Benchmarking the Mixed Linear Model for Longitudinal Data

THUAN NGUYEN AND JIMING JIANG
Oregon Health and Science University and University of California, Davis

We propose a new method of handling missing values in longitudinal data under the linear mixed model assumption. The new method combines the complete-data linear mixed model with benchmark equations that involve both the complete and incomplete data. Simulation studies show that the new method improve the efficiency of inference when a significant proportion of the data are missing.

Key Words. Longitudinal Data, Linear Mixed Models, Missing Values

1 Introduction

Mixed linear models are widely used in longitudinal studies (e.g., Laird & Ware 1982, Diggle *et al.* 1996). It is not uncommon that some of the data, either the responses or the covariates, are missing at certain time points. Standard treatments of the missing values include the throwing-out strategy, in which an entire record (including the response and covariates) is thrown out if at least one value is missing, imputation and modeling the missing data mechanism. See, for example, Diggle *et al.* (1996, chapter 13), Little and Rubin (2002). However, with the exception of the throwing-out method, which is obviously very inefficient, these strategies are not simple and often require knowledge, and assumptions, beyond the linear mixed models.

In this paper, we propose a simple approach to missing values in longitudinal studies incorporating the linear mixed models. Suppose that for the ith subject, $i = 1, \dots, m$, the data $y_{it}, x_{itk}, 1 \leq k \leq p$ are supposed to be collected at the time points $t \in T_i$, where y_{it} is the response and $x_{itk}, 1 \leq k \leq p$ are the covariates. However, due to the missing data, only a subset of the supposed values are collected. Our basic model is a (conditional) linear mixed model that assumes that, given the complete covariates $x_{itk}, 1 \leq i \leq m, t \in T_i, 1 \leq k \leq p$, the complete responses $y_{it}, 1 \leq i \leq m, t \in T_i$ satisfies

$$y_{it} \quad = \quad x'_{it}\beta + z'_{it}u_i + e_{it}, \tag{1}$$

where $x_{it} = (x_{itk})_{1 \leq k \leq p}$, β is a vector of unknown fixed effects, z_{it} is a known $b \times 1$ vector, u_i is a b-dimensional vector of subject-specific random effects, and e_{it} is an error. Write $y_i = (y_{it})_{t \in T_i}$, $X_i = (x'_{it})_{t \in T_i}$, $Z_i = (z'_{it})_{t \in T_i}$ and $e_i = (e_{it})_{t \in T_i}$. Then, (1) can be expressed as

$$y_i \quad = \quad X_i\beta + Z_iu_i + e_i, \tag{2}$$

$i = 1, \dots, m$. It is assumed that the u_i's and e_i's are independent with $u_i \sim N(0, G_i)$ and $e_i \sim N(0, R_i)$, where the covariance matrices G_i and R_i depend on some vector ψ of variance components. The expression (2) and the assumptions below are the same as the longitudinal linear mixed model discussed in Datta and Lahiri (2000) and Jiang (2007).

Due to the missing data, only some of the equations (1) are practically usable—namely, those in which none of the $y_{it}, x_{itk}, 1 \leq k \leq p$ are missing. The question is how to use the data more efficiently in case some of the $y_{it}, x_{itk}, 1 \leq k \leq p$ are missing. Here, we are not interested in the cases in which no data is available, so assume that at least one of the records $y_{it}, x_{itk}, 1 \leq k \leq p$

is collected. As noted, throwing out the entire records because of one or more missing values is inefficient. To make more efficient use of the data we make the following assumption that has something to do with the marginal distributions of the responses and covariates. Let $I_{t,0} = \{1 \le i \le m, y_{it}$ is observed$\}$ and $I_{t,k} = \{1 \le i \le m, x_{itk}$ is observed$\}$, $1 \le k \le p$, and $m_{t,k} = |I_{t,k}|$ (hereafter $|S|$ denotes the cardinality of the set S), $0 \le k \le p$. Define

$$\bar{y}_{\cdot t} = \begin{cases} 0, & \text{if } m_{t,0} = 0, \\ m_{t,0}^{-1} \sum_{i \in I_{t,0}} y_{it}, & \text{if } m_{t,0} > 0, \end{cases}$$
$$\bar{x}_{\cdot tk} = \begin{cases} 0, & \text{if } m_{t,k} = 0, \\ m_{t,k}^{-1} \sum_{i \in I_{t,k}} x_{itk}, & \text{if } m_{t,k} > 0, \end{cases} \quad 1 \le k \le p. \tag{3}$$

Furthermore, let $I_t = \cap_{k=0}^{p} I_{t,k}$ and $m_t = |I_t|$. I_t is the subset of indexes i such that there are no missing values among $y_{it}, x_{itk}, 1 \le k \le p$. Define $\bar{y}_{\cdot t}^{\mathrm{a}}, \bar{x}_{\cdot tk}^{\mathrm{a}}, 1 \le k \le p$ the same way as (3) except with $I_{t,k}$ and $m_{t,k}$ replaced by I_t and m_t, $0 \le k \le p$, respectively. Here the superscript a refers to "all-observed". Let T denote the set of index t such that at least one of the values $y_{it}, x_{itk}, 1 \le k \le p$ is not missing. We are not interested in the cases in which no observation, either the response or the covariates, is available, so assume that $T \ne \emptyset$. Let $T_{\mathrm{a}} = \{t \in T, m_t \ne 0\}$. The additional assumption we make is that

$$\begin{aligned} \mathrm{E}\left(\frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{y}_{\cdot t}^{\mathrm{a}}\right) &= \mathrm{E}\left(\frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{y}_{\cdot t}\right), \\ \mathrm{E}\left(\frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{x}_{\cdot tk}^{\mathrm{a}}\right) &= \mathrm{E}\left(\frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{x}_{\cdot tk}\right), \quad 1 \le k \le p; \end{aligned} \tag{4}$$

in other words, the mean of the all-observed average is equal to the mean of the observed average for the response and each covariate. Note that, here, $|T_{\mathrm{a}}|$ is considered a random variable, so cannot be canceled from the two sides of the equations.

To see that (4) is a reasonable assumption, consider the following special case. Let M_{it} be the $(p+1)$-dimensional vector whose first component is 1 if y_{it} is missing, and 0 otherwise, and whose $(k+1)$th component is 1 is x_{itk} is missing, and 0 otherwise, $1 \le k \le p$. M_{it} is the vector of indicators for the missing values for the given i and t. Let M denote the array $M_{it}, 1 \le i \le m, t \in T_i$. Suppose that the following hold:

(i) The complete data is independent of M.

(ii) The marginal means of y_{it} and $x_{itk}, 1 \le k \le p$ are finite and do not depend on i.

It is shown in Appendix that conditions (i) and (ii) imply (4).

Now, let us see how (4) can help. Let $S_{\mathrm{a}} = \{(i,t) : t \in T_{\mathrm{a}}, i \in I_t\}$. For each $(i,t) \in S_{\mathrm{a}}$, (1) holds, where $y_{it}, x_{itk}, 1 \le k \le p$ are all observed. It follows that

$$\begin{aligned} \frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{y}_{\cdot t}^{\mathrm{a}} &= \frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \frac{1}{m_t}\sum_{i \in I_t} y_{it} \\ &= \frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \frac{1}{m_t}\sum_{i \in I_t} \left(\sum_{k=1}^{p} x_{itk}\beta_k + z_{it}' u_i + e_{it}\right) \\ &= \sum_{k=1}^{p}\left(\frac{1}{|T_{\mathrm{a}}|}\sum_{t \in T_{\mathrm{a}}} \bar{x}_{\cdot tk}^{\mathrm{a}}\right)\beta_k + \cdots , \end{aligned} \tag{5}$$

Benchmarking mixed models

where $\cdots$ has mean zero. Thus, by taking expectations on both sides, and observing (4), we get

$$\mathrm{E}\left(\frac{1}{|T_\mathrm{a}|}\sum_{t\in T_\mathrm{a}}\bar{y}_{\cdot t}\right) = \sum_{k=1}^{p}\mathrm{E}\left(\frac{1}{|T_\mathrm{a}|}\sum_{t\in T_\mathrm{a}}\bar{x}_{\cdot tk}\right)\beta_k. \tag{6}$$

Equation (6) leads to a method-of-moments equation, by removing the expectation signs, that is,

$$\frac{1}{|T_\mathrm{a}|}\sum_{t\in T_\mathrm{a}}\bar{y}_{\cdot t} = \sum_{k=1}^{p}\left(\frac{1}{|T_\mathrm{a}|}\sum_{t\in T_\mathrm{a}}\bar{x}_{\cdot tk}\right)\beta_k. \tag{7}$$

Although (7) is similar to (5) without the term $\cdots$, the difference is that the averages inside the summations are now taken over larger data sets. The factor $|T_\mathrm{a}|^{-1}$ can now be canceled from both sides of the equation. The result is what we call a *benchmark* equation, or simply benchmark:

$$\sum_{t\in T_\mathrm{a}}\bar{y}_{\cdot t} = \sum_{k=1}^{p}\left(\sum_{t\in T_\mathrm{a}}\bar{x}_{\cdot tk}\right)\beta_k. \tag{8}$$

The idea is to fit the linear mixed model (1) for the all-observed data, that is, with $(i,t)\in S_\mathrm{a}$, subject to the benchmark (8). This may not work, however, if $p=1$ (because then β is completely determined by the benchmark). Nevertheless, in most applications we have $p>1$.

A simple numerical procedure is developed in section 2 for fitting the linear mixed model subject to the benchmark. In section 3 we discuss estimation of the variations of the benchmarked maximum likelihood estimators. In section 4 we carry out a simulation study to investigate the gain of efficiency by the benchmarking. Technical derivations are deferred to Appendix.

2 Computation

Let I denote the subset of indexes i such that $(i,t)\in S_\mathrm{a}$ for at least one t, and $T_{\mathrm{a},i}=\{t\in T_i:(i,t)\in S_\mathrm{a}\}$, $i\in I$, so that $S_\mathrm{a}=\cup_{i\in I}\{(i,t):t\in T_{\mathrm{a},i}\}$. Define $\tilde{y}_i=(y_{it})_{t\in T_{\mathrm{a},i}}$, $\tilde{X}_i=(x'_{it})_{t\in T_{\mathrm{a},i}}$, $\tilde{Z}_i=(z'_{it})_{t\in T_{\mathrm{a},i}}$ and $\tilde{e}_i=(e_{it})_{t\in T_{\mathrm{a},i}}$. Then we have the all-observed version of (2):

$$\tilde{y}_i = \tilde{X}_i\beta+\tilde{Z}_iu_i+\tilde{e}_i, \quad i\in I, \tag{9}$$

where β and the u_i's are the same as in (2), the $\tilde{e}_i$'s are independent with $\tilde{e}_i\sim N(0,\tilde{R}_i)$, $\tilde{R}_i$ being dependent on ψ, and the u_i's and $\tilde{e}_i$'s are independent, conditional on the complete covariates. It follows that, conditional on the complete covariates, $\tilde{y}=(\tilde{y}_i)_{i\in I}\sim N(\tilde{X}\beta,\tilde{V})$, where $\tilde{X}=(\tilde{X}_i)_{i\in I}$ and $\tilde{V}=\mathrm{diag}(\tilde{V}_i, i\in I)$ with $\tilde{V}_i=\tilde{Z}_iG_i\tilde{Z}'_i+\tilde{R}_i$. The conditional log-likelihood function based on $\tilde{y}$ is given by

$$\begin{aligned} l &= -\frac{1}{2}\{n\log(2\pi)+\log(|\tilde{V}|)+(\tilde{y}-\tilde{X}\beta)'\tilde{V}^{-1}(\tilde{y}-\tilde{X}\beta)\} \\ &= -\frac{n}{2}\log(2\pi)-\frac{1}{2}\sum_{i\in I}\{\log(|\tilde{V}_i|)+(\tilde{y}_i-\tilde{X}_i\beta)'\tilde{V}_i^{-1}((\tilde{y}_i-\tilde{X}_i\beta)\}, \end{aligned} \tag{10}$$

where $n=\dim(\tilde{y})$. The estimators of the parameters β and ψ are obtained by maximizing (10) subject to the benchmark constraint (8). Write $g=(g_k)_{1\le k\le p}$ with $g_k=\sum_{t\in T_\mathrm{a}}\bar{x}_{\cdot tk}$ and $h=\sum_{t\in T_\mathrm{a}}\bar{y}_{\cdot t}$. Using the method of Lagrange multipliers, we consider

$$L = l+\lambda(g'\beta-h), \tag{11}$$

where λ is the Lagrange multiplier, and find a stationary point of L without a constraint. To do the latter, we first let $\partial L/\partial\beta = \tilde{X}'\tilde{V}^{-1}\tilde{y} - \tilde{X}'\tilde{V}^{-1}\tilde{X}\beta + \lambda g = 0$. This gives

$$\beta \;=\; (\tilde{X}'\tilde{V}^{-1}\tilde{X})^{-1}(\tilde{X}'\tilde{V}^{-1}\tilde{y} + \lambda g), \tag{12}$$

assuming existence of the inverse matrices. We then plug (12) into the equation $\partial L/\partial\lambda = g'\beta - h = 0$ to obtain a solution for λ as a function of ψ:

$$\tilde{\lambda} \;=\; \frac{h - g'(\tilde{X}'\tilde{V}^{-1}\tilde{X})^{-1}\tilde{X}'\tilde{V}^{-1}\tilde{y}}{g'(\tilde{X}'\tilde{V}^{-1}\tilde{X})^{-1}g}. \tag{13}$$

We then plug (13) back to (12) to obtain a solution for β as a function of ψ:

$$\tilde{\beta} \;=\; (\tilde{X}'\tilde{V}^{-1}\tilde{X})^{-1}(\tilde{X}'\tilde{V}^{-1}\tilde{y} + \tilde{\lambda} g). \tag{14}$$

Finally, we replace β by $\tilde{\beta}$ in the equations

$$\frac{\partial l}{\partial\psi_r} \;=\; -\frac{1}{2}\left\{\mathrm{tr}\left(\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\psi_r}\right) - (\tilde{y} - \tilde{X}\beta)'\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\psi_r}\tilde{V}^{-1}(\tilde{y} - \tilde{X}\beta)\right\} \;=\; 0,$$

$1 \le r \le q$, where ψ_r is the rth component of ψ and $q = \dim(\psi)$, to obtain q equations that only involve ψ. By solving the latter equations we obtain the estimator of ψ, denoted by $\hat{\psi}$. The estimator of β, $\hat{\beta}$, is obtained by replacing ψ in (14) by $\hat{\psi}$ (note that $\tilde{\lambda}$ also depends on ψ). $\hat{\beta}$ and $\hat{\psi}$ are called the benchmarked maximum likelihood estimators (BMLE). An estimator of λ, $\hat{\lambda}$, is also obtained by replacing ψ by $\hat{\psi}$ in (13). Although $\hat{\lambda}$ may not be of direct interest, it is needed in the estimation of the standard errors. See the next section.

3 Estimation of variation

Recall that the BMLE, $\hat{\theta} = (\hat{\beta},' \hat{\psi}', \hat{\lambda})'$, is a solution to $\partial L/\partial\theta = 0$. It is easy to show that L can be expressed as $L = c + \sum_{i=1}^m L_i$, where c is a constant and

$$\begin{aligned} L_i \;=\; & \lambda \sum_{t\in T_{\mathrm{a}}}\left\{\sum_{k=1}^p \beta_k m_{t,k}^{-1} 1_{(i\in I_{t,k})} x_{itk} - m_{t,0}^{-1} 1_{(i\in I_{t,0})} y_{it}\right\} \\ & -\frac{1}{2} 1_{(i\in I)}\{\log(|\tilde{V}_i|) + (\tilde{y}_i - \tilde{X}_i\beta)'\tilde{V}_i^{-1}(\tilde{y}_i - \tilde{X}_i\beta)\}. \end{aligned} \tag{15}$$

We assume that the following assumptions hold in addition to assumptions (i) and (ii) in section 1:

(iii) $(y_i, X_i), i = 1, \ldots, m$ are independent.

(iv) The total number of the time points t is bounded.

By arguments similar to White (1982), it can be shown that, under assumptions (i)—(iv) and some additional regularity conditions (identifiability, moments, etc.), the BMLE is root-m consistent in the sense that $\sqrt{m}(\hat{\theta} - \theta)$ is bounded in probability, where $\theta = (\beta', \psi', \lambda)'$ is the unique solution to $\mathrm{E}(\partial L/\partial\theta) = 0$ so that β and ψ are the true parameters for the linear mixed model. In the following, partial derivatives such as $\partial L/\partial\theta$ are understood as evaluated at θ. By Taylor expansion, we have

$$\begin{aligned} 0 &= \left.\frac{\partial L}{\partial\theta}\right|_{\hat{\theta}} \\ &\approx \frac{\partial L}{\partial\theta} + \left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)(\hat{\theta}-\theta) \\ &\approx \frac{\partial L}{\partial\theta} + \mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)(\hat{\theta}-\theta). \end{aligned}$$

Hereafter, $\approx$ means that the remaining term is of lower order, in a suitable sense. Thus, we have $\mathrm{E}(\partial^2 L/\partial\theta\partial\theta')(\hat{\theta}-\theta) \approx -\partial L/\partial\theta$, or

$$\hat{\theta}-\theta \approx -\left\{\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)\right\}^{-1}\frac{\partial L}{\partial\theta}. \tag{16}$$

Write $\Sigma = \mathrm{Var}(\hat{\theta})$. It follows by (16) that

$$\Sigma \approx \left\{\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)\right\}^{-1}\mathrm{Var}\left(\frac{\partial L}{\partial\theta}\right)\left\{\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)\right\}^{-1}. \tag{17}$$

It is shown in Appendix that the right side of (17) can be further approximated by

$$\hat{\Sigma} = \left(\sum_{i=1}^m \hat{Q}_i\right)^{-1}\left(\sum_{i=1}^m \hat{P}_i\right)\left(\sum_{i=1}^m \hat{Q}_i\right)^{-1}, \tag{18}$$

where $\hat{P}_i = \hat{p}_i\hat{p}_i'$ with $\hat{p}_i = (\partial L_i/\partial\theta)|_{\theta=\hat{\theta}}$ and $\hat{Q}_i = (\partial^2 L_i/\partial\theta\partial\theta')|_{\theta=\hat{\theta}}$. Detailed expressions for the partial derivatives are also given in Appendix. We call (18) the "sandwich" estimator of Σ.

It should be pointed out that the "breads" of the "sandwich" are well defined, that is, the matrix $\sum_{i=1}^m \hat{Q}_i$ is nonsingular, at least asymptotically. From the expressions (A.4) in Appendix, we have

$$\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right) = -\begin{pmatrix} A & 0 & C \\ 0 & B & 0 \\ C' & 0 & 0 \end{pmatrix},$$

where A, B are nonsingnular matrices and C is a nonzero vector. It follows that

$$\begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -C'A^{-1} & 0 & I \end{pmatrix}\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)\begin{pmatrix} I & 0 & -A^{-1}C \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} = -\begin{pmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & -C'A^{-1}C \end{pmatrix},$$

which is nonsingular; therefore, $\mathrm{E}(\partial^2 L/\partial\theta\partial\theta')$ is nonsingular. In fact, we have

$$\liminf \lambda_{\min}\left\{\frac{1}{m}\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right)\right\} > 0$$

($\lambda_{\min}$ denotes the smallest eigenvalue), under regularity conditions. Then, because $\sum_{i=1}^m \hat{Q}_i \approx \mathrm{E}(\partial^2 L/\partial\theta\partial\theta')$ (see subsection A.3), $\sum_{i=1}^m \hat{Q}_i$ is nonsingular asymptotically.

In particular, the standard error of $\hat{\beta}_k$ is obtained by the square root of the kth diagonal element of $\hat{\Sigma}$, $1 \le k \le p$, and the standard error of $\hat{\psi}_s$ is obtained by the square root of the $(p+s)$th diagonal element of $\hat{\Sigma}$, $1 \le s \le q$.

Table 1: **Simulated Mean, Variance and MSE: Scenario (1)**

Parameter	True	Mean		Variance		MSE	
	Value	AMLE	BMLE	AMLE	BMLE	AMLE	BMLE
β_0	0.20	0.189	0.192	0.227	0.211	0.227	0.211
β_1	0.10	0.099	0.098	0.158	0.156	0.158	0.156
β_2	0.20	0.200	0.201	0.168	0.167	0.168	0.167
β_3	0.50	0.518	0.511	0.300	0.285	0.300	0.285
β_4	0.50	0.496	0.495	0.080	0.078	0.080	0.078
σ^2	0.25	0.212	0.223	0.116	0.114	0.117	0.115
τ^2	0.09	0.079	0.079	0.032	0.032	0.032	0.032

4 Simulation studies

We first compare finite sample performance of the BMLE with that of the "all-observed" maximum likelihood estimator (AMLE) in terms of the mean squared error (MSE). The AMLE is obtained by using only the complete data records, that is, those in which there are no missing values for either the response or the covariates (in other words, the response and covariates are all observed). Consider the following linear mixed model:

$$y_{it} \quad = \quad \beta_0 + \beta_1 x_{it1} + \beta_2 x_{it2} + \beta_3 x_i + \beta_4 x_{it4} + u_i + e_{it}, \tag{19}$$

$i = 1, \dots, m, t = 1, 2, 3$, where x_{it1} and x_{it2} are two indicators of the time points such that $x_{itk} = 1_{(t=k+1)}, k = 1, 2$ ($t = 1, 2, 3$ correspond to the baseline measure and two measures at different time points after the treatment); $x_i, i = 1, \dots, m$ are generated independently from a Bernoulli(0.5) distribution, the x_{it4}'s are generated independently from a $N(0, 1)$ distribution and are independent with the x_i's; the u_i's and e_{it}'s are generated independently such that $u_i \sim N(0, \sigma^2)$ and $e_{it} \sim N(0, \tau^2)$. The true parameters are $\beta_0 = 0.2, \beta_1 = 0.1, \beta_2 = 0.2, \beta_3 = 0.5, \beta_4 = 0.5, \sigma = 0.5$ and $\tau = 0.3$. Next, we generate, independent from the data, the missing value indicators. These are expressed as M_i and $M_{itk}, i = 1, \dots, 25, t = 1, 2, 3, k = 0, 4$ such that $M_i \sim \text{Bernoulli}(p_2)$ and $M_{itk} \sim \text{Bernoulli}(p_k), k = 0, 4$. The response y_{it} is missing iff $M_{it0} = 1$; x_i is missing iff $M_i = 1$; and x_{it4} is missing iff $M_{it4} = 1$. (The reason that there are no M_{it1} and M_{it2} is because x_{it1} and x_{it2} are always observed; also $M_{it3} = M_i$ because $x_{it3} = x_i$, which does not depend on t.) Two scenarios of the p_k's are considered: (1) $p_2 = 0.1, p_0 = p_4 = 0.3$; (2) $p_2 = 0.2, p_0 = p_4 = 0.5$. The simulated means, variances and MSEs for the parameter estimation based on 500 simulations are reported in Tables 1 and 2. It is seen that the gain of BMLE over AMLE much more substantial under Scenario (2) than under Scenario (1). This makes sense because more missing data are expected when the p's are larger; therefore, more information can be recovered by using the BMLE. In particular, under Scenario (2), the percentage MSE reduction by BMLE is in the range of 30-50% for most of the β parameters. Another interesting observation is that there is apparently no gain by BMLE for estimating τ^2 under either scenarios.

We next compare the performance of AMLE and BMLE in terms of testing the hypothesis $\text{H}_0 : \beta_3 = 0.5$ versus $\text{H}_1 : \beta_3 > 0.5$. The covariate corresponding to β_3 is an indicator, which may be the characteristic of treatment and control in practice. Thus, such a test is of practical interest. A standard test statistic is the t-statistic, obtained by dividing the difference between the estimator

Benchmarking mixed models

Table 2: **Simulated Mean, Variance and MSE: Scenario (2)**

Parameter	True Value	Mean AMLE	Mean BMLE	Variance AMLE	Variance BMLE	MSE AMLE	MSE BMLE
β_0	0.20	0.164	0.172	0.497	0.369	0.498	0.370
β_1	0.10	0.139	0.132	0.511	0.360	0.512	0.361
β_2	0.20	0.235	0.230	0.354	0.343	0.355	0.344
β_3	0.50	0.536	0.506	0.685	0.470	0.686	0.470
β_4	0.50	0.492	0.499	0.269	0.199	0.269	0.199
σ^2	0.25	0.177	0.189	0.153	0.146	0.158	0.149
τ^2	0.09	0.100	0.101	0.065	0.066	0.065	0.066

Table 3: **Size and Power of (Large Sample) t-Test**

Scenario	Size AMLE	Size BMLE	Power AMLE	Power BMLE
(1)	0.066	0.056	0.846	0.832
(2)	0.086	0.050	0.630	0.562

and value of the parameter under H_0 by the standard error (SE) of the estimator. The SE is the square root of the corresponding diagonal element of the estimated asymptotic covariance matrix (EACM). For both AMLE and BMLE, the EACM is in the form of (18), except that the L_i's are different. The L_is for BMLE is given by (15), while the L_i for AMLE is given by (15) without the first term (the one multiplied by λ). The t-statistic is then compared with the asymptotic critical value of the standard normal distribution. As the test is appropriate under large sample, we increase the number of subjects to $m = 50$. We consider the size and power of the t-test under the level of significance $\alpha = 0.05$. The power is considered at the alternative $\beta_3 = 1.0$, with the rest of the parameters unchanged. It appears that the BMLE t-test is more accurate in size, although its power is somewhat lower at the alternative. It may be argued, however, that the higher power of the AMLE t-test is due to its overrejecting. To see this, we adjust the critical values of the AMLE t-test so that it has the same sizes as the BMLE t-test, that is, 0.056 under Scenario (1) and 0.050 under Scenario (2). The adjusted critical values are 1.754 and 1.977, respectively (as compared to the nominal critical value of 1.645). With the adjusted critical values, the simulated powers of the AMLE t-test drop to 0.814 and 0.516, under (1) and (2), which are both lower than the simulated powers of the BMLE t-test.

Also note that both tests have higher power under Scenario (1) than under Scenario (2); and the difference between the two tests is more significant under Scenario (2) than under Scenario (1). These are, of course, reasonable.

Research is underway to implement our method in real-data applications.

Appendix

A.1 That conditions (i) and (ii) of section 1 imply (4)

We have $\mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\bar{y}^{\mathrm{a}}_{\cdot t}) = \mathrm{E}\{\mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\bar{y}^{\mathrm{a}}_{\cdot t}|M)\} = \mathrm{E}\{|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\mathrm{E}(\bar{y}^{\mathrm{a}}_{\cdot t}|M)\}$, and, similarly, for $t\in T_{\mathrm{a}}$ we have $\mathrm{E}(\bar{y}^{\mathrm{a}}_{\cdot t}|M) = m_t^{-1}\sum_{i\in I_t}\mathrm{E}(y_{it}|M) = \mu_{t0}$, where $\mu_{t0} = \mathrm{E}(y_{it}) = \mathrm{E}(y_{it}|M)$. Thus, we have $\mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\bar{y}^{\mathrm{a}}_{\cdot t}) = \mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\mu_{t0})$. On the other hand, note that $m_t \neq 0$ implies $m_{t,0} > 0$, thus, by a similar argument, it can be shown that $\mathrm{E}(\bar{y}_{\cdot t}|M) = \mu_{t0}$; hence, similarly, $\mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\bar{y}_{\cdot t}) = \mathrm{E}(|T_{\mathrm{a}}|^{-1}\sum_{t\in T_{\mathrm{a}}}\mu_{t0})$. By the same arguments, it can be shown that the rest of the equations in (4) hold.

A.2 Partial derivatives

Recall that L_i is defined by (15). We have

$$\begin{aligned}
\frac{\partial L_i}{\partial\beta} &= \lambda\left[\sum_{t\in T_{\mathrm{a}}} m_{t,k}^{-1}1_{(i\in I_{t,k})}x_{itk}\right]_{1\le k\le p} + 1_{(i\in I)}\tilde{X}_i'\tilde{V}_i^{-1}(\tilde{y}_i - \tilde{X}_i\beta),\\
\frac{\partial L_i}{\partial\psi_s} &= -\frac{1}{2}1_{(i\in I)}\left\{\mathrm{tr}\left(\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\right) - (\tilde{y}_i - \tilde{X}_i\beta)'\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\tilde{V}_i^{-1}(\tilde{y}_i - \tilde{X}_i\beta)\right\},\ 1\le s\le q,\\
\frac{\partial L_i}{\partial\lambda} &= \sum_{t\in T_{\mathrm{a}}}\left\{\sum_{k=1}^{p}\beta_k m_{t,k}^{-1}1_{(i\in I_{t,k})}x_{itk} - m_{t,0}^{-1}1_{(i\in I_{t,0})}y_{it}\right\}.
\end{aligned}$$

By the above expressions and assumptions (i) and (ii), we have

$$\mathrm{E}\left(\left.\frac{\partial L_i}{\partial\beta}\right|M\right) = \lambda\left[\sum_{t\in T_{\mathrm{a}}} m_{t,k}^{-1}1_{(i\in I_{t,k})}\mu_{tk}\right]_{1\le k\le p},$$

where $\mu_{tk} = \mathrm{E}(x_{itk})$. It follows that

$$\sum_{i=1}^{m}\mathrm{E}\left(\left.\frac{\partial L_i}{\partial\beta}\right|M\right) = \lambda\left[\sum_{t\in T_{\mathrm{a}}}\mu_{tk}\right]_{1\le k\le p}. \qquad (A.1)$$

Similarly, it can be shown that

$$\mathrm{E}\left(\left.\frac{\partial L_i}{\partial\psi}\right|M\right) = 0,\quad 1\le i\le m, \qquad (A.2)$$

$$\sum_{i=1}^{m}\mathrm{E}\left(\left.\frac{\partial L_i}{\partial\lambda}\right|M\right) = \sum_{t\in T_{\mathrm{a}}}\left\{\sum_{k=1}^{p}\beta_k\mu_{tk} - \mu_{t0}\right\} \qquad (A.3)$$

[recall $\mu_{t0} = \mathrm{E}(y_{it})$].

Benchmarking mixed models

Furthermore, we have

$$
\begin{aligned}
\frac{\partial^2 L_i}{\partial\beta\partial\beta'} &= -1_{(i\in I)}\tilde{X}_i'\tilde{V}_i^{-1}\tilde{X}_i, \\
\frac{\partial^2 L_i}{\partial\beta\partial\psi_s} &= -1_{(i\in I)}\tilde{X}_i'\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\tilde{V}_i^{-1}(\tilde{y}_i-\tilde{X}_i\beta), \quad 1\le s\le q, \\
\frac{\partial^2 L_i}{\partial\beta\partial\lambda} &= \left[\sum_{t\in T_{\mathrm{a}}} m_{t,k}^{-1}1_{(i\in I_{t,k})}x_{itk}\right]_{1\le k\le p}, \\
\frac{\partial^2 L_i}{\partial\psi_r\partial\psi_s} &= -\frac{1}{2}1_{(i\in I)}\left\{2(\tilde{y}_i-\tilde{X}_i\beta)'\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_r}\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\tilde{V}_i^{-1}(\tilde{y}_i-\tilde{X}_i\beta)\right. \\
&\quad -(\tilde{y}_i-\tilde{X}_i\beta)'\tilde{V}_i^{-1}\frac{\partial^2\tilde{V}_i}{\partial\psi_r\partial\psi_s}\tilde{V}_i^{-1}(\tilde{y}_i-\tilde{X}_i\beta)-\mathrm{tr}\left(\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_r}\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\right) \\
&\quad \left.+\mathrm{tr}\left(\tilde{V}_i^{-1}\frac{\partial^2\tilde{V}_i}{\partial\psi_r\partial\psi_s}\right)\right\}, \quad 1\le r,s\le q, \\
\frac{\partial^2 L_i}{\partial\psi\partial\lambda} &= 0, \quad \text{and} \\
\frac{\partial^2 L_i}{\partial\lambda^2} &= 0.
\end{aligned}
$$

The expressions, the linear mixed model, and the assumptions imply

$$
\mathrm{E}\left(\frac{\partial^2 L}{\partial\theta\partial\theta'}\right) = -\begin{pmatrix} A & 0 & C \\ 0 & B & 0 \\ C' & 0 & 0 \end{pmatrix}, \qquad (A.4)
$$

where $A=\mathrm{E}(\sum_{i\in I}\tilde{X}_i'\tilde{V}_i^{-1}\tilde{X}_i)$, $C=-\mathrm{E}(\sum_{t\in T_{\mathrm{a}}}\mu_{tk})_{1\le k\le p}$ and

$$
B = \frac{1}{2}\left[\mathrm{E}\left\{\sum_{i\in I}\mathrm{tr}\left(\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_r}\tilde{V}_i^{-1}\frac{\partial\tilde{V}_i}{\partial\psi_s}\right)\right\}\right]_{1\le r,s\le q}.
$$

Note that A, B are positive definite matrics in typical situations. Furthermore, if the linear mixed model includes an intercept, which means that $x_{it1}=1$ and therefore $\mu_{t1}=1$, and there is a positive probability that $T_{\mathrm{a}}\ne\emptyset$, then the vector C is guaranteed not to be a zero vector.

A.3 That the right side of (17) can be approximated by (18)

First, we have $\mathrm{Var}(\partial L/\partial\theta)=\mathrm{Var}\{\mathrm{E}(\partial L/\partial\theta|M)\}+\mathrm{E}\{\mathrm{Var}(\partial L/\partial\theta|M)\}$. By (A.1)—(A.3) and assumptions (ii) and (iv), we see that $\mathrm{E}(\partial L/\partial\theta|M)$ is a bounded random vector. It follows that $\mathrm{Var}\{\mathrm{E}(\partial L/\partial\theta|M)\}=O(1)$. On the other hand, assumptions (i) and (iii) imply that, given M,

$\partial L_i/\partial\theta, 1 \le i \le m$ are independent. Thus, we have

$$\begin{aligned}\mathrm{Var}\left(\frac{\partial L}{\partial\theta}\bigg| M\right) &= \mathrm{Var}\left(\sum_{i=1}^m \frac{\partial L_i}{\partial\theta}\bigg| M\right)\\ &= \sum_{i=1}^m \mathrm{Var}\left(\frac{\partial L_i}{\partial\theta}\bigg| M\right)\\ &= \sum_{i=1}^m \mathrm{E}\left(\frac{\partial L_i}{\partial\theta}\frac{\partial L_i}{\partial\theta'}\bigg| M\right) + \sum_{i=1}^m \mathrm{E}\left(\frac{\partial L_i}{\partial\theta}\bigg| M\right)\mathrm{E}\left(\frac{\partial L_i}{\partial\theta'}\bigg| M\right)\\ &= S_1 + S_2.\end{aligned}$$

Now, again by the results of subsection A.2, it can be shown that

$$S_2 = \left(\begin{array}{ccc} \lambda^2\sum_{i=1}^m \xi_i\xi_i' & 0 & \lambda\sum_{i=1}^m \eta_i\xi_i' \\ 0 & 0 & 0 \\ \lambda\sum_{i=1}^m \eta_i\xi_i' & 0 & \sum_{i=1}^m \eta_i^2 \end{array}\right),$$

where $\xi_i = [\sum_{t\in T_\mathrm{a}} m_{t,k}^{-1}1_{(i\in I_{t,k})}\mu_{tk}]_{1\le k\le p}$ and

$$\eta_i = \sum_{t\in T_\mathrm{a}}\left\{\sum_{k=1}^p \beta_k m_{t,k}^{-1}1_{(i\in I_{t,k})}\mu_{tk} - m_{t,0}^{-1}1_{(i\in I_{t,0})}\mu_{t0}\right\}.$$

It is now easy to show that the elements of S_2 are bounded random variables. For example, we have

$$\begin{aligned}\sum_{i=1}^m \eta_i^2 &= \sum_{t,t'\in T_\mathrm{a}}\sum_{k,l=1}^p \beta_k\beta_l m_{t,k}^{-1}m_{t,l}^{-1}m_{t,l}^{-1}\mu_{tk}\mu_{tl}|I_{t,k}\cap I_{t,l}|\\ &\quad -2\sum_{t,t'\in T_\mathrm{a}}\sum_{k=1}^p \beta_k m_{t,k}^{-1}m_{t,0}^{-1}\mu_{tk}\mu_{t0}|I_{t,k}\cap I_{t,0}|\\ &\quad +\sum_{t,t'\in T_\mathrm{a}} m_{t,0}^{-1}\mu_{t0}^2,\end{aligned}$$

which is a bounded random variable by assumptions (i) and (iv). Note that $|I_{t,k}\cap I_{t,l}| \le m_{t,k}\wedge m_{t,l}$. Therefore, combining the results, we have $\mathrm{Var}(\partial L/\partial\theta) = \mathrm{E}(S_1)+O(1) = \sum_{i=1}^m \mathrm{E}(p_ip_i')+O(1) = \mathrm{E}(\sum_{i=1}^m p_ip_i') + O(1) \approx \sum_{i=1}^m \hat{p}_i\hat{p}_i'$, where $p_i = \partial L_i/\partial\theta$.

Similarly, we have $\mathrm{E}(\partial^2 L/\partial\theta\partial\theta') = \mathrm{E}(\sum_{i=1}^m Q_i) \approx \sum_{i=1}^m \hat{Q}_i$, where $Q_i = \partial^2 L_i/\partial\theta\partial\theta'$.

References

[1] Datta, G. S. & Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statist. Sinica* 10, 613–627.

[2] Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1996). *Analysis of Longitudinal Data*. Oxford Univ. Press.

[3] Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer, New York.

[4] Laird, N. M. & Ware, J. M. (1982), Random effects models for longitudinal data, *Biometrics* 38, 963–974.

[5] Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd ed., Wiley, Hoboken, NJ.

[6] White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrika* 50, 1-25.