

BAYESIAN INFERENCE FOR NONRESPONSE TWO-PHASE SAMPLING

Yue Zhang, Henian Chen and Nanhua Zhang

*Shanghai Jiao Tong University, University of South Florida
and Cincinnati Children's Hospital Medical Center*

Abstract: Nonresponse is an important practical problem in epidemiological surveys and clinical trials. Common methods for dealing with missing data rely on untestable assumptions. In particular, non-ignorable modeling, which derives inference from the likelihood function based on a joint distribution of the variables and the missingness indicators, can be sensitive to misspecification of this distribution and may also have problems with identifying the parameters. Nonresponse two-phase sampling (NTS), which re-contacts and collects data from a subsample of the initial nonrespondents, has been used to reduce nonresponse bias. The additional data collected in phase II provide important information for identifying the parameters in the non-ignorable models. We propose a Bayesian selection model which utilizes the additional data from phase II and develop an efficient Markov chain Monte Carlo algorithm for the posterior computation. We illustrate the proposed model on simulation studies and a Quality of Life (QOL) dataset.

Key words and phrases: Bayesian selection model, Markov chain Monte Carlo, missing not at random, quality of life, two-phase sampling.

1. Introduction

Nonresponse at an appreciate rate exists in such applications as epidemiological surveys and clinical trials. Commonly used approaches to handling missing data include complete-case (CC) analysis, ignorable likelihood (IL) methods and nonignorable models (NIM). CC, which discards the incomplete cases, can result in substantial loss of information or biased estimation of the key parameters. IL methods are based on the observed likelihood which does not include a model for the missing data indicator, and these models provide valid inference if the missingness does not depend on the missing values. Such missing data mechanism is called missing at random (MAR) (Rubin (1976); Little and Rubin (2002)). Examples of IL methods include ignorable maximum likelihood (IML) (e.g., Dempster, Laird and Rubin (1977)), Bayesian inferences (e.g., Sugden and Smith (1984)), and multiple imputation (e.g., Rubin (2004)). When the data

Table 1. Four patterns in two-phase sampling.

Pattern	Observation, i	y_i	$R_{1,i}$	$S_{2 1,i}$	$R_{2 1,i}$	$R_{2,i}$
1	$i = 1, \dots, m$	\checkmark	1	-	-	1
2	$i = m + 1, \dots, m + r$	\times	0	1	1	1
3	$i = m + r + 1, \dots, m + s$?	0	1	0	0
4	$i = m + s + 1, \dots, n$?	0	0	0	0

Key: \checkmark denotes observed; ? denotes at least one entry missing; \times denotes at least one entry missing in phase I, but observed in phase II.

are missing not at random (MNAR), the missingness can depend on the missing values, non-ignorable models (NIM) are developed based on the joint distribution of the variables and the missing data indicators. The nonignorable models are less common in practice, because of the difficulty in specifying the models for the missing data mechanism, sensitivity to model misspecification, and problems with identifying the parameters (e.g., Little and Rubin (2002); Heckman (1979); Little (1993, 1994); Nandram and Choi (2002, 2010)).

All three methods (CC, IL, and NIM) rely on untestable assumptions about the missing data mechanism. Sensitivity analyses have been proposed to systematically examine the effect of perturbations to model assumptions (e.g., Little (1993, 1994); Troxel, Ma and Heitjan (2004); Zhu, Ibrahim and Tang (2014)). Another alternative is to use a study design to relax to some degree the assumptions required under IL and NIM. One such design is two-phase sampling, in which a subsample of non-respondents to the original survey (phase I) is randomly selected for further interview attempts (phase II). This method is called nonresponse two-phase sampling (NTS). The general missing data structure for NTS is listed in Table 1.

Let $\{y_i, i = 1, 2, \dots, n\}$ denote n independent observations on an outcome variable \mathbf{Y} , where \mathbf{Y} has missing values; if missing, y_i denotes the underlying missing value of the outcome for the i -th subject. The response indicator for phase I is denoted as $R_{1,i}$, which equals 1 if y_i is observed and 0 otherwise. $S_{2|1,i}$ is used to denote whether a subject is sampled among the nonrespondents in phase I. Let $R_{2|1,i}$ denote the phase II response indicator among the nonrespondents in phase I, and $R_{2,i}$ be the overall response indicator after completion of phase II. There are four patterns in Table 1. Pattern 1 consists of subjects for whom y_i is fully observed after first phase data collection. Pattern 2 consists of cases that were missing in phase I, but subsequently observed in phase II sampling. Pattern 3 consists of cases that were sampled in phase II, but did not respond, and Pattern 4 were those phase I nonrespondents not sampled in phase II.

NTS was first proposed by Hansen and Hurwitz (1946) to reduce the non-response bias in mail questionnaires by doing personal interviews on a fraction of the nonrespondents. This sampling scheme was referred to as "call-back" and Cochran (1977) studied the effects of call-backs and the optimal sampling fractions among the nonrespondents. Some other examples include the National Comorbidity Survey (Elliott and Little (2000)) and the 2003 Survey of Small Business Finances (Harter et al. (2007)). Different from those approaches relying on using case weights (e.g., Hansen and Hurwitz (1946); Srinath (1971); Harter et al. (2007)), a method called nonrespondent subsample multiple imputation (NSMI) was proposed by Zhang, Chen and Elliott (2016) to reduce bias using data from the NTS. NSMI performs multiple imputation within the subsample of nonrespondents in phase I by using additional data collected in phase II. It works well if MAR assumption holds in phase II within the sample of nonrespondents in phase I regardless of the missingness mechanism in phase I. However, this assumption is usually untestable and the phase II response mechanism from phase I nonrespondents may be related to outcome values, in which case the NSMI methods yield biased estimates. We propose a nonrespondents subsample Bayesian selection model (NSBSM), which makes use of the additional data from phase II when jointly modeling the outcome and the phase I missing data indicator. The rationale of using the additional data in modeling the phase I indicator is that these data provide important identifying information for modeling the phase I missingness indicator. For model comparison purpose, we also apply the Bayesian selection model without considering the phase II data (BSM).

The rest of the paper is organized as follows. Section 2 reviews NSMI method that has been proposed for NTS (Zhang, Chen and Elliott (2016)). The method works well when phase I missingness is MNAR but the missingness in the phase II among the nonrespondents is MAR. Section 3 introduces the NSBSM method, including the model setup and posterior inference. We illustrate the properties of the NSBSM and compare the performance of different methods in Section 4 using simulation studies, while Section 5 applies the method to a quality of life (QOL) dataset. Section 6 concludes the paper with discussion.

2. Nonrespondent Subsample Multiple Imputation (NSMI)

Data with the structure in Table 1 are considered. NSMI applies the multiple imputation method to the cases in Patterns 2, 3, and 4 of Table 1. Subjects in Pattern 1 are excluded when the missing values in Patterns 3 and 4 are imputed,

and then the imputed datasets from Patterns 2, 3, 4 are combined with data from Pattern 1 for statistical analyses (Zhang, Chen and Elliott (2016)). The key assumption of NSMI is that within the nonrespondent in phase I (Patterns 2, 3, and 4), the missingness after phase II is MAR. Let $\mathbf{Y}_{obs} = (\mathbf{Y}_{obs,1}, \mathbf{Y}_{obs,2})$ represent observed data in phase I and phase II, and \mathbf{Y}_{mis} to represent the data missing after phase II sampling. The assumption for NSMI to be valid can be expressed as follows based on fully observed covariate vector \mathbf{Z} .

$$\Pr(\mathbf{R}_{2|1} = 1 | \mathbf{R}_1 = 0, \mathbf{Y}_{obs,2}, \mathbf{Y}_{mis}, \gamma; \mathbf{Z}) = \Pr(\mathbf{R}_{2|1} = 1 | \mathbf{R}_1 = 0, \mathbf{Y}_{obs,2}, \gamma; \mathbf{Z}),$$

where γ is the parameter associated with the distribution of the response indicator \mathbf{R} . The missingness mechanism is called nonrespondent subsample missing at random (NS-MAR). This assumption does not confine the missing data mechanisms in the whole sample (\mathbf{R}_2) or the missing data mechanism in phase I (\mathbf{R}_1) to a certain missing data mechanism, and therefore NSMI can be applied even under the MNAR missingness mechanism in phase I as long as phase II is MCAR or MAR.

When the phase II missing data mechanism is MNAR, the NSMI method fails to yield unbiased estimates because the NS-MAR assumption is violated, it usually leads to estimates with large variance, due to the increased variability in the imputed values by using subjects from Patterns 2, 3, and 4, but not subjects from Pattern 1. These motivate the proposed method in the next section.

3. Nonrespondents Subsample Bayesian Selection Model (NSBSM)

3.1. Selection model

In this section, data with structure in Table 1 are also considered. This paper proposes nonrespondents subsample Bayesian selection model (NSBSM), a Bayesian approach based on selection model to address the identifiability issue (Little and Rubin (2002)) by utilizing additional data in phase II. A selection model contains a regression equation for the outcome, and a regression equation for the sample selection mechanism. Suppose the regression equation for the outcome of primary interest is

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \eta_i, \quad (3.1)$$

and the sample selection mechanism is driven by the latent linear regression equation

$$u_i = \mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon_i, \quad (3.2)$$

where $i = 1, 2, \dots, N$, \mathbf{x}_i and \mathbf{z}_i are the covariates and they may overlap with

each other. Assume that the outcome y_i is observed if and only if $u_i > 0$. The missing indicator for sample selection of phase I is

$$R_{1,i} = I(u_i > 0),$$

where $I(\cdot)$ is indicator function.

Heckman (1979) assumed a bivariate normal distribution for η_i and ϵ_i in equations (3.1) and (3.2).

$$\begin{pmatrix} \eta_i \\ \epsilon_i \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}_2, \mathbf{\Sigma}),$$

where $\mathbf{0}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$. The second diagonal element of $\mathbf{\Sigma}$ is set to 1 for full identification, a typical constraint for a binary choice model. To facilitate posterior inference, we factor the bivariate normal distribution (ϵ_i, η_i) into the product of marginal distribution of ϵ_i and conditional distribution of $\eta_i|\epsilon_i$, and obtain

$$\begin{aligned} y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + E(\eta_i|\epsilon_i) + \xi_i, \\ u_i &= \mathbf{z}_i^T \boldsymbol{\gamma} + \epsilon_i, \end{aligned}$$

where $\epsilon_i \sim \mathcal{N}(0, 1)$, $\xi_i \sim \mathcal{N}(0, \sigma^2)$, $\eta_i|\epsilon_i = \sigma_{12} \cdot \epsilon_i$, and $\sigma^2 = \sigma_{11} - \sigma_{12}^2$. By utilizing this re-parameterization, our parameters of interest for the covariance structure are σ^2 and σ_{12} , for which we assign independent priors. This re-parameterization was first proposed by Koop and Poirier (1997) to address the complication in estimating the two free parameters $(\sigma_{11}, \sigma_{12})$ in the covariance matrix $\mathbf{\Sigma}$. An efficient Gibbs sampling algorithm was developed by Li (1998) following the re-parameterization. Under the aforementioned bivariate normal assumption, the model at (3.1) and (3.2) implies that

$$\begin{aligned} \Pr(R_{1,i} = 1|\mathbf{z}_i) &= \Phi(\mathbf{z}_i^T \boldsymbol{\gamma}), \\ E(y_i|R_{1,i} = 1, \mathbf{x}_i, \mathbf{z}_i) &= \mathbf{x}_i^T \boldsymbol{\beta} + \sigma_{12} \lambda(\mathbf{z}_i^T \boldsymbol{\gamma}), \end{aligned} \quad (3.3)$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio (Little and Rubin (2002)).

3.2. Model identification

Choosing the appropriate covariates $(\mathbf{x}_i^T, \mathbf{z}_i^T)$ plays an important role in the selection model. Little and Rubin (2002) point out that the presence of the inverse Mills ratio in (3.3) often results in multicollinearity that can lead to profound identification problem for estimating $\boldsymbol{\beta}$ and σ_{12} in (3.3). One possible solution to this problem includes at least one of the elements of \mathbf{z}_i that are not

in \mathbf{x}_i that is associated with the selection process but not the outcome. With a valid exclusion restriction assumption, the inverse Mills ratio and the \mathbf{x}_i vector in (3.3) will be less correlated, reducing multicollinearity among predictors and therefore facilitating model identification.

In practice, it can be difficult to identify variables that satisfy the exclusion restriction assumption. With the additional data from phase II, the model can be identified without assuming exclusion restriction. The inverse Mills ratio is estimated by the non-linear probit model, the correction term λ is not perfectly correlated with \mathbf{x}_i , even in the absence of exclusion restriction. The proposed two-phase sampling method provides additional information of nonrespondents of phase I. The model for the initial nonrespondents takes the form (Little and Rubin (2002))

$$E(y_i | R_{1,i} = 0, \mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \sigma_{12} \{-\lambda(-\mathbf{z}_i^T \boldsymbol{\gamma})\}, \quad (3.4)$$

where $\mathbf{z}_i = \mathbf{x}_i$. With the additional data from phase II, equations (3.3) and (3.4) imply that the new correction term λ is a vector with both $\lambda(\mathbf{z}_i^T \boldsymbol{\gamma})$ and $-\lambda(-\mathbf{z}_i^T \boldsymbol{\gamma})$ making it less linearly correlated with \mathbf{x}_i , and therefore we can estimate $\boldsymbol{\beta}$ and σ_{12} even without exclusion restrictions. We show the benefit of the proposed method in identifying parameters in the simulation study. In the following sections, we use notation \mathbf{x}_i as the common covariate in both (3.1) and (3.2). For simplicity, we suppress the conditioning on \mathbf{x}_i from all equations in the following sections.

3.3. Likelihood function

Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$ denote all the coefficients. It follows from the bivariate normality distribution for η_i and ϵ_i that, when an outcome is observed in phase I,

$$\Pr\{R_{1,i} = 1 | y_i, \boldsymbol{\theta}, \sigma^2, \sigma_{12}\} = 1 - \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right).$$

When the outcome is missing in phase I, the probability that this occurs is

$$\Pr\{R_{1,i} = 0 | y_i, \boldsymbol{\theta}, \sigma^2, \sigma_{12}\} = \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right).$$

Without incorporating the additional phase II data, the likelihood function of the traditional Bayesian selection model (BSM) takes the form

$$L(\boldsymbol{\theta}, \sigma^2, \sigma_{12} | \mathbf{y}, R_{1,i}, R_{2|1,i}) \\ \propto \prod_{i=1}^m \phi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \left\{ 1 - \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right) \right\}$$

$$\times \prod_{i=m+1}^n \int \phi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right) dy_i.$$

With additional data from phase II, the likelihood is given by

$$\begin{aligned} & L(\boldsymbol{\theta}, \sigma^2, \sigma_{12} | \mathbf{y}, R_{1,i}, R_{2|1,i}) \\ & \propto \prod_{i=1}^m \phi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \left\{ 1 - \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right) \right\} \\ & \times \prod_{i=m+1}^{m+r} \phi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right) \\ & \times \prod_{i=m+r+1}^n \int \phi \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sqrt{\sigma^2 + \sigma_{12}^2}} \right) \times \Phi \left(\mathbf{x}_i^T \boldsymbol{\gamma} \sqrt{1 + \frac{\sigma_{12}^2}{\sigma^2}} + \frac{\sigma_{12}(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma \sqrt{\sigma^2 + \sigma_{12}^2}} \right) dy_i. \end{aligned}$$

In a traditional selection model when no data in phase II are available, identification of the parameter could be a problem. The rationale of the NSBSM methods is that the additional information provides valuable information for identifying the model parameters. In this study, we propose a probit model for $R_{1,i}$ but not for $R_{2|1,i}$. Generally, the NSBSM methods are based on a partial likelihood (Cox, 1972) with the component regarding the selection process $R_{2|1}$ discarded from the analysis. The identification problems remains if the selection model also includes a probit model for $R_{2|1,i}$; we leave this for future research.

3.4. Prior

Assume the prior distribution

$$f(\boldsymbol{\theta}, \sigma_{12}, \sigma^2) = f(\boldsymbol{\theta})f(\sigma_{12})f(\sigma^2),$$

where

$$\begin{aligned} f(\boldsymbol{\theta}) & \sim \mathcal{MVN}(\boldsymbol{\theta}_0, \boldsymbol{\Psi}_0^{-1}), \\ f(\sigma_{12}) & \sim \mathcal{N}(c_0, d_0^{-1}), \\ f(\sigma^{-2}) & \sim \mathcal{G} \left(\frac{v_0}{2}, \left(\frac{w_0}{2} \right)^{-1} \right). \end{aligned}$$

Here, \mathcal{MVN} , \mathcal{N} , and \mathcal{G} denote the multivariate normal, the univariate normal, and the gamma distribution, respectively. $\mathcal{G}(v_0/2, (w_0/2)^{-1})$ denotes a gamma distribution with shape parameter $v_0/2$ and scale parameter $(w_0/2)^{-1}$. Prior parameters $\{\boldsymbol{\theta}_0, \boldsymbol{\Psi}_0, c_0, d_0, v_0, w_0\}$ are specified in simulation studies and data applications. Proper priors are used to ensure that the resulting posterior densities have closed form.

3.5. Posterior

Let y_i^* denote augmented outcomes. The conditional means and variances of the bivariate normal variables (u_i, y_i^*) have the forms

$$\begin{aligned}\mu_{y^*|u} &= \mathbf{x}_i^T \boldsymbol{\beta} + \sigma_{12}(u_i - \mathbf{x}_i^T \boldsymbol{\gamma}), \quad \sigma_{y^*|u} = \sigma^2, \\ \mu_{u|y^*} &= \mathbf{x}_i^T \boldsymbol{\gamma} + \frac{\sigma_{12}}{\sigma^2 + \sigma_{12}^2}(y_i^* - \mathbf{x}_i^T \boldsymbol{\beta}), \quad \sigma_{u|y^*}^2 = 1 - \frac{\sigma_{12}^2}{\sigma^2 + \sigma_{12}^2}.\end{aligned}$$

The Gibbs sampling algorithm including data augmentation (imputation) is summarized in the following steps. Additional details and expressions for the parameters of the various posteriors are given in the Appendix.

1. If $R_{1,i} = 1$, $y_i^* = y_i$ and $u_i|(y_i^*, \boldsymbol{\theta}, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{\mu|y^*}^2; 0, \infty)$, where \mathcal{TN} denotes truncated normal distribution.
2. If $R_{1,i} = 0$ and $R_{2|1,i} = 1$, $y_i^* = y_i$ and $u_i|(y_i^*, \boldsymbol{\theta}, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{\mu|y^*}^2; -\infty, 0)$.
3. If $R_{1,i} = 0$ and $R_{2|1,i} = 0$, $y_i^*|(\boldsymbol{\theta}, \sigma_{12}, \sigma^2) \sim \mathcal{N}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2 + \sigma_{12}^2)$ and $u_i|(y_i^*, \boldsymbol{\theta}, \sigma_{12}, \sigma^2) \sim \mathcal{TN}(\mu_{u|y^*}, \sigma_{\mu|y^*}^2; -\infty, 0)$.
4. Sample $\boldsymbol{\theta}$ from $\mathcal{MVN}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\Psi}}^{-1})$, where

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \tilde{\boldsymbol{\Psi}}^{-1} \{ \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \tilde{\mathbf{y}} + \boldsymbol{\Psi}_0 \boldsymbol{\theta}_0 \}, \\ \tilde{\boldsymbol{\Psi}} &= \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n) \mathbf{Z} + \boldsymbol{\Psi}_0, \\ \mathbf{Z} &= \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{pmatrix},\end{aligned}$$

$$\text{with } \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}^* \\ \mathbf{u} \end{pmatrix}, \quad \mathbf{Z}_1 = \mathbf{Z}_2 = (\mathbf{1}, \mathbf{x}).$$

5. Sample σ_{12} from $\mathcal{N}(\tilde{c}, \tilde{d}^{-1})$, where $\tilde{c} = \tilde{d}^{-1}(\sigma^{-2} \boldsymbol{\eta}' \boldsymbol{\epsilon} + c_0 d_0)$ and $\tilde{d} = \sigma^{-2} \boldsymbol{\epsilon}' \boldsymbol{\epsilon} + d_0$.
6. Sample σ^{-2} from $\mathcal{G}(\tilde{v}/2, (\tilde{w}/2)^{-1})$, where $\tilde{v} = n + v_0$ and $\tilde{w} = (\boldsymbol{\eta} - \boldsymbol{\epsilon} \sigma_{12})^T (\boldsymbol{\eta} - \boldsymbol{\epsilon} \sigma_{12}) + w_0$.
7. Return to Step 1 and repeat.

4. Simulation Studies

This section illustrates the properties of the NSBSM method using simulation

studies and compares the performance of NSBSM to other methods under different missing data mechanisms in phase II. For each simulation study, six methods were applied to estimate regression coefficient of simple linear regression model based on outcome \mathbf{Y} and covariate \mathbf{X} :

- (1) BD: estimates use the full data generated from simulation before missing values are created, as a benchmark method.
- (2) CC: complete case analysis uses respondents from both phase I and phase II, discarding cases where are still missing after phase II.
- (3) IL: ignorable likelihood method through multiple imputation uses data from both phase I and phase II, assuming ignorable missingness.
- (4) NSMI: multiple imputation in the nonrespondent subsample in phase I uses only additional data from phase II.
- (5) BSM: Bayesian selection model uses data from only phase I.
- (6) NSBSM: Bayesian selection model uses data from both phase I and phase II.

All six methods except method (5) utilized all the observed data in phase I and II. Method (5) uses only information in phase I to estimate parameters in a regression model. We compared the performance of each of the methods using empirical bias, standard error (SE), root mean square error (RMSE), and the coverage probability of the 95% highest posterior density (HPD) interval.

The outcome was generated by the linear regression model

$$y_i = 1 + x_i + z_i + x_i \times z_i + \eta_i, \quad \eta_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

with x_i sampled from the standard normal, z_i from the Bernoulli with probability 0.5, and $x_i \times z_i$ is the interaction term, for $i = 1, 2, \dots, 1,000$. The response was subject to missingness, while x_i and z_i were fully observed. Phase I missing values in \mathbf{Y} were generated based on the MNAR mechanism

$$\Pr(R_{1,i} = 0 | y_i, x_i, z_i) = \Phi(-3 \times y_i + 2.5).$$

This missing data generation scheme results in approximately 35% of the values \mathbf{Y} being missing in phase I.

Let $R_{2|1,i}$ denote the response indicator in the subsample of nonrespondents in phase I. Phase II responses in y_i were generated under three missing data mechanisms:

- (1) MAR: $\Pr(R_{2|1,i} = 0 | y_i, x_i, z_i) = \Phi(-x_i - z_i - x_i \times z_i + c_1)$,

$$(2) \text{ MNAR: } \Pr(R_{2|1,i} = 0 | y_i, x_i, z_i) = \Phi(-0.3 \times y_i + x_i + z_i + x_i \times z_i + c_2),$$

$$(3) \text{ MNAR: } \Pr(R_{2|1,i} = 0 | y_i, x_i, z_i) = \text{expit}(-0.3 \times y_i + x_i + z_i + x_i \times z_i + c_3),$$

where c_1 , c_2 and c_3 were assigned different values for various selection proportions in phase II. Appropriate c_1 , c_2 and c_3 were set so that the corresponding selection proportions of phase II responses in \mathbf{Y} were approximately $\{40\%, 30\%, 20\%, 10\%\}$, and $\text{expit}(\cdot) = \exp(\cdot) / \{1 + \exp(\cdot)\}$. The third scenario is included to evaluate how the NSBSM method performs when data are not simulated from the same model as the NSBSM method. Let k denote the dimension of coefficients. Prior parameters were set as $\{\boldsymbol{\theta} = \mathbf{0}, \boldsymbol{\Psi}_0^{-1} = 100 \cdot \mathbf{I}_k, c_0 = 0, d_0 = 1, v_0 = 10, w_0 = 10\}$ to obtain a balanced variance-covariance matrix $\boldsymbol{\Sigma}$ and comparable variability for both σ^2 and σ_{12}^2 (Li (1998)). Results are based on 200 repetitions for each simulated condition. For each data set the Markov chain Monte Carlo (MCMC) algorithm was run for 12,000 iterations, where the first 2,000 draws were discarded as burn-in period. The Gelman-Rubin statistics and trace plot suggested that the Markov chain was mixing well. The biases, RMSEs, and coverage probabilities of the 95% HPD intervals from MAR and MNAR of phase II response are reported. CC and IL yield biased estimates of the regression coefficient since phase I missingness is MNAR; therefore we focus on the results from NSMI, BSM and NSBSM.

For the data simulated based on MAR mechanism in phase II (Table 2), both NSMI and NSBSM yield approximately unbiased estimates of the regression coefficients. The NSMI method has moderately larger SEs and RMSEs because of increased variability in the imputed values, which uses subjects from Patterns 2, 3 and 4, but not subjects from Pattern 1 (Zhang, Chen and Elliott (2016)). The BSM method shows significant bias in estimating the regression coefficients. This is not surprising because of the identification problem with the traditional selection model. By varying the proportion of sampled in phase II, we see the precision increases as the sampling proportion in phase II increases. Even sampling 10% of the nonrespondents in phase I is enough to distinguish the NSBSM and NSMI method from other competing methods.

When the phase II missing data mechanism is MNAR (Table 3), the NSBSM is the only method that provides unbiased estimates of the regression. All other methods show significant biases because the MAR assumptions are violated. We again see that precision increases as sampling proportion in phase II increases, and the improvement is substantial even if we only collect data from 10% of the nonrespondents in phase I.

Table 2. Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MAR in phase II (200 replicates, $\Pr(R_{21,i} = 0|y_i, x_i, z_i) = \Phi(-x_i - z_i - x_i \times z_i + c_1)$).

Phase II Sampling Proportion	Method	b_0				b_1				b_2				b_3			
		Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR
40%	BD	14	31	433	0.965	28	32	447	0.94	17	44	623	0.96	-15	49	694	0.935
	CC	2,897	32	2,932	0	496	32	669	0.785	-1,047	48	1,243	0.62	-870	51	1,125	0.69
	IL	2,909	33	2,946	0	501	32	674	0.78	-1,060	48	1,260	0.645	-877	51	1,137	0.685
	NSMI	168	39	586	0.96	170	33	524	0.935	179	51	808	0.955	-112	52	755	0.92
	BSM	1,123	45	1,289	0.6	-498	38	728	0.89	-332	53	824	0.93	-111	60	860	0.925
	NSBSM	235	33	520	0.92	67	32	457	0.955	-47	46	655	0.955	-95	51	719	0.93
30%	BD	-24	32	446	0.955	5	32	455	0.95	-9	45	641	0.96	-46	45	643	0.97
	CC	3,697	35	3,730	0	128	33	487	0.935	-1,555	46	1,687	0.365	-816	48	1,061	0.71
	IL	3,700	35	3,734	0	129	34	495	0.94	-1,561	47	1,697	0.375	-818	48	1,063	0.735
	NSMI	169	47	730	0.965	190	34	567	0.93	226	56	992	0.93	-207	47	722	0.96
	BSM	1,061	40	1,201	0.645	-488	39	740	0.88	-341	48	764	0.96	-187	53	766	0.95
	NSBSM	297	35	571	0.96	31	33	462	0.95	-113	46	664	0.98	-147	45	657	0.96
20%	BD	20	31	440	0.935	-61	32	455	0.955	-11	40	564	0.955	71	43	608	0.96
	CC	4,590	31	4,611	0	-537	37	747	0.78	-2,050	43	2,139	0.105	-438	53	866	0.84
	IL	4,593	32	4,615	0	-535	37	749	0.79	-2,057	44	2,146	0.13	-443	54	873	0.835
	NSMI	331	63	961	0.955	148	39	625	0.94	242	64	1,213	0.94	-156	49	748	0.96
	BSM	1,138	44	1,294	0.595	-576	41	814	0.81	-376	47	761	0.945	-37	51	725	0.975
	NSBSM	516	34	707	0.86	-120	36	526	0.945	-200	42	630	0.955	24	47	662	0.95
10%	BD	-10	32	456	0.95	36	31	441	0.955	-15	47	660	0.945	-60	47	660	0.945
	CC	5,329	38	5,357	0	-1,141	39	1,269	0.385	-2,458	53	2,570	0.065	-249	55	813	0.89
	IL	5,336	39	5,365	0	-1,149	40	1,281	0.415	-2,472	54	2,591	0.08	-233	57	826	0.9
	NSMI	309	40	1,420	0.95	258	44	705	0.945	440	101	1,756	0.96	-352	62	992	0.93
	BSM	1,183	117	1,336	0.565	-518	42	790	0.855	-441	54	885	0.905	-129	60	859	0.925
	NSBSM	678	44	890	0.74	-131	36	522	0.925	-324	53	816	0.895	-77	51	721	0.95

b_0 is the intercept. b_1, b_2, b_3 are the coefficients of $x_i, z_i, x_i \times z_i$, respectively. The true value of these parameters is 1.

Table 3. Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MNAR in phase II (200 replicates, $\Pr(R_{2|1,i} = 0|y_i, x_i, z_i) = \Phi(-0.3 \times y_i + x_i + z_i \times z_i + c_2))$.

Phase II Sampling Proportion	Method	b_0				b_1				b_2				b_3			
		Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR ($\times 10^4$)	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR ($\times 10^4$)	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR ($\times 10^4$)	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR ($\times 10^4$)
40%	BD	18	33	459	0.94	1	31	443	0.95	2	46	642	0.93	-3	44	622	0.96
	CC	2,654	43	2,721	0.005	-1,875	44	1,975	0.14	-1,051	56	1,312	0.735	185	61	881	0.96
	IL	2,661	43	2,729	0.01	-1,882	45	1,989	0.165	-1,062	57	1,332	0.725	195	64	920	0.95
	NSMI	-614	45	880	0.765	512	50	873	0.84	335	62	931	0.915	-348	73	1,089	0.93
	BSM	1,171	46	1,336	0.575	-540	40	777	0.83	-370	52	822	0.93	-108	54	765	0.96
	NSBSM	-187	39	586	0.91	166	39	576	0.955	92	52	745	0.93	-37	54	762	0.97
30%	BD	17	34	475	0.915	23	34	479	0.945	-8	46	655	0.935	-81	47	667	0.935
	CC	3,340	36	3,378	0	468	36	688	0.795	-1,376	49	1,540	0.49	-965	49	1,184	0.63
	IL	3,340	36	3,378	0	471	36	695	0.815	-1,368	50	1,537	0.485	-973	49	1,193	0.64
	NSMI	-578	51	920	0.89	239	38	592	0.905	645	70	1,179	0.93	-99	52	737	0.96
	BSM	1,135	42	1,278	0.605	-498	40	752	0.87	-383	53	846	0.92	-171	55	798	0.945
	NSBSM	15	36	510	0.945	210	35	534	0.91	16	51	718	0.935	-92	48	707	0.945
20%	BD	10	32	452	0.925	-12	30	419	0.97	39	47	659	0.94	39	42	596	0.955
	CC	4,719	47	4,766	0	-794	72	1,292	0.7	-2,015	53	2,148	0.2	-449	51	849	0.89
	IL	4,730	49	4,780	0	-804	74	1,313	0.705	-2,030	54	2,168	0.245	-435	51	845	0.89
	NSMI	-1,077	89	1,653	0.87	515	49	861	0.885	880	16	1,862	0.91	-136	75	1,068	0.95
	BSM	1,161	44	1,319	0.615	-577	37	773	0.855	-347	57	882	0.89	-64	53	754	0.96
	NSBSM	201	42	624	0.925	104	34	490	0.96	-32	54	759	0.94	-56	47	671	0.975
10%	BD	16	31	444	0.955	-53	31	447	0.965	13	46	645	0.945	8	46	650	0.94
	CC	4,949	36	4,974	0	-744	37	905	0.685	-2,286	49	2,390	0.09	-299	52	795	0.885
	IL	1,954	37	4,982	0	-755	38	924	0.68	-2,298	52	2,414	0.11	-281	53	802	0.89
	NSMI	-1,013	116	1,925	0.91	454	44	772	0.905	900	2,244	1,522	0.92	-56	61	868	0.975
	BSM	1,160	43	1,311	0.59	-609	39	818	0.83	-407	48	793	0.945	16	53	781	0.945
	NSBSM	313	38	616	0.935	61	35	491	0.975	-163	49	705	0.95	-12	47	698	0.955

b_0 is the intercept. b_1, b_2, b_3 are the coefficients of $x_i, z_i, x_i \times z_i$, respectively. The true value of these parameters is 1.

Table 4. Empirical bias (Bias), standard error (SE), root mean square error (RMSE), and 95% HPD interval coverage probabilities (CR) under MNAR in phase I and MNAR in phase II (200 replicates, $\logit(\Pr(R_{2|1,i} = 0|y_i, x_i, z_i)) = -0.3 \times y_i + x_i + z_i + x_i \times z_i + c_3$)).

Phase II Sampling Proportion	Method	b_0				b_1				b_2				b_3			
		Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR	Bias ($\times 10^4$)	SE ($\times 10^4$)	RMSE ($\times 10^4$)	CR
40%	BD	-24	34	479	0.93	7	33	467	0.955	15	43	613	0.960	-33	46	646	0.940
	CC	2,133	39	2,203	0.045	-1,035	39	1,175	0.555	-803	54	1,105	0.8	-96	59	843	0.955
	IL	2,125	40	2,199	0.045	-1,028	41	1,179	0.590	-807	55	1,117	0.84	-91	62	881	0.945
	NSMI	-666	41	886	0.79	500	43	783	0.860	193	56	820	0.935	-84	64	913	0.94
	BSM	1,116	40	1,250	0.665	-557	40	797	0.850	-366	49	777	0.96	-93	56	802	0.93
	NSBSM	-332	37	621	0.925	258	38	594	0.920	101	49	702	0.965	-56	55	772	0.95
30%	BD	46	34	480	0.92	-2	32	458	0.95	-62	44	629	0.945	11	46	647	0.955
	CC	3,070	41	3,124	0	-1,525	44	1,647	0.285	-1,230	53	1,441	0.655	-65	61	861	0.955
	IL	3,063	42	3,121	0	-1,514	45	1,644	0.325	-1,223	55	1,451	0.67	-71	62	879	0.965
	NSMI	-671	45	927	0.79	498	51	873	0.845	193	63	903	0.935	-127	77	1,089	0.905
	BSM	1,146	42	1,291	0.615	-534	41	790	0.865	-426	54	873	0.93	-70	56	796	0.955
	NSBSM	-185	40	594	0.94	168	40	586	0.94	97	52	727	0.95	18	55	780	0.955
20%	BD	18	30	427	0.96	48	32	455	0.965	-10	42	600	0.955	-31	50	701	0.945
	CC	3,874	39	3,913	0	-1,818	37	1,910	0.12	-1,466	53	1,644	0.55	-218	65	941	0.925
	IL	3,885	39	3,923	0	-1,833	37	1,922	0.14	-1,478	52	1,653	0.55	-205	66	952	0.93
	NSMI	-924	50	1,164	0.695	717	39	1,049	0.84	490	69	1,087	0.91	-402	93	1,368	0.915
	BSM	1,116	45	1,281	0.61	-491	41	752	0.875	-411	53	849	0.91	-67	60	854	0.92
	NSBSM	-152	40	579	0.94	177	36	584	0.935	59	50	714	0.955	-53	60	847	0.95
10%	BD	26	30	424	0.96	-13	32	446	0.960	-28	40	566	0.975	33	40	569	0.975
	CC	5,434	37	5,459	0	-2,558	41	2,624	0.02	-2,059	50	2,178	0.195	-210	64	930	0.9
	IL	5,445	38	5,471	0	-2,565	43	2,637	0.03	-2,072	52	2,198	0.26	-199	67	961	0.92
	NSMI	-926	66	1,316	0.85	585	78	1,243	0.885	376	103	1,494	0.92	-205	129	1,830	0.895
	BSM	1,196	42	1,337	0.55	-553	39	782	0.855	-435	49	821	0.910	-42	56	793	0.95
	NSBSM	306	40	641	0.905	-138	40	580	0.96	-142	50	715	0.950	36	58	813	0.935

b_0 is the intercept. b_1, b_2, b_3 are the coefficients of $x_i, z_i, x_i \times z_i$, respectively. The true value of these parameters is 1.

Table 5. Posterior means, lower and upper 95% HPD intervals of QOL analysis.

	CC	IL	NSMI	BSM	NSBSM
Intercept	4.195 (4.024, 4.365)	4.193 (4.014, 4.373)	4.167 (3.935, 4.399)	4.193 (4.010, 4.396)	4.194 (4.006, 4.387)
Sex (male vs. female)	0.009 (-0.030, 0.047)	0.009 (-0.027, 0.046)	0.003 (-0.048, 0.055)	0.015 (-0.030, 0.061)	0.009 (-0.034, 0.053)
Race (White vs. non-white)	0.091 (0.023, 0.159)	0.090 (0.021, 0.160)	0.086 (0.005, 0.168)	0.094 (0.014, 0.172)	0.091 (0.016, 0.169)
Education (\geq HS vs. < HS)	0.089 (0.049, 0.130)	0.089 (0.048, 0.130)	0.081 (0.021, 0.141)	0.092 (0.044, 0.137)	0.089 (0.043, 0.133)
Age	-0.003 (-0.010, 0.005)	-0.003 (-0.010, 0.005)	-0.001 (-0.009, 0.008)	-0.003 (-0.011, 0.006)	-0.003 (-0.011, 0.006)

When the phase II missing data mechanism is simulated based on the logit model (Table 4) rather than a probit model, the NSBSM still outperforms competing methods. This implies that the NSBSM method is robust to slight violation of the model assumptions.

5. Application to QOL Dataset

We applied the proposed method to a quality of life (QOL) dataset from a community-based study — the Children in the Community study (CIC) (Cohen et al. (2005)). A brief description of this dataset can be found in Chen and Cohen (2006). The 750 participants sample was originally drawn from 100 neighborhoods in two upstate New York counties in 1975. From 1991 to 1994 (T1), these 750 youths (mean age of 22.0 years and SD of 2.8 years) were interviewed at home by trained interviewers. QOL was assessed using the young adult quality of life instrument (YAQOL) (Chen et al. (2004)). In 2001-2004 (T2) at mean age of 32.0 years (SD = 2.8 years), the same group of participants was surveyed using the same YAQOL instrument. Of the 750 participants assessed for QOL at T1, 603 (80.4%) completed the survey at T2 while 147 did not respond to the survey at T2 (phase I). For those 147 subjects who did not respond to the survey, an abridged version of the YAQOL instrument was mailed to their home address (phase II). Subjects were paid for their participation upon return of the completed surveys. Of the 147 eligible subjects, 39 (26.5%) returned their YAQOL instrument. Since phase II data collection was completed within three months of phase I, it was assumed that the YAQOL outcomes remained unchanged from phase I. We focused our analysis on the resources subscale of the YAQOL instrument.

The goal of the QOL analysis is to determine whether the resources subscale at T2 is related to major demographic variables — age, gender, race and education. We regressed the resources on gender (male versus female), age (in years), race (White vs. non-White), and education (High school or above vs. less than high school). We applied the CC analysis, IL method using multiple imputation, the NSMI method, the BSM method, and NSBSM method to the dataset. The NSMI method is valid if, among the 147 nonrespondents, the missingness after phase II is MAR, meeting the assumptions for NSMI. When phase I missingness is MNAR, the NSBSM method is more efficient than the BSM because the additional data collected from phase II provide valuable information for modeling the missing data mechanism and improve the identification of the model parameters.

When both phase I and phase II missingness are MNAR, the NSBSM method is the only method that provides valid estimation of the regression model for the resources subscale.

To correct for the skewness of the outcome and improve the posterior estimation, the resources subscales was log-transformed. The results from all five methods are shown in Table 5. The results are quite consistent across five different methods, although the effect of sex and age are a little weaker for the NSMI method. All methods showed a significant effect of age with White race having significantly more resources than non-White participants. Those with at least high-school education also had significantly more resources at mean age of 33 compared to those with less than high school education.

6. Discussion

Two-phase sampling has been proposed and used in surveys to adjust for non-response bias for more than five decades. The traditional methods (i.e., weighting) fail to make full use of the additional data collected in the second phase. Little research has demonstrated the utility of NTS sampling and answered the question of what proportion of the nonrespondents should be surveyed in the second phase. Zhang, Chen and Elliott (2016) provided an efficient NSMI method that yields valid estimates when the missing data mechanism in the subsample of nonrespondents is MAR, regardless of the missing data mechanism in phase I. We proposed the NSBSM method that improved over the NSMI methods by yielding valid inference even when the missing data mechanism in phase II is MNAR. Our simulation studies also showed that it is beneficial even by collecting data from a small proportion of the nonrespondents.

Prior literature in missing data has primarily focused on preventing and minimizing nonresponse in the data collection stage (Groves and Couper (2012)), and developing methods to handle missing data after the data collection (Little and Rubin (2002)). Many of the methods rely on assumptions that are untestable, which motivated the research in sensitivity analyses (Troxel, Ma and Heitjan (2004); Scharfstein et al. (2014)). Two-phase sampling provides an alternative way to remedy nonresponse. The utilities of NTS sampling are two-fold. It minimizes nonresponse by collecting additional data from nonrespondents; the additional data from the initial nonrespondents provide valuable information regarding the missing data mechanism and therefore improves the modeling of the missingness. The proposed NSBSM methods and the NSMI methods are effective

means to make use of the additional data from phase II sampling.

In theory, the NSBSM method could be extended to a full selection model specification by including a selection model for the missingness indicator in phase II, conditional on missingness in phase I. Such models are subject to the same identification issues with the traditional selection models. Other approaches, such as the multiply robust estimators (Han (2014)), the instrumental variable approach (Wang, Shao and Kim (2014)), and the pattern mixture model developed for repeated attempt design (Daniels et al. (2015)), may be alternative methods to a full selection model specification. The proposed NSBSM could be extended by modeling the covariance structure following Barnard, McCulloch and Meng (2000) to incorporate prior information in the posterior inference. The parametric assumptions of the NSBSM model could be extended to a t -model (Marchenko and Genton (2012)) for the error distributions or be relaxed through the nonparametric approach studied in Chib, Greenberg and Jeliazkov (2009). For data arising from complex survey setting, the weights can be incorporated in the pseudo-likelihood to reflect the design feature (Chambers et al. (2012)). These extensions will be the subject of future work.

Acknowledgment

We thank an Associate Editor and the three referees for their thoughtful and constructive comments which greatly improved the paper.

Appendix: MCMC Sampling Algorithm

We provide some additional details about the MCMC sampling algorithm regarding to the augmented likelihood

$$L(\boldsymbol{\theta}, \sigma^2, \sigma_{12} | \tilde{\mathbf{y}}) \propto |\boldsymbol{\Sigma} \otimes \mathbf{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})^T (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1} (\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})\right),$$

where $\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}^* \\ \mathbf{u} \end{pmatrix}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 + \sigma_{12}^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$, and \mathbf{I}_n is identity matrix with dimension n .

Gibbs sampling on $\boldsymbol{\theta}$

The full conditional distribution of $\boldsymbol{\theta}$ is of the form

$$p(\boldsymbol{\theta} | \tilde{\mathbf{y}}, \sigma^2, \sigma_{12})$$

$$\begin{aligned}
&\propto |\boldsymbol{\Sigma} \otimes \mathbf{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})^T(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})\right) \\
&\quad \times |\boldsymbol{\Psi}^{-1}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right) \\
&\propto \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{\Psi}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right) \\
&\propto \exp\left(-\frac{1}{2}(-\tilde{\mathbf{y}}^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n \mathbf{Z}\boldsymbol{\theta} - \boldsymbol{\theta}^T \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n \tilde{\mathbf{y}} + \boldsymbol{\theta}^T \mathbf{Z}^T \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n \mathbf{Z}\boldsymbol{\theta} \right. \\
&\quad \left. + \boldsymbol{\theta}^T \boldsymbol{\Psi}_0 \boldsymbol{\theta} - \boldsymbol{\theta}^T \boldsymbol{\Psi}_0 \boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^T \boldsymbol{\Psi}_0 \boldsymbol{\theta}_0)\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \tilde{\boldsymbol{\Psi}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right),
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\Psi}}^{-1}(\mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\tilde{\mathbf{y}} + \boldsymbol{\Psi}_0\boldsymbol{\theta}_0)$ and $\tilde{\boldsymbol{\Psi}} = \mathbf{Z}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\mathbf{Z} + \boldsymbol{\Psi}_0$.

Gibbs sampling on σ_{12}

The full conditional distribution of σ_{12} is given by

$$\begin{aligned}
p(\sigma_{12}|\cdot) &\propto |\boldsymbol{\Sigma} \otimes \mathbf{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})^T(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})\right) \\
&\quad \times \exp\left(-\frac{1}{2}d_0(\sigma_{12} - c_0)^2\right)
\end{aligned}$$

Since $\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\epsilon} \end{pmatrix}$ and $\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma^2 & -\sigma_{12}/\sigma^2 \\ -\sigma_{12}/\sigma^2 & 1 + \sigma_{12}^2/\sigma^2 \end{pmatrix}$,

$$\begin{aligned}
p(\sigma_{12}|\cdot) &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\epsilon} \end{pmatrix}^T \begin{pmatrix} 1/\sigma^2 & -\sigma_{12}/\sigma^2 \\ -\sigma_{12}/\sigma^2 & 1 + \sigma_{12}^2/\sigma^2 \end{pmatrix} \otimes \mathbf{I}_n \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\epsilon} \end{pmatrix}\right) \\
&\quad \times \exp\left(-\frac{1}{2}d_0(\sigma_{12} - c_0)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \sigma_{12}^2}{\sigma^2} - \frac{2\boldsymbol{\eta}^T \boldsymbol{\epsilon} \sigma_{12}}{\sigma^2} + d_0(\sigma_{12} - c_0)^2\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\tilde{d}(\sigma_{12} - \tilde{c})^2\right),
\end{aligned}$$

where $\tilde{c} = \tilde{d}^{-1}[\sigma^{-2}\boldsymbol{\eta}'\boldsymbol{\epsilon} + c_0d_0]$ and $\tilde{d} = \sigma^{-2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} + d_0$.

Gibbs sampling on σ^{-2}

The full conditional distribution of σ_{12} takes the form

$$p(\sigma^{-2}|\cdot) \propto |\boldsymbol{\Sigma} \otimes \mathbf{I}_n|^{-1/2} \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})^T(\boldsymbol{\Sigma} \otimes \mathbf{I}_n)^{-1}(\tilde{\mathbf{y}} - \mathbf{Z}\boldsymbol{\theta})\right)$$

$$\begin{aligned}
& \times (\sigma^{-2})^{v_0/2-1} \exp\left(-\frac{w_0}{2}\sigma^{-2}\right) \\
& \propto (\sigma^{-2})^{n/2} \exp\left(-\frac{1}{2}\sigma^{-2}(\boldsymbol{\eta}^T \boldsymbol{\eta} - 2\boldsymbol{\eta}^T \boldsymbol{\epsilon}\sigma_{12} + \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}\sigma_{12}^2)\right) \\
& \quad \times (\sigma^{-2})^{v_0/2-1} \exp\left(-\frac{w_0}{2}\sigma^{-2}\right) \\
& \propto (\sigma^{-2})^{(v_0+n)/2-1} \exp\left(-\frac{w_0 + (\boldsymbol{\eta} - \boldsymbol{\epsilon}\sigma_{12})^T(\boldsymbol{\eta} - \boldsymbol{\epsilon}\sigma_{12})}{2}\sigma^{-2}\right).
\end{aligned}$$

References

- Barnard, J., McCulloch, R. and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.
- Chambers, R. L., Steel, D. G., Wang, S. and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. CRC Press.
- Chen, H. and Cohen, P. (2006). Using individual growth model to analyze the change in quality of life from adolescence to adulthood. *Health and Quality of Life Outcomes* **4**, 1.
- Chen, H., Cohen, P., Kasen, S., Gordan, K., Dufur, R. and Smailes, E. (2004). Construction and validation of a quality of life instrument for young adults. *Quality of Life Research* **13**, 747–759.
- Chib, S., Greenberg, E. and Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics* **18**, 321–348.
- Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons.
- Cohen, P., Crawford, T. N., Johnson, J. G. and Kasen, S. (2005). The children in the community study of developmental course of personality disorder. *Journal of Personality Disorders* **19**, 466.
- Daniels, M. J., Jackson, D., Feng, W. and White, I. R. (2015). Pattern mixture models for the analysis of repeated attempt designs. *Biometrics* **71**, 1160–1167.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B: Statistical Methodological* **39**, 1–22.
- Elliott, M. R. and Little, R. J. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics* **16**, 191.
- Groves, R. M. and Couper, M. P. (2012). *Nonresponse in Household Interview Surveys*. John Wiley & Sons.
- Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109**, 1159–1173.
- Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association* **41**, 517–529.
- Harter, R., Mach, T., Wolken, J. and Chapline, J. (2007). Determining subsampling rates for nonrespondents. In: *Third International Conference on Establishment Surveys*, Montréal, Canada.

- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.
- Koop, G. and Poirier, D. J. (1997). Learning about the across-regime correlation in switching regression models. *Journal of Econometrics* **78**, 217–227.
- Li, K. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics* **85**, 387–400.
- Little, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.
- Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.
- Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. John Wiley & Sons.
- Marchenko, Y. V. and Genton, M. G. (2012). A heckman selection-t model. *Journal of the American Statistical Association* **107**, 304–317.
- Nandram, B. and Choi, J. W. (2002). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association* **97**, 381–388.
- Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association* **105**, 120–135.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. **81**, John Wiley & Sons.
- Scharfstein, D., McDermott, A., Olson, W. and Wiegand, F. (2014). Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Statistics in Biopharmaceutical Research* **6**, 338–348.
- Srinath, K. (1971). Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association* **66**, 583–586.
- Sugden, R. and Smith, T. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika* **71**, 495–506.
- Troxel, A. B., Ma, G. and Heitjan, D. F. (2004). An index of local sensitivity to nonignorability. *Statistica Sinica* **14**, 1221–1237.
- Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.
- Zhang, N., Chen, H. and Elliott, M. R. (2016). Nonrespondent subsample multiple imputation in two-phase sampling for nonresponse. *Journal of Official Statistics* **32**, 769–785.
- Zhu, H., Ibrahim, J. G. and Tang, N. (2014). Bayesian sensitivity analysis of statistical models with missing data. *Statistica Sinica* **24**, 871.

Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, PR China.

E-mail: yue.zhang@sjtu.edu.cn

Department of Epidemiology and Biostatistics, University of South Florida, Tampa, FL 33612, USA.

E-mail: hchen1@health.usf.edu

Division of Biostatistics and Epidemiology, Cincinnati Childrens Hospital Medical Center,

Cincinnati, OH 45229, USA.

Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA.

E-mail: Nanhua.Zhang@cchmc.org

(Received June 2016; accepted December 2017)