Journal of Mathematics and Statistics 10 (3): 293-303, 2014
ISSN: 1549-3644

doi:10.3844/jmssp.2014.293.303 Published Online 10 (3) 2014 (http://www.thescipub.com/jmss.toc)

A MIXTURE DROPOUT MECHANISM IN A LONGITUDINAL STUDY WITH TWO TIME POINTS: A METHADONE STUDY

[1]Zohreh Toghrayee, [2]Parvin Jalili, [1]Heng Chin Low and [3]Ardavan Taghva

[1]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia
[2]Research and Studies Demonstration, Central Bank of Iran, Iran
[3]Shamim Clinic, Tehran, Iran

Received 2014-03-10; Revised 2014-04-04; Accepted 2014-04-26

ABSTRACT

One of the most important issues that confront statisticians in longitudinal studies is dropouts. A variety of reasons may lead to withdrawal from a study and produce two different missingness mechanisms, namely, missing at random and non-ignorable dropouts. Nevertheless, none of these mechanisms is tenable in most studies. In addition, it may be that not all of dropouts are nonignorable. Many dropout handling methods have been employed by assuming only one of these dropout mechanisms. In this study, the dropout indicator is improved to take into account both dropout mechanisms. In this two-stage approach, a selection model is combined with an imputation method for dropout process in a longitudinal study with two time points. Simulation studies in a variety of situations are conducted to evaluate this approach in estimating the mean of the response variable at the second time point. This parameter is estimated by using maximum likelihood method. The results of the simulation studies indicate the superiority of the proposed method to the existing ones in estimating the mean of the variable with dropouts. In addition, this method is performed on a methadone dataset of 161 patients admitted to an Iranian clinic to estimate the final methadone dose.

Keywords: Longitudinal Data, Dropout Mechanism, Imputation Method, Ignorability, Non-Ignorability

1. INTRODUCTION

A longitudinal data study is designed to measure the variables of every subject during a specific period. However, dropouts still occur because of different reasons. A dropout is a type of missingness wherein the subject leaves a study after a certain time. One of the most important questions is whether variables with dropouts are related to the values of the outcome variable that describes the dropout mechanism. The dropout mechanism has a principal role in data analysis because parameters that are related to the probability of dropout may affect the parameter estimation of the response's distribution. Hence, choosing an appropriate method to handle dropouts depends on the dropout mechanism. Little and Rubin (1987; Diggle and Kenward, 1994) studied these mechanisms in detail and introduced the following classifications of dropout processes: Completely at Random Dropout (CRD), Random Dropout (RD) and Informative (not at random) dropout (ID).

In CRD, observed data can be considered a random sample and can be analyzed by using common statistical models. Dropouts in CRD are uncorrelated with study variables. Thus, a chance mechanism causes dropout. In RD, the probability of dropout depends on several observed variables but not the response variable. Finally, ID is a situation wherein dropouts are related to the outcome. Both the observed responses and dropout mechanism are modeled in ID. If the dropout mechanism is RD or CRD, the mechanism is called ignorable and if it is ID the dropout mechanism is called nongnorable. Studies generally accept only one of these mechanisms in analyzing data with dropouts.

Two most common methods applied to handle dropouts under ignorable mechanism are imputation

Corresponding Author: Zohreh Toghrayee, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

methods and maximum likelihood method. In imputation methods, dropouts are replaced by values drawn from a specified distribution based on the observed values. A comprehensive detailed texts are provided by (Allison, 2001; Schafer and Graham, 2002; Durrant, 2005; Donders *et al*., 2006; Baraldi and Enders, 2010; Collins *et al*., 2001; Wei and Shih, 2001; Brick *et al*., 2004; Carpenter and Kenward, 2013). On the other hand, linear mixed effect model is a fascinating model for coping with dropouts (missing values) in longitudinal studies which applies maximum likelihood to estimate parameters (Chakraborty and Gu, 2009; Atif *et al*., 2014).

Different models have been recommended for handling non-ignorable dropouts, wherein a joint model of dropout indicator and the outcome variable is assumed. In addition, it was shown that biased estimates can be obtained if the non-ignorability assumption is not considered in the parameter estimation procedure (Wu and Bailey, 1989; Diggle and Kenward, 1994). In selection models, a dropout indicator is applied as:

$$R = \begin{cases} 1, & \text{non} - \text{response} \\ 0, & \text{response} \end{cases}$$

In the selection model, the joint distribution of the outcome and dropout indicators can be factorized as follows Equation 1:

$$f(R, Y \mid X, \gamma, \varphi) = f(Y \mid X, \gamma) f(R \mid X, Y, \varphi) \quad (1)$$

Where:
R = The dropout or missingness indicator,
Y = The response variable and
X = The covariate matrix. Another model is pattern-mixture model in which the factorization is as follows Equation 2:

$$f(R, Y \mid X, \gamma, \varphi) = f(Y \mid R, X, \gamma) f(R \mid X, \varphi) \quad (2)$$

Where:
γ = The parameter of the response variable and
φ = The parameter of the dropout indicator

The cause of dropouts in the selection model is assumed to be a latent variable, R^*; thus, the response is observed and not observed if $R^*>0$ and $R^*\leq$, respectively. In most studies, one of the dropout mechanisms is accepted and the appropriate method is chosen to analyze the data.

Heckman (1976) introduced a selection model to manage non-ignorability. However, this model has several limitations. First, this model assumes that the response variable has a normal distribution and a departure from this assumption will create substantial problems. Second, this model is sensitive to misspecification. Crouchley and Ganjali (2002) introduced a Generalized Heckman selection model and showed that the models proposed by (Wu and Carroll, 1988; Follmann and Wu, 1995; Diggle and Kenward, 1994; Ridder, 1990) can be written by this model.

In this study, we introduce a dropout mechanism indicator instead of a dropout indicator to account for both dropout at random and dropout not at random (nonignorable) mechanisms in a dataset when doubts exist with regard to the real dropout mechanism in a longitudinal study with two time points. To consider all different reasons of leaving a study in the data analysis, researchers should use a model based on a mixture dropout mechanism. Simulation studies are conducted under different conditions to assess the proposed model and the methadone data study is applied to illustrate the new model.

The paper is organized as follows. Section 2 explains the Heckman model and the Generalized Heckman Model (GHM) and introduces the new model with a mixture dropout mechanism. Section 3 presents the simulation studies and the results of three methods under three different dropout mechanisms in the methadone data. Section 4 provides the discussion and conclusion is given in section 5.

2. MATERIAL AND METHODS

Heckman (1976) proposed a joint model for outcome and missingness indicators wherein the missingness indicator is constructed based on a latent variable R^*. This variable is continuous, thus the response observable and unobservable is determined when R^* is positive and negative, respectively. As a special case of missingness, the dropout indicator is expressed as follows:

$$R = \begin{cases} 1, & R^* \geq 0 \\ 0, & R^* < 0 \end{cases}$$

The Heckman (1976) model is identified as follows Equation 3 and 4:

$$R_i^* = W_i'\gamma_i + u_i \quad (3)$$

$$Y_i = X_i'\beta + \varepsilon_i \quad (4)$$

where, W_i and X_i are different covariates, $(u_i, \varepsilon_i) \sim N(0, \Sigma)$, $\Sigma = \begin{bmatrix} \sigma_Y^2 & \sigma_{RY}^2 \\ \sigma_{RY}^2 & \sigma_R^2 \end{bmatrix}$ and $\sigma_R^2 = 1$.

Crouchley and Ganjali (2002) proposed the Generalized Heckman model for longitudinal data. This model for a bivariate normal vector is expressed as follows Equation 5 to 7:

$$Y_{i1} = X'_{i1}\beta + u_{i1} \quad (5)$$

$$Y_{i2} = X'_{i2}\beta + u_{i2} \quad (6)$$

$$R^* = W'_i\gamma + u_{i3} \quad (7)$$

The variance-covariance matrix is unstructured and given as:

$$\Sigma_{GHM} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 \\ \rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 \end{bmatrix}$$

The log-likelihood function for this model is defined as:

$$\begin{aligned} L &= \prod_{i=1}^{r} f\left(Y_1, Y_2, R^*\right) \prod_{i=r+1}^{n} f\left(Y_1, R^*\right) \\ &= \prod_{i=1}^{r} f\left(Y_1\right) f\left(Y_1 \mid Y_2\right) P\left(R^* > 0 \mid Y_1, Y_2\right) \\ &= \prod_{i=1}^{r} f\left(Y_1\right) f\left(Y_1 \mid Y_2\right) P\left(R^* > 0 \mid Y_1, Y_2\right) \end{aligned}$$

Likewise $Y_1, Y_2 | Y_1, R^* | y_1, R^* | y_1, y_2$, have normal distributions with the following parameters:

$$\mu_{Y_1} = X_1'\beta,\ \mu\left(R^* \mid Y_1\right) = \mu_{R^*} + \frac{\rho_{13}}{\sigma}\left(Y_1 - \mu_1\right),$$

$$\mu\left(Y_2 \mid Y_1\right) = X_2'\beta + \frac{\rho_{12}\sigma_2}{\sigma_1}\left(Y_1 - \mu_1\right),$$

$$\mu\left(R^* \mid Y_1, Y_2\right) = W'_\gamma + \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sigma_1\left(1 - \rho_{12}^2\right)}\left(Y_1 - \mu_1\right)$$

$$+ \frac{\rho_{23} - \rho_{12}\rho_{13}}{\sigma_1\left(1 - \rho_{12}^2\right)}\left(Y_2 - \mu_2\right),$$

$$\sigma_{Y_1}^2 = \sigma_1^2, \sigma_{(Y_2|Y_1)}^2 = \sigma_2{}^2\left(1 - \rho_{12}^2\right),$$

$$\sigma_{(R^*|Y_1)}^2 = \left(1 - \rho_{13}^2\right) \text{ and}$$

$$\sigma_{(R^*|Y_1,Y_2)}^2 = 1 - \frac{\rho_{13}^2 + \rho_{23}^2 - 2\rho_{12}\rho_{13}\rho_{23}}{1 - \rho_{12}^2}$$

In most studies, there is no strong proof to test dropout mechanism. If the dropout reasons do not relate to the response values but are associated with the other variables in the study, the ignorable dropout occurs. However, in non-ignorable dropout mechanism, the causes of leaving the study are related to the response values dropped. Hence, it is of special importance to find the reasons of withdrawing from the study. Nevertheless, the researcher cannot prove that the dropout mechanism is non-ignorable unless he or she knows the real causes of dropouts. On the other hand, in some studies, these reasons may vary from one subject to the other. For instance, in a clinical trial, patients do not return to the study because of the side effects of the medicine or they moved out of the area. In this case, considering a single dropout mechanism may lead to invalid results.

In this study, a longitudinal study with two time points is considered in which both variables follow a bivariate normal distribution and the second variable has dropouts. Then, a variety of reasons issue is handled by a two-stage approach based on determining two distinct groups of dropouts: Dropout at random and non-ignorable. In the first group, a stochastic regression model is applied to impute the dropout at random part then the remaining data is assumed to be under non-ignorable mechanism. In other words, dropouts in each part need to be coped with by its own appropriate method. Two groups are specified based on the standard deviation of the R^* distribution. This can be seen in **Figure 1**. To specify two groups, two different classes need to be taken into account: Subjects with observed values and subjects with dropouts displayed in **Fig. 2**.

According to the non-ignorable dropout, the probability of dropout is related to the dropped values. Suppose that the response variable is depression score for patients under a new treatment. During the study, the researcher noticed that patients with high depression score do not return to the study after a certain time. These patients have greater potential to be different from other patients. This is true in other situations in general. Therefore, if the researcher can find this group, the non-ignorable part can be determined. In theory, since a latent variable generates dropouts, it may help us to find this group addressed here.

Suppose that the dropout latent variable is denoted by R^* and its negative values generate the dropouts. In order to improve the dropout indicator, it is assumed that more discrepancy in the left part of the distribution of R^* leads to larger variance in the response variable. It means that, in this part, the values of the response variable are strongly different from the other part.

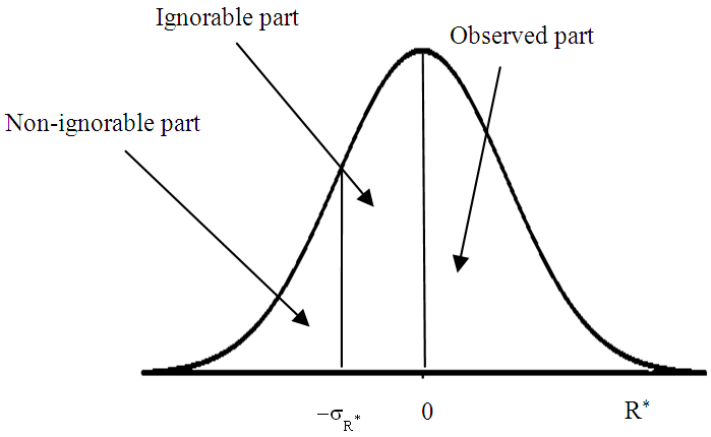

Fig. 1. R^* model distribution

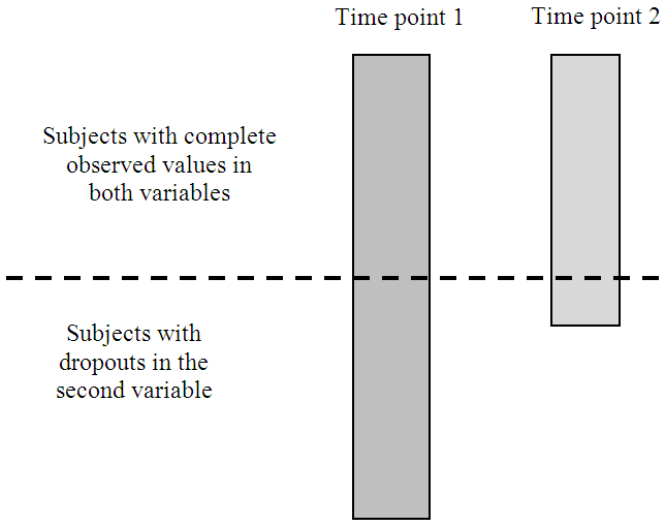

Fig. 2. Two groups: Complete observed values and dropouts in the second variable

Hence, in our model, the new indicator is defined as follows:

$$R_i = \begin{cases} 0, & R^* > 0 \quad y_i \text{ is observed} \\ 1, & -\sigma_{R^*} < R^* < 0 \quad y_i \text{ is missing at random} \\ 2, & R^* < -\sigma_{R^*} \quad y_i \text{ is missing not at random} \end{cases}$$

In fact, it is supposed that values of the R^* with less distance from zero, the threshold value or mean of the distribution of the error term in R^* model, leads to response values which are more similar to complete observed values. The distance is determined based on the standard deviation of the R^* distribution. This can be seen in **Fig. 1**.

Therefore, the likelihood function of the data can be written as follows:

$$L(\theta,\varphi \mid R,Y) = \prod_{\text{observed}} f(Y_1) f(Y_2 \mid Y_1) P(R^* > 0 \mid Y_1, Y_2)$$
$$\times \prod_{\text{missing}} f(Y_1) f(Y_2 \mid Y_1) P(-\sigma_{R^*} < R^* < 0 \mid Y_1, Y_2, \text{ignorable})$$
$$\times \prod_{\text{missing}} f(Y_1) P(R^* < -\sigma_{R^*} \mid Y_1, \text{nonignorable})$$

The likelihood function in terms of the new indicator variable is as given:

$$L(\theta,\varphi \mid R,Y) = \prod_{\text{observed}} f(Y_1) f(Y_2 \mid Y_1) P(R=0 \mid Y_1, Y_2)$$
$$\times \prod_{\text{missing}} f(Y_1) f(Y_2 \mid Y_1) P(R=1 \mid Y_1, Y_2, \text{ignorable})$$
$$\times \prod_{\text{missing}} f(Y_1) P(R=2 \mid Y_1, \text{nonignorable})$$

where, θ is the parameter of the response variable and φ is that of the dropout indicator.

As in most of the selection models, it is assumed that the distribution of the latent variable is normal, it needs to specify a selection model with this assumption in the new model. In this study, imputing the dropouts in the dropout at random part is performed by a regression model. However, we know that using this method leads to underestimation of the standard error of the estimator which is applied as the initial value in the iterative method to obtain maximum lilkelihood. Nevertheless, our goal is to show that even in this simple case of imputing, the final results of estimation of the parameters are plausible. To show how this mixture dropout mechanism is applied, the generalized Heckman model is used as the selection model for a bivariate normal distribution with dropouts in the second variable. Based on this model and the new indicator, the probabilities in the likelihood function based on normal distribution of the dropout indicator are calculated as follows:

Observed part of data, returning to the equation:

$$P(R^* > 0 \mid Y_1, Y_2) = P(\mu_{R^*} + u_3 > 0 \mid Y_1, Y_2) = P(u_3 \geq -\mu_{R^*} \mid Y_1, Y_2)$$
$$= P(u_3 \leq \mu_{R^*} \mid Y_1, Y_2) = P\left(Z \leq \frac{\mu_{R^*}}{\sigma_{R^*}} \mid Y_1, Y_2\right)$$
$$= P\left(Z \leq \frac{\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sigma_1(1-\rho_{12}^2)}(Y_1-\mu_1) + \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sigma_1(1-\rho_{12}^2)}(Y_2-\mu_2)}{\sqrt{1-\frac{\rho_{13}^2+\rho_{23}^2-2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}}}\right)$$
$$= \Phi\left(\frac{\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sigma_1(1-\rho_{12}^2)}(Y_1-\mu_1) + \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sigma_1(1-\rho_{12}^2)}(Y_2-\mu_2)}{\sqrt{1-\frac{\rho_{13}^2+\rho_{23}^2-2\rho_{23}-\rho_{12}\rho_{13}}{1-\rho_{12}^2}}}\right)$$

Dropout at random part:

$$P(R=1 \mid Y_1, Y_2) = P(-\sigma_{R^*} < R^* < 0 \mid Y_1, Y_2)$$
$$= P(-\sigma_{R^*} < \mu_{R^*} + u_3 < 0 \mid Y_1, Y_2)$$
$$= P(-\sigma_{R^*} - \mu_{R^*} < u_3 < -\mu_{R^*} \mid Y_1, Y_2)$$
$$= P(u_3 < -\mu_{R^*} \mid Y_1, Y_2) - P(u_3 < -\sigma_{R^*} - \mu_{R^*} \mid Y_1, Y_2)$$
$$= P\left(Z > \frac{\mu_{R^*}}{\sigma_{R^*}} \mid Y_1, Y_2\right) - P\left(Z > \frac{\sigma_{R^*} + \mu_{R^*}}{\sigma_{R^*}} \mid Y_1, Y_2\right)$$
$$= \Phi\left(1 + \frac{\mu_{R^*}}{\sigma_{R^*}}\right) - \Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\right)$$

Or, we can write:

$$P(R=1 \mid Y_1, Y_2)$$
$$= \Phi\left(1 + \frac{\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sigma_1(1-\rho_{12}^2)}(Y_1-\mu_1) + \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sigma_1(1-\rho_{12}^2)}(Y_2-\mu_2)}{\sqrt{1-\frac{\rho_{13}^2+\rho_{23}^2-2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}}}\right)$$
$$-\Phi\left(\frac{\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sigma_1(1-\rho_{12}^2)}(Y_1-\mu_1) + \frac{\rho_{23}-\rho_{12}\rho_{13}}{\sigma_1(1-\rho_{12}^2)}(Y_2-\mu_2)}{\sqrt{1-\frac{\rho_{13}^2+\rho_{23}^2-2\rho_{12}\rho_{13}\rho_{23}}{1-\rho_{12}^2}}}\right)$$

Non-ignorable part:

$$P(R=2) = P(R^* < -\sigma_{R^*} \mid Y_1) = P(\mu_{R^*} + u_3 < -\sigma_{R^*} \mid Y_{12})$$
$$= P(u_3 < -\sigma_{R^*} - \mu_{R^*} \mid Y_1) = P\left(Z > 1 + \frac{\mu_{R^*}}{\sigma_{R^*}} \mid Y_1\right)$$
$$= 1 - \Phi\left(1 + \frac{\mu_{R^*}}{\sigma_{R^*}} \mid Y_1\right) = = 1 - \Phi\left(1 + \frac{\frac{\rho_{13}}{\sigma_1}(Y_1-\mu_1)}{\sqrt{1-\rho_{13}^2}}\right)$$

Now to obtain the likelihood function, let n be the sample size, r is the number of subjects with complete observed values, $\frac{n-r}{2}$ is the ignorable part size and non-ignorable size too. Therefore the likelihood function is as follows:

$$L(\theta,\varphi \mid Y,R) \propto \prod_{i=1}^{r} \frac{1}{\sigma_1} \exp\left(-\frac{(y_{i1}-\mu_1)^2}{2\sigma_1^2}\right) \times \frac{1}{\sigma_1\sqrt{1-\rho_{12}^2}}$$
$$\exp\left(-\frac{(y_{i2}-\mu_{Y_2|Y_1})^2}{2\sigma_1^2(1-\rho_{12}^2)}\right) \Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}} \mid Y_1, Y_2\right)$$

$$\prod_{i=r+1}^{\frac{n+r}{2}}\frac{1}{\sigma_1}\exp\left(-\frac{\left(y_{i1}-\mu_1\right)^2}{2\sigma_1^2}\right)\times\frac{1}{\sigma_1\sqrt{1-\rho_{12}^2}}\exp\left(-\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)^2}{2\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}\right)$$

$$\times\Phi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)$$

$$\times\prod_{i=\frac{n+r}{2}+1}^{n}\frac{1}{\sigma_1}\exp\left(-\frac{\left(y_{i1}-\mu_1\right)^2}{2\sigma_1^2}\right)\times\left(1-\Phi\left(1+\frac{\mu_{R^*}}{\sigma R}|Y_1\right)\right)$$

Taking the logarithm of this equation to get log-likelihood function, we have:

$$\ln L\left(\theta,\varphi\mid Y_1,Y_2\right)=-r\ln\left(\sigma_1+\sigma_1\sqrt{1-\rho_{12}^2}\right)$$

$$-\frac{1}{2}\sum_{i=1}^{r}\frac{\left(y_{i1}-\mu_1\right)^2}{\sigma_1^{\,2}}+\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)^2}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}+$$

$$\ln\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\left(\frac{n-r}{2}\right)\ln\left(\sigma_1+\sigma_1\sqrt{1-\rho_{12}^2}\right)$$

$$-\frac{1}{2}\sum_{i=r+1}^{\frac{n+r}{2}}\frac{\left(y_{i1}-\mu_1\right)^2}{\sigma_1^{\,2}}+\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)^2}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}$$

$$+\ln\left(\Phi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)\right)-\left(\frac{n-r}{2}\right)$$

$$\ln\left(\sigma_1\right)-\frac{1}{2}\sum_{i=\frac{n+r}{2}+1}^{n}\frac{\left(y_{i1}-\mu_1\right)^2}{\sigma_1^{\,2}}+\ln\left(1-\Phi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1\right)\right)$$

As the main goal is the estimation of mean of Y_2, the derivatives are taken only with respect to μ_2 and σ_2^2, respectively, as follows:

$$\sum_{i=1}^{r}\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}+\frac{\varphi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}{\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}+\sum_{i=r+1}^{\frac{n+r}{2}}\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}$$

$$+\frac{\varphi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\varphi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}{\Phi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}=0$$

$$\sum_{i=1}^{r}\frac{\rho_{12}}{\sigma_1}\times\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}+\frac{\varphi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}{\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}+\sum_{i=r+1}^{\frac{n+r}{2}}\frac{\rho_{12}}{\sigma_1}$$

$$\times\frac{\left(y_{i2}-\mu_{Y_2|Y_1}\right)^2}{\sigma_1^{\,2}\left(1-\rho_{12}^2\right)}+\frac{\varphi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\varphi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}{\Phi\left(1+\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)-\Phi\left(\frac{\mu_{R^*}}{\sigma_{R^*}}\mid Y_1,Y_2\right)}=0$$

The two equations do not have closed forms to yield estimates of μ_2 and σ_2 directly. Hence, these equations need to be estimated by numerically solving the nonlinear system of equations. One of the most convenient nonlinear optimization methods to achieve this computation and maximize the log-likelihood function is the Newton-Raphson algorithm. The properties of the estimates such as bias and efficiency can be evaluated by investigating the behavior of the proposed likelihood function. Simulation studies are used for this model.

In practical situations, determining the two groups, dropout at random and non-ignorable, can be performed through an observed variable highly positively correlated to the response variable such that subjects who dropped out are classified into two different groups based on this variable. Since it is based on this correlation, it is expected that the distributional behaviors of the response variable and this covariate to be similar. In order to find these two groups, group 1 and 2, K-means cluster analysis is used. After classifying all dropouts, there are three distinct groups: Subjects with observed response variable, subjects who dropped out placed in group 1 and subjects who dropped out placed in group 2. To find the non-ignorable part, each of the group 1 and 2 are statistically compared to the first group through Kolmogorov-Smirnov test. The group with more similarity, which is determined by the larger p-value, is considered as dropout at random part and the other one as the non-ignorable group.

In this study, suggested methods are performed in terms of dropout rate and correlation coefficient in three different sample sizes in a bivariate normal distribution.

For simplicity, we consider a bivariate normal distribution as follows:

$$(Y_1, Y_2) \sim N(\mu, \Sigma)\mu = (8,16), \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where, ρ is the correlation coefficient.

Diggle and Kenward (1994) simulated 1000 data sets to evaluate their selection model. Crouchley and Ganjali (2002) introduced the GHM and investigated this model by simulating 100 sets of data with 1000 cases. In the selection model with augmented Gibbs sampling under non-ignorable dropouts suggested by (Yang and Li, 2011), 100 sets of data were generated in the simulation study. Austin and Escobar (2005) conducted a Bayesian modeling of missing data and generated 100 data sets of 1000 subjects. In the current study, we determine the number of simulations to investigate the properties of our methods.

The next step consists of the generation of dropouts. Different methods to create missing data in simulation studies have been proposed. Van Buuren (2007) generated different missing values under different mechanisms. They deleted observations in a completely random method to produce CRD and deleted larger values in both tails of the distribution of independent variables to produce RD. Allison (2000) introduced a simple method to create non-ignorable dropout wherein negative values of outcomes are deleted. Malfert *et al.* (2008) applied a logit model by using outcome and independent variables to create RD and CRD mechanisms. Yang and Li (2011) created non-ignorable dropouts in their selection model by applying the following logistic regression model:

$$\text{logit}(P(R_{12} = 1) \mid y_{i1}, y_{i2}, \phi) = \phi_0 + \phi_1 y_{i1} + \phi_2 y_{i2}$$

If $\phi_2 = 0$ and $\phi_1 = 0$, the dropout mechanism will be non-ignorable and the MAR mechanism will be generated, respectively. Diggle and Kenward (1994) used this model to produce RD and ID mechanisms.

Austin and Escobar (2005) applied logistic regression to generate non-ignorable missing data. In their study, the probability of missingness was estimated by using a logit model. In the current study, the method should consist of ignorable and non-ignorable dropout mechanisms. According to the definition of RD and ID mechanisms, the dropout probability is strongly related to outcome values under ID and to other variables under RD. Hence, dropout probabilities in the second time point are computed for all subjects by a logit function. The different values of the parameters of this model are examined to obtain a data set with 30 and 50% missingness.

In the next step, a stochastic regression imputation method is used to impute the ignorable part. The overall dropout mechanism is then considered as non-ignorable and two models, namely, the mixture and Heckman model, fitted to the obtained data to estimate the mean of Y_2. These models are programmed by R and Splus.

According to the fascinating statistical properties of the new procedure, it is applied to a methadone data. The harmful consequences of drug abuse have been well recognized in recent decades. A number of studies have focused on addressing this problem. One of the most common approaches in investigating this issue is the implementation of clinical treatment wherein a medicine such as methadone is given to drug addicts during a certain time. Medical investigations have shown that methadone treatment is an appropriate treatment to reduce drug use and prevent various diseases such as HIV transmission (Hubbard *et al.*, 1989). A longitudinal study was convened in 1999 by the Institute of Medicine to investigate several changes in treatment practices. Roy and Lin (2002) applied multivariate longitudinal outcomes under non-ignorable dropouts in a methadone study.

One of the most important aspects of a methadone study is the use of adequate dose levels. Therefore, a longitudinal study is necessary to evaluate the response to different methadone dose levels, attain a stable dose and ensure treatment efficacy. An appropriate estimate of a stable methadone dose level as the final dose is obtained based on the dose levels at previous times. However, several patients did not continue with the treatment after some time because of different reasons. So achieving the final dose becomes problematic. Given that we were not able to determine these reasons, the best option is to try different methods.

In this study, a study of an Iranian clinic where drug addicts are treated is considered. Different addicts received methadone at different doses according to their history of drug use. After some time, the dose was changed to achieve a specified stable dose. This dosage was continued until full treatment was completed. In this study, the two last doses are recognized as the most important doses in the treatment to find the final dose level. In the methadone study, a random sample of 161 methadone treatment units was taken from this clinic in 2012. The patients completed the first two time points of the treatment. However, 59 subjects did not return to the study for the final time points (the time point before time of stable dose) to receive the final dose level of methadone. Given that we were not aware of the reasons for the withdrawal of these subjects from the study, all methods are tested to determine the best estimate of the

mean of the final dose of methadone. The methods are tested because determining the average dose level wherein a patient achieved stability is necessary for effective treatment. In addition the new model and two other models that consider all three dropout mechanisms in the data are tested to compare the models precisely. The results show that a lesser MSE value allows the model to capture the behavior of dropout mechanisms better. The performances of these models are obtained using R and Splus programs.

3. RESULTS

The results of the new model with a mixture dropout mechanism and GHM are shown in **Table 1-3**. Four measures are obtained from these studies: The mean estimate of the second variable Y_2, absolute bias, Relative Bias (RB) and mean square error.

The absolute biases in estimating the mean estimate of Y_2 for large sample sizes (n = 100) by the new model and under two different values of dropout rate and correlation coefficients are shown in **Table 1**. The biases with 50% dropouts (i.e., 0.0942 and 0.0943) are more than biases with 25% dropouts (i.e., 0.0828 and 0.0831). Furthermore, RB is computed for all situations in both methods. Previous simulation studies show that RB values <5%, between 5 and 10 and >10% are indicative of minor bias, moderate bias and significant bias, respectively. RBs in all situations are less than 5% for the new model.

The new model with a mixture dropout mechanism, in addition to stronger assumption, is insensitive to dropout rates and correlation coefficients in large samples. This fact is also true for mean square errors.

When the sample size is large, the bias does not change significantly in different conditions. For small and moderate sample sizes, few biases are observed for high correlation cases compared with low correlation cases. The same results can be obtained for Mean Square Errors (MSEs).

In this study, the two last dose levels are considered such that the first one is considered the variable with a high correlation with the second dose level. Therefore, cluster analysis is performed on this variable. The methadone dose at this time point in the non-ignorable group significantly differs with the methadone dose in dropout at random and observed groups (p-value = 7.179479e-12<0.05). By contrast, the second methadone dose in the dropout at random part is similar to the second methadone dose in the observed subjects (p-value = 0.17>0.05).

To investigate the application of the proposed method, we analyze methadone data such that these two doses of methadone are supposed to follow a bivariate normal distribution. The first dose is completely given to participants but several patients do not return to receive the second one. The main goal is to estimate the average methadone dose for the second time where patients achieved a stable dose. In addition to the two existing methods, the GHM and Diggle and Kenward model are used for these data under three different dropout mechanisms.

The MSE of the parameters are computed by using the bootstrap method. The results are shown in **Table 4**.

Table 4 shows that the mean estimates obtained by the GHM and Diggle and Kenward model are close to each other under the ID mechanism. The GHM is sensitive to the dropout mechanism: The estimates of the mean are 100.38, 75.08 and 88.80 mg under ID, RD and CRD, respectively. This sensitivity is not observed in the results obtained from the Diggle and Kenward model. However, the mean of the second dose obtained by using the new model is 78.82 mg, which differs from the other estimates. A comparison of the MSE of the estimates indicates that the new model had an estimate with less MSE. The conclusion that can be drawn from these results is that the new model obtains a superior estimate when the reasons why participants leave a study are unknown.

Table 1. Mean estimate, absolute bias, RB and MSE of Y_2 obtained from the new model and GHM for large samples

		Variables included: (n, m, ρ)[a]			
	Evaluation measures	(100, 50%, 0.9)	(100, 50%, 0.5)	(100, 25%, 0.9)	(100, 25%, 0.5)
The new model	Mean estimate	15.9965	15.9966	16.0206	16.0209
	Bias	0.0942	0.0943	0.0828	0.0831
	RB	-0.0203	-0.0212	0.1291	0.1311
	MSE	0.0141	0.0141	0.0108	0.0109
GHM	Mean estimate	15.0537	15.1297	15.6183	15.6468
	Bias	0.9462	0.8702	0.3820	0.3533
	RB	-5.9139	-5.4391	-2.3855	-2.2071
	MSE	0.9150	0.7766	0.1614	0.1398

a) n is the sample size, m is the missing rate, ρ is the correlation coefficient between Y_1 and Y_2

Table 2. Mean estimate, absolute bias, RB and MSE of Y_2 obtained from the new model and GHM for medium samples

		Variables included: (n, m, ρ)[a]			
	Evaluation measures	(50, 50%, 0.9)	(50, 50%, 0.5)	(50, 25%, 0.9)	(50, 25%, 0.5)
The new model	Mean estimate	16.03240	15.97030	16.0136	16.0429
	Bias	0.14020	0.17020	0.0928	0.1293
	RB	0.20250	-0.18510	0.0854	0.2681
	MSE	0.03208	0.04636	0.0139	0.0248
GHM	Mean estimate	15.01300	15.15820	15.7523	15.6181
	Bias	0.98690	0.84180	0.4263	0.3818
	RB	-6.16830	-5.26150	-3.5874	-2.3864
	MSE	1.00210	0.74360	0.2307	0.1708

a) n is the sample size, m is the missing rate, ρ is the correlation coefficient between Y_1 and Y_2

Table 3. Mean estimate, absolute bias, RB and MSE of Y_2 obtained from the new model and GHM for small samples

		Variables included: (n, m, ρ)[a]			
	Evaluation measures	(25, 50%, 0.9)	(25, 50%, 0.5)	(25, 25%, 0.9)	(25, 25%, 0.5)
The new model	Mean estimate	16.05140	15.9507	16.0231	16.058100
	Bias	0.18250	0.2371	0.1108	0.142300
	RB	0.89210	-0.5320	0.2261	0.423600
	MSE	0.04710	0.0523	0.0173	0.036700
GHM	Mean estimate	15.01460	15.1712	15.7192	15.580200
	Bias	0.99760	0.8523	0.4469	0.419100
	RB	-7.22513	-7.8215	-5.2106	-4.895203
	MSE	1.03140	0.7516	0.2831	0.201800

a) n is the sample size, m is the missing rate, ρ is the correlation coefficient between Y_1 and Y_2

Table 4. The mean estimate of the last dose of Methadone and its mean square error in the mixture-mechanism models and the Generalized Heckman model and Diggle and Kenward model under three different dropout mechanisms

		ID		RD		CRD	
Mechanism method	The new model Generalized Heckman model	Generalized Heckman model	Diggle and Kenward model	Generalized Heckman model	Diggle and Kenward model	Generalized Heckman model	Diggle and Kenward model
$\hat{\mu}_2$	78.82	100.38	93.17	75.08	100.04	88.80	100.04
MSE	7.270142e-30	1.352076e-28	1.92963e-27	6.009227e-29	5.998372e-28	1.446231e-27	9.406882e-28

4. DISCUSSION

We compared the proposed procedure with generalized Heckman model to handle dropouts in a longitudinal data analysis with two time points. These methods were assessed in twelve different settings based on sample size, dropout rate and correlation coefficient between variables at two time points. Computing bias, relative bias and Mean Square Error (MSE) are applied to assess the model performance. In all twelve situations, the results of simulation studies indicate superiority of the proposed method to the existing one. All absolute biases in the new approach are considerably smaller than those of generalized Heckman model. Furthermore, relative biases in all settings are indicative of minor bias in the proposed method compared to moderate and significant bias in the other method. Mean square errors also confirmed the preference of the new method.

In most longitudinal studies with dropouts, only one dropout mechanism is assumed, either dropout at Random (RD) mechanism or dropout not at random (ID) mechanism. Imputation methods and selection models are two widely used methods to handle dropouts at random. Carpenter and Kenward (2013) discussed a variety of imputation methods. Furthermore, this approach was addressed by (Allison, 2001; Schafer and Graham, 2002; Durrant, 2005; Donders *et al*., 2006; Baraldi and Enders, 2010). On the other hand, linear mixed model is one of the advanced methods to deal with dropouts in longitudinal studies (Molenberghs and Kenward, 2007; Chakraborty and Gu, 2009; Atif *et al*., 2014).

On the contrary, in longitudinal studies such as clinical trials with dropouts, it is assumed subjects withdraw from the study because of side effects. Therefore, nonignorable dropout mechanism is introduced. Different selection models were proposed to handle nonignorable dropouts by (Diggle and Kenward, 1994; Heckman, 1976; Little, 1995; Follmann and Wu, 1995; Crouchley and Ganjali, 2002; Yang and Li, 2011). In addition this issue was addressed in detail in advanced statistical books (Molenberghs and Kenward, 2007; Daniels and Hogan, 2008; Fitzmaurice *et al*., 2008; Enders, 2010).

However, in practice, the subjects may withdraw from the study for different kinds of reasons, related to the response variable and unrelated. In this situation, assuming only one dropout mechanism may lead to biased estimation and invalid inference. In all selection models, nonignorable dropouts are generated by negative values of a latent variable on which the dropout indicator is constructed. In the proposed method the dropout indicator was improved bydividing the negative values of the latent variable into two groups, dropouts at random and nonignorable.

In addition to applicability of the proposed method, it is free from restrictive and untestable dropout mechanism. The dropout indicator in this method carries all information about the reasons of loss to follow-up in a longitudinal study.

5. CONCLUSION

Dropouts in longitudinal studies are common when repeated measurements are recorded for the same subjects during experiments. In this study, the subjects withdrew from the study for different reasons such as treatment side effects, movement to a new location and disinterest. In the non-ignorable dropout mechanism, we assume that the dropping out of subjects is caused by the measurement at the dropout time. However, some of the subjects may have undergone the study for this reason. Therefore, considering that non-ignorability is not appropriate a method that accounts for all kinds of reasons is needed.

In this study, along with introducing a mixture dropout mechanism we improved the generalized Heckman model. The findings showed that when in doubt with the dropout mechanism, applying the improved generalized Heckman model is appropriate even with a simple imputation method for handling the dropout at random part. All situations showed that the new model has high preference rather than the existing model.

There are a few limitations in this research. The new model applied a stochastic regression imputation to handle dropouts at random. In addition, it is constructed for a bivariate normal distribution.

In future studies, we can use other methods for imputation or apply covariates in the imputation model. Furthermore, the model can be extended for a multivariate normal distribution or even other distributions.

6. REFERENCES

Allison, P.D., 2000. Multiple imputation for missing data. Sociol. Methods Res., 28: 301-309.

Allison, P.D., 2001. Missing Data. 1st Edn., SAGE Publications, SAGE Publications, ISBN-10: 0761916725, pp: 93.

Atif, M., Z. Toghrayee, S.S. Sulaiman, A.A. Shafie and H.C. Low *et al*., 2014. Missing data analysis in longitudinal studies: Findings from a quality of life study in malaysian tuberculosis patients. Applied Res. Q. Life. DOI: 10.1007/s11482-014-9302-x

Austin, P.C. and M.D. Escobar, 2005. Bayesian modeling of missing data in clinical research. Comput. Stat. Data Anal., 49: 821-836. DOI: 10.1016/j.csda.2004.06.006

Baraldi, A.N. and C.K. Enders, 2010. An introduction to modern missing data analyses. J. School Psychol., 48: 5-37. DOI: 10.1016/j.jsp.2009.10.001

Brick, J.M., G. Kalton and J.K. Kim, 2004. Variance estimation with hot deck imputation using a model. Survey Methodol., 30: 57-66.

Carpenter, J.R. and M.G. Kenward, 2013. Multiple Imputation and its Application. 1st Edn., John Wiley and Sons, Chichester, ISBN-10: 111844261X, pp: 368.

Chakraborty, H. and H. Gu, 2009. A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values. 1st Edn., Research Triangle Park, RTI P. Press, pp: 9.

Collins, L.M., J.L. Schafer and C.M. Kam, 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol. Meth., 6: 330-351. PMID: 11778676

Crouchley, R. and M. Ganjali, 2002. The common structure of several models for non-ignorable dropout. Stat. Modell., 2: 39-62. DOI: 10.1191/1471082x02st022oa

Daniels, M.J. and J.W. Hogan, 2008. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. 1st Edn., CRC Press, ISBN-10: 1420011189, pp: 328.

Diggle, P. and M.G. Kenward, 1994. Informative drop-out in longitudinal data analysis. Applied Stat., 43: 49-93.

Donders, A.R.T., V.D. Heijden, J.M.G. Geert, T. Stijnen and K.G.M. Moons, 2006. Review: A gentle introduction to imputation of missing values. J. Clin. Epidemiol., 59: 1087-1091. PMID: 16980149

Durrant, G.B., 2005. Imputation methods for handling item-nonresponse in the social sciences: A methodological review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute.

Enders, C.K., 2010. Applied Missing Data Analysis. 1st Edn., Guilford Publications, ISBN-10: 1606236407, pp: 377.

Fitzmaurice, G., M. Davidian, G. Verbeke and G. Molenberghs, 2008. Longitudinal Data Analysis. 1st Edn., CRC Press, ISBN-10: 142001157X, pp: 632.

Follmann, D. and M. Wu, 1995. An approximate generalized linear model with random effects for informative missing data. Biometrics, 51: 151-168. PMID: 7766771

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann. Econom. Soc. Measure., 5: 475-492.

Hubbard, L., E. Marsden and V. Racholl, 1989. Drug Abuse Treatment. 1st Edn., The University of North Carolina Press Chapel Hill.

Little, R.J. and D.B. Rubin, 1987. Statistical Analysis with Missing Data. 1st Edn., Wiley, New York, ISBN-10: 0471802549, pp: 278.

Little, R.J.A., 1995. Modeling the drop-out mechanism in repeated-measures studies. J. Am. Stat. Assoc., 90: 1112-1121. DOI: 10.1080/01621459.1995.10476615

Malfert, M., D. Jonson and E.H. Garmo, 2008. Multiple imputing simulated missing data. Sweden, Appsala Univrisity.

Molenberghs, G. and M. Kenward, 2007. Missing Data in Clinical Studies. 1st Edn., John Wiley and Sons, Chichester, ISBN-10: 0470510439, pp: 526.

Ridder, G., 1990. Attrition in Multi-wave Panel Data. 1st Edn., Institute of Economic Research, Faculty of Economics, University of Groningen, pp: 26.

Roy, J. and X. Lin, 2002. Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: Changes in methadone treatment practices. J. Am. Stat. Assoc., 97: 40-52. DOI: 10.1198/016214502753479211

Schafer, J.L. and J.W. Graham, 2002. Missing data: Our view of the state of the art. Psychol. Meth., 7: 147-177. PMID: 12090408

Van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Meth. Med. Res., 16: 219-242. DOI: 10.1177/0962280206074463

Wei, L. and W.J. Shih, 2001. Partial imputation approach to analysis of repeated measurements with dependent drop-outs. Stat. Med., 20: 1197-1214. PMID: 11304736

Wu, M.C. and K.R. Bailey, 1989. Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. Biometrics, 45: 939-955. PMID: 2486189

Wu, M.C. and R.J. Carroll, 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics, 44: 175-188.

Yang, X. and J. Li, 2011. Selection models with augmented Gibbs samplers for continuous repeated measures with nonignorable dropout. University of California-Davis.