Statistica Sinica **23** (2013), 1043-1070
doi:http://dx.doi.org/10.5705/ss.2011.290

TESTS FOR RANDOM EFFECTS IN LINEAR MIXED MODELS USING MISSING DATA

Xianzheng Huang

University of South Carolina

Abstract: We propose novel methods to assess assumptions on random effects in linear mixed models. A key ingredient to the proposed methods involves creating missing data strategically from the observed data in order to detect multiple sources of misspecification on random effects. Random-effects assumptions traditionally tested separately by tests for variance components and tests for normality can now be assessed simultaneously using the proposed methods. The rationale underlying the new methods is applicable to other types of mixed effects models.

Key words and phrases: Ignorable missingness, missingness mechanism, model misspecification, nonignorable missingness, random effects.

1. Introduction

Mixed effects models are routinely used to model correlated data that naturally arise in a broad range of applications. Including random effects in a statistical model yields a practically meaningful and mathematically elegant way to characterize various correlation structures in data such as repeated measurements from longitudinal studies, spatially correlated data in geostatistics, and multivariate observations. In this article, we restrict attention to linear mixed models (LMM) for methodological development, but the underlying idea is applicable to other types of mixed effects models. Verbeke and Molenberghs (2000) give a comprehensive survey on application and inference based on LMM.

A major concern in drawing inference based on mixed effects models lies in the assumptions on random effects. In particular, one often faces two questions regarding random effects: (i) which predictor should have a random effect associated with it; (ii) whether or not the random effect normally distributed. Statisticians have developed diagnostic tests to address these questions, usually separately. To tackle the first problem, tests on variance components have been developed that are nonstandard testing problems in that the null space (zero variance) is on the boundary of the parameter space. Self and Liang (1987) and Stram and Lee (1994) used a mixture of chi-squares as the null distribution of the likelihood ratio when testing a variance component, which is shown to be the asymptotic null distribution of the test statistic. There has been empirical evidence suggesting that finite-sample performance of this test can be unsatisfactory. For more general data structures than that considered in these two articles, Crainiceanu and Ruppert (2004) proposed an algorithm to simulate the finite sample null distribution of the likelihood ratio based on spectral decomposition. This algorithm is developed for models with only one variance component. To allow multiple variance components, Crainiceanu (2008) discussed use of the parametric bootstrap and two other less computationally demanding algorithms to approximate the null distribution of the likelihood ratio. Saville and Herring (2009) proposed to use the Bayes factor to test a variance component. Underlying the development of the aforementioned tests is the normality of random effects. Operating characteristics of these tests when the normality assumption is violated are unclear.

Suppose random effects are appropriately allocated in the model and the attention is on the second question. Verbeke and Lesaffre (1994, 1997) showed that the maximum likelihood estimators (MLE) for the fixed effects in an LMM remain consistent and asymptotically normal even when the random-effects distribution is misspecified. They also pointed out that the sandwich variance estimator should be used in replace of the inverse Fisher information for more reliable standard error estimation. However, it was found that, when the true random-effects distribution deviates from normal, such as when the truth is a lognormal (Litière (2007)), it often requires an impractically large sample until the asymptotics kick in to produce a reasonably accurate sandwich variance estimator. Poor standard error estimation can compromise hypotheses testing and confidence intervals (Litière, Alonso, and Molenberghs (2007, 2008)). Moreover, the convergence rate of MLE can depend heavily on the shape of the true random-effects distribution. Finally, in studies such as surrogate marker evaluation and psychometric properties evaluation, distributional characteristics of random effects are the research focal points, and thus the second question is of direct interest.

Nonparametric and semiparametric approaches have been proposed to estimate random-effects distributions. Nonparametric approaches for random effects (Laird (1978)) rarely yield a continuous or smooth estimation. Verbeke and Lesaffre (1996) used mixtures of normals to model random effects, resulting in the so-called heterogeneity model. Zhang and Davidian (2001) proposed a flexible semiparametric distribution family for random effects. These non-/semiparametric approaches are useful when a random effect may not follow a familiar distribution. But before falling back on these potentially computationally intensive options, parsimonious parametric models for random effects are more appealing provided the validity of such models can be justified. Such justification calls for effective model diagnostic tools.

To the best of our knowledge, none of the existing methods tackle both problems simultaneously in a unified framework. The new methods presented in this article can achieve this. The idea stems from the results, shown in later sections, that in the presence of model misspecification, likelihood inference based on the observed data can disagree with the counterpart inference drawn from an induced incomplete data set. Realization of the discrepancy between these two sets of inferences depends on a user-designed mechanism according to which missing data are created. Statistical tests based on a cleverly designed missingness mechanism can solve both problems at the same time. The idea of using the discrepancy between two sets of inferences to detect model misspecification bares a slight resemblance with the specification tests for regression models in econometrics discussed in Hausman (1978), and the test for a certain type of generalized linear mixed model (GLMM) proposed by Tchetgen and Coull (2006). In both articles the authors compared two estimators for a parameter of interest, with one sensitive to the model assumption in question and the other robust to it. Test statistics were constructed based on the difference between these two estimators with adjustment for variation, with a significant difference indicating violation of model assumptions. Two pitfalls of these tests are in first, finding an estimator robust to model misspecification can be difficult depending on the model settings and assumptions, and second, deriving a variance estimator for the difference between two estimators can also be challenging. Tchetgen and Coull (2006) were able to overcome both difficulties for a special type of GLMM, but such luck does not carry over to more general contexts. Our proposed strategy circumvents both obstacles by comparing two (possibly biased) MLEs, one derived from the raw data and the other derived from the induced incomplete data. A variance estimator for the difference between them can be easily constructed based on influence functions associated with two MLEs. Hence, calculating the proposed test statistics only involves routine maximum likelihood calculations.

Verbeke and Molenberghs (2010) underscored the unverifiable nature of random-effects assumptions in a mixed effects model without assuming the other parts of the hierarchical modeling correct. We do not intend to test the unidentifiable part of a mixed effects model. Rather, we strive for more informative diagnostic procedures with fewer assumptions on the other parts of the model than those needed in the existing methods. Even when we cannot conclude whether or not a random-effects assumption is appropriate for the current data, we look to find out if the likelihood inference enjoys certain desirable robust features under the assumed model.

To set the stage for theoretical developments, we first formulate LMM for the observed data and the induced model for the incomplete data in Section 2. Test statistics used in the proposed diagnostic procedures are also defined in this section. In Section 3 we investigate different types of missingness mechanisms to facilitate testing for a random intercept and its distributional assumption. The parallel results for testing random slopes are derived in Section 4. In Section 5 we illustrate the implementation and performance of the proposed methods via simulation studies and application to a data set. Lastly, discussions on the implication of the underlying ideas, connections with existing methods, and future research are given in Section 6.

2. Models and Test Statistics

2.1. Models for complete data

Let $Y_i = (Y_{i1}, \ldots, Y_{in_i})^t$ be the ith observed response vector, for $i = 1, \ldots, m$. An LMM consists of two component models. The first component model is a conditional model of Y_i given the covariates, the fixed effects, and the random effects,

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{2.1}$$

where β is the p-dimensional fixed effects, b_i is the q-dimensional random effects, X_i and Z_i are the $n_i \times p$ and $n_i \times q$ design matrices for the fixed effects and random effects, respectively, and it is assumed throughout that $\epsilon_i \sim N(0_{n_i\times 1}, \Sigma_i)$, independent of b_i, with $n_i \times n_i$ variance-covariance matrix Σ_i that depends on i only through the dimension, but not through the parameters in the matrix. In order to focus on the two problems regarding random effects raised in Section 1, we assume that $\Sigma_i = \sigma_\epsilon^2 I_{n_i}$, where I_{n_i} is an $n_i \times n_i$ identity matrix and, for most of the article, we assume the fixed-effects part of (2.1) is correctly specified. The second component model is a model for b_i, with density $f_b(b_i; \Sigma_b)$, where Σ_b includes the variance components and other parameters in the random-effects model if needed. Let Ω be the $r \times 1$ parameter vector that includes all unknown parameters in the two component models. Combining the models, one has the likelihood of the ith observed response vector as $f_Y(Y_i; \Omega) = \int f_{Y|b}(Y_i|b_i; \Omega) f_b(b_i; \Sigma_b)\, db_i$.

2.2. Models for incomplete data

A common thread running through the article is missing data strategically created from the raw data. Incomplete data due to missingness is an example of coarsened data. Before we consider induced missingness per se, it is instructive to first introduce the generic notion of coarsened data. Let (Δ_i, Y_i^*) be the ith data vector in the coarsened data set, where Δ_i is the coarsening variable, $Y_i^* = C_{\Delta_i}(Y_i)$ is the coarsened response, and $C_{\Delta_i}(Y_i)$ is a many-to-one coarsening function that maps Y_i to Y_i^*. Designing a coarsening mechanism includes specifying a probability model for Δ_i indexed by parameters λ, $f_{\Delta|Y}(\Delta_i|Y_i; \lambda)$, and defining a coarsening function $C_{\Delta_i}(\cdot)$. Besides incomplete data due to missingness, examples of coarsened data include grouped data collected in pooled testing, and censored data considered in survival analysis. For the incomplete data of interest, Δ_i refers to missingness indicators. Denote by $Y_{\mathrm{obs},i}$ and $Y_{\mathrm{mis},i}$ the observed subvector and the missing subvector of Y_i, respectively, so $Y_i^* = Y_i$ if there is no missingness, and $Y_i^* = Y_{\mathrm{obs},i}$ otherwise.

In the missing data literature there is an important distinction between ignorable missingness and nonignorable missingness (Little and Rubin (2002, Sec. 15.1)). For ignorable missingness, Δ_i is independent of $Y_{\mathrm{mis},i}$. If Δ_i is also independent of $Y_{\mathrm{obs},i}$, then the data are missing completely at random (MCAR); otherwise, the data are called missing at random (MAR). When the missingness is ignorable, either MCAR or MAR, Ω and λ are distinct (Little and Rubin (2002, Def. 6.4)), which may not be true for nonignorable missingness. For nonignorable missingness, also referred to as "not missing at random" (NMAR), Δ_i depends on $Y_{\mathrm{mis},i}$. In the proposed methods, we specify the missingness mechanism and thus λ, known to us, is necessarily distinct from the unknown Ω. This eliminates the notorious issue of nonidentifiability in NMAR (Little and Rubin (2002, Chap. 15)).

The likelihood of (Δ_i, Y_i^*), denoted by $f_{\Delta, Y^*}(\Delta_i, Y_i^*; \Omega, \lambda)$, can be easily derived based on $f_{\Delta|Y}(\Delta_i|Y_i; \lambda)$ and $f_Y(Y_i; \Omega)$, where the former cannot be misspecified as it is user-designed, but the latter may be a wrong model for Y_i due to inadequate assumptions on random effects. Likelihood inference based on the two likelihood functions, $f_Y(Y_i; \Omega)$ and $f_{\Delta, Y^*}(\Delta_i, Y_i^*; \Omega, \lambda)$, leads to two sets of estimators for Ω. Denote by $\hat{\Omega}$ and $\hat{\Omega}^*$ the MLE for Ω based on the complete data and the estimator based on the incomplete data, respectively, and by $\tilde{\Omega}$ and $\tilde{\Omega}^*$ their limiting counterparts as m tends to infinity with $\max_{1\le i\le m} n_i$ bounded. These estimators are the basis of the test statistics we define for model diagnosis. The (limiting) MLE of a parameter in Ω is denoted following the same notational convention. For concise notations, the dependence of likelihood functions on the known λ is suppressed in the sequel.

2.3. Test statistics

Let θ be a parameter in Ω and consider testing the null hypothesis, $H_{0,\theta}$: $\tilde{\theta}^* = \tilde{\theta}$. Under regularity conditions, the MLEs are consistent and $H_{0,\theta}$ is true when all models are correctly specified. In the presence of model misspecification, we aim at designing a missingness mechanism so that $\tilde{\theta}^*$ differs from $\tilde{\theta}$. With a strategically designed missingness mechanism, violations on model assumptions can be revealed by the discrepancy between $\hat{\theta}^*$ and $\hat{\theta}$. In light of this reasoning, we define a test statistic that assesses the significance of the discrepancy between two estimators as

$$t_{1,\theta} = (\hat{\theta}^* - \hat{\theta})\hat{\nu}^{-1}, \tag{2.2}$$

where $\hat{\nu}$ is an estimator for the standard error of $\hat{\theta}^* - \hat{\theta}$, constructed based on influence functions as elaborated in Huang (2009). There it is shown that $t_{1,\theta}$ follows a t distribution with $m-r$ degrees of freedom asymptotically under $H_{0,\theta}$.

When missingness is severe, the numerical procedure used to obtain $\hat{\Omega}^*$ can be unstable. To avoid estimating Ω using incomplete data, another test statistic is constructed based on mismatching the estimator and the estimating equations as

$$t_2 = \left\{ m^{-1/2} \sum_{i=1}^{m} \psi(\Delta_i, Y_i^*; \hat{\Omega}) \right\} \odot \sqrt{\text{vecdiag}(\hat{V}_2^{-1})},$$

where $\psi(\Delta_i, Y_i^*; \Omega) = (\partial/\partial\Omega) \log f_{\Delta,\,Y^*}(\Delta_i, Y_i^*; \Omega)$ is the score function associated with the incomplete data, $\hat{V}_2$ is an estimator for the variance-covariance of $m^{-1/2} \sum_{i=1}^{m} \psi(\Delta_i, Y_i^*; \hat{\Omega})$, of which the derivation is given in Huang (2011), and $\odot$ is the elementwise multiplication operator. By the definition of $\tilde{\Omega}^*$, $\lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0\{\psi(\Delta_i, Y_i^*; \tilde{\Omega}^*)\} = 0$, where $E_0(\cdot)$ is the expectation with respect to the true model with Ω evaluated at its true value denoted by $\mho$. Consequently, if $\tilde{\Omega}^* = \tilde{\Omega}$, then $\lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0\{\psi(\Delta_i, Y_i^*; \tilde{\Omega})\} = 0$. If $t_{2,\theta}$ is the element in t_2 associated with θ, then $t_{2,\theta}$ is expected to be close to zero under $H_{0,\theta}$, and a significant deviation from zero signals violation of model assumptions. Huang, Stefanski, and Davidian (2009) showed that the second test statistic is asymptotically equivalent to the first one. Henceforth, t_θ is used to refer to either one of the two test statistics associated with θ when we do not distinguish these two.

With the test statistics constructed, the remaining task is to design a missingness mechanism that can separate $\tilde{\Omega}^*$ from $\tilde{\Omega}$ in the presence of model misspecification on the random intercept or/and random slopes in LMM. This task is tackled in the following two sections, where we compare $\tilde{\Omega}^*$ with $\tilde{\Omega}$ when ignorable missingness mechanisms and nonignorable missingness mechanisms are used, respectively, to generate missing data.

3. Testing for a Random Intercept

Consider a one-way analysis-of-variance model with m levels and n_i observations at level i, $M_2 : Y_{ij} = \mu + b_{i0} + \epsilon_{ij}$, and a null model without the random subject effect b_{i0}, $M_1 : Y_{ij} = \mu + \epsilon_{ij}$, where μ is the overall mean, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. Suppose M_1 is the assumed model and M_2 is the true model for the observed data.

Denote by $\mho = (\breve{\mu},\, v_0,\, v_\epsilon)^t$ the true parameter values under M_2, where $\breve{\mu} = E_0(Y_{ij})$, $v_0 = \text{Var}(b_{i0})$, and $v_\epsilon = \text{Var}(\epsilon_{ij})$. It is straightforward to show that $\tilde{\Omega} = (\breve{\mu},\, \tilde{\sigma}_\epsilon^2)^t$, where $\tilde{\sigma}_\epsilon^2 = v_\epsilon + v_0$. We next investigate likelihood inference based on incomplete data generated according to different missingness mechanisms.

The goal is to find a mechanism that leads to at least one element in $\tilde{\Omega}^*$ different from its counterpart in $\tilde{\Omega}$ when certain model assumptions in M_1 are violated.

To obtain a more tractable incomplete-data likelihood, we define a scalar missingness indicator, $\Delta_i = I(Y_i \text{ is not fully observed})$, where $I(\cdot)$ is the indicator function, with the missingness mechanism characterized by $P(\Delta_i = 1|Y_i)$. Moreover, when $\Delta_i = 1$, the index set, $\{j : 1 \leq j \leq n_i \text{ and } Y_{ij} \text{ is missing}\}$, is the same across $i = 1, \ldots, m$, and $Y_{\text{mis},i}$ is $s \times 1$, for $i = 1, \ldots, m$. In the next two subsections, we show that an ignorable missingness mechanism always yields $\tilde{\Omega}^* = \tilde{\Omega}$ regardless of the true model for b_{i0}, but it is not the case for a nonignorable missingness mechanism. More importantly, we show that using different nonignorable missingness mechanisms can reveal different sources of misspecification on b_{i0}. When it is not feasible to derive $\tilde{\Omega}^*$ or $\tilde{\Omega}$ explicitly, the strategy often used to compare $\tilde{\Omega}^*$ with $\tilde{\Omega}$ is to check if $\tilde{\Omega}$ also solves the score equations associated with the incomplete data. In the sequel, it is assumed that the root to a set of normal score equations is unique.

3.1. Ignorable missingness mechanisms

The essence of the proposition proved next is that an ignorable missingness mechanism does not interact with model misspecification on b_{i0} to result in $\tilde{\Omega}^* \neq \tilde{\Omega}$. Henceforth, "$A \perp B$" means that A is independent of B.

Proposition 1. *If $\Delta_i \perp Y_{\text{mis},i}$ for $i = 1, \ldots, m$, then $\tilde{\Omega}^* = \tilde{\Omega}$ regardless of the true model for b_{i0} in M_2.*

Proof. Let $l_i = \log f_Y(Y_i; \Omega)$, $l_{i1} = \log f_Y(Y_{\text{obs},i}; \Omega)$, and $l_{i2} = \log f_Y(Y_{\text{mis},i}; \Omega)$. Under M_1, because $\{Y_{ij}, j = 1, \ldots, n_i\}$ are independent and identically distributed (i.i.d.), $E_0(\partial l_{i1}/\partial\Omega) \propto E_0(\partial l_{i2}/\partial\Omega) \propto E_0(\partial l_i/\partial\Omega)$. By the definition of $\tilde{\Omega}$, $E_0(\partial l_i/\partial\Omega|_{\tilde{\Omega}}) = 0$, and thus

$$E_0\left(\frac{\partial l_{i1}}{\partial\Omega}\Big|_{\tilde{\Omega}}\right) = E_0\left(\frac{\partial l_{i2}}{\partial\Omega}\Big|_{\tilde{\Omega}}\right) = 0. \tag{3.1}$$

For the incomplete data, the likelihood under M_1 is given by

$$\begin{aligned} & f_{\Delta, Y^*}(\Delta_i, Y_i^*; \Omega) \\ & \quad = \{P(\Delta_i = 0|Y_{\text{obs},i}) f_Y(Y_i; \Omega)\}^{1-\Delta_i} \{P(\Delta_i = 1|Y_{\text{obs},i}) f_Y(Y_{\text{obs},i}; \Omega)\}^{\Delta_i}, \end{aligned}$$

and the corresponding log likelihood, denoted by l_i^*, is

$$\begin{aligned} \log f_{\Delta, Y^*}(\Delta_i, Y_i^*; \Omega) = & (1 - \Delta_i)\{\log P(\Delta_i = 0|Y_{\text{obs},i}) + l_i\} \\ & + \Delta_i \{\log P(\Delta_i = 1|Y_{\text{obs},i}) + l_{i1}\}. \end{aligned}$$

It follows that the score function associated with the incomplete data is

$$\frac{\partial l_i^*}{\partial \Omega} = (1-\Delta_i)\frac{\partial l_i}{\partial \Omega} + \Delta_i \frac{l_{i1}}{\partial \Omega} = \frac{\partial l_{i1}}{\partial \Omega} + (1-\Delta_i)\frac{\partial l_{i2}}{\partial \Omega}.$$

Because $\Delta_i \perp Y_{\text{mis},i}$ and by (3.1),

$$E_0\left(\frac{\partial l_i^*}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) = E_0\left(\frac{\partial l_{i1}}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) + E_0(1-\Delta_i)E_0\left(\frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) = 0.$$

Hence, $\tilde{\Omega}^* = \tilde{\Omega}$.

Proposition 1 suggests that neither MAR nor MCAR is useful for testing M_1 versus M_2, and this motivates us to consider NMAR for diagnostic purposes.

3.2. Nonignorable missingness mechanisms

Proposition 2. *If $P(\Delta_i = 1|Y_i) = P(\Delta_i = 1|Y_{\text{mis},i})$ for $i = 1,\ldots,m$, then $\tilde{\Omega}^* = \tilde{\Omega}$ if and only if*

$$E_0(\Delta_i)E\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}; \tilde{\Omega}\right) = E(\Delta_i; \tilde{\Omega})E_0\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}\right), \tag{3.2}$$

where $E(\cdot;\tilde{\Omega})$ is the expectation with respect to the assumed model with $\Omega = \tilde{\Omega}$.

Proof. The incomplete-data log likelihood is

$$\begin{aligned} l_i^* = {} & (1-\Delta_i)\left\{\log P(\Delta_i = 0|Y_{\text{mis},i}) + l_i\right\} \\ & +\Delta_i\left\{\log \int P(\Delta_i = 1|Y_{\text{mis},i}) f_{\scriptscriptstyle Y}(Y_{\text{mis},i};\Omega)dY_{\text{mis},i} + l_{i1}\right\}. \end{aligned}$$

Assuming interchangeability of differentiation and integration, the score function is

$$\frac{\partial l_i^*}{\partial \Omega} = (1-\Delta_i)\frac{\partial l_i}{\partial \Omega} + \Delta_i\left\{\frac{\partial l_{i1}}{\partial \Omega} + \frac{\int P(\Delta_i = 1|Y_{\text{mis},i})\frac{\partial}{\partial \Omega} f_{\scriptscriptstyle Y}(Y_{\text{mis},i};\Omega)dY_{\text{mis},i}}{\int P(\Delta_i = 1|Y_{\text{mis},i}) f_{\scriptscriptstyle Y}(Y_{\text{mis},i};\Omega)dY_{\text{mis},i}}\right\}.$$

Therefore,

$$E_0\left(\frac{\partial l_i^*}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) = -E_0\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) + E_0(\Delta_i)\frac{E\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}; \tilde{\Omega}\right)}{E(\Delta_i; \tilde{\Omega})}. \tag{3.3}$$

Setting (3.3) equal to zero gives the sufficient and necessary condition for $\tilde{\Omega}^* = \tilde{\Omega}$, as in (3.2).

Denote by $g_{Y}(\cdot;\Omega)$ the density of Y under the true model. A close examination of (3.2) reveals that a sufficient condition for $\tilde{\Omega}^{*}=\tilde{\Omega}$ to hold is

$$g_{Y}(Y_{\text{mis},i};\mho)=f_{Y}(Y_{\text{mis},i};\tilde{\Omega}). \tag{3.4}$$

As a case of particular interest, if $b_{i0}\sim N(0,\,\sigma_{b0}^{2})$ in M_2, then

(i) if $s=1$, then the densities in (3.4) are both of $N(\breve{\mu},\,\tilde{\sigma}_{\epsilon}^{2})$, and thus $\tilde{\Omega}^{*}=\tilde{\Omega}$;

(ii) if $s>1$, then (3.4) is violated because the correlation among the elements in $Y_{\text{mis},i}$ under the true model is ignored in $f_{Y}(Y_{\text{mis},i};\tilde{\Omega})$. In most practical situations, $\tilde{\sigma}_{\epsilon}^{2*}\neq\tilde{\sigma}_{\epsilon}^{2}$ is expected in this case.

On the other hand, if b_{i0} is not normally distributed, then (3.4) is not satisfied regardless of the size of s since $g_{Y}(Y_{\text{mis},i};\mho)$ is no longer a normal density but $f_{Y}(Y_{\text{mis},i};\tilde{\Omega})$ still is. Even though mathematically not impossible, it is expected in most practical situations that $\tilde{\Omega}^{*}$ does not coincide with $\tilde{\Omega}$ when (3.4) is not satisfied. Additionally, following a similar proof as above, one can show that if Δ_i also depends on $Y_{\text{obs},i}$ besides $Y_{\text{mis},i}$, then $\tilde{\Omega}^{*}\neq\tilde{\Omega}$ even when b_{i0} is normal and $s=1$.

Based on Proposition 2 and the follow-up remarks, we propose a three-step procedure to test the variance component σ_{b0}^{2} and the normality assumption on b_{i0} simultaneously.

Step 1: Create missing data according to a mechanism that depends only on a scalar $Y_{\text{mis},i}$, that is, $s=1$.

Step 2: Compute $t_{\sigma_{\epsilon}^{2}}$ to test $H_{0,\sigma_{\epsilon}^{2}}:\tilde{\sigma}_{\epsilon}^{2*}=\tilde{\sigma}_{\epsilon}^{2}$. If $H_{0,\sigma_{\epsilon}^{2}}$ is rejected, then one finds sufficient evidence to support M_2 with a nonnormal random intercept. If the test result is insignificant, it can be interpreted as lack of evidence to reject M_1 or some evidence of M_2 being the true model with a normal b_{i0}. To tell which is more plausible for the current observed data, one proceeds to Step 3.

Step 3: Create missing data according to another mechanism, such as, $P(\Delta_i=1|Y_i)=P(\Delta_i=1|Y_{\text{mis},i})$ with $s>1$, or $P(\Delta_i=1|Y_i)=P(\Delta_i=1|Y_{\text{mis},i},Y_{\text{obs},i})$. Repeat the test in Step 2. If the test result is significant, then the current observed data supports M_2 with a normal random intercept. Otherwise, there is insufficient evidence to reject M_1.

It is worth noting that this testing strategy kills two birds with one stone if the test at Step 2 yields a significant result, because one detects $\sigma_{b0}^{2}\neq 0$ and the nonnormality of b_{i0} simultaneously, two tasks traditionally being tackled separately. If the normality assumption seems appropriate, Step 3 still gives one a chance to detect a random intercept. If one is not concerned about the distributional assumption and is merely interested in whether or not $\sigma_{b0}^{2}=0$, then one just needs to implement Step 3, with a significant test supporting $\sigma_{b0}^{2}\neq 0$.

4. Testing for Random Slopes

Now consider testing, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$,

$$\begin{aligned} M_3 &: \; Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \qquad \text{versus} \\ M_4 &: \; Y_{ij} = \beta_0 + (\beta_1 + b_{i1}) X_{ij} + \epsilon_{ij}, \end{aligned}$$

where b_{i1} is a random slope with variance σ^2_{b1}, $\beta = (\beta_0, \beta_1)^t$ is the vector of fixed effects, and X_{ij} is the covariate, assumed to be a scalar for ease of exposition. Let $X_i = (X_{i1}, \ldots, X_{in_i})^t$, for $i = 1, \ldots, m$. Suppose M_3 is the assumed model and M_4 is the true model with true parameter values $\mho = (\check{\beta}_0, \check{\beta}_1, v_1, v_\epsilon)^t$, where $v_1 = \text{Var}(b_{i1})$. Parallel with the development in Section 3, we compare $\tilde{\Omega}^*$ with $\tilde{\Omega}$ when ignorable missingness and nonignorable missingess is created, respectively, from the raw data. To further simplify notations, we assume $n_i = n$ and set $Y_{\text{mis},i} = Y_{in}$ for $i = 1, \ldots, m$.

4.1. Ignorable missingness mechanisms

We focus on the simplest MCAR in this subsection, a case in which $\tilde{\Omega}^*$ can be derived analytically.

Proposition 3. *If $P(\Delta_i = 1|Y_i) = \lambda$, where $\lambda \in (0, 1)$ is fixed for all Y_i, $i = 1, \ldots, m$, then $\tilde{\Omega}^* = \tilde{\Omega}$ if and only if*

$$E(X^2_{1n}) = \frac{1}{n-1} \sum_{j=1}^{n-1} E(X^2_{1j}). \tag{4.1}$$

When $\tilde{\Omega}^* \neq \tilde{\Omega}$, the discrepancy lies only in the estimators for σ^2_ϵ.

Proof. With $P(\Delta_i = 1|Y_i) = \lambda$, the incomplete-data likelihood under M_3 is

$$\begin{aligned} f_{\Delta, Y^*}(\Delta_i, Y^*_i | X_i; \Omega) &= (1-\lambda)^{1-\Delta_i} \lambda^{\Delta_i} \, (2\pi\sigma^2_\epsilon)^{-(n_i-1)/2} \\ &\quad \times \exp\left\{ -\frac{\sum_{j=1}^{n_i-1} (Y_{ij} - \beta_0 - \beta_1 X_{ij})^2}{2\sigma^2_\epsilon} \right\} \\ &\quad \times \left[(2\pi\sigma^2_\epsilon)^{-1/2} \exp\left\{ -\frac{(Y_{in_i} - \beta_0 - \beta_1 X_{in_i})^2}{2\sigma^2_\epsilon} \right\} \right]^{1-\Delta_i}, \end{aligned}$$

and the corresponding log likelihood function is

$$\begin{aligned} l^*_i =& (1-\Delta_i) \log(1-\lambda) + \Delta_i \log\lambda - \frac{n_i - 1}{2} \log(2\pi\sigma^2_\epsilon) \\ & - \frac{1}{2\sigma^2_\epsilon} \sum_{j=1}^{n_i-1} (Y_{ij} - \beta_0 - \beta_1 X_{ij})^2 + (1-\Delta_i) \left\{ -\frac{1}{2} \log(2\pi\sigma^2_\epsilon) - \frac{(Y_{in_i} - \beta_0 - \beta_1 X_{in_i})^2}{2\sigma^2_\epsilon} \right\}. \end{aligned}$$

It follows that the score vector, $(\partial/\partial\Omega)l_i^*$, has three elements,

$$\frac{\partial l_i^*}{\partial \beta_0} = \frac{1}{\sigma_\epsilon^2}\Big\{\sum_{j=1}^n (Y_{ij}-\beta_0-\beta_1 X_{ij}) - \Delta_i(Y_{in}-\beta_0-\beta_1 X_{in})\Big\}, \tag{4.2}$$

$$\frac{\partial l_i^*}{\partial \beta_1} = \frac{1}{\sigma_\epsilon^2}\Big\{\sum_{j=1}^n (Y_{ij}-\beta_0-\beta_1 X_{ij})X_{ij} - \Delta_i(Y_{in}-\beta_0-\beta_1 X_{in})X_{in}\Big\}, \tag{4.3}$$

$$\frac{\partial l_i^*}{\partial \sigma_\epsilon^2} = \frac{1}{2\sigma_\epsilon^4}\Big\{\sum_{j=1}^n (Y_{ij}-\beta_0-\beta_1 X_{ij})^2 - \Delta_i(Y_{in}-\beta_0-\beta_1 X_{in})^2\Big\} \\ -\frac{1}{2\sigma_\epsilon^2}(n-\Delta_i). \tag{4.4}$$

Because $E_0(Y_i|X_i) = \check\beta_0 + \check\beta_1 X_i$ and $\Delta_i \perp Y_i$, (4.2) and (4.3) imply $\tilde\beta^* = \check\beta$. By (4.4),

$$E_0\Big(\frac{\partial l_i^*}{\partial \sigma_\epsilon^2}\Big|X_i\Big) = \frac{n-\lambda}{2\sigma_\epsilon^2}\Big(\frac{v_\epsilon}{\sigma_\epsilon^2}-1\Big) - \frac{v_1}{2\sigma_\epsilon^4}\Big(\lambda X_{in}^2 - \sum_{j=1}^n X_{ij}^2\Big). \tag{4.5}$$

Solving $\lim_{m\to\infty} m^{-1}\sum_{i=1}^m E_0(\partial l_i^*/\partial\sigma_\epsilon^2|X_i) = 0$ for σ_ϵ^2 yields

$$\tilde\sigma_\epsilon^{2*} = v_\epsilon + \frac{v_1}{n-\lambda}\Big\{\sum_{j=1}^n E(X_{1j}^2) - \lambda E(X_{1n}^2)\Big\}. \tag{4.6}$$

Now that $\tilde\Omega^*$ is derived, $\tilde\Omega$ can be obtained by setting $\Delta_i = 0$ in (4.2) and (4.3) and evaluating (4.6) at $\lambda = 0$, which yields $\tilde\Omega = (\check\beta^t,\, \tilde\sigma_\epsilon^2)^t$, where

$$\tilde\sigma_\epsilon^2 = v_\epsilon + \frac{v_1}{n}\sum_{j=1}^n E(X_{1j}^2). \tag{4.7}$$

Equating (4.6) to (4.7) reveals that $\tilde\sigma_\epsilon^{2*} = \tilde\sigma_\epsilon^2$ if and only if (4.1) holds.

Remark 1. Proposition 3 still holds if M_3 and M_4 involve extra covariates which no random slopes are associated with, i.e., $M_3 : Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2^t W_{ij} + \epsilon_{ij}$ and $M_4 : Y_{ij} = \beta_0 + (\beta_1 + b_{i1})X_{ij} + \beta_2^t W_{ij} + \epsilon_{ij}$, where β_2 is the (vector of) fixed effect(s) and W_{ij} is a (vector of) covariate(s).

Remark 2. Besides extra covariates, if some of them also have random slopes associated with them in M_4, then one still has $\tilde\beta^* = \tilde\beta = \check\beta$, but the sufficient and necessary condition for $\tilde\sigma_\epsilon^{2*} = \tilde\sigma_\epsilon^2$ is no longer (4.1). For example, with a scalar β_2 and M_3 given in *Remark 1*, but M_4 as $Y_{ij} = \beta_0 + (\beta_1 + b_{i1})X_{ij} + (\beta_2 + b_{i2})W_{ij} + \epsilon_{ij}$, where $\mathrm{Var}(b_{i2}) = v_2$ and $\mathrm{Cov}(b_{i1},\, b_{i2}) = v_{12}$ (true parameter values), then the new condition replacing (4.1) is

$$v_1 E\left(X_{1n}^2\right) + v_2 E\left(W_{1n}^2\right) + 2v_{12} E\left(X_{1n}W_{1n}\right)$$

$$= \frac{1}{n-1} \sum_{j=1}^{n-1} \left\{ v_1 E\left(X_{1j}^2\right) + v_2 E\left(W_{1j}^2\right) + 2v_{12} E\left(X_{1j} W_{1j}\right) \right\}. \tag{4.8}$$

Clearly, (4.1) is a special case of (4.8) when $v_2 = v_{12} = 0$.

Remark 3. Under M_3 and M_4 as in Remark 2, considering a scalar W_{ij} as an example, one can show that

$$\tilde{\sigma}_\epsilon^2 = v_\epsilon + \frac{1}{n} \sum_{j=1}^{n} \left\{ v_1 E\left(X_{1j}^2\right) + v_2 E\left(W_{1j}^2\right) + 2v_{12} E\left(X_{1j} W_{1j}\right) \right\}. \tag{4.9}$$

As one would expect, (4.7) is a special case of (4.9) with $v_2 = v_{12} = 0$. According to (4.9), $\tilde{\sigma}_\epsilon^2$ depends on the second moments of a covariate if and only if this covariate has a random slope associated with it. This result suggests a way to detect random slope for one covariate at a time without generating missing data. The idea is to create two subsets of data from the observed data. Both subsets have m experimental units (or clusters) but they differ in the covariates values (and also probably the size of a cluster). If X is the covariate of interest in terms of testing for a random slope, then the subsets are created to satisfy the following: two subsets have the same value of $\sum_j E(W_{1j}^2)$ and also agree in the value of $\sum_j E(X_{1j}W_{1j})$, but disagree in $\sum_j E(X_{1j}^2)$. Now one has two estimates of σ_ϵ^2 based on the two subsets. These two should be similar if $v_1 = 0$, and are expected to differ (at least in limit) if $v_1 \neq 0$. A test similar to those defined in Section 2.3 can be used to quantify the discrepancy between two estimates. Of course, controlling the moment conditions of the covariates is more convenient to implement at the design stage than after data are collected. An example illustrating this strategy is given in the Supplementary Materials (Part I).

All discussions in this subsection are free on the distributional assumption on b_{i1}. Except for the specific forms of $f_{\Delta,Y^*}(\Delta_i, Y_i^*|X_i;\Omega)$ and l_i^*, the proof carries over to MAR. In order to test the distributional assumption on the random slope, we turn to NMAR.

4.2. Nonignorable missingness mechanisms

Proposition 4. *If $P(\Delta_i = 1|Y_i) = P(\Delta_i = 1|Y_{\mathrm{mis},i})$ for $i = 1,\ldots,m$, then $\tilde{\Omega}^* = \tilde{\Omega}$ if and only if*

$$\lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}\right) = \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0(\Delta_i) \frac{E\left(\Delta_i \frac{\partial l_{i2}}{\partial \Omega}\Big|_{\tilde{\Omega}}; \tilde{\Omega}\right)}{E(\Delta_i;\tilde{\Omega})}, \tag{4.10}$$

where $E_0(\cdot)$ is the expectation with respect to the true model of Y_i given X_i with $\Omega = \mho$, and $E(\cdot;\tilde{\Omega})$ is the expectation with respect to the assumed model of Y_i given X_i with $\Omega = \tilde{\Omega}$. Moreover, a sufficient condition for (4.10) is

$$g_{Y}(Y_{\mathrm{mis},i}|X_i;\mho) = f_{Y}(Y_{\mathrm{mis},i}|X_i;\tilde{\Omega}). \tag{4.11}$$

The proof is very similar to the one for Proposition 2 and thus is omitted. Even though the sufficient condition in (4.11) is in the same spirit as (3.4), now that covariates are involved, (4.11) is actually a more stringent condition than (3.4). Consequently, the three-step testing procedure in Section 3.2 is not applicable for testing a random slope, and a new procedure for diagnosing model assumptions on b_{i1} is needed. This is elaborated upon next.

If $b_{i1} \sim N(0, v_1)$ in the true model and $Y_{\mathrm{mis},i} = Y_{in}$ for $i = 1, \ldots, m$, then

$$g_{Y}(Y_{\mathrm{mis},i}|X_i;\mho) = \frac{1}{\sqrt{v_{Y_{in}}}}\phi\left(\frac{Y_{in} - \check{\mu}_{Y_{in}}}{\sqrt{v_{Y_{in}}}}\right), \tag{4.12}$$

$$f_{Y}(Y_{\mathrm{mis},i}|X_i;\tilde{\Omega}) = \frac{1}{\tilde{\sigma}_\epsilon}\phi\left(\frac{Y_{in} - \check{\mu}_{Y_{in}}}{\tilde{\sigma}_\epsilon}\right), \tag{4.13}$$

where $\phi(\cdot)$ denotes the standard normal density function, $\check{\mu}_{Y_{in}} = \check{\beta}_0 + \check{\beta}_1 X_{in}$, $v_{Y_{in}} = v_\epsilon + X_{in}^2 v_1$, and $\tilde{\sigma}_\epsilon^2$ is given as (4.7). Clearly, if $v_{Y_{in}} = \tilde{\sigma}_\epsilon^2$, then (4.12) coincides with (4.13) and $\tilde{\Omega}^* = \tilde{\Omega}$ follows. However, comparing $v_{Y_{in}}$ with $\tilde{\sigma}_\epsilon^2$ makes it evident that (4.11) is generally not satisfied even with a normal random slope. A special case in which (4.12) and (4.13) agree is when the design points are fixed, that is, $X_{ij} = X_{1j}$ for $i = 1, \ldots, m$, $j = 1, \ldots, n$, and

$$X_{1n}^2 = \frac{1}{n-1}\sum_{j=1}^{n-1} X_{1j}^2. \tag{4.14}$$

If design points are random and (4.1) is satisfied, then (4.12) and (4.13) agree almost surely.

We now zoom in on a particular nonignorable missingness mechanism under which (4.10) can be elaborated on further. Suppose that the true model for b_{i1} is normal and one generates missing data among $\{Y_{in}\}_{i=1}^m$ according to

$$P(\Delta_i = 1|Y_{in}) = \Phi(\lambda_0 + \lambda_1 Y_{in}), \tag{4.15}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. Assuming $\{X_i\}_{i=1}^m$ i.i.d, and by the Weak Law of Large Numbers, $\lim_{m\to\infty} m^{-1}\sum_{i=1}^m h(X_{in}) = E\{h(X_{1n})\}$, where $h(\cdot)$ is a generic function and $E(\cdot)$ is the expectation with respect to the distribution of X_{1n}. Now (4.10) can be simplified and made explicit for each element in Ω. In what follows, the necessary and sufficient condition for each element in $\tilde{\Omega}^*$ to agree with its counterpart in $\tilde{\Omega}$ is given. To attain concise notations, for $a = v_{Y_{in}}$ or $\tilde{\sigma}_\epsilon^2$, let

$$D(a) = \sqrt{1 + \lambda_1^2 a}, \qquad R(a) = \frac{\lambda_0 + \lambda_1 \check{\mu}_{Y_{in}}}{D(a)}.$$

Now one has $\tilde{\beta}_0^* = \tilde{\beta}_0$ if and only if

$$\lambda_1 E\left[\frac{1}{D(v_{Y_{in}})}\frac{v_{Y_{in}}}{\tilde{\sigma}_\epsilon^2}\phi\left\{R(v_{Y_{in}})\right\}\right] = \lambda_1 E\left[\frac{1}{D(\tilde{\sigma}_\epsilon^2)}\frac{\Phi\left\{R(v_{Y_{in}})\right\}}{\Phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}}\phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}\right]. \quad (4.16)$$

It is obvious from (4.16) that $\lambda_1 = 0$ leads to $\tilde{\beta}_0^* = \tilde{\beta}_0$, which is expected according to Proposition 3 as $\lambda_1 = 0$ corresponds to MCAR.

We have $\tilde{\beta}_1^* = \tilde{\beta}_1$ if and only if

$$\lambda_1 E\left[\frac{1}{D(v_{Y_{in}})}\frac{v_{Y_{in}}}{\tilde{\sigma}_\epsilon^2}X_{in}\phi\left\{R(v_{Y_{in}})\right\}\right] = \lambda_1 E\left[\frac{1}{D(\tilde{\sigma}_\epsilon^2)}\frac{\Phi\left\{R(v_{Y_{in}})\right\}}{\Phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}}X_{in}\phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}\right]. \quad (4.17)$$

This condition differs from (4.16) only in the extra multiplicative factor X_{in} within the expectations. Naturally, $\lambda_1 = 0$ leads to $\tilde{\beta}_1^* = \tilde{\beta}_1$. Moreover, $X_{in} \equiv 0$ also results in $\tilde{\beta}_1^* = \tilde{\beta}_1$. Therefore, it does not help model diagnosis to choose the component in Y_i as the potentially missing component if the covariate associated with it is fixed at zero.

As well, $\tilde{\sigma}_\epsilon^{2*} = \tilde{\sigma}_\epsilon^2$ if and only if

$$\begin{aligned}
&\lambda_1^2 E\left[\frac{1}{D(v_{Y_{in}})}\frac{v_{Y_{in}}^2}{\tilde{\sigma}_\epsilon^4}\frac{R(v_{Y_{in}})}{D(v_{Y_{in}})}\phi\left\{R(v_{Y_{in}})\right\}\right] - \frac{1}{\tilde{\sigma}_\epsilon^4}E\left[\left(v_{Y_{in}} - \tilde{\sigma}_\epsilon^2\right)\Phi\left\{R(v_{Y_{in}})\right\}\right] \\
&\quad = \lambda_1^2 E\left[\frac{1}{D(\tilde{\sigma}_\epsilon^2)}\frac{\Phi\left\{R(v_{Y_{in}})\right\}}{\Phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}}\frac{R(\tilde{\sigma}_\epsilon^2)}{D(\tilde{\sigma}_\epsilon^2)}\phi\left\{R(\tilde{\sigma}_\epsilon^2)\right\}\right]. \qquad (4.18)
\end{aligned}$$

Unlike the conditions in (4.16) and (4.17), now $\lambda_1 = 0$ does not suffice to yield $\tilde{\sigma}_\epsilon^{2*} = \tilde{\sigma}_\epsilon^2$, which is also expected according to Proposition 3. In fact, when $\lambda_1 = 0$, (4.18) can be simplified to (4.1).

With the conditions explicitly spelled out in (4.16)–(4.18), it becomes evident that $v_{Y_{in}} = \tilde{\sigma}_\epsilon^2$ guarantees all three conditions to hold. In summary, we have the following conclusions regarding the comparison between $\tilde{\Omega}^*$ and $\tilde{\Omega}$ when the true model contains a normal random slope and the missingness mechanism is given by (4.15).

(L1) If $\lambda_1 = 0$, then $\tilde{\beta}^* = \tilde{\beta} = \check{\beta}$, but $\tilde{\sigma}_\epsilon^{2*} \neq \tilde{\sigma}_\epsilon^2$ unless (4.1) is true. This is consistent with Proposition 3.

(L2) If $\lambda_1 \neq 0$, then the sufficient and necessary conditions for $\tilde{\Omega}^* = \tilde{\Omega}$ are given in (4.16)-(4.18), which mainly depends on the distribution of X_{in}.

(L3) If the design points are fixed and (4.14) is satisfied, then $\tilde{\Omega}^* = \tilde{\Omega}$. In this case the missingness mechanism in (4.15) cannot reveal the existence of a normal random slope even if $\lambda_1 \neq 0$.

If the random slope does not follow a normal distribution, then (4.11) is immediately violated since (4.12) is no longer true even though (4.13) still holds. Although, mathematically, (4.11) is not a necessary condition for $\tilde{\Omega}^* = \tilde{\Omega}$, for most practical situations it is virtually impossible to have $\tilde{\Omega}^* = \tilde{\Omega}$ when (4.11) is violated. In other words, a nonnormal random slope practically assures that $\tilde{\Omega}^*$ disagrees with $\tilde{\Omega}$. This claim, along with (L3), suggests the following two-step procedure for simultaneously testing the variance component σ_{b1}^2 and assessing the normality assumption on b_{i1}.

Step 1: Transform the design points, if necessary, so that (4.14) holds. Generate missing data according to (4.15) and calculate $t_{\sigma_\epsilon^2}$. A significant value of $t_{\sigma_\epsilon^2}$ at this step suggests a nonnormal random slope. If it is insignificant, it can be due to a normal random slope in the true model or to lack of evidence to reject M_3. To decide which is more plausible for the current data, one proceeds to Step 2.

Step 2: Reset the design points so that (4.14) is violated. Repeat the test in Step 1. A significant value of $t_{\sigma_\epsilon^2}$ at this step provides evidence of a normal random slope in the true model. An insignificant value suggests lack of evidence to reject M_3.

Like the three-step procedure in Section 3.2, only Step 2 in this two-step procedure is needed if one just wants to test the variance component. In Step 1, when a transformation on design points is needed, it can be shown that the following transformed points satisfy (4.14),

$$X_{1j}^* = X_{1j} - \frac{X_{1n}^2 - \sum_{j=1}^{n-1} X_{1j}^2/(n-1)}{2\left\{X_{1n} - \sum_{j=1}^{n-1} X_{1j}/(n-1)\right\}}, \qquad j = 1, \ldots, n,$$

assuming that $X_{1n} \neq \sum_{j=1}^{n-1} X_{1j}/(n-1)$. Under the same assumption, it is just as easy to reset the design points in Step 2 to violate (4.14) when needed. For instance, a non-zero constant shift applied to $\{X_{1j},\, j = 1, \ldots, n\}$ will do. If it so happens that $X_{1n} = \sum_{j=1}^{n-1} X_{1j}/(n-1)$, then one can create missingness among $\{Y_{ij'},\, i = 1, \ldots, m\}$ instead of $\{Y_{in},\, i = 1, \ldots, m\}$, where $j' \in \{1, \ldots, n-1\}$ satisfies $X_{1j'} \neq \sum_{j=1, j\neq j'}^{n} X_{1j}/(n-1)$. Such j' always exists unless all the design points within a cluster are equal, a very unnatural setting in practice.

4.3. Additional random effects and fixed effects

We have focused on testing for a random intercept only (without considering random slopes) or testing for random slopes only (without a random intercept in either assumed or true models) in order to highlight the rationale of using missing data to develop diagnostic tests. Once the rationale is well understood, one can certainly consider other assumed and true model combinations that include both random intercept and random slopes in one or both models. As an example, we have looked into the null and alternative models, $M_5 : Y_{ij} = \beta_0 + b_{i0} + \beta_1 X_{ij} + \epsilon_{ij}$ and $M_6 : Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})X_{ij} + \epsilon_{ij}$. Discussions parallel to those in Section 3 and Section 4 are given in the Supplementary Materials (Part II). The procedure used to test M_3 versus M_5 is essentially identical to those described in Section 3 and thus is omitted here.

Lastly, all the above considered pairs of assumed and true models have the same fixed-effects structure. In the Supplementary Materials (Parts III and IV), focusing on testing for b_{i1}, we study the limiting MLEs in the presence of additional covariates that may be excluded or misspecified in the assumed model. Also included in the Supplementary Materials is empirical evidence of the effects of such extra misspecification on the proposed tests. An overall impression gained from the empirical evidence is that the proposed tests are usually fairly robust to fixed-effects misspecification, especially when the additional covariates are identically distributed across $i = 1, \ldots, m$, $j = 1, \ldots, n$. But, through some exploration, we do realize that the power of the tests for b_{i1} can increase or lessen depending on how random-effects misspecification interacts with fixed-effects misspecification. The comforting news is that it requires very specific covariate-parameter configurations (see details in the Supplementary Materials), a coincidence rarely expected in practice, for such interaction to occur in a way to negate the power of the test to zero.

5. Numerical Evidence

5.1. Three examples

In this subsection, the results of two sets of simulation studies are presented to empirically justify the theoretical development in Section 3 and Section 4, and to demonstrate the operating characteristics of the test statistics. We also apply the methods to data from a longitudinal study.

Example 1 (Testing M_1 versus M_2)**.** For each of 2,000 Monte Carlo (MC) replicates, a data set $\{(Y_{i1}, \ldots, Y_{in_i})\}_{i=1}^m$ was generated from M_2 with $\mu = 1$, $\sigma_\epsilon^2 = 1$, $\sigma_{b0}^2 = 0.5$ or 1, $m = 50$, 100, 200, or 300, and $\{n_i\}_{i=1}^m$ generated from a Poisson distribution with mean 10, among which any simulated n_i falling below two was replaced by two. Realizations of b_{i0} were generated from

(i) (C1) $N(0, \sigma_{b0}^2)$;

(ii) (C2) a skewed distribution with mean zero and variance σ_{b0}^2 obtained by shifting and scaling a gamma distribution with shape parameter 1.5.

Table 1. Averages of MLEs across 2,000 MC replicates from Example 1 with $m = 200$. True parameter values are $(\mu, \sigma_\epsilon^2) = (1, 1)$. Average missing rate under each setting is given by r. Numbers in parentheses are MC standard errors associated with the averages. Incomplete-data MLEs that are expected to converge to the same limits as their complete-data counterparts are underlined.

	μ	σ_ϵ^2	μ	σ_ϵ^2
(C1) $b_{i0} \sim N(0, \sigma_{b0}^2)$	$\sigma_{b0}^2 = 0.5$		$\sigma_{b0}^2 = 1$	
Complete-data Limiting MLE	1	1.5	1	2
Complete-data MLE	1.000 (0.001)	1.497 (0.002)	1.000 (0.002)	1.994 (0.003)
Incomplete-data MLE				
$P(\Delta_i = 1 \mid Y_i) = \Phi(-6Y_{i1}Y_{in_i})$	$\underline{1.002}$ (0.001)	1.505 (0.002)	$\underline{1.004}$ (0.002)	2.011 (0.003)
	$[r = 0.271]$		$[r = 0.255]$	
$P(\Delta_i = 1 \mid Y_i) = \Phi(6Y_{i1}Y_{in_i})$	$\underline{0.997}$ (0.001)	1.486 (0.002)	$\underline{0.996}$ (0.002)	1.972 (0.003)
	$[r = 0.729]$		$[r = 0.745]$	
$P(\Delta_i = 1 \mid Y_i) = 0.5$	$\underline{1.000}$ (0.001)	$\underline{1.497}$ (0.002)	$\underline{1.000}$ (0.002)	$\underline{1.993}$ (0.003)
	$[r = 0.501]$		$[r = 0.500]$	
$P(\Delta_i = 1 \mid Y_i) = \Phi(-2Y_{in_i})$	$\underline{1.002}$ (0.001)	$\underline{1.497}$ (0.002)	$\underline{0.999}$ (0.002)	$\underline{1.997}$ (0.003)
	$[r = 0.224]$		$[r = 0.252]$	
(C2) $b_{i0} \sim$ shifted gamma	$\sigma_{b0}^2 = 0.5$		$\sigma_{b0}^2 = 1$	
Complete-data Limiting MLE	1	1.5	1	2
Complete-data MLE	1.001 (0.001)	1.496 (0.002)	1.002 (0.002)	1.996 (0.004)
Incomplete-data MLE				
$P(\Delta_i = 1 \mid Y_i) = \Phi(-6Y_{i1}Y_{in_i})$	$\underline{1.003}$ (0.001)	1.507 0.002)	$\underline{1.005}$ (0.002)	2.020 (0.004)
	$[r = 0.295]$		$[r = 0.305]$	
$P(\Delta_i = 1 \mid Y_i) = \Phi(6Y_{i1}Y_{in_i})$	$\underline{0.998}$ (0.001)	1.483 (0.002)	$\underline{0.996}$ (0.002)	1.969 (0.004)
	$[r = 0.705]$		$[r = 0.694]$	
$P(\Delta_i = 1 \mid Y_i) = 0.5$	$\underline{1.001}$ (0.001)	$\underline{1.496}$ (0.002)	$\underline{1.002}$ (0.002)	$\underline{1.997}$ (0.004)
	$[r = 0.500]$		$[r = 0.500]$	
$P(\Delta_i = 1 \mid Y_i) = \Phi(-2Y_{in_i})$	$\underline{1.000}$ (0.001)	1.506 (0.002)	$\underline{0.995}$ (0.002)	2.012 (0.004)
	$[r = 0.225]$		$[r = 0.256]$	

When implementing the proposed diagnostic procedures, missing data were created among $\{Y_{in_i}\}_{i=1}^m$. Under each of (C1) and (C2), the following missingness mechanisms were considered, $P(\Delta_i = 1|Y_i) = \Phi(\lambda_1 Y_{i1} Y_{in_i})$ with $\lambda_1 = -6, 0, 6$, and $P(\Delta_i = 1|Y_i) = \Phi(-2Y_{in_i})$, all of which lead to NMAR except for the one with $\lambda_1 = 0$ that corresponds to MCAR, as now the missingness mechanism is simply $P(\Delta_i = 1|Y_i) = 0.5$. In each setting, we computed $\hat{\Omega}$, $\hat{\Omega}^*$, $t_{1\theta}$, and $t_{2\theta}$. Results summarizing the MLEs when $m = 200$ are presented in Table 1. The empirical power of t_{1,σ_ϵ^2} is shown in Figure 1. The picture for t_{2,σ_ϵ^2} is very similar and thus is omitted here.

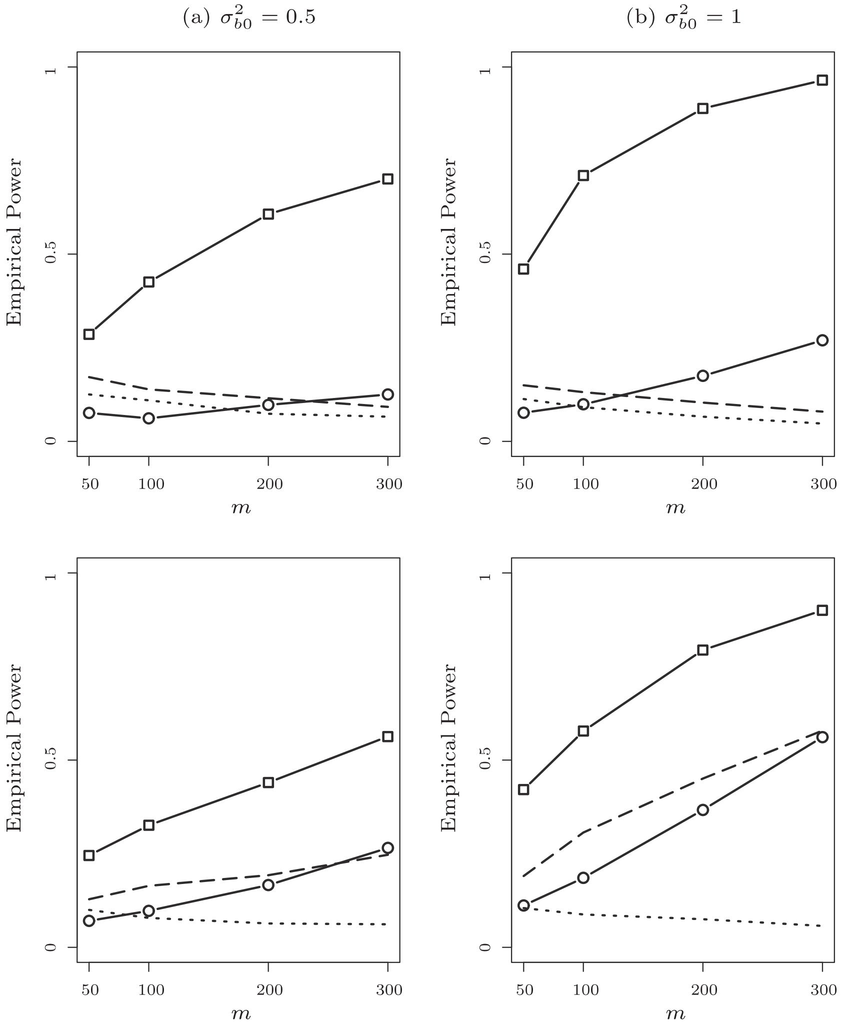

Figure 1. Empirical power of t_{1,σ_ϵ^2} versus sample size m in Example 1. Upper panel is for (C1): $b_{i0} \sim N(0, \sigma_{b1}^2)$; lower panel is for (C2): $b_{i0} \sim$ shifted gamma. Four curves within each grid correspond to the missingness mechanisms $P(\Delta_i|Y_i) = \Phi(\lambda_1 Y_{i1} Y_{in_i})$ (solid lines, with squares and circles for $\lambda_1 = -6$ and 6, respectively), $P(\Delta_i|Y_i) = 0.5$ (dotted lines), and $P(\Delta_i|Y_i) = \Phi(-2Y_{in_i})$ (dashed lines).

As stated in Proposition 1, when the missingness is MCAR (see dotted lines in Figure 1, even though both $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\epsilon^{2*}$ are biased for σ_ϵ^2, both $\hat{\Omega}$ and $\hat{\Omega}^*$ are consistent for $\tilde{\Omega}$, and consequently the t_{1,σ_ϵ^2} are mostly insignificant at the 0.05 significance level. More interestingly, as proved in Section 3.2, $\hat{\Omega}$ and $\hat{\Omega}^*$ are still consistent for $\tilde{\Omega}$ when the missingness is NMAR as long as b_{i0} is normally distributed and Δ_i only depends on Y_{in_i} (see dashed lines in the upper panel of Figure 1). However, if the nonignorable missingness also depends on Y_{i1}, then $\hat{\Omega}^*$ is biased for $\tilde{\Omega}$ and substantial deviation from $\hat{\Omega}$ emerges (see solid lines in the upper panel of Figure 1). In this case, the test statistics show moderate to high power to detect the discrepancy between $\hat{\Omega}^*$ and $\hat{\Omega}$, with power increasing as the misspecification in M_1 becomes more severe (as σ^2_{b0} increases). In contrast, if b_{i0} is not normally distributed, $\tilde{\Omega}^* \neq \tilde{\Omega}$ is expected when the missingness is NMAR, and t_{1,σ^2_ϵ} exhibits promising power to detect a nonnormal random intercept (see solid lines and dashed lines in the lower panel of Figure 1). The observed phenomena in this example reinforce the results of Section 3.

As for existing methods, we consider the likelihood ratio test (LRT) in Self and Liang (1987) and the LRT in Crainiceanu and Ruppert (2004), with the latter providing a more accurate approximation for the null distribution of LRT than the former. Let SL-LRT and CR-LRT denote these methods. Because n_i varies across $i = 1, \ldots, m$, the condition of identically distributed Y_i's ($i = 1, \ldots . m$) required for SL-LRT is not satisfied, hence their test is not included for comparison under this particular simulation setting. In other simulation settings with the n_i fixed that we looked into (not reported here), SL-LRT usually has higher power in detecting existence of a random intercept. Like our test, CR-LRT is applicable whether n_i varies across $i = 1, \ldots, m$, or not. The empirical power of their test reaches one much faster as m grows even when $\sigma^2_{b0} = 0.5$. Both LRTs are very robust to the normality assumption. This robustness is a virtue of these methods especially when identifying a random effect is of main interest and its distributional assumption is secondary. But if the distributional assumption is also of interest, our test has the advantage in detecting deviations from normal.

Example 2 (Testing M_3 versus M_4)**.** For each of 2,000 MC replicates, a data set of size $m = 50$, 100, or 200 was generated from M_4 with $\beta_0 = 0$, $\beta_1 = 1$, $\sigma^2_\epsilon = 1$, and σ^2_{b1} ranging from 0 to 0.5. The missingness mechanism used to create missing data among $\{Y_{in}\}_{i=1}^m$ was defined by $P(\Delta_i = 1|Y_i) = \Phi(\lambda_1 Y_{in})$ with $\lambda_1 = -2$, 0, 2. Varying the true distribution of b_{i1} and the value of X_i, we considered four settings. For $i = 1, \ldots, m$,

(D1) $b_{i1} \sim N(0, \sigma^2_{b1})$ and $X_i = (1, 2)$;

(D2) b_{i1} follows the shifted gamma distribution described in (C2) and $X_i = (1, 2)$;

(D3) $b_{i1} \sim N(0, \sigma^2_{b1})$ and $X_i = (1, 2, \sqrt{2.5})$;

(D4) b_{i1} follows the shifted gamma distribution described in (C2) and $X_i = (1, 2, \sqrt{2.5})$.

In (D3) and (D4), X_{i3} was chosen to satisfy (4.14). Table 2 summarizes $\hat{\Omega}$ and $\hat{\Omega}^*$ when $m = 100$. Figure 2 presents the empirical power of t_{1,σ^2_ϵ} under (D1) and (D2) when $\lambda_1 = -2$, 0, and 2. Figure 3 depicts the empirical power of t_{1,σ^2_ϵ} and t_{2,σ^2_ϵ} when $\lambda_1 = -2$ and 2. Also included in Figure 2 are the empirical power of SL-LRT, CR-LRT, and the score test proposed by Lin (1997), which considers a two-sided test (of no random slope) with a test statistic following a chi-square distribution under the null. Unlike the two LRTs, the derivation of Lin's test only depends on assumptions on the first two moments of the random effects and does not require a normality assumption. As in the previous example, both LRTs have higher empirical power across the considered range of $\sigma^2_{b1} > 0$, but t_{1,σ^2_ϵ} quickly catches up in power as m or σ^2_{b1} increases (see solid lines crossing circles in Figure 2). The power of Lin's test is much lower when $n = 2$ (see dash-dotted lines in Figure 2) but gains much more power when $n = 3$ (though not plotted in Figure 3). This observation is in line with Lin's comment (Lin (1997, p.322)) that the critical value of her test statistic can be less accurate when the cluster size is small.

Under (D1) and (D2), the operating characteristics of all tests considered are very robust to the normality assumption on b_{i1} (comparing the upper and lower panels of Figure 2). Such robustness is retained for the three existing methods under (D3) and (D4) (though not plotted in Figure 3), but not for our method. As observed in Figure 3, with X_i satisfying (4.14), the power of t_{1,σ^2_ϵ} and t_{2,σ^2_ϵ} remains mostly lower or around 0.05 when $b_{i1} \sim N(0, \sigma^2_{b1})$ (see the upper panel of Figure 3); when b_{i1} is not normally distributed (see the lower panel of Figure 3), both tests reject H_0 more often, with empirical power increasing as m or σ^2_{b1} increases. Such sensitivity of the proposed tests to the random-slope distribution allows one to not only detect a random slope but also assess its normality assumption, which is an intriguing feature not possessed by the other three methods. Finally, as stated in Section 2.3, t_{1,σ^2_ϵ} and t_{2,σ^2_ϵ} are asymptotically equivalent, and Figure 3 provides empirical evidence for this statement (comparing lines crossing squares with lines crossing circles in Figure 3.

In summary, the empirical evidence in this example concurs with the theoretical development in Section 4. We also investigated the scenarios when $\{X_i\}_{i=1}^m$ were randomly generated with mean equal to the fixed values in (D1)–(D4), and observed similar patterns to those in Table 2, Figure 2, and Figure 3.

Example 3 (Rat data)**.** We now consider a data set from an experiment designed to investigate the effect of Decapeptyl, an inhibitor for testosterone production, in male Wistar rats on their craniofacial growth (Verdonck et al. (1998)). This data set consists of the distance (in pixels) between well-defined points on X-ray pictures of the skull of each rat, a measurement scheduled to be collected after

Table 2. Averages of MLEs across 2,000 MC replicates from Example 5.2 with $m = 100$ and $\sigma^2_{b1} = 0.2,\ 0.5$. True parameter values are $(\beta_0,\ \beta_1,\ \sigma^2_\epsilon) = (0,\ 1,\ 1)$. Missingness is created among $\{Y_{in}\}_{i=1}^m$ according to $P(\Delta_i = 1|Y_i) = \Phi(\lambda_1 Y_{in})$. Average missing rate under each setting is given by r. Numbers in parentheses are MC standard errors associated with the averages. Incomplete-data MLEs that are expected to converge to the same limits as their complete-data counterparts are underlined.

	Complete-data		Incomplete-data MLE		
	Limiting MLE	MLE	$\lambda_1 = -2$	$\lambda_1 = 0$	$\lambda_1 = 2$
		(D1) $b_{i1} \sim N(0,\ \sigma^2_{b1})$, $X_i = (1,\ 2)$			
$\sigma^2_{b1} = 0.2$			$[r = 0.081]$	$[r = 0.500]$	$[r = 0.919]$
β_0	0	−0.008 (0.005)	−0.019 (0.005)	$\underline{-0.008}$ (0.006)	0.313 (0.007)
β_1	1	1.002 (0.003)	1.013 (0.003)	$\underline{1.002}$ (0.005)	0.682 (0.006)
σ^2_ϵ	1.5	1.483 (0.004)	1.457 (0.004)	1.374 (0.004)	1.217 (0.004)
$\sigma^2_{b1} = 0.5$			$[r = 0.134]$	$[r = 0.500]$	$[r = 0.867]$
β_0	0	−0.002 (0.005)	−0.034 (0.005)	$\underline{-0.004}$ (0.006)	0.557 (0.006)
β_1	1	0.999 (0.003)	1.031 (0.003)	$\underline{1.001}$ (0.005)	0.440 (0.005)
σ^2_ϵ	2.25	2.235 (0.006)	2.139 (0.006)	1.974 (0.006)	1.620 (0.005)
		(D2) $b_{i1} \sim$ shifted gamma, $X_i = (1,\ 2)$			
$\sigma^2_{b1} = 0.2$			$[r = 0.070]$	$[r = 0.501]$	$[r = 0.930]$
β_0	0	−0.000 (0.005)	0.004 (0.005)	$\underline{0.003}$ (0.006)	0.197 (0.008)
β_1	1	1.000 (0.003)	0.996 (0.003)	$\underline{0.997}$ (0.004)	0.803 (0.007)
σ^2_ϵ	1.5	1.492 (0.004)	1.503 (0.004)	1.384 (0.004)	1.193 (0.004)
$\sigma^2_{b1} = 0.5$			$[r = 0.112]$	$[r = 0.499]$	$[r = 0.888]$
β_0	0	0.001 (0.005)	0.020 (0.005)	$\underline{-0.001}$ (0.006)	0.386 (0.007)
β_1	1	0.999 (0.004)	0.980 (0.004)	$\underline{1.001}$ (0.005)	0.614 (0.006)
σ^2_ϵ	2.25	2.229 (0.008)	2.286 (0.009)	1.971 (0.008)	1.480 (0.005)
		(D3) $b_{i1} \sim N(0,\ \sigma^2_{b1})$, $X_i = (1,\ 2,\ \sqrt{2.5})$			
$\sigma^2_{b1} = 0.2$			$[r = 0.116]$	$[r = 0.501]$	$[r = 0.884]$
β_0	0	0.009 (0.005)	$\underline{0.009}$ (0.005)	$\underline{0.009}$ (0.005)	$\underline{0.009}$ (0.005)
β_1	1	0.996 (0.003)	$\underline{0.996}$ (0.003)	$\underline{0.996}$ (0.003)	$\underline{0.996}$ (0.003)
σ^2_ϵ	1.5	1.489 (0.003)	$\underline{1.489}$ (0.003)	$\underline{1.489}$ (0.003)	$\underline{1.487}$ (0.003)
$\sigma^2_{b1} = 0.5$			$[r = 0.159]$	$[r = 0.501]$	$[r = 0.841]$
β_0	0	−0.002 (0.005)	$\underline{-0.002}$ (0.005)	$\underline{-0.001}$ (0.005)	$\underline{-0.003}$ (0.005)
β_1	1	1.002 (0.003)	$\underline{1.002}$ (0.003)	$\underline{1.003}$ (0.003)	$\underline{1.002}$ (0.003)
σ^2_ϵ	2.25	2.238 (0.005)	$\underline{2.239}$ (0.005)	$\underline{2.238}$ (0.005)	$\underline{2.234}$ (0.005)
		(D4) $b_{i1} \sim$ shifted gamma, $X_i = (1,\ 2,\ \sqrt{2.5})$			
$\sigma^2_{b1} = 0.2$			$[r = 0.109]$	$[r = 0.499]$	$[r = 0.890]$
β_0	0	0.000 (0.005)	−0.001 (0.005)	$\underline{0.000}$ (0.005)	0.004 (0.005)
β_1	1	1.000 (0.003)	0.999 (0.003)	$\underline{1.000}$ (0.003)	1.002 (0.003)
σ^2_ϵ	1.5	1.482 (0.004)	1.498 (0.004)	$\underline{1.481}$ (0.004)	1.456 (0.004)
$\sigma^2_{b1} = 0.5$			$[r = 0.148]$	$[r = 0.498]$	$[r = 0.853]$
β_0	0	0.003 (0.005)	−0.003 (0.005)	$\underline{0.003}$ (0.005)	0.013 (0.005)
β_1	1	0.997 (0.004)	0.993 (0.004)	$\underline{0.997}$ (0.004)	1.003 (0.004)
σ^2_ϵ	2.25	2.237 (0.008)	2.298 (0.008)	$\underline{2.234}$ (0.008)	2.156 (0.007)

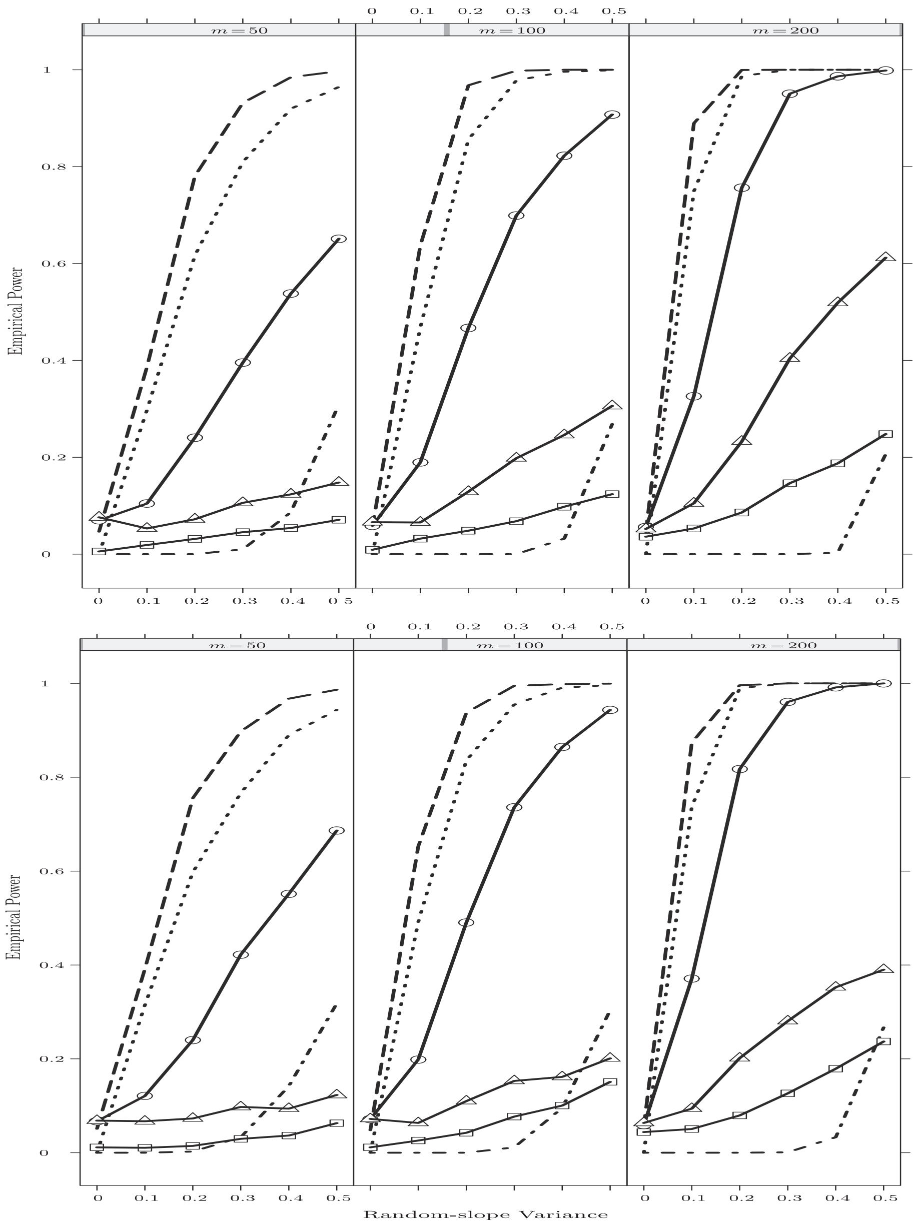

Figure 2. Empirical power of tests for a random slope versus σ_{b1}^2 for (D1) and (D2) with $m = 50$, 100, 200. Upper panel is for (D1): $b_{i1} \sim N(0, \sigma_{b1}^2)$; lower panel is for (D2): $b_{i1} \sim$ shifted gamma. Six curves within each grid are SL-LRT (dotted lines), Lin's test (dash-dotted lines), CR-LRT (dashed lines), and ours based on t_{1,σ_ϵ^2} (solid lines, with squares, triangles, and circles for $\lambda_1 = -2$, 0, and 2, respectively).

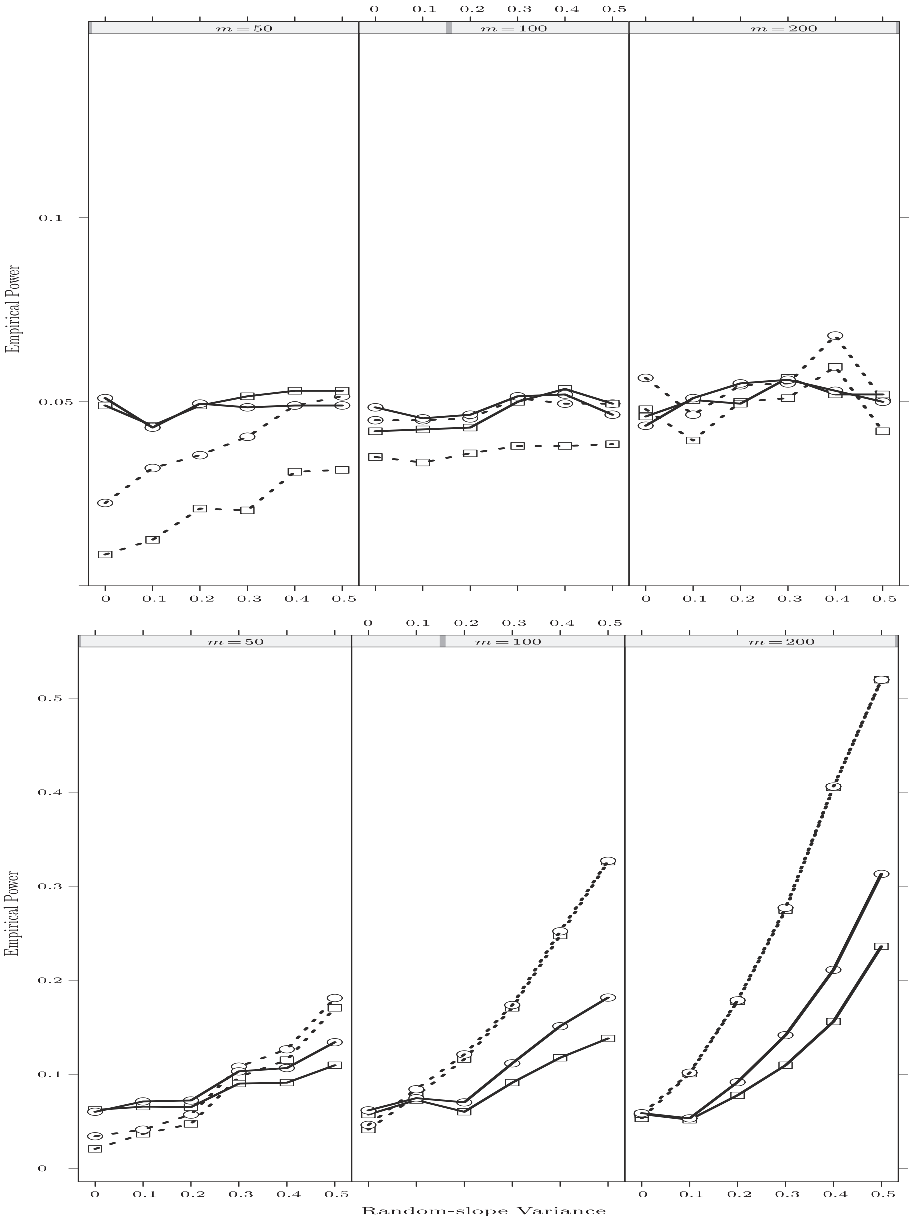

Figure 3. Empirical power of t_{1,σ_ϵ^2} and t_{2,σ_ϵ^2} versus σ_{b1}^2 for (D3) and (D4) with m = 50, 100, 200. Upper panel is for (D3): $b_{i1} \sim N(0, \sigma_{b1}^2)$; lower panel is for (D4): $b_{i1} \sim$ shifted gamma. Four curves within each grid are for $\lambda_1 = -2$ (dotted lines) and $\lambda_1 = 2$ (solid line), with squares and circles symbolizing t_{1,σ_ϵ^2} and t_{2,σ_ϵ^2}, respectively.

Table 3. Maximum likelihood estimates for the parameters in the null model (M_3) and the alternative model (M_5) in Example 5.3, along with the values of test statistics and the associated p-values. The standard errors (s.e.) are estimated according to the sandwich variance estimation.

	β_0	β_{11}	β_{12}	β_{13}	σ_{b0}^2	σ_ϵ^2
Assume M_3, $P(\Delta_i = 1 \mid Y_i) = \Phi(Y_{i1}Y_{i2}/6000)$						
Observed data	68.742	7.643	6.493	7.175	(NA)	4.630
(s.e.)	(0.356)	(0.370)	(0.324)	(0.311)	(NA)	(0.871)
Incomplete date	68.329	8.345	5.862	8.295	(NA)	3.028
(s.e.)	(0.376)	(0.263)	(0.342)	(0.829)	(NA)	(0.526)
$t_{1,\theta}$	1.967	-1.707	1.759	-1.3672	(NA)	2.391
p-value	0.056	0.095	0.086	0.179	(NA)	0.022
Assume M_5, $P(\Delta_i = 1 \mid Y_i) = \Phi(0.005Y_{i1})$						
Observed data	68.665	7.502	6.855	7.303	3.372	1.427
(s.e.)	(0.319)	(0.164)	(0.250)	(0.203)	(0.788)	(0.133)
Incomplete data	68.991	7.272	6.613	7.172	3.645	1.412
(s.e.)	(0.407)	(0.226)	(0.298)	(0.259)	(0.839)	(0.142)
$t_{1,\theta}$	-1.789	1.422	1.606	0.971	-1.541	0.322
p-value	0.081	0.163	0.116	0.338	0.131	0.749

the rat was anesthetized every ten days starting at the age of 50 days. Each rat was randomly assigned to one of three groups, control group, a low-dose group, and a high-dose group. For the latter two groups, treatment started at the age of 45 days. Some rats did not survive anesthesia and the number of measurements collected from a rat before it dropped out ranged from one to seven (see Verbeke and Molenberghs (2000, Table 2.1) for a summary of the data information). To facilitate missing response generation, we excluded four rats (out of a total of 50) that only had one measurement in the experiment, leaving data from 46 rats for analyses. Verbeke and Lesaffre (1999) and Verbeke and Molenberghs (2000, 2003) analyzed the observed data set with 50 rats using LMM. The model they considered was, for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$,

$$Y_{ij} = \beta_0 + b_{i0} + X_{ij}(\beta_1 + b_{i1}1_3) + \epsilon_{ij}, \tag{5.1}$$

where $\beta_1 = (\beta_{11}, \beta_{12}, \beta_{13})^t$, 1_3 is a 3×1 vector of ones, Y_{ij} is the response of rat i on the jth occasion, $X_{ij} = (L_i t_{ij}, H_i t_{ij}, C_i t_{ij})$, with L_i, H_i, and C_i being the indicators that take value one if the rat was in the low-dose group, the high-dose group, and the control group, respectively, and $t_{ij} = \log\{1 + (\text{age}_{ij} - 45)/10\}$. To illustrate the proposed diagnostic method, we first assumed a model without random effects, $M_3 : Y_{ij} = \beta_0 + X_{ij}\beta_1 + \epsilon_{ij}$. A relatively simple alternative model considered as a potential improved model when M_3 is in doubt is $M_5 : Y_{ij} = \beta_0 + b_{i0} + X_{ij}\beta_1 + \epsilon_{ij}$. To test M_3 versus M_5, we adopted a missingness mechanism such that, for rat i ($i = 1, \ldots, 46$), only Y_{i1} was missing with probability $\Phi(0.005Y_{i1})$, resulting in around 54% missing rate. Following the three-step testing procedure described in Section 3.2, we computed t_{1,σ_ϵ^2} to find $t_{1,\sigma_\epsilon^2} = 1.82$ with a p-value of 0.08, insignificant at the 0.05 significance level. The result from this test could imply that M_3 is an appropriate model for the current data or that M_5 with a normal b_{i0} is more adequate. To explore further, we created missing data using a different mechanism according to which Y_{i1} was missing with probability $\Phi(Y_{i1}Y_{i2}/6,000)$, leading to a missing rate of around 72%. Using this mechanism and repeating the same test, we found $t_{1,\sigma_\epsilon^2} = 2.39$, corresponding to a p-value of 0.02. The second significant test result provides evidence that the model with a normal random intercept is more preferred than M_3. Lastly, we considered whether the data provide sufficient evidence to support including a random slope as in (5.1). That is, we tested M_5 versus (5.1). Using the first missingness mechanism again to create missing data, we found that the tests associated with all parameters were insignificant at the 0.05 significance level. Hence, the observed data do not provide sufficient evidence that a random slope in necessary. This is also the conclusion reached by the analyses in Verbeke and Molenberghs (2003), when normal random effects were assumed throughout. The results from the latter two tests and the parameter estimates are given in Table 3.

5.2. Practical considerations

Now we are in the position to make some practical remarks on two issues, one about designing missingness mechanisms, and the other regarding multiple testing. If one focuses on the mechanism defined by (4.15), then the first issue boils down to the choice of λ there. It is theoretically desirable to choose λ so that $|\tilde{\Omega}^* - \tilde{\Omega}|$ is large when the model is misspecified, which usually leads to tests with high power provided data information is not too scarce. However, because $|\tilde{\Omega}^* - \tilde{\Omega}|$ rarely happens to be a trivial or monotone function of λ, it is unrealistic to expect one to know how to tune λ to increase this discrepancy, not to mention that it also depends on the unknown true model. In practice, the best one can do is to try various λ to attain several different missing rates, while bearing in mind that, first, the power of the test is typically not monotone in missing rate, and second, too large a missing rate can result in unreliable inference based on the incomplete data. A conservative rule of thumb we took from extensive simulation study is to conduct the test with a missing rate of $10\% \sim 20\%$, then repeat the test with a missing rate of $80\% \sim 90\%$, except when sacrificing this much data yields incomplete data with too little information, in which case one can lower the missing rate. If the test does not reject the null hypothesis at either missing rate, then the current data do not provide sufficient evidence against the null.

Otherwise, the data suggest some evidence against the null. One may certainly consider any missingness mechanisms other than (4.15), but we believe it overkill to consider too many (seemingly) different mechanisms. An overloaded toolbox can put one in an unnecessarily difficult situation where one does not know which tool to pick for a particular task. With the theoretical support provided in Sections 3 and 4, we recommend a simpler and lighter toolbox. The recipe for doing this is the following. First, choose one fixed index to be the index of the potentially missing response within each cluster; second, use results proved in Sections 3 and 4 as guidelines to decide on MCAR, MAR, or NMAR, depending on the goal of a test; third, set the parameter(s) in the mechanism chosen in the previous step to attain certain missing rate(s) following the aforementioned rule of thumb.

Turning to the second issue of multiple testing, even though we computed t_θ for all parameters in Ω in the above examples, we only reported results associated with $t_{\sigma_\epsilon^2}$, which sufficed for those problems. When the model under testing is more complex and one wishes to assess more model assumptions, information from tests associated with other parameters can be useful. In that case, adjustment for multiplicity in testing is the right thing to pursue. But not doing the right thing may be less disastrous here than in other statistical analyses, such as in some variable selection procedures. The reason is that the *relative* significance of the t_θ associated with different parameters may already contain useful information regarding the source of misspecification.

6. Discussion

We propose a novel diagnostic method to assess random-effects assumptions in LMM. The novelty of our proposal lies in the use of strategically created incomplete data. The incompleteness of the user-generated data is relative to the raw observed data set, which itself can be incomplete data due to missingness during the data collecting process, as in Example 5.3 where many rats missed some measurements due to early drop-out. Unwise as it appears, sacrificing some data and making inference based on incomplete data can reveal valuable information that is hard to discover when only the raw data are used. An important lesson we have learned from this research is that combing multiple sets of potentially inconsistent inference, or, combing multiple sets of “wrong-model analyses”, can reveal more useful information than one set of (potentially-)wrong-model analyses. The use of wrong-model analyses (repeatedly) distinguishes our method from those in Hausman (1978) and Tchetgen and Coull (2006), which require some form of consistent inference robust to the model assumption being tested. Not relying on some form of consistent inference actually allows more flexibility in developing diagnostic methods.

Although the context of the theoretical development in this article is LMM for responses that can be partitioned into independent subvectors, the underlying idea of using missing data to detect model misspecification is applicable in a wide range of models and to aspects of model selection other than random effects. We have experimented on several other data coarsening strategies for more complex models. These experiments produce empirical evidence that the patterns of changes in estimators before and after data coarsening are closely related to the source of misspecification. Analytic explanations of these phenomena are then more involved. One possible direction that can lead to some theoretical insight is to explore the dependence of some "distance" between two sets of inferences on the "distance" between the assumed model and the true model.

If the raw observed data are not as complete or rich as one would like, we plan to explore the enrichment of data as opposed to the coarsening of data. This opposite direction is related to the idea of imputation for model checking presented in Gelman et al. (2005). We conjecture that, when additional data generated according to a user-designed mechanism are added to the observed data, inference based on the enriched data can differ from inference based on the raw observed data in a way that is informative for model diagnosis.

Acknowledgement

I am grateful to two referees for their insightful reviews that not only lead to an improved presentation but also pointed me toward several exciting future research topics. I thank an associate editor and the Editor for their constructive comments. This research was supported by the U. S. National Science Foundation.

References

Crainiceanu, C. M. (2008). Likelihood ratio testing for zero variance components in linear mixed models. *Random Effect and Latent Variable Model Selection* (Edited by D. B. Dunson). Springer Science+Business Media.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. Roy. Statist. Soc. Ser. B.* **66**, 165-185.

Gelman, A., Mechelen, I. V., Verbeke, G., Heitjan, D. F. and Meulders, M. (2005). Multiple imputation for model checking: completed-data plots with missing and latent data. *Biometrics* **61**, 74–85.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251-1271.

Huang, X. (2009). Diagnosis of random-effect model misspecification in generalized linear mixed models for binary response. *Biometrics* **65**, 361-368.

Huang, X. (2011). Detecting random-effects model misspecification via coarsened data. *Comput. Statist. Data An.* **55**, 703-714.

Huang, X., Stefanski, L. A. and Davidian, M. (2009). Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics* **65**, 719-727.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Assoc.* **73**, 805-811.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309-326.

Litière, S. (2007). The impact of a misspecified random-effects distribution on estimation in generalized linear mixed models. Doctoral dissertation. Center for Statistics, Hasselt University, Agoralaan, Belgium.

Litière, S., Alonso, A. and Molenberghs, G. (2007). Type I and type II error under random-effects misspecification in generalized linear mixed models. *Biometrics* **63**, 1038-1044.

Litière, S., Alonso, A. and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statist. Med.* **27**, 3125-3144.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, Second Edition. Wiley, Hoboken, New Jersey.

Saville, B. R. and Herring, A. H. (2009). Testing random effects in the linear mixed models using approximate Bayes factors. *Biometrics* **65**, 369-376.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects models. *Biometrics* **50**, 1171-1177.

Tchetgen, E. J. and Coull, B. A. (2006). A diagnostic test for the mixing distribution in a generalised linear mixed model. *Biometrika* **93**, 1003-1010.

Verbeke, G. and Lesaffre, E. (1994). Large sample properties of the maximum likelihood estimators in linear mixed models with misspecified random-effects distributions. Report #1996.1, Biostatistical Centre for Clinical Trials, Catholic University of Leuven, Belgium.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* **91**, 217-221.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Comput. Statist. Data Anal.* **23**, 541-556.

Verbeke, G. and Lesaffre, E. (1999). The effect of drop-out on the efficiency of longitudinal experiments. *Appl. Statist.* **48**, 363-375.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer Verlag, New York.

Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254-262.

Verbeke, G. and Molenberghs, G. (2010). Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statist. Modelling* **10**, 391-419.

Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J. P., Carels, C., Kuhn, E. R., Darras, V. and de Zegher, F. (1998). Comparative effects of neonatal and prepubertal castration on craniofacial growth in rats. *Archives. Oral Biol.* **43**, 861-871.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distribution of random effects for longitudinal data. *Biometrics* **57**, 795-802.

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, USA.

E-mail: huang@stat.sc.edu

(Received November 2011; accepted September 2012)