# Estimation of Random-Effects Model for Longitudinal Data with Nonignorable Missingness using Gibbs Sampling
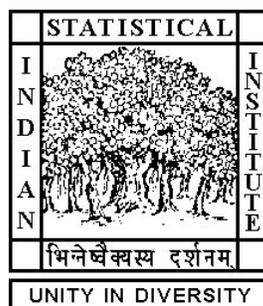
**Prajamitra Bhuyan**

Applied Statistics Unit,

Indian Statistical Institute,

Kolkata 700108

*bhuyan.prajamitra@gmail.com*

STATISTICAL
INDIAN
INSTITUTE
भिन्नेष्वैक्यस्य दर्शनम्
UNITY IN DIVERSITY

# Estimation of Random-Effects Model for Longitudinal Data with Nonignorable Missingness using Gibbs Sampling

**Prajamitra Bhuyan**

Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108.

### Abstract

The missing data problem is common in longitudinal or repeated measurements data. When the missingness mechanism is nonignorable, the distribution of the observed response and indicators of missingness should be modelled jointly using either 'shared random-effects model' or 'correlated random-effects model'. However, computational challenges arise in the model fitting due to intractable numerical integration involved in the log-likelihood function. We provide an alternative modeling of 'correlated random-effects model' using latent variables and propose simple algorithm based on Gibbs sampling for estimation of associated parameters. The method is illustrated through simulation and the analysis of a real data set arising from an autism study.

## 1   Introduction

In designed longitudinal studies, the aim is to estimate the mean response at a certain time based on fixed or time-varying covariates. For studies with long follow up periods, the proportion of individuals with missing data can be substantial. Inference based on the observed data may lead to biased and unreliable results. Ample amount of research work on modeling of longitudinal data with ignorable missingness are available in the literature (Little and Rubin, 2002). In this paper, we focus on modeling of longitudinal

data when the missingness mechanism is nonignorable. In these cases, the distribution of the observed response and indicators of missingness should be modelled jointly. Such joint models can be classified into either pattern mixture models or selection models. Little (1995) provided a detailed Overview of pattern mixture models and selection models for longitudinal studies with missing data due to informative dropout. Siddiqui and Ali (1998) considered random-effects pattern-mixture models and provided estimates of the associated parameters by averaging the estimates obtained from different subsets of the data depending on the missing data-patterns. Daniels and Hogan (2000) proposed a reparameterization of the pattern mixture model that allows consideration of a wide range of nonignorable missing-data mechanisms. Diggle and Kenward (1994) proposed the use of a selection model with a logistic regression form to deal with informative dropout. Baker (1995) considered repeated binary data with nonignorable and non-monotone missingness. Troxel et al. (1998) extended those of Diggle and Kenward (1994) for non-monotone and nonignorable missing data. However, its implementation is computationally challenging. Rotnitzky et al. (1998) considered inverse probability weighted estimating equations for the nonignorable missing and provided simultaneous estimation of the dropout probability and mean response based on selection model. Minini and Chavance (2004) considered a log-linear model and provided sensitivity analysis for the longitudinal binary data with nonignorable missingness.

In the context of survival analysis, shared random-effects (SRE) model proposed by Wu and Carrol (1998), is popularly used as an alternative to the selection models. De and Tu (1994), and Schluchter (1994) suggested some extensions of SRE models. Have et al. (1998) and Pulkstenis et al. (1998) adopted the SRE model for binary longitudinal data with the informative dropouts. Tsonaka et al. (2009) considered semi-parametric shared parameter model for the modelling of the response variable with non-monotone and nonignourable missingness. In the aforementioned SRE models, the selection model and the response model have exactly the same random component.

In many situations, the latent factors affecting the missingness could be different from those affecting the response; however, they are correlated due to common risk

factors. In order to model such phenomena, Lin et al. (2010) considered an interesting generalization of SRE model by using correlated random effects model. An underlying assumption for the random-effects model is that, conditional on the random effects, the missingness is independent of the response. Note that ignorable missingness is a special case of the correlated random-effects model if the two random effects are independent. A main concern in the random-effects models is computational challenges arise in the model fitting due to intractable numerical integration involved in the log-likelihood function. In order to overcome such difficulties, well-known approximation methods like the Gauss-Hermite quadrature (Pinheiro and Bates, 1995) and the Laplace approximation (Breslow and Clayton, 1993) are exploited for estimation purposes. Lin et al. (2010) expressed the likelihood as a ratio of two integrals and then approximated the numerator and denominator using the Laplace formula. In order to estimate the associated model parameters, one needs to evaluate first and second derivatives and maximize the two integrands.

We propose an alternative modeling of the observed response and indicators of missingness based on correlated latent variables. In particular, we develop regression models with the covariates having time-varying effect and time-invariant effects on the latent variables involving correlated random effects. A simple Gibbs sampler is developed following Albert and Chib (1993), where in each iteration, we sample the model parameters as well as the latent variables. Our method is simple because it is based on the Gibbs sampler, and it is fast since we estimate the parameters for both the models simultaneously in an automated manner and avoids the computational challenges posed by intractable log-likelihood functions typically encountered in the frequentist method. The rest of the paper is organized as the follows. In Section 2, we discuss our proposed model and the Bayesian estimation method in detail. We analyze data from longitudinal study of the social development of children with autism in Section 3. Simulation studies are performed for assessing the effectiveness of the proposed method and the results are finally summarized in Section 4. In Section 5, we provide outlines of possible future work and some concluding remarks.

## 2 Proposed Model

In the following presentation, we consider a continuous response measured over $m$ different time points from $n$ subjects. We consider a set of covariates, some of which possibly have time-varying effect on the response. The response for the $i$-th subject at the $t$-th time point, which we denote by $Y_i(t)$, can thus be modelled as the following:

$$Y_i(t) = \sum_{j=1}^{J} \beta_j(t) X_{ji}(t) + \sum_{j'=1}^{J'} \gamma_{j'} Z_{j'i}(t) + \sum_{l=1}^{L} u_{li} \tilde{Z}_{li}(t) + e_i(t), \tag{1}$$

where we consider $J$, $J'$ and $L$ denote the number of covariates with time-varying fixed effects, time-invariant fixed effects and random effects on response, respectively. Subject-specific random effects $\mathbf{u}_i = (u_{1i}, \ldots, u_{Li})$'s capture the longitudinal dependence and are assumed to be iid $N_L(\mathbf{0}, \Sigma_u)$. The residuals $e_i(t)$'s are assumed to be iid $N(0, \sigma_e^2)$. Note that the above model is a special case of generalized varying coefficient model for longitudinal data, introduced by Sentrk et al. (2013).

In general, the data from longitudinal studies involving large number of people possesses some missing responses. In many situations, the missing data mechanism is ignorable and it can be well handled using several available methodologies under the assumption of missing at random (MAR) (Little, 1995). However, when there are nonignorable missing values in the response variable, the models in equation (1) will produce biased estimates under the MAR assumption. Such data are not so uncommon, specially in biomedical studies and social sciences (Little and Rubin, 1987, Ch-1). In order to address such issues, we first define $U_i(t)$ as

$$U_i(t) = \begin{cases} 1, \text{ if } & Y_i(t) \quad \text{is observed}, \\ 0, \text{ if } & Y_i(t) \quad \text{is missing}, \end{cases}$$

and rewrite $Y_i(t)$ as

$$Y_i(t) = \begin{cases} Y_i^*(t), \text{ if } & U_i(t) = 1, \\ \text{missing, if } & U_i(t) = 0, \end{cases}$$

4

where $Y_i^*(t)$ is a latent random variable. We then rewrite the regression model given in equation (1) in terms of the latent random variables as follows

$$Y_i^*(t) = \sum_{j=1}^{J} \beta_j(t)X_{ji}(t) + \sum_{j'=1}^{J'} \gamma_{j'}Z_{j'i}(t) + \sum_{l=1}^{L} u_{li}\tilde{Z}_{li}(t) + e_i(t). \tag{2}$$

In addition, we consider the latent variable $U_i^*(t)$ and write

$$U_i(t) = \begin{cases} 1, \text{ if } & U_i^*(t) > 0, \\ 0, \text{ if } & U_i^*(t) \leq 0. \end{cases}$$

Now we consider the following model for missing data mechanism with some covariates as

$$U_i^*(t) = \sum_{k=1}^{K} \theta_k(t)W_{ki}(t) + \sum_{k'=1}^{K'} \delta_{k'}S_{k'i}(t) + \sum_{l=1}^{L} v_{li}\tilde{Z}_{li}(t) + \epsilon_i(t), \tag{3}$$

where the random effects $\mathbf{v}_i = (v_{1i}, \ldots, v_{Li})$'s are assumed to be iid $N_L(\mathbf{0}, \Sigma_v)$, and the residuals $\epsilon_i(t)$ are iid $N(0,1)$. In order to incorporate the possible correlation between the response variable $Y_i(t)$ and the missing indicator $U_i(t)$, we consider $\mathbf{u}_i$ and $\mathbf{v}_i$ are correlated random vectors following a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma = \left( \begin{smallmatrix} \Sigma_u & \Sigma_{uv} \\ \Sigma_{uv} & \Sigma_v \end{smallmatrix} \right)$. For the aforementioned models, we can use the usual models for multivariate longitudinal data (Diggle et al., 2002, p. 332) but this requires values of the latent variables at each step. We propose a Bayesian estimation method for simultaneous estimation of the parameters associated with the joint model using Gibbs sampling.

## 2.1 Modelling Time-varying Coefficients

One of the major advantages of the longitudinal studies is to incorporate age effect in the modeling and its capacity to distinguish changes in the response within and across individuals over time (Diggle et al., 2002, p. 1). In order to model the effects of the respective covariates over time, we have considered the time-varying coefficients

$\beta_j(t)$ and $\theta_k(t)$ in equations (2) and (3), respectively. Since parametric nature of $\beta_j(t)$ and $\theta_k(t)$ is not known in advance, we consider semi-parametric approach of modelling the time-varying coefficients using Legendre Polynomials (LP) basis functions. These Polynomials have already been proven as powerful tool by several authors for semi-parametric regression (Marie and Sen, 1985; Meyer, 2000; Cui and Zhu, 2006).

The general form of a Legendre Polynomial of order $r$ is given by the following sum

$$P_r(x) = \sum_{l=0}^{L}(-1)^l \frac{(2r-2l)!}{2^r l!(r-l)!(r-2l)!}x^{r-2l},$$

where $L=\frac{r}{2}$ or $\frac{r-1}{2}$, whichever is an integer. These polynomials are defined over [-1, 1] and are orthogonal to each other in this interval in the sense that the inner product $\int_{-1}^{1}P_r(x)P_s(x)dx = 0$, for $r \neq s$. First few LPs are as the following:
$P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$, $P_3(x) = \frac{1}{2}(5x^3 - 3x)$.

In our context, we denote the $r$-th order Legendre Polynomial (LP) at time $t$ by $P_r(t)$. We transform the original time points $t$ to the adjusted time points $t' = -1 + 2(\frac{t-t_{min}}{t_{max}-t_{min}})$, for fitting the orthogonal LP over the range [-1, 1], where $t_{min}$ and $t_{max}$ are the smallest and the highest time points respectively. Let $P^{(r)}(t') = [P_0(t'), P_1(t'), \ldots, P_r(t')]^T$ denote the family of the first $r+1$ basis functions and express the functions $\beta_j(t')$ and $\theta_k(t')$ as some linear combinations of these basis functions:

$$\beta_j(t) = \mathbf{a}_j^T P^{(r_1)}(t); \quad \theta_k(t) = \mathbf{b}_k^T P^{(r_2)}(t),$$

where $\mathbf{a}_j = (a_{j0}, a_{j1}, \ldots, a_{jr_1})^T$, and $\mathbf{b}_k = (b_{k0}, b_{k1}, \ldots, b_{kr_2})^T$ are called base vectors. The optimal orders $r_1$ and $r_2$ may be chosen by the information criteria, e.g. AIC/BIC etc. Unless great flexibility is required, very low values of $r_l$ such as 1 or 2 will suffice, for $l = 1, 2$. For example, let us consider $J = 2$, $J' = 0$, $L = 1$, $r_1 = r_2 = 1$, $X_{1i}(t) = 1$, $\tilde{Z}_{1i}(t) = 1$ and $X_{2i}(t) \equiv X_{2i}$ is a time-invariant covariate. Then equation (2) and (3) reduces to

$$Y_i^*(t) = \alpha_{10} + \alpha_{11}t + \alpha_{20}X_{2i} + \alpha_{21}tX_{2i} + u_{1i} + e_i(t), \qquad (4)$$

and

$$U_i^*(t) = \lambda_{10} + \lambda_{11}t + \lambda_{20}X_{2i} + \lambda_{21}tX_{2i} + v_{1i} + \epsilon_i(t), \tag{5}$$

respectively, where $\boldsymbol{\alpha}_j = (\alpha_{j0}, \alpha_{j1})^T$'s and $\boldsymbol{\lambda}_j = (\lambda_{j0}, \lambda_{j1})^T$'s are suitably adjusted for the change in location and scale in time, for $j = 1, 2$. The models (4) and (5) account for the age effects and its interaction with the covariate $X_{2i}(t)$. Note that the parameters involved in LPs possesses interesting interpretations and it is easy to implement. Moreover, computational complexities (e.g., knot selection, knot location) related to the other basis functions can be avoided.

## 2.2 Bayesian Estimation using Gibbs Sampler

We employ a Bayesian approach of estimating the model parameters for the equations (2) and (3) using the Gibbs sampler. Let $\Theta = [\mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma_e^2, \Sigma]$ denote the set of all the model parameters involved in the equations (2) and (3), where the bold symbols denote the vector of the respective coefficients. Note that, one needs to sample from the joint posterior of the model parameters and unknown latent variables.

Let us denote $Y^* = (Y_{11}^*(t), \ldots, Y_{nm}^*(t))$, $U^* = (U_{11}^*(t), \ldots, U_{nm}^*(t))$, $Y = (Y_{11}(t), \ldots, Y_{nm}(t))$, and $U = (U_{11}(t), \ldots, U_{nm}(t))$, and write the joint posterior density for the latent variables and the parameters associated with the proposed model as

$$\pi(\Theta, Y^*, U^*|Y, U) \propto \pi(\mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma_e^2, \Sigma) \times \prod_{i=1}^{n} \int \prod_{t=1}^{m} f(Y_i^*(t), U_i^*(t)|u_i, v_i)$$

$$\times \left\{ 1(U_i^*(t) > 0)1(U_i(t) = 1) + 1(U_i^*(t) \leq 0)1(U_i(t) = 0) \right\} g(u_i, v_i) du_i dv_i,$$

where $\pi(\cdot)$ is the joint prior for $\Theta$, and $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ are the joint distribution of $(Y_i^*(t), U_i^*(t))$, and $(\mathbf{u}_i, \mathbf{v}_i)$, respectively. Following the traditional Bayesian regression, we consider the non-informative priors for $(\mathbf{a}, \boldsymbol{\gamma}, \sigma_e^2)$ and $(\mathbf{b}, \boldsymbol{\delta})$. In addition, we consider maximal data information prior for $\Sigma$. Therefore the joint prior distribution $\pi(\mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma_e^2, \Sigma)$, can be expressed as the product of $\pi(\mathbf{a}, \boldsymbol{\gamma}, \sigma_e^2) \propto \frac{1}{\sigma_e^2}$, $\pi(\mathbf{b}, \boldsymbol{\delta}) \propto 1$,

and $\pi(\Sigma) \propto \frac{1}{|\Sigma|}$.

The posterior distributions of $\Theta$ conditional on $Y^*, U^*, Y, U$ can be derived routinely and hence we skip the details. The full conditionals for $(\mathbf{a}, \mathbf{b}, \boldsymbol{\gamma}, \boldsymbol{\delta})$, $\sigma_e^2$, and $\Sigma$ are normal, inverse gamma and inverse-Wishart, respectively. The latent variable $Y_i^*(t)$ and $U_i^*(t)$ are sampled from the conditional densities

$$
\begin{cases}
Y_i(t) \quad \text{with probability 1,} & \text{if } U_i(t) = 1, \\
N\left(\sum_{j=1}^{J} \beta_j(t) X_{ji}(t) + \sum_{j'=1}^{J'} \gamma_{j'} Z_{ji}(t) + \sum_{l=1}^{L} u_{li} \tilde{Z}_{li}(t), \sigma_e^2\right), & \text{if } U_i(t) = 0,
\end{cases}
\tag{6}
$$

and

$$
\begin{cases}
N\left(\sum_{k=1}^{K} \theta_k(t) W_{ki}(t) + \sum_{k'=1}^{K'} \delta_{k'} S_{k'i}(t) + \sum_{l=1}^{L} v_{li} \tilde{Z}_{li}(t), 1\right) \text{ left truncated at 0,} & \text{if } U_i(t) = 1, \\
N\left(\sum_{k=1}^{K} \theta_k(t) W_{ki}(t) + \sum_{k'=1}^{K'} \delta_{k'} S_{k'i}(t) + \sum_{l=1}^{L} v_{li} \tilde{Z}_{li}(t), 1\right) \text{ right truncated at 0,} & \text{if } U_i(t) = 0,
\end{cases}
\tag{7}
$$

respectively. Note that the full conditional densities of the latent variables, given by (6) and (7), are also from standard densities, and hence, one can directly apply Gibbs sampler algorithm for estimation of model parameters $\Theta$.

## 3  Data Analysis

We apply our proposed method on the data arising from a prospective longitudinal study of the social development of children with autism. A total of 214 children participated in the study who were divided into three diagnostic groups at 2 years of age: autism, pervasive developmental disorder (PDD), and nonspectrum children. We consider a subset of 158 autism spectrum disorder (ASD) children, and social development information was collected for each child at ages 2, 3, 5, 9 and 13 years based on a parent-reported survey. The objective was to assess the development trajectories of these children's socialization for different language proficiency groups. The response variable, Vineland Socialization Age Equivalent (VSAE), was a combined score that included assessments

8

of interpersonal relationships, play/leisure time activities and coping skills. However, the measurements corresponding to all children at each age are not available, which resulted in 22% missing in the response variable. The childrens language development was assessed by the Sequenced Inventory of Communication Development (SICD) score at age 2, and children were categorized into three groups (SICDEGP) based on their SICD scores. The data were collected by researchers at the University of Michigan and analyzed in West et al. (2007, Ch-6) under MAR assumption.

Let $Y_i(t)$ be the $\log(VSAE)$ score of the $i$th child at time $t$, where $t = \log(Age)$. Lin et al. (2010) observed that the general trend of VSAE score is increasing with age, while there is a substantial variation of the VSAE scores among the children, and hence logarithmic transformation has been considered. In order to incorporate the categorical variable SCIDEGP, we introduce two dummy variables SCI2 and SCI3 representing the second and third level of SICD group, respectively, and we take SICDEGP=1 as the reference group. As discussed in the Subsection 2.1, we consider first order LP for modeling of time-varying coefficients and the models (4) and (5) can be rewritten as

$$Y_i^*(t) = \alpha_{10} + \alpha_{11}t + \alpha_{20}SIC2_{2i} + \alpha_{21}tSIC2_{2i} + \alpha_{30}SIC3_{2i} + \alpha_{31}tSIC3_{2i} + u_{1i} + e_i(t),$$

and

$$U_i^*(t) = \lambda_{10} + \lambda_{11}t + \lambda_{20}SIC2_{2i} + \lambda_{21}tSIC2_{2i} + \lambda_{30}SIC3_{2i} + \lambda_{31}tSIC3_{2i} + v_{1i} + \epsilon_i(t),$$

respectively. The model parameters are estimated by the Gibbs sampler, as discussed in Section 2.2. We run MCMC for 50,000 iterations and discard the first 10,000 iterations as burn-in. We also thin the chains by saving every 10-th iteration. The convergence of the chains are monitored graphically. The results are summarized in Table 1, which are consistent with the finding of Lin et al. (2010) based on frequentist approach. Our primary interest is to estimate the interaction effect between SICDEGP and age. Note that SICDEGP group 1 has been taken as the reference group and the estimate of the

9

interaction effect of the $\log(Age)$ with corresponding SIC3 is larger than that of with SIC2 in magnitude. One can interpret that children in higher SICD groups can easily be socialized as they grow up. As expected, we find that age is positively associated with the missingness of VSAE score. Interestingly, the children in higher SICD group are more likely to be observed as age increases compared to those in lower SICD group. The estimates of $\sigma_e^2$ and $\Sigma$ are 0.18 and $\left( \begin{smallmatrix} 0.151 & -0.056 \\ -0.056 & 0.038 \end{smallmatrix} \right)$, respectively. The negative estimate for $\Sigma_{uv}$ suggests that children with higher VSAE scores are more likely to have missing outcomes. In order to asses the usefulness of the joint analysis, we test $H_0 : |\rho_{uv}| \leq \delta$, where $\rho_{uv}$ is the correlation coefficient between the random effects $u_{1i}$ and $v_{1i}$. We calculate the posterior probability $P(|\rho_{uv}| > \delta | \text{Data})$ for different choices of $\delta$, and the results are presented in Figure 1. It is evident that the posterior probability against the null hypothesis is very high and hence, we can conclude that the joint modeling is useful for our analysis.

Table 1: The estimated parameters for the autism study

| Predictor | $U_i^*(t)$ Mean | SD | 95% C.I. | $\log(VSAE)$ Mean | SD | 95% C.I. |
|---|---|---|---|---|---|---|
| Intercept | 2.279 | 0.309 | (1.712,2.909) | 1.503 | 0.085 | (1.332,1.669) |
| $\log(Age)$ | -0.873 | 0.153 | (-1.171,-0.586) | 0.681 | 0.045 | (0.593,0.769) |
| SIC2 | -0.207 | 0.390 | (-0.978, 0.568) | 0.074 | 0.113 | (-0.152,0.295) |
| SIC3 | -0.309 | 0.442 | (-1.195, 0.560) | 0.328 | 0.126 | (0.084,0.580) |
| SIC2$\times \log(Age)$ | 0.119 | 0.196 | (-0.287,0.491) | 0.118 | 0.059 | (-0.002, 0.231) |
| SIC3$\times \log(Age)$ | 0.253 | 0.222 | (-0.172, 0.685) | 0.312 | 0.067 | (0.183,0.439) |

## 4   Simulation Study

In order to study the performance of the proposed method, we generate data from the following model:

$$Y_i^*(t) = \beta_1 + \beta_2(t)X_{2i}(t) + \gamma_1 Z_{1i} + u_{1i} + e_i(t),$$
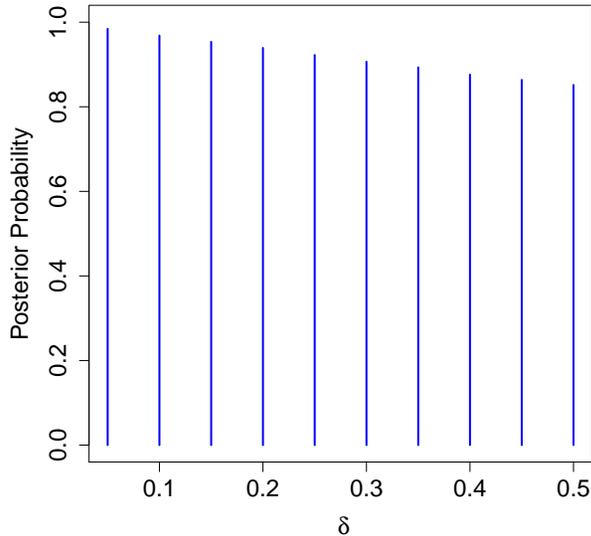
Figure 1: Posterior probability $P\left(|\rho_{uv}| > \delta|\text{Data}\right)$ for different choices of $\delta$

and

$$U_i^*(t) = \theta_1 + \theta_2(t)X_{2i}(t) + \delta_1 Z_{1i} + v_{1i} + \epsilon_i(t),$$

where $\beta_2(t) = \beta_{20} + \beta_{21}t$, and $\theta_2(t) = \theta_{20} + \theta_{21}t$. We first generate data for $n = 100$ subjects at $m = 5$ evenly spaced time points, with $\beta_1 = 10$, $\beta_{20} = 2$, $\beta_{21} = 5$ $\gamma_1 = 15$, $\theta_1 = 0$, $\theta_{20} = 15$, $\theta_{21} = -3$, and $\delta_1 = -1$. The time dependent covariate $X_{2i}(t)$ is generated from uniform distribution with support $(0, 2)$ and the time-invariant covariate $Z_{1i}$ is generated from Bernoulli distribution with mean 0.6. We consider the error variance $\sigma_e^2 = 9$, and the covariance matrix associated with the random effect parameters $\Sigma = \left(\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right)$. Under the above choice of the parameters almost 20% observations corresponding to the response variable are missing, which is in line with the data in the real example discussed in the previous section. For the purpose of camparison, we also generate complete data from the model

$$Y_i(t) = \beta_1 + \beta_2(t)X_{2i}(t) + \gamma_1 Z_{1i} + u_{1i} + e_i(t),$$

11

with the same parameter choices. For each simulated data, we generates samples from the posterior distribution of the parameters $\beta_1$, $\beta_{20}$, $\beta_{21}$, and $\gamma_1$ using the Gibbs sampler, and computed posterior mean and posterior standard deviations. This is repeated 1000 and the estimates are averaged over the 1000 simulations. We have also performed the same exercise under MAR assumption. In order to compare the performance of these models, we presented relative bias (RB), posterior standard deviation (SD) along with coverage probability in Table 2. Comparing the results from the complete and missing data analysis, we find that the estimates from our proposed method based on missing data mechanism is as good as complete data estimates. It is also evident that the proposed model performs better compared to MAR in terms of relative bias. Next, we generate data for $n = 200$ subjects at $m = 10$ evenly spaced time with the same parameter choice, which resulted in 40% missing in the response variable. The results are summarized in Table 3. As expected, posterior standard deviation decreases for the estimates for both the complete and missing data analysis with increase in the number of observations. Even with this higher percentage (40%) of missing observations, our method seems to perform reasonably well in comparison with complete data analysis. Here also, the proposed model is superior compared to MAR with respect to relative bias and coverage probability.

In order to study the sensitivity of the proposed method under mis-specified selection model, we consider the aforementioned response model with the missing data mechanism given by

$$U_i^*(t) = \theta_1 + \theta_2(t)X_{2i}(t) + \delta_1 Z_{1i} + \eta(Y_i(t) - \xi) + \epsilon_i(t),$$

with $\eta = 0.5$ and $\xi = 35$. Note that the missing indicator $U_i(t)$ depends on the response $Y_i(t)$ through the fixed effect parameter $\eta$. Next, we generate data for $n = 100$, $m = 5$, and $n = 200$, $m = 10$, which resulted in 30% missing in the response variable. The summarized results are presented in Tables 4-5. It is not surprising that the estimates from the proposed method are biased. However, it is interesting to note that the proposed estimates performs better compared to MAR under the mis-specified selection mechanism.

12

Table 2: Results of the simulation study with 20% missing observations ( $n = 100$ and $m = 5$)

| Parameter | Complete Data | | | Proposed | | | MAR | | |
|---|---|---|---|---|---|---|---|---|---|
| | RB | SD | CP | RB | SD | CP | RB | SD | CP |
| $\beta_1$ | -0.791 | 0.402 | 0.96 | 0.287 | 0.426 | 0.95 | -1.270 | 0.436 | 0.94 |
| $\beta_{20}$ | -0.510 | 0.354 | 0.95 | 1.267 | 0.409 | 0.95 | 10.825 | 0.408 | 0.85 |
| $\beta_{21}$ | 0.393 | 0.084 | 0.96 | -0.383 | 0.102 | 0.94 | -0.802 | 0.102 | 0.93 |
| $\gamma_1$ | 0.067 | 0.403 | 0.93 | -0.296 | 0.392 | 0.93 | -0.546 | 0.411 | 0.94 |
| $\sigma_u^2$ | 1.997 | 0.589 | 0.94 | 5.538 | 0.660 | 0.94 | -13.042 | 0.620 | 0.80 |
| $\sigma_e^2$ | 1.328 | 0.657 | 0.95 | 2.669 | 0.719 | 0.95 | 4.825 | 0.774 | 0.90 |

Table 3: Results of the simulation study with 40% missing observations ( $n = 200$ and $m = 10$)

| Parameter | Complete Data | | | Proposed | | | MAR | | |
|---|---|---|---|---|---|---|---|---|---|
| | RB | SD | CP | RB | SD | CP | RB | SD | CP |
| $\beta_1$ | 0.239 | 0.226 | 0.94 | 0.589 | 0.279 | 0.95 | -2.173 | 0.297 | 0.93 |
| $\beta_{20}$ | -0.511 | 0.165 | 0.94 | -2.059 | 0.280 | 0.94 | 12.71 | 0.273 | 0.85 |
| $\beta_{21}$ | 0.019 | 0.020 | 0.96 | -0.004 | 0.071 | 0.95 | -0.913 | 0.070 | 0.88 |
| $\gamma_1$ | 0.001 | 0.246 | 0.94 | -0.295 | 0.254 | 0.93 | -0.670 | 0.294 | 0.95 |
| $\sigma_u^2$ | -1.079 | 0.295 | 0.96 | 2.731 | 0.441 | 0.95 | -8.465 | 0.445 | 0.88 |
| $\sigma_e^2$ | 0.238 | 0.301 | 0.94 | -0.135 | 0.491 | 0.94 | 0.975 | 0.513 | 0.90 |

Table 4: Results of the simulation study under mis-specified selection model ( $n = 100$ and $m = 5$)

| Parameter | Proposed | | | MAR | | |
|---|---|---|---|---|---|---|
| | RB | SD | CP | RB | SD | CP |
| $\beta_1$ | 14.556 | 0.646 | 0.40 | 17.219 | 0.672 | 0.31 |
| $\beta_{20}$ | -38.817 | 0.469 | 0.65 | -49.398 | 0.482 | 0.42 |
| $\beta_{21}$ | -0.202 | 0.085 | 0.96 | 0.207 | 0.087 | 0.94 |
| $\gamma_1$ | -1.648 | 0.415 | 0.85 | -2.627 | 0.447 | 0.83 |
| $\sigma_u^2$ | -6.261 | 0.511 | 0.83 | -18.670 | 0.669 | 0.83 |
| $\sigma_e^2$ | 1.143 | 0.753 | 0.92 | 1.633 | 0.840 | 0.92 |

Table 5: Results of the simulation study under mis-specified selection model ( $n = 200$ and $m = 10$)

| Parameter | Proposed | | | MAR | | |
|---|---|---|---|---|---|---|
| | RB | SD | CP | RB | SD | CP |
| $\beta_1$ | 17.473 | 0.340 | 0.01 | 18.943 | 0.363 | 0 |
| $\beta_{20}$ | -51.309 | 0.232 | 0.03 | -56.239 | 0.236 | 0 |
| $\beta_{21}$ | 0.245 | 0.021 | 0.92 | 0.326 | 0.021 | 0.85 |
| $\gamma_1$ | -2.378 | 0.222 | 0.65 | -2.687 | 0.271 | 0.65 |
| $\sigma_u^2$ | -5.605 | 0.263 | 0.80 | -6.058 | 0.334 | 0.86 |
| $\sigma_e^2$ | -1.873 | 0.355 | 0.87 | -2.460 | 0.369 | 0.89 |

# 5   Discussion

In this paper, a Bayesian methodology has been developed for estimation of correlated random-effects model for longitudinal data with nonignorable missingness. Unlike the existing frequentist methods that require approximation of the intractable log-likelihood function, we provide a simple estimation methodology using Gibbs sampler. Our method is easy to implement, and as a special case, it can be applicable for various other models. For example, traditional SRE models are a special cases of the correlated random-effects models, which are popularly used for modeling nonignourable missing data. Moreover, the proposed method is also readily applicable for the data with ignorable missingness. The simulation results indicate that the estimates from our proposed method with missing information in response, are as good as compared to the estimates from complete data analysis. Moreover, the performance of the proposed method is superior compared to MAR even under the mis-specified selection mechanism.

In this paper, we have considered longitudinal data with continuous response variable. Since, the underlying latent response is assumed to be continuous, one can also consider our approach for the purpose of modeling binary or count response variable with minor modification. In this work, our proposed method deals with missingness in the response variable only. Sometimes, the data sets from longitudinal studies may possess missing information not only in response but also in the covariates. It may be

an interesting problem to deal with such data and develop models and corresponding estimation methodology.

**Acknowledgement**

The author is thankful to Prof. Murari Mitra and Mr. Jayabrata Biswas for many helpful comments and suggestions.

# References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Baker, S. G. (1995). Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics*, 51:1042–1052.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. journal of the american statistical association. *Journal of the American Statistical Association*, 88:9–25.

Cui, Y. and Zhu, J. (2006). Functional mapping for genetic control of programmed cell death. *Physiological Genomics*, 25(3):458–69.

Daniels, M. J. and Hogan, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout authors. *Biometrics*, 56(4):12411248.

De, V. G. and Tu, X. M. (1994). Modeling progression of cd4+ lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014.

Diggle, P. J., Heagerty, P. J., Liang, K., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Inc.

Diggle, P. J. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43:49–93.

Have, T. R. T., Kunselman, A. R., Pulkstenis, E. P., and Landis, J. R. (1998). Mixed effects logistic regression models for longitudinal binary response data with informative dropout. *Biometrics*, 54:367–383.

Lin, H., Liub, D., and Zhou, X. (2010). A correlated random-effects model for normal longitudinal data with nonignorable missingness. *Statistics in Medicine*, 29:236–247.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121.

Little, R. J. A. and Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.

Marie, H. and Sen, P. K. (1985). On sequentially adaptive asymptotically efficient rank statistics. *Sequential Analysis*, 4(3):125–151.

Meyer, K. (2000). Random regressions to model phenotypic variation in monthly weights of australian beef cows. *Livestock Production Science*, 65:19–38.

Minini, P. and Chavance, M. (2004). Sensitivity analysis of longitudinal binary data with non-monotone missing values. *Biometrics*, 5:531–544.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4:12–35.

Pulkstenis, E. P., Have, T. R. T., and Landis, J. R. (1998). Model for the analysis of binary longitudinal pain data subject to informative dropout through remedication. *Journal of the American Statistical Association*, 93:438–450.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93:321–1339.

Schluchter, M. D. (1994). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, 11:1861–1870.

Sentrk, D., Dalrymple, L. S., Mohammed, S. M., Kaysen, G. A., and Nguyen, D. V. (2013). Modeling time-varying effects with generalized and unsynchronized longitudinal data. *Statistics in Medicine*, 32:29712987.

Siddiqui, O. and Ali, M. W. (1998). A comparison of the random-effects pattern mixture model with last-observation-carried-forward(locf) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, 8(4):61–106.

Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 74:425–438.

Tsonaka, R., Verbeke, G., and Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Roula Tsonaka, Geert Verbeke and Emmanuel Lesaffre*, 65(1):81–87.

West, B., Welch, K., and Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall, CRC Press.

Wu, M. C. and Carrol, R. J. (1998). Estimation and comparison of changes in the presence of informative censoring by modeling the censoring process. *Biometrics*, 44:175–188.