Thailand Statistician
January 2009; 7(1) : 1-11
http://statassoc.or.th
Contributed paper

Application of Auxiliary Variable in Response Mean Estimation for Incomplete Longitudinal Data

Juthaphorn Saekhoo* [a] and Pachitjanut Siripanich [b]

[a] Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand.

[b] School of Applied Statistics, National Institute of Development Administration, Bangkapi, Bangkok, 10240, Thailand.

***Author for correspondence**; e-mail: fscijps@ku.ac.th

Received: 12 June 20082008
Accepted: 14 November 2008.

Abstract

This paper proposes an application of estimator for response mean Y given the value of an auxiliary variable X under simple linear regression model for incomplete longitudinal data and monotone missing data patterns. The proposed estimator, called Conditional Maximum Likelihood Estimator (CondMLE), is adapted from an Anderson's factored likelihood function. Monte Carlo simulations was repeated 2,000 times for each situations in comparison of the coefficient of variations (CV) derived from CondMLE and Anderson's estimator which generally regards no auxiliary variable in estimations. Essentially, regarding the results of the simulation study, CondMLE presented smaller CV than Anderson's estimator, for sample size of 20, 30 and 50, regardless differences in the percentages of missing data and correlation coefficients of the response variables in the two occasions.

Keywords: factored likelihood function, longitudinal data, monotone missing data pattern, occasion.

1. Introduction

Longitudinal studies are measurements of response variables which are continually operated on the same subjects over a period of time. Usually, this type of studies is occupied in measurements of changes in human body, typically in the identification of possible diseases, as well as in other studies where particular series of continual data are exclusively required. Nevertheless, since it is usually run on over years, a longitudinal study always entails a great deal of operational efforts and expenses. Yet, missing data are also always occurred, resulting in bias and inefficiency in the estimation of means, and also affecting the estimated mean of response variables in the study [1]. Consequently, data analysis for longitudinal studies should be carefully designed, so that incomplete data can be effectively analyzed.

Usually, estimation of response mean for incomplete longitudinal data is analyzed using standard statistical methods, through which missing observations are simply ignored. However, a mention has to be noted that those missing-observation-ignored methods are only appropriate for studies in which only a small number of missing data are occurred. Nonetheless, when analyzing a large number of missing data, standard statistical approaches usually cause imprecision and also escalation in biases.

In addition, this may either reduce or exaggerate statistical power (see [1-7]), and each of these distortions can also lead to invalid conclusions. In fact, a range of methods are used to impute non-response items in studies. However, the imputation methods usually cause under estimation of variance, and wrong inference of the response mean [1]. As a result, likelihood function based methods are usually employed to administer missing data. For example, Lord [8] proposed a maximum likelihood estimator for trivariate normal population from file matching incomplete data which is shown in the form of

$$\begin{array}{ll} y_{11}, \cdots, y_{1r}, & y_{1,r+1}, \cdots, y_{1n} \\ y_{21}, \cdots, y_{2r} & \\ & y_{3,r+1}, \cdots, y_{3n}. \end{array}$$

Furthermore, Edgett [9] found the maximum likelihood estimators for parameters of a trivariate normal distribution from univariate incomplete data which can be shown in the form of

$$\begin{array}{ll} y_{11}, \cdots, y_{1r}, & y_{1,r+1}, \cdots, y_{1n} \\ y_{21}, \cdots, y_{2r}, & y_{2,r+1}, \cdots, y_{2n} \\ y_{31}, \cdots, y_{3r}. & \end{array}$$

Moreover, Anderson [2] first introduced a factored likelihood function for normal data and estimated parameters of a multivariate normal population for monotone incomplete data. Particularly, in this paper another estimator of response mean, given application of auxiliary information for incomplete longitudinal data, under simple linear regression model with classical assumptions, is proposed. The proposed estimator is developed from Anderson's factored likelihood function, and is called CondMLE.

2. Methodology

2.1 Estimators

In this study, assume that missing completely at random data (MCAR), a drop out happening when probability is depended neither on the observed responses nor the responses with no dropped out [1], is assumed in estimation of response mean. The methods used in estimation of response mean for monotone missing data are as following:

2.1.1 Anderson's estimator

Let f_1 and $f_{2|1}$ be, respectively, the density of Y_1 and conditional density of Y_2 given Y_1 where Y_i is a random variable observed in occasion i, i = 1 or 2.

Anderson's factored likelihood function method is shown in equation (1)

$$L(\underline{\theta} \mid Y) = \prod_{j=1}^{n} f_1(y_{1j} \mid \mu_1, \sigma_1^2) \prod_{j=1}^{r} f_{2|1}(y_{2j} \mid \beta_0 + \beta_{2|1} y_{1j}, \sigma_{2|1}^2) \quad (1)$$

where $\beta_0 = \mu_2 - \beta_{2|1}\mu_1$

$$\beta_{2|1} = \rho_{12}\frac{\sigma_2}{\sigma_1}$$

$$\sigma_{2|1}^2 = (1-\rho_{12}^2)\sigma_2^2 .$$

In this method, the mean of the response variable Y from multivariate normal distribution is computed when some observations are missed, and no auxiliary variable X is concerned: $y_{11}, y_{12}, \dots, y_{1n}$ are assumed as data completely observed in the first occasion. The Anderson's estimator of response mean and variance in the first occasion are shown in equation (2). Then, in the second occasion, when $n-r$ observations are lost, there are only r subjects which can be observed, namely $y_{21}, y_{22}, \dots, y_{2r}$ where $0 < r < n$. The layout of this dataset is shown in Figure 1.

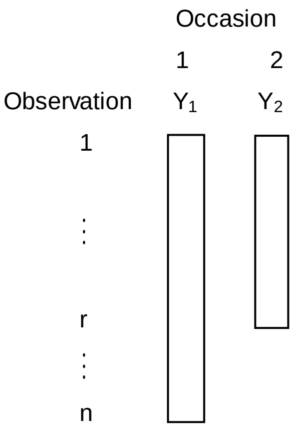

Figure 1. The monotone incomplete longitudinal dataset of Y.

The Anderson's estimator of response mean and variance in the second occasion are shown in equation (3).

First occasion:

$$\hat{\mu}_{Y_1 \text{Anderson}} = \bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}$$

$$\hat{\sigma}^2_{Y_1 \text{Anderson}} = \frac{1}{n}\sum_{j=1}^{n} (y_{1j} - \bar{y}_1)^2 \qquad (2)$$

Second occasion:

$$\hat{\mu}_{Y_2 \text{Anderson}} = \bar{y}'_2 - \hat{\beta}_{2|1}(\bar{y}'_1 - \bar{y}_1)$$

$$\hat{\sigma}^2_{Y_2 \text{Anderson}} = \hat{\sigma}^2_{2|1} + \hat{\beta}^2_{2|1}\hat{\sigma}^2_{Y_1 \text{Anderson}} \qquad (3)$$

where $$\hat{\beta}_{2|1} = \frac{\sum_{j=1}^{r}(y_{1j} - \bar{y}'_1)(y_{2j} - \bar{y}'_2)}{\sum_{j=1}^{r}(y_{1j} - \bar{y}'_1)^2}$$

$$\hat{\sigma}^2_{2|1} = \frac{1}{r}\left(\sum_{j=1}^{r}(y_{2j} - \bar{y}'_2)^2 - \hat{\beta}^2_{2|1}\sum_{j=1}^{r}(y_{1j} - \bar{y}'_1)^2\right)$$

$\bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}$, $\bar{y}'_1 = \frac{1}{r}\sum_{j=1}^{r} y_{1j}$ and $\bar{y}'_2 = \frac{1}{r}\sum_{j=1}^{r} y_{2j}$

2.1.2 The Proposed Estimator

In estimation of response mean for incomplete longitudinal data in this study, auxiliary information is required for both occasions. Rueda et al. [10] suggested the use of the auxiliary variable information, provided by one or several auxiliary variables, in the estimation of mean as a very usual technique and typically advantageous, particularly since powerful auxiliary information encourages reduction of biases and sampling errors. Therefore, the present study proposes an estimator which generally applies auxiliary variable- which is normally ignored in Anderson's estimator- in the estimation of response mean, under the name of Conditional Maximum Likelihood Estimator, or CondMLE: $\hat{\mu}_{Y_2|X_2=x_{2j_{\text{CondML}}}}$

In the first occasion, both response variable Y_{1j} and auxiliary variable X_{1j} where j = 1, 2, … , n are completely observed, that is, (x_{11}, y_{11}), (x_{12}, y_{12}), …, (x_{1n}, y_{1n}) are observations from the first occasion. For the second occasion, only r subjects can be observed, and (x_{21}, y_{21}), (x_{22}, y_{22}), … , (x_{2r}, y_{2r}) where $0 < r < n$ are observations from the second occasion. The layout of this dataset is shown in Figure 2.

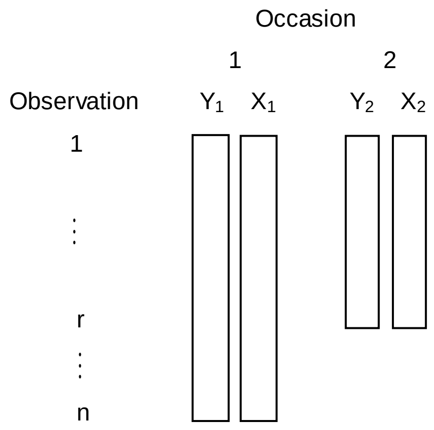

Figure 2. The monotone incomplete longitudinal dataset of (X,Y).

Moreover, a linear model is assumed and written as

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{4}$$

where $\underline{Y}$ is a $(n+r)\times 1$ vector of response which is distributed as normal and have no outliers, X is a $(n+r)\times 4$ matrix of auxiliary variables, $\underline{\beta} = \left[\ \underline{\beta}_1 \ \ \underline{\beta}_2\ \right]'$ is a 4×1 vector of coefficient of simple linear regression model, and $\underline{\varepsilon}$ is a $(n+r)\times 1$ vector of errors which is multivariate normally distributed with mean $\underline{0}_{(n+r)\times 1}$ and variance covariance matrix of $\underline{\varepsilon}$ is as followed.

$$\Sigma_\varepsilon = \begin{bmatrix} \sigma_1^2\mathbf{I}_r & \mathbf{0}_{r\times(n-r)} & \sigma_{12}\mathbf{I}_r \\ \mathbf{0}_{(n-r)\times r} & \sigma_1^2\mathbf{I}_{(n-r)} & \mathbf{0}_{(n-r)\times r} \\ \sigma_{12}\mathbf{I}_r & \mathbf{0}_{r\times(n-r)} & \sigma_2^2\mathbf{I}_r \end{bmatrix}_{(n+r)\times(n+r)}$$

Note that $\mathbf{I}_m$ is an $m\times m$ identity matrix, and $\mathbf{0}_{m\times p}$ is a $m\times p$ zero matrix. Under model (4), the likelihood function of parameter θ can be written as

$$L(\underline{\theta}\,|\,\underline{\varepsilon}) = \left[\left(2\pi\sigma_1^2\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2\sigma_1^2}\sum_{j=1}^{n}\varepsilon_{1j}^2\right)\right]\times\left[\left(2\pi\sigma_{2|1}^2\right)^{-\frac{r}{2}}\exp\left(-\frac{1}{2\sigma_{2|1}^2}\sum_{j=1}^{r}\left\{\varepsilon_{2j}-\tau_{12}\varepsilon_{1j}\right\}^2\right)\right] \qquad (5)$$

where $\underline{\varepsilon} = (\varepsilon_{11},\ \varepsilon_{12}, \ldots,\ \varepsilon_{1n},\ \varepsilon_{21},\ \varepsilon_{22}, \ldots,\ \varepsilon_{2r})'$, $\tau_{12} = \rho_{12}\dfrac{\sigma_2}{\sigma_1}$, $\sigma_{2|1}^2 = (1-\rho_{12}^2)\sigma_2^2$,

$V(Y_{1j}) = \sigma_1^2$, $V(Y_{2j}) = \sigma_2^2$, $Cov(Y_{1j},Y_{2j}) = \sigma_{12}$, and $Corr(Y_{1j},Y_{2j}) = \rho_{12}$.

The conditional maximum likelihood estimator of E(Y|X), $\hat{\mu}_{Y_2|X_2=x_{2j}{}_{CondML}}$, can be derived by solving $\dfrac{\partial}{\partial\underline{\theta}}\ln L(\underline{\theta}\,|\,\varepsilon) = 0$ where $\underline{\theta} = (\beta_{10},\beta_{11},\sigma_1^2,\beta_{20},\beta_{21},\sigma_{2|1}^2,\tau_{12})'$ and β_{10}, β_{11}, σ_1^2 are assumed to be known. Therefore, the estimator of E(Y|X) under simple linear regression model for incomplete longitudinal data for the second occasion 〈1 is

$$\hat{\mu}_{Y_2|x_{2j}{}_{CondML}} = \hat{\beta}_{20_{CondML}} + \hat{\beta}_{21_{CondML}}x_{2j} \qquad , \quad \text{for } j = 1,2,\ldots,r$$

where

〈1 Detail is shown in dissertation of Juthaphorn Saekhoo, **Simple Linear Regression Analysis for Incomplete Longitudinal Data**, Doctoral dissertation, National Institute of Development Administration.

$$\hat{\beta}_{21_{\text{CondML}}} = \frac{\left(\sum_{j=1}^{r} x_{2j}y_{2j} - r\bar{x}'_2\bar{y}'_2\right)}{\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right)} - \hat{\tau}_{12}\frac{\left(\sum_{j=1}^{r} x_{2j}y_{1j} - r\bar{x}'_2\bar{y}'_1\right)}{\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right)} + \hat{\tau}_{12}\hat{\beta}_{11}\frac{\left(\sum_{j=1}^{r} x_{1j}x_{2j} - r\bar{x}'_1\bar{x}'_2\right)}{\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right)},$$

$$\hat{\beta}_{20_{\text{CondML}}} = (\bar{y}'_2 - \hat{\beta}_{21_{\text{CondML}}}\bar{x}'_2) - \hat{\tau}_{12}(\bar{y}'_1 - \hat{\beta}_{10} - \hat{\beta}_{11}\bar{x}'_1),$$

$$\hat{\beta}_{11} = \frac{\left(\sum_{j=1}^{n} x_{1j}y_{1j} - n\bar{x}_1\bar{y}_1\right)}{\left(\sum_{j=1}^{n} x_{1j}^2 - n\bar{x}_1^2\right)}, \qquad \hat{\beta}_{10} = \bar{y}_1 - \hat{\beta}_{11}\bar{x}_1,$$

$$\hat{\tau}_{12} = \frac{A}{B},$$

$$\begin{aligned} A = {} & \left(\sum_{j=1}^{r} y_{1j}y_{2j} - r\bar{y}'_1\bar{y}'_2\right)\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right) - \hat{\beta}_{11}\left(\sum_{j=1}^{r} x_{1j}y_{2j} - r\bar{x}'_1\,\bar{y}'_2\right)\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right) \\ & - \left(\sum_{j=1}^{r} x_{2j}y_{2j} - r\bar{x}'_2\bar{y}'_2\right)\left(\sum_{j=1}^{r} x_{2j}y_{1j} - r\bar{x}'_2\bar{y}'_1\right) + \hat{\beta}_{11}\left(\sum_{j=1}^{r} x_{2j}y_{2j} - r\bar{x}'_2\bar{y}'_2\right)\left(\sum_{j=1}^{r} x_{1j}x_{2j} - r\bar{x}'_1\bar{x}'_2\right), \end{aligned}$$

$$\begin{aligned} B = {} & \left\{\sum_{j=1}^{r}(y_{1j} - \hat{\beta}_{10} - \hat{\beta}_{11}x_{1j})^2 - \frac{1}{r}\left(\sum_{j=1}^{r}(y_{1j} - \hat{\beta}_{10} - \hat{\beta}_{11}x_{1j})\right)^2\right\}\left(\sum_{j=1}^{r} x_{2j}^2 - r\bar{x}'^2_2\right) - \left(\sum_{j=1}^{r} x_{2j}y_{1j} - r\bar{x}'_2\bar{y}'_1\right)^2 \\ & + 2\hat{\beta}_{11}\left(\sum_{j=1}^{r} x_{1j}x_{2j} - r\bar{x}'_1\bar{x}'_2\right)\left(\sum_{j=1}^{r} x_{2j}y_{1j} - r\bar{x}'_2\bar{y}'_1\right) - \hat{\beta}_{11}^2\left(\sum_{j=1}^{r} x_{1j}x_{2j} - r\bar{x}'_1\bar{x}'_2\right)^2 \end{aligned}$$

$$\bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n} x_{1j}\ ,\quad \bar{y}_1 = \frac{1}{n}\sum_{j=1}^{n} y_{1j}\ ,$$

$$\bar{x}'_1 = \frac{1}{r}\sum_{j=1}^{r} x_{1j}\ ,\quad \bar{y}'_1 = \frac{1}{r}\sum_{j=1}^{r} y_{1j}\ ,$$

$$\bar{x}'_2 = \frac{1}{r}\sum_{j=1}^{r} x_{2j} \text{ and } \bar{y}'_2 = \frac{1}{r}\sum_{j=1}^{r} y_{2j}\ .$$

2.2 Comparison between the Proposed Estimator and Anderson's Estimator

In order to empirically evaluate the proposed estimator, a simulation study was conducted. A population for the first occasion (X_1,Y_1) and the second occasion (X_2,Y_2) of size N = 1,000,000 was generated and the models for (X_1,Y_1) and (X_2,Y_2) are in the form of $Y_1 = 1 + 2X_1 + \varepsilon_1$ and $Y_2 = 3 + 4X_2 + \varepsilon_2$ where $\varepsilon_1 \sim N(0,4)$, $\varepsilon_2 \sim N(0,9)$ respectively. The correlations of (Y_1,Y_2) were set at 0, 0.2, …, 1. Samples of size n = 20, 30, 50 were repeatedly drawn at random with replacement for 2,000 times. In each samples, 0, 10, 20, … , 50 percent of cases were randomly dropped out which were represented as missing.

In this study, 108 situations [2] are employed to compare the coefficient of variation (CV) [3] of CondMLE and Anderson's estimator. Since the value of response means for CondMLE and Anderson's estimator are dissimilar, thus CV of the two estimators are instead compared.

[2] Six values of correlation, three values of sample size, and six percentage level of missing make 108 situations.

[3] $CV_{\text{Anderson}} = \dfrac{\hat{\sigma}_{Y_{2\text{Anderson}}}}{\hat{\mu}_{Y_{2\text{Anderson}}}}$, $CV_{\text{CondML}} = \dfrac{\hat{\sigma}_{Y_2|\bar{x}'_2\ \text{CondML}}}{\hat{\mu}_{Y_2|\bar{x}'_2\ \text{CondML}}}$

Table 1. Comparison of CV for CondMLE and Anderson's estimator (Unit: percent).

n	ρ_{12}	Non-missing data		Missing 10 %		Missing 20 %		Missing 30%		Missing 40%		Missing 50%	
		CV_{CondML}	$CV_{Anderson}$	CV_{CondML}	$CV_{Anderson}$	CV_{CondML}	$CV_{Anderson}$	CV_{CondML}	$CV_{Anderson}$	CV_{CondML}	$CV_{Anderson}$	CV_{CondML}	$CV_{Anderson}$
20	0	3.88	17.72	4.08	17.69	4.40	17.61	4.65	17.54	5.08	17.48	5.50	17.40
	0.2	3.96	17.75	4.18	17.68	4.44	17.63	4.80	17.56	5.14	17.49	5.58	17.34
	0.4	3.99	17.76	4.20	17.70	4.46	17.63	4.80	17.60	5.03	17.58	5.72	17.42
	0.6	4.03	17.70	4.20	17.67	4.54	17.57	4.75	17.51	5.16	17.49	5.63	17.39
	0.8	4.07	17.83	4.29	17.80	4.52	17.69	4.90	17.62	5.25	17.60	5.71	17.47
	1.0	3.99	17.74	4.23	17.68	4.45	17.63	4.76	17.58	5.20	17.41	5.68	17.33
30	0	3.35	17.74	3.50	17.72	3.73	17.69	3.92	17.64	4.26	17.59	4.69	17.49
	0.2	3.19	17.77	3.35	17.73	3.56	17.71	3.83	17.64	4.19	17.61	4.55	17.45
	0.4	3.24	17.72	3.40	17.68	3.59	17.65	3.81	17.63	4.13	17.53	4.58	17.44
	0.6	3.20	17.76	3.36	17.75	3.56	17.71	3.87	17.61	4.14	17.53	4.43	17.58
	0.8	3.13	17.75	3.26	17.74	3.58	17.66	3.71	17.62	4.12	17.58	4.44	17.52
	1.0	3.20	17.77	3.40	17.72	3.55	17.71	3.86	17.68	4.22	17.64	4.57	17.49
50	0	2.55	17.72	2.71	17.70	2.85	17.67	3.01	17.66	3.27	17.60	3.59	17.50
	0.2	2.47	17.76	2.63	17.73	2.78	17.72	2.96	17.69	3.20	17.63	3.57	17.59
	0.4	2.52	17.75	2.66	17.74	2.80	17.71	3.00	17.69	3.28	17.61	3.58	17.60
	0.6	2.48	17.77	2.59	17.74	2.79	17.73	2.97	17.71	3.26	17.68	3.48	17.60
	0.8	2.57	17.79	2.71	17.78	2.81	17.75	2.97	17.71	3.28	17.65	3.57	17.62
	1.0	2.46	17.81	2.59	17.80	2.77	17.78	2.93	17.72	3.22	17.68	3.56	17.61

Remark : $\rho_{12} = \mathrm{Corr}(Y_{1j}, Y_{2j})$ is the correlation of response variable between the first and second occasion.

Results of the simulation study show that CV of CondMLE is smaller than that of Anderson's estimator for sample size 20, 30 and 50, at whatever percentage of missing and correlation coefficient of response variable between two occasions. Furthermore, CondMLE produces smaller CV when the dataset contains large sample size and smaller number of missing data. The results of simulation study are shown in Table 1.

3. Conclusion and Comments

In this study, a response mean estimator for longitudinal studies with incomplete data, MCAR and monotone missing data pattern, is proposed and examined. For the first occasion, both response variable Y_{1j} and auxiliary variable X_{1j} where $j = 1, 2, ..., n$ are completely observed. Moreover, classical assumption for linear model of the relationship between response variable Y and auxiliary variable X is assumed. For the second occasion, when $n - r$ observations are lost, and there are only r observations, $0 < r < n$, collected from the second occasion. The estimator of response mean Y given the value of an auxiliary variable X for the second occasion is proposed. It can be verified that such estimator is unbiased. [4] Results from the Monte Carlo simulations showed that CondMLE presented lower CV than Anderson's estimator which is an auxiliary-information-ignored estimator. In summary, it can be assumed that application of auxiliary information in the estimation of response mean Y can reduce bias and sampling error and thus can be more effective than estimations with no application of auxiliary information. In addition, CV of CondMLE also decreased though the data set contained large sample size with any percentage of missing data, while CV of Anderson's estimator presented no change though there were different percentages of missing data and sample sizes. However, this simulation study showed that CondMLE became more efficiency when the dataset contains larger sample size and smaller amount of missing data are observed.

References

[1] Little, R.J.A. and Rubin, D.B. *Statistical analysis with missing data.* New Jersey: Wiley, 2002.

[2] Anderson, T.W. Maximum likelihood estimates for the multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association*, 1957; 52,200-203.

[3] Acock, A.C. Working with missing values, *Journal of marriage and family*, 2005; 67,1012-1028.

[4] Fitzmaurice, G. Missing data: implications for analysis, *Nutrition*, 2008; 24,200-202.

[4] Detail is shown in dissertation of Juthaphorn Saekhoo, **Simple Linear Regression Analysis for Incomplete Longitudinal Data**, Doctoral dissertation, National Institute of Development Administration.

[5] Gorelick, M.H. Bias arising from missing data in predictive models, *Journal of clinical epidemiology,* 2006; 59, 1115-1123.

[6] Rotnitzky, A. and Wypij, D. A note on the biased of estimators with missing data, *Biometrics,* 1994; 50,1163-1170.

[7] Roth, P.L., Campion, J.E., and Jones, S.D. The impact of four missing data techniques on validity estimates in human resource management, *Journal of business and psychology,* 1996; 11,101-112.

[8] Lord, F.M. Estimation of Parameter from Incomplete Data, *Journal of the American Statistical Association*, 1955; 50,870-876.

[9] Edgett, G.L. Multiple Regression with Missing Observations among the Independent Variables, *Journal of the American Statistical Association*, 1956; 51,122-131.

[10] Rueda, M., Martinez, S., Martinez, H. and Arcos, A. Mean estimation with calibration techniques in presence of missing data, *Computational statistics & data analysis*, 2006; 50,3263-3277.