*Journal of Statistical Research* *https://doi.org/10.3329/jsr.v56i2.67468*
*2022, Vol. 56, No. 2, pp. 155-183* *ISSN 0256 - 422 X*

# APPROXIMATE METHODS FOR ANALYZING SEMIPARAMETRIC LONGITUDINAL MODELS WITH NONIGNORABLE MISSING RESPONSES

NAJLA ALORAINI
*School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada*
*Department of Mathematics, Collage of Sciences, Qassim University, Qassim, KSA*
*Email: arienie@qu.edu.sa*

SANJOY SINHA
*School of Mathematics and Statistics, Carleton University, Ottawa, ON, Canada*
*Email: sinha@math.carleton.ca*

SUMMARY

We often encounter missing data in longitudinal studies. When the missingness in longitudinal data is nonignorable, it is necessary to incorporate the missing data mechanism into the observed data likelihood function for a valid statistical inference. In this article, we propose and explore a novel semiparametric approach to estimating the regression parameters and variance components using a partially linear mixed model with nonignorable and nonmonotone missing responses. The finite sample properties of the proposed method are studied using Monte Carlo simulations, where our method is found to be very effective in capturing any curvilinear pattern in the mean response. The method is also illustrated using some actual longitudinal data obtained from a public health survey, referred to as the Health and Retirement Study (HRS).

*Keywords and phrases:* EM algorithm, Gibbs sampling, Mote Carlo EM, Regression spline, Semiparametric method, Nonignorable missingness.

*AMS Classification:* MSC 2000: Primary 62F10; secondary 62F35

## 1 Introduction

During the past few decades, missing data problems have been studied extensively in the literature, with a focus on the case when data are missing at random (MAR), i.e., when the missingness depends only on the observed values, but not on the values that are missing (Little and Rubin, 2002). By contrast, research into nonignorable missing data or data that are not missing at random (NMAR) is quite limited. The main problem with analyzing nonignorable missing data is that often it is difficult to specify a parametric model for the missing data that can be incorporated into the observed data likelihood function. Also, the likelihood inference with nonignorable missingness often requires intensive computation involving multi-dimensional integration.

⋆ Corresponding author

Missing data are common in longitudinal studies because of study dropouts, mistimed measurements, subjects being too sick to come to the clinic to be measured, etc. A subject's response can be missing at one follow-up time and then be measured at the next follow-up time, resulting in an arbitrary and nonmonotone missingness pattern. Often missing data are nonignorable in the sense that the missingness depends on the missing values themselves. For example, side effects of a treatment may make a patient's health worse and thereby affect the patient participation (Ibrahim et al., 2001). For nonignorable missing data, methods developed under the MAR assumption may result in estimation bias and incorrect inference (Zhao and Shao, 2015). Methods for analyzing nonignorable missing data have been studied by a number of authors in recent years. Ibrahim et al. (2001) proposed an approximate Monte Carlo EM method for estimating parameters in generalized linear mixed models with nonignorable missing data.

A parameteric linear mixed model may not be adequate for analyzing longitudinal data when the response variable and associated covariates tend to have a complex nonlinear relationship (Grace et al., 2009). In this regard, semiparametric partially linear mixed models for longitudinal data have become increasingly popular. Recently, there has been much research into nonparametric and semiparametric regression methods for clustered longitudinal data. For more details, we refer to Grace et al. (2009), Wang et al. (2005), Fan and Li (2004), Carroll et al. (1997), Heckman (1986), Ruppert and Carroll (2009), Parise et al. (2001), Coull et al. (2001a), Coull et al. (2001b), and Opsomer et al. (2008), among others.

Also, an attention has been paid to semiparametric approaches to analyzing incomplete longitudinal data with nonignorable missing responses, where the missing data model is considered parametric and the response model semiparametric. Robins et al. (1994, 1995) discussed semiparametric methods for nonignorable missing responses based on weighted estimating equations for repeated outcomes. Many researchers considered a parametric model for the missing data and a nonparametric model for the response variable (Qin et al., 2002; Chang and Kott, 2008; Kott and Chang, 2010; Morikawa and Kim, 2016; Morikawa et al., 2017; Ai et al., 2018).

This research is motivated by a longitudinal public health survey from the Health and Retirement Study (HRS, 2019) conducted by the Institute for Social Research at the University of Michigan. We consider analyzing a subset of the RAND HRS data (RAND, 2019) concerning the physical health of aged people, which includes the most recent surveys for the years 2010, 2012, 2014 and 2016. The HRS data feature nonmonotone missingness in the response variable BMI (body mass index). The missing values in BMI is considered nonignorable (NMAR) in the sense that the missingness probability may depend on the missing values themselves. Also, the relationship between the response BMI and covariate Age is considered nonlinear, as evident from an initial analysis. To model the nonlinear relationship, we investigate a flexible semiparametric approach in the framework of the partially linear mixed model with nonignorable and nonmonotone missing responses. To reduce the computational burden involving high-dimensional integration, we adopt an approximate Monte Carlo EM approach following Ibrahim et al. (2001).

The paper is organized as follows. Section 2 introduces the model and notation to describe a partially linear mixed model for the mean response and a logistic regression model for the missing data. Section 3 introduces the proposed MCEM method for estimating the model parameters. The asymptotic properties of the MCEM estimators are also discussed in Section 3. Section 4 presents results from a simulation study, which was carried out to investigate the finite-sample properties of the proposed MCEM estimators. Section 5 provides an application of the proposed method using the longitudinal survey data from the Health and Retirement Study (HRS) introduced earlier. Section 6 provides conclusions of the paper with some directions for future research.

# 2 Model and Notation

## 2.1 Partially linear mixed effect model

Consider first a simple response model with a single covariate $x$, given by

$$y = m_0(x) + \varepsilon, \tag{2.1}$$

where the random error $\varepsilon$ has the mean 0 and variance $\sigma^2_\varepsilon$, and $m_0(\cdot)$ is an unknown smooth function representing the mean response of $y$ given the covariate $x$. We assume that $m_0(\cdot)$ can be approximated sufficiently well by a penalized spline (P-spline) regression function, given by

$$m(x, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \mathbf{c}^t \boldsymbol{\nu} = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \delta_k)^p_+, \tag{2.2}$$

for $k = 1, \ldots, K$, where $\mathbf{c} = \{1, x, \ldots, x^p, (x-\delta_1)^p_+, \ldots, (x-\delta_K)^p_+\}^t$ is the truncated polynomial basis consisting of piecewise continuous $p$-th degree polynomials with a fixed set of knots $\delta_1 < \ldots < \delta_K$ and $\boldsymbol{\nu}^t = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)$, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^t$ denoting the coefficient vectors corresponding to the parametric and spline portions of the regression function, respectively. Here $p \geq 1$ is the degree of the spline. When $p = 1, 2$, or 3, the truncated power basis is called the "linear", "quadratic" or "cubic", respectively. We denote

$$(x - \delta_k)_+ = \begin{cases} (x - \delta_k) & \text{if} \quad x > \delta_k \\ 0 & \text{if} \quad x \leq \delta_k, \end{cases}$$

for $k = 1, \ldots, K$. The knots $\delta_1 < \ldots < \delta_K$ are typically chosen as the quantiles of the distribution of the covariate $x$ (Opsomer et al., 2008).

The choice of the number of knots $K$ has been a subject of extensive research in spline regression. Too many knots may lead to overfitting the data, whereas too few knots may lead to underfitting (Eilers and Marx, 1996). Also, there is a trade-off between the number of knots and bias-variance of the regression estimator. A large number of knots may lead to an estimate with a small bias and a large variance. On the other hand, a small number of knots may lead to an estimate with a large bias and a small variance. The reduced-knot specification may offer considerable computational advantages (Harezlak et al., 2005). More knots should be put in places where there is more curvature in the response function (Minggao and Taylor, 1996). In practice, one may consider several sets of knots and choose the one that gives the best result. Often a spline model with a number of knots between 5 and 10 give satisfactory results (Minggao and Taylor, 1996).

Furthermore, the class of regression functions $m(\cdot, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is very large and can approximate most smooth functions $m_0(\cdot)$ with a high degree of accuracy (Opsomer et al., 2008). As is commonly done in the context of P-spline regression, it is assumed that the lack-of-fit error $m_0(\cdot) - m(\cdot, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is negligible relative to the estimation error $m(\cdot, \boldsymbol{\beta}, \boldsymbol{\gamma}) - m(\cdot, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ (Opsomer et al., 2008). Ruppert (2002) provided a simulation-based evidence that this lack-of-fit error is indeed negligible in the case of univariate semi-parametric regression.

In the next step, consider a partially linear mixed model (PLMM) for longitudinal data, given by

$$y_{ij} = \mathbf{w}_{ij}^t \boldsymbol{\alpha} + m_0(x_{ij}) + \mathbf{z}_{ij}^t \mathbf{u}_i + \varepsilon_{ij}, \tag{2.3}$$

where subscript $i$ represents the $i$th subject ($i = 1, \cdots, m$) and $j$ represents observations from a given subject. We assume that each subject is measured at a fixed number of time-points $j = 1, \cdots, n$. Also in Eq. (2.3), $\mathbf{w}_{ij}$ is a vector of covaiates that are observed along with the response $y_{ij}$ and $\boldsymbol{\alpha}$ is a vector of regression coefficients. In this sitting, the fixed effects of $\mathbf{w}_{ij}$ are modeled parametrically, while the effect of $x_{ij}$ is modeled nonparametrically. We consider approximating model (2.3) by the P-spline model

$$y_{ij} = \mathbf{w}_{ij}^t \boldsymbol{\alpha} + m(x_{ij}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \mathbf{z}_{ij}^t \mathbf{u}_i + \varepsilon_{ij}, \tag{2.4}$$

where

$$m(x_{ij}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x_{ij} + \ldots + \beta_p x_{ij}^p + \sum_{k=1}^{K} \gamma_k (x_{ij} - \delta_k)_+^p = \mathbf{x}_{ij}^t \boldsymbol{\beta} + \mathbf{d}_{ij}^t \boldsymbol{\gamma},$$

with $\mathbf{x}_{ij} = (1, x_{ij}, \ldots, x_{ij}^p)^t$ and $\mathbf{d}_{ij} = \{(x_{ij} - \delta_1)_+^p, \ldots, (x_{ij} - \delta_K)_+^p\}^t$.

Model (2.4) can be written in the matrix form

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\alpha} + \mathbf{X}_i \boldsymbol{\beta} + \mathbf{D}_i \boldsymbol{\gamma} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i, \tag{2.5}$$

for $i = 1, \ldots, m$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})^t$ represents the response vector from the $i$th subject, and $\mathbf{W}_i$, $\mathbf{X}_i$, $\mathbf{D}_i$, and $\mathbf{Z}_i$ are corresponding design matrices with their $j$th ($j = 1, \ldots, n$) rows $\mathbf{w}_{ij}^t$, $\mathbf{x}_{ij}^t$, $\mathbf{d}_{ij}^t$, and $\mathbf{z}_{ij}^t$, respectively. We assume that the random cluster effects $\mathbf{u}_i$ and random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})^t$ are mutually independent and follow normal distributions, given by

$$\mathbf{u}_i \sim N_q(\mathbf{0}, \mathbf{V}_u) \qquad \text{and} \qquad \boldsymbol{\varepsilon}_i \sim N_n\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n\right), \tag{2.6}$$

where $N_q(\mathbf{0}, \mathbf{V}_u)$ denotes the $q$-dimensional multivariate normal distribution with the mean vector $\mathbf{0}$ and a covariance matrix $\mathbf{V}_u$, and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

Model (2.5) can be rewritten as

$$\mathbf{y}_i = \mathbf{X}_i^* \boldsymbol{\beta}^* + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i, \tag{2.7}$$

where $\mathbf{X}_i^* = (\mathbf{W}_i^t, \mathbf{X}_i^t, \mathbf{D}_i^t)^t$ denotes a combined design matrix and $\boldsymbol{\beta}^* = (\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)^t$ is a combined regression parameter vector. From (2.7), the marginal mean and variance of $\mathbf{y}_i$ are obtained as

$$E(\mathbf{y}_i) = \mathbf{X}_i^* \beta^* \text{ and } V(\mathbf{y}_i) = \mathbf{V}_{y_i} = \mathbf{Z}_i \mathbf{V}_u \mathbf{Z}_i^t + \sigma_\varepsilon^2 \mathbf{I}_n.$$

Under the normality of the random effects $\mathbf{u}_i$ and random errors $\boldsymbol{\varepsilon}_i$, the conditional distribution of $\mathbf{y}_i$ given $\mathbf{u}_i$ is obtained as

$$\mathbf{y}_i|\mathbf{u}_i \sim N_n(\mathbf{X}_i^*\boldsymbol{\beta}^* + \mathbf{Z}_i\mathbf{u}_i, \sigma_\varepsilon^2\mathbf{I}_n). \tag{2.8}$$

The corresponding marginal distribution of $\mathbf{y}_i$ is obtained as

$$\mathbf{y}_i \sim N_n(\mathbf{X}_i^*\boldsymbol{\beta}^*, \mathbf{Z}_i\mathbf{V}_u\mathbf{Z}_i^t + \sigma_\varepsilon^2\mathbf{I}_n). \tag{2.9}$$

For a complete set of longitudinal data, given the variance components $\mathbf{V}_u$ and $\sigma_\varepsilon^2$, the regression parameters $\boldsymbol{\beta}^*$ may be estimated using the method of generalized least squares (GLS). The variance components may be estimated using the method of maximum likelihood (ML) or the restricted maximum likelihood (REML). For incomplete longitudinal data, however, the estimation may require maximizing an adjusted log-likelihood that incorporates a missing data model into the observed data likelihood function, as described in the next section.

## 2.2 Models for missing data

The missing data model is defined as the distribution of the binary missingness indicators $\mathbf{r}_i = (r_{i1}, \ldots, r_{in})^t$, with

$$r_{ij} = \begin{cases} 1 & \text{if} \quad y_{ij} \text{ is observed,} \\ 0 & \text{if} \quad y_{ij} \text{ is missing,} \end{cases}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The conditional distribution of $\mathbf{r}_i|\mathbf{y}_i$ is a multinomial distribution with $2^n$ cell probabilities. To model the missing data indicators, Diggle and Kenward (1994) proposed a binomial model whose probability mass function is given by

$$f(\mathbf{r}_i|\mathbf{y}_i, \boldsymbol{\phi}) = \prod_{j=1}^{n} \{p(r_{ij} = 1|\boldsymbol{\phi})\}^{r_{ij}} \{1 - p(r_{ij} = 1|\boldsymbol{\phi})\}^{1-r_{ij}}, \tag{2.10}$$

where $p(r_{ij} = 1|\boldsymbol{\phi})$ is modelled by a logistic regression including all previous outcomes and the current outcome. This model is defined by

$$\text{logit}\{p(r_{ij} = 1|\boldsymbol{\phi})\} \equiv \log\left\{\frac{p(r_{ij} = 1|\boldsymbol{\phi})}{1 - p(r_{ij} = 1|\boldsymbol{\phi})}\right\} = \phi_0 + \phi_1 y_{ij} + \sum_{k=2}^{j} \phi_k y_{j+1-k},$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Here the missing data indicators $r_{ij}$'s for a given subject are assumed independent. A more general multinomial missing data model, which includes a general correlation structure, can be constructed by identifying the joint distribution of $\mathbf{r}_i = (r_{i1}, \ldots, r_{in})^t$ through a sequence of one-dimensional conditional distributions, as suggested by Ibrahim et al. (2001), i.e.,

$$p(\mathbf{r}_i|\mathbf{y}_i, \boldsymbol{\phi}) = p(r_{i1}|\mathbf{y}_i, \boldsymbol{\phi}_1)\, p(r_{i2}|r_{i1}, \mathbf{y}_i, \boldsymbol{\phi}_2) \ldots p(r_{in}|r_{i1}, \ldots, r_{i,n-1}, \mathbf{y}_i, \boldsymbol{\phi}_n), \tag{2.11}$$

where $\boldsymbol{\phi}_j$ is a vector of indexing parameters for the $j$th conditional distribution and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^t, \ldots, \boldsymbol{\phi}_n^t)^t$ represents the vector of all missing data parameters. As each $r_{ij}$ is a binary random variable, we can use a sequence of binary logistic regression models for three conditional densities in (2.11). This modeling approach can potentially reduce the number of nuisance parameters that have to be specified for the missing data mechanism. In addition, it yields general correlation structures between the $r_{ij}$'s, and allows more flexibility in the specification of the missing data model. Finally, it provides a natural way to specify the joint distribution of the missing data indicators when knowledge about the missingness of one response affects the probability of missingness of another (Ibrahim and Molenberghs, 2009).

# 3 Estimation under missing data

We extend the Monte Carlo EM (MCEM) approach of Ibrahim et al. (2001) for estimating the parameters of the partially linear mixed effects model (2.3) with nonignorable and nonmonotone missing responses. The MCEM approach approximates the conditional expectations involving the E-step of the EM algorithm.

## 3.1 Complete data log-likelihood

Treating $\{(\mathbf{y}_i, \mathbf{r}_i, \mathbf{u}_i), i = 1, \ldots, m\}$ as "complete data", the complete data log-likelihood is obtained as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \ell(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{u}_i, \mathbf{r}_i) \equiv \sum_{i=1}^{m} \Big\{ \log f(\mathbf{y}_i|\boldsymbol{\beta}^*, \sigma_\varepsilon^2, \mathbf{u}_i) + \log f(\mathbf{u}_i|\mathbf{V}_u) + \log f(\mathbf{r}_i|\boldsymbol{\phi}, \mathbf{y}_i) \Big\}, \qquad (3.1)$$

where $f\left(\mathbf{y}_i|\boldsymbol{\beta}^*, \sigma_\varepsilon^2, \mathbf{u}_i\right)$ is the conditional density of $\mathbf{y}_i|\mathbf{u}_i$ as defined in (2.8), $f(\mathbf{u}_i|\mathbf{V}_u)$ is the density of the random effects $\mathbf{u}_i$, and $f\left(\mathbf{r}_i|\boldsymbol{\phi}, \mathbf{y}_i\right)$ is the density of the missing data indicators $\mathbf{r}_i$. The vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^*, \sigma_\varepsilon^2, \mathbf{V}_u, \boldsymbol{\phi}\right)$ represents all model parameters, where our focus is on the estimation of the regression parameters $\boldsymbol{\beta}^*$ and variance components $(\sigma_\varepsilon^2, \mathbf{V}_u)$, with $\boldsymbol{\phi}$ being considered as a vector of nuisance parameters.

## 3.2 E-step

The E-step of the EM algorithm consists of computing the expected value of the complete data log-likelihood given the observed data $(\mathbf{y}_{\text{obs},i}, \mathbf{r}_i)$ and current parameter estimates $\boldsymbol{\theta}^{(l)}$. Assuming that the patterns of missing data in $\mathbf{y}_i$ are arbitrary and nonmonotone, some permutation of the indices of $\mathbf{y}_i$ can be written as $\mathbf{y}_i = (\mathbf{y}_{\text{mis},i}, \mathbf{y}_{\text{obs},i})$, where $\mathbf{y}_{\text{obs},i}$ denotes the observed values and $\mathbf{y}_{\text{mis},i}$ the missing values of the response vector $\mathbf{y}_i$. Since both $\mathbf{u}_i$ and $\mathbf{y}_{\text{mis},i}$ are unobserved, they are integrated out in the E-step. For the $i$th observation and at the $(l+1)$st iteration, the E-step calculates

$$\begin{aligned} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) &= E\left\{\ell(\boldsymbol{\theta};\mathbf{y}_i,\mathbf{u}_i,\mathbf{r}_i)|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right\} \\ &= \int\int\Big\{\log f(\mathbf{y}_i|\boldsymbol{\beta}^*,\mathbf{u}_i)+\log f(\mathbf{u}_i|\mathbf{V}_u)+\log f(\mathbf{r}_i|\mathbf{y}_i,\boldsymbol{\phi})\Big\} \\ &\quad\times f\left(\mathbf{y}_{\text{mis},i},\mathbf{u}_i|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)d\mathbf{u}_i\ d\mathbf{y}_{\text{mis},i} \\ &\equiv \mathbf{I}_1+\mathbf{I}_2+\mathbf{I}_3, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\theta}^{(l)} = \left(\boldsymbol{\beta}^{*(l)},\sigma_\varepsilon^{2(l)},\mathbf{V}_u^{(l)},\boldsymbol{\phi}^{(l)}\right)$, and $f\left(\mathbf{y}_{\text{mis},i},\mathbf{u}_i|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)$ represents the conditional distribution of the "missing data" $(\mathbf{y}_{\text{mis},i},\mathbf{u}_i)$, given the observed data $(\mathbf{y}_{\text{obs},i},\mathbf{r}_i)$. Note that Eq. (3.2) involves integration with respect to the random effects $\mathbf{u}_i$, which can be found analytically. For this, we can write the conditional density $f\left(\mathbf{y}_{\text{mis},i},\mathbf{u}_i|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)$ as

$$f\left(\mathbf{y}_{\text{mis},i},\mathbf{u}_i|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right) = f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)f\left(\mathbf{u}_i|\mathbf{y}_i,\boldsymbol{\theta}^{(l)}\right),$$

where the conditional distribution of $\mathbf{u}_i$, given $\mathbf{y}_i$ and current estimator $\boldsymbol{\theta}^{(l)}$, may be obtained as

$$\mathbf{u}_i|(\mathbf{y}_i,\boldsymbol{\theta}^{(l)}) \sim N\left(\mathbf{u}_i^{(l)},\boldsymbol{\Sigma}_i^{(l)}\right),$$

with

$$\mathbf{u}_i^{(l)} = \boldsymbol{\Sigma}_i^{(l)}\boldsymbol{Z}_i^t\left(\mathbf{y}_i-\mathbf{X}_i^*\boldsymbol{\beta}^{*(l)}\right)/\sigma_\varepsilon^{2(l)}, \quad \text{and} \quad \boldsymbol{\Sigma}_i^{(l)} = \left(\sigma_\varepsilon^{-2(l)}\mathbf{Z}_i^t\mathbf{Z}_i+(\mathbf{V}_u^{(l)})^{-1}\right)^{-1}.$$

Then after some algebra, the term $\mathbf{I}_1$ in Eq. (3.2) may be obtained as

$$\begin{aligned} \mathbf{I}_1 = &-\frac{n}{2}\log(\sigma_\varepsilon^2)-\frac{1}{2\sigma_\varepsilon^2}\left\{\text{tr}\left(\mathbf{Z}_i^t\mathbf{Z}_i\boldsymbol{\Sigma}_i^{(l)}\right)+\int\left(\mathbf{y}_i-\mathbf{X}_i^*\boldsymbol{\beta}^*-\mathbf{Z}_i\mathbf{u}_i^{(l)}\right)^t\left(\mathbf{y}_i-\mathbf{X}_i^*\boldsymbol{\beta}^*-\mathbf{Z}_i\mathbf{u}_i^{(l)}\right)\right. \\ &\left.\times f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)d\mathbf{y}_{\text{mis},i}\right\}. \end{aligned} \tag{3.3}$$

The term $\mathbf{I}_2$ in (3.2) may be obtained as

$$\mathbf{I}_2 = -\frac{1}{2}\log(|\mathbf{V}_u|)-\frac{1}{2}\text{tr}(\mathbf{V}_u^{-1}\boldsymbol{\Sigma}_i^{(l)})-\frac{1}{2}\int\left(\mathbf{u}_i^{(l)t}\mathbf{V}_u^{-1}\mathbf{u}_i^{(l)}\right)f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)d\mathbf{y}_{\text{mis},i}. \tag{3.4}$$

Also, the term $\mathbf{I}_3$ in (3.2) leads to

$$\mathbf{I}_3 = \int\log f\left(\mathbf{r}_i|\mathbf{y}_i,\boldsymbol{\phi}\right)f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)d\mathbf{y}_{\text{mis},i}. \tag{3.5}$$

Note that the terms $\mathbf{I}_1$–$\mathbf{I}_3$ in (3.2) still require integration with respect to the conditional density $f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)$. If the dimension of the vector of missing data $\mathbf{y}_{\text{mis},i}$ is large, it may be difficult to perform the integration numerically. For this reason, we adopt a Monte Carlo approach that can approximate the conditional expectations by generating random draws from the conditional distribution of $\mathbf{y}_{\text{mis},i}|(\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)})$, as described in the next section.

### 3.3 Monte Carlo EM algorithm

To approximate the expectations in the E-step of the EM algorithm, we draw random samples from the conditional density $f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right)$, which can be written up to a constant of proportionality as

$$\begin{aligned} f\left(\mathbf{y}_{\text{mis},i}|\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right) &= f\left(\mathbf{y}_i|\boldsymbol{\beta}^{*(l)},\sigma_\varepsilon^{2(l)},\mathbf{V}_u^{(l)}\right) f\left(\mathbf{r}_i|\mathbf{y}_i,\boldsymbol{\phi}^{(l)}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mathbf{y}_i-\mathbf{X}_i^*\boldsymbol{\beta}^{*(l)}\right)^t\left(\mathbf{Z}_i\mathbf{V}_u^{(l)}\mathbf{Z}_i^t+\sigma_\varepsilon^{2(l)}\mathbf{I}_n\right)^{-1}\left(\mathbf{y}_i-\mathbf{X}_i^*\boldsymbol{\beta}^{*(l)}\right)\right) \\ &\quad\times f\left(\mathbf{r}_i|\mathbf{y}_i,\boldsymbol{\phi}^{(l)}\right), \end{aligned} \tag{3.6}$$

where the normal density for $\mathbf{y}_i$ and the logistic regression for $\mathbf{r}_i$ are log-concave in the components of $\mathbf{y}_i$. So the sampling from (3.6) can be done using the adaptive rejection sampling algorithm of Gilks and Wild (1992), based on the conditional density

$$f\left(\mathbf{y}_{\text{mis},ih},|\mathbf{y}_{\text{mis},it},t\neq h,\mathbf{y}_{\text{obs},i},\mathbf{r}_i,\boldsymbol{\theta}^{(l)}\right), \tag{3.7}$$

where $\mathbf{y}_{\text{mis},ih}$ denotes the $h$-th component of $\mathbf{y}_{\text{mis},i}$. We use the R function "arms" to generate samples from the conditional density (3.6), which uses the Gibbs sampler along with the adaptive rejection sampling algorithm.

Let $(\mathbf{y}_{i1}^*,\ldots,\mathbf{y}_{im_i}^*)$ be a random sample of size $m_i$ drawn from the conditional distribution (3.6). Define

$$\mathbf{y}_i^{(k)}=\left(\mathbf{y}_{ik}^*,\mathbf{y}_{\text{obs},i}\right)^t, \quad \text{and} \quad u_i^{(lk)}=\boldsymbol{\Sigma}_i^{(l)}\mathbf{Z}_i^t\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^*\boldsymbol{\beta}^{*(l)}\right)/\sigma_\varepsilon^{2^{(l)}},$$

for $k=1,\ldots,m_i$. Then the Monte Carlo E-step for the $i$th subject at the $(l+1)$st EM iteration approximates the objective function (3.2) by

$$\begin{aligned} Q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right) &= -\frac{n}{2}\log(\sigma_\varepsilon^2)-\frac{1}{2\sigma_\varepsilon^2}\Bigg\{\text{tr}\left(\mathbf{Z}_i^t\mathbf{Z}_i\boldsymbol{\Sigma}_i^{(l)}\right) \\ &+\frac{1}{m_i}\sum_{k=1}^{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^*\boldsymbol{\beta}^*-\mathbf{Z}_i\mathbf{u}_i^{(lk)}\right)^t\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^*\boldsymbol{\beta}^*-\mathbf{Z}_i\mathbf{u}_i^{(lk)}\right)\Bigg\} \\ &-\frac{1}{2}\log(|\mathbf{V}_u|)-\frac{1}{2}\text{tr}(\mathbf{V}_u^{-1}\boldsymbol{\Sigma}_i^{(l)})-\frac{1}{2}\frac{1}{m_i}\sum_{k=1}^{m_i}\left(\mathbf{u}_i^{(l)t}\mathbf{V}_u^{-1}\mathbf{u}_i^{(l)}\right) \\ &+\frac{1}{m_i}\sum_{k=1}^{m_i}\log f\left(\mathbf{r}_i|\mathbf{y}_i^{(k)},\boldsymbol{\phi}\right). \end{aligned} \tag{3.8}$$

Here the approximate E-step takes a complete data weighted form in which $\mathbf{y}_{\text{mis},i}$ is replaced by a set of $m_i$ values, each contributing a weight of $1/m_i$. By the law of large numbers, the "estimator" (3.8) converges to the theoretical expectation (3.2) (Levine and Casella, 2001). The values of $m_i$'s may be equal, $m_i=m^*$, for all $i$ and also for each EM iteration. However, if the $m_i$'s are chosen to be different, this may speed up the convergence in EM iterations, as discussed in Booth and Hobert (1999).

The M-step of the EM algorithm consists of maximizing the objective function for all $m$ subjects, defined by

$$\mathbf{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right) = \sum_{i=1}^{m} Q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right).$$

We can use the iterative Newton-Raphson method for numerically estimating the model parameters. In particular, for estimating the regression parameters $\boldsymbol{\beta}^*$, the Newton-Raphson method leads to the iterative equations

$$\boldsymbol{\beta}^{*(l+1)} = \left(\sum_{i=1}^{m} \mathbf{X}_i^{*t}\mathbf{X}_i^{*}\right)^{-1} \sum_{i=1}^{m}\left(\mathbf{X}_i^{*t}\frac{1}{m_i}\sum_{k=1}^{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{Z}_i u_i^{(lk)}\right)\right). \tag{3.9}$$

Details about the computational algorithm for estimating all model parameters are discussed in Section 4.3.

## 3.4 Asymptotic variance

The asymptotic variance of the estimators $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_\varepsilon^2, \hat{\mathbf{V}}_u, \hat{\boldsymbol{\phi}}\right)$ may be obtained from the inverse of the observed Fisher information matrix, as $V(\hat{\boldsymbol{\theta}}) = [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}$, where the Fisher information matrix $\mathcal{I}(\hat{\boldsymbol{\theta}})$ may be calculated from the observed data log-likelihood following Louis (1982); see also McCulloch et al. (2008) and McLachlan and Krishnan (2008) for details. We have

$$\begin{aligned}\frac{\partial^2 \ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t} &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t} E\left(\log f_{y,u,r}(\mathbf{y},\mathbf{u},\mathbf{r})|\mathbf{y}_{\text{obs}},\mathbf{r}\right)\\ &+ E\left(S(\boldsymbol{\theta};\mathbf{y},\mathbf{u},\mathbf{r})S\left(\boldsymbol{\theta};\mathbf{y},\mathbf{u},\mathbf{r}\right)^t\middle|\mathbf{y}_{\text{obs}},\mathbf{r}\right)\\ &- \frac{\partial}{\partial\boldsymbol{\theta}} E\left(\log f_{y,u,r}(\mathbf{y},\mathbf{u},\mathbf{r})|\mathbf{y}_{\text{obs}},\mathbf{r}\right)\frac{\partial}{\partial\boldsymbol{\theta}} E\left(\log f_{y,u,r}(\mathbf{y},\mathbf{u},\mathbf{r})|\mathbf{y}_{\text{obs}},\mathbf{r}\right)^t,\end{aligned} \tag{3.10}$$

where $S(\boldsymbol{\theta};\mathbf{y},\mathbf{u},\mathbf{r}) = \partial\log f_{y,u,r}(\mathbf{y},\mathbf{u},\mathbf{r})/\partial\boldsymbol{\theta}$ denotes the score vector for the complete data.

From (3.10) and by using the $\mathbf{Q}$ function in the E-step, we obtain

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - E\left(S(\hat{\boldsymbol{\theta}};\mathbf{y},\mathbf{u},\mathbf{r})S(\hat{\boldsymbol{\theta}};\mathbf{y},\mathbf{u},\mathbf{r})^t\middle|\mathbf{y}_{\text{obs}},\mathbf{r}\right) + \frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}). \tag{3.11}$$

By the Monte Carlo approach, the observed Fisher information (3.11) can be approximated as

$$\begin{aligned}\mathcal{I}(\hat{\boldsymbol{\theta}}) \approx &-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\mathbf{Q}(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}) - \left\{\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}S_i(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i)S_i(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i)^t\right\}\\ &+\sum_{i=1}^{m}\frac{\partial}{\partial\boldsymbol{\theta}}Q_i(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})\frac{\partial}{\partial\boldsymbol{\theta}}Q_i(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}})^t,\end{aligned} \tag{3.12}$$

where $\hat{\boldsymbol{\theta}}$ is the MCEM estimate of $\boldsymbol{\theta}$.

# 4 Simulation Study

To study the empirical properties of our proposed semiparametric MCEM method (Method 1), we ran two sets of simulations with two different sample sizes ($m = 100$ and $200$ subjects; $n = 2$ observations per subject). In the first set, the estimators were studied for the case when the true model is linear, while in the second set these were studied when the true model is nonlinear. We also compared our proposed estimators to those obtained by the ordinary MCEM method (Method 2) of Ibrahim et al. (2001), where the mean response function was considered linear. Each simulation run was based on 1000 replications of data sets.

## 4.1 Response model for simulations

The data were generated from the semiparametric mixed model

$$y_{ij} = \alpha_1 \text{trt}_{A_i} + m_0(x_{ij}) + u_i + \varepsilon_{ij},\ i = 1, \dots, m,\ j = 1, 2, \tag{4.1}$$

where the covariate $\text{trt}_{A_i}$ is a group indicator of treatment A, which takes the value 0 for the first 50% subjects under the control group and 1 for the remaining subjects under the treatment group. The values of the covariate $x_{ij}$ were generated from the normal $N(1, 1)$. We used two different forms of the mean response $m_0(x)$: a) $m_0(x) = 1 + x$ (linear) and b) $m_0(x) = 1 + x + x^2$ (quadratic). The intercept random effects $u_i$ were generated from $N(0, \sigma_u^2)$ and the random errors $\varepsilon_{ij}$ were generated from $N\left(0, \sigma_\varepsilon^2\right)$, with $\sigma_u^2 = \sigma_\varepsilon^2 = 1$.

For the semiparametric method, we used the linear P-spline approximation with $p = 1$. The number of knots $K$ and their locations $(\delta_1, \dots, \delta_K)$ were chosen as per the recommendations of Ruppert and Carroll (2009) and Opsomer et al. (2008). We chose $K = 5$, with the knots being placed at the empirical quantiles of the $x$ distribution that gave roughly equal number of $x$-values between the knots.

In our numerical studies, we focused on estimating the variance parameters $\sigma_u^2$ and $\sigma_\varepsilon^2$, and the mean response at five different $x$-values, $m_0(x_1), \dots, m_0(x_5)$, with $x_1, \dots, x_5$ being chosen as normal quantiles with probabilities $0.1, 0.3, 0.5, 0.7, 0.9$, respectively. Our proposed semiparametric approach estimates the spline regression parameters $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^t, \boldsymbol{\gamma}^t)^t$ that leads to an approximation $\hat{m}(x) = \mathbf{x}_{ij}^t \hat{\boldsymbol{\beta}} + \mathbf{d}_{ij}^t \hat{\boldsymbol{\gamma}}$ to the true mean response $m_0(x)$.

## 4.2 Missing data model

We considered a nonignorable missing data model that was functionally dependent on the current and previous values of the response variable. For computational simplicity, without loss of generality, we considered the outcome at the first time point always observed. The outcome at the second time point may be missing with the missingness probability $p_{i2} = P(r_{i2} = 1|\boldsymbol{\phi})$ that follows a binary logistic regression model, given by

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = \phi_0 + \phi_1 y_{i1} + \phi_2 y_{i2} + \phi_3 x_{i2} = \boldsymbol{\omega}_{i2}^t \boldsymbol{\phi}, \tag{4.2}$$

where $\boldsymbol{\omega}_{i2} = (1,\ y_{i1},\ y_{i2},\ x_{i2})^t$ and $\boldsymbol{\phi} = (\phi_0,\ \phi_1,\ \phi_2,\ \phi_3)^t$. In this setting, the joint density of the missing data indicators is given by

$$p(\mathbf{r}|\mathbf{y},\boldsymbol{\phi}) = \prod_{i=1}^{m} p_{i2}^{r_{i2}} \left(1 - p_{i2}\right)^{1-r_{i2}} .$$

Note that a non-zero value of $\phi_2$ in model (4.2) would lead to missing data that are not missing at random (**NMAR**). We chose $\boldsymbol{\phi} = (\phi_0,\ \phi_1,\ \phi_2,\ \phi_3)^t = (-2.5, 0.2, 0.3, 1)^t$, which resulted in roughly $23\%$ missing values in the response variable. We also considered a higher proportion ($30\%$) of missing data by choosing the parameter values $\boldsymbol{\phi} = (\phi_0,\ \phi_1,\ \phi_2,\ \phi_3)^t = (-3, 0.2, 0.3, 1)^t$.

## 4.3 Estimation

We estimate all regression parameters, variance components and parameters of the missing data model using the Monte Carlo EM (MCEM) algorithm, as described below.

### 4.3.1 The MCEM algorithm

1. Set $l = 0$. Choose initial values $\boldsymbol{\beta}^{*(l)} = (\alpha_1^{(l)}, \boldsymbol{\beta}^{(l)t}, \boldsymbol{\gamma}^{(l)t})^t, \sigma_u^{2(l)}, \sigma_\varepsilon^{2(l)}$ and $\boldsymbol{\phi}^{(l)}$.
2. Approximate the nonlinear true mean response $m(x)$ by the linear spline

$$\hat{m}^{(l)}(x) = \hat{\beta}_0^{(l)} + \hat{\beta}_1^{(l)} x + \sum_{k=1}^{5} \hat{\gamma}_k^{(l)} (x - \delta_k)_+.$$

3. Using the Gibbs sampling algorithm described earlier in Section 3.3, generate random draws $\mathbf{y}_{i1}^*, \ldots, \mathbf{y}_{im_i}^*$ from the conditional distribution (3.6). Define $\mathbf{y}_i^{(k)} = \left(\mathbf{y}_{ik}^*, \mathbf{y}_{\text{obs},i}\right)^t$, for $k = 1, \ldots, m_i$. Use these $\mathbf{y}_i^{(k)}$ to find the MCEM estimates. Specifically,

   (a) Compute $\boldsymbol{\beta}^{*(l+1)}$ from

$$\boldsymbol{\beta}^{*(l+1)} = \left( \sum_{i=1}^{m} \mathbf{X}_i^{*t} \mathbf{X}_i^* \right)^{-1} \sum_{i=1}^{m} \left( \mathbf{X}_i^{*t} \frac{1}{m_i} \sum_{k=1}^{m_i} \left( \mathbf{y}_i^{(k)} - \mathbf{Z}_i u_i^{(lk)} \right) \right).$$

   (b) Compute $\sigma_u^{2(l+1)}$ from

$$\sigma_u^{2(l+1)} = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{m_i} \sum_{k=1}^{m_i} u_i^{(lk)t} u_i^{(lk)} + \boldsymbol{\Sigma}_i^{(l)} \right).$$

   (c) Compute $\sigma_\varepsilon^{2(l+1)}$ from

$$\sigma_\varepsilon^{2(l+1)} = \frac{1}{mn} \sum_{i=1}^{m} \Bigg\{ \frac{1}{m_i} \sum_{k=1}^{m_i} \left( \mathbf{y}_i^{(k)} - \mathbf{X}_i^* \boldsymbol{\beta}^{*(l+1)} - \mathbf{Z}_i u_i^{(lk)} \right)^t \times \left( \mathbf{y}_i^{(k)} - \mathbf{X}_i^* \boldsymbol{\beta}^{*(l+1)} - \mathbf{Z}_i u_i^{(lk)} \right) + n \boldsymbol{\Sigma}_i^{(l)} \Bigg\}.$$

(d) Compute $\boldsymbol{\phi}^{(l+1)}$ using a Newton-Raphson iterative method, as described in (4.3).

4. Set $l = l + 1$ and continue Steps 2–3.

5. If a convergence is achieved, then declare the current estimates to be the semiparametric MCEM estimates, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2, \hat{\boldsymbol{\phi}})^t$, otherwise return to Step 2.

To compute the asymptotic variance of $\hat{\boldsymbol{\theta}}$, we evaluate the observed Fisher information (3.12) at the MCEM estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2, \hat{\boldsymbol{\phi}})^t$. Details are given in the Appendix. In our experience, the above algorithm is not very sensitive to the choice of initial values. We found the initial estimates by treating that the data were missing at random (MAR). In most cases, we obtained convergence in the MCEM estimation in fewer than 25 iterations when $m_i = 2000$ samples were used for the Monte Carlo approximation.

### 4.3.2 Estimation of missing data model parameters

To estimate the parameters $\boldsymbol{\phi}$, we solve the score equation

$$S(\boldsymbol{\phi}) = \sum_{i=1}^{m} \frac{1}{m_i} \sum_{k=1}^{m_i} \boldsymbol{\omega}_{i2}^{(k)} \left( r_{i2} - p_{i2}^{(k)} \right) = \mathbf{0}.$$

An approximate Fisher information is given by

$$I(\boldsymbol{\phi}) = \sum_{i=1}^{m} \frac{1}{m_i} \sum_{k=1}^{m_i} \boldsymbol{\omega}_{i2}^{(k)} p_{i2}^{(k)} \left( 1 - p_{i2}^{(k)} \right) \boldsymbol{\omega}_{i2}^{t(k)},$$

where

$$\boldsymbol{\omega}_{i2}^{(k)} = \left( 1, \mathbf{y}_i^{(k)}, x_{i2} \right)^t = \left( 1, y_{\text{obs},i1}, y_{ik}^*, x_{i2} \right)^t,$$

and

$$p_{i2}^{(k)} = \frac{e^{\boldsymbol{\omega}_{i2}^{(k)} \boldsymbol{\phi}}}{1 + e^{\boldsymbol{\omega}_{i2}^{(k)} \boldsymbol{\phi}}} = \frac{e^{\phi_0 + \phi_1 y_{\text{obs},i1} + \phi_2 y_{ik}^* + \phi_3 x_{i2}}}{1 + e^{\phi_0 + \phi_1 y_{\text{obs},i1} + \phi_2 y_{ik}^* + \phi_3 x_{i2}}}.$$

We use the Newton-Raphson iterative method to estimate the missing data model parameters $\boldsymbol{\phi}$, given by

$$\boldsymbol{\phi}^{(l+1)} = \boldsymbol{\phi}^{(l)} + \left( I(\boldsymbol{\phi}) \right)^{-1} \Big|_{\boldsymbol{\phi} = \boldsymbol{\phi}^{(l)}} \left( S(\boldsymbol{\phi}) \right) \big|_{\boldsymbol{\phi} = \boldsymbol{\phi}^{(l)}}, \tag{4.3}$$

for $l = 0, 1, 2, \ldots$, where $(l)$ indicates the $l$th iteration.

### 4.4 Results

Tables 1 and 2 present simulated biases and mean squared errors (MSEs) of the estimators of the regression parameter $\alpha_1$, mean response $m_0(x)$ and variance components $(\sigma_u^2, \sigma_\varepsilon^2)$. Table 1 shows the empirical results for the case when $m_0(x)$ is linear: $m_0(x) = 1 + x$, and Table 2 repeats those for the case when $m_0(x)$ is quadratic: $m_0(x) = 1 + x + x^2$.

It is clear from Table 1 that when $m_0(x)$ is linear, both Methods 1 and 2 provide roughly unbiased estimates of the mean response and variance components. As expected, the biases and mean squared errors of the estimators tend to decrease as the sample size increases. When comparing the two methods, our proposed semiparametric method (Method 1) produces slightly larger MSEs of the estimators of $m_0(x)$ for the case when $m_0(x)$ is, in fact, linear, as shown in Table 1. However, our focus is on the estimation of $m_0(x)$ when it is nonlinear. In the case of the quadratic $m_0(x)$, Table 2 shows that the proposed semiparametric method (Method 1) performs much better than the parametric MCEM method (Method 2). For example, when estimating $m_0(x_2)$ with sample size $m = 200$, Table 2 shows that Method 1 provides an empirical bias of 0.0107 and an MSE of 0.0557, whereas Method 2 provides a larger bias of 0.5438 and also a much larger MSE of 0.3498.

We also ran simulations for larger proportions (more than $30\%$) of missing responses (not shown here), where we observed a similar behaviour among the estimates of the regression parameters and variance components. The Simulation code can be found using this link

## 5 Application: HRS Longitudinal Data

In this section, we present an analysis of the Health and Retirement Study (HRS) data introduced earlier in Section 1. The HRS is a longitudinal household survey conducted by the Institute for Social Research at the University of Michigan. The survey includes groups of individuals over age 50 and their spouses in the USA. Its main goal is to provide grouped data (panel data) that allow research and analysis in support of policies and rules on retirement, health insurance, saving, and financial security. The HRS data, available at https://hrs.isr.umich.edu/data-products, were collected from 13 waves of interviews across 15 survey years (1992, 1993, 1994, 1995), and biennially (1996 - 2016).

We consider a subset of the data concerning the physical health of individuals over age 50 and their spouses. The subset includes the most recent surveys for the years 2010, 2012, 2014 and 2016. In our study, the response is the respondent's body mass index (BMI), collected longitudinally over those four time points (years). The BMI is one of the important aspects of the physical health that indicates whether or not the respondent is considered in a healthy category.

In our analysis, we consider the baseline covariates of individuals for the year 2010, which include binary indicators of their health conditions: high blood pressure (HBP), heart disease (Heart), Stroke, and diabetes (Diab). Also, the data include binary indicators of their medical care utilization (Hosp, overnight hospital stay in the last two years), smoking status (Smoke, whether the respondent ever smoked cigarettes or not), and race (White, if the respondent is white or not) and a continuous covariate Age in years. An initial analysis of the data indicated a nonlinear relationship between the response BMI and covariate Age.

Table 1: Comparison of proposed semiparametric MCEM (Method 1) with ordinary MCEM (Method 2) of Ibrahim et al. (2001) when true response is linear. Simulated biases and mean squared errors (MSEs) are shown for the estimators of the mean response $m_0(x)$ at different $x$-values. Missing data parameters $\boldsymbol{\phi} = (-2.5, 0.2, 0.3, 1)^t$ lead to **NMAR** with $23\%$ missing responses. Simulations are based on 1000 replicates of data sets.

| | | | Fitted Model | | | |
|---|---|---|---|---|---|---|
| | | | Method 1 | | Method 2 | |
| True Model | Parameter | True Value | Bias | MSE | Bias | MSE |
| | | m = 100 | | | | |
| Linear | $\alpha_1$ | 1.2 | -0.0136 | 0.0653 | 0.0111 | 0.0643 |
| | $m_0(x_1)$ | 0.72 | 0.0119 | 0.1025 | -0.0029 | 0.0545 |
| | $m_0(x_2)$ | 1.48 | 0.0294 | 0.1103 | -0.0029 | 0.0375 |
| | $m_0(x_3)$ | 2 | 0.0182 | 0.1053 | -0.0029 | 0.0333 |
| | $m_0(x_4)$ | 2.52 | 0.0189 | 0.0893 | -0.0029 | 0.0352 |
| | $m_0(x_5)$ | 3.28 | 0.0157 | 0.0673 | -0.003 | 0.0488 |
| | $\sigma_u^2$ | 1 | -0.0481 | 0.0717 | -0.0142 | 0.069 |
| | $\sigma_\varepsilon^2$ | 1 | -0.055 | 0.0376 | -0.0233 | 0.0356 |
| | | m = 200 | | | | |
| Linear | $\alpha_1$ | 1.2 | -0.013 | 0.0345 | 0.0040 | 0.033 |
| | $m_0(x_1)$ | 0.72 | 0.013 | 0.0477 | -0.0054 | 0.0301 |
| | $m_0(x_2)$ | 1.48 | 0.0153 | 0.0548 | -0.0053 | 0.0208 |
| | $m_0(x_3)$ | 2 | 0.0174 | 0.0598 | -0.0052 | 0.0183 |
| | $m_0(x_4)$ | 2.52 | 0.0131 | 0.045 | -0.0052 | 0.019 |
| | $m_0(x_5)$ | 3.28 | 0.0039 | 0.0398 | -0.0051 | 0.0257 |
| | $\sigma_u^2$ | 1 | -0.0318 | 0.0359 | -0.0072 | 0.0365 |
| | $\sigma_\varepsilon^2$ | 1 | -0.0281 | 0.018 | -0.0087 | 0.0179 |

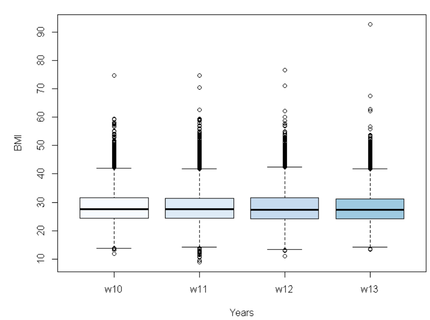
Table 2: Comparison of proposed semiparametric MCEM (Method 1) with ordinary MCEM (Method 2) of Ibrahim et al. (2001) when true response is nonlinear (quadratic). Simulated biases and mean squared errors (MSEs) are shown for the estimators of the mean response $m_0(x)$ at different $x$-values. Missing data parameters $\boldsymbol{\phi} = (-2.5, 0.2, 0.3, 1)^t$ lead to **NMAR** with $23\%$ missing responses. Simulations are based on 1000 replicates of data sets.

| | | | Fitted Model | | | |
|---|---|---|---|---|---|---|
| | | | Method 1 | | Method 2 | |
| True Model | Parameter | True Value | Bias | MSE | Bias | MSE |
| | | m = 100 | | | | |
| Quadratic | $\alpha_1$ | 1.2 | 0.0007 | 0.066 | -0.024 | 0.1205 |
| | $m_0(x_1)$ | 0.79 | 0.0986 | 0.11 | -1.0251 | 1.3002 |
| | $m_0(x_2)$ | 1.71 | 0.0054 | 0.113 | 0.5403 | 0.4072 |
| | $m_0(x_3)$ | 3 | -0.0132 | 0.099 | 0.9459 | 0.9685 |
| | $m_0(x_4)$ | 4.83 | 0.0141 | 0.0919 | 0.8106 | 0.7301 |
| | $m_0(x_5)$ | 8.47 | 0.1221 | 0.0877 | -0.3599 | 0.2749 |
| | $\sigma_u^2$ | 1 | -0.0167 | 0.0727 | 0.105 | 0.1347 |
| | $\sigma_\varepsilon^2$ | 1 | -0.0388 | 0.038 | 1.707 | 3.323 |
| | | m = 200 | | | | |
| Quadratic | $\alpha_1$ | 1.2 | -0.0031 | 0.035 | -0.0111 | 0.0573 |
| | $m_0(x_1)$ | 0.79 | 0.1015 | 0.0641 | -1.0203 | 1.1604 |
| | $m_0(x_2)$ | 1.75 | 0.0107 | 0.0557 | 0.5438 | 0.3498 |
| | $m_0(x_3)$ | 3 | -0.0329 | 0.0543 | 0.9485 | 0.9342 |
| | $m_0(x_4)$ | 4.83 | 0.0107 | 0.0429 | 0.8123 | 0.6955 |
| | $m_0(x_5)$ | 8.47 | 0.1142 | 0.0516 | -0.3595 | 0.2035 |
| | $\sigma_u^2$ | 1 | 0.0832 | 0.073 | 0.0257 | 0.064 |
| | $\sigma_\varepsilon^2$ | 1 | 0.01007 | 0.0172 | 1.755 | 3.304 |

A total of 9760 respondents were surveyed in years 2010, 2012, 2014 and 2016. Many individuals have BMI missing on at least one occasion. The percentage of subjects with at least one response missing is $32\%$, and the overall percentage of missing observations is $5632/(4 \times 9760) = 14.4\%$, where 5632 represents the total number of missing observations. The percentages of respondents with 1, 2 and 3 missing responses were $14\%$, $10.3\%$ and $7.7\%$, respectively. The amount of missing observations is $0.24\%$ at baseline (year 2010) and $2.38\%$, $4.64\%$ and $7.17\%$ at the second (year 2012), third (year 2014) and fourth wave (year 2016), respectively.

In this study, compliance was not mandatory and respondents sometimes refused or missed an interview on one occasion and then were interviewed at the next follow-up time, resulting in non-monotone patterns of missing data. It is reasonable to conjecture that a respondent may not come to an interview if his/her physical health is poor (e.g., unhealthy BMI) and therefore the outcome may be nonignorably missing, or not missing at random (NMAR). So we consider analyzing the data assuming a nonignorable missing data model. Also, for comparison purposes, we consider analyzing the data assuming that the missing data are MAR (missing at random) for which the likelihood-based inference does not depend on the missing data mechanism.

The left panel in Figure 1 shows four boxplots of the response BMI obtained in four waves (w10=year 2010, w11=year 2012, w12=year 2014, w13=year 2016). It is evident that the median BMI levels are similar across the four waves. The right panel exhibits the scatter plot of the mean BMI at a given age versus Age/10, which suggests a curvilinear relationship between the response and covariate Age.



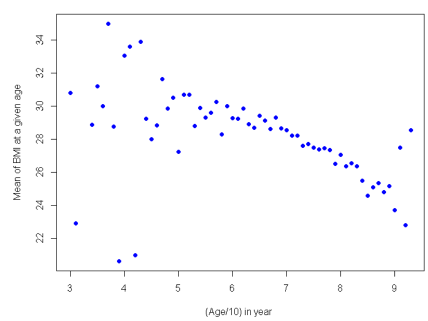




Figure 1: Boxplots of BMI (left). Scatter plot of average BMI vs. age/10 (right).

Figure 2 displays boxplots of the BMI for four different groups of individuals. Starting from the left to the right, the first group consists of those individuals who completed all four interviews (no missing values). The second, third and fourth groups consist of the individuals who missed one, two and three interviews, respectively. It is evident that the median BMI levels are similar across the four waves with a value of 27 for the first group with no missing values. In contrast, for the last group with three missing values, the median BMI levels clearly differ across the four waves, with larger BMI values (e.g., 36 at w12 and 33 at w13). Also, for the second and third groups, the medians are different, with generally higher values of BMI (e.g., 28 at w13 in group 2 and 28.5 at w13 in group

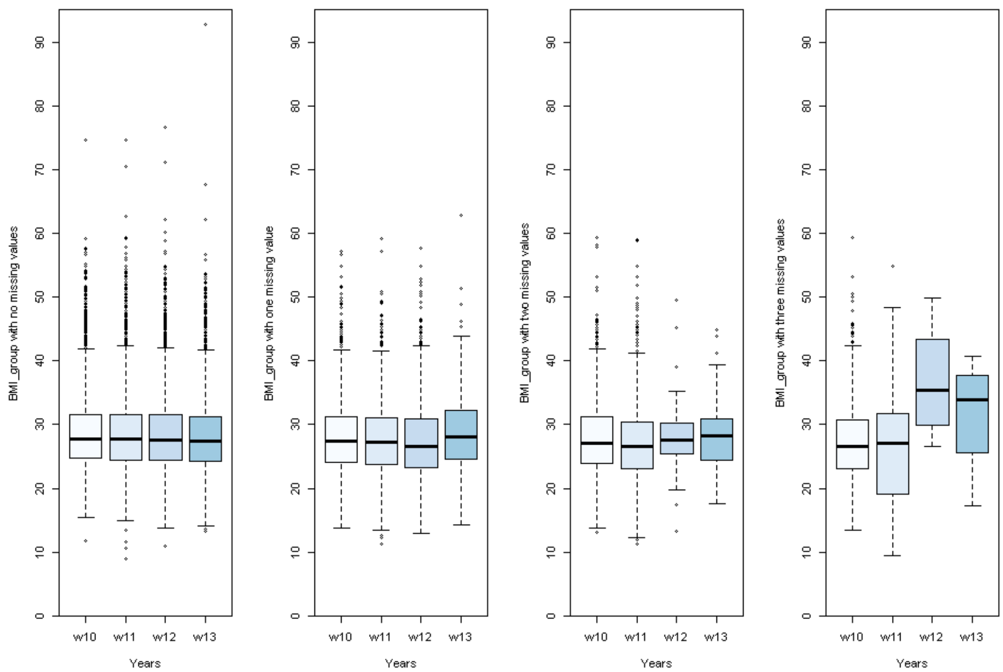



Figure 2: Boxplots of BMI for HRS Data

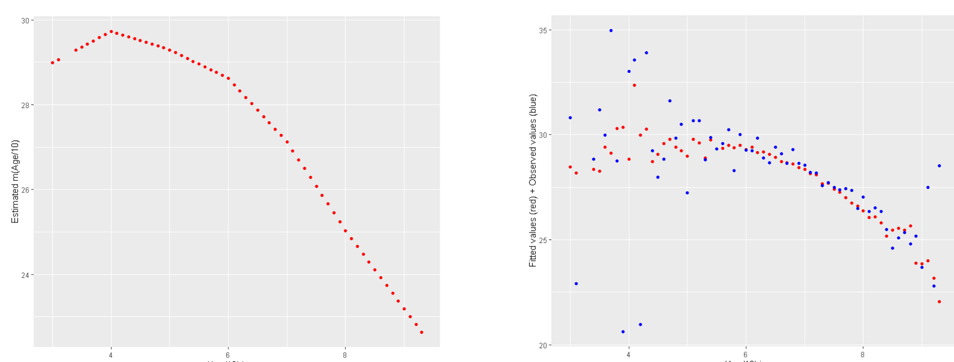



Figure 3: Estimated nonlinear curve $m_0(\text{age})$ for HRS data (left). Scatter plot of average BMI vs. age for HRS data (right).

3), as compared to the first group.

Here we study the relationship between the BMI of respondents and other baseline covariates, assuming a nonignorable missing data mechanism. Let $y_{it}$ denote the BMI of the $i$th individual observed at the $t$th time point ($t = 1, 2, 3, 4$). We consider a semiparametric partially linear mixed model for the BMI, given by

$$y_{it} = \alpha_0 + \alpha_1 t + \alpha_2 \text{HBP}_i + \alpha_3 \text{Heart}_i + \alpha_4 \text{Strok}_i + \alpha_5 \text{Diab}_i + \alpha_6 \text{Smok}_i \\ + \alpha_7 \text{Hosp}_i + \alpha_8 \text{White}_i + m_0((\text{Age}/10)_i) + u_i + \varepsilon_{it}, \tag{5.1}$$

for $i = 1, \ldots, m$, and $t = 1, \ldots, n$, with $m = 9760$ and $n = 4$. Here $m_0(\cdot)$ is assumed an unknown smooth function of the covariate Age. The random effects $u_i$ are assumed to be independent $N(0, \sigma_u^2)$. Also, the random errors $\varepsilon_{it}$ are assumed to be independent $N(0, \sigma_\varepsilon^2)$.

We consider a logistic regression model for the missingness probability $p_{it} = P(r_{it} = 1|\boldsymbol{\phi})$, given by

$$\log\left(\frac{p_{it}}{1 - p_{it}}\right) = \phi_0 + \phi_1 (y_{it}/10) + \phi_2 t + \phi_3 \text{HBP}_i + \phi_4 \text{Heart}_i + \phi_5 \text{Stroke}_i \\ + \phi_6 \text{Diab}_i + \phi_7 \text{Smoke}_i + \phi_8 \text{Hosp}_i + \phi_9 \text{White}_i + \phi_{10} (\text{Age}/10)_i. \tag{5.2}$$

The missing data indicators $r_{it}$ are assumed independent for all $(i, t)$, so that the joint density of the missing data indicators is given by

$$f(\boldsymbol{r}|\boldsymbol{\phi}) = \prod_{i=1}^{m} \prod_{t=1}^{n} \{P(r_{it} = 1|\boldsymbol{\phi})\}^{r_{it}} \{1 - P(r_{it} = 1|\boldsymbol{\phi})\}^{1 - r_{it}}.$$

For a comparative study, we estimate the model parameters by considering four different models:

M1. Proposed partially linear mixed model (5.1) for the mean response $E(y_{it})$ and nonignorable model (5.2) for the missing data.

M2. Partially linear mixed model (5.1) for the mean response $E(y_{it})$, but a misspecified MAR model for the missing data.

M3. Ordinary linear mixed model for the mean response $E(y_{it})$ and nonignorable model (5.2) for the missing data.

M4. Ordinary linear mixed model for the mean response $E(y_{it})$, but a misspecified MAR model for the missing data.

The MCEM methods M1 and M3 for nonignorable missing data were based on $m_i = 500$ Gibbs samples generated from the conditional distribution of the missing data given the observed data. For the spline approximation to the nonlinear function $m_0(\cdot)$, we used a linear spline $(p = 1)$ with $K = 5$ knots. We also considered quadratic $(p = 2)$ and cubic $(p = 3)$ splines (not shown here) to approximate $m_0(\cdot)$, but the linear spline appeared to be sufficient to capture the curvature in the mean response.

The left panel in Figure 3 displays the fitted nonlinear function $\hat{m}_0(\cdot)$ by our proposed semiparametric MCEM approach (M1) that uses a nonignorable missing data model. The right panel in Figure 3 displays a scatter plot of the observed and fitted values of the average BMI against (Age/10), which clearly indicates that our proposed semiparametric MCEM approach is very effective in modelling the nonlinear mean response function.

Table 3 reports the estimates, their corresponding standard errors, and z-values for the regression parameters and variance components of the partially linear mixed model (5.1) obtained by the four methods M1–M4. We observe that the estimates and their standard errors are generally close to each other for the nonignorable models M1 and M3. We also note that the estimates under M1 and M3 generally have smaller standard errors, as compared to those obtained under ignorable (MAR) models M2 and M4, which justifies the use of a nonignorable (NMAR) model for the missing data.

Furthermore, under all four models M1–M4, covariates time, high blood pressure and diabetes appear to be highly significant with p-values less than $10^{-7}$, indicating a strong relationship with the response BMI. Also, covariates Heart, Stroke and Smoke are all significant by all four methods, but the semiparametric approach with nonignorable missing data (M1) provides smaller p-values as compared to M3, justifying the need for modelling the mean response as a nonlinear function.

From Table 3, it appears that the BMI values are higher among respondents with high blood pressure, heart disease and/or diabetes, and lower among respondents with stroke. White respondents have lower BMI values as compared to others. The BMI appears to increase with increased overnight stay at hospital and decrease with an increased smoking habit. The BMI values also decrease over time. The variance components are also highly significant under all models.

Table 4 shows estimated mean response $\hat{m}_0(\cdot)$ at different values of age under the proposed MCEM method (M1) with nonignorable missingness. For example, estimates of $m_0(\cdot)$ are 29.5, 29.81 and 27.86 corresponding to the (age/10) values of 3.6, 4 and 6.5, respectively, indicating a curvilinear relationship.

Table 5 presents the MCEM estimates of the missing data model parameters, their standard errors, and corresponding z-values. Results in Table 5 suggest that the missingness rate varies across all covariates as well as the response BMI at the current time point. Respondents are likely to have exp(0.4448) = 1.56 times higher odds to miss the interview if their BMI values are increased by one unit at the current time point. Also, the study suggests that respondents suffering from any of the diseases, high blood pressure, heart disease, stroke or diabetes are less likely to miss an interview, as compared to those who are otherwise healthy. Also, the results show that respondents who ever smoked are likely to have lower odds to miss an interview as compared to nonsmokers. White respondents are likely to have somewhat higher odds to miss an interview as compared to others. Younger respondents are more likely to miss an interview as compared to older respondents. The missingness probability increases with decreased overnight stays at the hospital. We also find that covariate time has a negative effect on the missingness probability, that is, the odds of missing an interview is higher at latter time points. The Application code can be found using this link

Table 3: Estimates and standard errors (SEs) of regression parameters and variance components for HRS data analysis.

| | Semiparametric approach | | | | | | Parametric approach | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nonignorable (M1) | | | Ignorable (M2) | | | Nonignorable (M3) | | | Ignorable (M4) | | |
| Covariate | Estimate | SE | Z-value | Estimate | SE | Z-value | Estimate | SE | Z-value | Estimate | SE | Z-value |
| Time ($\alpha_1$) | -0.2303 | 0.0104 | -22.09 | -0.178 | 0.0102 | -17.42 | -0.229 | 0.0104 | -22.004 | -0.178 | 0.0102 | -17.39 |
| HBP ($\alpha_2$) | 2.1115 | 0.116 | 18.05 | 2.124 | 0.118 | 17.87 | 2.141 | 0.116 | 18.32 | 2.152 | 0.118 | 18.12 |
| Heart ($\alpha_3$) | 0.362 | 0.128 | 2.82 | 0.396 | 0.13 | 3.03 | 0.345 | 0.128 | 2.68 | 0.376 | 0.13 | 2.88 |
| Stroke ($\alpha_4$) | -0.518 | 0.188 | -2.74 | -0.4608 | 0.192 | -2.39 | -0.536 | 0.188 | -2.83 | -0.484 | 0.192 | -2.51 |
| Diab ($\alpha_5$) | 2.533 | 0.128 | 19.67 | 2.592 | 0.1308 | 19.8 | 2.553 | 0.128 | 19.83 | 2.608 | 0.1309 | 19.92 |
| Smoke ($\alpha_6$) | -0.297 | 0.109 | -2.71 | -0.275 | 0.1114 | -2.47 | -0.3007 | 0.109 | -2.74 | -0.282 | 0.1115 | -2.52 |
| Hosp ($\alpha_7$) | 0.202 | 0.121 | 1.67 | 0.229 | 0.123 | 1.85 | 0.194 | 0.121 | 1.6 | 0.216 | 0.123 | 1.75 |
| White ($\alpha_8$) | -0.217 | 0.1373 | -1.57 | -0.227 | 0.139 | -1.63 | -0.204 | 0.1374 | -1.48 | -0.213 | 0.139 | -1.53 |
| $\sigma_u^2$ | 26.718 | 0.4019 | 66.46 | 27.692 | 0.416 | 66.45 | 26.769 | 0.4027 | 66.47 | 27.748 | 0.417 | 66.45 |
| $\sigma_\varepsilon^2$ | 4.139 | 0.038 | 108.61 | 3.948 | 0.036 | 108.64 | 4.139 | 0.038 | 108.6 | 3.948 | 0.036 | 108.64 |

Table 4: Fitted values of nonlinear function $\hat{\mathrm{m}}_0(\mathrm{age})$ at different values of age by the proposed semiparametric approach with nonignorable missing responses.

| Estimated mean response at six age-values | | | | | | |
|---|---|---|---|---|---|---|
| Age/10 | 3.6 | 4 | 5.8 | 6.5 | 7.4 | 8 |
| $\hat{m}$(Age/10) | 29.5 | 29.81 | 28.76 | 27.86 | 26.28 | 25.04 |

Table 5: ML estimates and standard errors (SEs) of nonignorable missing data model parameters for HRS data analysis.

| Covariate | Estimate | SE | Z-value |
|---|---|---|---|
| Intercept $(\phi_0)$ | 6.445 | 0.219 | 29.409 |
| $y_{it}/10$ $(\phi_1)$ | 0.4448 | 0.029 | 14.938 |
| Time $(\phi_2)$ | -0.895 | 0.016 | -53.801 |
| HBP $(\phi_3)$ | -0.189 | 0.035 | -5.387 |
| Heart $(\phi_4)$ | -0.266 | 0.035 | -7.6041 |
| Stroke $(\phi_5)$ | -0.459 | 0.047 | -9.763 |
| Diab $(\phi_6)$ | -0.328 | 0.0362 | -9.068 |
| Smoke $(\phi_7)$ | -0.3101 | 0.0325 | -9.534 |
| Hosp $(\phi_8)$ | -0.4306 | 0.0332 | -12.947 |
| White $(\phi_9)$ | 0.102 | 0.039 | 2.615 |
| Age/10 $(\phi_{10})$ | -0.382 | 0.025 | -15.067 |

# 6 Discussion

The purpose of this research was to suggest a flexible semiparametric approach to analyzing incomplete longitudinal data with nonignorable and nonmonotone patterns of missing responses. We have developed a semiparametric Monte Carlo EM (MCEM) method in the framework of the penalized regression spline (P-spline) for approximating the mean response parameters and variance components in a partially linear mixed model (PLMM). The proposed method appeared to be very efficient for jointly estimating the parameters of the mean response function and the missing data model. Our simulation study demonstrated that the proposed semiparametric approach generally provides unbiased and efficient estimators when the missing data are NMAR. We have studied the proposed method under different proportions of missing responses in longitudinal data. In all cases, the proposed approach was found to be very flexible and effective in capturing the nonlinear curve when fitting the mean response function.

We have investigated nonignorable and nonmonotone patterns of missing data in the response variable, where in practice there can be covariates which are also nonignorably missing. For this, we need to incorporate the covariates distribution into the observed data likelihood function. Future research is suggested for exploring nonignorable and nonmonotone patterns of missing data in both responses and covariates in longitudinal data for a valid statistical inference.

As there is no practical way of assessing the missing data model used for analyzing incomplete longitudinal data, in a future study, we wish to perform a sensitivity analysis to investigate if the predictors of the mean response function are affected by a misspecified missing data model.

# Acknowledgements

The research of Najla Aloraini is supported by Qassim university, KSA. The research of Sanjoy Sinha is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

# References

Ai, C., Linton, O., and Zhang, Z. (2018), "A simple and efficient estimation method for models with nonignorable missing Data," *arXiv:1801.04202v1*.

Booth, J. G. and Hobert, J. P. (1999), "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm," *Journal of the Royal Statistical Society, Series B*, 61, 265–285.

Carroll, R. J., Fan, J., I., G., and Wand, M. P. (1997), "Generalized partially linear single-index models," *Journal of the American Statistical Association*, 92, 477–489.

Chang, T. and Kott, P. S. (2008), "Using calibration weighting to adjust for nonresponse under a plausible model," *Biometrika*, 95, 557–571.

Coull, B. A., J., S., and Wand, M. P. (2001b), "Respiratory health and air pollution: additive mixed model analyses," *Biostatistics*, 2, 337–349.

Coull, B. A., Ruppert, D., and Wand, M. P. (2001a), "Simple incorporation of interactions into additive models," *Biometrics*, 57, 539–545.

Diggle, P. and Kenward, M. G. (1994), "Informative drop-out in longitudinal data analysis," *Applied Statistics*, 43, 49–93.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible smoothing with B-splines and penalties," *Statistical Science*, 11, 89–121.

Fan, J. and Li, R. (2004), “New estimation and model selection procedures for semi-parametric modelling in longitudinal data analysis,” *Journal of the American Statistical Association*, 99, 710–723.

Gilks, W. R. and Wild, P. (1992), “Adaptive rejection sampling for Gibbs sampling,” *Applied Statistics*, 41, 337–48.

Grace, Y. Y., He, W., and Liang, H. (2009), “Analysis of correlated binary data under partially linear single-index logistic models,” *Journal of Multivariate Analysis*, 100, 278–290.

Harezlak, J., Ryan, L. M., Giedd, J. N., and Lange, N. (2005), “Individual and population penalized regression splines for accelerated longitudinal designs,” *Biometrics*, 61, 1037–1037.

Heckman, N. (1986), “Spline smoothing in a partly linear model,” *Journal of the Royal Statistical Society, Series B*, 48, 244–248.

HRS (2019), “Health and Retirement Study, (RAND HRS Longitudinal File 2016) public use dataset,” *Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740), Ann Arbor, MI.*

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001), “Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable,” *Biometrika*, 88, 551–564.

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

Kott, P. S. and Chang, T. (2010), “Using calibration weighting to adjust for nonignorable unit non-response,” *Journal of the American Statistical Association*, 105, 1265–1275.

Levine, R. A. and Casella, G. (2001), “Implementations of the Monte Carlo EM algorithm,” *Journal of Computational and Graphical Statistics*, 10, 422–439.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, Hoboken, New Jersey: John Wiley and Sons.

Louis, T. (1982), “Finding the observed information matrix when using the EM algorithm,” *Journal of the Royal Statistical Society, Series B*, 44, 226–33.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, Hoboken, New Jersey: John Wiley & Sons.

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions*, Hoboken, New Jersey: John Wiley & Sons.

Minggao, R. E. W. and Taylor, J. M. G. (1996), “An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves,” *Applied Statistics*, 45, 151–163.

Morikawa, K. and Kim, J. K. (2016), "Semiparametric adaptive estimation with nonignorable nonresponse data," *arXiv preprint arXiv:1612.09207*.

Morikawa, K., Kim, J. K., and Kano, Y. (2017), "Semiparametric maximum likelihood estimation with data missing not at random," *The Canadian Journal of Statistics*, 45, 393–409.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008), "Nonparametric small area estimation using penalized spline regression," *Journal of the Royal Statistical Society, Series B*, 70, 265–286.

Parise, H., Wand, M. P., D., R., and Ryan, L. (2001), "Incorporation of historical controls using semiparametric mixed models," *Applied Statistics*, 50, 31–42.

Qin, J., Leung, D., and Shao, J. (2002), "Estimation with survey data under nonignorable nonresponse or informative sampling." *Journal of the American Statistical Association*, 97, 193–200.

RAND, H. (2019), "RAND HRS Longitudinal File 2016. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration." .

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 90, 106–21.

— (1995), "Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data," *Journal of the American Statistical Association*, 89, 846–66.

Ruppert, D. (2002), "Selecting the number of knots for penalized splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D. M. W. and Carroll, R. J. (2009), "Semiparametric regression during 2003-2007," *Electronic Journal of Statistics*, 3, 1193–1256.

Wang, N., Carroll, R. J., and Lin, X. (2005), "Efficient semi-parametric marginal estimation for longitudinal/clustered data," *Journal of the American Statistical Association*, 100, 1090–1095.

Zhao, J. and Shao, J. (2015), "Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data," *Journal of the American Statistical Association*, 110, 1577–1590.

*Received: July 12, 2022*

*Accepted: January 31, 2023*

# A Appendix

The estimated Fisher information matrix of $\boldsymbol{\theta}$ may be written in matrix form as

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} = -H(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}, \mathbf{r})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = - \left.\begin{pmatrix} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^* \partial \boldsymbol{\beta}^{*t}} & \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^* \partial \sigma_u^2} & \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^* \partial \sigma_\varepsilon^2} \\ \frac{\partial^2 \ell}{\partial \sigma_u^2 \partial \boldsymbol{\beta}^{*t}} & \frac{\partial^2 \ell}{\partial (\sigma_u^2)^2} & \frac{\partial^2 \ell}{\partial \sigma_u^2 \partial \sigma_\varepsilon^2} \\ \frac{\partial^2 \ell}{\partial \sigma_\varepsilon^2 \partial \boldsymbol{\beta}^{*t}} & \frac{\partial^2 \ell}{\partial \sigma_\varepsilon^2 \partial \sigma_u^2} & \frac{\partial^2 \ell}{\partial (\sigma_\varepsilon^2)^2} \end{pmatrix}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} .$$

That is,

$$\mathcal{I}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathcal{I}\left(\hat{\boldsymbol{\beta}^*}\right) & \mathcal{I}\left(\hat{\boldsymbol{\beta}^*}, \hat{\sigma_u^2}\right) & \mathcal{I}\left(\hat{\boldsymbol{\beta}^*}, \hat{\sigma_\varepsilon^2}\right) \\ \mathcal{I}\left(\hat{\sigma_u^2}, \hat{\boldsymbol{\beta}^*}\right) & \mathcal{I}\left(\hat{\sigma_u^2}\right) & \mathcal{I}\left(\hat{\sigma_u^2}, \hat{\sigma_\varepsilon^2}\right) \\ \mathcal{I}\left(\hat{\sigma_\varepsilon^2}, \hat{\boldsymbol{\beta}^*}\right) & \mathcal{I}\left(\hat{\sigma_\varepsilon^2}, \hat{\sigma_u^2}\right) & \mathcal{I}\left(\hat{\sigma_\varepsilon^2}\right) \end{pmatrix},$$

where the components of $\mathcal{I}(\hat{\boldsymbol{\theta}})$ are shown below. Calculating the estimated observed FIM of $\hat{\boldsymbol{\beta}^*}$, we get

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}^*}\right) = -\frac{\partial^2}{\partial \boldsymbol{\beta}^* \boldsymbol{\beta}^{*t}} \boldsymbol{Q}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left\{ \sum_{i=1}^m \sum_{k=1}^{m_i} \frac{1}{m_i} S\left(\hat{\boldsymbol{\theta}}; \mathbf{b}_{ik}, \mathbf{r}_i\right) \left(S\left(\hat{\boldsymbol{\theta}}; \mathbf{b}_{ik}, \mathbf{r}_i\right)\right)^t \right\}$$
$$+ \sum_{i=1}^m \frac{\partial}{\partial \boldsymbol{\beta}^*} Q_i(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \left(\frac{\partial}{\partial \boldsymbol{\beta}^*} Q_i(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^t . \tag{A.1}$$

By plugging the score function and second derivatives w.r.t. $\boldsymbol{\beta}^*$ into (A.1), we obtain

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}^*}\right) = \sum_{i=1}^m \mathbf{X}_i^{*t} \hat{\mathbf{V}}_{y_i}^{-1} \mathbf{X}_i^* - \sum_{i=1}^m \sum_{k=1}^{m_i} \frac{1}{m_i} \left\{ \left(\mathbf{X}_i^{*t} \hat{\mathbf{V}}_{y_i}^{-1} \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^* \hat{\boldsymbol{\beta}}^*\right)\right) \times \left(\mathbf{X}_i^{*t} \hat{\mathbf{V}}_{y_i}^{-1} \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^* \hat{\beta}^*\right)\right)^t \right\}$$
$$+ \sum_{i=1}^m \left\{ \left(\mathbf{X}_i^{*t} \hat{\mathbf{V}}_{y_i}^{-1} \sum_{k=1}^{m_i} \frac{1}{m_i} \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^* \hat{\boldsymbol{\beta}}^*\right)\right) \times \left(\mathbf{X}_i^{*t} \hat{\mathbf{V}}_{y_i}^{-1} \sum_{k=1}^{m_i} \frac{1}{m_i} \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^* \hat{\boldsymbol{\beta}}^*\right)\right)^t \right\}.$$

For the diagonal component $\mathcal{I}\left(\hat{\sigma}_\varepsilon^2\right)$, we get

$$\mathcal{I}\left(\hat{\sigma_\varepsilon^2}\right) = -\frac{\partial^2}{\partial \left(\sigma_\varepsilon^2\right)^2} \boldsymbol{Q}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left\{ \sum_{i=1}^m \sum_{k=1}^{m_i} \frac{1}{m_i} S\left(\hat{\boldsymbol{\theta}}; \mathbf{b}_{ik}, \mathbf{r}_i\right) \left(S\left(\hat{\boldsymbol{\theta}}; \mathbf{b}_{ik}, \mathbf{r}_i\right)\right)^t \right\}$$
$$+ \sum_{i=1}^m \frac{\partial}{\partial \sigma_\varepsilon^2} Q_i(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \left(\frac{\partial}{\partial \sigma_\varepsilon^2} Q_i(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^t . \tag{A.2}$$

By using the score function and second derivatives w.r.t. $\sigma_\varepsilon^2$ in (A.2), we get

$$\mathcal{I}\left(\hat{\sigma}_\varepsilon^2\right) = \frac{1}{2}\sum_{i=1}^{m}\left\{\mathrm{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}\right) + \frac{1}{m_i}\sum_{k=1}^{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\left(-2\hat{\mathbf{V}}_{y_i}^{-3}\right)\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}$$
$$- \sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(-\frac{1}{2}\left\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right) - \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}\right)^2$$
$$+ \sum_{i=1}^{m}\left(-\frac{1}{2}\left\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right) - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}\right)^2.$$

The diagonal term $\mathcal{I}\left(\hat{\sigma_u^2}\right)$ is obtained as

$$\mathcal{I}\left(\hat{\sigma_u^2}\right) = -\frac{\partial^2}{\partial\left(\sigma_u^2\right)^2}\boldsymbol{Q}\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left\{\sum_{i=1}^{m}\sum_{k=1}^{m_i}S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\left(S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\right)^t\right\}$$
$$+ \sum_{i=1}^{m}\frac{\partial}{\partial\sigma_u^2}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\left(\frac{\partial}{\partial\sigma_u^2}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^t. \tag{A.3}$$

By using the score function and second derivatives w.r.t. $\sigma_u^2$ in (A.3), we get

$$\mathcal{I}\left(\hat{\sigma}_u^2\right) = \frac{1}{2}\sum_{i=1}^{m}\left\{\mathrm{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}(\mathbf{Z}_i\mathbf{Z}_i^t)^2\right) + \right.$$
$$\left. - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\left(-2\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\right)\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}$$
$$- \sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(-\frac{1}{2}\left\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\right) - \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}\right)^2$$
$$+ \sum_{i=1}^{m}\left(-\frac{1}{2}\left\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\right) - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}\right)^2.$$

For the off diagonal term $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^2\right)$, we have

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^*, \hat{\sigma}_u^2\right) = -\frac{\partial}{\partial\sigma_u^2\partial\boldsymbol{\beta}^*}\boldsymbol{Q}\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left\{\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\left(S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\right)^t\right\}$$
$$+ \sum_{i=1}^{m}\frac{\partial}{\partial\sigma_u^2}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\left(\frac{\partial}{\partial\boldsymbol{\beta}^*}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^t = \mathcal{I}\left(\hat{\sigma_u^2}, \hat{\boldsymbol{\beta}^*}\right).$$

This leads to

$$
\begin{aligned}
\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*},\hat{\sigma_u^2}\right) =& \sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{X}_i \\
&-\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\Bigg\{-\frac{1}{2}\Bigg\{\operatorname{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\right)-\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\Bigg\} \\
&\times\left(\mathbf{X}_i^{*t}\hat{\mathbf{V}}_{y_i}^{-1}(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*})\right)^{t}\Bigg\} \\
&+\sum_{i=1}^{m}\Bigg\{-\frac{1}{2}\Bigg\{\operatorname{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\right)-\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\boldsymbol{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\Bigg\} \\
&\times\left(\mathbf{X}_i^{*t}\hat{\mathbf{V}}_{y_i}^{-1}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\right)^{t}\Bigg\}.
\end{aligned}
$$

For the off diagonal term $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*},\hat{\sigma_\varepsilon^2}\right)$, we have

$$
\begin{aligned}
\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*},\hat{\sigma_\varepsilon^2}\right) =& -\frac{\partial}{\partial\sigma_\varepsilon^2\partial\boldsymbol{\beta}^{*}}\boldsymbol{Q}\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}-\left\{\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\left(S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\right)^{t}\right\} \\
&+\sum_{i=1}^{m}\frac{\partial}{\partial\sigma_\varepsilon^2}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\left(\frac{\partial}{\partial\boldsymbol{\beta}^{*}}Q_i\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^{t}=\mathcal{I}\left(\hat{\sigma_\varepsilon^2},\hat{\boldsymbol{\beta}}^{*}\right).
\end{aligned}
$$

This yields

$$
\begin{aligned}
\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*},\hat{\sigma}_\varepsilon^2\right) =& \sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-2}\mathbf{X}_i \\
&-\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\Bigg\{-\frac{1}{2}\Bigg\{\operatorname{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right)-\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\Bigg\} \\
&\times\left(\mathbf{X}_i^{*t}\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\right)^{t}\Bigg\} \\
&+\sum_{i=1}^{m}\Bigg\{-\frac{1}{2}\Bigg\{\operatorname{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right)-\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t}\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\Bigg\} \\
&\times\left(\mathbf{X}_i^{*t}\hat{\mathbf{V}}_{y_i}^{-1}\sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)}-\mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)\right)^{t}\Bigg\}.
\end{aligned}
$$

For the off diagonal term $\mathcal{I}\left(\hat{\sigma_u^2}, \hat{\sigma_\varepsilon^2}\right)$, we have

$$\mathcal{I}\left(\hat{\sigma_u^2}, \hat{\sigma_\varepsilon^2}\right) = -\frac{\partial}{\partial\sigma_u^2\partial\sigma_\varepsilon^2}\boldsymbol{Q}\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \left\{\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\left(S\left(\hat{\boldsymbol{\theta}};\mathbf{b}_{ik},\mathbf{r}_i\right)\right)^t\right\} + \sum_{i=1}^{m}\frac{\partial}{\partial\sigma_u^2}Q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\left(\frac{\partial}{\partial\sigma_\varepsilon^2}Q_i\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^t = \mathcal{I}\left(\hat{\sigma_\varepsilon^2}, \hat{\sigma_u^2}\right).$$

This gives

$$\begin{aligned}
\mathcal{I}\left(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2\right) =& \frac{1}{2}\sum_{i=1}^{m}\Bigg\{\mathrm{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}\mathbf{Z}_i\mathbf{Z}_i^t\right) - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t((-1)\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-2} \\
& -\hat{\mathbf{V}}_{y_i}^{-2}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1})\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\Bigg\} \\
& -\sum_{i=1}^{m}\sum_{k=1}^{m_i}\frac{1}{m_i}\Bigg\{-\frac{1}{2}\Bigg\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\right) - \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\Bigg\} \\
& \times -\frac{1}{2}\Bigg\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right) - \left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\Bigg\}\Bigg\} \\
& +\sum_{i=1}^{m}\Bigg\{-\frac{1}{2}\Bigg\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\right) - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\Bigg\} \\
& \times -\frac{1}{2}\Bigg\{\mathrm{tr}\left(\hat{\mathbf{V}}_{y_i}^{-1}\right) - \sum_{k=1}^{m_i}\frac{1}{m_i}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\hat{\mathbf{V}}_{y_i}^{-2}\left(\mathbf{y}_i^{(k)} - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\Bigg\}\Bigg\}.
\end{aligned}$$

Now, in case there is no missing data, the estimated observed information matrix of $\boldsymbol{\theta}$ can be simplified as follows.
The Fisher information $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^*\right)$ may be obtained as

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^*\right) = \sum_{i=1}^{m}\mathbf{X}_i^{*t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{X}_i^*.$$

The term $\mathcal{I}\left(\hat{\sigma}_\varepsilon^2\right)$ may be obtained as

$$\mathcal{I}\left(\hat{\sigma}_\varepsilon^2\right) = \frac{1}{2}\sum_{i=1}^{m}\left\{\mathrm{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}\right) + \left(\mathbf{y}_i - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)^t\left(-2\hat{\mathbf{V}}_{y_i}^{-3}\right)\left(\mathbf{y}_i - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*\right)\right\}.$$

The term $\mathcal{I}\left(\hat{\sigma}_u^2\right)$ may be obtained as

$$\mathcal{I}\left(\hat{\sigma}_u^2\right) = \frac{1}{2}\sum_{i=1}^{m}\left\{\mathrm{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}(\mathbf{Z}_i\mathbf{Z}_i^t)^2\right) + (\mathbf{y}_i - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*)^t(-2\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^t\hat{\mathbf{V}}_{y_i}^{-1})(\mathbf{y}_i - \mathbf{X}_i^*\hat{\boldsymbol{\beta}}^*)\right\}.$$

The off diagonal term $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*}, \hat{\sigma}_{\varepsilon}^{2}\right)$ may be obtained as

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*}, \hat{\sigma}_{\varepsilon}^{2}\right) = \sum_{i=1}^{m} \left(\mathbf{y}_i - \mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t} \hat{\mathbf{V}}_{y_i}^{-2}\mathbf{X}_i^{*}.$$

The term $\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*}, \hat{\sigma_u^2}\right)$ may be obtained as

$$\mathcal{I}\left(\hat{\boldsymbol{\beta}}^{*}, \hat{\sigma_u^2}\right) = \sum_{i=1}^{m} \left(\mathbf{y}_i - \mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t} \hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{X}_i^{*}.$$

The term $\mathcal{I}\left(\hat{\sigma}_u^2, \hat{\sigma}_{\varepsilon}^2\right)$ may be obtained as

$$\begin{aligned}\mathcal{I}\left(\hat{\sigma}_u^2, \hat{\sigma}_{\varepsilon}^2\right) =& \frac{1}{2}\sum_{i=1}^{m} \Bigg\{ \operatorname{tr}\left(-\hat{\mathbf{V}}_{y_i}^{-2}\mathbf{Z}_i\mathbf{Z}_i^{t}\right) - \left(\mathbf{y}_i - \mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right)^{t} ((-1)\hat{\mathbf{V}}_{y_i}^{-1}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\mathbf{V}}_{y_i}^{-2} \\ & -\hat{\mathbf{V}}_{y_i}^{-2}\mathbf{Z}_i\mathbf{Z}_i^{t}\hat{\mathbf{V}}_{y_i}^{-1}) \left(\mathbf{y}_i - \mathbf{X}_i^{*}\hat{\boldsymbol{\beta}}^{*}\right) \Bigg\}.\end{aligned}$$