Journal of Statistical Research *https://doi.org/10.47302/jsr.2021550112*
2021, Vol. 55, No. 1, pp. 175-186 *ISSN 0256 - 422 X*

A MULTIPLE IMPUTATION METHOD FOR NONLINEAR MIXED EFFECTS MODELS WITH MISSING DATA

XINZHE DONG
Department of Statistics, University of British Columbia
Vancouver, BC, V6T 1Z2, Canada
Email: dong.hannah@hotmail.com

LANG WU*
Department of Statistics, University of British Columbia
Vancouver, BC, V6T 1Z2, Canada
Email: lang@stat.ubc.ca

SUMMARY

Multiple imputation methods are widely used in practice for missing data. An important consideration for a multiple imputation method is the choice of an imputation model which generates the imputations for each missing value, especially when the missing rate is not low. Mixed effects models are commonly used for modelling longitudinal data which exhibit large between-individual variations. In this case, a good imputation model should generate imputations at the *individual level* to incorporate the large between-individual variations. In this article, we propose a multiple imputation method for nonlinear mixed effects models with missing responses. We consider an iterative linearization method where the imputations are generated based on a "working" linear mixed effects model. We evaluate the proposed method via simulations and apply the method to a real dataset.

Keywords and phrases: Imputation, linearization, longitudinal data, multi-level data.

1 Introduction

Missing data arise frequently in longitudinal studies. Sometimes we need to impute the missing values, especially when the missing data are not missing completely at random (Rubin, 1976). For example, if a longitudinal variable is used as a time-dependent covariate in a Cox proportional hazards model for survival analysis, the missing values in the time-dependent covariate need to be imputed at event times. Multiple imputation methods are widely used in practice to impute missing data and incorporate missing data uncertainty (e.g., Rubin, 1996; Murray, 2018). For a multiple imputation method, each missing value is imputed by several plausible predicted values based on an imputation model, leading to several "complete datasets". Each "complete dataset" is analyzed using a method for complete data. The results from the complete-data analyses are then combined to form

* Corresponding author

an overall conclusion. A main advantage of a multiple imputation method is that the missing data uncertainty is incorporated. Moreover, the imputed complete datasets may be analyzed by different data analysts in various ways using existing software.

Mixed effects models are popular in the analysis of longitudinal data, especially when the between-individual variations are large. Here we focus on nonlinear mixed effects (NLME) models, which include linear mixed effects (LME) models as special cases. NLME models are often mechanistic or scientific models in the sense that they are usually derived based on the underlying data-generation processes. These models have been used for modelling some important longitudinal processes, such as studies of growth and decay, HIV viral dynamics, and pharmacokinetics analysis (e.g., Davidian and Giltinan, 1995; Wu, 2009). In this article, we consider a multiple imputation method for missing data in NLME models. Wu and Wu (2002) considered a multiple imputation method for missing covariates in NLME models. Here we consider a multiple imputation method for missing responses in NLME models. The method can be useful when the response of the NLME model is used as a time-dependent covariates in a survival model or in another longitudinal model such as a generalized linear mixed model. We assume that the missing data are missing at random.

There is an extensive literature on missing data and multiple imputation methods (e.g., Rubin, 1976, 1996; Little and Rubin, 2002; Sinha et al., 2014; Murray, 2018). A key step for a multiple imputation method is to build a good imputation model which is used to generate plausible imputed values for each missing data. Mixed effects models are usually used for modelling longitudinal data with large between-individual variations. In this case, a desirable imputation model should generate imputations at the *individual level*, rather than the population level, since imputations generated from a population-average model do not reflect the large variations between individuals. In this article, we propose a multiple imputation method which generates imputations at the individual levels for missing responses in an NLME model. The basic idea of the proposed method is first to linearize the NLME model in a way similar to that of Lindstrom and Bates (1990), and then we generate multiple imputations based on the resulting working linear mixed effects (LME) model from the linearization. Parameter estimates are also obtained based on this working LME model. This process is iterated until convergence. At convergence, we combine the final estimates and their standard errors using standard formulas for multiple imputations.

This article is organized as follows. In Section 2, we describe the proposed method in details. In Section 3, we evaluate the proposed method via simulations. A real dataset is analyzed in Section 4. We conclude the article with some discussions in Section 5.

2 A Multiple Imputation Method for NLME Models with Missing Data

2.1 NLME models

NLME models are extensions of LME models to nonlinear regressions for modelling longitudinal data. Random effects are introduced in the nonlinear regression models to incorporate between-individual variations and within-individual correlations. The nonlinear regressions are usually derived based on understandings of the underlying scientific or biological processes in a given application. For example, the viral load trajectories during an anti-HIV treatment can be modelled by two-compartment exponentially decay models based on some biological arguments represented by a set of differential equations to describe the virus production and elimination process (Wu and Ding, 1999). We describe a general NLME model as follows.

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be n_i repeated measurements of the response y for individual i, $i = 1, 2, \ldots, r$. A general NLME model can be written as follows

$$y_{ij} = g(t_{ij}, \boldsymbol{\beta}_{ij}) + e_{ij}, \tag{2.1}$$

$$\boldsymbol{\beta}_{ij} = h(\mathbf{x}_{ij}, \boldsymbol{\beta}, \mathbf{b}_i), \quad i = 1, 2, \ldots, r, \;\; j = 1, 2, \ldots, n_i, \tag{2.2}$$

$$\mathbf{b}_i \sim N(0, D), \qquad \mathbf{e}_i \sim N(0, R_i), \tag{2.3}$$

where $g(\cdot)$ is a known nonlinear function, $h(\cdot)$ is usually a linear function, $\boldsymbol{\beta}_{ij}$ and $\boldsymbol{\beta}$ are individual-specific time-varying parameters and fixed-effects parameters respectively, $\mathbf{x}_{ij}$ contains possibly time-varying covariates for individual i, R_i is a covariance matrix for the repeated observations within individual i, D is a unstructured covariance matrix for the random effects $\mathbf{b}_i$, $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})^T$ are random errors for observations within individual i, and $\mathbf{b}_i$'s are random effects. We assume that $\mathbf{e}_i$ and $\mathbf{b}_i$ are independent. For simplicity, we choose $R_i = \sigma^2 I_{n_i}$, where I_{n_i} is the identity matrix, i.e., the within-individual repeated measurements are assumed to be conditionally independent given the random effects. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, D)$ denote all parameters.

Parameter estimation and inference for a NLME model is usually based on the likelihood method. The likelihood is given by

$$L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{r} \int f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma, \mathbf{b}_i) f(\mathbf{b}_i|D)\, d\mathbf{b}_i, \tag{2.4}$$

where the unobservable random effects are integrated out. For a NLME model, the likelihood (2.4) usually does not have analytic or closed form expressions, since the NLME model is *nonlinear* in the unobserved random effects $\mathbf{b}_i$. Monte Carlo or stochastic Expectation-Maximization (EM) algorithms have been proposed, but they are often computationally intensive (Wu, 2009). A commonly used and computationally efficient approach is to use the linearization method of Lindstrom and Bates (1990), which is implemented in the R package `nlme` and `lme4`.

We can rewrite NLME model (2.1) and (2.2) as a single equation

$$y_{ij} = g\big(t_{ij}, h(\mathbf{x}_{ij}, \boldsymbol{\beta}, \mathbf{b}_i)\big) + e_{ij} \equiv u_{ij}(\mathbf{x}_{ij}, \boldsymbol{\beta}, \mathbf{b}_i) + e_{ij}, \quad i = 1, \ldots, r, \; j = 1, \ldots, n_i, \tag{2.5}$$

where $u_{ij}(\cdot)$ is a nonlinear function. Let $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})^T$. Beginning with some starting values, the linearization method of Lindstrom and Bates (1990) iterates the following steps until convergence. At each iteration, denote the current estimates of $(\boldsymbol{\beta}, \mathbf{b}_i)$ by $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i)$, suppressing the iteration number, where $\widehat{\mathbf{b}}_i$ is the empirical Bayesian estimate of $\mathbf{b}_i$. The procedure of Lindstrom and Bates (1990) is equivalent to *iteratively* solving the following "working" LME model

$$\widetilde{\mathbf{y}}_i = W_i \boldsymbol{\beta} + T_i \mathbf{b}_i + \mathbf{e}_i, \tag{2.6}$$

where

$$\begin{aligned}\widetilde{\mathbf{y}}_i &= \mathbf{y}_i - \mathbf{u}_i\big(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i\big) + W_i\widehat{\boldsymbol{\beta}} + T_i\widehat{\mathbf{b}}_i,\\ W_i &= \frac{\partial \mathbf{u}_i\big(\mathbf{x}_i, \boldsymbol{\beta}, \widehat{\mathbf{b}}_i\big)}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}, \qquad T_i = \frac{\partial \mathbf{u}_i\big(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \mathbf{b}_i\big)}{\partial \mathbf{b}_i^T}\Big|_{\mathbf{b}_i=\widehat{\mathbf{b}}_i}.\end{aligned}$$

At each iteration we obtain the updated estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i)$ of the parameters and random effects from the working LME model (2.6) using standard methods (Laird and Ware, 1982) and then proceed with next iteration until convergence.

2.2 A multiple imputation method

In this section, we propose a multiple imputation method for missing responses in NLME models when the missing data are missing at random in the sense of Rubin (1976). The method may be useful when the response of the NLME model is used as a time-dependent covariate in a survival model in a joint model setting, since in this case the missing response values may need to be imputed at event times. The method may also be used when the response values are (left) censored, such as viral load values below a detection limit in HIV studies.

We can write the response vector for individual i as $\mathbf{y}_i = (\mathbf{y}_{obs,i}, \mathbf{y}_{mis,i})$, where $\mathbf{y}_{obs,i}$ and $\mathbf{y}_{mis,i}$ are the observed components and missing components of $\mathbf{y}_i$ respectively and y_{ij} is the response value at time t_{ij}. The basic idea of the proposed method is to first linearize the NLME model using the first-order Taylor approximation about the estimated parameters and random effects, then use a multiple imputation method to impute the missing data based on the working LME model, and finally iterate the procedure until convergences. For the working LME model, we can use existing software for multiple imputations, such as the `mice` package (van Buuren and Groothuis-Oudshoorn, 2011) in `R`.

Specifically, we first choose starting values for the unknown parameters $\boldsymbol{\beta}$ and random effects $\mathbf{b}_i$, denoted by $\boldsymbol{\beta}^{(0)}$ and $\mathbf{b}_i^{(0)}$. For example, we may choose $\boldsymbol{\beta}^{(0)}$ and $\mathbf{b}_i^{(0)}$ to be the estimates based on complete data. At iteration k, $(k = 1, 2, 3, \ldots)$, we proceed with the following steps.

Step 1. Take a first-order Taylor expansion of the NLME model (2.5) about the current estimates of the parameters and random effects $\boldsymbol{\beta}^{(k)}$ and $\mathbf{b}_i^{(k)}$ respectively, and obtain the following working LME model,

$$\widetilde{\mathbf{y}}_i^{(k)} = W_i^{(k)}\boldsymbol{\beta} + T_i^{(k)}\mathbf{b}_i + \mathbf{e}_i, \tag{2.7}$$

where

$$\begin{aligned}\widetilde{\mathbf{y}}_i^{(k)} &= \mathbf{y}_i - \mathbf{u}_i\big(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\mathbf{b}}_i^{(k)}\big) + W_i\widehat{\boldsymbol{\beta}}^{(k)} + T_i\widehat{\mathbf{b}}_i^{(k)},\\ W_i^{(k)} &= \frac{\partial \mathbf{u}_i\big(\mathbf{x}_i, \boldsymbol{\beta}, \widehat{\mathbf{b}}_i\big)}{\partial \boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(k)}}, \qquad T_i^{(k)} = \frac{\partial \mathbf{u}_i\big(\mathbf{x}_i, \widehat{\boldsymbol{\beta}}, \mathbf{b}_i\big)}{\partial \mathbf{b}_i^T}\Big|_{\mathbf{b}_i=\widehat{\mathbf{b}}_i^{(k)}}.\end{aligned}$$

Suppose that y_{ij} is missing, then $\widetilde{y}_{ij}^{(k)}$ is also missing. Let $\widetilde{\mathbf{y}}_i^{(k)} = (\widetilde{y}_{i1}^{(k)}, \ldots, \widetilde{y}_{in_i}^{(k)})$ where some $\widetilde{y}_{ij}^{(k)}$'s may be missing.

Step 2. We generate multiple imputations for the missing y_{ij}'s based on the following imputation model

$$\widetilde{\mathbf{y}}_i^{(k)} = U_i\boldsymbol{\alpha}_1 + W_i^{(k)}\boldsymbol{\alpha}_2 + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, r, \tag{2.8}$$

where U_i is a design matrix containing information about the missing $\widetilde{y}_{ij}^{(k)}$'s, such as the fitted slopes of individual i based on observed data, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are unknown parameter vectors, and ϵ_i is a vector of random errors. We may also use the working LME model (2.7) as the imputation model. However, model (2.8) can be more general than model (2.7) in the sense that it can contain additional information useful for creating imputations, and model (2.7) may be considered as a special case of model (2.8). As noted in Little and Rubin (2002), it is desirable to create multiple imputations based on a more general imputation model than the data analysis model. For each missing $\widetilde{y}_{ij}^{(k)}$, we impute m values (say, $m = 5$), so we obtain m "complete datasets".

Step 3. For each "complete dataset" from Step 2, we fit the working LME model (2.7) in Step 1 and obtain updated parameter and random effects estimates $\boldsymbol{\beta}^{(k+1)}$ and $\mathbf{b}_i^{(k+1)}$. The m estimates are then combined by simply taking averages of the m estimates. For simplicity, we still denote the *combined estimates* by $\boldsymbol{\beta}^{(k+1)}$ and $\mathbf{b}_i^{(k+1)}$. Then, we go back to Step 1 for next iteration.

Iterating Steps 1 to 3 until convergence, we obtain a sequence of (combined) estimates $\{\boldsymbol{\beta}^{(k)},\ k = 1, 2, \ldots\}$. We may claim convergence when two consecutive estimates are close, say $|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}|$ is very small. Suppose that $\hat{\boldsymbol{\beta}}$ is the (combined) estimate from the last iteration at convergence. The standard error of $\hat{\boldsymbol{\beta}}$ can be obtained using standard formula for multiple imputation method (Little and Rubin, 2002), i.e.,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (1 + \frac{1}{m})\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\boldsymbol{\beta}}^{(i)} - \hat{\boldsymbol{\beta}}\right)^2 + \frac{1}{m}\sum_{i=1}^{m}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}^{(i)}\right) \tag{2.9}$$

and $SE(\hat{\boldsymbol{\beta}}) = \sqrt{\mathrm{Var}(\hat{\boldsymbol{\beta}})}$.

3 Simulation Study

In this section, we conduct a simulation study to evaluate the performance of the proposed multiple imputation method, and compare it with the naive complete-case (CC) method which simply deletes all incomplete observations.

We choose sample size $r = 50$, and let each individual has $n_i = 10$ (or $n_i = 20$) repeated measurements over time. When $n_i = 10$, the measurement times are set to be $\{0, 1, 2, 3, 5, 9, 18, 35, 50, 70\}$. When $n_i = 20$, the measurement times are set to be $\{0, 1, 2, 3, 5, 7, 9, 12, 14, 19, 21, 25, 30, 35, 40, 46, 51, 57, 65, 70\}$. We choose the models and its parameter values to be similar to that in the data analysis application presented in next section. That is, we consider the following NLME model

$$y_{ij} = \log_{10}\left(e^{P_{1i} - \lambda_{1ij}t_{ij}} + e^{P_{2i} - \lambda_{2i}t_{ij}}\right) + e_{ij}, \quad i = 1, 2, ..., r,\ j = 1, 2, ..., n_i, \tag{3.1}$$

$$P_{1i} = P_1 + b_{1i}, \quad \lambda_{1ij} = \lambda_1 + \beta x_{ij} + b_{2i}, \quad P_{2i} = P_2 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \tag{3.2}$$

where y_{ij} is the response value for patient i at time t_{ij}, $\boldsymbol{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$ are random effects, and e_{ij} is the random error. We assume that e_{ij} i.i.d. $\sim N(0, \sigma^2)$, and $\boldsymbol{b}_i \sim N(0, D)$. The true values of the fixed effects are set to be $(P_1,\ \lambda_1,\ \beta,\ P_2,\ \lambda_2) = (12, 0.3, 0.1, 7.5, 0.02)$, and the covariance matrix for the random effects is set to be

$$D = \begin{bmatrix} 0.636 & -1.418\times10^{-2} & 0.613 & 3.736\times10^{-3} \\ -0.014 & 6.601\times10^{-4} & -0.024 & -3.853\times10^{-5} \\ 0.613 & -2.385\times10^{-2} & 1.218 & 1.487\times10^{-2} \\ 0.004 & -3.853\times10^{-5} & 0.015 & 5.322\times10^{-4} \end{bmatrix}.$$

The value of σ will have two different values: 0.2 and 0.5.

The values of the time-varying covariate x_{ij} are generated from the following LME model:

$$x_{ij} = a_{1i} + a_{2i}\times t_{ij} + a_{3i}\times t_{ij}^2 + \varepsilon_{ij}, \quad i = 1, 2, ..., r,\ j = 1, 2, ..., n_i, \tag{3.3}$$

$$a_{1i} = a_1 + \alpha_{1i}, \quad a_{2i} = a_2 + \alpha_{2i}, \quad a_{3i} = a_3 + \alpha_{3i}, \tag{3.4}$$

where x_{ij} is the measured covariate for patient i at time t_{ij}, $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{in_i})^T$ are the measurement errors, $\boldsymbol{a} = (a_1, a_2, a_3)^T$ is the vector of fixed effects, and $\boldsymbol{\alpha_i} = (\alpha_{1i}, \alpha_{2i}, \alpha_{3i})^T$ is the vector of random effects. We assume that ε_{ij} i.i.d. $\sim N(0, \sigma_1^2)$, and $\boldsymbol{\alpha_i} \sim N(0, A)$. The true values of the fixed effects are set to be $\boldsymbol{a} = (1.37, 0.015, -0.00015)^T$, the value of σ_1 is set to be 0.15, and the covariance matrix for the random effects is set to be

$$A = \begin{bmatrix} 7.173\times10^{-2} & -1.028\times10^{-3} & 6.658\times10^{-6} \\ -1.028\times10^{-3} & 7.147\times10^{-5} & -6.916\times10^{-7} \\ 6.658\times10^{-6} & -6.916\times10^{-7} & 7.121\times10^{-9} \end{bmatrix}.$$

The missing data mechanism is assumed to be MAR: we assume that subjects with a smaller increase in covariate value during the first month will tend to have a larger chance of missing responses. The rationale is that, in the AIDS study, increasing value of CD4 indicates decreased inflammation.

We evaluate the methods by comparing the bias (in %) and mean square error (MSE) of the parameter estimates with respect to their true values, as well as the estimated coverage probabilities of the 95% confidence intervals of the true parameter values. Tables 1 and 2 show the simulation results for the selected cases. The other cases, including more repeated measurements and larger within-individual variations, are not presented here since the results are similar. All simulations are repeated 1000 times. The simulation results show that, compared with the complete case method, the multiple imputation method tends to estimate most parameters with a lower bias, smaller MSE, and better coverage rate. The performance of the multiple imputation method improves as the missing rate is higher, compared to the complete case method. Note that, the parameter λ_2 may be unstable or poorly estimated due to its small true values and large between-individual variations in the later period.

4 Real Data Example

In an AIDS study designed to evaluate an anti-HIV treatment, 45 HIV infected patients were treated with an antiviral regimen. Viral load (copies/mL) was repeatedly quantified for each patient in the

Table 1: Simulation results ($n_i = 10$, missing rate = 20%, $\sigma = 0.2$)

True Parameters		CC Method	MI Method
$P_1 = 12$	Estimate	12.018	12.002
	S.E.	0.123	0.124
	Bias (%)	0.152	0.017
	Coverage	0.940	0.933
	MSE	0.015	0.015
$\lambda_1 = 0.3$	Estimate	0.364	0.336
	S.E.	0.073	0.072
	Bias (%)	21.197	12.007
	Coverage	0.845	0.932
	MSE	0.009	0.006
$\beta = 0.1$	Estimate	0.058	0.070
	S.E.	0.048	0.046
	Bias (%)	-41.703	-30.496
	Coverage	0.831	0.933
	MSE	0.004	0.003
$P_2 = 7.5$	Estimate	7.503	7.432
	S.E.	0.186	0.194
	Bias (%)	0.045	-0.911
	Coverage	0.946	0.928
	MSE	0.035	0.042
$\lambda_2 = 0.02$	Estimate	0.020	0.018
	S.E.	0.004	0.004
	Bias (%)	-0.351	-8.010
	Coverage	0.945	0.925
	MSE	1.558×10^{-5}	1.979×10^{-5}

next three months after initiation of the treatment. Immunologic marker known as CD4 cell counts (cells/μL) was also measured along with viral load. Table 3 contains summary statistics of viral load measured in the first, second, and third month respectively. The dataset contains missing values in the viral load measurements, but the CD4 counts are all available. The viral load missing rate is approximately 15%. Most of the missing values occur during the third month. Since we do not know the reason for these missing viral load measurements, it might be inappropriate to assume MCAR, but a MAR may be more reasonable.

Table 2: Simulation results ($n_i = 10$, missing rate = 40%, $\sigma = 0.2$)

True Parameters		CC Method	MI Method
$P_1 = 12$	Estimate	12.028	12.002
	Sample S.E.	0.132	0.135
	Bias (%)	0.237	0.019
	Coverage	0.936	0.922
	MSE	0.018	0.018
$\lambda_1 = 0.3$	Estimate	0.376	0.301
	Sample S.E.	0.100	0.101
	Bias (%)	25.433	0.476
	Coverage	0.859	0.959
	MSE	0.016	0.010
$\beta = 0.1$	Estimate	0.053	0.086
	Sample S.E.	0.062	0.062
	Bias (%)	-47.390	-14.193
	Coverage	0.852	0.959
	MSE	0.006	0.004
$P_2 = 7.5$	Estimate	7.512	7.419
	Sample S.E.	0.209	0.229
	Bias (%)	0.166	-1.083
	Coverage	0.936	0.908
	MSE	0.044	0.059
$\lambda_2 = 0.02$	Estimate	0.020	0.018
	Sample S.E.	0.005	0.005
	Bias (%)	0.089	-10.809
	Coverage	0.936	0.881
	MSE	2.160×10^{-5}	3.010×10^{-5}

HIV viral dynamic models are useful to describe the virus elimination and production processes

Table 3: Summary statistics of viral loads (in $\log_{10}$ scale)

	n	Median	Q1-Q3	Missing data (%)
1^{st} month	246	3.998	3.376-4.826	14 (5.69%)
2^{nd} month	56	2.716	2.204-3.255	15 (26.79%)
3^{rd} month	62	2.230	1.699-2.857	25 (40.32%)

Note: n is the number of measurements; Q1 indicates the first quartile, and Q3 indicates the third quartile.

Table 4: Data analysis results

Parameter	MI method		CC method	
	Estimate	S.E.	Estimate	S.E.
P_1	11.699	0.193	11.740	0.200
P_2	7.346	0.265	7.724	0.295
λ_1	0.318	0.051	0.311	0.053
λ_2	0.016	0.005	0.023	0.006
β	0.007	0.019	0.023	0.021

Note: MI is multiple imputation; CC is complete case; S.E. is standard error.

during antiviral treatments (Ho et al., 1995; Perelson et al., 1996, 1997; Wu and Ding, 1999). These models provide good understanding of the parthenogenesis of HIV infection and evaluation of antiretroviral therapies. NLME models have been used in these studies to account for inter-patient and intra-patient variations in viral load measurements (Wu and Ding, 1999). We consider the following HIV viral dynamic model (Wu and Ding, 1999). Let y_{ij} be the $\log_{10}$-transformed viral load measurement for patient i at time t_{ij}. The following NLME model has been shown to model HIV viral dynamics well:

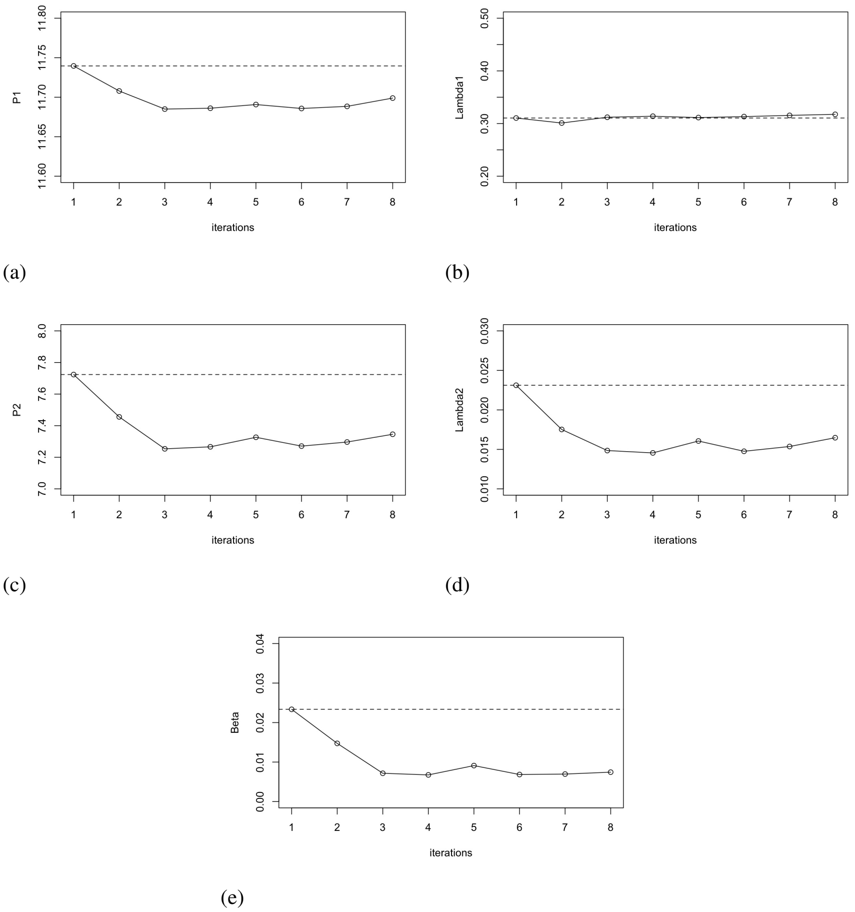
$$y_{ij} = \log_{10}\left(e^{P_{1i}-\lambda_{1ij}t_{ij}} + e^{P_{2i}-\lambda_{2i}t_{ij}}\right) + e_{ij}, \tag{4.1}$$

$$P_{1i} = P_1 + b_{1i}, \quad \lambda_{1ij} = \lambda_1 + \beta CD4_{ij} + b_{2i}, \tag{4.2}$$

$$P_{2i} = P_2 + b_{3i}, \quad \lambda_{2i} = \lambda_2 + b_{4i}, \quad i = 1, 2, ..., r,\ j = 1, 2, ..., n_i, \tag{4.3}$$

where $\boldsymbol{b}_i = (b_{1i}, b_{2i}, b_{3i}, b_{4i})^T$ are random effects, and e_{ij} is a random error. We assume that e_{ij} i.i.d. $\sim N(0, \sigma^2)$, $\boldsymbol{b}_i \sim N(0, D)$, and e_{ij} and $\mathbf{b}_i$ are independent. When viral loads y_{ij} are used as time-dependent covariates in a survival model, such as time to viral rebound, the missing viral loads at event times must be addressed. We consider two methods for comparison: the proposed multiple imputation method and the naive complete-case method.

Parameter estimates and standard errors obtained using the proposed multiple imputation method



(a) (b)

(c) (d)

(e)

Figure 1: Parameter estimates in each iteration of the multiple imputation method. Dashed lines indicate complete-case estimates (starting values).

for missing responses in NLME models and the complete case method are presented in Table 4. In Figure 1, the parameter estimates of the proposed multiple imputation method are plotted at each iteration, with the dashed lines indicating the complete-case estimates. It takes 8 iterations for the parameters to converge: the average absolute change of the five estimated parameters is less than 0.05 in two consecutive iterations. These results show that the multiple imputation method estimates all parameters with a smaller standard error compared with the complete case method, which indicates that the proposed method is more efficient than the complete case method and may lead to shorter confidence intervals for the parameters. Moreover, the complete case method seems over-estimating most parameters. Based on the simulation results presented in the last section, the results in Table 4 based on the multiple imputation method should be more reliable.

5 Conclusions and Discussion

Although there has been extensive research on missing data problems in the past few decades, research in this area is expected to remain active in the future. This is because missing data problems are very common in practice, so any new statistical models and methods may need to address missing data problems in practice. From a practical point of view, multiple imputation methods are perhaps most useful, while other missing data methods such as the EM algorithms and methods of weighting have the disadvantage of limited available software and their implementations are often restricted to specific models. Developments of new multiple imputation methods for models and methods useful in practice are important for these models and methods to be more widely used by applied statisticians.

We have proposed a multiple imputation method for missing responses of NLME models where the imputations are generated at individual levels. In principle, the method may be extended to missing responses of NLME models where the missing data mechanism may be non-ignorable. In this case, we may introduce a non-ignorable missing data model in the imputation model to generate imputations. The method may also be extended to missing data in other mixed effects models such as missing responses in generalized linear mixed models or mixed effects models with missing time-dependent covariates.

An advantage of the proposed method is that the imputations are generated at individual levels, which is desirable if the data exhibit large between-individual variations. Since nonlinear models are often mechanistic models in the sense that they are derived based on the underlying data generation mechanism, the proposed method should provide better "predictions" of the missing data than those based on empirical models such as linear mixed effects or nonparametric mixed effects models. A limitation of the proposed method is that theoretical properties of the method remain to be developed. While simulation results show its good performance under certain simulation settings, its performance in other settings needs to be investigated. Another limitation of the proposed method is that convergence is not guaranteed.

References

Davidian, M. and Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurements Data*, Chapman and Hall/CRC.

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995), "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection," *Nature*, 373, 123–126.

Laird, N. M. and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lindstrom, M. J. and Bates, Douglas, M. (1990), "Nonlinear Mixed Effects Models for Repeated Measures Data," *Biometrics*, 46, 673–687.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical Analysis with Missing Data, 2nd edition*, Wiley.

Murray, J. S. (2018), “Multiple Imputation: A Review of Practical and Theoretical Findings,” *Statistical Science*, 33, 142–159.

Perelson, A. S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M., and Ho, D. D. (1997), “Decay characteristics of HIV-1-infected compartments during combination therapy,” *Nature*, 387, 188–191.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996), “HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time,” *Science*, 271, 1582–1586.

Rubin, D. B. (1976), “Inference and Missing Data,” *Biometrika*, 63, 581–592.

— (1996), “Multiple Imputation After 18 Years,” *Journal of the American Statistical Association*, 91, 473–489.

Sinha, S. K., Kaushal, A., and Xiao, W. (2014), “Inference for longitudinal data with nonignorable nonmonotone missing responses,” *Computational Statistics & Data Analysis*, 72, 77–91.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011), “MICE: Multivariate Imputation by Chained Equations in R,” *Journal of Statistical Software*, 45, 1–67.

Wu, H. and Ding, A. A. (1999), “Population HIV-1 Dynamics In Vivo: Applicable Models and Inferential Tools for Virological Data from AIDS Clinical Trials,” *Biometrics*, 55, 410–418.

Wu, L. (2009), *Mixed effects models for complex data*, Chapman and Hall/CRC.

Wu, L. and Wu, H. (2002), “Missing Time-Dependent Covariates in Human Immunodeficiency Virus Dynamic Models,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 51, 297–318.

Received: March 13, 2021

Accepted: April 7, 2021