مجله‌ی پژوهش‌های آماری ایران

J. Statist. Res. Iran **3** (2006): 75-89

دوره‌ی ۳، شماره‌ی ۱، بهر و تابستان ۱۳۸۵، صص ۷۵-۸۹

A Comparative Review of Selection Models in Longitudinal Continuous Response Data with Dropout

Elaheh Vahidi-Asl and Mojtaba Ganjali*

Shahid Beheshti University

Abstract. Missing values occur in studies of various disciplines such as social sciences, medicine, and economics. The missing mechanism in these studies should be investigated more carefully. In this article, some models, proposed in the literature on longitudinal data with dropout are reviewed and compared. In an applied example it is shown that the selection model of Hausman and Wise (1979, *Econometrica* **47**, pp. 455-473) and the shared parameter model of Follmann and Wu (1995, *Biometrics* **51**, pp. 151-168), two of the most used models for longitudinal data with dropout in economics and medical researches, respectively, cannot sufficiently consider the relation between response variables and missing mechanism. In this paper, the Follmann and Wu's (1995) dropout model is also generalized by adding a previous time outcome component to the model. Having modified this model, in the case of longitudinal data with two time periods, a general form of this model is obtained, which is able to consider all relations between response and missing mechanism. This is proven in an implicit way. A test for missing at random in the generalized Heckman model (Crouchley and Ganjali, 2002, *Stat. Model.* **2**, pp. 39-62) is also introduced where one has to use δ-method to find the variance of the test statistic.

Keywords. longitudinal data; continuous response; missing values; selection bias; dropout; random effect model.

* Corresponding author

1 Introduction

In longitudinal studies, each subject is measured, for some responses, repeatedly at different times. In these studies the missing values commonly occur. Rubin (1976) and Little and Rubin (2002) have done an important classification on missing response mechanism for modeling longitudinal data. In accordance with their taxonomy, missing data mechanism is classified into three different types. These three types are called missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). We shall discuss these mechanisms in section 2. The pattern of missing values may be dropout (monotone) or intermittent (nonmonotone). In the dropout pattern, some subjects withdraw and never come back to the study, but in intermittent missing pattern, observed values may be available even after a missing value occurs. In section 3, we review the Hausman and Wise (1979, hereafter HW), Diggle and Kenward (1994, hereafter DK) and Follmann and Wu (1995, hereafter FW) models for longitudinal data with dropout. Crouchley and Ganjali (2002) explain why these models can not completely describe relations between response and missing mechanisms. In this paper we shall propose an extension of FW model and will show that this extension will lead to the special case of the two-period longitudinal data of the generalized Heckman model (hereafter, GH), introduced by Crouchley and Ganjali (2002). We shall conclude that between the four above mentioned models only GH model can, properly, asses the relationship between response and missing mechanism due to the use of multivariate normal distribution to obtain various conditional distributions, but needs to be used along with a sensitivity analysis. A test for MAR in GH model will also be presented where a δ-method approach is used to find the variance of the test statistic. In section 4, in an applied example of two period longitudinal data, the above mentioned points can be seen practically. The lack of fit of FW model in analyzing these data is what Crouchley and Ganjali (2002) have not covered. In section 5, we have some conclusions.

2 Some Basic Definitions

2.1 Response Indicator Variable

Let R denote a variable that indicates whether the value of the response is observed or not, i.e. $R = 1$ if the response value is observed and $R = 0$ otherwise. It is important to note that R is generated by a latent variable denoted by R^*, that is $R = 1$ when R^* passes a particular threshold point (such as 0, without losing any generality) and $R = 0$ otherwise. The latent variable R^* can be interpreted as the propensity to response of the individual.

2.2 Selection Model

The selection model for the ith response is defined as

$$f(\mathbf{y}_i, \mathbf{r}_i|\psi, \beta, X_i) = f(\mathbf{y}_i|\beta, X_i)P(\mathbf{r}_i|\mathbf{y}_i, \psi, X_i), \quad (1)$$

where $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT})$ is the indicator response vector for T subsequent periods, $\mathbf{y} = (y_{i1}, \ldots, y_{iT})$ denotes the vector of observations, and $f(\mathbf{y}_i \mid \beta, X_i)$ is the probability density function of $\mathbf{y}_i$. Here, X_i is a vector of explanatory variables. The vector of parameters are β and ψ, where β is the parameter of interest and ψ is the missing mechanism parameter. It is assumed that β and ψ are distinct which means these vector of parameters are not functionally related. The expression of selection model in equation (1), in view of $P(\mathbf{r}_i \mid \mathbf{y}_i, \psi, X_i)$, means that, observing response is due to a probabilistic model, which is conditioned on values of the response and explanatory variables.

2.3 The Mechanism of Missing Responses in Longitudinal Studies

Assume that the observed and missed components of $\mathbf{Y}_i$ are denoted as $\mathbf{Y}_{i\text{obs}}$ and $\mathbf{Y}_{i\text{miss}}$. Let $\mathbf{R}_i$ be the vector of response indicators. Under MCAR mechanism, the probability of observing an observation is independent of any responses whether observed or missing. That is

$$P(\mathbf{R}_i = \mathbf{r}_i \mid \mathbf{y}_i, \psi, X) = P(\mathbf{R}_i = \mathbf{r}_i \mid \psi, X_i).$$

Under MAR mechanism we have,

$$P(\mathbf{R}_i = \mathbf{r}_i \mid \mathbf{y}_i, \psi, X_i) = P(\mathbf{R}_i = \mathbf{r}_i \mid \mathbf{y}_{i\text{obs}}, \psi, X_i)$$

where the missing mechanism, given the observed values of responses ($\mathbf{y}_{\text{obs}}$), does not depend on missed responses.

The expression of informative dropout was used by Diggle and Kenward (1994) to describe the NMAR mechanism where the missing mechanism depends on the response values, the values which should have been observed if not missing. In practice, the missing data mechanism is usually NMAR and ignoring this, leads to the bias estimates of parameters (Little and Rubin, 2002, ch. 15).

3 Joint Modeling of Response and Nonresponse in Longitudinal Studies

In this section, four wide-used models, HW, DK, FW, and GH models for longitudinal data are reviewed and compared. Some of these revisions and

comparisons are also given by Crouchley and Ganjali (2002).

3.1 Hausman and Wise Selection Model in Economic and Social Applications

There are two statistical models that, commonly, are used to study the behavior of subjects: the random effect and fixed effect models. In this subsection, we use the random effect model of Hausman and Wise (1979). In particular, we assume a model with two time periods. The regression model for behavior of the subjects is given by

$$Y_{it} = \mathbf{x}'_{it}\beta + U_{it} \quad (i = 1, 2, \ldots, n; t = 1, 2), \tag{2}$$

where i indexes subjects, t denotes the time period, and $\mathbf{x}_{it}$ is a vector of explanatory variables. Errors, U_{it}, are partitioned into two perpendicular components. The first component, μ_i, is the individual effect. These individual effects are assumed to be independent and identically distributed with zero mean and variance σ^2_μ. The second component, v_{it}, is the measurement error. These are independent of μ_i and also assumed to be independent and identically distributed with zero mean and variance σ^2_ν. So U_{it} is partitioned as $U_{it} = \mu_i + v_{it}$ with $E(U_{it}) = 0$ and $\text{var}(U_{it}) = \sigma^2_\mu + \sigma^2_v = \sigma^2$.

In practice, it is often observed that $\sigma^2_\mu > \sigma^2_v$, and this is due to the large differences between the subjects. The correlation between U_{i1} and U_{i2} is

$$\rho_{12} = \frac{\sigma^2_\mu}{\sigma^2_\mu + \sigma^2_v}.$$

Here, it is assumed that y_{i1} is always observed, but y_{i2} is unobserved for some individuals. Suppose that the probability of missingness of y_{i2} depends on y_{i2}, so the mechanism of missingness is not at random. R_i (the response indicator variable) is one if y_{i2} is observed and zero if y_{i2} is not observed. Hausman and Wise (1979) applied the latent variable R^*_i so that R_i is zero if and only if $R^*_i \leqslant 0$, where R^*_i is defined as

$$R^*_i = \alpha y_{i2} + \mathbf{x}'_{i2}\boldsymbol{\theta} + \mathbf{W}'_i\boldsymbol{\gamma} + \omega_i,$$

where $\mathbf{W}_i$ is a vector of variables that do not affect conditional expectation of Y but affect probability of missingness in y_{i2}. The vectors of parameters are $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, the scale parameter is α, and the variables ω_i for $i = 1, 2, \ldots, n$ are independent and identically distributed random variables. By substituting y_{i2} from equation (2) in the model for R^*_i, we reach the following equation:

$$R^*_i = \mathbf{x}'_{i2}(\alpha\boldsymbol{\beta} + \boldsymbol{\theta}) + \mathbf{W}'_i\boldsymbol{\gamma} + \alpha U_{i2} + \omega_i. \tag{3}$$

In (3) let $\boldsymbol{\epsilon} = \alpha\boldsymbol{\beta} + \boldsymbol{\theta}$ and $U_{i3} = \alpha U_{i2} + \omega_i$, then it turns out that

$$R_i^* = \mathbf{x}_{i2}'\boldsymbol{\epsilon} + \mathbf{W}_i'\boldsymbol{\gamma} + U_{i3}.$$

Assuming U_{i2} and ω_i are independent and normally distributed, and defining $\mathbf{Z}_i' = [\mathbf{x}_{i2}', \mathbf{W}_i']$ and $\boldsymbol{\delta} = [\epsilon, \boldsymbol{\gamma}]$, then $R_i^{**} = \mathbf{Z}_i'\boldsymbol{\delta}^* + U_{i3}^*$, where

$$R_i^{**} = \frac{R_i^*}{\sqrt{\text{var}(U_{i3})}} = \frac{R_i^*}{\sqrt{\alpha^2\sigma^2 + \sigma_\omega^2}},$$
$$U_{i3}^* = \frac{U_{i3}}{\sqrt{\text{var}(U_{i3})}} = \frac{U_{i3}}{\sqrt{\alpha^2\sigma^2 + \sigma_\omega^2}},$$

and

$$\boldsymbol{\delta}^* = \frac{\boldsymbol{\delta}}{\sqrt{\text{var}(U_{i3})}} = \frac{\boldsymbol{\delta}}{\sqrt{\alpha^2\sigma^2 + \sigma_\omega^2}}.$$

Since $\text{var}(U_{i3}^*) = 1$, specifying the binary regression model for the indicator of response is possible and the parameters become identifiable (Long, 1997, p. 47). Therefore, the probabilities of observing or not observing y_{i2}, respectively, are defined with probit models as

$$P(R_i = 1) = \Phi(\mathbf{Z}_i'\boldsymbol{\delta}^*),$$
$$P(R_i = 0) = 1 - \Phi(\mathbf{Z}_i'\boldsymbol{\delta}^*),$$

where Φ is the cumulative distribution function of standard normal distribution. The conditional expectation of Y_{i2} given that Y_{i2} is observed can be obtained as (for more details, see Johnson and Kotz, 1972)

$$E(Y_{i2} \mid \mathbf{x}_{i2}, R_i = 1) = \mathbf{x}_{i2}'\boldsymbol{\beta} + \rho_{23}\sigma\frac{\phi(\mathbf{Z}_i'\boldsymbol{\delta}^*)}{\Phi(\mathbf{Z}_i'\boldsymbol{\delta}^*)}, \tag{4}$$

where ρ_{23} is the correlation between U_{i2} and U_{i3}^*, and ϕ is the probability density function of the standard normal distribution. The conditional expectation of Y_{i1} given that Y_{i2} is observed can be obtained as:

$$E(Y_{i1} \mid \mathbf{x}_{i1}, R_i = 1) = \mathbf{x}_{i1}'\boldsymbol{\beta} + \rho_{12}\rho_{23}\sigma\frac{\phi(\mathbf{Z}_i'\boldsymbol{\delta}^*)}{\Phi(\mathbf{Z}_i'\boldsymbol{\delta}^*)}. \tag{5}$$

Due to the assumptions of the HW model, the correlations are related as

$$\rho_{13} = \rho_{12}\rho_{23},$$

where ρ_{13} is the correlation between U_{i1} and U_{i3}^*. From equations (4) and (5), it is realized that the important parameter in determining the bias of selection

is the correlation between U_{i2} and U_{i3}^*. The hypothesis to test whether the dropout mechanism is completely at random or not, is defined as

$$H_0 : \rho_{23} = 0.$$

This hypothesis can be tested by using, for example, a generalized likelihood ratio test. Variables U_{i1}, U_{i2}, and U_{i3}^* have joint normal distribution with zero mean and the covariance matrix

$$\Sigma_{\mathrm{HW}} = \begin{pmatrix} \sigma^2 & \rho_{12}\sigma^2 & \rho_{12}\rho_{23}\sigma \\ \rho_{12}\sigma^2 & \sigma^2 & \rho_{23}\sigma \\ \rho_{12}\rho_{23}\sigma & \rho_{23}\sigma & 1 \end{pmatrix}.$$

It is important to note that $\mathrm{cov}(U_{i1}, U_{i3}^*)$ and $\mathrm{cov}(U_{i2}, U_{i3}^*)$ depend on ρ_{23}. Our criticism of HW model is that it lacks distinction between MAR and MCAR mechanisms. This can be corrected by adding a previous response as a covariate in the model of R_i^* (see next subsection).

3.2 Diggle and Kenward Model for Dropout in Clinical Applications

Diggle and Kenward (1994) propose a model as

$$Y_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta} + U_{i1}, \tag{6}$$
$$Y_{i2} = \mathbf{x}_{i2}'\boldsymbol{\beta} + U_{i2}, \tag{7}$$
$$R_i^* = \gamma_0 + \gamma_1 y_{i1} + \gamma_2 y_{i2} + U_{i3}. \tag{8}$$

which can better describe the relation between the missing mechanism and response variables in comparing with HW model. That is, this model considers the effect of current response (y_{i2}) and previous response $(y_{i,1})$ in R_i^*. The errors in equations (6) to (8) have zero mean and the covariance matrix as follows.

$$\Sigma_{\mathrm{DK}} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 \\ \sigma_{12} & \sigma_2^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Furthermore, Diggle and Kenward (1994) consider the logistic link function for dropout mechanism. It is important to note that this model distinguishes MCAR and MAR mechanisms. In subsection 3.4 we shall see that this model gives a different form of GH model for two-period longitudinal data. So, by adding a previous outcome in the model for missing mechanism, we may improve HW model with the property that having the same missing mechanism as DK model, which itself is the same as GH model.

3.3 The Shared-parameter Random Effect Model of Follmann and Wu

In this section, similar to HW model, we consider the random effect model. FW model is defined as

$$\begin{aligned} Y_{i1} &= \mathbf{x}_{i1}'\boldsymbol{\beta} + \mu_i + v_{i1}, \\ Y_{i2} &= \mathbf{x}_{i2}'\boldsymbol{\beta} + \mu_i + v_{i2}, \\ R^* &= \mathbf{W}_i'\boldsymbol{\alpha} + \theta\mu_i + v_{i3}. \end{aligned}$$

The variance covariance matrix of (Y_{i1}, Y_{i2}, R_i^*) is given by

$$\Sigma_{\mathrm{FW}} = \begin{pmatrix} \sigma_\mu^2 + \sigma_{v_1}^2 & \sigma_\mu^2 & \theta\sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_{v_2}^2 & \theta\sigma_\mu^2 \\ \theta\sigma_\mu^2 & \theta\sigma_\mu^2 & 1 \end{pmatrix}.$$

For the identifiability of parameters in model R^*, it is necessary to impose the condition

$$\theta^2\sigma_\mu^2 + \sigma_{v_3}^2 = 1.$$

In this model, the missing mechanism is ignorable when $\theta = 0$. This model cannot also distinguish the MAR mechanism from the MCAR mechanism (see subsection 3.5).

3.4 Generalized Heckman Model

Crouchly and Ganjali (2002), using Heckman model (1979) for cross sectional studies, have proposed a more general model for longitudinal data that is known as GH model. This model is defined as

$$\begin{aligned} Y_{i1} &= \mathbf{x}_{i1}'\boldsymbol{\beta} + U_{i1}, \\ Y_{i2} &= \mathbf{x}_{i2}'\boldsymbol{\beta} + U_{i2}, \\ R_i^{**} &= \mathbf{Z}_i'\boldsymbol{\delta}^* + U_{i3}^*. \end{aligned}$$

Here, there is no structural dependency between errors and, therefore, the variance-covariance matrix is defined as

$$\Sigma_{\mathrm{GH}} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2 \\ \rho_{13}\sigma_1 & \rho_{23}\sigma_2 & 1 \end{pmatrix}.$$

The GH model is more general than HW model, since variances of responses could be different and also there is no restriction on the correlation of responses

with R_i^{**}. If we evaluate the adjusted form of DK model for two time periods, by substituting Y_{i1} and Y_{i2} of equations (6) and (7) into equation (8), then it can be seen that the two models (DK and GH) are identical. In GH model the missing mechanism is MCAR if

$$\rho_{13} = \rho_{23} = 0,$$

and it is MAR if

$$\rho_{23} = \rho_{12}\rho_{13}$$

(for more details, see Ganjali and Rezaee, 2005). Suppose $\rho_{12} \neq 0$ and $\rho_{13} \neq 0$, and let $h = \rho_{23} - \rho_{12}\rho_{13}$. The function h may be estimated by using the invariant property of maximum likelihood as $\hat{h} = \hat{\rho}_{23} - \hat{\rho}_{12}\hat{\rho}_{13}$. The δ-method is then used to find an estimate for the variance of $\hat{h}$ (see Appendix A). This is useful for testing MAR mechanism against NMAR.

3.5 Generalized Follman and Wu Model

Consider the FW model. As previously noted, there is no relationship between the missing mechanism and responses if $\theta = 0$, and in this case the mechanism is MCAR. We may add a component to this model in order to distinguish between the two mechanisms (MCAR and MAR) when $\theta = 0$. To do this we assume that the mechanism of missingness is related to the past response (y_{i1}). Then, the full model would be

$$\begin{aligned} Y_{i1} &= \mathbf{x}_{i1}'\boldsymbol{\beta} + \mu_i + v_{i1}, \\ Y_{i2} &= \mathbf{x}_{i2}'\boldsymbol{\beta} + \mu_i + v_{i2}, \\ R^* &= \mathbf{W}_i'\boldsymbol{\alpha} + \theta\mu_i + \gamma y_{i1} + v_{i3}. \end{aligned}$$

In this model, for testing MCAR mechanism, the hypotheses $\gamma = 0$ and $\theta = 0$ are required to be tested. To test MAR, we just need to test that $\theta = 0$. In the following, we show, in an implicit way, that including this component to the FW model, makes it equivalent to the GH model for two-period longitudinal data.

By substituting Y_{i1} in R_i^*, we have the following equations:

$$\begin{aligned} Y_{i1} &= \mathbf{x}_{i1}'\boldsymbol{\beta} + \mu_i + \nu_{i1}, \\ Y_{i2} &= \mathbf{x}_{i2}'\boldsymbol{\beta} + \mu_i + \nu_{i2}, \\ R_i^* &= \mathbf{W}_i'\boldsymbol{\alpha} + \gamma(\mathbf{x}_{i1}'\boldsymbol{\beta}) + (\theta + \gamma)\mu_i + \gamma\nu_{i1} + \nu_{i3}. \end{aligned}$$

By defining

$$\nu_{i3}^* = \frac{\gamma\nu_{i1} + \nu_{i3}}{\sqrt{Var(\gamma\nu_{i1} + \nu_{i3})}} = \frac{\gamma\nu_{i1} + \nu_{i3}}{\sqrt{\gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2}},$$

the equations can be rewritten as

$$\begin{aligned} Y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + \mu_i + \nu_{i1}, \\ Y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + \mu_i + \nu_{i2}, \\ R_i^* &= \frac{\mathbf{W}'_i\boldsymbol{\alpha} + \gamma(\mathbf{x}'_{i1}\boldsymbol{\beta})}{\sqrt{\gamma^2\sigma^2_{\nu_1} + \sigma^2_{\nu_3}}} + \frac{(\theta+\gamma)}{\sqrt{\gamma^2\sigma^2_{\nu_1} + \sigma^2_{\nu_3}}}\mu_i + \nu^*_{i3}. \end{aligned}$$

We shall consider this model in the form of GH model as

$$\begin{aligned} Y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + U_{i1}. \\ Y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + U_{i2}. \\ R_i^* &= \mathbf{Z}'_i\boldsymbol{\alpha}^* + U_{i3}, \end{aligned}$$

where

$$\begin{aligned} U_{i1} &= \mu_i + \nu_{i1}, \\ U_{i2} &= \mu_i + \nu_{i2}, \\ U_{i3} &= \frac{(\theta+\gamma)}{\sqrt{\gamma^2\sigma^2_{\nu_1} + \sigma^2_{\nu_3}}}\mu_i + \nu^*_{i3}, \\ \mathbf{Z}'_i\boldsymbol{\alpha}^* &= \frac{\mathbf{W}'_i\boldsymbol{\alpha} + \gamma(\mathbf{x}'_{i1}\boldsymbol{\beta})}{\sqrt{\gamma^2\sigma^2_{\nu_1} + \sigma^2_{\nu_3}}}. \end{aligned}$$

and

$$\mathrm{var}(U_{i3}) = \left(\frac{\theta+\gamma}{\sqrt{\gamma^2\sigma^2_{\nu_1} + \sigma^2_{\nu_3}}}\right)^2 \sigma^2_\mu + 1.$$

Defining

$$U^*_{i3} = \frac{U_{i3}}{\sqrt{\mathrm{var}(U_{i3})}},$$

where $\mathrm{var}(U^*_{i3}) = 1$, the FW model becomes

$$\begin{aligned} Y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + U_{i1}. \\ Y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + U_{i2}. \\ R_i^{**} &= \mathbf{Z}'_i\boldsymbol{\alpha}^{**} + U^*_{i3}, \end{aligned}$$

where

$$\alpha^{**} = \frac{\alpha^*}{\sqrt{\mathrm{var}(U_{i3})}}.$$

To show that the above model is a form of GH model, we may write the covariance matrix of $(Y_{i1}, Y_{i2}, R_i^{**})$ as

Table 1. Number and percentage of cows in the first and the second periods of mastitis data for jth selected year ($j = 1, 2, \ldots, 5$)

Selected year	1	2	3	4	5
No. of cows in the first period	9	27	25	23	23
Percentage of cows in the first period	8.4	25.2	23.4	21.5	21.5
No. of cows in the second period	6	19	19	15	21
Percentage of cows in the second period	66.7	70.4	76.0	65.2	91.3

$$\Sigma_{\mathrm{GH}} = \begin{pmatrix} \sigma_\mu^2 + \sigma_{\nu_1}^2 & \sigma_\mu^2 & \frac{(\theta+\gamma)\sigma_\mu^2 + \gamma\sigma_{\nu_1}^2}{\sqrt{(\theta+\gamma)^2\sigma_\mu^2 + \gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2}} \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_{\nu_2}^2 & \frac{(\theta+\gamma)\sigma_\mu^2}{\sqrt{(\theta+\gamma)^2\sigma_\mu^2 + \gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2}} \\ \frac{(\theta+\gamma)\sigma_\mu^2 + \gamma\sigma_{\nu_1}^2}{\sqrt{(\theta+\gamma)^2\sigma_\mu^2 + \gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2}} & \frac{(\theta+\gamma)\sigma_\mu^2}{\sqrt{(\theta+\gamma)^2\sigma_\mu^2 + \gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2}} & 1 \end{pmatrix}.$$

With respect to the covariance matrix, the dropout mechanism is completely at random if $\theta = \gamma = 0$, and in the case of $\theta = 0$ the relation between correlations is

$$\rho_{23} = \rho_{12}\rho_{13},$$

which means that the dropout mechanism is at random. It is important to know that for identifiability of parameters in the probit model the following restriction should be held:

$$(\theta + \gamma)^2\sigma_\mu^2 + \gamma^2\sigma_{\nu_1}^2 + \sigma_{\nu_3}^2 = 1.$$

4 An Application

4.1 Mastitis Data

The mastitis disease in cows could decrease the amount of milking. Diggle and Kenward (1994) studied the amount of milking of 107 cows in two successive time periods. The aim was to find the relation between amount of milking and the disease of mastitis. In each of five years, a group of cows in their third lactation (which may happen in any of these 5 years) and free of mastitis is selected and the amount of milking are recorded for two successive years. In this study, 27 of 107 chosen cows became infected. The yield of milking of these 27 cows for second period are supposed to be missed. Table 1 shows the number and the percentage of chosen cows in the jth year ($j = 1, 2, \ldots, 5$) and also denotes the number and percentage of cows, which have no infection in second period.

It is seen that the percentage of cows which were present (did not have mastitis) in the second time period, is minimum in the 4th selected year (with value of 65.2%) and it has the maximum in the 5th selected year (with value of 91.3%).

4.2 Model Building

The following GH model can be used for the mastitis data to find the explanatory variable and time effects on the mean of responses and also the relation between the responses and the missing mechanism

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_{i1}, \\ Y_{i2} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \eta + \epsilon_{i2}, \\ R_i^* &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + v_{i3}, \end{aligned}$$

where the explanatory variables are defined as

$$x_{ij} = \begin{cases} 1, & \text{if the } i\text{th cow is chosen in the } j\text{th year} \\ 0, & \text{otherwise.} \end{cases}$$

Here η shows the time effect on the second response mean. The **NAG** (1996) program, or function **optim** in R may be used to maximize the logarithm of the likelihood function given by Crochley and Ganjali (2002). We have also fitted DK model to the data. The form of this model is

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_{i1}, \\ Y_{i2} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \eta + \epsilon_{i2}, \\ R_i^* &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \gamma_1 y_{i1} + \gamma_2 y_{i2} + e_{i3}, \end{aligned}$$

where it is assumed that there is no correlation between the missing mechanism error term (e_{i3}) and response errors (ϵ_{i1} and ϵ_{i2}). The form of the HW model that we use, is a special case of DK model, where the previous outcome is not included in the model for R_i^*. The following FW model is also fitted:

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \mu_i + \nu_{i1}, \\ Y_{i2} &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \eta + \mu_i + \nu_{i2}, \\ R_i^* &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \theta\mu_i + \nu_{i3}. \end{aligned}$$

Results of fitting these models are given below.

Table 2. Parameter estimates and their standard errors for fitting different models with mastitis data (* significant at 0.01 level, ** significant at 0.05 level, GH: Generalized Heckman model, DK: Diggle and Kenward model, HW: Hausman and Wise model, FW: Follmann and Wu model)

	GH model		DK		HW		FW	
Parameters	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**	**Est.**	**S.E.**
β_0	5.624**	0.189	5.624**	0.189	5.624**	0.196	5.581**	0.187
β_1	0.158	0.356	0.158	0.356	0.110	0.366	0.235	0.353
β_2	−0.031	0.252	−0.031	0.252	0.052	0.262	0.032	0.252
β_3	0.324	0.262	0.324	0.262	0.232	0.266	0.359	0.261
β_4	0.277	0.264	0.277	0.264	0.298	0.274	0.337	0.263
η	0.293*	0.143	0.293*	0.143	0.762**	0.118	0.736**	0.108
ρ_{12}	0.467**	0.088	0.467**	0.088	0.562**	0.072		
ρ_{13}	−0.148**	0.132						
ρ_{23}	0.729**	0.101			−0.165	0.193		
σ_1	0.911**	0.630	0.911**	0.063	1.020**	0.059	0.436**	0.107
σ_2	1.310**	0.122	1.310**	0.122			0.846**	0.088
α_0	0.965*	0.335	0.509**	1.855	1.378**	0.371	1.536**	0.269
α_1	−0.031	0.597	−0.114	1.476	−0.965	0.574	−0.925	0.571
α_2	−0.449	0.376	−1.100	0.595	−0.829	0.447	−0.823	0.447
α_3	−0.165	0.371	−0.483	0.941	−0.689	0.464	−0.647	0.465
α_4	−0.697	0.370	−1.789	1.213	−0.985*	0.458	−0.965	0.457
γ_1			−1.696*	0.794				
γ_2			1.926*	0.807				
θ							−0.191	0.201
σ_μ							0.784	0.108
$-\log l$	304.441		304.441		310.717		307.075	

4.3 Results

Results of fitting the four models, GH, DK, HW (where $\rho_{13} = \rho_{12}\rho_{23}$ and $\sigma_1 = \sigma_2$), and FW models are presented in Table 2. In this table the logarithm of likelihood includes the constant value $1/\sqrt{2\pi}$. As seen in Table 2, the value of log likelihood for the GH model is equivalent to that of DK model. This shows that GH and DK models give the same results for $\boldsymbol{\beta}$ and, in general, for the relationship between responses and missing mechanism. Table 2 also shows that for the GH model, the missing mechanism is not ignorable, since the missing process is related to the response of the second period ($\hat{\rho}_{23} = .729$). The value

of $\hat{\rho}_{23}$ says that, the larger the value of the response in the second period, the more is the probability of response in the second period. Of course, a better test for MAR against MNAR is to test $h = \rho_{23} - \rho_{12}\rho_{13} = 0$ against $h \neq 0$. Using Appendix A, we find $\hat{h} = \hat{\rho_{23}} - \hat{\rho_{12}}\hat{\rho_{13}} = 0.798$ and $S.E.(\hat{h}) = 0.084$, which gives a p-value of 0.00. This shows that the missing mechanism is NMAR.

Also the probability of not having mastitis for the selected cows in the 4th year is the lowest ($\alpha_4 = -0.697$). Whereas, the probability of observing the cows in the 5th year is the largest. Results for DK model also show that the missing mechanism is not at random, but they do not claim that the missing of responses depends on covariates. From the results for HW model we can conclude that the missing mechanism is completely at random ($\rho_{23} = -.165,\ S.E. = 0.193$). All the models show the same significant value for the effect of time in the average of response in the second period. But the HW model overestimates the parameters. The FW model does not reject hypothesis of missing completely at random, and gives nearly the same results as the HW model. This application as well as our theoretical views in previous sections confirm that the restrictions, which are imposed on HW and FW model, causes these models not to properly investigate the relationship between responses and missing mechanisms.

5 Recommendation and Concluding Remarks

Dealing with missing data should be done with care. Assuming that missing data are at random, without any testing approach, and using methods such as expectation maximization algorithm or multiple imputation to estimate the parameters gives biased estimates if missing data, in fact, are not at random. Although the use of joint modeling to, simultaneously, model response and missing mechanism can give a way to test for missing at random, it can be misleading if a wrong model is chosen. In this paper, for example, we show that FW or HW models can not properly assess the relation between response and missing mechanisms. On the other hand, GH model, in our example, did the job perfectly well and, in general, can be preferred to the other joint modeling approaches. This is due to having a nonstructural covariance matrix for relations between response and missing mechanisms in GH model. However, like any other model, GH model needs to be done along with some sensitivity analysis. For one of these analyzes for GH model see Ganjali and Rezaei (2005). Future works on sensitivity analysis for data with missing values needs to be investigated.

References

Crouchley, R.; Ganjali, M. (2002). The common structure of several models for non-ignorable dropout. *Stat. Model.* **2**, 39-62.

Diggle, P.J.; Kenward, M.G. (1994). Informative dropout in longitudinal data analysis. *Appl. Statist.* **43**, 49-93.

Follmann, D.; Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151-168.

Ganjali, M.; Rezaei, M. (2005). An influence approach for sensitivity analysis of non-random dropout based on the covariance structure. *Iran. J. Sci. Technol. Trans. A Sci.* **29**. 287-294.

Hausman, J.A.; Wise, D.A. (1979). Attrition bias in experimental and panel data: The Gary income maintaince experiment. *Econometrica* **47**, 455-473.

Heckman, J.(1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimation for such models. *Ann. Econ. Soc. Meas.* **5**, 475-492.

Heckman, J. (1979). Sample selection bias as a specification error. *Biometrics* **47**.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage, Thousand Oaks, CA.

Johnson, N.L.; Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

Little, R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90**, 1112-1121.

Little, R.J.A.; Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Wiley, New York.

Mood, A.M.; Graybill, F.A. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill, New York.

NAG (1996). Numerical Algorithms Group Manual, Mark 16. NAG, Oxford, UK.

Rubin. D.B. (1976). Inference and missing data. *Biometrica* **63**, 581-592.

Appendix A: δ-method

Suppose $\boldsymbol{\beta}$ is the parameter vector to be estimated and the ML estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}$. Let $\text{cov}(\hat{\boldsymbol{\beta}})$ denote the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. Under regularity conditions (Mood et al., 1974, pp. 315-316) $\text{cov}(\hat{\boldsymbol{\beta}})$ is evaluated by the inverse of information matrix. The (j,k) element of the information matrix is given by

$$-E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\beta_j \partial\beta_k}\right),$$

where $\ell(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta}))$ is the log-likelihood function. Standard errors are the square roots of diagonal elements of the inverse information matrix. ML estimates have large-sample normal distribution; they are asymptotically consistent and asymptotically efficient when the model is correctly specified. Furthermore, any subset of β has also large-sample normal distribution. For our example, the vector

$$\hat{\gamma} = \begin{pmatrix} \hat{\rho}_{12} \\ \hat{\rho}_{13} \\ \hat{\rho}_{23} \end{pmatrix} = \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}$$

has large-sample normal distribution with mean $\gamma = (\rho_{12}, \rho_{13}, \rho_{23})'$ and covariance matrix which can be found by partitioning of the inverse information matrix. Let show this matrix by Σ_γ. Now, assume that $\rho_{12} \neq 0$ and $\rho_{13} \neq 0$. One may use the δ-method to find the variance of $h = \rho_{23} - \rho_{12}\rho_{13}$. This variance can be used to test MAR against NMAR in the generalized Heckman model. As $h(t_1, t_2, t_3)$ has nonzero differential $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$ at γ, where

$$\phi_i = \left.\frac{\partial h}{\partial t_i}\right|_{t=\gamma}.$$

then

$$h(t_1, t_2, t_3) \xrightarrow{\text{D}} N(h(\rho_{12}, \rho_{13}, \rho_{23}), (\phi_1, \phi_2, \phi_3)'\Sigma_\gamma(\phi_1, \phi_2, \phi_3)).$$

This distribution is used to test MAR ($h = 0$) against NMAR.

Elaheh Vahidi-Asl
Department of Statistics,
Faculty of Mathematical Sciences,
Shahid Beheshti University,
Evin,
Tehran, Iran.
e-mail: *elahehva@yahoo.com*

Mojtaba Ganjali
Department of Statistics,
Faculty of Mathematical Siences,
Shahid Beheshti University,
Evin,
Tehran, Iran.
e-mail: *m-ganjali@sbu.ac.ir*