistically insignificant in the equation but including these variables increase the value of R^2 that means, contribution of these variables in the equation was significant. For median regression, frequency of residuals computed from both mean and median of the data. It showed that the calendar age of Pakistani children can be estimated within ±1 year with 100% of accuracy by all the three methods. For Karachi, the calendar age of the children can be estimated within ± 0.5 year with 89.3% accuracy by method 1 and 2, 78.6% vs.

82.1% accuracy by method 3 (error from mean vs. error from median). For Larkana, the calendar age of the children can be estimated within ± 0.5 year with 92.9% and 85.7% accuracy by method 1 and 2, 89.3% vs. 67.9% accuracy by method 3 (error from mean vs. error from median). For Quetta, the calendar age of the children can be estimated within ± 0.5 year with 100% accuracy by method 1, method 2 and method 3 (error from mean), 75% accuracy with method 3 (error from median). From the above results it can be concluded that median regression- Method 3 is not a suitable statistical method to estimate the calendar age of Pakistani children. **Conclusion:** Median regression does not show any superiority than simple regression both methods (1 &2). It could be due to the type of data, we have used for this exercise. It could give better result if different types of data are used to compare this method with simple regression methods.

INTRODUCTION

Dental age is very much important in the areas where birth records are not properly maintained as it helps presuming age of not only alive but also children at death (Khan N; 2010). From the point of justice and legislations, age assessment is often required (Singh K). In forensic science, estimation of age is of very much importance for identification purposes of deceased person (Willems G; 2001). Knowledge of age is also noteworthy in school attendance, marriage, employment, community health project and many social benefits (Khan N; 2010, Willems G; 2001). Estimation of age with teeth is one of the acceptable methods of age determination not only because of low variations in dental indicators (Singh K, Willems G; 2001) but also due to a typical chronological pattern in teeth development and evolution of teeth shows aging changes (Kim Y U; 2000).

There have been several methods available to estimate dental age 1) Number of teeth at the time of emergence of permanent teeth 2) regression analysis 3) probit analysis 4) Demirjian methods (Khan N; 2010 and Gillet RM; 1997). The Demirjian method has few limitations (Khan N; 2010, Foti B; 2003, Fruchi S; 2000, Davis PJ; 1994) and is not applicable on Pakistani Population (Firdos T; 2009). Willem G; 2001 reported that atlas approach and scoring system may be the two techniques used to estimate age in children based on dental maturation. Median regression is not being used yet anywhere in the world to estimate age using number of erupted teeth.

In this paper, age will be estimated through regression analysis and median regression; their accuracy will be assessed by comparing with original age recorded while collecting the data.

OBJECTIVE

To estimate age of school going children of Pakistan by number of teeth eruptedusing median regression and to compare its results with simple linear regression analysis.

METHODOLOGY

This study is a part of a larger national clinical survey for “Time of eruption of permanent teeth”. A sample of 4400 students from 102 schools of all the towns of Karachi, 1200 students from Larkana and 2680 students from Quetta was collected using systematic random sampling. The data was entered in SPSS software and analyzed using

SAS 9.2 software. Self administered questionnaire was used to obtain information from the children who had ‘just erupted teeth’ about age, gender, height and weight along with other required dental information. The age range was 4 to 15 years. Permission to conduct this survey in the respective schools was obtained from Principal.

Procedure:

The sampling design and methodology for Karachi is defined in detailed below and the same procedure is applied to collect data in the other four cities.

There are 18 administrative towns in Karachi containing 3948 public and 2560 private registered schools. Public schools further registered as different schools separately for both boys and girls as primary, secondary and high schools. Most of the time these three types of schools are situated in one building, so if we take in a high school then primary and secondary schools attached with it includes automatically in the sample. For this reason, we decided to pick up only high schools from public school list for sampling. But this system is opposite for private schools, one registered private school not only covers all primary, secondary and high school sections for both gender but also cover up its various campuses located at different areas. In city government schooling system, there are 529 and 2560 public and private registered schools. Moreover, private school’s attendance ratio is pretty better than public school. Thus, it was decided to divide the number of cases into the ratio of 3:1 in private and public schools. From literature (Khan N; 2008, Khan N.B.; 2006, Chohan A.N.; 2007) it is found that about 30% of the total children have at least one tooth just erupted. Keeping this percentage in mind, it was planned that at least 4000 cases should be collected for the research, the distribution of sample size was determined on this basis. Ten percent more is added into this sum to make sure that minimum number of cases is obtained. Using OpenEpi software the sample size was found to be 13900 with maximum error 1% and confidence interval 99% but it was planned to examine 25,000 children to obtain 4400 cases of just erupted teeth (3400 from private and 1000 from public schools). Assuming that each school enrolls 250 students on average, we were needed to visit 100 schools. Schools were randomly selected from the list of schools, using systematic random sampling procedure and the number of schools was doubled in case of non-cooperation from the administration of the school.

A team of dentists (1 male and 1 female) and assistants (1 male and 1 female) were hired and make clear the objectives and methodology in detail. To obtain permission from the administration of the school, letters were posted at the addresses mentioned in the list. Very poor response rate was obtained for this mailing sampling procedure, either due to ‘no response’ or wrong addresses, so the investigators visited the selected school to explain the purpose of the project and got the permission. A planned calendar was geared up for the investigators after arranging time and date with the administrators. All the present students were examined for the general check up and clinical form was duly filled for dental caries and oral hygiene status. Among those students if a child has have just erupted tooth, then the child was taken away from the class and the questionnaire was completed for the selected child. A child whose birth date is unknown or not a citizen of Pakistan or has no ‘just erupted’ tooth is excluded from the study.

Definition:

The criterion of just erupted teeth was defined as: a tooth deemed to have emerged if any part of it was visible in the mouth·

QUANTILE REGRESSION

Quantile regression is an extension of ordinary least squares(Chen C;2005, Koenker R; 2001).Ordinary least square regression examines location of response variable only; on the other hand quantile regression gives a complete picture of location, shape and scale of a distribution of response variable (Koenker R; 2001, Hao L; 2007).Ordinary least square can only provide information whether specific covariate is important or not but cannot address the influence of that covariate on particular percentile (or quantile) of the response variable. Quantile regression is a very helpful approach in the fields like environmental studies, public health, economic status, ecology, survival analysis, social sciences etc where extreme values and inequalities are of more importance (Chen C; 2005, Koenker R; 2001, Hao L; 2007).

RESULTS

A total of 8280 children were examined and only 6423 students were found suitable for the study: 4093 from Karachi, 1135 from Larkana and 1195 from Quetta. Four models (2 of simple linear regression and 2 of Median regression) were employed to estimate the age. The regression equations (Table 1) were as follows.

MODEL # 1 (Simple regression):

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where, $\hat{Y}$ =age

X_1=Total teeth erupted; X_2=Total teeth erupted-Square; X_1=Total teeth erupted-cube

MODEL # 2 (Simple regression):

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where, $\hat{Y}$ =age

X_1=Total teeth erupted; X_2=Height; X_1=Weight

MODEL # 3 (Median regression):

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where, $\hat{Y}$ =age

X_1=Total teeth erupted; X_2=Total teeth erupted-Square; X_3=Total teeth erupted-cube

MODEL # 4 (Median regression):

$$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where, $\hat{Y}$ =age

X_1=Total teeth erupted; X_2=Total teeth erupted-Square;
X_3=Total teeth erupted-Cube; X_4=Height; X_5=Weight

Table 2 shows the frequency of residual computed from the data using these four models of estimation for Karachi, Larkana and Quetta. It showed that the calendar age of Karachi students can be estimated within ± 0.5 years with 33.7%, 35.5%, 33.7% and 35.8% with these 4 models respectively, the calendar age of Larkana students can be estimated within ± 0.5 years with 7.2%, 7.2%, 7.1% and 7.5% with these 4 models respectively and the calendar age of Quetta students can be estimated within ± 1 years with 13.5%, 13.3%, 14.2% and 14.1% with these 4 models respectively.

Overall the calendar age of Pakistani students can be estimated within ± 1 years with 63%, 64%, 63% and 65% with these four models respectively. Furthermore, the calendar age of Pakistani students can be estimated within ± 0.5 years with 35%, 36%, 36% and 37% with these four models respectively.

Table 1:
Regression coefficient of explanatory variables

KARACHI							
	Intercept	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**R^2**
Model 1	5.74077**	0.19151**	0. 00672*	-0.00020674**	-		0.6748
Model 2	-0.1976	0.15768**	0.05768**	-0.01299*	-		0.7033
Model 3	5.6364**	0.2061**	0.0043**	-0.0001**	-		-
Model 4	0.2192	0.0665*	0.0090*	-0.0002*	0.0550**	-0.0100	-
LARKANA							
Model 1	5.57301**	0.06501	0.01595*	-0.00041397**			0.5368
Model 2	1.56073*	0.16111**	0.02897**	0.03259**			0.5653
Model 3	6.0627**	-0.0893	0.0272*	-0.0006**			-
Model 4	2.2978*	-0.0314	0.0183*	-0.0004*	0.0262**	0.0314*	-
QUETTA							
Model 1	5.04342**	0.16955*	0.01380**	-0.00034236**			0.8457
Model 2	4.23932**	0.26845**	0.00187	0.03500**			0.8516
Model 3	4.7739**	0.2167**	0.0097**	-0.0002*			-
Model 4	4.3671**	0.1691*	0.0101*	-0.0003*	0.0025	0.0287*	-

** $P<0.0001$ & * $P<0.05$

Table 2:
Error in estimation by simple regression and median regression

Karachi								
Error	**M1**		**M2**		**M3**		**M4**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
<=-2	218	5.2	171	4.1	204	4.9	162	3.9
>-2 to -1.5	278	6.6	242	5.8	223	5.3	225	5.4
>-1.5 to -1	326	7.8	396	9.4	434	10.4	353	8.4
>-1 to -0.5	710	16.9	589	14.0	535	12.8	600	14.3
>-0.5 to 0.00	665	15.9	768	18.3	842	20.1	767	18.3
>0.00 to 0.5	746	17.8	721	17.2	572	13.6	733	17.5
>0.5 to 1	457	10.9	580	13.8	619	14.8	593	14.1
>1.00 to 1.5	366	8.7	319	7.6	289	6.9	330	7.9
>1.5 to 2.00	162	3.9	189	4.5	222	5.3	197	4.7
> 2.00	265	6.3	218	5.2	253	6.0	233	5.6
Total	4193	100.0	4193	100.0	4193	100.0	4193	100.0
Within ±0.5	1411	33.7	1489	35.5	1414	33.7	1500	35.8
Within ±1	2578	61.5	2658	63.4	2568	61.2	2693	64.2
Larkana								
Error	**M1**		**M2**		**M3**		**M4**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
<=-2	78	7.4	73	6.9	86	8.1	75	7.1
>-2 to -1.5	76	7.2	66	6.2	80	7.6	56	5.3
>-1.5 to -1	91	8.6	112	10.6	107	10.1	106	10.0
>-1 to -0.5	142	13.4	143	13.5	129	12.2	129	12.2
>-0.5 to 0.00	146	13.8	150	14.2	162	15.3	167	15.8
>0.00 to 0.5	154	14.5	150	14.2	137	12.9	146	13.8
>0.5 to 1	110	10.4	113	10.7	126	11.9	125	11.8
>1.00 to 1.5	109	10.3	109	10.3	92	8.7	106	10.0
>1.5 to 2.00	68	6.4	65	6.1	59	5.6	65	6.1
> 2.00	85	8.0	78	7.4	81	7.6	84	7.9
Total	1059	100.0	1059	100.0	1059	100.0	1059	100.0
Within ±0.5	300	7.2	300	7.2	299	7.1	313	7.5
Within ±1	552	13.2	556	13.3	554	13.2	567	13.5

Table 2 (continued)

Quetta								
Error	**M1**		**M2**		**M3**		**M4**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**
<=-2	6	.5	5	.4	6	.5	4	.3
>-2 to -1.5	27	2.3	23	1.9	11	.9	13	1.1
>-1.5 to -1	87	7.3	89	7.4	96	8.0	68	5.7
>-1 to -0.5	227	19.0	235	19.7	146	12.2	202	16.9
>-0.5 to 0.00	321	26.9	319	26.7	396	33.1	314	26.3
>0.00 to 0.5	247	20.7	239	20.0	198	16.6	276	23.1
>0.5 to 1	111	9.3	124	10.4	169	14.1	142	11.9
>1.00 to 1.5	84	7.0	70	5.9	61	5.1	78	6.5
>1.5 to 2.00	36	3.0	45	3.8	61	5.1	50	4.2
> 2.00	49	4.1	46	3.8	51	4.3	48	4.0
Total	1195	100.0	1195	100.0	1195	100.0	1195	100.0
Within ±0.5	568	13.5	558	13.3	594	14.2	590	14.1
Within ±1	906	21.6	917	21.9	909	21.7	934	22.3

Table 3:
Number of Permanent Teeth

Age	**Karachi**		**Larkana**		**Quetta**	
	N	**# of teeth Min-max**	**N**	**# of teeth Min-max**	**N**	**# of teeth Min-max**
4-5	96	1-16	61	1-16	21	1-4
6-7	959	1-20	323	1-20	223	1-14
8-9	1199	2-26	1518	3-26	342	5-23
10-11	1210	8-28	1478	8-28	318	9-28
12-13	520	10-28	808	10-28	239	10-28
14-15	109	12-28	14	12-28	52	16-28

CONCLUSION

Median Regression (Model 4) showed better age estimation procedure than the other three procedures.

REFERENCES

1. Khan, N. and Chohan AN. (2010). Age estimation of female school children by number of permanent teeth erupted; a study from Riyadh, Saudi Arabia. *J. Pak. Dent. Assoc.*; 19(3), 180-183.
2. Singh, K. (2005). *Age Estimation from Eruption of Temporary and Permanent Teeth from 6 Months To 25 Years*. A thesis for MD (forensic medicine) govt. Medical College, Patiala.

3. Willems, G. (2001). A review of the most commonly used dental age estimation techniques. *J. Forensic Odontostomatol* 19, 9-17.
4. Kim, Y.U., Kho, H.S. and Lee, K.H. (2000). Age Estimation by Occlusal Tooth Wear. *J. Forensic Sci*; 45(2), 303-309.
5. Gillet, R.M. (1997). Dental Emergence among Urban Zambian School Children: An Assessment of the Accuracy of three methods in Assigning Ages. *American Journal of Physics Anthropology*, 102, 447-54.
6. Firdos, T., Bashir, M.Z. and Aziz, K. (2009). Dental maturation in Peshawar: Applicability of Demirjian's standards. *Pakistan Journal of Medical & Health Sciences*.
7. Foti, B., Lal, L., Adalian, P., Giustimian, J., Maczel, M., Signoli, M., Dutour, O. and Leonetti, G. (2003). New Forensic Approach to Age Determination in Children Based on Teeth Eruption. *Forensic Science International*, 132, 49-56.
8. Fruchi, S., Schnegelsberg, C., Schulte-Monting, J., Rose, E. and Jones, I. (2000). Dental Age in SouthWest Germany; a Radiographic Study. *Journal of Orofacial Orthopedics*, 61, 318-29.
9. Davis, P.J. and Hagg, U. (1994). The Accuracy and Precision of the Demirjian System when Used for Age Determination in Chinese Children. *Swedish Dental Journal*, 18, 113-6.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 593-600

COMMON FUNCTIONAL PRINCIPAL COMPONENT (CFPC) MODELS FOR COHERENT MORTALITY FORECASTING

Farah Yasmeen[1], **Rob J. Hyndman**[2] and **Sidra Zaheer**[1]
[1] Department of Statistics, University of Karachi, Karachi, Pakistan.
Email: riazfarah@yahoo.com; sidraz.ku@gmail.com
[2] Monash University, Australia.

ABSTRACT

The functional time series (FTS) models are used for analyzing, modeling and forecasting age-specific mortality rates. However, the application of these models in presence of two or more groups within similar populations needs some modification. In these cases, it is desirable for the disaggregated forecasts to be *coherent* with the overall forecast.

The *'coherent'* forecasts are the non-divergent forecasts of sub-groups within a population. In this paper, we relate some of the functional models to the common principal components (CPC) and partial common principal components (PCPC) models introduced by Flury 1988 and provide the methods to estimate these models. We call them common functional principal component (CFPC) models and use them for coherent mortality forecasting. Here, we propose a sequential procedure to estimate the model parameters.

KEY WORDS

Mortality; forecast; coherent forecasts; functional data; life expectancy; sex-ratio.

1. INTRODUCTION

Functional time series (FTS) encompasses data in the form of curves that are observed at regular intervals in time. Recently these models are applied for demographic forecasting and breast cancer mortality forecasting (see Hyndman and Ullah 2007 and Yasmeen et al. 2010).

However, the application of these models in presence of two or more groups within similar populations needs some modification. In this paper, we are relating some of the functional models to the common principal components (CPC) and providing the methods to estimate these models. We will call them common functional principal component (CFPC) models.

2. CPC AND CFPC MODELS

The CPC analysis assumes that the space spanned by the eigenvectors is identical across groups whereas the variances associated with them vary (Flury 1988). CPC models have never been used for mortality forecasting.

Suppose there are J related functional time series corresponding to age-specific mortality rates of J groups. Let $m_{t,j}(x)$ denote the mortality rate for age x and year *t*, *t*= 1,...,*n* , for the *j*th group. We will model the log mortality, $y_{t,j}(x_i) = \log[m_{t,j}(x)]$, and assume that there is an underlying smooth function $g_{t,j}(x)$ that we are observing with error. Thus,

$$y_{t,j}(x_i) = g_{t,j}(x_i) + \sigma_{t,j}(x_i)\varepsilon_{t,j,i} \tag{1}$$

Since the mean function can vary across groups, the following model termed as CFPC-I can be used:

$$g_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \varepsilon_{t,j}(x) \tag{2}$$

We can define another model with same basis functions but different coefficients

$$g_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,j,k}\phi_k(x) + \varepsilon_{t,j}(x) \tag{3}$$

We define a functional linear model with more than one, say K common and L specific components, we have

$$g_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{l=1}^{L} \gamma_{t,j,l}\psi_{j,l}(x) + \varepsilon_{t,j}(x) \tag{4}$$

where each $\gamma_{t,j,l}$ is a stationary time series, but $\beta_{t,k}$ may be non-stationary. This model will be termed as Partial Common Functional Principal Component (PCFPC)(*p*,*q*) model.

3. ESTIMATION OF PARAMETERS

Let ***M*** be a $p \times n$ matrix with (*i* , *t*)th element m_{it} , i.e.

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \ldots & m_{1n} \\ m_{21} & m_{22} \ldots & m_{2n} \\ \vdots & \vdots \;\; \vdots & \vdots \\ \vdots & \vdots \;\; \vdots & \vdots \\ m_{p1} & m_{p2} \ldots & m_{pn} \end{bmatrix}$$

For CFPC-I, first we obtain the weighted average of the log mortality rates after smoothing. Let $f_{t,j}(x_i)$ is the value of smoothed log mortality rate at age x_i and time *t* for the *j*th group and $p_{t,j}(x_i)$ is the (*i* , *t*)th element of $\mathbf{P}_j$, the total number of exposures for the *j*th population, *j* = 1, 2, . . . , *J*.

$$P_j(x) = [P_{1,j}(x) \quad P_{2,j}(x) \quad \ldots \quad P_{n,j}(x)]$$

where

$$P_{t,j}(x) = \begin{bmatrix} p_{t,j}(x_1) \\ p_{t,j}(x_2) \\ \vdots \\ \vdots \\ p_{t,j}(x_p) \end{bmatrix}.$$

For CFPC-I, the matrix $\boldsymbol{M}$ can be obtained as

$$m_{it} = \frac{\sum_{j=1}^{J} [f_{t,j}(x_i) - \mu_j(x_i)] \times p_{t,j}(x_i)}{\sum_{j=1}^{J} p_{t,j}(x_i)} \qquad (5)$$

After obtaining $\boldsymbol{M}$ from the elements given in equation (5), we apply SVD to get the common basis functions and coefficients for model given in (2).
For CFPC-II, the matrix $\mathbf{M}$ will be

$$\mathbf{M} = [M_1 \quad |M_2| \quad \ldots \quad |M_J],$$

where

$$\mathbf{M}_k = \left(m_{(it)k}\right)_{p\times n}$$

and

$$m_{(it)j} = \frac{[f_{t,j}(x_i) - \mu_j(x_i)] \times p_{t,j}(x_i)}{\sum_{j=1}^{J} p_{t,j}(x_i)}. \qquad (6)$$

Applying SVD to $\boldsymbol{M}$ gives $\boldsymbol{M} = \boldsymbol{\Psi}\,\boldsymbol{\Lambda}\,\boldsymbol{P}$,where $\boldsymbol{\Psi}$ is of order $p\times p$ and $\boldsymbol{B} = \boldsymbol{M}'\,\boldsymbol{\Psi}$ is of order $Jn\times p$. The common basis function $\hat{\psi}_j(x_i)$ is the (i, j)*th* element of $\boldsymbol{\Psi}$. The specific coefficients $\hat{\beta}_{t,j,k}$ can be obtained from $\boldsymbol{B} = \boldsymbol{M}'\,\boldsymbol{\Psi}$.

Finally, for the PCFPC model, we first apply SVD to the matrix $\boldsymbol{M}$ for CFPC-I given in equation (5) so that, the terms are estimated.

$$\mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x)$$

Next, the problem is to estimate the specific coefficients and basis functions for each group. For this, we apply SVD to the residual matrix for the *j*th group defined as

$$\mathbf{R}_j = [\mathbf{R}_{1,j}(x) \quad \mathbf{R}_{2,j}(x) \quad \ldots \quad \mathbf{R}_{n,j}(x)],$$

where

$$\mathbf{R}_{t,j}(x) = f_{t,j}(x) - \left[\mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x)\right]. \qquad (7)$$

Applying SVD to $\boldsymbol{R}_j$ gives $\boldsymbol{R}_j = \boldsymbol{A}_j\boldsymbol{\Psi}_j\boldsymbol{B}_j$. The specific basis function $\psi_{j,l}(x_i)$ is the (*i* , *l*)th element of $\boldsymbol{A}_j$ where as the coefficient $\gamma_{t,j,l}$ can be obtained from $\boldsymbol{R'}_j\boldsymbol{A}_j$, as described for the common components.

4. FORECASTING FROM CFPC MODELS

To obtain the forecast value of mortality rates from CFPC model, we first obtain the forecast for each of the common (or non-common) components in the model.
are estimated.

Let $\hat{\beta}_{n,h,j}$ and $\hat{\beta}_{n,h,j,k}$ denote the *h*-step ahead forecast of $\beta_{n+h,j}$ and $\beta_{n+h,j,k}$ respectively. Also, let $\hat{g}_{n,h,j}(x)$ denote the *h*-step ahead forecast of $g_{n+h/n,j}(x)$ for *j* = 1, 2, . . . , J. Then the forecasts for CFPC models are given by are estimated.

$$\hat{g}_{n+h/n,j}(x) = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n,h,k}\hat{\phi}_k(x) \qquad (8a)$$

$$\hat{g}_{n+h/n,j}(x) = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n,h,j,k}\hat{\phi}_k(x) \qquad (8b)$$

$$\hat{g}_{n+h/n,j}(x) = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n,h,k}\hat{\phi}_k(x) + \sum_{l=1}^{L} \hat{\gamma}_{n,h,j,l}\hat{\psi}_{j,l}(x) \qquad (8c)$$

where $\hat{\mu}_j(x)$ are the estimated values of $\mu_j(x)$, j = 1, 2, . . . , J and $\hat{\phi}_k(x)$, $k = 1,2,\dots,K$ are the estimated common basis functions. Here $\hat{\gamma}_{n,h,j,l}$ denote the h-step ahead forecast of $\gamma_{n+h,j,l}$ for PCFPC model given in equation (4).

5. FORECASTING FROM PCFPC MODELS

For forecasting the specific coefficients $\gamma_{t,j,l}$ of a PCFPC model, we use vector autoregressive (VAR) models with cointegration (see Engle and Granger 1987).

Cointegration when cointegrating vector is pre-specified

For the two-sex data, suppose the males and females time series coefficients are cointegrated with prespecified vector $\beta = (1, -1)'$ i.e the long-run relationship between the male series $\mathcal{Y}_{t,M}$ and female series $\mathcal{Y}_{t,F}$ is $\mathcal{Y}_{t,M} = \mathcal{Y}_{t,F}$. We can test it for its cointegration using a unit root test for the difference series $d = \mathcal{Y}_{t,M} - \mathcal{Y}_{t,F}$, using Augmented Dickey Fuller ADF test (see Said and Dickey 1984). We have to test

$$H_0 = \beta'\mathcal{Y}_t = \mathcal{Y}_{t,M} - \mathcal{Y}_{t,F} \sim I(1) \quad \text{(no cointegration)}$$
$$H_1 = \beta'\mathcal{Y}_t = \mathcal{Y}_{t,M} - \mathcal{Y}_{t,F} \sim I(0) \quad \text{(cointegration)}$$

Cointegration is found if the unit-root test rejects the no-cointegration null (Hamilton 1994). For forecasting the specific components, we use Johnsen Methodology (1988).

Johansen Methodology

Johansen methodology starts with a VAR model. A VAR process with p lags is defined as

$$\mathcal{Y}_t = \mathbf{\Phi}_0 + \mathbf{\Phi}_1\mathcal{Y}_{t-1} + \mathbf{\Phi}_2\mathcal{Y}_{t-2} + \cdots + \mathbf{\Phi}_p\mathcal{Y}_{t-p} + \varepsilon_t \qquad (9)$$

where $\mathbf{\Phi}_i$'s are (K× K) coefficient matrices for i = 1, 2, . . . , p and ε_t is K-dimensional white noise i.e. $\varepsilon_t \sim N(0, \Sigma)$.

One can obtain the following form of vector error correction models (VECM):

$$\Delta\mathcal{Y}_t = \mathbf{\Phi}_0 + \mathbf{\Pi}\mathcal{Y}_{t-1} + \mathbf{\Gamma}_1\Delta\mathcal{Y}_{t-1} + \mathbf{\Gamma}_2\Delta\mathcal{Y}_{t-2} + \cdots + \mathbf{\Gamma}_{p-1}\Delta\mathcal{Y}_{t-p+1} + \varepsilon_t \qquad (10)$$

with

$$\mathbf{\Gamma}_{\boldsymbol{i}} = -\sum_{j=i+1}^{p}\mathbf{\Phi}_{\boldsymbol{j}} \quad for\ \ i = 1,2,\dots,p-1$$

and

$$\mathbf{\Pi} = -\left(I - \sum_{i=1}^{p}\mathbf{\Phi}_{\boldsymbol{i}}\right)$$

Here matrix $\mathbf{\Pi}$ is called the long run impact matrix and matrices $\mathbf{\Gamma}_{\boldsymbol{i}}$ are called short-run impact matrices. The matrix $\mathbf{\Pi}\mathcal{Y}_{t-1}$ represents the cointegration relation.

Forecasting from VECM

To forecast from VECM, we first convert it into appropriate VAR model . e.g. in bivariate case

$$\Delta\mathcal{Y}_t = \mathbf{\Phi_0} + \mathbf{\Gamma_1}\Delta\mathcal{Y}_{t-1} + \mathbf{\Pi}\,\mathcal{Y}_{t-1} + \varepsilon_t \qquad (11)$$

After estimating the matrices $\mathbf{\Pi}$ and $\mathbf{\Gamma}_{\boldsymbol{i}}$, we can easily convert this model into VAR(2) using

$$\mathbf{\Phi}_2 = -\mathbf{\Gamma}_1 \tag{12a}$$
$$\mathbf{\Phi}_{\mathbf{1}} = \mathbf{\Gamma}_{\mathbf{1}} - \mathbf{\Pi} + \boldsymbol{I}_2 \tag{12b}$$

Once the parameters $\mathbf{\Phi}_{\boldsymbol{i}}$'s are estimated, forecasts can be obtained in the usual manner as forecast from a VAR model.

Choosing among the Best CFPC model

We define

K: the number of common components and *L*: the number of non-common or specific components

$PV_{1,k}$: Percentage variance explained by the first *K* common components;

$PV_{2,j,l}$: Percentage variance explained by the first *L* non-common components in *j*th group,

where

$$PV_{1,k} = \frac{\lambda_k}{\sum_{k=1}^{k=p} \lambda_k}, \qquad PV_{2,j,l} = \frac{\lambda_{j,l}}{\sum_{l=1}^{l=p} \lambda_{j,l}} \tag{13}$$

where λ_k and $\lambda_{j,l}$ are the eigenroots corresponding to the *k*th common and *l*th non-common factor for the *j*th group respectively. Also, one can define the cumulative variation

CP_1: Total variation explained by the first *K* common components;
$CP_{2,j}$: Total variation explained by the first *L* non-common components in *j*th group; with

$$CP_1 = \frac{\sum_{k=1}^{K} \lambda_k}{\sum_{k=1}^{k=p} \lambda_k}, \qquad CP_{2,j} = \frac{\sum_{l=1}^{l=L} \lambda_{j,l}}{\sum_{l=1}^{l=p} \lambda_{j,l}} \tag{14}$$

Coefficient of Explanation ($\boldsymbol{CE_j}$)

To measure the performance of a CFPC model, we define

$$CE_j = 1 - \frac{\sum_{t=1}^{n} \int_x [\mathcal{Y}_{t,j}(x) - \hat{g}_{t,j}(x)]^2 dx}{\sum_{t=1}^{n} \int_x [\mathcal{Y}_{t,j}(x) - \hat{\mu}_j(x)]^2 dx} \tag{15}$$

where $\mathcal{Y}_{t,j}(x)$ are the observed value of log mortality rate and $\hat{g}_{t,j}(x)$ are fitted values obtained from a CFPC model.

6. EMPIRICAL APPLICATION

In this section, we will illustrate the procedure of fitting and forecasting through CFPC models using an application to the age-sex specific data of Australia. The data are obtained from Hyndman (2008).

Table 1 shows the coefficient of explanation for different CFPC models, each with six common components. CFPC-I gives considerably higher values of CE_j and the model is explaining about 96.82% variability in females and 96.38% in males.

The values of CE_j for CFPC-II are slightly lower in both groups (74.4% in female and 73.88% in males).

Table 1
Coefficient of Explanation for Australian Sex Data

Model	Female	Male
CFPC-I (6)	0.9682	0.9638
CFPC-II (6)	0.7439	0.7388
PCFPC(6,6)	0.9931	0.9904
Independent	0.9886	0.9857

Table 2
p-value for the Augmented Dickey-Fuller unit root test for specific coefficient $\gamma_{t,j,l}$

	Specific Coefficients					
Series	1	2	3	4	5	6
Male	0.6099	0.9566	0.0100	0.0118	0.0317	0.0259
Female	0.5156	0.9004	0.0308	0.0215	0.0265	0.0154

Table 3
p-values for Augmented Dickey Fuller unit root test for Engle-Granger Cointegration Method.

	Engle-Granger: Co-integration Test					
Coefficient Number	1	2	3	4	5	6
Residuals of male model	0.0166	0.0204	0.0255	0.0183	0.0466	0.0453
Residuals of female Model	0.0100	0.0395	0.0100	0.0100	0.0306	0.0278

If we apply PCFPC model, we will get the best fitted model as now it is capturing about 99% variability in both groups. We use *K*=6 and *L*=6 and the value of CP_1 is 99.4% for common components and the values of $CP_{2,j}$ are 83.2% and 79.9% for females and males respectively, which are sufficiently high.

Estimation and Forecasting from PCFPC Models

Table 2 represents the *p*-value of Augmented Dickey Fuller (ADF) unit root test for the specific coefficients for males and females. For the first two coefficients, we are unable to reject the hypothesis about the presence of unit-root. The next step is to determine whether the individual series are cointegrated. For this, we can use Engle-Granger cointegration test based on the residuals from OLS regression of one variable (say $\mathcal{Y}_{1t}$) on other variable ($\mathcal{Y}_{2t}$).

$$\mathcal{Y}_M = \alpha_M + \beta_M \mathcal{Y}_F + z_{1t} \tag{16a}$$

$$\mathcal{Y}_F = \alpha_F + \beta_F \mathcal{Y}_M + z_{2t} \tag{16b}$$

Table 3 represents the *p*-value of ADF test for the static

Table 4
p-values of ADF test for the difference of coefficients $\gamma_{t,j,l}$

	Difference of Male and Female Coefficients					
Coefficient Number	1	2	3	4	5	6
p-value for differences	0.01418	0.0205	0.0165	0.0100	0.0144	0.0265

residuals z_{1t} and z_{2t}). It is interesting that for both males and females, the p-values are smaller than 0.05. After confirming the cointegration between the first two coefficients of males and females, the next step is to determine the nature of cointegration relation. For bivariate time series, the cointegrating vector $\boldsymbol{\beta}$ will be

$$\boldsymbol{\beta} = (1, -\beta^*)'$$

In our case, suppose that the demographers assume that there is a long-run equilibrium in the male and female series with pre-specified cointegrating vector $\beta = (1, -1)'$. For this, we first check the presence of unit-root in the differences of observed series. Table 4 shows the p-values of Augmented Dickey Fuller test for the difference of specific coefficients $d = \mathcal{Y}_M - \mathcal{Y}_F$. The results in tables (2) and (4) confirm that the original series were I(1) and their differences are stationary.

7. CONCLUSION

In this paper, we introduced a new class of functional linear models for coherent mortality forecasting. We developed the methods for estimating their parameters and forecasting from these models.

It is found that the specific time series coefficients among different subgroups are highly correlated; hence we used vector autoregressive (VAR) and vector error correction models (VECM) to forecast them. For the purpose of illustration, the models are applied to the two-sex data of Australia. It is found that all new methods work well and mortality rates and life expectancy can be forecast in a coherent way. Also, PCFPC model provides some extra information about the age-groups that are most responsible for the difference. We found that the difference of male and female series $d = \mathcal{Y}_M - \mathcal{Y}_F$ is a stationary process. It means that there is long run equilibrium among male and female coefficients.

REFERENCES

1. Engle, R. and Granger, C. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55, 251-276.
2. Erbas, B., Hyndman, R.J. and Gertig, D.M. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 26, 458-470.
3. Flury, B. (1988). *Common Principal Components & Related Multivariate Models*, Wiley Series in Probability and Mathematical Statistics.
4. Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.
5. Hyndman, R.J. (2008). *addb: Australian Demographic Data Bank*. R package version 3.222. URL: *robjhyndman.com/software/addb*

6. Hyndman, R.J., Booth, H. and Yasmeen, F. (2012). Coherent forecasting of mortality rates using functional time series models. *Demography*, (To appear).
7. Hyndman, R.J. and Ullah, M.S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics and Data Analysis*, 51, 4942-4956.
8. Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economics Dynamics and Control*, 12, 231-254.
9. Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, Springer Series in Statistics.
10. Yasmeen, F., Hyndman, R.J. and Erbas, B. (2010). Forecasting age-related changes in breast cancer mortality among white and black US women: A functional data approach. *Cancer Epidemiology*, 34(5), 542-549.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 601-606

THE PERFORMANCE OF ROBUST-DIAGNOSTIC F IN THE IDENTIFICATION OF MULTIPLE OUTLIERS

Habshah Midi[1] and **Nor Mazlina Abu Bakar**[2]
[1] Institute of Mathematical Research & Faculty of Science,
Universiti Putra Malaysia, Serdang, Malaysia.
Email: habshahmidi@gmail.com
[2] Universiti Sultan Zainal Abidin, Terengganu, Malaysia.
Email: normazlina@gmail.com

ABSTRACT

It is now evident that outliers have an unduly effect on the least squares estimates and they are responsible for misleading conclusion about the fitting of regression model and collinearity measures. In this respect, outlier detection for multivariate data plays significant role in applied sciences. Classically, these outliers are detected by Mahalanobis Distance (MD_i) but the measure is found to mask a few outliers. To remedy masking, robust estimates of location and scatter are utilized to obtain the Robust Mahalanobis Distance, (RMD_i). Unfortunately, the robustness of RMD_i comes with a tendency to swamp a few inliers. In this regard, we propose the combine use of robust and diagnostic method that we call *RDF* for the identification of multiple outliers. The *RDF* consists of three steps whereby in the first step the suspected outlying data points are determined by the robust distances. The second step checks each observation in the suspected group by using F statistics (Djauhari 2010) to confirm our suspicion. In the third step, any inlier detected in Step 2 will be added back to the initial clean data set and be updated. The performance of *RDF* is evaluated through real data and simulations study and compared to other methods. The numerical results indicate that the proposed *RDF* is very successful in identifying the correct outliers in multivariate data. We relate the effectiveness of the method to the key factor - F statistic that detects the change in data covariance structure upon the entrance of outlying values into the data set. The effect of outliers on the least squares estimates and collinearity diagnostics measures are also illustrated.

KEYWORDS

Robust, Diagnostic, Frobenius Norm, Outliers.

1. INTRODUCTION

Real data are expected to contain a small percentage of outliers (Hampel 2011). These outliers arise from many different sources such as mechanical or human errors. Very more often, their occurrences provide useful information for example fraud detection or an indication that machine needs calibration. At the same time, these outliers can cause an analysis or result to become unreliable or bias. Therefore, the identification of outliers in a dataset is vital so that corrective measures can be taken up to produce a more reliable and unbiased analysis.

Classically, multivariate outliers are detected by calculating Mahalanobis distances of the n observations, Let X be an $n \times p$ matrix representing a random sample of size n from a p-dimensional population. For Mahalanobis Distance (MD_i) of the i-th observation is defined as:

$$MD_i = \sqrt{(X_i - T)C^{-1}(X_i - T)'} \tag{1}$$

for $i = 1, \dots, n. T$ and C are the classical measures of location and shape respectively represented by the arithmetic mean, $\bar{X}$ and the sample covariance matrix, S. Any observation with MD_i values greater than $\chi^2_{p,0.975}$ cutoff point may give an indication of outlyingness. Unfortunately, MD_i are largely affected by multiple outliers due to the non-robust property of both classical measures of location and shape (Hadi 1992). In order to resolve the problem, robust location and scatter estimates are used since they are resistant to outliers. Minimum Volume Ellipsod (MVE) and Minimum Covariance Determinant (MCD) by Rousseeuw and Van Zomeren (1990) can provide the robust location and scatter estimates to obtain the Robust Mahalanobis Distance, RMD_i. However, Imon (2002) pointed out that the RMD_i with MVE (or MCD) often suffer from swamping effect. From the work of Fung (1993), the complementary use of robust-diagnostic method is found to give satisfactory results. This is further supported by Habshah et. al (2009) who proposed Diagnostic Robust Generalized Potential (DRGP) which combines RMD_i and the Generalized Potential. The initial stage of DRGP involves dividing the dataset into two subsets; a clean dataset and a suspicious dataset by using RMD_i. Each suspicious data is later checked with the generalized potential to confirm the true outliers. As a result, DRGP is found to successfully identify the true outliers and largely reduce the swamping effect of RMD_i. Hadi (1992) pointed out that the main factor to a successive identification of real outliers is to find the correct initial basic subset. In DRGP, the swamping property of RMD_i provides a new advantage in obtaining a real clean initial dataset.

This study is aimed to propose another robust-diagnostic method called Robust-Diagnostic F(RDF). The next section describes the proposed procedure which involves the combine use of RMD_i and a diagnostic measure taken by F (Djauhari 2010). The performance of RDF in detecting multiple outliers for multivariate data is studied and comparisons are made with the classical and robust detection methods. Collinearity diagnostics are also studied using numerical examples.

2. THE PROPOSED METHOD – ROBUST DIAGNOSTIC F (RDF)

The proposed method, RDF consists of three steps:

First Step: RMD_i is used to identify the initial basic subset. Based on the RMD_i values, the dataset can be categorized into two subsets – initial basic subset and suspicious subset. The initial basic subset will only contain clean data and the suspicious subset may contain potential outliers due to the swamping property of RMD_i. As suggested by Imon (2002), cutoff value for RMD_i is taken as:

$$\text{Median}\,(RMD_i) + 3(RMD_i) \tag{2}$$

Second Step: A diagnostic procedure is performed whereby each member of the suspicious subset with the lowest RMD_i value is tested with F statistic as proposed by Djauhari (2010). Consider $X_1, X_2, \ldots, X_n, X_{n+1}$ to be a random sample from p-variate normal distribution with covariance matrix Σ and let:

$$SS_k = \sum_{i=1}^{k}(X_i - \bar{X}_k)(X_i - \bar{X}_k)^T \tag{3}$$

with $$\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i \tag{4}$$

where SS_k is the scatter matrix from the clean initial subset when $k = n$ and from the suspicious subset if $k = n + 1$. If $D = SS_{n+1} - SS_n$ then $F = \sqrt{Tr(D)}$ represents the effect of X_{n+1} on the scatter matrix of the initial subset. Any outlier will immediately change the sample variance structure and be detected by the F statistics. Thus, it can be tested whether the addition of another data point can change the covariance structure of the initial subset. From Djauhari (2010), when Σ is unknown, the distribution of F can be approximated by $c\chi_r^2$ where:

$$c = \frac{Tr(S_n^2)}{Tr(S_n)} \text{ and } r = \frac{\{Tr(S_n)\}^2}{Tr(S_n^2)} \tag{5}$$

where S_n is the sample covariance structure of the remaining dataset. Thus, the cutoff point is taken as the $(1 - \alpha)$-th quantile of $c\chi_r^2$, where α = 0.025. Any data point which exceeds the cutoff value is considered as an outlier.

Third Step: Update the initial subset. In this step, any inlier detected in Step 2 will be added into the basic initial subset and form a new basic subset. Then the procedure is repeated until every data in the suspicious group is checked.

3. NUMERICAL EXAMPLES

We consider three well-known data sets in applied sciences to illustrate RDF and compare its performance with existing measures to identify multiple outliers. The datasets are Hawkins-Bradu-Kass data, stackloss data and aircraft data. Hawkin *et al.* (1984) reported that the Hawkin –Bradu-Kass data has 14 high leverage points while stackloss data has 4 high leverage points (Imon, 2005). High leverage points are outliers in the x directions. The results are exhibited in Table 1 which summarizes the ability of RDF, MD_i and RMD_i to identify outliers in all datasets. The MD performs poorly for all the three data sets due to masking effects. The RMD swamped few high leverage points for aircraft and stackloss data but able to detect all the high leverage points for Hawkin Bradu- Kass data. The RDF successfully identify all the high leverage points in all the three data sets. It has the ability to detect even small changes in the structure of the dataset. Thus, any aberrant data introduced in a clean dataset is detected as outliers by RDF whereas MD_i and RMD_i are prone to mask and swamp outliers, respectively.

Table 1
Outliers detected by different methods in well-known datasets

Data	n	No. of variables	MD	RMD	RDF
1. Aircraft Data	23	4	14, 22	14, 20, 22	22
2. Stackloss Data	21	3	21	1, 2, 3, 4, 13, 21	1, 2, 3
3. HBK Data	75	3	12, 14	1, 2, …, 14	1, 2, …, 14

Collinearity diagnostics such as Pearson's correlation and Variance Inflation Factor (VIF) for selected datasets are presented in Table 2 to see the effect of high leverage points on the collinearity measures. The results of Table 2 show that the collinearity measures for the three data sets are small in the absence of outliers. On the other hands when high leverage points are present in all the data sets, these measures increased. Hence, these high leverage points are collinearity enhancing observations.

Table 2
Collinearity diagnostics for selected datasets

Data		Pearson's correlation coefficient		VIF	
		With Outliers	Without Outliers	With Outliers	Without Outliers
Aircraft Data	*1*	r_{12}=-0.151	-0.024	1.927	2.476
	2	r_{23}=0.336	0.050	1.431	1.511
	3	r_{24}=0.463	0.337	6.501	3.419
	4	r_{34}=0.914	0.807	8.433	5.638
HBK Data	*1*	r_{12}=0.946	0.044	13.432	1.012
	2	r_{23}=0.979	0.127	23.853	1.017
	3	r_{13}=0.962	0.107	33.432	1.027

Table 3 exhibits the Ordinary Least Squares (OLS) and the MM estimates in the presence and absence of outliers in the data. When all assumptions are met including no outliers in the data, the OLS is the best estimator. It can be observed from Table 3 that the OLS parameter estimates and the standard errors of OLS estimates are different in the situation when outliers are present and not present in the data. We observe from the table that the OLS estimates are easily affected by outliers. To rectify this problem, we recommend using the robust MM estimator which is outlier resistant. It is interesting to point out that the MM estimates are reasonably closed to the OLS estimates in the absence of outliers but very far from the OLS estimates when outliers are present in the data. Thus, the OLS estimates are not reliable when outliers are present in the data and the use of robust MM estimator will produce very efficient estimates.

Table 3
Least Squares and MM estimates and standard errors in parenthesis for selected datasets

Data	Coef.	With Outliers		Without Outliers
		LS	MM	LS
Aircraft Data	***X1***	-3.853 (1.763)	-3.049 (1.003)	-3.353 (1.109)
	X2	2.488 (1.187)	1.210 (0.708)	1.525 (0.645)
	X3	0.004 (0.001)	0.001 (0.001)	0.002 (0.001)
	X4	0.002 (0.001)	-0.001 (0.001)	-0.001 (0.001)
HBK Data	***X1***	0.240 (0.263)	0.081 (0.073)	0.062 (0.069)
	X2	-0.335 (0.155)	0.040 (0.044)	0.012 (0.068)
	X3	0.383 (0.129)	-0.052 (0.040)	-0.107 (0.071)

5. CONCLUSIONS

We proposed a robust diagnostic measure, RDF to identify multiple outliers for multivariate data. Its performance is compared to the classical and robust MD_i. We presented evidences that the proposed method surpass the existing methods in the detection of outliers. Although RMD_i tends to swamp some data points, RDF is able to identify the outliers correctly. We relate the effectiveness of the method to the key factor-F statistic that detects the change in data covariance structure upon the entrance of outlying values into the data set. The RDF works very well in small dataset and applications to large datasets or large p are assumed to produce more interesting results.

REFERENCES

1. Djauhari, M. (2010). A multivariate process variability monitoring based on individual observations. *Modern Applied Science*, Vol 4, No 10.
2. Fung, W. (1993). Unmasking outliers and leverage points: a confirmation. *Journal of the American Statistical Association*, 88, 515-519.
3. Habshah, M., Norazan M.R. and Imon, A.H.M.R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36, 507-520.
4. Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society*, B, 54, 761-771.
5. Hampel, F.R. (2001). *Robust statistics: A brief introduction and overview*. Research Report No. 94, Seminar für Statistik, Eidgenössische Technische Hochschule.
6. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics*, pp. 197-208.
7. Imon A.H.M.R. (2002). Identifying multiple high leverage points in linear regression, *Journal of Statistical Studies, 3*, 207-218.

8. Imon A.H.M.R. (2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*, 32, 929-946.
9. Rousseeuw P.J. and Van Zomeren B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633-639.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 607-618

MODELLING THE AGE SPECIFIC FERTILITY RATES OF RURAL AND URBAN AREAS OF PAKISTAN DURING 1984-2007

Muhammad Zakria[1], **Muhammad Yaseen**[2], **Faqir Muhammad**[3]
and **Sajid Mahmood**[4]

[1] Department of Mathematics & Statistics, Allama Iqbal Open University, Islamabad, Pakistan. Email: zakriauaf@yahoo.com

[2] University of Nebraska, Lincoln-NE, USA. Email: myaseen208@gmail.com

[3] Department of Business Administration, Air University, Islamabad, Pakistan Email: aioufsd@yahoo.com

[4] Government Degree College, Khalabat Township, Haripur, Pakistan. Email: aqsmkts@gamil.com

ABSTRACT

This paper investigates the fertility pattern of Pakistan and its major breakdown rural and urban areas. The analysis of the age specific fertility rates (ASFRs) illustrated the greater fertility in the rural population as compared to the urban and of Pakistan population during the period 1984-2007. Approximately, 44.98% decrease in total fertility rate (TFR) of urban population during the past 23 years has been seen whereas 48.72% and 46.76% decrease in TFR of urban area and of Pakistan respectively with reference to the 1984. Different parametric models were fitted on the ASFRs of rural, urban and Pakistan during the year 2007. But for the sake of comparison, the outcomes of only two highly significant parametric models i.e. Hadwiger function and Peristera and Kostaki model (P-K model-I) were discussed thoroughly in this paper. Consequently, the P-K model-I is proposed as a parsimonious model. To examine the disparity among the ASFRs of rural, urban and Pakistan, Gini coefficients& Lorenz curves were also computed for Pakistan as well as its major geographical regions. The Lorenz curves explained more concentration within the urban population as compared to the Pakistan and its rural population during 2007.

KEYWORDS

Hadwiger function; Peristera and Kostaki; Gini coefficients and Lorenz curves

1. INTRODUCTION

The population studies especially the focus on fertility is the most interesting area for national and international demographers and researchers. In demography, fertility indicates the product or output of reproduction, rather than the ability to have children. It is directly proportional to the population of a territory. In literature, different measures of fertility are used to estimate the reproductivity of the population. The most common and direct measures of fertility are TFR, gross reproduction rate (GRR), net reproduction rate (NRR) and ASFRs etc. These measures vary from country to country with respect to own country's health status and circumstances (Siegel & Swanson, 2004).TFR of Pakistan was reported

4.1 (Pakistan Demographic Survey ([PDS], Federal Bureau of Statistics [FBS], (2006))whereas TFR was stated 3.4 children per women in 2011-12 as compared to 3.5 in 2010-11 [Pakistan Economic survey 2010-11]. A number of national and international scientists always insist to model the reproductively pattern of the population of a territory. Kabir & Mosleh Ud din (1987) used the age specific fertility rates to examine the fertility trend of Bangladesh during 1953-1986. In this study, the fitted model showed an increasing trend during the period 1953-1973 whereas the model started to decrease from 1974 to 1986. Bairagi & Datta (2001) also studied the fertility pattern of Bangladesh and revealed that TFR of Bangladesh decreased till 1966, whereas from 1967 to 1974, TFR was increased and after that TFR again started to decrease till 1998. Paraskevi & Kostaki (2007) also modelled the fertility in modern population. Ishida, Stupp & Melian (2009) discussed the fertility decline in Paraguay. Sathar (1993) reported the fertility decline specifically among the women of age 15-19 years old as that of the rising age at marriage. It is an accepted reality that female age at marriage increases with the increase of female education. Sathar et al. (1988) also publicized the three significant determinants of fertility i.e. age at marriage, female education and workforce participation. Islam & Nesa (2009) examined the fertility transition in Bangladesh through educational differentials and established that secondary level educated rich women have lowest TFR i.e. closer to the replacement level as compared to rich household women. It was also recommended that higher education should be provided free of cost to bring the further reduction in fertility in Bangladesh. Zakria et al. (1999) publicized the strong association between women education and her contraceptive behaviour. Islam & Ali (2004) worked out demographic cohort measures of the rural Bangladesh using the age specific fertility rates during 1980-1998. Polynomial regression models were also fitted on the ASFRs and age of the mother. The third degree polynomial model was suggested as a parsimonious model for Bangladesh fertility. Nasir, Akhtar & Tahir (2009) also used the age specific fertility rates of Pakistan to estimate the almost same demographic cohort measures, parsimonious polynomials and evaluated their goodness of fit as that of Islam & Ali (2004). In both these studies, the analysis and evaluation techniques are exactly the same but the countries are different. These statistics were only the estimation of the parameters of polynomials, not the interpretation of the parameters estimates. The issue is that, if the interpretation of the estimates of model parameters is not essential, then it is ok. On the other hand, if the testing and interpretation of an estimate is an essential, then this part seems missing in these articles. This important point became the novelty of this study. The main objectives of this study are almost the same as that of Islam & Ali (2004) and Nasir et al. (2009), but these objectives have been achieved through interpretable parsimonious parametric models using the ASFRs for Pakistan and its rural-urban breakdown during 1984-2007. In addition to, the concentration/inequality among age specific fertility rates corresponding to different age groups has also been examined for some specified years through Gini coefficients and Lorenz curves.

2. MATERIAL AND METHODS

2.1 About the Data

The empirical data on ASFR of Pakistan and its rural-urban breakdown had been taken from Federal Bureau of Statistics [FBS], now re-named as Pakistan Bureau of Statistics [PBS] 2006. The data were provided by Pakistan Demographic Surveys (PDS) 1984-1986, 1988-1992, 1995-1997, 1999-2001, 2003, 2005, 2006, and 2007. Furthermore, the ASFRs are not available for the years 1987, 1993-94, 1998, 2002 and 2004. PDS is the most reliable source of population dynamics data in Pakistan since 1984.

2.2 Methodology

First of all, a summary statistics was obtained by computing the commonly used demographic cohort measures i.e. TFR, GRR, Mean age of childbearing (MAC) and Standard deviation of age of childbearing (SDAC) (see Table-1). These measures are usually used to examine the fertility pattern of population data of Pakistan. Different parametric models were also fitted on the ASFRs of 2007 to propose a parsimonious model for the fertility behaviour of rural-urban areas (Montgomery et al., 2012; Kostaki et al., 2009). The algebraic expressions i.e. *f (x)* of only these two models are given at this juncture:

The Hadwiger function (Hadwiger 1940; Gilje 1969) is expressed by

$$f(x) = \frac{ab}{c}\left(\frac{c}{x}\right)^{\frac{3}{2}} \exp\left\{-b^2\left(\frac{c}{x}+\frac{x}{c}-2\right)\right\}$$

where x is the age of the mother at birth and a, b, c are the parameters to be estimated. Chandola et al. (1999) argued that the parameters of the model may have a demographic interpretation as follows. Parameter a is associated with total fertility, parameter b determines the height of the curve and the parameter c is related to the mean age of motherhood, while the term $\frac{ab}{c}$ is associated to the maximum age-specific fertility rate (Kostaki et al., 2009).

Peristera and Kostaki (2007) proposed a flexible and simpler model for describing the fertility pattern (hereafter P-K model-I) is

$$f(x) = c_1 \exp\left(-\left(\frac{x-\mu}{\sigma_{(x)}}\right)^2\right)$$

where $f(x)$ is the age-specific fertility rate at mother age x, whereas c_1, μ, σ are the parameters to be estimated, while $\sigma_{(x)} = \sigma_{11}$ if $x \le \mu$, and $\sigma_{(x)} = \sigma_{12}$ if $x > \mu$. The parameter c_1 describes the base level of the fertility curve and is associated with the TFR; μ reflects the location of the distribution, i.e., the modal age, while σ_{11}, σ_{12} reflect the spread of the distribution before and after its peak respectively (Kostaki et al., 2009).

Furthermore, the most popular technique known as Gini coefficient was used to measure the disparity in ASFRs with respect to the different age groups during some specific years. The Lorenz curves of the ASFRs of different age groups of Pakistan and its rural-urban breakdown were also drawn to display the inequality.

3. RESULTS AND DISCUSSION

Initially, the most commonly used demographic measures i.e. TFR, GRR, MAC and SDACs were calculated from ASFRs for Pakistan and its major geographical regions i.e. urban and rural during the years 1984-2007 (see Table-1). In 1984, TFR of rural area of

Pakistan was 7.27 whereas it reduced 4.00 in 2007; which indicated that about 44.98% decrease in rural fertility during the past 23 years. Similarly, the TFRs were decreased up to 48.72% and 46.76% in urban area and in Pakistan respectively as compared to the year 1984. Moreover, almost the same decrease had been seen in GRRs within the above mentioned three regions. The MAC and SDAC of Pakistan were 30.05 and 7.13 years respectively during the year 1984 which were slightly decreased up to 29.75 and 6.69 years in 2007 respectively, these results indicate the diminutive decrease in MAC and SDAC. Nasir et al. (2009) also reported some demographic estimates i.e. TFR, GRR and MAC for Pakistan only.

The fertility parameters of Hadwiger function and of P-K model-I were estimated for the ASFRs of rural Pakistan during the year 2007 (see Table-2). The estimates of parameters of both the models are highly significant (p-value <0.001). The estimated value i.e. 2312.486 of parameter "a" designated the total fertility per thousand women with that of the modal age of motherhood i.e. 30.095 years. The estimated dispersion of the parameterized fertility schedule was 2.761. Collectively, the estimated rural model ASFR per thousand women i.e. *ab/c* was 212.154 which differed significantly from the observed rural ASFR 228.4 corresponding to the age group 25-29 of the motherhood. On the other hand, the estimate of base level fertility parameter of P-K model-I was 230.783 that is closer to the observed rural ASFR. The estimated value of the parameter μ was 24.684 years which points out the modal age of the motherhood. The estimate of the variance parameter within the fertility was approximately tripled above the modal age as compared to below the model age of the motherhood.

The parameters of Hadwiger function and P-K model-I were estimated for the ASFRs of urban Pakistan during the year 2007 (see Table-3). The estimated total fertility per thousand women was 1797.255 with that of the age of the mother i.e. 29.432 years. The estimated dispersion of the parameterized fertility schedule was 3.409 which was around 12.4% greater than the rural ASFRs of Pakistan. The P-K model-I fertility parameter estimate corresponds to the model age of mother i.e. 25.787 years was 222.008 which was more close to the observed urban ASFRs i.e. 219.9. The dispersion parameter estimates were little bit smaller than the estimates of rural ASFRs of Pakistan during the year 2007.

The estimates of fertility parameters using the ASFRs of Pakistan during the year 2007 are given in Table-4. In 2007, the total fertility per thousand women of overall Pakistan was 2113.562 at the age of the motherhood 29.86 years. The estimate of parameterized fertility schedule dispersion was 2.957 which was almost in the middle of the rural and Urban ASFRs dispersion. Similarly, the base level of fertility per thousand women using P-K model-I was 222.609 at the age 25.145 years of the motherhood. The dispersion above the model age was in the centre of rural and urban dispersion whereas the dispersion below the model age was approximately close to rural and urban below dispersion. The P-K model-I fitting to the ASFRs of Pakistan and its geographical regions i.e. rural and urban seems more suitable as compared to the Hadwiger function.

The disparity among ASFRs of Pakistan and its rural, urban regions were also manifested using Gini coefficients along with their confidence limits for some specific years 1990, 1995, 2000, 2005 and 2007. The Gini coefficients gradually increased in successive years which signify the increased variation in the fertility pattern. The Gini coefficients of the year 2007 significantly increased as compared to the earlier years. It is an indication about the fertility decline at both ends i.e. 15-19 and 45-49 years ages.

Furthermore, graphical fertility behaviour of rural, urban and of Pakistani women during the past twenty three years is presented in Figures 1-3 respectively. The fertility and fertility period of rural women are wider than the fertility of Pakistan and its urban area women. The greater fertile period seems around 20-34 years than 15-49 years among women of all three regions whereas the most fertile period of women age is about 25-29 years. It might be a sign of the improved health sciences and contraceptive measures, especially the increased female literacy rate etc. The fertility pattern of Pakistan and its rural urban regions behaves like the reciprocal of V shape.

The comparison of TFRs of Pakistan and of two major geographical regions i.e. rural and urban is given in Figure 4. The fertility pattern of Pakistan has more smooth declined than the rural TFRs whereas TFRs of urban population fluctuates greatly during the period 1986-1995. In the last one and half decade, a similar fertility reduction pattern has been visualized in the urban region as that of the rural areas of Pakistan.

The predicted ASFRs by Hedwiger function along with confidence limits of rural, urban and overall Pakistan’s ASFRs during the year 2007 are presented in Figures 5, 7, 9 corresponding to the age of the motherhood. Figure 5 demonstrates the wider confidence limits than the Figures 7 and 9 but all three figures reveal the age group 25-30 years of highest fertility. P-K model-I also predicted the ASFRs with respect to the age of the motherhood along with confidence limits of rural, urban and of overall Pakistan for the year 2007 (see Figures 6, 8, 10).These figures have almost the same pattern as that of Figures 5, 7, and 9 but with lesser variation as compared to the Hadwiger estimate. Although, different parametric models were fitted on the empirical data consist of ASFRs of Pakistan and of rural and urban areas, but for the comparison purposes, the analysis of only two highly significant parametric models i.e. Hadwiger and P-K model-1 is thoroughly discussed here. Keeping in view the behaviour of the figures of these two models, P-k model-I is proposed as a parsimonious model for the observed ASFRs of rural, urban and Pakistan during the year 2007.

The disparity among the ASFRs of 2007 is manifest through the Lorenz curves of Pakistan and its urban and rural regions respectively (see Figures 11-13). The Lorenz curve of 2007 indicates the greater variation in the ASFRs of urban area than the rural area as well as of overall Pakistan. These results are consistent with the results given in Table-1.

Although, The Government of Pakistan launched different contraceptive measures to decrease the fertility in the country. Fertility decreased within the country but not as desired by the Government of Pakistan. The least reduction has been seen in the rural areas as compared to overall Pakistan and its urban region. The reasons might be lack of educational facilities, exposure to mass media, early marriages, poor health facilities, lack of information about the family planning methods, greater influence of religious leaders over the innocent rural inhabitants as well as male dominant societies etc., of the far flung areas. On the other hand, the urban women with that of the above facilities have a good understanding about the fertility and its related issues. That is why; there is greater reduction in fertility in urban area as compared to the other areas.

4. CONCLUSION AND RECOMMENDATIONS

Although, overall fertility has decreased in Pakistan and its breakdown rural and urban regions during the year 2007 as compared to 1984 but greater decline in fertility was seen in urban population. Regarding the performance of the fitted models P-K model-I estimates are closer to the empirical rates than the estimates of Hadwiger model parameters. So, P-K model-1 is proposed as parsimonious model to graduate the ASFRs of the rural, urban and overall Pakistan. As it has an adequate and feasible interpretation of the model parameters. The Gini coefficient of the year 2007 indicates the greater inequality in urban fertility as compared to Pakistan and its rural population. It might be because of the increased mean age of child bearing which seems only due to increased female literacy rate. Keeping in view this evidence and to increase the mean age of childbearing, the Government should facilitate the young generation for higher education, eventually; the greater reduction in fertility will be achieved. Moreover, to overcome the unavailable fertility and mortality at national level, the Government should educate and encourage the community for birth deliveries in hospital through electric and print media. The Government should take serious action against the indoor deaths and deliveries owing to the untrained and novice midwives. Furthermore, the Government as well as private hospitals should bound to send the weekly report of fertility and mortality to the concerned registration department or at least one representative of registration department must be appointed in each hospital or medical superintendent (MS) must be bound to update the registration department regarding the births and deaths within the hospital under his jurisdiction.

REFERENCES

1. Bairagi, R. and Datta, A.K. (2001). Demographic transition in Bangladesh: What happened in the twentieth century and what will happen next? *Asia Pacific Population Journal.* 16(4), 3-16.
2. Chandola, T., Coleman, D. A., and Horns, R. W. (1999). Recent European fertility patterns: Fitting curves to ‘distorted’ distributions. *Population Studies*, 53(3), 317-329.
3. Federal Bureau of Statistics (2006). *Population Demographic Survey-2006*. Islamabad, Government of Pakistan.
4. Finance Division (2011). *Pakistan Economic Survey (2011-12).* Economic Adviser’s Wing, Islamabad. Government of Pakistan.
5. Gilji, E. (1969). Fitting curves to age-specific fertility rates: some examples. *Statistical Review of the Swedish National Central Bureau of Statistics*, III (7), 118-134.
6. Hadwiger, H. (1940). Eineanalytische reproductions-funktion fur biologischegesamtheiten. *SkandinaviskAktuarietidskrift*, 23, 101-113.
7. Ishida, K., Stupp, P. and Melian, M. (2009). Fertility Decline in Paraguay. *Studies in Family Planning*, 40(3), 227-234.
8. Islam, M.R. and Ali, M.K. (2004). Mathematical modeling of age specific fertility rates and study the reproductivity in the rural area of Bangladesh during 1980-1998. *Pak. J. Statist.*, 20(3), 379-392.
9. Islam, S. and Nesa, M.K. (2009). Fertility transition in Bangladesh: The role of education. *Proceeding Pakistan Academy of Sciences*, 46(4), 195-201.

10. Kabir, M. and Mosleh Ud din M. (1987). Fertility transition of Bangladesh: Trends and determinants. *Asia Pacific Population Journal*, 2(4), 54-74.
11. Kostaki, A., Moguerza, J.M., Olivares, A. and Psarakis, S. (2009). Graduating the age-specific fertility pattern using Support Vector Machines. *Demographic Research,* 20(25), 599-622.
12. Montgomery, D.C. and Peck, E.A. (2012). *Introduction to linear regression analysis.* New York: John Wiley & Sons.
13. Nasir, J.A., Akhtar, M. And Tahir, M.H. (2009). Reproductivity and age specific fertility rates in Pakistan after 1981. *Pak. J. Statist.*, 25(3), 251-263.
14. Sathar, Z.A., Crook, N., Callum, C. and Kazi, S. (1988). Women‘s status and fertility change in Pakistan. *Population and Development Review*, 14(3), 415-432.
15. Sathar, Z.A. (1993). The much-awaited fertility decline in Pakistan: wishful thinking or reality. *International Family Planning Perspectives,* 19(4), 142-146.
16. Siegel, J.S. and Swanson, D.A. (2004). *The methods and materials of demography*. Academic Press.
17. Zakria, M., Muhammad, F., Zafar, M.I., and Asif, F. (1999). Modelling the Contraceptive Behaviour of Married Females.*Pak. J. Agri. Sci*., 36(3, 4), 149-153.

Table 1:
TFR, GRR, MAC and SDAC of Pakistan and its Rural, Urban Regions

	TFR			GRR			MAC			SDAC		
Year	Rural	Urban	Pakistan	Rural	Urban	Pakistan	Rural	Urban	Pakistan	Rural	Urban	Pakistan
1984	7.27	6.24	6.95	3.53	2.93	3.33	30.15	29.75	30.05	7.18	6.99	7.13
1990	6.71	5.16	6.21	3.18	2.47	2.96	29.55	29.14	29.47	7.41	6.74	7.25
1995	5.94	4.89	5.59	2.87	2.39	2.70	29.66	29.19	29.53	7.33	6.85	7.19
2000	4.91	3.69	4.34	2.34	1.80	2.09	29.75	29.01	29.46	7.19	6.33	6.86
2005	4.09	3.29	3.79	1.94	1.54	1.79	29.83	29.59	29.75	6.99	6.17	6.69
2006	4.10	3.20	3.70	1.91	1.53	1.84	29.99	28.57	29.87	6.93	5.98	6.63
2007	4.00	3.20	3.70	1.88	1.65	1.76	29.83	29.58	29.75	6.99	6.17	6.68

Table 2:
Estimates of Hadwiger Function and P-K Model-I
for ASFRs of Rural Pakistan (2007)

Hadwiger Function			
Parameters	Estimate	Std. Error	Pr(>|t|)
a	2312.486	106.185	<0.000
b	2.761	0.147	<0.000
c	30.095	0.448	<0.000

Peristera and Kostaki Model-1			
Parameters	Estimate	Std. Error	Pr(>|t|)
c_1	230.783	3.628	<0.000
μ	24.684	0.233	<0.000
σ_{11}	4.910	0.340	<0.001
σ_{12}	14.919	0.387	<0.000
$\sigma_{(x)}= \sigma_{11}$ if $x \le \mu$, and $\sigma_{(x)}= \sigma_{12}$ if $x > \mu$			

Table-3:
Estimates of Hadwiger Function and P-K Model-I
for ASFRs of Urban Pakistan (2007)

Hadwiger Function			
Parameters	Estimate	Std. Error	Pr(>|t|)
a	1797.255	31.224	<0.000
b	3.409	0.068	<0.000
c	29.432	0.131	<0.000

Peristera and Kostaki Model-1			
Parameters	Estimate	Std. Error	Pr(>|t|)
c_1	222.008	4.974	<0.000
μ	25.787	0.482	<0.000
σ_{11}	4.810	0.634	<0.005
σ_{12}	11.373	0.509	<0.000
$\sigma_{(x)}= \sigma_{11}$ if $x \le \mu$, and $\sigma_{(x)}= \sigma_{12}$ if $x > \mu$			

Table-4:
Estimates of Hadwiger Function and P-K Model-I
for ASFRs of Pakistan (2007)

Hadwiger Function			
Parameters	Estimate	Std. Error	Pr(>|t|)
a	2113.562	66.535	<0.000
b	2.957	0.108	<0.000
c	29.860	0.281	<0.000

Peristera and Kostaki Model-1			
Parameters	Estimate	Std. Error	Pr(>|t|)
c1	226.609	3.341	<0.000
μ	25.145	0.251	<0.000
σ11	4.931	0.363	<0.001
σ12	13.580	0.342	<0.000
σ (x)= σ11 if x ≤ μ, and σ(x)= σ12 if x >μ			

Table 5:
Gini coefficients & 95% C.I. of ASFRs of Pakistan and its Rural-Urban breakdown during 1990-2007

Years	Rural	Urban	Pakistan
1990	0.3594 (0.2540-0.5286)	0.4492 (0.3239- 0.5982)	0.3827 (0.2809- 0.5490)
1995	0.3828 (0.2917- 0.5189)	0.4499 (0.3431- 0.6102)	0.4018 (0.3044- 0.5569)
2000	0.4015 (0.2553- 0.4974)	0.4962 (0.3369- 0.6664)	0.4403 (0.2959- 0.5685)
2005	0.4346 (0.2816- 0.5756)	0.5152 (0.3029- 0.6916)	0.4612 (0.2970- 0.6109)
2007	0.4342 (0.2638- 0.5915)	0.5361 (0.3939- 0.7048)	0.4678 (0.3131- 0.6249)

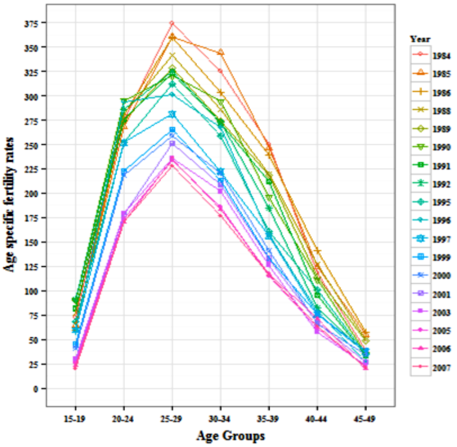


Fig. 1: ASFRs Trend of Rural Pakistan during 1984-2007

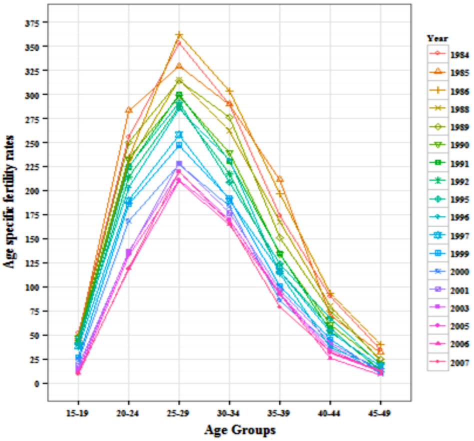


Fig. 2: ASFRs Trend of Urban Pakistan during 1984-2007

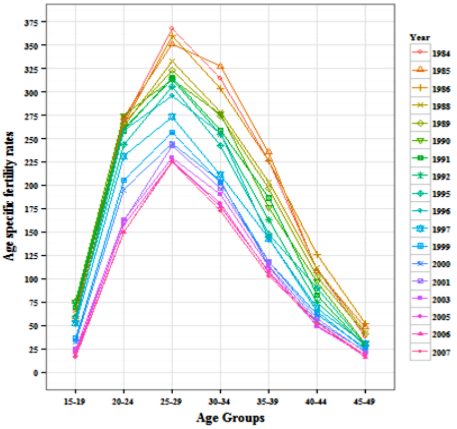


Fig. 3: ASFRs Trend of Pakistan during 1984-2007

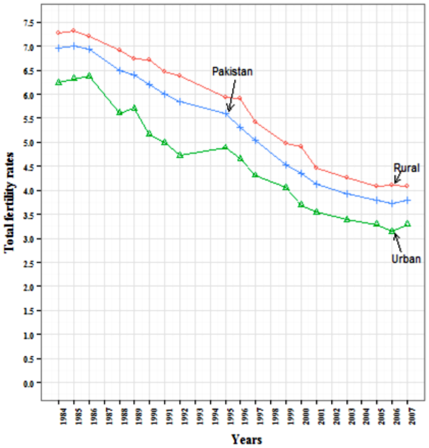


Fig. 4: Trend of TFRs of Rural, Urban and Pak. 1984-2007

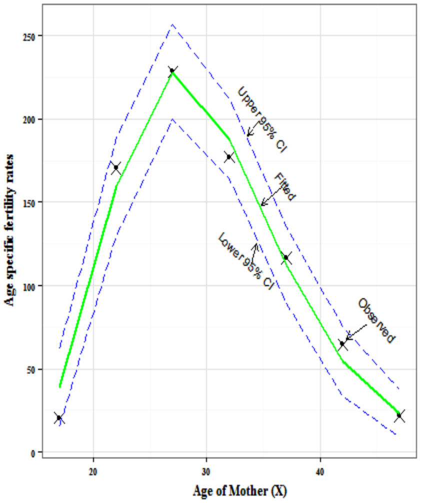


Fig. 5: ASFRs trend of Rural Pak. using Hadwiger Function

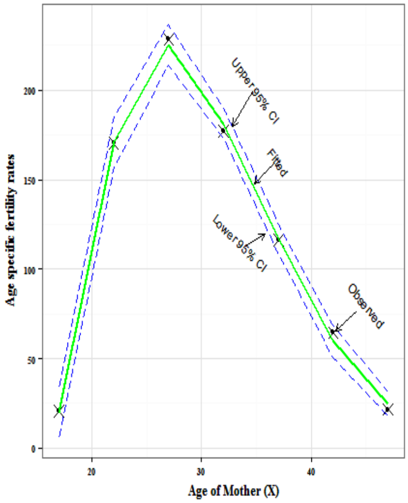


Fig. 6: ASFRs trend of Rural Pak. using P-K model-I

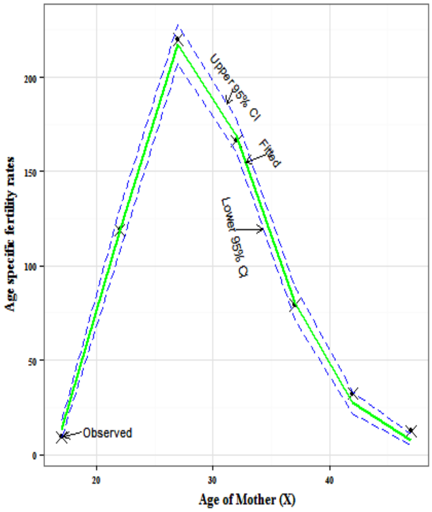


Fig. 7: ASFRs trend of Urban Pak. using Hadwiger Function

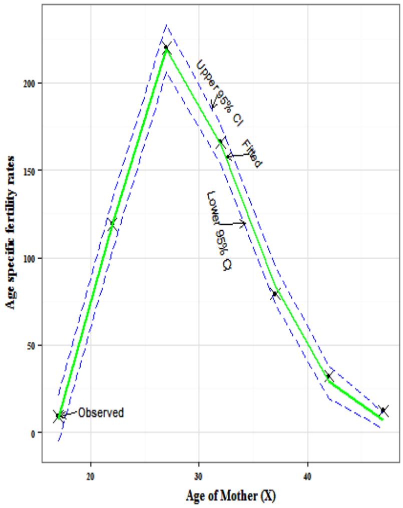


Fig. 8: ASFRs trend of Urban Pak. using P-K model-I

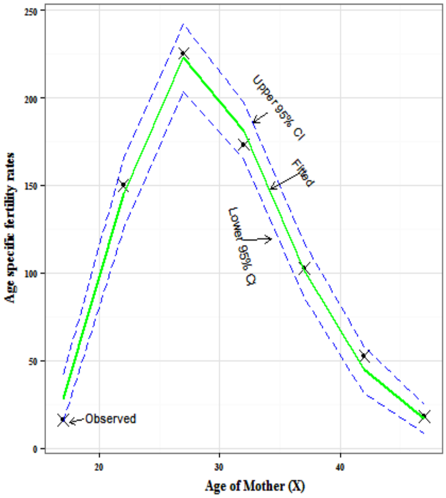


Fig. 9: ASFRs trend of Pakistan using Hadwiger Function

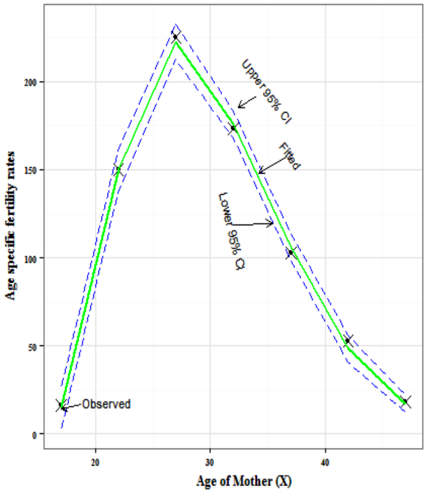


Fig. 10:ASFRs trend of Pakistan using P-K model-I

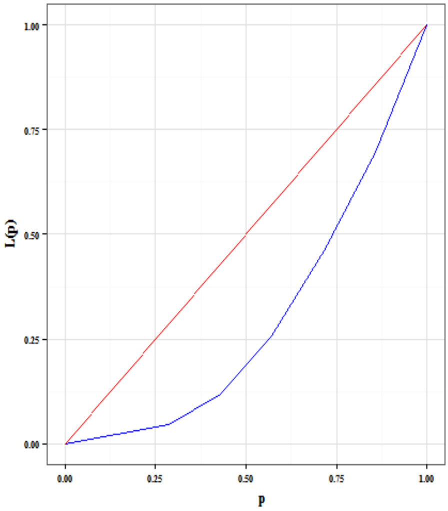


Fig. 11: Lorenz curve of ASFRs of Pakistan during 2007

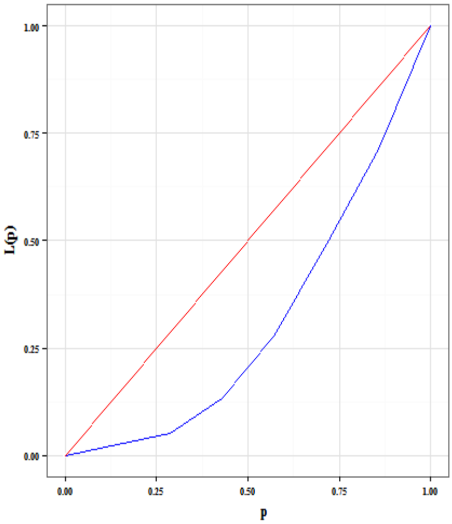


Fig. 12: Lorenz curve of ASFRs of Rural Pak. during 2007

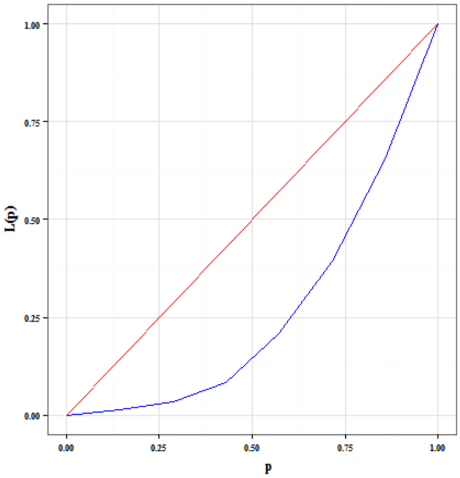


Fig. 13: Lorenz curve of ASFRs of Urban Pak. during 2007

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 619-624

PHASE ZERO INTRON INTERRUPT THE CODING SEQUENCES IN DNA BY CYCLIC PERMUTATION

Naila Rozi[1] and **Nasir Uddin Khan**[2]

[1] Sir Syed University of Engineering & Technology,
Karachi, Pakistan. Email: nrozi@yahoo.com

[2] University of Karachi, Karachi, Pakistan
Email: drkhan.prof@yahoo.com

ABSTRACT

The phase of an intron refers to the position at which it interrupts a coding DNA sequences. There are three types of intron. Phase zero intron interrupt the coding sequences between adjacent codons. The internal exon can be classified into various groups, depending on the phase of the two flanking introns. Exons where the total number of nucleotides is exactly divisible by three will fall into three groups (0,0:1,1: and 2,2) Exon duplication with a gene and exon shifting between genes involves such exon because when inserted they do not alter the translational reading frame unlike the other six exon group(0,1;0,2;1,0;1,2;2,0; and 2,1).

INTRODUCTION

An important component of the exon theory of genes was the idea that exons in polypeptide –encoding genes represented the functional or structural unit.

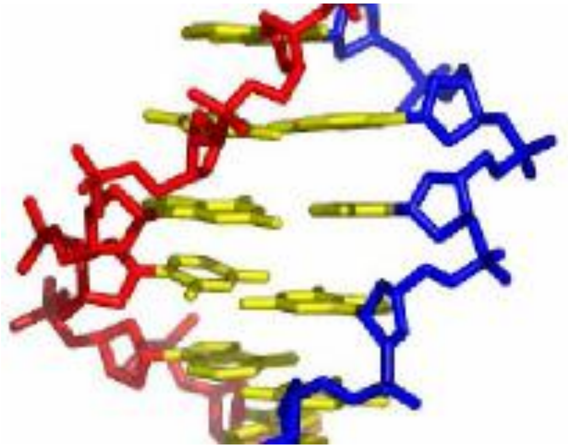


Fig 1: Building block unit of DNA

The exon theory of genes was supported by the apparent conservation. The exon theory of gene are the descendants of ancient time mingenes and introns are the descendants of the space between then which were present in primordial cells. By an exon database from available protein sequences and searching for homologous exons, Dorit et al. 1990 estimated that only 1000-7000 different exons were needed to construct all proteins

HYPERVARIABLE DNA

Hypervariable DNA Sequences are highly polymorphic and are organized in over 1000 arrays (from 0.1 to 20kb long) of short tandom repeats. The repeat unit in different hyper variable array vary considerably in size, but share a common core sequences. GGGCAGGAXG(where X is any nucleotide which is similar in size and in G content to the chi sequences a signal for generalized recombination in E.Coli.DNA sequence occur at other chromosomal locations.

SEQUENCE TAGGED SITE

Sequence tagged sites are important mapping tool simply because the presence of that sequence can be assayed very conveniently by polymer chain reaction (PCR). Most STS are non-polymorphic. An example of how an STS is developed from a DNA sequence is shown below

Rough sequence: 200 nucleotides

Primers: Chosen from underlined sequences, both 16 nucleotides long.

A(Forward primer in bold, identical to the sense sequence from +50 to +66),

B(Reverse primer, in bold, corresponding to ant sense strand for the sequence from + 175 to + 190.

The sequence from +1 to +50 and from +191 to +200 are extremely CG rich with multiple runs of G and C, which are not optimal sequence for designing PCR primers. Hence decision to use internal primers.

STS: 141 nucleotides defined by primers ends +50 to +190.

1 40

5’ CCAGCGGC CCGCGGCGCA GGGGCCCGGC GGGGCCTGG

41 5’_PrimerA →3’ 80

GGCCGCCCGG CAGTGAGCAT CAGATTCAGA ACCTAGACGsA

80 120

ACCTAGGACC AGTACCTACA AGGTACTCTA GATGATCTAT

121

The fidelity of DNA replication in vivo is extremely high during replication of human genomes. For example only one base in about 3x10^9 is copied in correctly. Misincorption occur at low frequency dependent on the relative free energies of correctly and incorrectly paired base Very minor changes in helix geometry can stabilize G-T base pair. In vivo copying normally shows an error rate much lower than these thermodynamics limitation would imply. This is achieve by proofreading mechanism, one of which is a common property of DNA polymerases. Unlike RNA polymerase, DNA polymerase require absolutely the 3'hydroxyl end of a base paired primer strand as substrate for chain extension.

CYCLIC PERMUTATION

Tendom trinucleotides repeats are not in frequent in the human genome. Although there are 64 possible trinucleotides sequences when allowance is made for cycle permutations $(CAG)_n$=$(AGC)_n$= $(GCA)_n$ and reading from either strand [5'$(CAG)_n$ on the one strand =5' $(CTG)_n$ on the other], there are only 10 different trinucleotides repeats.

AAC/GTT
AAG/CTT
AAT/ATT
ACC/GGT
AGG/CCT
ATC/GAT
CAG/CTG
CCG/CGG

The 10 possible trinucleotides repeats are cyclic permutationof one or another of these10. From the relation to unit fraction, it can be shown that cyclic numbers are of the form

Table 1
A simple procedure of multiplication get 8mapping,thus a part of this cyclic permutation any two permutation acting on mutually.

*	A	G	C	T	SEQUENCE
A	AA	AG	AC	AT	AAAGACT
G	GA	GG	GC	GT	GAGGGCT
C	CA	CG	CC	CT	CACGCCT
T	TA	TG	TC	TT	TATGTCT
Sequence	AAGACAT	AGGGCGTG	ACGCCCTC	ATGTCTTT	8permutation

This cyclic permuatation can be write as follows.

$$(b^{p-1}-1)/p \qquad (1)$$

where b is the number base 10 for decimal, and p is a prime that does not divide b prime p that give cyclic number are called full reputed primes or long primes. The case b=10, p=7 gives cyclic number 142857. Not all values of p will yield a cyclic number using this formula if p=13 gives 076923076923.These failed cases will always contain a repetition of digits(possibly several). The first value of p for which this formula produces cyclic numbers in decimal are sequence A001913.

The member of some gene families may not very obviously related at the DNA sequences level, but nevertheless encoded gene products that are characterized by common General function and the presence of very short conserved sequences. In some types of gene family, the gene encoded products that are known to be functionally related in a general sense and show only vey week sequences homology over a large segments without very significance conserved amino acid as in DEAD box below.

a) DEAD box:
NH$_2$______ AXXGXGKT$^{22-42}$_PTRELA$^{19-29}$____GG$^{17-29}$______TPGR$^{17-23}$_______
DEAD$^{19-51}$________SAT$^{115-192}$_____ARGXD$^{20-25}$______HRIGR___OOH

b) WD repeat ($^{6-94}$______GH$^{23-41}$_______WD) n=4-8
Fig(a)Gene families

In the figure

(a) DEAD box family the gene family encoded products implicated in cellular process involving alteration of RNA secondary structure such translation initiation and spicing Eight very highly conserved amino acid motif are evident including the DEAD box.

(b) In WD repeat family the gene family encoded products are involved in a variety of regulatory function such as regulation of cell division transcription. Eight tandom repeat containing a core sequence of fixed length.

IDENTIFYING CODING DNA

From the outset of the Human Genome Project there has been much debate over whether to go for an all out assault of in discriminate sequencing of all 3 billion base of the human genome. Tandom trinucleotides repeat are not infrequent in the human genome. Although there are 64 possible trinucleotides sequences. when allowance is made for cyclic permutations(CAG)$_n$=(AGC)$_n$=(GCA)$_n$ and reading from either strand [5'(CAG)$_n$ on one strand=5'(CTG)n on the other], there are only 10 different trinucleotides repeats. Most of them are known as usefully polymorphic microsatellite markers but in addition certain repeats of CAG/CTG and CCG/GGC show anomalous behavior.

Several genes contains (CAG)$_n$ repeats within the coding sequences, translated a polyglutamine tracts in the protein product. Typically, pthalogical all stable and non-pathological allels have 10-30 repeats, while unstable pathological have modest expansions, often in range of 40-100 repeats.

COMPETITION HYBRIDIZATION AND COT-1 DNA

Competition (or suppression) hybridization involves blocking a potentially strong repetitive DNA signal which can be obtained when using a complex DNA probe. The labeled probe DNA is denatured and allowed to re-associate in the presence of unlabeled total genomic DNA in solution, or preferably a fraction that is enriched for highly repetitive DNA sequences. In either case, the highly repetitive DNA within the unlabeled DNA is present in large excess over the repetitive elements in the labeled probe. As a result, such sequences will readily associate with complementary strands of the repetitive sequences within the labeled probe, thereby effectively blocking their hybridization to target sequences. Instead of using total genomic DNA as a blocking agent in hybridization, it is more effectively to use a fraction of total genomic DNA that is enriched for highly repetitive DNA sequences, such as the *Alu*, LINE-1 and repeats o human DNA. From human DNA, and other mammalian DNA where the genome size id much the same as that of the human genome, the latter usually involves preparing a fraction of DNA known as COT-1 DNA. Total purified human genomic DNA is solicited to an average length of about 400bp, denatured by heating, then allowed to renature in 0.3 M NaCl at 65°C at a starting concentration of χ moles of nucleotides per liter for time of t sec, where χt =1.0.

CONCLUSION

- Several genes contain (CAG) repeats within the coding sequence, translated as polyglutamine tracts in the protein product. Typically, the stable and non-pathologic alleles have 10-30 repeats, while unstable pathological alleles have modest expansions, often in the range of 40-100 repeats. Transcription and translation of the gene are not affected by the expansion.
- Some (CGG) repeats in non-coding sequences can expand massively from a normal copy number of 10-50 up to hundreds or thousands of repeats. By unknown means, the expanded repeats affect DNA methylation and chromatin structure, producing inducible chromosomal fragile sites.
- Uniquely, a (CTG) repeat in the 3’ untranslated region of the myotonic dystrophy kinase gene(DMK) at 19q13 has 5-35 repeats units in normal people, but up to 2000 units in people with myotonic dystrophy (DM, MIM 160900). There is perfect correlation between the repeat expansion and the disease, even though the repeat has no evident effect on transcription or the structure of the gene product.

REFERENCES

1. Naila Rozi and Nasir Uddin Khan (2009). Geometric Patterns of DNA Help in Signal Processing. *2nd ABRC 2009 Conference of Comsat Institute* Lahore on 22-23 October, 2009.
2. Naila Rozi and Nasir Uddin Khan (2011). Termination of DNA replication and role of enzymes in recombination. *Journal of Life Sciences*, U.S.A (ISSN-1934-7391), 5(2).
3. Naila Rozi and Nasir Uddin Khan (2011). Improve Hidden Markov Model In DNA Sequences HMM). *Journal of IJCRB*, 2, 528. ISSN# 2073 7122.
4. Naila Rozi and Nasir Uddin Khan (2009). Number Patterns Geometric representation of DNA. *2ND ICECS Conference, Proceeding IEEE*, DUBAI, U.A.E, ISBN 978-7695-3937-9, 343-344.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 625-632

DETERMINANTS OF MALE PARTICIPATION IN REPRODUCTIVE HEALTHCARE SERVICES: A CROSS-SECTIONAL STUDY

Md Shahjahan
Bangladesh Institute of Health Sciences,
Birdem 125/1, Darus Salam, Mirpur, Dhaka-1216, Bangladesh
Email: mdshahjahan@agnionline.com

ABSTRACT

Background: The role of male's participation in reproductive healthcare is now well-recognized. The present study investigated the role of men in some selected reproductive health issues, characterizing their involvement, including factors influencing their participation in reproductive healthcare.

Methods: This study was conducted in the working areas of urban and rural implemented by NGOs. The sample-size was determined scientifically. The systematic sampling procedure was used for selecting the sample. The study included 615 men aged 00-00 years. Bivariate analysis was performed between male's involvements as the dependent variable with several independent variables. Logistic regression analysis was applied to assess the effects of risk factors on the participation of men in reproductive health.

Results: The mean age of the respondents was little over 34 years while their mean years of schooling was 3.7, and their mean monthly income was about Tk 3,400 (US$ 1=Tk 70 at the time of the study. Rickshaw-pulling and driving was main the occupation of the respondents from the urban while farming were main the rural area respectively. About two-thirds of the respondents discussed reproductive health issues with their wives and accompanied them to healthcare facilities. The current contraceptive-use rate was 63% among the men who attended the evening clinics. Results of bivariate analysis showed a significant association with education, occupation, income, access to media, and number of living children. Results of logistic regression analysis showed that secondary to higher education level, number of living children, paid employment status, long marital duration, and access to media were important correlates of males' involvement in reproductive healthcare services.

Conclusions: The results imply that a greater integration of reproductive healthcare matters with the Millennium Development Goals and increasing perception of men through enrollment in various components of reproductive activities will produce synergistic effects.

KEYWORDS

Cross-sectional studies; Male participation; Reproductive health; Bangladesh.

1. INTRODUCTION

The Program of Action of the ICPD clearly set a new agendum when it emphasized on male's responsibilities and participation in reproductive healthcare [1]. Although consensus was reached on involvement of men in reproductive health, and the policy environments generally support that notion in many countries including Bangladesh, healthcare service for reproductive health is still largely female-oriented. The reproductive health programmes have traditionally focused on women and the exclusion of men. However, results of recent studies revealed that men might serve as gatekeepers to women's access to reproductive health services significantly [2].

The concept of reproductive healthcare is that men, women, and young people have the right to be informed and have access to safe, effective, affordable, and acceptable reproductive healthcare services [3]. Although there is a tendency to overlook the relevance of men in matters relating to reproductive healthcare, they have substantial reproductive health influence. So, reproductive health in its broader sense should be a concern for all, not just that of women.

A large number of articles [4-6] and the growing number of conferences, research projects, and debates on this subject bear testimony to the importance of this issue, both from the programmatic point of view and as a process for bringing about a gender balance in men's and women's reproductive rights and responsibilities. This renewed interest in male's involvement is not unconnected with the HIV/AIDS pandemic that has spurred an intense interest in the promotion of condom-use and the need to address men's roles in the abuse of reproductive rights and sexual violence directed towards the female partners and relatives. Effective family planning is important in spacing childbirth so that both mother and child can gain the maximum quality of life, especially the mothers at high -risk. Birth spacing will also give the mother ample time to recuperate from her previous pregnancy.

Men, especially in Africa, are dominant and are the major decision-makers in family affairs, including reproductive healthcare matters [7]. The dominance of male in this respect is reinforced by the cultural institution of patriarchy, religion, and the economic power that men wield. Ezeh reported that, in Ghana, spousal influence in respect of reproductive goals, rather than being mutual or reciprocal, is an exclusive right exercised only by the husband [8]. In Ilorin, Nigeria, one of the major reasons for not adopting modern contraceptive method by women is the husband's resistance [9]. In northern Nigeria, women cannot practise family-planning method without the formal consent of their husbands [10].

Reasons for involving men in reproductive health matters are multifaceted. First of all, men have their own reproductive health concerns and their involvement should not be seen only as a means to achieve women's better reproductive healthcare. Second, men's sexual and reproductive well-being and behaviours directly affect their partners. Third, decisions on the matters of reproductive healthcare occur within relationships that affect both men and women [11]. The involvement of men in reproductive healthcare matters should be seen as an important measure for achieving the MDGs that include the reduction of maternal mortality and reducing the prevalence and impact of HIV/AIDS.

In the present context of Bangladesh, involving men and bringing positive influences in reproductive healthcare services are the crucial aspects of enhancement of couples' reproductive healthcare services. Therefore, identification of demographic variables relating to men's involvement and the contribution of different influencing factors to demographic change would help formulate future policies for achieving the demographic target through men's involvement. In this paper, an attempt has been made to assess the relationship between the level of men's involvement and the demographic variables in order to measure the contribution of different factors for increasing the participation of men in reproductive activities.

2. METHODS

The concept of male's involvement in reproductive health care services can be considered as the main aspect for the success of family planning program in Bangladesh. Although there is a strong argument that socioeconomic and demographic differentials, such as education, occupation, access to information, number of living children, income, and other related issues, influence the involvement of men in reproductive healthcare services, there are other factors, such as behavioural factors, that directly affect the involvement of men. Towards ensuring an effective participation of males in reproductive healthcare services, spousal communication of men, accompanying wife during visits to clinics, and their delivery care are essential preconditions of male's involvement. Among the spousal communicating men, those who are visiting clinics with their wives, and of them, those attending delivery care are considered that they participated or were involved in reproductive healthcare services.

This cross-sectional study was carried out among males who visited some selected NGOs working in both urban slums and rural areas of Bangladesh. Married males who attended an evening clinic constituted the sampling frame. In total, 615 men were randomly selected for the study. The sample-size was determined using the statistical cluster sampling technique. The cluster was NGO evening clinics. Six study sites were randomly selected from NGOs working in urban slums and rural areas located in Agargoan (Dhaka), Narayanganj, Narsingdi, Tangail, Narail, and Gaibandha. From each of these six sites, at least 100 men were interviewed employing a systematic sampling technique. A pretested structured questionnaire was used for collecting information on sociodemographic characteristics, cultural factors, and on the usage of family-planning methods.

Data were analyzed using the SPSS for Windows software (version 17). The associations between the variables were measured using the appropriate statistical techniques. Bivariate analysis was performed between male's involvement as the dependent variable and each independent variable. Linear logistic regression analysis was done to determine the factors affecting men's participation in reproductive healthcare services.

3. RESULTS

The distribution of the socioeconomic and demographic characteristics of the respondents is shown in Table 1. The mean age of the respondents was 34 [standard deviation (SD]±7.6)] years. Forty-four percent of the respondents had no education. The mean years of schooling of the respondents were 3.7 (SD±4.1). Rickshaw-pulling and driving were the primary occupations of the men living in the urban slums, followed by business, monthly salaried job, and day laborer. The mean income was Tk 3,438 (US$ 1=Tk 70), and the mean land holding was 37.5 decimals. In the case of access to media, around 34% of the men had no access to any media, 35% had access to one, 21% to any two, and 10% had access to all three media, such as newspapers, radio, and television (TV).

The distribution of respondents' inter-spousal communication is shown in Table 2. More than half (58%)of the husbands accompanied their wives during visit s to clinics. Two-thirds of the husbands discussed about reproductive healthcare issues with their wives. Most (95%) couples were approving any family-planning methods. The proportion of couple currently using any contraceptive method was 63.1%.

Table 3 presents the results of bivariate analysis between the male's involvement and the demographic variables. A significant association was found among education ($p<0.001$), occupation ($p<0.001$), income ($p<0.001$), access to media ($p<0.001$), and number of living children ($p<0.003$).

Table 4 shows the results of logistic regression analysis, which was performed for identifying the factors affecting the involvement of males in reproductive healthcare services. The fitness of model was significant; chi-square was 81.472 ($p<0.001$), and -2Log likelihood was 604.049. Men having secondary and higher-level education were more likely to be involved in reproductive healthcare services than men who had no or primary education. Men who were paid employees were more likely to be involved in reproductive healthcare services compared to farming professionals. Men who had two children had higher odds of involvement in reproductive healthcare services than those with no or one children whereas men having three or more children had lower odds of involved in reproductive healthcare services. Men with the marital duration of 5-10 years were significantly more likely to be involved in reproductive healthcare services than those with the marital duration of less than five years. Men who had access to media were more likely to be involved in reproductive healthcare services than their counterparts.

4. DISCUSSIONS

In reproductive health matters, most people reviewed women as the target group, and little attention is given to the role of men. However, in patriarchal society where decisions are largely made by men, the needs to include them in all matters that require joint spousal decisions are crucial to achieving the reproductive health goals. This paper aimed to determine the factors that influence the involvement of males in reproductive healthcare. The results of the study revealed that when men had a higher level of education, their involvement in reproductive healthcare was more. Men are more exposed to radio, TV, newspapers, and diversified personal communication than women as men generally have more free time, more education, more disposable income, and, in many cultures, more freedom of movement than women [12]. Men who have exposure on mass media exposure have effects in changing their attitudes to use of family planning and their spousal communications improve. This exposure of men obviously increased contraceptive-use and following the use of mass media has other behavioral change [13].

The results of the study documented that the majority 66% of the men discussed the reproductive health-related matter with their wives and accompanied their wives for seeking reproductive healthcare services. The proportion of couple currently using any contraceptive methods was 63%. The results also revealed that most married couples were approving family-planning methods in achieving reproductive health benefits.

The results of logistic regression analysis showed that men having secondary and higher level of education, those were paid employees and involved in business activities, and had access to media were more likely to be involved in reproductive healthcare services. Again, those who had two children were more likely to be involved in reproductive healthcare services probably because their wives would be at greater risk of health hazards in bearing another child. These were important correlates of male’s involvement in reproductive healthcare services. The above results are consistent with those of a similar study in Mambwe district in Zambia [13], which relates increased access to mass media and counseling by health care providers help receive to information, knowledge, and awareness facilitates good choices.

The participation of men in reproductive health issues leads to a better understanding between husband and wife [14, 15], and it reduces the number of unwanted pregnancies and the unmet need for family planning [16]. The results of the study also showed that only one-fourth of the men were involved in reproductive healthcare services, which indicates the lower level of men participation. The results indicate that the low level of male’s involvement in reproductive healthcare services are not attributed to national, regional and international experiences [17] where the need for women’s healthcare issues could draw more attention in today’s world. In this study, we have found that men with the marital duration of 5-10 years were significantly more likely to be involved in reproductive healthcare services compared to their counterparts. The results support the argument that “it is the male who takes major decisions and imposes their views/choices on females including basic health services”. Again, the increase in the marital duration increases men’s participation in reproductive healthcare services, which is also consistent with the result of the study.

The results of the study also documented that little over one-third of the respondents had access to media. So, it is expected that men in Bangladesh is likely to have misinformation and require factual information from relevant sources to promote and ensure men’s participation in reproductive health services. To overcome this problem, media can play a key role in promoting participation of men in reproductive healthcare services. Finally, the findings of the study suggest that more pragmatic and target-oriented programmes will still be required to increase the involvement of in reproductive health matters.

5. COMMENTS AND CONCLUSION

The findings of the study revealed a lower level of men’s participation in reproductive health services in Bangladesh, which suggests that a greater integration of reproductive health matters with the MDGs is still needed. Increasing the perceptions of men through involvement in various reproductive activities is believed to produce synergistic effects.

The involvement of the growing private sector in providing men-friendly reproductive health services also deserves due attention (18).

One of the findings of the study indicate that their wives reproductive-related matter with is critically important to design programs involving men. This may foster an environment where contributions of men to reproductive healthcare decisions are valued and desired while protecting the rights and integrity of women. Another important aspect is inter-spousal communication as women bear the major health hazards during pregnancy, which should be shared by men as well. To protect and ensure the safety of women's lives, it is indeed required to increase men's participation in the area of reproductive life of women. It is, therefore, a very crucial area that needs continuous strengthening and increasing male's participation in reproductive health services to reduce the maternal pregnancy- related risks.

REFERENCES

1. Odu, O.O., Jadunola, K.T.I. and Parakoyi, D.B. (2005). Reproductive behaviour and determinants of fertility among men in a semi-urban Nigerian community. *Community Primary Health Care*, 17(1), 13-19.
2. Reproductive Health Outlook, PATH (RHO). (2003). Men and reproductive health. [http://www.rohtml/menrh.htm].
3. Kaushalendra, K.S., Shelah, S.B. and Amy, O.T. (1998). Husbands' reproductive health, knowledge, attitudes and behavior in Uttar Pradesh, India. *Stud Fam Plann*, 29(4), 388-399.
4. Estborn, B. (1995). Gendering men shared concern. Women's empowerment base. Planning and sexual health. Technical Report, No. 28.
5. United Nations Population Fund. Male involvement in reproductive health, including family [http://www.qweb.kvinnoforum.se/papers/maleinvolv.html]
6. Khan, M.E., Khan, M.I. and Mukerjee, N. (1997). Men's attitude towards sexuality and their sexual behaviour: observations from rural Gujarat. *In Proceedings of the National Seminar on Male Involvement in Reproductive Health and Contraception*: April 30 - May 2, Baroda.
7. Berhane, Y. (2006). Male involvement in reproductive health *Ethiopian J Health Dev*, 20(3), 135-136.
8. Ezeh, A.C. (1993). The influence of spouses on each other's contraceptive attitudes in Ghana. *Stud Fam Plann*, 24(3), 163-174.
9. Fakeye, O. and Babaniyi, O. (1989). Reasons for non-use of family planning methods at Ilorin, Nigeria; male opposition and fear of methods. *Trop Doct*, *1*, 114-117.
10. Central Statistics Agency [Ethiopia] and ORC Macro. (2005). IMTIAZ Ethiopia Demographic and Health Survey. Addis Ababa, Ethiopia and Calverton, Maryland, USA: Central Statistical Agency and ORC Macro.
11. Kaushalendra, K.S., Shelah, S.B. and Amy, O.T. (1998). Husbands' reproductive health, knowledge, attitudes and behavior in Uttar Pradesh, India. *Stud Fam Plann*, 29(4), 388-399.
12. Dudgeon, M.R. and Inhorn, M.C. (2004). Men's influences on women's reproductive health: medical anthropological perspective. *Soc Sci. Med.* 59, 1379-1395.

13. Haile, A., Enqueselassie, F. (2006). Influence of women's autonomy on couple's contraception use in Jimma town, Ethiopia. *Ethiop J Health Dev*. 20(3), 145-151.
14. Berhane, Y. (2006). Male involvement in reproductive health. *Ethiop J Health Dev*, 20(3), 103-205.
15. David, OO., Akinrinola, B. (1994). The impact of mass media family planning promotion on contraceptive behavior of women in Ghana. *Popul Res Policy Rev*, 13(2), 161-177.
16. Tshibumbu, D. (2006). Factors influencing men's involvement in prevention of mother-to-child transmission (PMTCT) of HIV programmes in Mambw district, Zambia. Pretoria, University of South Africa.
17. Helzner, JF. (1996). Men’s involvement in family planning. *Reprod Health Matters* 7, 146-154.
18. Mullick, S., Kunene, B., Wanjiru, M. (2005). Involving men in maternity care: health service delivery issues. *Agenda Special Focus*, 124-135.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 633-646

A SURVEY OF R SOFTWARE FOR PARALLEL COMPUTING

Esam Mahdi
Department of Mathematics, Islamic University of Gaza, Palestine
Email: emahdi@iugaza.edu.ps

ABSTRACT

This article provides a summary of a selection of some of the high-performance parallel packages (libraries) available from the Comprehensive **R** Archive Network (CRAN) using the statistical software **R**[1]. These packages can utilize multicore systems often found in modern personal computers as well as computer cluster or grid computing in order to provide linear speed up the computations in many of the advanced statistical modern applications. Some illustrative **R** parallel codes are given in order to introduce the reader to some basic ideas about parallel programming in **R** packages.

KEYWORDS

R, high performance computing, network of workstations, message passing interface, parallel computing, computer cluster, grid computing, multicore systems.

1. PRELIMINARY INTRODUCTION

Many of the recent computational statistical analysis involve advanced algorithms with massive datasets and large numbers of parameters need to be estimated. In particular; DNA sequence analysis in bioinformatics (Vera, Jansen and Suppi 2008), bootstrap and Monte-Carlo simulations in multivariate time series analysis (Mahdi and McLeod 2012), neural network (Seiffert 2002), and cross-validation are computationally very intensive. Researchers have more and more need for parallel computation in order to speed up the computational bottlenecks. This can accomplish by dividing the calculations into smaller tasks and distributing them in simultaneous processing of multiple systems in order to obtain the statistical results much faster than those obtained based on the sequential computation procedures.

In general, the parallel computing aims to do three things: splitting the calculations into chunks, executing the chunks simultaneously in parallel mode using the so called slaves nodes (alternative name is workers), and combining the results back together to the so called master node. In other words, the algorithm of the parallel computing illustrated in Figure 1 bellow can be described into steps:

Step 1: Set up the cluster by initializing the master node and the worker processes.

Step 2: Export all needed variables/objects and data to all slaves.

[1] http://cran.r-project.org/

Step 3: Do the needed parallel calculations using calculation functions. Repeat as many times as needed. For example 1000 repetitions on a cluster of 10 CPUs will be distributed so that each slave simultaneously deals with 100 repetitions.

Step 4: Wait for all slaves to complete their tasks and then to send the results back to the master node.

Step 5: End the parallel execution and shut down the worker processes.

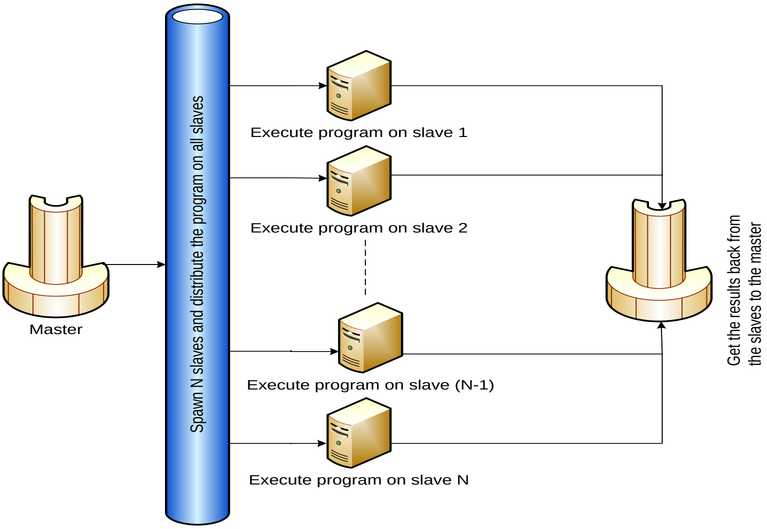

Spawn N slaves and distribute the program on all slaves
Execute program on slave 1
Execute program on slave 2
Master
Execute program on slave (N-1)
Execute program on slave N
Get the results back from the slaves to the master

Figure 1: Schematic diagram illustrating the algorithm of a parallel program.

Super-fast computers with multiple processors are useful for high-performance computing as they have the ability to provide a linear speed up with respect to the number of processors. Basically, high-performance computing can be achieved using: a computer cluster (Bader & Pennington, 2001) or a Grid computing (Rossini, Li, & Tierney, 2007) or a workstation personal desktop/laptop with multicore systems.

A cluster constitutes single machines, called nodes, in which the worker processes communicate and managed by a master node using the standardized Message Passing Interface (MPI)[2] or Parallel Virtual Machine (PVM)[3] or network sockets (SOCKETS) (Saltzer, Clarrk, Romkey, & Gramlich, 1985). In this sense, a personal computer with multicore/CPU systems may be seen as a computer cluster where each core/CPU behaves like a node in the cluster.

[2] http://www.mpi-forum.org/
[3] http://www.csm.ornl.gov/pvm/

MPI and network SOCKETS are more popular than PVM and have become the standard in parallel computing. MPI was designed by a group of researchers from academia and industry to function on a wide variety of parallel computers such as Beowulf clusters[4] (Sterling, et al. 1999). It can be implemented by some software such as MPICH/MPICH2[5], LAM/MPI[6], and OpenMPI[7]. Most of the implementations use C or C++ or FORTRAN in order to run the computations in a parallel mode.

A network socket is defined as an endpoint of an inter-process communication flow across a computer network. The popular communication between computer sockets is based on the Internet Protocol such as TCP/IP. The term socket refers to an entity that is uniquely identified by the socket number called IP address.

Lack of knowledge about MPI, SOCKET, C, C++, and FORTRAN could be a barrier for lots of statisticians who are dealing with intensive statistical computations and trying to speed up the calculations by implementing parallel algorithms.

The **R** software as a free high quality open source project has established itself as the choice of many researchers in such cases. It is a high level programming language in which some packages listed on the Comprehensive **R** Archive Network (CRAN) provide the communications layer required for interfaces high-performance parallel computing.

R itself does not do the parallel and users do not have to know C or FORTRAN in order to run parallel jobs. They need to know how to setup and configure their computers (cluster or grid or personal multicore workstation desktop or laptop) in order to implement the parallel computing using **R**.

A good introduction to parallel programming for statistical purposes can be found in Rossini, Tierney, & Li, (2007) and in Sevcikova, (2004). Eexcellent instructions to install and run **R** parallel package **Rmpi** under Linux and Windows are given by the maintainer of this package[8]. Knaus, (2010) breifly discussed the requirements for the **R** parallel packages **snowfall** and **snow**.

In Section 2 we provide a summary of a selection of some of the high-quality published computational **R** packages that can utilize multicore/CPUs often found in modern personal computer as well as computer cluster or grid computing. We stress on the most widely used **R** parallel packages: **Rmpi**, **snow**, **snowfall**, and **parallel**. In Section 3 we briefly compare between some of the contributed loop functions in **R**. An introduction to parallel bootstrapping and Monte-Carlo simulations is given in Section 4 and simple example with illustrated **R** parallel codes is given in Section 5.

[4] http://www.beowulf.org/
[5] http://www.mcs.anl.gov/research/projects/mpich2/
[6] http://www.lam-mpi.org/
[7] http://www.open-mpi.org/
[8] http://www.stats.uwo.ca/faculty/yu/Rmpi/

2. SOME CONTRIBUTED PARALLEL PACKAGES TO CRAN

We provide a summary of high performance computing **R** packages that might be of most general of research interest, **rpvm**[9], **Rmpi**[10], **nws**[11], **snow**[12], **snowfall**[13], **foreach**[14], **multicore**[15], and **parallel**[16]. A more complete overview is given in the task views[17].

The first **R** package introduced to CRAN was **rpvm**. It was proposed by Li and Rossini in 2001so it can distribute the computation over many computers (a computer cluster) using the PVM standard. This package is no longer available on CRAN. More details about this package can be found in Schmidberger, Morgan, Eddelbuettel, Yu, Luke, & Mansmann, (2009).

2.1 RMPI: INTERFACE (WRAPPER) TO MPI

The **R** package **Rmpi** was introduced by Hao Yu in 2002 to run initially under LAM/MPI. It has been developing over the years to work as an interface (wrapper) under other implementations of MPI such as MPICH/MPICH2, OpenMPI, and Deino MPI environments. It can be run under various versions of Linux and Windows.

The common **Rmpi** codes running on computer cluster and multicore systems in which MPICH2 is properly installed can be summarized into steps:

Step 1: Start cluster and use the command ***mpi.spawn.Rslaves()*** to detect the number of the slaves automatically. The command ***mpi.spawn.Rslaves(nslaves=8)*** means that a cluster with 8 slaves will start **R** with connection to the running MPI.

Step 2: Send objects to all slaves using the function ***mpi.bcast.Robj2slave()***. In this step, one can use the function ***mpi.scatter.Robj2slave()*** in order to distribute a list of objects from the master node to all slave workers. The function ***mpi.setup.rngstream()*** maybe used (optional) to initialize the random number stream as we will discuss in Section 4.

Step 3: Do the parallel calculations on all slaves simultaneously using suitable functions implemented in **Rmpi** such as ***mpi.remote.exec()***, ***mpi.apply()***, ***mpi.parSapply()***, ***mpi.parReplicate()***, and others.

Step 4: Stop the cluster using the ***mpi.close.Rslaves()*** command.

Some other functions from this package are briefly summarized in the Appendix of this article. More details can be found in Schmidberger, Morgan, Eddelbuettel, Yu, Luke, & Mansmann, (2009).

[9] http://cran.r-project.org/web/packages/rpvm/index.html

[10] http://cran.r-project.org/web/packages/Rmpi/index.html

[11] http://cran.r-project.org/web/packages/nws/index.html

[12] http://cran.r-project.org/web/packages/snow/index.html

[13] http://cran.r-project.org/web/packages/snowfall/index.html

[14] http://cran.r-project.org/web/packages/foreach/index.html

[15] http://cran.r-project.org/web/packages/multicore/index.html

[16] This package becomes a part of **R** >=2.14.0.

[17] http://cran.r-project.org/web/views/HighPerformanceComputing.html

2.2 NWS: R Functions for NetWorkSpaces and Sleigh

The **nws** package (stands for Net Work Spaces) was submitted to CRAN in 2006 by REvolution Computing so it provides coordination and parallel execution facilities using Sleigh. Sleigh is a part of NetWorkSpaces (NWS) allows users to execute tasks in parallel. More details about this package can be found in Schmidberger, Morgan, Eddelbuettel, Yu, Luke, & Mansmann, (2009).

2.3 SNOW: Simple Network of Workstations

The **snow** package proposed to literature by Rossini, Tierney and Li, (2007). It uses the PVM, MPI, NWS standards as well as direct SOCKETS. This package is widely used in parallel computing in **R**. It should be noted that **snow** requires the packages **Rmpi**, **nws**, and **rpvm** as interfaces to use the MPI, NWS, PVM standards respectively.

Similar to what we have discussed about **Rmpi**, the common **snow** codes for parallel computing can be summarized into steps:

Step 1: Start the cluster using any of the contributed functions ***makeCluster()***, ***makePVMcluster()***, ***makeMPIcluster()***, ***makeNWScluster()***, or ***makeSOCKcluster()***. The functions ***makePVMcluster()***, ***makeMPIcluster()***, ***makeNWScluster()***, and ***makeSOCKcluster()*** are used to start cluster under PVM, MPI, NWS, and SOCKETS using the corresponding **R** packages **rpvm ,** **Rmpi**, **nws**, and **snow** itself respectively. By default, the function ***makeCluster()*** uses **Rmpi** to start the job under MPI but others ("SOCK", "PVM", or "NWS".) can be also used via its argument type.

Step 2: Send a list of objects to all workers using the function ***clusterExport()***.

Step 3: Do the calculations at all slaves using any of the contributed functions such as ***clusterEvalQ()***, ***parLapply()***, ***parSapply()***, ***parApply()***, ***parRapply()***, ***parCapply()***, and others. In this step, one can use the functions ***clusterCall()*** or/and ***clusterApply()*** to call or/and distribute a list of objects from the master to the slave workers.

Step 4: Stop the cluster using the function ***stopCluster()***.

Some other functions from this package are briefly summarized in the Appendix of this article and more discussion can be found in Tierney, Rossini and Li (2009).

2.4 SNOWFALL: Easier Cluster Computing (Based on Snow)

The **snowfall** package was developed by Knaus, Porzelius, Binder, & Schwarzer, (2009) in order to improve the package **snow**. It provides an easier parallel programming using SOCKETS, MPI, PVM and NWS support. The **sfCluster** package is a UNIX management tool was developed by Knaus based on LAM/MPI to help the users of the **snowfall** package setting up the cluster and shutting it down. Functions available from **snowfall** can be used in sequential or parallel modes using the build in **snowfall** function ***sfInit()***. The function ***sfStop()*** is used to stop the cluster.

More details about these packages can be found in Knaus, Porzelius, Binder, & Schwarzer, (2009) and Schmidberger, Morgan, Eddelbuettel, Yu, Luke, & Mansmann, (2009).

2.5 MULTICORE: Parallel Processing of R Code on Machines with Multiple Cores or Cpus

The **multicore** package proposed by Simon Urbanek in 2009 and provides a way of running parallel computations in **R** using the forking[18] techniques on machines running with POSIX operating systems with multiple cores. It should be noted that this package is not compatible with Windows operating system and it is only running on MacOS X.

Recently, the R-Core team starts to include a new library **parallel** in **R** software releases >=2.14.0. This library contains a slightly revised copy of some useful functions taken from the **multicore** package as we will discuss in Section 2.7.

2.6 FOREACH: Foreach Looping Construct for R

In **R**, repeated executions can be accomplish using one of the looping functions such as ***for()***, ***repeat()***, ***while()***, and ***replicate()***. The family of ***apply()*** function which includes the functions ***apply()***, ***lapply()***, ***sapply()***, ***eapply()***, ***mapply()***, and ***rapply()*** can be also implemented to evaluate some statistical expressions repeatedly.

The **foreach** package provides a new looping algorithm for executing the **R** code repeatedly converting the ***for()*** loop statement in **R** to a ***foreach()*** loop. This algorithm allows general iteration over list of values in a collection (without the use of the body of the loop counter) on personal computer with multicore systems or multiple nodes of a cluster computer.

For parallel execution, the **foreach** package has been adapting the **R** parallel packages **doMC**[19](based on the package **multicore** on single workstations**)**, **doSNOW**[20](based on the package **snow** with SOCKETS**)**, and **doMPI**[21](based on the package **Rmpi**).

2.7 PARALLEL: Parallel R Package

The package **parallel** was introduced by Luke Tierney[22] and R-Core team in order to support parallel computation in **R**. The first version of this package was included in **R** version 2.14.0 and since then it becomes a part of the core **R** packages. This package is using *coarse-grained parallelization* and can be installed in the usual ways and is ready to use after typing:

```
R> library("parallel")
```

This package builds on the work done for CRAN packages **multicore** and **snow** with slight revised of copies of these packages by using forking taking from the package **multicore** and sockets taken from the package **snow**.

[18] Fork creates a new process (child) as a copy of the current **R** process that can work in parallel to the master process (parent).

[19] http://cran.r-project.org/web/packages/doMC/index.html

[20] http://cran.r-project.org/web/packages/doSNOW/index.html

[21] http://cran.r-project.org/web/packages/doMPI/index.html

[22] Luke Tierney is the maintainer of the **R** package **snow**.

The steps of computational parallel mode using **parallel** are almost exactly as those we discussed in Section 2.3. The **parallel** package provides new two functions ***makePSOCKcluster()*** and ***makeForkCluster()*** in order to spawn the slaves running on the same host as the master or optionally elsewhere. The ***makePSOCKcluster()*** is very similar to ***makeSOCKcluster()*** in the package **snow** whereas the function ***makeForkCluster()*** starts SOCKET cluster by forking on Unix-alike platforms only.

Some other functions from this package are briefly summarized in the Appendix of this article.

3. ANALOGUES OF APPLY R FUNCTIONS IN PARALLE PACKAGES

The **base R** library has a set of useful contributed functions for evaluating some expressions repeatedly that we have discussed in Section 2.6. In this article, we have named this group by the family of ***apply()*** function. The functions ***apply()***, ***lapply()***, ***sapply()***, and ***replicate()*** are used in a sequential (not in a parallel) mode in many applications. For example, selecting tuning parameters for neural network models often uses cross validation methodology based on iteration techniques. Usually the structure codes of such techniques include some looping functions taken from the members of this family. For parallel computing these functions aren’t useful, hence developers of **R** parallel packages provide some alternative parallel functions in order to parallelize the loops and speed up the calculations. Some of these functions are listed in Table 1.

Table 1
List to some of sequential and parallel loop functions in some R **libraries.**

base	**Rmpi**	**snow**	**snowfall**	**parallel**
apply	mpi.parApply	parApply	sfApply	parApply
apply for a row of matrix	mpi.parRapply	parRapply	-	parRapply
apply for a column of matrix	mpi.parCapply	parCapply	-	parCapply
lapply	mpi.parLapply	parLapply	sfLapply	parLapply/mclapply[23]
sapply	mpi.parSapply	parSapply	sfSapply	parSapply
replicate	mpi.parReplicate	-	-	-

Roughly, the speed up calculation time of running **R** code simultanousely on a cluster with n workers is approximately:

$$\text{Speed up} = \frac{\text{time with 1 CPU}}{\text{time with } n \text{ CPUs}}$$

4. PARALLEL MONTE-CARLO SİMULATİON

Bootstrapping is a nonparametric technique used for deriving estimates of standard errors and confidence intervals for estimates, such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient, based on selecting samples (with

[23] ***mclapply()*** uses forking from packages **multicore** and it is not working on Windows.

replacement) from the original dataset (observed dataset). Each selected sample (tested dataset) has the same number of elements as the observed dataset.

Monte-Carlo simulation (Barnard, 1963, Dufour and Khalaf, 2001) is more general than bootstrapping in the sense that it uses the random numbers for simulating the samples from a fitted model to the original dataset. Parametric bootstrap can be seen as a Monte-Carlo technique used to assess the performance of the bootstrapping estimators or predictors.

Monte-Carlo techniques involve massive computation due to the replications in simulations. These techniques have been used in many statistical applications such as multivariate portmanteau tests (Mahdi and McLeod 2012), DNA analysis (Hsiao and Stewart 2008). Parallel computing is a very useful tool aims to speed up the computation in such cases.

Results based on bootstrap and Monte-Carlo simulations aren't reproducible unless **R** users set the random seed up in their **R** code. In CRAN, Random Number Generators (RNG) for parallel computing are derived from the libraries **rsprng**[24] and **rlecuyer**[25] based on the algorithms discussed in L'Ecuyer, (1999) and L'Ecuyer, Richard, Chen, & Kelton, (2002).

The function ***mpi.setup.rngstream()*** from the package **Rmpi** contains support for multiple RNG streams based on the **rlecuyer** package, whereas the functions ***clusterSetupRNG()***, ***sfClusterSetupRNG()***, and ***clusterSetRNGStream()*** from the packages **snow**, **snowfall**, and **parallel** respectively contain multiple RNG streams based on **rsprng** and **rlecuyer** (users can choose either **rsprng** or **rlecuyer** with these functions). These functions from **Rmpi**, **snow**, **snowfall**, and **parallel** share the same idea that each separate worker generates random numbers independently from the others.

5. APPLICATIONS

Many packages available from CRAN have many applications with support for parallel processing using some contributed parallel functions from the packages that we have discussed above. For example, the package **portes**: Portmanteau Tests for Time Series Models[26], implements the Monte-Carlo significance test of portmanteau time series discussed in Lin and McLeod, (2006) and Mahdi and McLeod, (2012) based on the sequential and parallel modes using the **parallel** package. The package **survey**: Analysis of Complex Survey Samples[27], (Lumley, 2004) and the package **adegenet:** an **R** Package for the Exploratory Analysis of Genetic and Genomic Data[28], (Jombart, 2008, & Jombart & Ahmed, 2011) both have some support for parallel processing using **multicore** package. The package **PCIT**: PCIT Algorithm-Partial Correlation Coefficient with Information Theory[29], (Watson-Haigh, Kadarmideen, & Reverter, 2010) uses **Rmpi**

[24] http://cran.r-project.org/web/packages/rsprng/index.html
[25] http://cran.r-project.org/web/packages/rlecuyer/index.html
[26] http://cran.r-project.org/web/packages/portes/index.html
[27] http://cran.r-project.org/web/packages/survey/index.html
[28] http://cran.r-project.org/web/packages/adegenet/index.html
[29] http://cran.r-project.org/web/packages/PCIT/index.html

package for performing the partial correlation coefficient with information theory algorithm developed by Reverter & Chan, (2008). The package **pensim**: Simulation of High-Dimensional Data and Parallelized Repeated Penalized Regression[30], (Waldron, Pintilie, Tsao, Shepherd, Huttenhower, & Jurisica, 2011) implements **snow** package in simulating of continuous, correlated high-dimensional data with time to event or binary response, and parallelized functions for Lasso, Ridge, and Elastic Net penalized regression with repeated starts and two-dimensional tuning of the Elastic Net. Many examples are given in the online reference manual and vignettes of these packages.

In this section, we give a simple example with some illustrated **R** codes using a personal computer with quad core systems. It should be noted that some of these codes are running in a sequential mode and maybe consider being computer intensive while others implement functions taken from the **Rmpi** package in parallel mode. One can modify these codes in order to utilize other packages that we have discussed before.

The example that we introduce make use of the univariate regular time series giving the luteinizing hormone in blood samples (lh) at 10 minutes intervals from a human female. The sample size is 48 samples and it is available from the **R** package **datasets** (Diggle 1990). The ARMA(1,1) model[31] is fitted to this data and the portmanteau diagnostic test based on the Monte-Carlo version of Ljung-Box test discussed in Lin and McLeod, (2006) and Mahdi and McLeod, (2012) is applied into steps:

Step 1: Fit an ARMA(1,1) model to the time series and then apply the observed Ljung-Box statistic (denoted by *obs.stat*) at lag 10 using the function ***Box.test()*** given in **base R** package on the residual of the fitted model:

```
R > fit <-arima(lh, order = c(1,0,1))
R > res <- ts(fit$residuals)
R > n <- length(res)
R > ans <- Box.test(res, lag =10, type = "Ljung-Box")
R > obs.stat <- as.vector(ans$statistic)
```

Step 2: Extract the estimates parameters from the fitted model. These estimators are then will be used to simulate models using Monte-Carlo simulations:

```
R>phi <- as.vector(fit$coef[1])
R>theta <- as.vector(fit$coef[2])
R>sigma <- fit$sigma2
R>demean <- as.vector(fit$coef[3])
```

[30] http://cran.r-project.org/web/packages/pensim/index.html

[31] ARMA model in time series stands for autoregressive moving average model.

Step 3: Encapsulate the Monte-Carlo procedures into a function that can be used for simulating models from ARMA(1,1). In this function, the Ljung-Box test is applied on each simulated fitted model (denoted by *OneSim.stat*):

```
R>MonteCarlo <- function(n, res, phi, theta, sigma, demean) {
+                  innov <- sample(x=res,size=n,replace = TRUE, prob = NULL)
+                   Sim.Data <- arima.sim(n = n, list(ar = phi, ma = theta), innov = innov,
+                     sd = sqrt(sigma), mean = demean)
+                     FitSimModel <- arima(Sim.Data, order = c(1,0,1))
+                 rboot <- FitSimModel$resid
+                OneSim.stat <- Box.test(rboot, lag =10, type = "Ljung-Box")$statistic
+              return(as.vector(OneSim.stat))
+   }
```

Step 4: Start the cluster by loading the **Rmpi** package and spawn all slaves. Then apply Monte-Carlo significance test in parallel mode:

```
R>library("Rmpi")
R>mpi.spawn.Rslaves()                    ## Spawn all available slaves
R>mpi.setup.rngstream(123)               ## Set RNG stream to all slaves
R>mpi.bcast.Robj2slave(n)                ##  Export sample size to all salves
R>mpi.bcast.Robj2slave(res)
R>mpi.bcast.Robj2slave(phi)
R>mpi.bcast.Robj2slave(theta)
R>mpi.bcast.Robj2slave(sigma)
R>mpi.bcast.Robj2slave(demean)
R>mpi.bcast.Robj2slave(MonteCarlo)     ##  Export MonteCarlo function to all salves
R>sim.stat <- mpi.parReplicate(1000,MonteCarlo(n,res,phi,theta, sigma,demean))
R>mpi.close.Rslaves()                  ## Close the cluster
```

Step 5: In the final step, calculate the Monte-Carlo portmanteau p-value (note that the p-value based on the asymptotic distribution of Ljung-Box test is 0.587):

```
R>pvalue <- (1 + sum(as.numeric(sim.stat >= obs.stat)))/(1000 + 1)
R>pvalue
0.5394605
```

Running the previous example in a computer with single core using the sequential function ***replicate()*** instead of using the parallel function ***mpi.parReplicate()*** as follows:

```
R> set.seed(123)
R> sim.stat <- replicate(1000,MonteCarlo(n, res, phi, theta, sigma, demean))
R> pvalue <- (1 + sum(as.numeric(sim.stat >= obs.stat)))/(1000 + 1)
R> pvalue
0.5214785
```

With respect to the computer time, the CPU time is 3.21 seconds to get the output of this example in the parallel mode whereas it takes 10.4 seconds in the sequential mode.

6. COMMENTS AND CONCLUSION

There are many more available **R** parallel packages than those we discussed in this article and many of these are briefly described in the CRAN Task Views.

The reader should be aware that the packages available from CRAN, including those in the task views, need only to obey **R** formatting rules with no computer errors. It is not guaranteed that all of these packages produce correct results. On the other hand packages published by major publishers with high impact factor such as the Journal of Statistical Software (JSS) or Bioinformatics or Springer-Verlag or Chapman & Hall/CRC have been carefully reviewed for correctness and quality.

We have selected those parallel packages that might be of most general interest, that have been most widely used and that we are most familiar with. We have implemented some useful functions available from **Rmpi** package in the applications section using our personal computer with four quad core CPUs, but one can easily modify the implemented applications' code in order to use it with other **R** parallel packages.

Reader should note that **Rmpi** is only working in parallel mode with MPI, whereas **snow**, **snowfall**, and **parallel** can be implemented in either sequential or parallel mode using PVM, MPI, NWS, and SOCKETS.

ACKNOWLEDGEMENTS

The author would like to thank Dr. Ayman Baklizi, the coordinator of the statistics program at Qatar University, and both Dr. Raid Salha and Dr. Hisham Mahdi, the professors of statistics and mathematics at the Islamic University of Gaza (IUG) for their suggestions and helps. This research was supported by the research affairs at IUG.

REFERENCES

1. Bader, D. and Pennington, R. (2001). Cluster computing: Applications. *The International Journal of High Performance Computing*, 15(2), 181-185.
2. Barnard, G.A. (1963). Discussion of "The spectral analysis of point processes" by M. S. Bartlett. *Journal of the Royal Statistical Society*, B, 25, 264-96.
3. Diggle, P. (1990). Time Series: A Biostatistical Introduction. Oxford.
4. Dufour, J.-M. and Khalaf, L. (2001). Monte–Carlo test methods in econometrics. In companion to theoretical econometrics (eds B. Baltagi). Oxford:Blackwell.
5. Hsiao, Y. and Stewart, R.D. (2008). Monte Carlo simulation of DNA damage induction by x-rays and selected radioisotopes. *Physics in medicine and biology*, 53(1), 233-244.
6. Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11), 1403-1405.
7. Jombart, T. and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27(21), 3070-3071.
8. Knaus, J. (2010). *Developing parallel programs using snowfall*. Retrieved from CRAN : cran.r-project.org/web/packages/snowfall/vignettes/snowfall.pdf
9. Knaus, J., Porzelius, C., Binder, H. and Schwarzer, G. (2009). Easier parallel computing in R with snowfall and sfCluster. *The R Journal*, 1(1).

10. L'Ecuyer, P., Richard, S., Chen, E.J. and Kelton, W.D. (2002). An object-oriented random-numberpackage with many long streams and substreams. *Operations Research*, 50, 1073-1075.
11. L'Ecuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators. *Operations Research*, 47, 159-164.
12. Lin, J.-W. and McLeod, A.I. (2006). Improved Pen˜a-Rodrıguez portmanteau test. Computational statistics and data analysis. 51(3), 1731-1738.
13. Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19.
14. Mahdi, E. and McLeod, I. (2012). Improved multivariate portmanteau test. *Journal of Time Series Analysis*, 33(2), 211-222.
15. Reverter, A. and Chan, E. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, 24(21), 2491-2497.
16. Rossini, A., Tierney, L. and Li, N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16(2), 399-420.
17. Saltzer, J., Clarrk, D., Romkey, J. and Gramlich, W. (1985). The desktop computer as a network. *1EEE Journal on selected areas in communications*, 3(3).
18. Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Luke, T. and Mansmann, U. (2009). State of the art in parallel computing with R. 31(1), 1-26.
19. Seiffert, U. (2002). Artificial neural networks on massively parallel computer hardware. ESANN'2002 proceedings - *European symposium on artificial neural network*s, (319-330). Bruges (Belgium).
20. Sevcikova, H. (2004). Statistical simulations on parallel computers. *Journal of Computational and Graphical Statistics*, 13(4), 886-906.
21. Sterling, T., Becker, D., Salmon, J. and Daniel, S. (1999). How to build a Beowulf - A guide to the implementation and application of PC clusters. Cambridge, Ma: The MIT Press.
22. Tierney, L., Rossini, A. and Li, N. (2009). Snow: A parallel computing framework for the R system. *International Journal of Parallel Programming*, 37, 78-90.
23. Vera, G., Jansen, R. and Suppi, R. (2008). R/parallel – speeding up bioinformatics analysis with R. *BMC Bioinformatics*, 9(390).
24. Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F., Huttenhower, C. and Jurisica, I. (2011). Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27(24), 3399-3406.
25. Watson-Haigh, N., Kadarmideen, H. and Reverter, A. (2010). PCIT: an R package for weighted gene co-expression networks based on partial correlation and information theory approaches. *Bioinformatics*, 26(3), 411-413.

APPENDIX

Rmpi

Table 2
Some contributed functions to CRAN taken from Rmpi **package.**

Function	**Purpose**
mpi.spawn.Rslaves	Start the cluster and spawn **R** slaves
mpi.setup.rngstream	Setups RNG Streams based on **rlecuyer** package on all slaves
mpi.bcast.Robj	Move a general **R** object around among master and all slaves
mpi.bcast.Robj2slave	Same as ***mpi.bcast.Robj()*** function
mpi.scatter.Robj2slave	Distributing a list of data from the master to all slaves
mpi.remote.exec	Do calculations at all slaves
mpi.parReplicate	Parallel version of the function ***replicate()*** in **base R**
mpi.apply	Scatter an array to slaves and then apply a function
mpi.iapplyLB	Parallel apply with no blocking features
mpi.close.Rslaves	Shut down the cluster

SNOW

Table 3
Some contributed functions to CRAN taken from snow **package.**

Function	**Purpose**
makeCluster	Start the cluster: The default is the MPI cluster
makeSOCKcluster	Start SOCKET clusters
makeMPIcluster	Start MPI cluster
makePVMcluster	Start PVM cluster
makeNWScluster	Start NWS cluster
clusterSetupRNG	Implementation of Pierre L'Ecuyer's RNG Streams on all slaves
clusterSetupSPRNG	Load the **rsprng** package and initialize separate streams on all slaves
parLapply	Parallel version of ***lapply()*** function
parSapply	Parallel version of ***sapply()*** function
parApply	Parallel version of ***apply()*** function
parRapply	Parallel row ***apply()*** function for a matrix
parCapply	Parallel column ***apply()*** function for a matrix
clusterCall	Call a function on each node and returns list of results
clusterEvalQ	Evaluate a literal expression on each node
clusterExport	Export global variables from master to slaves
stopCluster	Shut down the cluster

SNOWFALL

Table 4
Some contributed functions to CRAN taken from snowfall **package.**

Function	**Purpose**
sfInit	Start the cluster
sfGetCluster	Get the snow cluster handler. Use for direct calling of snow functions
sfClusterSetupRNG	Setups RNG Streams on all slaves (L'Ecuyer is default)
sfLibrary	Load an R libraries on all nodes, including master
sfExport	Export variables from the master to all slaves
sfRemove	Remove previous exported variables from slaves and (optional) master
sfExportAll	Export global variables from master to slaves with exception of given list
sfRemoveAll	Remove all global variables from the slaves
sfClusterCall	Call a function on each node and returns list of results
sfClusterEvalQ	Evaluate a literal expression on all nodes
sfLapply	Parallel version of lapply() function
sfSapply	Parallel version of sapply() function
sfApply	Parallel version of apply() function
sfStop	Shut down the cluster

PARALLEL

Table 5
Some contributed functions to CRAN taken from parallel **package**

Function	Purpose
detectCores	Detect the number of cores/ CPU automatically
clusterSetRNGStream	Implementation of Pierre L'Ecuyer's RngStreams on all slaves
makeCluster	Same as ***makeCluster()*** in **snow** package
makePSOCKcluster	Create a parallel socket cluster
makeForkCluster	Create socket cluster by forking (On Unix-alike platforms only)
clusterCall	Same as ***clusterCall()*** in **snow** package
clusterEvalQ	Same as ***clusterEvalQ()*** in **snow** package
clusterExport	Same as ***clusterExport()*** in **snow** package
parLapply	Same as ***parLapply()*** in **snow** package
parSapply	Same as ***parSapply()*** in **snow** package
parApply	Same as ***parApply()*** in **snow** package
mcmapply	Parallel version ***mapply()*** function using forking (not for Windows)
stopCluster	Same as ***stopCluster()*** in **snow** package

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 647-654

ESTIMATING THE POPULATION MEAN USING STRATIFIED EXTREME RANKED SET SAMPLE

Mahmoud Ibrahim Syam[1], **Kamarulzaman Ibrahim**[2]
and **Amer Ibrahim Al-Omari**[3]

[1] Department of Mathematics, Foundation Program, Qatar University,
Doha, Qatar. Email: m.syam@qy.edu.qa

[2] School of Mathematical Sciences, University Kebangsaan Malaysia,
Malaysia. Email: kamarulz@ukm.my

[3] Department of Mathematics, Faculty of Sciences, Al al-Bayt University,
Jordan. Email: alomari_amer@yahoo.com

ABSTRACT

Stratified extreme ranked set sampling (SERSS) method is suggested for estimating the population mean. The SERSS is compared with the simple random sampling (SRS), stratified simple random sample (SSRS) and stratified ranked set sampling (SRSS). The authors showed that SERSS estimator is an unbiased of the population mean and more efficient than SRS, SSRS and SRSS when the underlying distribution is symmetric about its mean. Also, by SERSS the efficiency of the mean estimator can be increased for specific value of the sample size. A collection of a real data is used to illustrate the method.

KEY WORDS

Ranked set sampling; extreme ranked set sampling; stratified extreme ranked set sampling; efficiency.

1. INTRODUCTION

In Last year’s, the ranked set sampling method which was proposed by McIntyre [2] to estimate mean pasture yields was developed and modified by many authors to estimate the mean of the population. Dell and Clutter [1] showed that the mean of the RSS is an unbiased estimator of the population mean, whatever or not there are errors in ranking. Samawi et al [7] investigated variety of extreme ranked set sample (ERSS) for estimating the population mean. Muttlak [5] suggested using median ranked set sampling (MRSS) to estimate the population mean. Muttlak [4] suggested quartile ranked set sampling (QRSS) to estimate the population mean and he showed using QRSS procedure will reduce the errors in ranking comparing to RSS since we only select and measure the first or the third quartile of the sample. Also, Samawi [8] suggested using stratified ranked set sampling.

In this paper, we suggest the stratified extreme ranked set sampling (SERSS) to estimate the population mean of symmetric and asymmetric distributions. The organization of this paper is as follows: In Section 2 we present some of sampling methods. Estimation of the population mean is given in Section 3. A simulation study is considered in Section 4. Finally, conclusions on the suggested estimator are introduced in Section 5.

2. SAMPLING METHODS

In stratified sampling the population of N units is first divided into L subpopulations of $N_1, N_2, \cdots, N_L$ units, respectively. These subpopulations are no overlapping and together they comprise the whole population, so that $N_1 + N_2 + \cdots + N_L = N$. The subpopulations are called strata. To obtain the full benefit from stratification, the values of the $N_h, h = 1, 2, \cdots, L$ must be known. When the strata have been determined, a sample is drawn from each, the drawings being made in different strata. The sample sizes within the strata are denoted by $n_1, n_2, \cdots, n_L$, respectively. If a simple random sample is taken in each stratum, the whole procedure is described as stratified simple random sampling (SSRS).

The ranked set sampling (RSS) suggested by McIntyre (1952) is conducted by selecting n random samples from the population of size n elements each, and ranking each element within each set with respect to the variable of interest. Then an actual measurement is taken of the element with the smallest rank from the first sample. From the second sample an actual measurement is taken from the second smallest rank, and the procedure is continued until the element with the largest rank is chosen for actual measurement from the n-th sample. Thus we obtain a total of n measured elements, one from each ordered sample of size n and this completed one cycle. The cycle may be repeated m times until nm elements have been measured.

The ERSS procedure proposed by Samawi et al. (1996) depend on selecting *n* random samples each of size m units from the population and rank each sample with respect to a variable of interest. If the number of samples *n* is even, then select for measurement from the first $n/2$ samples the smallest rank unit (minimum) and from the second $n/2$ samples the largest rank unit (maximum). If the number of the samples n is odd, then select for measurement from the first $(n-1)/2$ samples the smallest rank unit (minimum) and from the last $(n-1)/2$ samples the largest rank unit (maximum), and the median from the middle sample. The cycle can be repeated *m* times if needed to get a sample of size *mn* units.

If we divide the population into L non overlapping strata, and apply the extreme ranked set sample in each stratum, the whole procedure is described as stratified extreme ranked set sampling (SERSS).

To illustrate the method, let us take the following two examples; first example if the number of samples in each stratum is even, and the second example if the number of samples in each stratum is odd. Please notice that the number of subpopulations (strata) is not important to be even or odd.

Example 1: Suppose we have two strata, i.e., $L = 2$ and $h = 1, 2$. Let $X_{hi(1)}$, $X_{hi(m)}$ and $X_{hi(\frac{m}{2})}$ be the minimum, maximum and median order statistics of the *i*th sample in the *h*th stratum, respectively. Assume that from the first stratum we draw 6 samples, each of size 6, and from the second stratum we draw 7 samples each of size 7 (number of samples is not necessary to be the same as number of elements) as the following

Stratum 1: Six samples are obtained and ranked as follows:

$$X_{11(1)}, X_{11(2)}, \ldots, X_{11(6)}; X_{12(1)}, X_{12(2)}, \ldots, X_{12(6)}; X_{13(1)}, X_{13(2)}, \ldots, X_{13(6)};$$
$$X_{14(1)}, X_{14(2)}, \ldots, X_{14(6)}; X_{15(1)}, X_{15(2)}, \ldots, X_{15(6)}; X_{16(1)}, X_{16(2)}, \ldots, X_{16(6)}$$

For h=1, select $X_{1i(1)}$ for I = 1,2.3, and select $X_{1i(6)}$ for i =4,5,6. Thus, the following units are obtained from the first stratum: $X_{11(1)}, X_{12(1)}, X_{13(1)}, X_{14(6)}, X_{15(6)}, X_{16(6)}$

Stratum 2: Seven samples, each of 7 units, as given below:

$$X_{21(1)}, X_{21(2)}, \ldots, X_{21(7)}; X_{22(1)}, X_{22(2)}, \ldots, X_{22(7)}; X_{23(1)}, X_{23(2)}, \ldots, X_{23(7)};$$
$$X_{24(1)}, X_{24(2)}, \ldots, X_{24(7)}; X_{25(1)}, X_{25(2)}, \ldots, X_{25(7)}; X_{26(1)}, X_{26(2)}, \ldots, X_{26(7)};$$
$$X_{27(1)}, X_{27(2)}, \ldots, X_{27(7)}$$

For $h = 2$, select $X_{2i(1)}$ for i=1,2,3, and select $X_{2i(7)}$ for i=5,6,7, and $X_{2i(4)}$ for $i = 4$. So, the selected elements are $X_{21(1)}, X_{22(1)}, X_{23(1)}, X_{24(4)}, X_{25(7)}, X_{26(7)}, X_{27(7)}$.

Therefore, the SERSS units consist of $X_{11(1)}, X_{12(1)}, X_{13(1)}, X_{14(6)}, X_{15(6)}, X_{16(6)}$, $X_{21(1)}, X_{22(1)}, X_{23(1)}, X_{24(4)}, X_{25(7)}, X_{26(7)}, X_{27(7)}$.

3. ESTIMATION OF THE POPULATION MEAN

In the case of stratified extreme ranked set sampling (SERSS), when n_h is even, the estimator of the population mean is defined as

$$\overline{X}_{SERSS1} = \sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h}{2}} X_{hi(1)} + \sum_{i=\frac{n_h}{2}+1}^{n_h} X_{hi(m)} \right) \tag{1}$$

where $W_h = \dfrac{N_h}{N}$, N_h is the stratum size and N is the total population size.

The variance of SERSS1 is given by

$$Var\ \overline{X}_{SERSS1} = Var\left(\sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h}{2}} X_{hi(1)} + \sum_{i=\frac{n_h}{2}+1}^{n_h} X_{hi(m)} \right) \right)$$
$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} \left(\sum_{i=1}^{\frac{n_h}{2}} Var(X_{hi(1)}) + \sum_{i=\frac{n_h}{2}+1}^{n_h} Var(X_{hi(m)}) \right)$$
$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} \left(\sum_{i=1}^{\frac{n_h}{2}} \sigma_{hi(1)}^2 + \sum_{i=\frac{n_h}{2}+1}^{n_h} \sigma_{hi(m)}^2 \right). \tag{2}$$

In the case of stratified extreme ranked set sampling (SERSS), when n_h is odd, the estimator of the population mean is defined as

$$\overline{X}_{SERSS\,2} = \sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h-1}{2}} X_{hi(1)} + \sum_{i=\frac{n_h+3}{2}}^{n_h} X_{hi(m)} + X_{h\frac{n_h+1}{2}(\frac{m+1}{2})} \right) \qquad (3)$$

The variance of SERSS2 is

$$Var\ \overline{X}_{SERSS\,2} = Var\left(\sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h-1}{2}} X_{hi(1)} + \sum_{i=\frac{n_h+3}{2}}^{n_h} X_{hi(m)} + X_{h\frac{n_h+1}{2}(\frac{m+1}{2})} \right)\right)$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} \left(\sum_{i=1}^{\frac{n_h-1}{2}} Var(X_{hi(1)}) + \sum_{i=\frac{n_h+3}{2}}^{n_h} Var(X_{hi(m)}) + Var(X_{h\frac{n_h+1}{2}(\frac{m+1}{2})}) \right)$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} \left(\sum_{i=1}^{\frac{n_h-1}{2}} \sigma^2_{hi(1)} + \sum_{i=\frac{n_h+3}{2}}^{n_h} \sigma^2_{hi(m)} + \sigma^2_{h\frac{n_h+1}{2}(\frac{m+1}{2})} \right). \qquad (4)$$

Property 1: $\overline{X}_{SERSS}$ is an unbiased estimator of the mean of symmetric distributions

Proof: If n_h is even, we have

$$E(\overline{X}_{SERSS1}) = E\left(\sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h}{2}} X_{hi(1)} + \sum_{i=\frac{n_h}{2}+1}^{n_h} X_{hi(m)} \right)\right)$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h}{2}} E(X_{hi(1)}) + \sum_{i=\frac{n_h}{2}+1}^{n_h} E(X_{hi(m)}) \right)$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left(\sum_{i=1}^{\frac{n_h}{2}} \mu_{h(1)} + \sum_{i=\frac{n_h}{2}+1}^{n_h} \mu_{h(m)} \right)$$

where $\mu_{h(1)}$ and $\mu_{h(m)}$ are the means of the order statistics which correspond to the minimum and maximum respectively. Since the distribution is symmetric about μ, then $\mu_{h(1)} + \mu_{h(m)} = 2\mu_h$. Therefore, we have

$$
\begin{aligned}
E(\overline{X}_{SERSS1}) &= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\frac{n_h}{2}\mu_{h(1)}+\frac{n_h}{2}\mu_{h(m)}\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\frac{n_h}{2}\ \ \mu_{h(1)}+\mu_{h(m)}\ \right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\frac{n_h}{2}\ 2\mu_h\ \right) = \sum_{h=1}^{L}W_h\mu_h\ =\mu
\end{aligned}
$$

If n_h is odd, then

$$
\begin{aligned}
E\ \ \overline{X}_{SERSS2}\ \ &= E\left(\sum_{h=1}^{L}\frac{W_h}{n_h}\left(\sum_{i=1}^{\frac{n_h-1}{2}}X_{hi(1)}+\sum_{i=\frac{n_h+3}{2}}^{n_h}X_{hi(m)}+X_{h\frac{n_h+1}{2}(\frac{m+1}{2})}\right)\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\sum_{i=1}^{\frac{n_h-1}{2}}E(X_{hi(1)})+\sum_{i=\frac{n_h+3}{2}}^{n_h}E(X_{hi(m)})+E(X_{h\frac{n_h+1}{2}(\frac{m+1}{2})})\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\sum_{i=1}^{\frac{n_h-1}{2}}\mu_{h(1)}+\sum_{i=\frac{n_h+3}{2}}^{n_h}\mu_{h(m)}+\mu_{h(\frac{m+1}{2})}\right),
\end{aligned}
$$

where $\mu_{h(1)}$ is the mean for the minimum in the first $(n_h-1)/2$ samples in stratum *h*. $\mu_{h(m)}$ is the mean for the maximum in the last $(n_h-1)/2$ samples in stratum *h*. μ_h is the mean for the stratum *h*. Since the distribution is symmetric about the μ, then we have $\mu_{h(1)}+\mu_{h(m)}=2\mu_h$. Also we know that the mean and median are equal for any symmetric distribution. Therefore,

$$
\begin{aligned}
E\ \ \overline{X}_{SERSS2}\ \ &= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\left(\frac{n_h-1}{2}\right)\mu_{h(1)}+\left(\frac{n_h-1}{2}\right)\mu_{h(m)}+\mu_h\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\left(\frac{n_h-1}{2}\right)\ \mu_{h(1)}+\mu_{h(m)}\ \ +\mu_h\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\left(\left(\frac{n_h-1}{2}\right)\ 2\mu_h\ \ +\mu_h\right)\\
&= \sum_{h=1}^{L}\frac{W_h}{n_h}\ \ n_h-1\ \ \mu_h\ +\mu_h\ \ = \sum_{h=1}^{L}\frac{W_h}{n_h}\ n_h\mu_h\ \ = \sum_{h=1}^{L}W_h\mu_h\ \ =\mu.
\end{aligned}
$$

Property 2. If the distribution is symmetric about μ, then

$$
Var\ \ \overline{X}_{SERSS1}\ < Var\ \ \overline{X}_{SRS}\quad \text{and } Var\ \ \overline{X}_{SERSS2}\ < Var\ \ \overline{X}_{SRS}
$$

Proof. If the sample size is even, the variance of $\overline{X}_{SQRSS1}$ is given by

$$Var \;\; \overline{X}_{SERSS1} \;\; = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} \left(\sum_{i=1}^{\frac{n_h}{2}} \sigma^2_{hi(1)} + \sum_{i=\frac{n_h}{2}+1}^{n_h} \sigma^2_{hi(m)} \right)$$

Since the distribution is symmetric, we have $\sigma^2_{hi(1)} = \sigma^2_{hi(m)}$

$$Var \;\; \overline{X}_{SERSS1} \;\; = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \sigma^2_{h(1)} \; .$$

But $\sigma^2_{h(1)} < \sigma^2_h$ for each stratum $h = 1, 2, \cdots, L$, which implies that

$$Var \;\; \overline{X}_{SERSS1} \;\; = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \sigma^2_{h(1)} < \sum_{h=1}^{L} \frac{W_h^2}{n_h} \sigma^2_h = Var \;\; \overline{X}_{SSRS} \;\; < Var \;\; \overline{X}_{SRS}$$

and the proof is the same for odd sample size.

4. SIMULATION STUDY

In this section a simulation study is conducted to investigate the performance of SERSS for estimating the population mean. Symmetric and asymmetric distributions have been considered for samples of sizes $n = 7, 12, 14, 15, 18,$ assuming that the population is partitioned into two or three strata. The simulation was performed for the SRSS, SSRS and SRS data sets from different distributions symmetric and asymmetric. The symmetric distributions are uniform, normal and student t, and the asymmetric distributions are geometric, beta and Weibull. Using 100000 replications, estimates of the means, variances and mean square errors were computed.

When the underlying distribution is symmetric, the efficiency of SERSS relative to $SRS, SSRS, SRSS$ is given by

$$eff \;\; \overline{X}_{SERSS}, \overline{X}_T \;\; = \frac{Var \;\; \overline{X}_T}{Var \;\; \overline{X}_{SERSS}}, \; T = SSRS, SRSS, SRS \tag{5}$$

and if the distribution is asymmetric, the efficiency of SERSS relative to SSRS and SRSS is defined in terms of mean square error (MSE) as the following

$$eff \;\; \overline{X}_{SERSS}, \overline{X}_T \;\; = \frac{MSE \;\; \overline{X}_T}{MSE \;\; \overline{X}_{SERSS}}, \; T = SSRS, SRSS \tag{6}$$

where

$$MSE \;\; \bar{X}_T \;\; = Var \;\; \bar{X}_T \;\; + \left[Bias \;\; \bar{X}_T \;\; \right]^2, \; T = SSRS, SRSS, SERSS$$

The values of the relative efficiency found under different distributional assumptions are provided in Tables 1, 2, 3, 4 and 5.

Table 1
The relative efficiency for estimating the population mean using SERSS with respect to SRSS, SSRS and SRS with sample size *n = 14* and *n = 7*.

Distribution	***n = 14:*** $n_1 = 8$, $n_2 = 6$			***n = 7:*** $n_1 = 4$ and $n_2 = 3$		
	SRSS	***SSRS***	***SRS***	***SRSS***	***SSRS***	***SRS***
Uniform (0,1)	1.1743	1.1416	1.1278	1.2032	1.7873	1.7423
Normal (0,1)	1.4531	1.4923	1.5002	1.8941	1.2007	1.1901
Student T (3)	2.0684	2.4270	2.3999	2.7326	3.0082	3.0118
Geometric (0.5)	1.0017	1.0991	1.2042	2.4573	2.3682	2.2014
Beta (5,2)	1.2865	1.1637	1.1048	1.1039	1.0074	1.0353
Weibull (1,2)	1.5113	1.7723	1.6476	1.0088	1.0032	1.0418

Table 2
The relative efficiency for estimating the population mean using SERSS with respect to SRSS, SSRS and SRS with sample size *n = 12* and *n = 18*.

Distribution	***n = 12:*** $n_1 = 5$, $n_2 = 7$			***n = 18:*** $n_1 = 4$, $n_2 = 6$, $n_3 = 8$		
	SRSS	***SSRS***	***SRS***	***SRSS***	***SSRS***	***SRS***
Uniform (0,1)	1.7825	1.6264	1.6314	1.7427	2.4325	2.3741
Normal (0,1)	2.5172	3.6534	3.6221	2.1421	3.5272	3.3034
Student T (3)	3.1415	3.1138	3.0892	2.7812	2.8333	2.8673
Geometric (0.5)	1.0523	1.4326	1.4271	1.7643	2.0063	2.0341
Beta (5,2)	1.5396	1.3931	1.3514	1.0996	1.0368	1.9932
Weibull (1,2)	1.1892	1.5663	1.7132	1.5834	2.4672	2.2476

Table 3
The relative efficiency for estimating the population mean using SERSS with respect to SRSS, SSRS and SRS with sample size *n = 15* and *n = 18*.

Distribution	***n = 15:*** $n_1 = 3$, $n_2 = 5$, $n_3 = 7$			***n = 18:*** $n_1 = 10$, $n_2 = 8$		
	SRSS	***SSRS***	***SRS***	***SRSS***	***SSRS***	***SRS***
Uniform (0,1)	1.0053	2.6435	2.4734	1.0531	1.0168	1.0002
Normal (0,1)	1.1764	3.5673	3.4756	1.0543	1.2764	1.2650
Student T (3)	2.0641	2.8648	2.7639	2.6463	2.6409	2.7412
Geometric (0.5)	1.3824	2.0118	1.9974	1.0076	1.4623	1.4586
Beta (5,2)	1.0305	2.2017	2.0508	1.1021	1.0001	1.8231
Weibull (1,2)	1.6136	2.4365	2.1327	1.1847	1.0592	1.0248

Tables 1-3, in general, indicate that greater efficiency is attained using SERSS method as opposed to the other contending methods that have been discussed when estimating the population mean of the variable of interest. To be more specific, based on the simulation results we can conclude that:

1. SERSS is more efficient than SRSS, SSRS and SRS based on the same number of measured units. For example, when $n = 12$, the efficiency of SERSS with respect to SRSS, SSSRS and SRS are 2.5172, 3.6534 and 3.6221 respectively for estimating the mean of the normal distribution.
2. When the performance of SERSS are compared to either SRSS, SSRS or SRS, it is found that SERSS is more efficient, as shown by all the values of relative efficiency which are greater than 1.
3. When the performances of the suggested estimators are compared, the efficiency of the suggested estimators is found to be more superior when the underlying distributions are symmetric as compared to some asymmetric.
4. When the sample sizes are odd the performance of SERSS estimator is better than even data sets based on SRSS, SSRS and SRS.

5. CONCLUSIONS

In this paper, we have suggested a new estimator of the population mean using SERSS. The performance of the estimator based on SERSS is compared with those found using SRSS, SSRS and SRS for the same number of measured units. It is found that SERSS produces estimator of the population mean which is unbiased and SERSS is more efficient than SRSS, SSRS and SRS. Thus, SERSS should be more preferred than SRSS, SSRS and SRS for both symmetric and asymmetric distributions.

REFERENCES

1. Dell, T.R., and Clutter, J.L. (1972). Ranked set sampling theory with order statistics background. *Biometrika*, 28, 545-555.
2. McIntyre, G.A. (1952). A method for unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*. 3, 385-390.
3. Muttlak, H.A. (2003b). Modified ranked set sampling methods. *Pak. J. Statist.*, 19(3), 315-323.
4. Muttlak, H.A. (2003). Investigating the use of quartile ranked set samples for estimating the population mean. *Journal of Applied Mathematics and Computation*, 146, 437- 443.
5. Muttlak, H.A. (1997). Median ranked set sampling. *J. App. Statist. Sci*., 6(4), 577-586.
6. Ohyama, T., Doi, J., and Yanagawa, T. (2008). Estimating population characteristics by incorporating prior values in stratified random sampling/ranked set sampling. *Journal of Statistical Planning and Inference*, 138, 4021-4032.
7. Samawi, H., Abu-Dayyeh, W., and Ahmed, S. (1996). Extreme ranked set sampling, *The Biometrical Journal* 30, 577-586.
8. Samawi, H.M. (1996). Stratified Ranked Set Sample. *Pak. J. Statist.*, 12(1), 9-16.
9. Takahasi K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, 20, 1-31.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 655-664

SPATIAL ANALYSIS OF MALARIA INCIDENCE IN AFGHANISTAN USING COMBINED ESTIMATING EQUATIONS

Oyelola A. Adegboye[1], **Denis H.Y. Leung**[2], **You-Gan Wang**[3] and **Danelle Kotze**[4]

[1] Department of Mathematics, American University of Afghanistan
Email: aadegboye@auaf.edu.af

[2] School of Economics, Singapore Management University, Singapore

[3] Department of Mathematics, University of Queensland, Queensland, Australia

[4] Department of Statistics and Population Studies,
University of the Western Cape, SA

ABSTRACT

Malaria epidemics often occur in areas with non-immune populations living in arid and desert-fringe zones. It is of high importance to properly identify the risk factors that are associated with the incidence of Malaria. A crucial step in modeling spatial data is the specification of the spatial dependency, by choosing the correlation function. However, often the choice for a particular application is unclear and diagnostic tests will have to be carried out following fitting of a model. To resolve this problem, we propose a more efficient method for modeling spatial correlation by combining different estimating equations using the platform of generalized method of moments. We applied the proposed method to malaria data from the 34 provinces of Afghanistan in 2009. A total of 521817 blood slides were examined and 242127 confirmed malaria cases were recorded, out which 94% were Plasmodium *vivax*. Forty five percent of the total malaria cases occurred in low lands (< 1000 m). Moran's I statistics confirms the presence of spatial autocorrelation in our data (Moran=0.0704, *p-value* =0.011) and a good fit was obtained by fitting a wave model to the semivariogram. Purely spatial scan statistics suggested clusters in 13 provinces of Afghanistan in 2009; Farah and Takhar province were identified as the most likely clusters with relative risk of 5.33 (*p-value* < 0.0001). Preliminary results showed that the proposed is superior to those using Generalized Estimating Equations (GEE) with three known working correlation. It shows an improvement with precise parameter estimates, at least when compared with independent, compound symmetry, user defined correlations.

1. INTRODUCTION

About 248 million people in the Eastern Mediterranean region are at risk of malaria transmission from both Plasmodium *falciparum* and Plasmodium *vivax*. Afghanistan is a malaria-endemic country and has the second highest burden of malaria in the region (Safi, *et al.*, 2009), with Plasmodium *vivax* accounting for nearly 90% of all the malaria cases. P. *vivax* malaria is associated with rice growing areas and transmitted by the endophilic and exophilic rice-field breeders Anopheles *pulcherrimus* and A. *hyrcanus* in Afghanistan (Faulde, *et al.*, 2007). The incidence of P. *falciparum* malaria in Afghanistan dropped to 7% in 2005 from 21% in 2001 World Health Organization, 2008). The use of vertical indoor residual spraying and treatment with chloroquine had almost eradicated the disease before the start of the armed conflicts in Afghanistan (Kolaczinski, *et al.*, 2007). The strain of malaria in Afghanistan is now chloroquine resistance (Centers for Disease Control and Prevention, 2012) and thus, the use of insecticide-treated mosquito nets (ITN) has been extensively debated for as a feasible option for protection against malaria in chronic emergencies (Rowland, *et al.*, 2002). Inequities in the usage of ITN within households in Afghanistan has been documented in Howard, *et al.*, (2010)

In the current study, we will consider only the spatial aspect of the transmission as we assume that malaria incidences across provinces are dependent. We handled the spatial dependence using Generalized Estimating Equations (GEE) (Liang and Zeger, 1986), wherein the covariance matrix is structured by using a working correlation matrix fully specified by the vector of parameters. An advantage of GEE as demonstrated by Liang and Zeger (1986), is that the parameter estimates and their robust variances are consistent even when the correlation structure is misspecified. However, choosing the working correlation structure closest to the true structure increases the statistical efficiency of the parameter estimates. To resolve this problem, we propose a more robust method for modeling spatial correlation by combining different estimating equations using the platform of generalized method of moments (Hansen, 1982). We illustrate this robust method by modeling the spatial correlation in the malaria incidence in Afghanistan in 2009.

2. DATA SOURCES

The data sets used in this study were monthly cases of malaria incidence reported to the Health Management Information System of the Ministry of Public Health by different health facilities across the 34 provinces of Afghanistan in 2009. Of a total of 521817 blood slides examined, 242127 were positive and clinical malaria cases; 94% were of P. *vivica*, 6% P. *falciparum* and 148602 were cases of malaria hemophagocytic syndrome

3. COMBINING ESTIMATING EQUATIONS

Let y_1, y_2,...y_{34} be the counts of malaria cases in the 34 provinces (s_1, s_2, ..., s_{34}) of Afghanistan. Associated with these provinces are the covariates x_1,....,x_n that measure the spatial location. In order to model spatial correlation and over dispersion, we assume there is a nonnegative weakly stationary latent process e_1,...,e_n such that conditional on the *e's*, the *y's* are independent and are assumed to follow a log-linear model given by

$$\mathrm{E}(y_i \mid e_i) = \exp(x_i^T \beta) e_i \quad \text{and} \quad \mathrm{Var}(y_i \mid e_i) = \mathrm{E}(y_i \mid e_i) \qquad 3.1$$

where β is a vector of unknown parameters that capture the association. We assume E(e_i) to be 1 so that exp($x_i\beta$) represents the marginal mean of y_i. The latent process e_i is assume to have a variance of σ^2 and the covariance between e_i, e_j is given by

$$\mathrm{Cov}(e_i, e_i) = \sigma^2 \rho(z_i, z_j, \alpha) \qquad 3.2$$

where z_i, z_j are covariates from s_i, s_j that jointly induce spatial correlation and α is an unknown vector of parameters. This is the same model as in Zeger (1988). Under these assumptions, it can be shown easily that

$$\mathrm{E}(y_i) = \exp(x_i^T\beta) = \mu_i(\beta),$$
$$\mathrm{Var}(y_i) = \mu_i(\beta) + \mu_i(\beta)^2\sigma^2,$$
$$\mathrm{Cor}(y_i, y_j) = \rho(z_i, z_j, \alpha)\{1 + \sigma^2\mu_i(\beta))^{-1}\}\{1 + \sigma^2\mu_j(\beta))^{-1}\}$$

If $\rho(z_i, z_j, \alpha) = 0$, then we have an over dispersion Poisson model; furthermore, if σ^2=0 then we have a standard Poisson model at each spatial location. A crucial step in modeling spatial data is the specify the spatial correlation. Cressie, 1991, pp 61-64) discussed some popular choices of the correlation function. However, often the choice for a particular application is unclear and diagnostic tests will have to be carried out following fitting of a model (e.g. McShane, *et al.,* 1997). To resolve this problem, we adopt a more robust method for modeling spatial correlation.

Following Zeger (1988}, a generalized estimating equation (GEE) type model can be used to estimate the parameters β, by solving the following estimating equation:

$$S(\beta,\alpha) \equiv D^T V^{-1}\{y - \mu\} = 0 \qquad 3.3$$

where μ= $(\mu_1, \ldots, \mu_n)^T$ and, $D = \partial\mu / \partial\beta^T$ and V is the $n \times n$ variance covariance matrix of $y_{=(}y_1, y_2,\ldots y_{34)}{}^T$. The matrix V can be expressed as $A + \sigma^2 AR(\alpha)A$, where A=diag $\mu_1, \ldots, \mu_n$), $R(\alpha)$ is a $n \times n$ correlation matrix, with the *i,j-th* element equal to $\rho(z_i, z_j, \alpha)$,and σ^2 is the scale parameter used to model over dispersion.

Consider different linearly independent choices of $R(\alpha)$, say $R^j(\alpha), j=1,\ldots,J$, and write $S^j(\beta, \alpha)$ for the estimating equation 3.3 using working correlation matrix $R^j(\alpha)$.

Let h(β, α)≡ $(S^1(\beta, \alpha)^T,\ldots, S^j(\beta, \alpha)^T,\ldots, S^J(\beta, \alpha)^T)^T$ and note that h≡ h(β, α) is a function of β, α only. Note that

$$h(\beta,\alpha) = (S^1(\beta,\alpha)^T, \ldots, S^J(\beta,\alpha)^T)^T = \begin{pmatrix} D^T\{A + \sigma^2 AR^1(\alpha)A\}^{-1}\{y-\mu\} \\ . \\ . \\ . \\ D^T\{A + \sigma^2 AR^J(\alpha)A\}^{-1}\{y-\mu\} \end{pmatrix}, \qquad 3.4$$

In general, the dimension of h is higher than the dimension of β. We can use generalized method of moments (Hansen, 1982, GMM) to combine the estimating equations $S^1,\ldots,S^J$. The idea of the GMM is to find *β, α* that jointly minimize the following quantity:

$$h(\beta,\alpha)W(\beta,\alpha)h^T(\beta,\alpha) \qquad 3.5$$

where W is a weight matrix. Hansen (1982) showed that the optimal choice of W^{-1} is the variance covariance matrix of the components of h, in the sense that the resulting estimate of β, α is semi-parametrically efficient under the assumptions of the model. The solution to 3.5 could easily be obtained by algorithms such as the Newton-Raphson method.

Our interest is if one of the $S^1,...,S^J$ is the correct estimating equation, in the sense that it solves 3.3 with $A + \sigma^2 AR(\alpha)A = V^{-1}$, then the GMM estimate will be optimal. If none of them is correct, then the GMM estimate is still consistent and combines optimally the information in $S^1,...,S^J$.

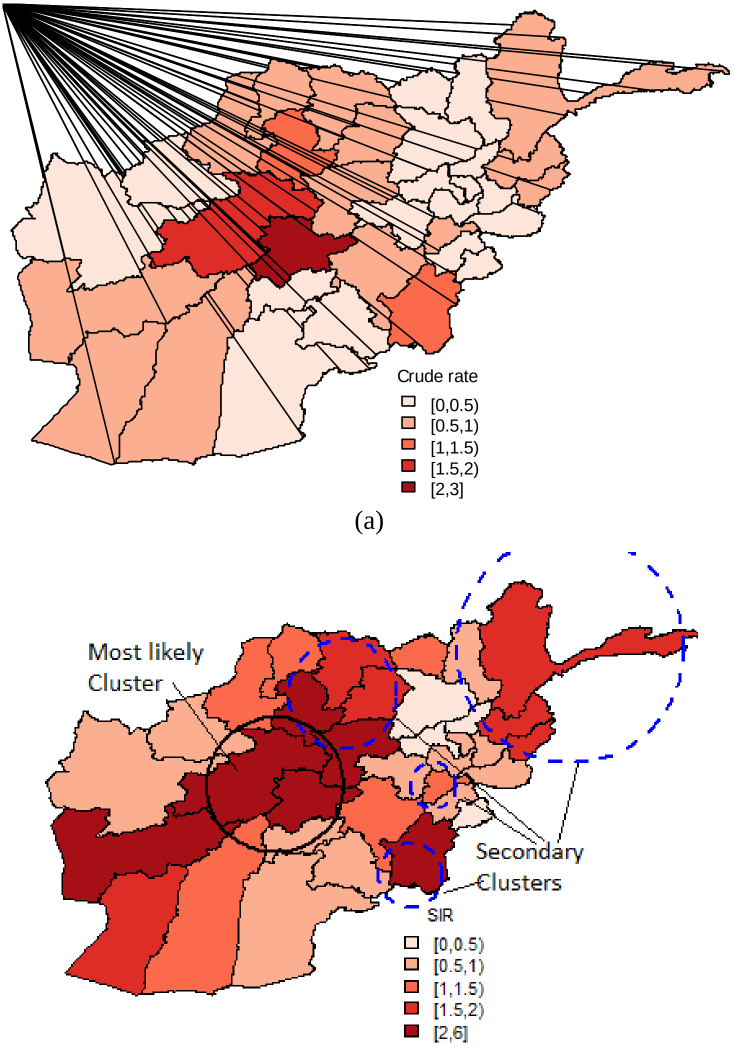
In practice, a few popular choices of $R(\alpha)$, for example, those discussed in Cressie (1991) may be used; we will also explore simpler forms of correlation matrices, for example, exchangeable, AR(1) and MA(1). These simpler forms of correlation matrices may become valuable in practice if they jointly can approximate unknown and difficult forms of the correlation structures.

4. COMMENTS AND DISCUSSION

This study is motivated by data from malaria cases in Afghanistan in 2009. The distribution of the crude rates across Afghanistan are presented in Figure 1a while Figure 1c indicates the 25th, 50th and 75th percentile of the malaria cases in the 34 provinces of Afghanistan in 2009. About 30% of total malaria incidence occurred in Nangarhar province, 13% in Kunar and 7% in Badakhshan province. Although some parts of Afghanistan that lies above 2000 m (above sea level) maybe free of malaria, cases of malaria predominantly occurred between the month of April and December in all areas less than 2,000 m (Centers for Disease Control and Prevention, 2012).. A few cases of P. *falciparum* have been documented in areas with high altitude (>2400 m) (AbdurRab, *et al.*, 2003). Malaria cases are more pronounced in areas that are less than 1000 m above sea level and accounted for about 45% of the total malaria cases (out of this, 95% are cases of P. *vivax*). Furthermore, 21% of the malaria incidence occurred in areas that have

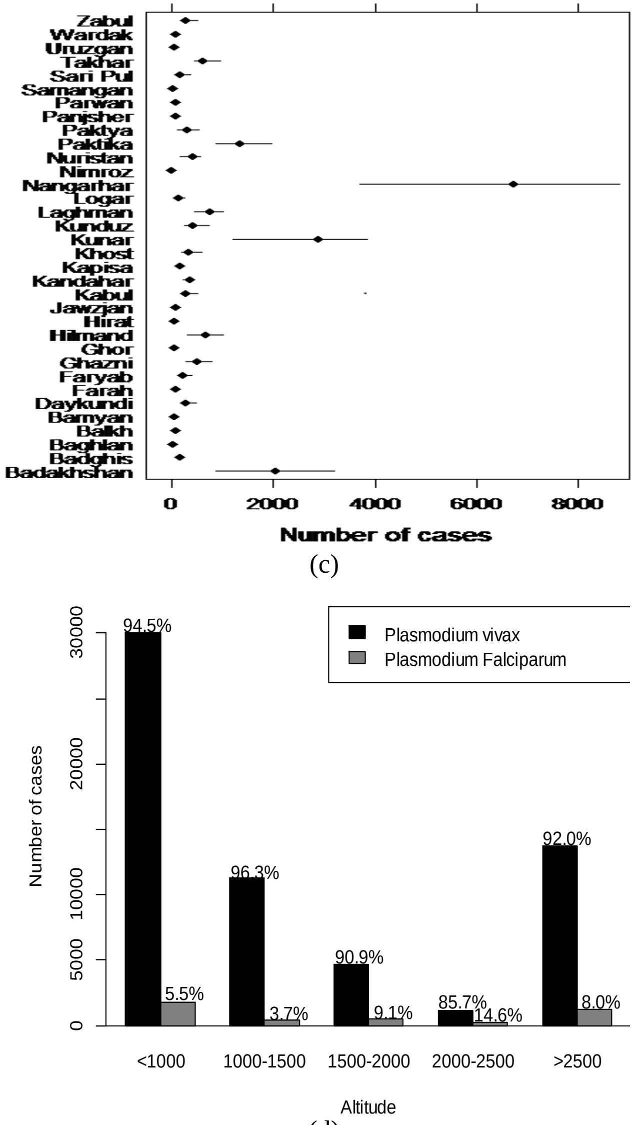
Table 1
Summary results of confirmed malaria incidence clusters in Afghanistan

Province	**Cases**	**Relative Risk**	***P-value***
Farah, Takhar	5339	5.33	<0.0000
Ghazni	16436	2.33	<0.0000
Badakhshan, Bamyan, Faryab, Kapisa, Hilmand, Kunduz, Zabul	76173	1.32	<0.0000
Baghlan, Balkh, Wardak	4511	2.17	<0.0000


Crude rate
[0,0.5)
[0.5,1)
[1,1.5)
[1.5,2)
[2,3]
Most likely Cluster
Secondary Clusters
SIR
[0,0.5)
[0.5,1)
[1,1.5)
[1.5,2)
[2,6]

(a)

(b)



(c)

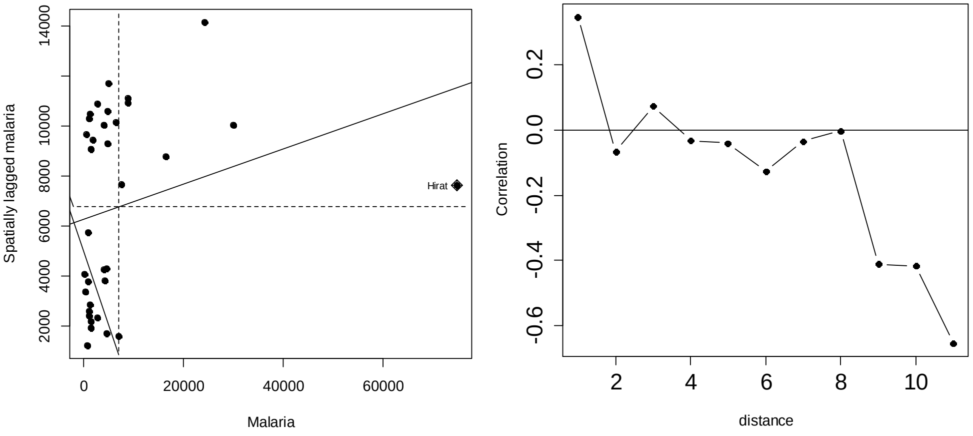
(d)

Fig. 1: Map of Afghanistan indicating

a) The crude rates of malaria incidence:
b) Locations of the confirmed malaria incidence clusters in Afghanistan (2009) by SaTScan:
c) Distribution of total number of reported malaria cases; the lines run from the 25th percentile to the 75th percentile and the dot indicates the 50th percentile (median):
d) Malaria species by altitude in the 34 provinces of Afghanistan an altitude above 2500 m. It can be seen that about 14000 cases of P. *vivax* malaria cases where reported in these areas (Figure 1d).

Figure 1b points out the presence of heterogeneity in the standard incidence rates of malaria across the provinces. The plot of the standard incidence rates (SIR) gave an impression of high incidence of malaria in the Central and North-Eastern region of Afghanistan. The Central area was identified as a primary cluster and the North-Eastern region as secondary clusters by SaTScan of Kullduf (1997), which is a circular spatial scan statistic that is an effective method of detecting and testing clusters. Table 1 shows the summary statistics of clusters of malaria cases detected by SaTScan in Afghanistan in 2009.

The analysis of the spatial autocorrelation was carried out using Moran's I (Moran, 1950), it measures the spatial autocorrelation that indicates how related the values of a variable are based on the locations where they were measured on a global scale. The Moran statistic was 0.0704 (Figure 2a) with a *p-value* = 0.011 confirming a significant spatial autocorrelation in our data. The structure of the spatial variation was estimated via semivariogarm (Cressie, 1991), to explore the spatial dependence between neighboring measures. In constructing a variogram, we only considered the values at smaller distances as the crucial portion of our analysis. Plot of the empirical semivariograms from smoothed malaria rates and the robust semivariogram as well as three fitted models (wave, gaussian and spherical) for the robust semivariogram (not shown here). We modeled the spatial dependence structure by fitting the wave model to the robust semivariogram with nugget=0.44, range=0.67 and sill=0.54.



(a) (b)

Fig. 2: Plots of spatial association for total malaria cases in Afghanistan, 2009

A crucial step in modeling spatial data is to specify the spatial correlation. The spatial correlation between measurements taken at two locations (provinces) were estimated using the following working correlations: independent (assuming no correlation), compound symmetry (constant correlation) and user define (distance dependence).

In order to analyze the spatial correlation of malaria cases, we proposed the method of combining Generalized Estimating Equations. Firstly, we used the three working correlation matrices mentioned before and fitted the data by GEE using these working correlations. The modified user define will be called user define 2 and is given as:

$$\mathrm{Corr}(Y_{ij}, Y_{ij+1}) = \begin{cases} \alpha^{d_{ij}} & d_{ij} \le d \\ 0 & d_{ij} > d \end{cases} \qquad 2.7$$

where α is the correlation, d_{ij} is the Euclidean distance and d is a specified maximum distance.

We set the correlation between communities that are more than 1 Euclidean distance apart to zero. This implied that we assumed no correlation between such communities. We begin by using different working correlations: independent, compound symmetry, autoregressive of order one, user defined correlation structure and those of the proposed method. Table 2 presents the preliminary results from the analysis implemented in **R**. These are point estimates together with their corresponding Akaike's Information Criteria (AIC) to assess the best model. The results assuming independent and distance dependent working are similar while that of compound symmetry and proposed method are similar. The merit of the models as assessed by AIC indicates that our proposed method has the lowest AIC implying its superiority.

Table 2
Parameter estimates from GEE Poisson model (and proposed method) of malariaincidence in Afghanistan

Risk factors	**GEE$_{\text{inde}}$**	**GEE$_{\text{CS}}$**	**GEE$_{\text{User define}}$**	**GEE$_{\text{GMM}}$**
Constant	9.00300	-0.5189	0.62131	-0.52620
Altitude	-0.00004	0.00004	-0.00016	0.00003
Population density	-0.0004	-0.00001	-0.00073	-0.00001
AIC	1326.714	626.701	701.246	623.512

The interpretation of the best model from combined GEE's shows positive effect of altitude and negative effect of population density on risk of malaria incidence. The positive effect of altitude is not surprising because about 21% of the total cases of malaria incidence occurred in highland areas (altitude > 2500) and P. *vivax* accounted for 92% of cases in this region. For example, Badakhshan province in the North-East (Figure 1b) which account for 7% of the total malaria cases in Afghanistan in 2009 is characterized by highlands and is associated with rice growing thereby providing a breeding environment for malaria vector Anopheles *pulcherrimus* and A. *hyrcanus* (Faulde, *et al.*, 2007).

Also the rural areas are characterized by low population densities, vacant lands and farms that provide suitable habitat for malaria vectors, thereby increasing the transmission of malaria. Studies on the effect of population density on malaria transmission have indicated that areas with low population density may not provide enough people to aid the transmission (Snow, *et al.*, 1999). High population density areas have reduced risk probably due to urbanization which implies access to preventative and curative measures, and better health care facilities that make the urban population less biologically or economically vulnerable to malaria infection (Hay, *et al.*, 2005; Tatem, *et al.*, 2008).

We have shown that combined GEE's method provide better results and precise parameter estimates, at least when compares with models with independent, compound symmetry, and user defined working correlations separately. Apart from the efficient parameter estimation and inferences thereof, our method also optimally combines the information in $S^1,...,S^j$. The merit of the proposed method was confirmed by AIC and can be implemented using the **geepack** package (Hjsgaard, *et al.*, 2006; Yan and Fine, 2004; Yan, 2002) in **R** with little additional coding.

5 REFERENCES

1. AbdurRab, M., Freeman, T., Rahim, S., Durrani, N., Simon-Taha, A. and Rowland, M. (2003). High altitude epidemic malaria in Bamian Province, central Afghanistan. *East Mediterranean Health Journal*, 9, 232-239.
2. Centers for Disease Control and Prevention (2012). URL http://wwwnc.cdc.gov.
3. Cressie, N. (1991). *Statistics for Spatial Data.* New York: Wiley.
4. Faulde, M., Ho_mann, R., Fazilat, K. and Hoerauf, A. (2007). Malaria Reemergence in Northern Afghanistan. *Emerging Infectious Diseases*, 13, 1402-1404.
5. Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.
6. Hay, S., Guerra, C., Tatem, A., Atkinson, P. and Snow, R. (2005). Urbanization, malaria transmission and disease burden in Africa. *Nature Reviews Microbiology*, 3, 81-90.
7. Hjsgaard, S., Halekoh, U. and Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2, 1-11.
8. Howard, N., Shafi, A., Jones, C. and Rowland, M. (2010). Malaria control under the Taliban regime: insecticide-treated net purchasing, coverage, and usage among men and women in eastern Afghanistan. *Malaria Journal*, 9, 14-18.
9. Kolaczinski, J., Graham, K., Fahim, A., Brooker, S. and Rowland, M. (2007). Malaria control in Afghanistan: progress and challenges. *Lancet*, 365, 1506-1512.
10. Kulldorff, M. (1997). A spatial scan statistic. *Communication in Statistics-Theory Methods.*, 26, 1481-1496.
11. Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
12. McShane, L.M., Albert, P.S. and Palmatier, M.A. (1997). A latent process regression model for spatially correlated count data. *Biometrics*, 53, 698-706.
13. Moran, P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37, 17-23.
14. Rowland, M., Webster, J., Saleh, P., Chandramohan, D., Freeman, T., Pearcy, B., Durrani, N., Rab, A. and Mohammed, N. (2002). Prevention of malaria in Afghanistan through social marketing of insecticide-treated nets: evaluation of coverage and effectiveness by cross-sectional surveys and passive surveillance. *Tropical Medicine and International Health*, 7, 813-822.
15. Safi, N., Hameed, H., Sediqi, W. and Himmat, E. (2009). NMLCP Annual Report, 2008. *Afghanistan Anuual Malaria Journal*, 1, 8-14.

16. Snow, R., Craig, M., Deichmann, U. and Marsh, K. (1999). Estimating mortality, morbidity and disability due to malaria among Africa's non-pregnant population. Bull *World Health Organ.*, 77, 624-640.
17. Tatem, A., Guerra, C., Kabaria, C., Noor, A. and Hay, S. (2008). Human population, urban settlement patterns and their impact on plasmodium falciparum malaria endemicity. *Malaria Journal*, 218.
18. World Health Organization (2006). Report on the Sixth intercountry meeting of national malaria programme managers, Cairo, Egypt. Tech. rep.
19. Yan, J. (2002). Geepack: Yet another package for generalized estimating equations. *R-News*, 2/3, 12-14.
20. Yan, J. and Fine, J.P. (2004). Estimating equations for association structures. *Statistics in Medicine*, 23, 859-880.
21. Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika*, 75, 621-629.

Proc. ICCS-12, Doha, Qatar
December 19-22, 2012, Vol. 23, pp. 665-668

2^n METRIC BINARY SYSTEM TO MEASURE DATA QUALITY

Redouane Betrouni
Mathematical Statistician, United States Census Bureau
Email: redouane.betrouni@census.gov; rbetroun@gmail.com

ABSTRACT

While data continues to grow, decision makers need answers faster than ever including the choice of using the right data to be analyzed. But the quality of the analysis conducted by their statisticians and scientists is only as good as the quality of the data they are analyzing.

1. CURRENTLY

Data providers conduct quality control on data using several techniques

- Frequency counts on categorical variables with not more than 5 categories.
- Compute the means at the most on numerical fields (variables).
- Visually explore the output.
- Comparison with last year output.
- Displaying the first 100 records.

2. BACKGROUND INFO

$$\begin{pmatrix} X_1X_2......X_p \\ \downarrow\downarrow.........\downarrow \\ x_{11}x_{12}...x_{1p} \\ x_{11}x_{12}...x_{1p} \\ .... \\ x_{n1}\ x_{n2}...x_{np} \end{pmatrix}$$

Data frequently gets to the analyst in a form of a table of m variables and p observations, It would be ideal if every cell is properly populated with information, but usually this is not necessarily the case.

- Missing information.
- Data entry errors.
- Data errors due to issues with FTP File transfer Protocols (Transferring files across servers)
- Unexpected modification to the file especially after encryption and decryption.
- Files created in main frame environment and moved to Unix/Linux

3. JUNK DATA

- ^M carriage return, ^Z end of file marker
- Linefeed marker, Non printable characters
- Non ASCII characters
- od -c filename | cut-c pos1-pos2 | more
- Analysts may unintentionally throw useful data "information" because the software said so.(Embedded spaces problem)

4. WEIGHTING BASED ON DATA QUALITY?

- Survey analysts employ weighting techniques to remove the bias that can be caused by non responses, non-interviews, opt-out cases, etc...
- If a measure of goodness can be made on every single record then this can be used as a weighting factor in order to produce a good estimate of the population total **Y**.

5. IMPLEMENTATION

- Data dictionary usage

With a good **data dictionary** in a place in addition to having a prior info about what my data should be then the researcher can develop a formula that measures a vector p*1 and then produce

$$S = \sum_{j=1}^{p} W_j Dict_{1\vee 0}(X_j)$$

where Dictionary take the value of either 1 or 0 depending on whether the data violates the dictionary

6. EXAMPLES

- If variable Gender/Sex is coded "F" for female and "M" for male

  1 will be assigned to all records with values other than F and M, & 0 otherwise.

- For the case of 3 variables X, Y, Z and assuming that X is more important for my analysis than Y and that Y is more important than Z

  I choose my weights using the power of 2 as follows

7. 2 TO THE n

$$S = 2^0 dictZ + 2^1 dictY + 2^2 dictX$$

- There are 2 to the n subsets that can be formed from the set formed by X, Y and Z including the empty set.
- The possible values of the score S computed above are 0,1, 2,3,4,5,6, and 7.

- There is a 1 to 1 relationship between the two sets

  $$\{0,1,2,3,4,5,6,7\}$$

  and

  $$\Omega_{\{X,Y,Z\}}$$

8. QC INPUT DATA USING S

Using Statistical data analysis techniques to ensure quality

- Detect any possible data anomalies using this score
- Modeling the score S to be a dependent variable
- the P(S=2) = probability of having a score of 2

  $$0*2^0+1*2^1+0*2^2=2$$

Corresponds with investigating if there is any anomaly with variable Y

9. NEED FOR EXPANSION

- The indicator Bernoulli variable

  $$Dict^j$$

Would not able to handle all situations

1 for data has violated the data dictionary, and 0 otherwise that can be expanded if Instead of using the 1 and 0 I can model my data using the probability of being not consistent with the data dictionary

- Dealing with variables like names, free text format variables In most cases We data users are not 100% sure if our data is wrong or not, it is much easier to know that the variable Gender must take two values Female and Male, but for a lot of variables it is not clear, this is the case with names for example We know that anybody can name his son Mickey Mouse but what is the probability that this can happened it is not zero but it is certainly very small and close to 0.

10. OPTIMUM WEIGHT ALLOCATIONS

- Assuming that I don’t have a preference of any variable over another.

  $$Score=S=\sum_{j=1}^{p}W_jS_j$$

- It is desirable to have a metric with good properties
- Having a smaller variance, more precision on the estimate.

  $$VAR(S)=W_j\sum_{j=1}^{p}VAR(S_j)+Double\sum W_iW_jCOV(S_i,S_j)$$

- By ignoring the second term

$$V^2 = \sum_{j=1}^{p} W_j V_j^2$$

- The Cauchy inequality states that for two set of real numbers Xi's and Yi's **(i=1,2,…,I)**

$$(\sum x_i^{\ 2})(\sum y_i^2) \geq (\sum x_i y_i)^2$$

- The minimum for the products is attainable at the equality and where the xi's and yi's are proportional(Kish Survey Sampling page 279)
- If The individual scores can be modeled as poisson distribution with parameters $\lambda_1, \lambda_2, \ldots, \lambda_p$

$$\frac{V_j^2}{W_j} = K$$

- If I have prior variance estimates of my lambdas the according to this result the appropriate weights I should use the line

$$k * (\lambda_1, \lambda_2, \ldots, \lambda_p).$$

11. COMMENTS AND CONCLUSION

This work is in ongoing and the next steps are

1. Expand the Score S to account for missing information.
2. Additional work conducting simulation to confirm the result regarding the optimum allocation (inch)
3. An alogrithm that uses something similar to backward elimiation process for regression but done on the dependence logistic variables p(y=n–1), p(y=n–2),...p(y=1) instead of the usual independent variables, and simulate the behaviour of the score S
4. Suggest Allocate a cut-off-point as a parameter for accepting /rejecting a file.
5. Develop an a java application that measure quality via a GUI on multiple type of files like Excel, SAS, SPSS, R. etc. assuming that there is a database of data dictionary.

REFERENCES

Leslie Kish (1965). *Survey Sampling*, page 279, New York: John Wiley and Sons.