A note on the robustness of a full Bayesian method for non-ignorable missing data analysis

Zhiyong Zhang and Lijuan Wang
University of Notre Dame

Abstract

A full Bayesian method utilizing data augmentation and Gibbs sampling algorithms is presented for analyzing non-ignorable missing data. The discussion focuses on a simplified selection model for regression analysis to demonstrate the rationale of the introduced method. Regardless of missing mechanisms, it is always assumed that missingness only depends on the missing variable itself. Simulation studies are conducted and demonstrate that the selection model can recover model parameters under both correctly specified situations and many mis-specified situations. The method is also applied to analyzing a training intervention data set.

Missing data problem is a big challenge in statistical inference even for a well designed study. Little & Rubin (2002) distinguished three kinds of missing data mechanisms – missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). For the MCAR mechanism, every subject (datum) has the same probability to be missing. For example, if a participant's test score is missing because his/her experimenter accidentally forgets to give the participant the test, then the resulting missing datum can be viewed as MCAR. If missingness can be fully predicted by observed data in a study, the missing mechanism is MAR. For example, in a pre-test post-test experiment, all subjects participated in the pre-test. Some subjects missed the post-test because they did not perform well in the pre-test. In this case, missingness during the post-test can be predicted using data from the pre-test. However, if missingness cannot be fully predicted by observed data, missing data are MNAR. For example, in survey research, when asking about salary, those with high salary often choose not to respond. Thus, missing data on income are often believed to be MNAR.

To some extent, MCAR and MAR data are ignorable because problems caused by them can be overcome through sophisticated statistical techniques such as the full information likelihood (FIML) method and multiple imputation (e.g., Little & Rubin, 2002; Schafer, 1997). MNAR data, however, are non-ignorable because without extra information/modeling, their influences cannot be well addressed. To deal with non-ignorable missing data, selection models (see recent reviews by Ibrahim et al., 2006 and Ibrahim & Molenberghs, 2009) and the MI method with auxiliary variables (e.g., Graham, 2009) can be used. However, both selection models and MI typically require extra/auxiliary variables (information) to account for non-ignorable missingness.

Best et al. (1996) proposed a selection model that simplified the modeling of missingness to study cognitive decline in the elderly. The model assumed that missingness in a variable was only related to itself and the model, therefore, did not require auxiliary variables to model the non-ignorable missing mechanism. Through empirical data analysis, they demonstrated that model parameters of interest were insensitive to different prior specifications on the parameters predicting missingness. In this article, we examine and extend the model by Best et al. in several ways with a focus on the robustness of the model. First, we derive the full Bayesian posterior distribution for the simplified selection model. Second, we evaluate the performance of the model under different conditions. Third, we apply the model to analyze a set of training intervention data to illustrate the use of the model.

In the remainder of the paper, we will first discuss selection models and the simplified version by Best et al. (1996). Then, we will discuss selection model estimation through a full Bayesian method by obtaining the full sets of conditional posterior distributions for model parameters utilizing the data augmentation algorithm. After this, we will evaluate the performance of the model and the estimation method through several simulation studies. Finally, we will present an empirical example to demonstrate how to apply the model and method in behavioral research.

Selection models

Consider a multiple regression model with a dependent variable y and a vector of independent variables $\boldsymbol{x}$. The regression model can be expressed as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + e_i \tag{1}$$

where y_i and $\boldsymbol{x}_i$ are observed data for the i-th person, $\boldsymbol{\beta}$ is a vector of regression coefficients, and e_i is residual that is often assumed to be normally distributed such that $e_i \sim N(0, \phi)$. If both the dependent and independent variables are fully observed, estimates of the regression coefficients can be obtained conveniently.

However, data are often incomplete or not fully observed for a variety of reasons such as non-response and dropout. For example, we consider a scenario that the dependent variable y is partially observed and independent variables $\boldsymbol{x}$ are fully observed. This scenario is very common in designed experiments and survey research. For example, in an experiment, the controlled factors are often pre-determined and thus are known. However, data of the outcome variables may not be always observed. In survey research, subjects may respond to non-sensitive questions but are less likely to answer sensitive questions.

Let m_i be an indicator variable where $m_i = 0$ if y_i is observed and $m_i = 1$ if y_i is missing. Then, the missing probability of y_i, in general, can be modeled as

$$\Pr(y_i \text{ is missing}) = \Pr(m_i = 1) = f(\gamma_0 + \gamma_1 y_i + \boldsymbol{\alpha}\boldsymbol{v}_i) \tag{2}$$

where $\boldsymbol{v}$ is a set of variables that may be related to the missingness of y. The variables in $\boldsymbol{v}$ may or may not be fully observed and could be latent variables. $\boldsymbol{v}$ may also include part of or all variables of $\boldsymbol{x}$. The γ_0, γ_1, and $\boldsymbol{\alpha}$ are regression coefficients. The link function f can be any function that maps its input to an output value from 0 to 1. For example, if f is a logistic function, Equation (2) becomes a logistic regression model.

Equation (2) models the missing mechanism of y. The missing mechanisms defined by Little & Rubin (2002) can be distinguished according to the parameter values in Equation (2). If $\gamma_1 = 0$ and $\boldsymbol{\alpha} = 0$, the missing probability is a constant and the missing mechanism is MCAR. If $\gamma_1 = 0$, and $\boldsymbol{\alpha} \neq 0$ and $\boldsymbol{v}$ are fully observed, the missing mechanism is MAR. If $\gamma_1 \neq 0$ or the coefficients in $\boldsymbol{\alpha}$ of the partially observed or unobserved variables in $\boldsymbol{v}$ are not equal to zero, missing data are MNAR.

Selection models (e.g., Heckman, 1976; Little & Rubin, 2002) focus on modeling the joint distribution of observed data and missing indicators as

$$p(y_i, m_i|\beta, \phi, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{x}_i, \boldsymbol{v}_i) = p(y_i|\boldsymbol{x}_i, \beta, \phi)p(m_i|y_i, \boldsymbol{\gamma}, \boldsymbol{v}_i, \boldsymbol{\alpha}) \tag{3}$$

where $p(.)$ represents the probability density function and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$. A possible choice for $p(m_i|y_i, \boldsymbol{\gamma}, \boldsymbol{v}_i, \boldsymbol{\alpha})$ is given in Equation (2). Model parameters in selection models can be estimated by maximizing the following likelihood function (e.g., Little, 1982; Little & Rubin, 2002)

$$L = \int_{\boldsymbol{v}_i^{miss}} \int_{y_i^{miss}} \prod_{i=1}^{n} p(y_i, m_i|\beta, \phi, \gamma, \boldsymbol{\alpha}, \boldsymbol{x}_i, \boldsymbol{v}_i) dy_i^{miss} \boldsymbol{v}_i^{miss} \tag{4}$$

where y_i^{miss} denotes a missing datum in y and $\boldsymbol{v}_i^{miss}$ denotes unobserved data in $\boldsymbol{v}$ for individual i. To identify selection models, $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ should not completely overlap or the values of λ or $\boldsymbol{\alpha}$ for some covariate(s) are known or pre-constrained (e.g., Little, 1985; Olsen, 1980; Tang et al., 2003).

To properly apply selection models, variables in $\boldsymbol{v}$ have to be determined carefully, which often hinders practical adoptions of selection models. Best et al. (1996) discussed a simplified selection model with $\boldsymbol{\alpha} \equiv 0$. Thus, the model assumes that missingness in y is only related to itself and no auxiliary variables need to be used in the model. One of the advantages of the model is avoiding the difficulty of selecting auxiliary variables. Although the function f in Equation 2 can take many different forms, the logistic function (the logit link) and the normal distribution function (the probit link) are most widely used (e.g., Ibrahim et al., 2006; Little & Rubin, 2002). Best et al. used the logistic function in Equation 2. Here we use the normal distribution function because of its convenience in obtaining posterior distributions and the aim to provide an alternative configuration on the missing mechanism. Therefor, the missing mechanism in the simplified selection model can be specified as

$$\begin{cases} m_i & \sim Bernoulli(p_i) \\ p_i & = \Phi(\gamma_0 + \gamma_1 y_i) = \Phi\left[\begin{pmatrix} 1 & y_i \end{pmatrix}\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix}\right] = \Phi(\boldsymbol{w}_i\boldsymbol{\gamma}) \end{cases} \tag{5}$$

where p_i is the probability that y_i is missing and Φ is the normal distribution function. m_i is a binary variable following a Bernoulli distribution. Furthermore, $\boldsymbol{w}_i = (1, y_i)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)'$.

Note that if the missing mechanism is MCAR or MNAR depending on y only, the missing mechanism is correctly specified. Otherwise, the missing mechanism is mis-specified, e.g., under the situation of MAR. For convenience, we refer to the model in Equations (1) and (5) together as the simplified selection model and the model in Equation (1) as the regular model in the remainder of the paper. Because of the focus of Best et al.'s study, they did not present the posterior distributions of the selection model. Furthermore, they did not investigate under what conditions the simplified selection model would/would not work. Thus, in this study, we will present posterior distributions of the simplified selection model and then evaluate its performance under a variety of conditions.

Full Bayesian estimation method

To estimate parameters in the simplified selection model, we will use the Bayesian estimation method based on data augmentation (Albert & Chib, 1993; Tanner & Wong, 1987) and Gibbs sampling (e.g., Casella & George, 1992). We first assume that there is an underlying normal variable $z_i \sim N(\boldsymbol{w}_i\boldsymbol{\gamma}, 1)$ for each m_i. If $z_i > 0$, then $m_i = 1$. Otherwise, $m_i = 0$. In other words, if $z_i > 0$, y_i is unobserved. By augmenting z_i with m_i, the joint distribution of m_i and z_i is

$$p(m_i, z_i|y_i, \boldsymbol{\gamma}) = p(m_i|z_i)p(z_i|y_i, \boldsymbol{\gamma}).$$

The distribution of z_i - $p(z_i|y_i, \boldsymbol{\gamma})$ -is known as a normal distribution with mean $\boldsymbol{w}_i\boldsymbol{\gamma}$ and variance 1 and we need to get the distribution for m_i conditional on z_i. Note that

$$\begin{aligned} p(m_i &= 1|z_i > 0) = 1, \;\; p(m_i = 1|z_i \leq 0) = 0, \\ p(m_i &= 0|z_i > 0) = 0, \;\; p(m_i = 0|z_i \leq 0) = 1. \end{aligned}$$

Thus, the distribution for $m_i|z_i$ can be expressed as

$$p(m_i|z_i) = \mathcal{I}(m_i = 1)\mathcal{I}(z_i > 0) + \mathcal{I}(m_i = 0)\mathcal{I}(z_i \leq 0)$$

where $\mathcal{I}(A)$ is an indicator function which takes 1 if the expression A is true and otherwise 0.

Furthermore, by augmenting missing data y_i^{miss} with observed data, m_i and z_i, the joint distribution of y_i, m_i and z_i is

$$\begin{aligned} p(y_i, z_i, m_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi, \boldsymbol{X}) &= p(y_i|\boldsymbol{\beta}, \phi, \boldsymbol{X})p(m_i|z_i)p(z_i|y_i, \boldsymbol{\gamma}) \\ &= \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{2\phi}\right] \\ &\quad \times [\mathcal{I}(m_i = 1)\mathcal{I}(z_i > 0) + \mathcal{I}(m_i = 0)\mathcal{I}(z_i \leq 0)] \\ &\quad \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(z_i - \gamma_0 - \gamma_1 y_i)^2}{2}\right]. \end{aligned} \tag{6}$$

Thus, the likelihood function for the selection model with augmented data can be expressed as

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{m}) = \prod_{i=1}^{n} p(y_i, z_i, m_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi, \boldsymbol{X}) \tag{7}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denotes a vector of data for the dependent variable and $\boldsymbol{X} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$ is the design matrix. Furthermore, $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and $\boldsymbol{m} = (m_1, \ldots, m_n)'$. By integrating out missing data y_i^{miss} and the underlying variable z_i, one can obtain the observed data likelihood for the maximum likelihood estimation method. However, the integration is not an easy task. Bayesian estimation procedure, however, can be implemented relatively easily through Gibbs sampling after obtaining the full set of conditional posterior distributions for the selection model.

To use the Bayesian method, we need to specify priors for unknown parameters. We first consider the following semi-conjugate priors and then discuss the use of uninformative priors (Gelman et al., 2003). For the regression coefficients $\boldsymbol{\beta}$, a multivariate normal prior is used as

$$p(\boldsymbol{\beta}) = MN(\boldsymbol{\beta}_0, \Sigma_0) \propto |\Sigma_0|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right] \tag{8}$$

where $\boldsymbol{\beta}_0$ and Σ_0 are predefined hyper-parameters representing the mean vector and covariance matrix of the multivariate normal distribution. For the residual variance parameter ϕ, an inverse gamma distribution is employed,

$$p(\phi) = IG(a_0, b_0) \propto \phi^{-a_0/2-1} \exp\left(-\frac{b_0}{2\phi}\right) \tag{9}$$

where a_0 and b_0 are assumed to be known shape and scale parameters. Finally, for parameters $\boldsymbol{\gamma}$, a multivariate normal prior is also used,

$$p(\boldsymbol{\gamma}) = MN(\boldsymbol{\gamma}_0, D_0) \propto |D_0|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)'D_0^{-1}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)\right] \tag{10}$$

where $\boldsymbol{\gamma}_0$ and D_0 are known mean vector and covariance matrix. These priors are called semi-conjugate priors because the corresponding conditional posteriors are from the same distribution family (Gelman et al., 2003).

With the likelihood function in (7) and priors in (8), (9), and (10), the joint posterior distribution of the unknown parameters is readily available. However, the marginal posterior distributions of the parameters are difficult to obtain explicitly. To avoid the difficulty of getting the marginal posterior distributions explicitly, we obtain the conditional distributions of the parameters and then utilize the Gibbs sampling method to generate Markov chains for the parameters and construct the Bayesian parameter estimates.

The conditional posterior distribution for $\boldsymbol{\beta}$ is a multivariate normal distribution defined by

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \phi \sim MN(\boldsymbol{\beta}_1, \Sigma_1)$$

where

$$\boldsymbol{\beta}_1 = \left(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\phi^{-1}\right)^{-1} \left(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}'\boldsymbol{y}\phi^{-1}\right),$$

and

$$\Sigma_1 = \left(\Sigma_0^{-1} + \boldsymbol{X}'\boldsymbol{X}\phi^{-1}\right)^{-1}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ are as defined earlier.

The conditional posterior distribution for ϕ is an inverse Gamma distribution given by

$$\phi|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X} \sim IG(a_1, b_1)$$

where

$$a_1 = \frac{a_0 + n}{2}$$

and

$$b_1 = \frac{b_0 + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2}.$$

The conditional posterior distribution for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}|\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{W} \sim MN(\boldsymbol{\gamma}_1, D_1)$$

where

$$\boldsymbol{\gamma}_1 = \left(D_0^{-1} + \boldsymbol{W}'\boldsymbol{W}\right)^{-1} \left(D_0^{-1}\boldsymbol{\gamma}_0 + \boldsymbol{W}'\boldsymbol{z}\right)$$

and

$$D_1 = \left(D_0^{-1} + \boldsymbol{W}'\boldsymbol{W}\right)^{-1}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1', \ldots, \boldsymbol{w}_n')'$.

The conditional posterior distribution for the underlying variable z_i is

$$z_i|\boldsymbol{\gamma}, y_i, m_i \sim \begin{cases} N(\boldsymbol{w}_i\boldsymbol{\gamma}, 1)I(0, +\infty) & m_i = 1 \\ N(\boldsymbol{w}_i\boldsymbol{\gamma}, 1)I(-\infty, 0] & m_i = 0 \end{cases}.$$

Thus, z_i follows a truncated normal distribution.

Finally, the conditional posterior distribution for missing data y_i^{miss} is

$$y_i^{miss}|\boldsymbol{\beta}, \boldsymbol{\gamma}, z_i, \phi, \boldsymbol{x}_i \sim N(\mu_1, \phi_1)$$

where

$$\mu_1 = \left[\frac{\boldsymbol{x}_i\boldsymbol{\beta}}{\phi} + \gamma_1(z_i - \gamma_0)\right]\left(\frac{1}{\phi} + \gamma_1^2\right)^{-1}$$

and

$$\phi_1 = \left(\frac{1}{\phi} + \gamma_1^2\right)^{-1}.$$

Note that if missing data are MCAR, one should expect that $\gamma_1 = 0$. Then the posterior for missing data y_i^{miss} reduces to

$$y_i^{miss}|\boldsymbol{\beta}, \boldsymbol{\gamma}, z_i, \phi, \boldsymbol{x}_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \phi)$$

which can be viewed as the multiple imputation method.

Uninformative priors can also be used in obtaining the conditional posterior distributions of the model parameters. One form of uninformative priors can be

$$p(\phi) \propto 1/\phi$$

and

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto 1.$$

Note that these priors are improper and can be viewed as carrying no information (Box & Tiao, 1973). With these priors, the conditional posterior distributions for z_i and y_i^{miss} remain the same. The conditional posterior distributions for the model parameters become

$$\begin{aligned}
\boldsymbol{\beta}|y_i, \boldsymbol{x}_i, \phi &\sim MN\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\phi\right] \\
\phi|\boldsymbol{\beta}, y_i, \boldsymbol{x}_i &\sim IG\left[\frac{n}{2}, \frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2}\right] \\
\boldsymbol{\gamma}|z_i, y_i &\sim MN\left[(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{Z}, (\boldsymbol{W}'\boldsymbol{W})^{-1}\right],
\end{aligned}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$ and $\boldsymbol{Z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_n')'$.

With the conditional posterior distributions, one can implement the following Gibbs sampling procedure.

1. Start with initial values $\boldsymbol{\beta}^{(0)}, \phi^{(0)}, \boldsymbol{\gamma}^{(0)}, z_i^{(0)}, y_i^{miss(0)}$.

2. Assume at the iteration t, one has $\boldsymbol{\beta}^{(t)}, \phi^{(t)}, \boldsymbol{\gamma}^{(t)}, z_i^{(t)}, y_i^{miss(t)}$.

3. At iteration $t+1$,

   (a) Generate $\boldsymbol{\beta}^{(t+1)}$ from $\boldsymbol{\beta}|y_i^{obs}, y_i^{miss(t)}, \boldsymbol{x}_i, \phi^{(t)}$,

   (b) Generate $\phi^{(t+1)}$ from $\phi|\boldsymbol{\beta}^{(t+1)}, y_i^{obs}, y_i^{miss(t)}, \boldsymbol{x}_i$,

   (c) Generate $z_i^{(t+1)}$ from $z_i|\boldsymbol{\gamma}^{(t)}, y_i^{obs}, y_i^{miss(t)}, m_i$ for $i = 1, \ldots, n$,

   (d) Generate $\boldsymbol{\gamma}^{(t+1)}$ from $\boldsymbol{\gamma}|z_i^{(t+1)}, y_i^{obs}, y_i^{miss(t)}$,

   (e) Generate $y_i^{miss(t+1)}$ from $y_i^{miss}|\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}, z_i^{(t+1)}, \phi^{(t+1)}, \boldsymbol{x}_i$ if y_i is missing.

The above Gibbs sampling procedure can be used to generate a Markov chain for each model parameter, underlying variable, and missing datum. After convergence, these Markov chains can be viewed as samples from the joint distribution and marginal distributions of the parameters and thus Bayesian parameter estimates can be constructed (e.g., Casella & George, 1992; Geman & Geman, 1984).

Simulation studies

In this section, we conduct several simulation studies to investigate the performance of the selection model and the Bayesian estimation method. The focus of the simulation studies is to evaluate whether the Bayesian selection model is robust to choices of priors, choices of link functions, and missing data mechanisms. In the first study, we compare the parameter estimates from the selection model and the regular regression model assuming MNAR and the missing mechanism is known to depend solely on $\boldsymbol{y}$. In the second simulation, we evaluate whether results from the selection model are influenced by different choices of priors and in the third simulation, we investigate whether results from the selection model are influenced by different choices of the link functions. In Simulation 4, the missing mechanism is related to an external unobserved variable. In Simulation 5, the missingness depends on both the dependent variable $\boldsymbol{y}$ and independent variables $\boldsymbol{X}$. In Simulation 6, the missing mechanism is MCAR.

General settings of the simulation studies

In all simulation studies, data are generated from a multiple regression model with two covariates as in

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i = \boldsymbol{x}_i \boldsymbol{\beta} + e_i. \tag{11}$$

The population parameters are $\beta_0 = \beta_1 = \beta_2 = 1$ and $e_i \sim N(0, \phi)$ with variance $\phi = .25$. Both covariates x_{1i} and x_{2i} are generated from the standard normal distribution. Complete data are first generated from this model and missing data are then generated according to each simulation condition. In all simulation studies, sample size is $n = 100$ and the missing data percentage is 40%. Furthermore, for each simulation study, a total of 1000 data sets with missing data are generated and analyzed.

For each simulation study, four statistics will be reported. The first one is the parameter estimate which is calculated as the average of parameter estimates of 1000 simulation replications. The second one is the average standard error (ASE) which is the average of standard errors of parameter estimates from the 1000 replications. The third one is the standard deviation of the 1000 sets of estimated parameters. Finally, the coverage probability of the 95% credible (confidence) interval of each parameter are also reported.

In each simulation study, the following priors are used without further elaboration. For $\boldsymbol{\beta}$, the trivariate normal prior is used with $\boldsymbol{\beta}_0 = (0, 0, 0)'$ and

$$\Sigma_0 = \begin{pmatrix} 10^6 & 0 & 0 \\ 0 & 10^6 & 0 \\ 0 & 0 & 10^6 \end{pmatrix}.$$

For ϕ, the inverse gamma prior is used with $a_0 = b_0 = 10^{-3}$. And for $\boldsymbol{\gamma}$, the bivariate normal prior is used with $\boldsymbol{\gamma}_0 = (0, 0)'$ and

$$D_0 = \begin{pmatrix} 10^6 & 0 \\ 0 & 10^6 \end{pmatrix}.$$

These priors can be considered as carrying little information (Congdon, 2003).

All simulations are conducted using SAS and WinBUGS (Zhang et al., 2008). The convergence of the Markov chains is monitored through the Geweke statistics (Geweke, 1992). The WinBUGS codes for the selection model are provided in the Appendix.

Simulation 1: Missingness depends on $\boldsymbol{y}$ only

Purpose. This simulation study investigates whether the regression model parameters can be recovered using the selection model and the Bayesian estimation method when the missing mechanism is correctly specified.

Missing data generation. In this simulation study, the probability that a datum is missing is assumed to depend on itself. Thus, the missingness is non-ignorable. Missing data are generated in the following way. Let c_α denote the 100α-th percentile of $\boldsymbol{y}$. Then, the probability that y_i is missing is set as

$$\Pr(y_i \text{ is missing}) = \Pr(m_i = 1) = \begin{cases} 0.9r/(1-\alpha) & \text{if } y_i > c_\alpha \\ 0.1r/\alpha & \text{otherwise} \end{cases},$$

where r is the predefined missing data rate. In this simulation study, we set $\alpha = 60\%$ and $r = .4$. Note that the missing probability function is a step function instead of a continuous function. It also indicates that when y_i is larger than the 60th percentile, its missing probability is 0.9. Otherwise, its missing probability is about 0.067.

Results. The simulated data are analyzed using the selection model in Equations (1) and (5). For the purpose of comparison, we also analyze the data by ignoring the missing mechanism, namely, the data are analyzed as MAR through a regular regression model. The results from the analysis are summarized in Table 1.

When the missing mechanism is MNAR and the data are analyzed ignoring the missing mechanism, the parameter estimates are biased - underestimated in this simulation study. Furthermore, the coverage probabilities are not correct, much smaller than 95% especially for the regression coefficients ($\boldsymbol{\beta}$). When the data are analyzed using the selection model, the parameters are well recovered, especially for the regression coefficients, with less than 0.6% bias. The coverage probabilities are also very close to .95. Finally, the ASE and SD are very close for the regression parameters, which indicates that the standard error estimates of parameters are also accurate. Note that although the missing mechanism is modeled using the normal distribution function and the missingness is not simulated from the normal distribution function but a step function, the regression model parameter estimates ($\boldsymbol{\beta}$) are still accurate. This indicates that the non-ignorable missing mechanism does not need to be perfectly specified.

Table 1 Estimate MNAR data using the selection model and the regular model

	Selection model				Regular model			
Parameters	Estimates	ASE	SD	Coverage	Estimates	ASE	SD	Coverage
β_0(Intercept)	0.998	0.082	0.081	0.947	0.855	0.075	0.076	0.477
$\beta_1(x_1)$	0.994	0.075	0.077	0.945	0.922	0.074	0.076	0.808
$\beta_2(x_2)$	0.997	0.075	0.076	0.932	0.924	0.074	0.077	0.823
ϕ	0.261	0.056	0.054	0.945	0.236	0.045	0.044	0.922
γ_0	-1.488	0.365	0.503	-	-	-	-	-
$\gamma_1(y)$	1.023	0.218	0.323	-	-	-	-	-

Note. The results are based on 1000 simulation replications. ASE: average standard error. SD: standard deviation of parameter estimates. Coverage: coverage probability of the 95% highest posterior density credible interval.

Simulation 2: Semi-conjugate priors vs. uninformative priors (sensitivity analysis of priors)

Purpose. During the discussion of the model estimation method, we used two types of priors - the semi-conjugate priors and the uninformative priors. This simulation study is to investigate whether the estimated model parameters are influenced by the choice of the two sets of priors.

Missing data generation. The data are generated using the same procedure of Simulation 1.

Results. The results from the analysis using two types of priors are given in Table 2. From the results, the parameter estimates, especially the regression coefficients ($\boldsymbol{\beta}$), are very close from two different types of priors. The results for using the uninformative priors can be viewed as the baseline data analysis. Prior information, when available, can be incorporated into the semi-conjugate priors.[1]

Simulation 3: Normal distribution function vs. logistic function

Purpose. In Equation (2), the f function can be any function that maps a value to be within 0 and 1. For convenience, we have used the normal distribution function. With it, the explicit forms of the conditional posterior distributions can be obtained conveniently. However, the other functions such as the

[1]Best et al. (1996) compared three sets of informative priors and found that parameter estimates were insensitive to the chosen priors.

Table 2 Selection model with different priors

	Semi-conjugate priors				Uninformative priors			
Parameters	Estimates	ASE	SD	Coverage	Estimates	ASE	SD	Coverage
β_0(Intercept)	0.998	0.082	0.081	0.947	1.004	0.084	0.082	0.954
$\beta_1(x_1)$	0.994	0.075	0.077	0.945	0.997	0.077	0.077	0.947
$\beta_2(x_2)$	0.997	0.075	0.076	0.932	1.000	0.076	0.077	0.933
ϕ	0.261	0.056	0.054	0.945	0.272	0.060	0.057	0.950
γ_0	-1.488	0.365	0.503	-	-1.514	0.385	0.553	-
$\gamma_1(y)$	1.023	0.218	0.323	-	1.036	0.229	0.356	-

Note. The same as the previous table.

logistic function can also be used as in Best et al. (1996). In this simulation study, we investigate whether the choice of the function f such as the normal distribution function or the logistic function influences parameter estimates.

Missing data generation. The data are generated using the same procedure of Simulation 1.

Results. The results from the selection model using normal distribution and logistic functions are provided in Table 3. The results are almost identical for the regression parameters ($\boldsymbol{\beta}$). For the γs, the estimates using the logistic distribution is about 1.83 times of those using the normal distribution. This simulation suggests that the selection model again does not require the link function to match the missing mechanism exactly.

Table 3 Selection model with different f

	Normal distribution function				Logistic function			
Parameters	Estimates	ASE	SD	Coverage	Estimates	ASE	SD	Coverage
β_0(Intercept)	0.998	0.082	0.081	0.947	0.996	0.080	0.078	0.958
$\beta_1(x_1)$	0.994	0.075	0.077	0.945	0.994	0.074	0.075	0.947
$\beta_2(x_2)$	0.997	0.075	0.076	0.932	0.997	0.074	0.074	0.933
ϕ	0.261	0.056	0.054	0.945	0.256	0.054	0.052	0.946
γ_0	-1.488	0.365	0.503	-	-2.721	0.686	0.663	-
$\gamma_1(y)$	1.023	0.218	0.323	-	1.903	0.439	0.420	-

Note. The same as the previous table.

Simulation 4: Missing data depend on an external and unobserved variable z

Purpose. In this simulation, we consider the situation that the missingness of $\boldsymbol{y}$ depends on an external variable $\boldsymbol{z}$ that is related to $\boldsymbol{y}$ but unobserved. The purpose of this simulation is to investigate whether we can recover model parameters using the selection model when the missing mechanism is mis-specified.

Missing data generation. To generate data, z_i is first generated using

$$z_i = ay_i + v_i$$

where $v_i \sim N(0, 1)$. To investigate whether the correlation between y and z influences parameter estimates, we set a at 1 and 0.2 corresponding to the correlations (ρ) between$y$ and z at 0.83 and 0.29, respectively.

Let c_α denote the 100α-th percentile of $\boldsymbol{z}$. Then, the probability that y_i is missing is set as

$$\Pr(y_i \text{ is missing}) = \Pr(m_i = 1) = \begin{cases} 0.9r/(1-\alpha) & \text{if } z_i > c_\alpha \\ 0.1r/\alpha & \text{otherwise} \end{cases},$$

where r is the missing data rate. As in the previous simulations, we set $\alpha = 60\%$ and $r = .4$.

Results. Results from this simulation are summarized in Table 4. Biases in parameter estimates ($\boldsymbol{\beta}$) are small and the coverage probabilities are close to 95% regardless of the size of the correlation between the external variable and the outcome variable. Thus, the regression model parameters can still be recovered well using the selection model even when the missing mechanism is mis-specified in the current situation.

Table 4 Missing mechanism is related to an external variable z

	$a = 1$ and $\rho = .83$				$a = .2$ and $\rho = .29$			
Parameters	Estimates	ASE	SD	Coverage	Estimates	ASE	SD	Coverage
β_0(Intercept)	1.009	0.078	0.078	0.945	1.003	0.067	0.067	0.952
$\beta_1(x_1)$	1.001	0.072	0.072	0.955	0.998	0.066	0.067	0.942
$\beta_2(x_2)$	1.000	0.072	0.071	0.946	1.003	0.066	0.069	0.944
ϕ	0.263	0.055	0.052	0.960	0.259	0.049	0.047	0.952
γ_0	-1.061	0.258	0.252	-	-0.557	0.174	0.127	-
$\gamma_1(y)$	0.680	0.150	0.162	-	0.191	0.099	0.103	-

Note. The same as the previous table.

Simulation 5: Missing data depend on both $\boldsymbol{y}$ and $\boldsymbol{X}$ (a mixture of MNAR and MAR)

Purpose. This simulation study investigates whether the selection model is robust to the missing mechanism that the missing probability depends on both $\boldsymbol{y}$ and $\boldsymbol{X}$. The missing mechanism can be viewed as a mixture of MNAR and MAR. Thus, the missing mechanism is also mis-specified in the selection model in this simulation study.

Missing data generation. To generate missing data, we first generate a variable $\boldsymbol{z}$ such that

$$z_i = .5y_i + x_{1i} - .5x_{2i}.$$

Then the missing probability is set by

$$\Pr(y_i \text{ is missing}) = \Pr(m_i = 1) = \begin{cases} 0.9r/(1-\alpha) & \text{if } z_i > c_\alpha \\ 0.1r/\alpha & \text{otherwise} \end{cases},$$

where c_α is the 100α-th percentile of $\boldsymbol{z}$. In this simulation, we have $\alpha = 60\%$ and $r = .4$.

Results. Results from this simulation study are given in Table 5. The biases of the regression coefficient parameters ($\boldsymbol{\beta}$) are slightly larger than the previous simulations. For example, for the intercept β_0, the relative bias is about 3.8%. However, the bias is still small (less than 5%). Furthermore, the coverage probabilities seem correct. Thus, the selection model appears to work well under the current condition.

Simulation 6: MCAR data analysis using the selection model

Purpose. This simulation study investigates whether the selection model can be applied when the missing mechanism is MCAR. Note that for MCAR, the selection model is still correctly specified and one would expect that $\gamma_1 = 0$.

Table 5 Missing probability is related to both $\boldsymbol{y}$ and $\boldsymbol{X}$

Parameters	Estimates	ASE	SD	Coverage
β_0(Intercept)	1.038	0.084	0.082	0.934
$\beta_1(x_1)$	1.026	0.080	0.077	0.954
$\beta_2(x_2)$	0.998	0.073	0.072	0.949
ϕ	0.271	0.059	0.055	0.953
γ_0	-1.438	0.364	0.476	-
$\gamma_1(y)$	0.948	0.209	0.289	-

Note. The same as the previous table.

Missing data generation. To generate missing data, the missing probability is set as

$$\Pr(y_i \text{ is missing}) = \Pr(m_i = 1) = r,$$

where r is a constant and is set as .4 in this simulation.

Results. Results from this simulation are given in Table 6. First, the estimate of γ_1 is almost 0. Thus, the selection model correctly identifies that the missing mechanism does not depend on $\boldsymbol{y}$. Second, the regression coefficient estimates are very close to the true values with the maximum relative bias about 0.3%. Third, the estimate of ϕ is not as accurate as the regression coefficients but close to the true value .25. Finally, the coverage probabilities are close to the nominal value .95. Clearly, if the missing mechanism is MCAR, the selection model can still be applied.

Table 6 Analyze MCAR data using the selection model

Parameters	Estimates	ASE	SD	Coverage
β_0(Intercept)	1.002	0.068	0.066	0.958
$\beta_1(x_1)$	1.002	0.067	0.066	0.944
$\beta_2(x_2)$	1.003	0.067	0.067	0.937
ϕ	0.259	0.051	0.049	0.951
γ_0	-0.258	0.159	0.096	-
$\gamma_1(y)$	-0.002	0.093	0.093	0.955

Note. The same as the previous table.

An empirical example

In this section, we illustrate the application of the simplified selection model through the analysis of a subset of data from the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) study. The ACTIVE study is a randomized and controlled study designed to determine whether cognitive training interventions can affect cognitively based measures of daily functioning (Jobe et al., 2001; Tennstedt, 2001). For the purpose of illustration, the analysis here focuses on whether booster training on memory can improve everyday problem solving ability (EPT) of the elderly.

The sample size for this data analysis is $N = 703$ with about 53% (372) participants selected randomly to receive the booster training on memory. Before and after the booster training, everyday problem sloving ability test was administered to the participants. The time interval between the two tests was about one year. In this data set, the change scores (ΔEPT) as the difference between test scores before and after training were available for 76.5% (583) participants (about 23.5% participants had missing data). It is hypothesized that the training group has a larger ΔEPT than the control group. To control possible confounding factors, we also included demographic variables, age and education level, in our data analysis. There are no missing data in the demographic variables.

Table 7 presents the summary statistics of each group in the data analysis. From Table 7, we can see that the control group had more missing data on EPT than the training group. There were no significant differences in age and education between the two groups. Both groups on average had negative change on EPT and the control group seemed to have more everyday functioning decline. Based on the two-sample t-test on ΔEPT with list-wise deletion, the difference on ΔEPT between two groups was not significant.

Table 7 Summary statistics for the ACTIVE sub-sample

Variable	Training group	Control group	Difference
Sample Size	372	331	
Missing rate	17.2%	30.5%	-13.3%
Age (years)	73.25 (5.78)	73.84 (6.28)	-0.59
Education (years)	13.71 (2.60)	13.45 (2.87)	0.26
ΔEPT	-0.11 (3.27)	-0.28 (3.13)	0.17

Note: Values in the parentheses are standard deviations.

Two models were fitted to the data. The first one is a regular regression model assuming that missingness on ΔEPT is MAR. Thus, the model can be written as

$$\Delta\text{EPT}_i = \beta_0 + \beta_1\text{Training}_i + \beta_2\text{Age}_i + \beta_3\text{Education}_i + e_i.$$

The second one is a selection model assuming that missingness of ΔEPT is related to change in EPT before and after training. Therefore, the model can be specified as

$$\begin{aligned} \Delta\text{EPT}_i &= \beta_0 + \beta_1\text{Training}_i + \beta_2\text{Age}_i + \beta_3\text{Education}_i + e_i \\ m_i &\sim Bernoulli(p_i) \\ p_i &= \Phi(\gamma_0 + \gamma_1\Delta\text{EPT}_i). \end{aligned}$$

In the two models, Training is a binary variable with 1 denoting that a participant is from the booster training group and 0 denoting the control group. The missingness indicator variable m was created with 1 indicating that the score of ΔEPT is missing. Both models were estimated through the Bayesian method as discussed earlier. For priors, each regression coefficient was given a normal distribution with mean 0 and variance 10^6 and the variance of the residuals was given an inverse gamma distribution with both scale and shape parameters equal to 10^{-3}. The results from the two models are summarized in Table 8.

First, the Geweke statistics show that the Markov chain for each model parameter may have converged to its marginal distribution because all Geweke statistics were in the range of -1 to 1. Second, for all model parameters, the ratios between the Monte Carlo error and the standard deviation are smaller than 5%. This indicates that the parameter estimates were accurate. Thus, we can make our inference based on the results in Table 8.

When missing data are assumed to be MAR, results from the regular regression model show that booster training did not improve everyday problem solving ability of the elderly after controlling effects of age and education level. Age did not predict change in everyday problem solving ability, which is not consistent with aging literature (e.g., Finkel et al., 2003; Hedden & Gabrieli, 2004). However, analyzing the

Table 8 Bayesian parameter estimates from the regular regression model and the selection model

		Estimate	s.d.	MC/s.d.	HPD		Geweke
Regular Model	Intercept	1.816	1.925	0.036	-2.011	5.506	-0.124
	Training	0.117	0.281	0.006	-0.429	0.671	-0.740
	Age	-0.041	0.024	0.035	-0.087	0.006	-0.018
	Education	0.070	0.051	0.018	-0.031	0.171	0.805
	ϕ	10.280	0.631	0.003	9.057	11.520	0.649
Selection Model	Intercept	1.384	2.083	0.040	-2.843	5.273	0.483
	Training	0.739	0.316	0.011	0.110	1.351	-0.819
	Age	-0.057	0.026	0.039	-0.105	-0.005	-0.426
	Education	0.090	0.054	0.020	-0.018	0.194	-0.277
	ϕ	13.010	1.288	0.021	10.460	15.500	-0.961
	γ_0	-1.312	0.230	0.028	-1.766	-0.866	0.994
	$\gamma_1(\Delta\text{EPT})$	-0.256	0.058	0.028	-0.366	-0.142	0.865

Note. s.d.: standard deviation, can also be viewed as standard error from a frequentist's perspective. MC/s.d.: the ratio between Monte Carlo error and standard deviation of a parameter. HPD: highest posterior density credible interval.

data as MNAR using the simplified selection model reveals a different picture. Training did have a positive effect in the change of everyday problem solving ability. Overall, participants in the training group had .739 (s.d.: .316; HPD: .110-1.351) more positive change than those in the control group after controlling effects of age and education level. In addition, age is negatively related to change in everyday problem solving ability (older adults had more negative change than younger adults), which is consistent with previous findings (e.g., see the review by Hedden & Gabrieli, 2004).

With different assumptions on missing mechanisms and consequently different models, our data analysis led to different conclusions. Empirical results indicate that it is more likely that missing data were MNAR in this example. From the selection model, ΔEPT was negatively related to the missingness of itself. This means that participants with a lower ΔEPT (less positive change) were more likely to have missing data on EPT. In other words, if a participant expected that he/she would not gain much on EPT through training, he/she was more likely to miss the post-test. Note that we have found that for the training group, participants on average had 0.739 more positive change in their EPT scores than the control group after controlling the effects of age and education from the selection model. Thus, participants in the training group would have a lower probability to have missing data than participants in the control group. Actually, from the empirical missing data rates of the ACTIVE study, about 30.5% percent of participants have missing data in the control group and about 17.2% of participants have missing data in the booster training group. In addition, participants who have complete data are more likely to have a higher ΔEPT. Therefore, the control group had a higher missing data rate, which means that relatively more lower ΔEPTs were missed and relatively more higher ΔEPTs were included in the analysis than the training group in terms of proportions.

Conclusion and discussion

To ease the choice of appropriate variables in explaining missing mechanisms in selection models, we examined and extended the selection model used by Best et al. (1996) in which missingness depends solely on the missing variable itself. We first derived the full conditional posterior distributions for the simplified selection model using the data augmentation algorithm. Then, we conducted seven simulation studies to evaluate the performance of the simplified selection model under a variety of conditions. Finally, we demonstrated the application of the simplified selection model through a real example using data from the ACTIVE study.

Our simulation studies clearly portrayed important features of the simplified selection model. Simulation study 1 showed that when MNAR data were analyzed as MAR data, parameter estimates were incorrect. When the simplified selection model was applied, parameter estimates were accurate. Simulation studies 2 and 3 demonstrated that the simplified selection model was insensitive to the choice of priors and link functions. For example, using either semi-conjugate priors or uninformative priors, model parameters were recovered equally well. Furthermore, although missing data were generated through a step function, both the normal distribution and logistic functions can be used to obtain correct model parameter estimates.

MNAR data may be resulted from different situations other than the simple situation that missingness depends on variables themselves of interest. For example, missingness could be related to an auxiliary and unobserved variable. Simulation study 4 investigated such a scenario and found that the simplified selection model was still able to recover model parameter very well. It is also possible that missingness not only depends on the variables with missing data themselves but also other variables. Simulation study 5 looked into the situation where missingness in $\boldsymbol{y}$ depends on both $\boldsymbol{y}$ and $\boldsymbol{X}$. The results showed that the simplified selection model again performed well in this situation.

Overall, it seems that the simplified selection model worked if missing data involved were MNAR no matter whether the exact missing mechanism was modeled or not. In other words, the performance of the simplified selection model was related to whether missing data were MNAR but not how MNAR data were generated. Therefore, the simplified selection model is robust in terms of analyzing MNAR data.

Our ACTIVE data analysis provided another example on how to judge missing mechanisms based on prior information. From previous research, we knew that with the increase of age, there was a accelerated decline in cognitive ability (e.g., Finkel et al., 2003; Hedden & Gabrieli, 2004). However, the regular data analysis assuming MAR missing data showed that age was not related to the change of EPT. This signaled that the missing mechanism here may not be MAR. On the other hand, when the simplified selection model was applied, the negative relation between age and the change in EPT showed up. Furthermore, other results from the selection model seemed aslo reasonable. It demonstrated that if a participant did not expect much help from booster training, she/he was more likely to miss a test. It should be clear that this is not a formal test for distinguishing missing mechanisms. Thus, we suggest whenever a selection model is used to analyze missing data, data analysis based on a corresponding regular model should be conducted and reported for comparison.

This study has its limitations. Admittedly, the regression model discussed in the current study is a relatively simple model. However, with a simple model, we can disentangle complexity of missing data analysis in a transparent way. For example, conditional posterior distributions of missing data are readily available and they clearly show the difference and connection between a regular model and a selection model. The method and strategy used in this study can be readily generalized to more complex models, e.g., growth curve models and growth mixture models.[2] By presenting the details of a simple model, it is our hope that future research on non-ignorable missing data can be conducted for other sophisticated models.

Although only the monotonously increasing link functions including the logistic function and the normal distribution function are considered in this study, we want to emphasize that the link function can take many different forms according to substantive soundness of data analysis. For example, if there is a higher missing probability for those participants who have either a relatively higher value or a relatively lower value, a U-shape link function can be adopted. Furthermore, even for the same link function, we can specify different relationship between missing probability and its predicting variables. In this study, we focused on a linear relationship. However, it is convenient to model a nonlinear relationship, for example, by adding a quadratic term into the model if previous substantive findings support this.

[2]Manuscripts on those models are under review and preparation and can be obtained from the first author.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Best, N. G., Spiegelhalter, D. J., Thomas, A., & Brayne, C. E. (1996). Bayesian analysis of realistically complex models. *Journal of Royal Statistical Society, A*, *159*, 323–342.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Wiley.

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistican*, *46*(3), 167-174.

Congdon, P. (2003). *Applied bayesian modelling*. Wiley.

Finkel, D., Reynolds, C. A., McArdle, J. J., Gatz, M., & Pedersen, N. L. (2003). Latent growth curve analyses of accelerating decline in cognitive abilities in late adulthood. *Developmental Psychology*, *39*(3), 535–550.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd, Ed.). Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Oxford, UK: Clarendon Press.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

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. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Hedden, T., & Gabrieli, J. D. E. (2004). Insights into the ageing mind: a view from cognitive neuroscience. *Nature Reviews Neuroscience*, *5*, 87–96.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2006). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*, 332–346.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test*, *18*, 1–43.

Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., et al. (2001). Active: A cognitive intervention trial to promote independence in older adults. *Controlled Clinical Trials*, *22*(4), 453–479.

Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of American Statistical Association*, *77*, 237–250.

Little, R. J. A. (1985). A note about model for selectivity bias. *Econometrica*, *53*, 1469–1474.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, N.Y.: Wiley-Interscience.

Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica*, *48*, 1815–1820.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Tang, G., Little, R. J. A., & Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, *90*(4), 747-764.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the american statistical association*, *82*, 528–540.

Tennstedt, S. (2001). *ACTIVE (advanced cognitive training for independent and vital elderly), 1999-2001 [UNITED STATES] [Computer file]. ICPSR04248-v1. Watertown, MA: New England Research Institute [producer], 2001. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2005-10-11.*

Zhang, Z., McArdle, J. J., Wang, L., & Hamagami, F. (2008). A sas interface for bayesian analysis with winbugs. *Structural Equation Modeling*, *15*(4), 705–728.

Appendix
WinBUGS codes for the selection model

```
## Model
model{
  for (i in 1:N){
    mu[i]<-b[1]+b[2]*x1[i]+b[3]*x2[i]
    y[i]~dnorm(mu[i], pre.phi)

    z[i]~dnorm(muz[i], 1)I(L[i],U[i])
    muz[i]<-b[4]+b[5]*y[i]

    L[i]<- -(1-m[i])*10000
    U[i]<- m[i]*10000
  }
  for (i in 1:5){
    b[i]~dnorm(0, 1.0E-6)
    Para[i]<-b[i]
  }
  pre.phi~dgamma(.001,.001)
  Para[6]<-1/pre.phi
}
## Starting values
list(b=c(0,0,0,0,0), pre.phi=1)
## Data are omitted for the sake of saving space
```