Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 21

5-2016

Impact of Serial Correlation Misspecification with the Linear Mixed Model

Brandon LeBeau
University of Iowa, brandon-lebeau@uiowa.edu

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 389-416.

ISSN 1538 − 9472

Impact of Serial Correlation Misspecification with the Linear Mixed Model

Brandon LeBeau
University of Iowa
Iowa City, IA

Linear mixed models are popular models for use with clustered and longitudinal data due to their ability to model variation at different levels of clustering. A Monte Carlo study was used to explore the impact of assumption violations on the bias of parameter estimates and the empirical type I error rates. Simulated conditions included in this study are: simulated serial correlation structure, fitted serial correlation structure, random effect distribution, cluster sample size, and number of measurement occasions. Results showed that the fixed effects are unbiased, but the random components tend to be overestimated and the empirical Type I error rates tend to be inflated. Implications for applied researchers were discussed.

Keywords: Longitudinal, simulation, linear mixed model

Introduction

Linear mixed models (LMM) have become much more prominent in educational research over the past couple decades, where they are commonly known as hierarchical linear models (HLM) (Raudenbush & Bryk, 2002) or multilevel models (Goldstein, 2010). The mixed portion in the linear mixed model indicates that the model has both fixed and random effects present in the model. These models have become more widely used for a couple of reasons: 1) the advancements in computing which allow for easier and quicker estimation, 2) the notice of the need to model the hierarchical or nested nature of the data, and 3) handles unbalanced data/designs well without any additional work. A few common data collection settings in education where LMM are used include: students nested within classrooms or students nested within schools. For some additional examples of how these models are used in education see Bryk and Raudenbush (1987) and Raudenbush (1988).

Dr. LeBeau is an Assistant Professor in the Educational Measurement and Statistics Program. Email him at: brandon-lebeau@uiowa.edu.

Research Problem

In longitudinal studies, the repeated measures for the same person are likely to be more similar due to the fact that the same person is being measured multiple times on the same measurement scale (Littell, Henry, & Ammerman, 1998). The multiple measurements brings about a dependency due to repeated measurements, or alternatively, there is less information available as the measurement occasions within an individual are correlated. This dependency can be accounted for in the LMM by specifying random effects at the cluster level, the level one covariance matrix, or a combination of the two. In most cases, researchers allow the random effects to account for the dependency due to repeated measures and assume that the variance is the same across the observations with no correlation between the observations (e.g. the correlation between observation one and observation two is zero) at level one. This level one structure is often called an independence structure. For certain repeated measures designs, especially when the repeated measures are collected close in time or correlations among the repeated measures do not decay quickly, random effects alone may not adequately account for the dependency due to the repeated measures and a more complex covariance structure at level one may be needed (Browne & Goldstein, 2010; Goldstein, Healy, & Rabash, 1994).

Unfortunately, few simulation studies have looked at these implications (Ferron, Dailey, & Yi, 2002; Kwok, West, & Green, 2007; Murphy & Pituch, 2009) in a LMM framework. The current study looks to add to this literature by exploring possible implications of misspecifying the level one covariance structure using a computer simulation. The primary question of interest will be the extent to which the misspecification of the variance matrix for the repeated measures biases the parameter estimates (and ultimately inferences as well) for the fixed and random portion of the LMM. Interactions to other assumption violations will also be explored.

The Model

A basic linear mixed model can be written as follows:

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_{j} + \mathbf{e}_{ij} \tag{1}$$

In this model, the $\mathbf{Y}_{ij}$ is the response variable for the i^{th} level 1 unit nested within the j^{th} level 2 unit. Next is the $\mathbf{X}_{ij}$, which is an $n_i \times p$ matrix of covariates in the model (also known as the design matrix) where n_i is the total number of observations for every individual and p is the number of covariates. This matrix

includes covariates at both level 1 and level 2 as well as covariates that are aggregated over the level 1 units. The $\boldsymbol{\beta}$ in the model is a $p \times 1$ vector representing the fixed effects. Next is the $\mathbf{Z}_{ij}$ which is the design matrix for the random effects. This term is commonly formed from a subset of the columns of $\mathbf{X}_{ij}$. The $\mathbf{b}_j$ are the random effects and are unique for each level 2 unit but are the same for each level 1 unit within a given level 2 unit. The random effects represent the deviation of the j^{th} subject from the group or average growth curve. Finally, the $\mathbf{e}_{ij}$ are the level 1 residuals (i.e. measurement or sampling error) similar to simple linear regression. These represent deviations from the individual growth curves.

This model can also be expressed in matrix form:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \mathbf{e}_j \tag{2}$$

Model Assumptions

Just like any statistical model, there are model assumptions that need to be satisfied (at least approximately) in order for parameter estimates and inferences to be unbiased. The model assumptions for the LMM are as follows (Raudenbush & Bryk, 2002):

1. The random effects $\mathbf{b}_j$ are independent across level 2 units, normally distributed (multivariate normal when more than one random effect is in the model), and each has a mean of 0 and a covariance matrix $\mathbf{G}$. This can be succintly written as: $\mathbf{b}_j \sim$ iid N(0, $\mathbf{G}$).
2. Each of the $\mathbf{e}_{ij}$ are independent and follow a normal distribution with mean 0 and variance σ^2 for every level 1 unit within level two. This can be summed up as: $\mathbf{e}_{ij} \sim$ iid N(0, σ^2).
3. The $\mathbf{e}_{ij}$ are independent of the random effects.

The models considered in this paper are assumed to have a continuous response variable with at least an interval scale of measurement and the within individual errors (i.e. level one errors) are assumed to be approximately normally distributed.

Violation of Model Assumptions

Simulation studies that have data conditions similar to longitudinal data have found little evidence of parameter bias in the fixed or random effects when the random effect distributions are non-normal. However, these studies have reported

confidence intervals for the variance of the random effects with poor coverage when the random effect distributions are not normal, specifically chi-square with one degree of freedom and Laplace distributions (Maas & Hox, 2004a; 2004b). This suggests that the standard errors are underestimated for the variance components of the random effects.

Sample size considerations for the LMM is an important consideration when planning a study. This is especially true since maximum likelihood is asymptotic and require large sample sizes for proper estimation (Maas & Hox, 2004a). Typically, the highest level sample size is of most concern as there are fewer numbers at this level (Maas & Hox, 2004a). This issue is commonly exacerbated for longitudinal studies as the level 1 sample size tends to also be small (i.e. few observations per subject); where 10 observations per subject is considered large (Snijders & Bosker, 1993). Unfortunately, there have been few studies that have studied small level 1 sample sizes commonly found in longitudinal studies.

Simulation studies that have looked at the sample size needed for unbiased estimates for the parameters in general have not found any problems with estimating the fixed effects at level 1 or level 2 (Maas & Hox, 2004a; 2005b; 2005; Browne & Draper, 2000). Additionally, the standard errors for the fixed effects are generally estimated accurately with at least 30 groups (Maas & Hox, 2004a; 2005).

Covariance Structures

The variance structure for the response variable is an important aspect of the LMM; this is where the dependency due to the repeated measures is taken into account. The equation for the variance of the response variable is

$$\mathrm{Var}\left(\mathbf{Y}_j\right) = \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_j\left(\theta\right) = \mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathrm{T}} + \sigma_e^2\mathbf{I}_{n1j} \tag{3}$$

As can be seen from the above equation, the variance is composed of two portions, $\mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathrm{T}}$ is the portion of the variance that is accounted for by the random effects and the $\sigma_e^2\mathbf{I}_{n1j}$ is the portion that is accounted for by the level 1 error.

Commonly, researchers choose a simple level 1 error structure. The most common structure specified by researchers assumes homogeneity of variance with no correlation between the time points, known as the independence structure. An example of such a matrix with four time points is as follows:

$$\begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix} \quad (4)$$

where σ_e^2 represents a common variance across the four time points.

Complex variance structures can be achieved by including multiple random effects (e.g. random effects for intercept, time, time2, etc.) and specifying a complex level one error structure. For example, if a researcher fits a model with a random effect for intercept and an independence level one error structure. The covariance structure for the model would look as follows (assuming four time points):

$$\begin{bmatrix} \sigma_e^2 + g_{11} & g_{11} & g_{11} & g_{11} \\ g_{11} & \sigma_e^2 + g_{11} & g_{11} & g_{11} \\ g_{11} & g_{11} & \sigma_e^2 + g_{11} & g_{11} \\ g_{11} & g_{11} & g_{11} & \sigma_e^2 + g_{11} \end{bmatrix} \quad (5)$$

Here σ_e^2 represents the error variance and g_{11} represents the variance of the random intercepts. As can be seen from (5) above when a random intercept is included in the model and an independence structure is assumed at level one, the covariance structure follows a compound symmetry structure (which is what is assumed by RM-ANOVA). Although this structure is not very complex and likely not justifiable for many longitudinal studies, adding more random effects (i.e. a random effect for time) or specifying a more complicated level one error structure (e.g. first order autoregressive, toeplitz, etc.) would produce a more complex covariance structure.

With the inclusion of more complicated error terms, it can be helpful to include additional notation for the level one residual to separate random error and serial correlation denoted as $e_j = e_{(1)j} + e_{(2)j}$. Here $e_{(1)j}$ represents random error and $e_{(2)j}$ represents serial correlation. Serial correlation can be thought of as a random process of an observed profile within an individual that usually decreases as the time lag increases (Diggle, 2002). More simply, serial correlation represents the correlation between two observations on the same individual that depends solely on the time lag between the observations. Explicitly showing the serial correlation

and random error separately in the variance of the response variable leads to the following expression:

$$\text{Var}\left(\mathbf{Y}_j\right) = \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_j\left(\theta\right) = \mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\text{T}} + \sigma_e^2\mathbf{I}_{n1j} + \tau^2\mathbf{H}_j \quad (6)$$

Different from (3) above, serial correlation is explicitly shown as $\tau^2\mathbf{H}_j$, where $\mathbf{H}_j$ is an $n_j \times n_j$ matrix where the $(j, k)^{\text{th}}$ element is the correlation between two time points within an individual.

Most researchers when using a LMM tend to assume the level one residual structure follows an independence structure without taking into account the type of data (i.e. cross sectional or longitudinal data). This may be chosen due to the parsimonious nature of the independence model or the researcher believes that including more random effects adequately accounts for the dependency due to repeated measures. However, the following question must be asked, after removing the variation due to the random effects are the within individual residuals independent from one another within an individual (Browne & Goldstein, 2010)? In other words, conditional on the random effects, is it tenable to assume that the within individual residuals are independent? This assumption may not hold in some data situations, especially if the time between observations is very short (i.e. daily or weekly observations) or if the correlation between observations does not decrease very quickly (Browne & Goldstein, 2010; Goldstein et al., 1994). If the level one residuals are not independent of one another, then the level one structure takes a form similar to time series models. See Box and Jenkins (1976) to explore time series models.

Misspecification of the Covariance Structure

There was quite a bit of interest earlier in the history of the LMM on adequately modeling the covariance structure (Chi & Reinsel, 1989; Diggle, 1988; Goldstein et al., 1994; Keselman, Algina, Kowalchuk, & Wolfinger, 1998; 1999; Núñez-Antón & Zimmerman, 2000; Wolfinger, 1996). However, only recently have simulation studies started exploring the impact of misspecification of the level one residual structure (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009). Kwok et al. (2007) defined three useful terms to use when talking about misspecification of the covariance structure: underspecified, overspecified, and general-misspecification. An underspecified covariance structure (US) occurs when the specified matrix is simpler but nested within the true covariance matrix (e.g. compound symmetry is chosen but the true structure is AR(1)). An

overspecified covariance structure (OS) occurs when the specified matrix is more complex than the true covariance matrix but the true covariance matrix is nested within the specified matrix (e.g. ARMA(1, 1) structure chosen but AR(1) is the true structure). Lastly, general-misspecification (GS) occurs when the specified and true covariance matrices are not nested (e.g. TOEP(2) structure chosen but AR(1) is the true structure).

Simulation studies have found little to no bias for fixed effect estimates, however there is evidence of bias in the estimates for the standard errors of the fixed effects (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009). When the covariance structure was US or GS the standard errors for the within-individual intercept and slope were overestimated (Kwok et al., 2007). Not suprisingly, the bias in the variance components can be quite substantial when the covariance structure is ignored. If the covariance structure was US or GS $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$ were overestimated (Ferron et al., 2002; Kwok et al., 2007); OS covariance structures produced the smallest estimates for $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$ (Kwok et al., 2007). As a result of the overestimated $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$, $\hat{\sigma}^2$ tended to be underestimated to compensate (Ferron et al., 2002). Murphy and Pituch (2009) even found that the variance components are biased even when the correct covariance structure was modeled.

These results produced the following general guidelines: if the researcher is only interested in estimates of the fixed effects (i.e. group level estimates) then the researcher may not need to model the covariance structure. However, if the researcher is interested in the variance components, individual growth curves, inferential statisics, or model predictions the researcher should explore alternative structures for the level one covariance structure (Ferron et al., 2002; Kwok et al., 2007; Verbeke & Molenberghs, 2000).

Selecting a Covariance Structure

In most cases when researchers use a LMM, they are interested in doing more than just looking at the fixed effect estimates and some care should be taken to select a covariance structure. However there are no strong descriptive or hypothesis testing procedures to detect serial correlation. The few studies that have explored methods of selecting and detecting serial correlation have found it difficult to empirically select the correct structure (Ferron et al., 2002; Keselman et al., 1998). Another study by Verbeke and Molenberghs (2000) showed that including the serial correlation regardless if it is correctly modeled, is more important than correctly modeling the serial correlation.

There are alternative criteria that can be used for selecting the best covariance structure based on the data, these are: Akaike's Information Criterion (AIC), Schwartz's Bayesian Criterion (SBC), or a likelihood ratio test (LRT). Ferron et al. (2002) found that the AIC on average identified the correct structure about 79% of the time. The SBC and LRT identified the correct model less frequently, on average 66% and 71% of the time respectively. However, the variability in correct identification was very large, the AIC ranged from 7% to 100%. Increasing the number of time points, increasing the sample size, and higher levels of autocorrelation improved correct identification (Ferron et al., 2002). In contrast to Ferron et al. (2002), Keselman et al. (1998) found that the AIC or SBC were only able to correctly identify the covariance structure 47% and 35% of the time respectively. The large variability and conflicting results leaves uncertainty in how the researcher should proceed when they desire a test to help decide if serial correlation is present and should be modeled.

Methodology

A factorial research design was used for the computer simulation study. Previous simulation work (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009) have assessed covariance misspecification under perfect model conditions (i.e. normally distributed random effects and residuals); however, a classic study by Micceri (1989), showed that real world data are rarely normally distributed and can deviate quite substantially from a normal distribution. Therefore, simulating conditions more representative of real world data can help inform researchers to the robustness of the estimation algorithm, specifically under small sample size conditions. In addition, missing data tends to be the rule rather than the exception for longitudinal data where the likelihood of missing data commonly increases as time increases (i.e. more likely to encounter more missing data further along in the study). Understanding the implications of covariance misspecification under more common real world data conditions would be helpful and this simulation attempts to inform this area.

In order to simulate conditions that are common in real world data and improve external validity but yet keep the simulation design manageable, the following data conditions were manipulated: the covariance structure (five levels: ID, AR(1), MA(1), MA(2), ARMA(1, 1)), the random effect distribution (three levels: Normal, Laplace, Chi–Square(1)), number of subjects (two levels: 25, 50), and the number of measurement occasions (two levels: 6, 8). This leaves a total of 5*3*2*2 = 60 simulated data conditions. To avoid finding a single extreme data

condition, five hundred replications were generated for each simulated data condition resulting in 60*500 = 30,000 total datasets. Statistics were averaged across the 500 replications within each of the 60 simulation conditions. For each dataset, all five of the covariance structures were fitted (i.e. ID, AR(1), MA(1), MA(2), ARMA(1, 1)), resulting in a total of 30,000*5 = 150,000 models.

Data

Population parameters were generated from data collected by the Minnesota Mathematics Achievement Project (MNMAP). The MNMAP project collected data exploring the relationship between high school mathematics curriculum and subsequent college mathematics grades and course taking for students graduating from a high school in an upper Midwestern state. A retrospective cohort design was used in collecting the data from three sources: high schools, universities or colleges, and the state. The resulting dataset contained student, high school, and college information on more than 20,000 students, from about 300 high schools, and approximately 35 two and four year colleges or universities. In this model, student semester GPA from a college mathematics course will serve as the dependent variable. Time was the primary within-subject variable, ACT score will serve as the single continuous student level predictor and difficulty of the college mathematics course will serve as a time varying covariate. The intercepts and the slope for time were allowed to vary for every student (i.e. a random intercept and a random slope for time were specified in the model). Additional information about the data collection procedures from the MNMAP project can be seen in Harwell et al. (2009) and Post et al. (2010).

Data were simulated according to the following model:

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{diff}_{ij} + \beta_3 \text{ACT}_j + \beta_4 \text{ACT}_j : \text{time}_{ij} \\ & + b_{0j} + b_{1j} \text{time}_{ij} + e_{(1)ij} + e_{(2)ij} \end{aligned} \tag{7}$$

In this equation, let i represent repeated measurements and j represent individuals. The fixed effects are represented by β_0, β_1, β_2, β_3, and β_4, time_{ij} represents the within subject time metric, diff_{ij} is a within subject time varying covariate representing the difficulty of the mathematics course, and ACT_j is a continuous subject level covariate representing the mathematics ACT score for each subject. The random components of the model are represented by b_{0j}, b_{1j}, $e_{(1)ij}$,

Table 1. Parameter values for all terms

Parameter	Value
β_0	2.639
β_1	-0.014
β_2	-0.187
β_3	0.095
β_4	0.003
Var b_{0j}	0.552
Var b_{1j}	0.015
Var e_{ij}	0.549
Φ_1	0.450
θ_1	0.500
θ_2	0.300
Var $diff_{ij}$	1.250
Var ACT_{1j}	4.905

Note: Var – Variance

and $e_{(2)ij}$ which represent subject specific deviations from the average intercept and slope, deviations from the subject specific growth curves, and serial correlation respectively. Data were simulated from the model shown in (7), where the $e_{(2)ij}$ and the distribution of the random components were the primary differences between the simulated data.

Table 1 shows the population values used to generate the data according to (7). Table 1 reveals that many parameter values are quite small and are reflective of the scale of the dependent variable ranging from zero to four. Of particular note are the small values for β_1, β_4, and Var b_{1j} representing the slope for time, the interaction between time and mathematics ACT score, and lastly the variance of the random slopes for time. These small values will have to be kept in mind later as the bias statistic chosen divides by the parameter value.

Analysis

Model convergence, relative bias, and type I error rates were generated for all 150,000 models fitted. Relative bias was computed for all of the fixed effects and the variance components. The formula for relative bias took the form of:

$$\text{Rel. Bias} = \frac{\hat{\theta} - \theta}{\theta} \tag{8}$$

where $\hat{\theta}$ is the parameter estimate (i.e. β_k or Var(b_{1j}) and θ is the parameter value set in the simulation.

The Type I error rate was computed as the proportion of significant fixed effect estimates out of the total number of replications. That is, a Wald test statistic was set up of the form:

$$Z = \frac{\hat{\beta} - \beta}{\text{SE}} \tag{9}$$

where $\hat{\beta}$ is the parameter estimate, β is the simulated paramater value shown in Table 1, and SE is the empirical standard error calculated from the model fit. The Wald test statistic was assumed to follow a standard normal distribution. If there is no bias and the type I error rate is accurate, approximately 5% of the parameter estimates should fall outside of ± 1.96 quantile of the standard normal distribution.

Since a simulation is similar to a completely randomized experiment, the relative bias and type I error rates served as dependent variables and the simulated conditions were treated as independent variables or factors. These variables were analyzed descriptively and inferentially to answer the research questions depicted above.

Inferential Analyses

All of the simulation factors are between-subject factors except for the covariance structure factor which was a within-subject factor as all five covariance structures were fitted to each simulated dataset. Due to the within-subject factor, repeated measures analysis of variance (RM-ANOVA) is a common analysis for this type of data. However, the RM-ANOVA procedure can make interpretation more difficult and increase the burden during estimation. Another data analysis option was to treat all the design factors as between-subject factors and use univariate analysis of variance (UANOVA) to estimate the effects. The UANOVA procedure has the disadvantage of reduced power of the within-subject and mixed interaction effects (i.e. the interaction between the within-subject and between-subject effects). However, with a large sample size in the study (30,000*5 = 150,000 total cases in the main analysis) statistical power was not deemed an issue and the UANOVA model was fitted to ease interpretation. A similar analysis was done by Kwok et al. (2007) in their article addressing misspecification of the covariance structure.

The initial UANOVA model that was fitted to the relative bias data took the following structure:

$$\begin{aligned} Y_{ijklmn} = \mu &+ \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(l)} + \alpha_{D(m)} + \alpha_{E(n)} + \alpha_{AB(jk)} + \alpha_{AC(jl)} + \alpha_{AD(jm)} \\ &+ \alpha_{AE(jn)} + \alpha_{BC(kl)} + \alpha_{BD(km)} + \alpha_{BE(kn)} + \alpha_{CD(lm)} + \alpha_{CE(\ln)} + \alpha_{DE(mn)} \\ &+ \alpha_{ABC(jkl)} + \alpha_{ABD(jkm)} + \alpha_{ABE(jkn)} + \alpha_{ACD(jlm)} + \alpha_{ACE(jln)} + \alpha_{ADE(jmn)} \\ &+ \alpha_{BCD(klm)} + \alpha_{BCE(kln)} + \alpha_{BDE(kmn)} + \alpha_{CDE(lmn)} + \alpha_{ABCD(jklm)} + \alpha_{ABCE(jkln)} \\ &+ \alpha_{ACDE(jlmn)} + \alpha_{BCDE(klmn)} + \alpha_{ABCDE(jklmn)} + e_{ijklmn} \end{aligned} \tag{10}$$

The above equation represents a factorial UANOVA that fits all possible interactions. In (10), the α represent cell means, μ is the grand mean, the first set of subscripts, *A*, *B*, *C*, *D*, and *E*, represent the five simulation conditions, the subscripts in parentheses, *j*, *k*, *l*, *m*, and *n*, index the factor categories, and *i* depicts the observation number. The model for the empirical type I error rates is simplified compared to (10) because there was only one observation per cell. As a result, all four and five-way interactions were pooled into the error term.

Lastly, significance tests were not used due to the large sample size and statistical power. Instead, effects sizes were computed to determine which factors explained the most variation in the dependent variable. An η^2 statistic was used as the effect size in this analysis and took the following form:

$$\eta^2 = \frac{\mathrm{SS}_{\mathrm{trt}}}{\mathrm{SS}_{\mathrm{total}}} \tag{11}$$

In the above equation, $\mathrm{SS}_{\mathrm{trt}}$ is the amount of variation attributable to the treatment of interest (e.g. covariance structure) and $\mathrm{SS}_{\mathrm{total}}$ is the total sum of squares or the total amount of variation in the dependent variable. η^2 values greater than .001 and .01 were deemed important predictors for the relative bias and empirical type I error rates respectively.

Software

Data generation, model fitting, and analyses were conducted with R (R Development Core Team, 2010). Data generation was undertaken via an author written program. In order to replicate the results, a random seed was chosen and to ensure independent replications, the random number generation was based on the

procedure by L'Ecuyer (L'Ecuyer, Simard, Chen, & Kelton, 2002). This procedure has the advantage of producing very large strings of random numbers without worrying about duplication and supports multiple threads of random number generation which allowed multiple cores of the processor to be used simultaneously improving the data simulation speed. Model fitting was done with the nlme package found in R (Pinheiro, Bates, DebRoy, & Sarkar, 2012). Lastly, in order to check the simulated data conditions, the sample autocorrelation function was plotted to see if the values approximately followed the theoretical autocorrelation function. In addition, the empirical skewness and kurtosis of the simulated random effect distribution was computed to check for accurate random effect simulation. No significant deviations were found.

Results

The convergence rates for study one can be seen in Table 2. This table breaks down the convergence rate of the estimation algorithm by the generated and fitted serial correlation structures. As can be seen from the table, convergence rates tended to be low ranging from a low of 41.6% to a high of 95.9%. Low convergence rates tended to occur when the serial correlation structure was overspecified (e.g. ARMA(1, 1) structure fitted to an AR(1) structure) or when a generally misspecified serial correlation structure was fitted (e.g. AR(1) structure fitted to a MA(1) structure). In general, the AR(1) and ARMA(1, 1) fitted structures had the worst convergence rate compared to the other fitted structures and the independent structure had the best convergence rate, which is not surprising as no additional terms were needed to be estimated with an independent structure.

Relative Bias

Summary statistics for the relative bias of the fixed effects can be seen in Table 3. The table shows that although the mean and median for all of the parameters were very close to zero, the slope terms (i.e. β_1 and β_4) had large amounts of variation as shown by the variance in Table 3. The large amount of variation in the relative bias for those two terms is likely attributable to the small parameter values as seen in Table 1 (i.e. to get the relative bias, the absolute bias was divided by the parameter value which are small for β_1 and β_4).

Table 2. Convergence rates by generated and fitted serial correlation structure

Gen SC	Fit SC	Convergence %
Ind	Ind	72.48
Ind	AR(1)	68.38
Ind	MA(1)	71.02
Ind	MA(2)	67.23
Ind	ARMA(1, 1)	65.10
AR(1)	Ind	93.88
AR(1)	AR(1)	64.88
AR(1)	MA(1)	81.37
AR(1)	MA(2)	70.78
AR(1)	ARMA(1, 1)	60.45
MA(1)	Ind	92.23
MA(1)	AR(1)	55.12
MA(1)	MA(1)	69.15
MA(1)	MA(2)	65.93
MA(1)	ARMA(1, 1)	63.68
MA(2)	Ind	95.62
MA(2)	AR(1)	61.98
MA(2)	MA(1)	84.50
MA(2)	MA(2)	68.83
MA(2)	ARMA(1, 1)	54.88
ARMA(1, 1)	Ind	98.37
ARMA(1, 1)	AR(1)	42.17
ARMA(1, 1)	MA(1)	88.02
ARMA(1, 1)	MA(2)	72.90
ARMA(1, 1)	ARMA(1, 1)	63.60

Note: Gen – generated, SC – serial correlation, Fit – fitted

Table 3. Summary statistics for relative bias of fixed effects

Term	Mean	Var	Med	Min	Max
β_0	0.0005	0.0054	0.0004	-0.3581	0.4424
β_1	0.0606	26.6853	0.1011	-26.8454	25.1670
β_2	0.0010	0.0905	0.0010	-1.5945	1.7359
β_3	-0.0016	0.1882	-0.0025	-2.4923	2.4803
β_4	0.0579	24.6815	0.0357	-28.2912	30.8497

Note: Var – variance, Med – median, Min – minimum, Max – maximum

The variation in the relative bias for the parameters was explored using ANOVA. No four or five-way interactions had $\hat{\eta}^2$ greater than .001 and were dropped from the models, however all two and three-way interactions were retained.

The results of these final ANOVAs and the resulting $\hat{\eta}^2$ can be seen in Table 4 for all five fixed effect parameters and the variance of the random components. The values in bold in the table are $\hat{\eta}^2$ statistics that are larger than .001.

Looking at the first five columns of Table 4 reveals there are no large $\hat{\eta}^2$ statistics for any of the fixed effects. This means that the simulation conditions do not explain a significant amount of variation in the relative bias of the fixed effects. This suggests that the grand mean relative bias for each of the fixed effects acts as an adequate summary measure for each fixed effect and can be seen in Table 3. Exploring the simple averages shows that relative bias for the two slope terms (i.e. β_1 and β_4) have the largest bias statistics. Even though the slope terms showed slight evidence of bias (.0606 and .0579 for β_1 and β_4 respectively), the relative bias statistic is quite small and would likely not seriously distort any findings.

Summary statistics for the relative bias of the random components can be seen in Table 5. The table shows that on average the variance of the random components tends to be biased and there was significant variation in the relative bias statistics for each term. Since variances can only be positive, it is not surprising that the minimum relative bias is small (approximately -1) compared to the maximum relative bias (approximately 10, 35, and 6.6 for variance of intercept, slope, and within cluster residuals respectively).

The variation in the relative bias statistics for the random components were explored with an ANOVA and the $\hat{\eta}^2$ can be seen in the last three columns of Table 4. These columns reveal that there are variables that explain variation in the relative bias of the random components (i.e. $\hat{\eta}^2 > 0.001$). The strongest effects were the simulated conditions related to the generated and fitted serial correlation structure.

The significant interaction between the generated and fitted serial correlation structures for the random effects are explored in Figure 1. These figures show that fitting an underspecified independence structure has severe consequences in terms of relative bias of the variance of the random effects. More specifically, when an AR(1), MA(1), MA(2), or ARMA(1, 1) structure underlie the data, the independence serial correlation structure produces significantly greater bias compared to fitting other serial correlation structures. For example, when an ARMA(1, 1) structure underlies the data and the serial correlation structure is underspecified as independent, the variance of the intercept and slope are overspecified by over 1.5 times and at least 6 times respectively.

Table 4. Eta-squared statistics for all terms from ANOVA models

Variable	$\hat{\eta}^2\beta_0$	$\hat{\eta}^2\beta_1$	$\hat{\eta}^2\beta_2$	$\hat{\eta}^2\beta_3$	$\hat{\eta}^2\beta_4$	$\hat{\eta}^2$Var b_0	$\hat{\eta}^2$Var b_1	$\hat{\eta}^2$Var Res
N	0.0000	0.0000	0.0001	0.0000	0.0002	**0.0023**	**0.0123**	**0.0014**
p	0.0001	0.0000	0.0000	0.0001	0.0001	0.0010	**0.0136**	**0.0031**
RE Dist	0.0001	0.0004	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000
Gen SC	0.0006	0.0008	0.0003	0.0003	0.0001	**0.0937**	**0.0930**	**0.1704**
Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0904**	**0.0862**	**0.1984**
N:*p*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0001
N:RE Dist	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0000
N:Gen SC	0.0002	0.0001	0.0003	0.0001	0.0001	0.0002	0.0006	0.0003
N:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0004
p:RE Dist	0.0002	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
p:Gen SC	0.0004	0.0004	0.0001	0.0004	0.0000	0.0002	**0.0013**	0.0005
p:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0001
RE Dist: Gen SC	0.0002	0.0003	0.0003	0.0004	0.0001	0.0002	0.0003	0.0001
RE Dist: Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gen SC:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0670**	**0.0548**	**0.1658**
N:*p*:RE Dist	0.0001	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0001
N:*p*:Gen SC	0.0002	0.0000	0.0002	0.0002	0.0001	0.0001	0.0002	0.0000
N:*p*:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
N:RE Dist:Gen SC	0.0002	0.0004	0.0001	0.0004	0.0006	0.0001	0.0002	0.0003
N:RE Dist:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N:Gen SC:Fit SC	0.0000	0.0001	0.0000	0.0000	0.0000	0.0006	0.0002	**0.0019**
p:RE Dist:Gen SC	0.0002	0.0001	0.0004	0.0001	0.0003	0.0003	0.0004	0.0001
p:RE Dist:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
p:Gen SC:Fit SC	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001	0.0006	0.0005
RE Dist:Gen SC:Fit SC	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Note: Bold numbers are > 0.001, *N* – cluster sample size, *p* – within cluster sample size, Gen – generated, RE Dist – random effects distribution, SC – serial correlation, Fit – fitted, ":" represents an interaction

Table 5. Summary statistics for relative bias of random components

Term	Mean	Var	Med	Min	Max
$\hat{\eta}^2$Var b_{0j}	0.4012	0.6942	0.2904	-1.0000	10.0186
$\hat{\eta}^2$Var b_{1j}	1.9116	9.2561	1.1211	-1.0000	35.4700
$\hat{\eta}^2$Var Res	0.1222	0.2645	-0.0151	-0.7943	6.6436

Note: Var – variance, Med – median, Min – minimum, Max – maximum

The AR(1) and ARMA(1, 1) fitted structures tend have the smallest bias statistics for the variance of the random effects compared to the other structures, which may suggest that the moving average component does not aid in modeling

serial correlation in longitudinal data. Lastly, even when the correct structure is modeled there is still evidence of bias in the variance of the random effects and in many cases the correct fitted structure does not produce the smallest average relative bias statistics.

Lastly, Figure 2 shows that the variance of the residuals tend to be underestimated when an underspecified independence structure is fit, however this underestimation is not as large as the overspecification found in the random effects. The largest amount of bias occurs when the underlying structure is ARMA(1, 1), which tends to produce average relative bias statistics for the residuals that are comparable to the average relative bias for the variance of the intercept. Except for the systematic underestimation when an independence structure was fitted when serial correlation was present, the average relative bias still tends to be positive suggesting that all of the random components are overestimated when serial correlation is present.

Type I Error Rate

Even though there was no evidence of bias in the fixed effects under any of the simulated data conditions, the random components did show evidence of bias; therefore, the standard errors of the fixed effects may not be accurate. This may cause the type I error rate to be too conservative (type I error rate smaller than the specified α) or too liberal (type I error rate greater than the specified α).

Box plots can be seen in Figure 3 and show the empirical type I error rates for each of the fixed effect parameters. This figure shows that the median empirical type I error rate for the fixed effects tends to be slightly above the expected $\alpha = 0.05$, however β_0 and β_3 both include 0.05 in the middle 50% of the distribution. β_0, β_1, and β_4 have median type I error rates around 0.06, whereas β_2 has a median around 0.07. The variability in the five box plots tend to be similar indicated by the size of the interquartile range. Since there does appear to be variability in the empirical type I error rates, these will be modeled inferentially. Table 6 shows the $\hat{\eta}^2$ statistics for the empirical type I error rates for all terms up to three-way interactions. All higher order interaction terms were pooled into the error.

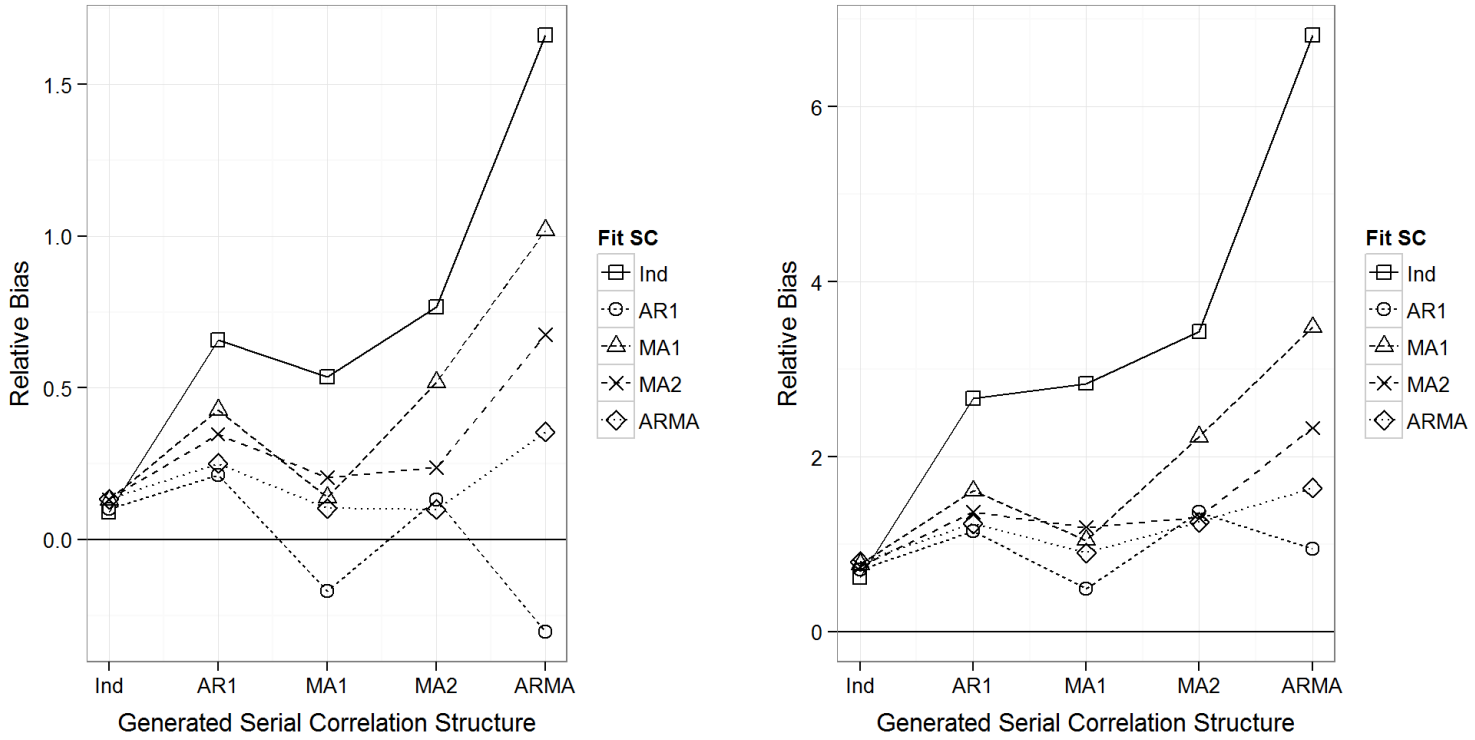

Figure 1. Relative bias of random effects by generated and fitted serial correlation structure; variance of b_{0j} (left) and b_{1j} (right)

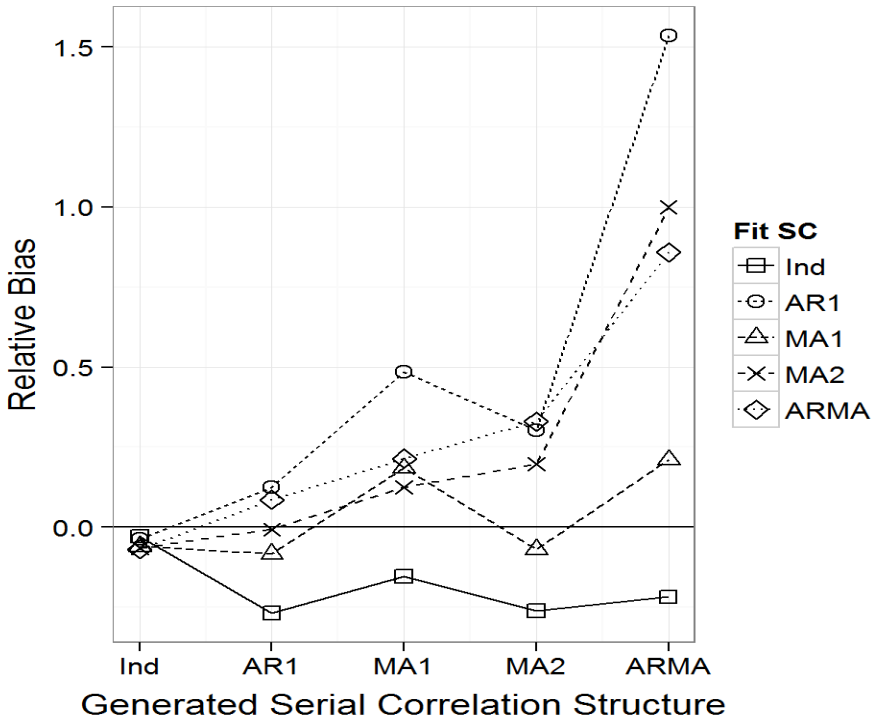

Figure 2. Relative bias of the variance of the residuals by generated and fitted serial correlation structures

As can be seen from the table there were numerous effect sizes greater than 0.01. Some of the largest effects were the cluster sample size, the interaction between the generated serial correlation structure and random effect distribution, and the three way interactions between the generated serial correlation structure, the random effect distribution, and the cluster sample size or the within cluster sample size. These large effects were around 0.10 suggesting that approximately 10% of the variation in the type I error rates can be explained by each of these terms.

The average empirical type I error rate for β_0 by the generated serial correlation structure, random effect distribution and the cluster sample size can be seen in Figure 4. From the figure, cluster sample sizes of 25 tend to have larger average type I error rates compared to cluster sample sizes of 50. There also was a lot of variability in the average type I error rate as the generated serial correlation structure differs, with the AR(1) structure having the smallest amount of variation. The empirical type I error rate was the smallest when the simulated random effect distribution was normally distributed.

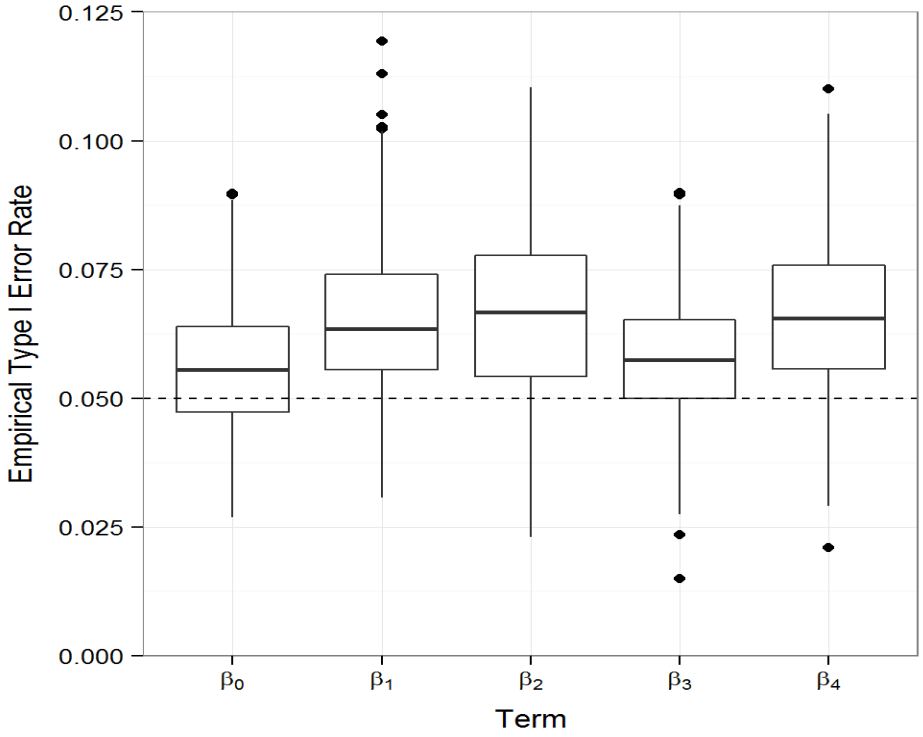

Figure 3. Box plot of type I error rates by parameter

Lastly, the scale of the *y*-axis should be taken into account. Although there is variability in the average type I error rates, this variability ranges from about 0.04 to just over 0.07 with an even smaller range when the cluster size is 50. Even though most conditions are inflated, they may not be inflated enough to significantly concern applied researchers.

Patterns for the empirical type I error rates were similar for the other parameters (i.e. $\beta_1,\ldots,\beta_4$) and are not presented graphically. In addition, the patterns were also similar for the three way interaction between the generated serial correlation structure, random effect distribution, and within cluster sample size and these graphs are not presented. The range of possible average empirical type I error rates were smaller for this second three way interaction compared to the one shown in Figure 4.

Table 6. Eta-squared statistics for all terms from ANOVA models

Variable	$\hat{\eta}^2\beta_0$	$\hat{\eta}^2\beta_1$	$\hat{\eta}^2\beta_2$	$\hat{\eta}^2\beta_3$	$\hat{\eta}^2\beta_4$
N	**0.0108**	**0.1111**	**0.1014**	**0.0150**	**0.0866**
p	0.0037	0.0005	**0.0152**	0.0000	0.0065
RE Dist	**0.1133**	**0.0119**	**0.0617**	**0.0282**	**0.0286**
Gen SC	**0.0416**	**0.0518**	**0.0338**	**0.0196**	**0.0857**
Fit SC	0.0086	**0.0145**	**0.1579**	0.0049	**0.0137**
N:*p*	**0.0476**	**0.0385**	**0.0240**	**0.0129**	0.0038
N:RE Dist	**0.0160**	**0.0147**	**0.0631**	0.0072	0.0066
N:Gen SC	**0.0300**	0.0090	**0.0352**	**0.1305**	**0.0755**
N:Fit SC	0.0037	0.0030	0.0079	0.0017	0.0024
p:RE Dist	**0.0102**	**0.0188**	0.0096	0.0075	**0.0638**
p:Gen SC	**0.0468**	**0.0306**	0.0027	**0.0581**	**0.0356**
p:Fit SC	0.0030	0.0025	0.0034	**0.0131**	0.0088
RE Dist: Gen SC	**0.0339**	**0.0525**	**0.0354**	**0.0814**	**0.0820**
RE Dist: Fit SC	0.0060	0.0038	0.0043	**0.0117**	0.0035
Gen SC:Fit SC	**0.0151**	**0.0412**	**0.0351**	**0.0180**	**0.0712**
N:*p*:RE Dist	**0.0196**	0.0051	0.0047	**0.0218**	**0.0338**
N:*p*:Gen SC	**0.1475**	**0.0156**	**0.0601**	**0.0269**	**0.0338**
N:*p*:Fit SC	0.0010	0.0021	**0.0115**	0.0012	0.0005
N:RE Dist:Gen SC	**0.0397**	**0.0713**	**0.0523**	**0.0747**	**0.0380**
N:RE Dist:Fit SC	0.0070	0.0084	**0.0132**	**0.0188**	0.0084
N:Gen SC:Fit SC	**0.0128**	**0.0109**	**0.0103**	**0.0191**	**0.0111**
p:RE Dist:Gen SC	**0.1112**	**0.0989**	**0.0792**	**0.0969**	**0.0961**
p:RE Dist:Fit SC	0.0023	0.0038	**0.0152**	0.0099	**0.0107**
p:Gen SC:Fit SC	0.0067	**0.0193**	**0.0147**	**0.0254**	**0.0103**
RE Dist:Gen SC:Fit SC	**0.0309**	**0.0205**	**0.0254**	**0.0355**	**0.0228**

Note: Bold numbers are > 0.01, *N* – cluster sample size, *p* – within cluster sample size, Gen – generated, RE Dist – random effect distribution, SC – serial correlation, Fit – fitted, ":" represents an interaction

Sensitivity Analysis

An arcsine transformation was done on the empirical type I error rates that were analyzed above. The transformation was performed for two reasons, first to remove the hard 0 and 1 boundaries of the proportion metric, and second to remove the mean and variance relationship of the proportion metric. This transformation took the following form:

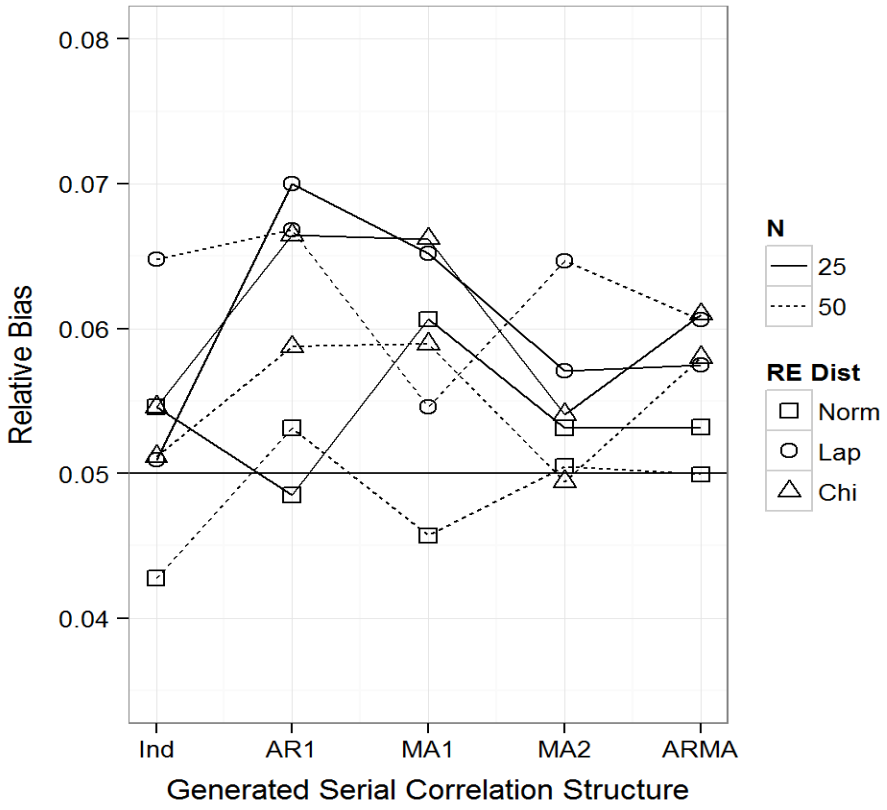

Figure 4. Mean type I error rate for β_0 by generated serial correlation structure, random effect distribution, and cluster sample size

$$\begin{aligned}
\hat{p}'_k &= 2\sqrt{\sin^{-1}\hat{p}_k}, 0 < \hat{p}_k < 1 \\
\hat{p}'_k &= 2\sqrt{\sin^{-1}\left(\frac{1}{4}R_k\right)}, \hat{p}_k = 0 \\
\hat{p}'_k &= 3.14 - 2\sqrt{\sin^{-1}\left(\frac{1}{4}R_k\right)}, \hat{p}_k = 1
\end{aligned} \tag{12}$$

where R refers to the number of simulation replications. After making the transformation, the transformed empirical type I error rates will be normally distributed with mean p'_k and variance $1/R_k$ (Marascuilo & McSweeney, 1977). After the transformation was performed, a similar model was fitted to the data as discussed above except now the average arcsine transformed empirical type I error

rate was used as the dependent variable. Just as before, η^2 served as the effect size to identify variables that explained significant variation in the dependent variable as opposed to *p*-values.

The effect sizes calculated from the arcsine transformed empirical type I error rates were similar to the model left in the original proportion metric with no additional variables identified as significant. Since the results were similar, interpretations made above in the original proportion metric are similar regardless of the scale of measurement which adds to the robustness of results.

Discussion

The current Monte Carlo study explored the implications for the LMM when model assumptions have not been adequately met. Five different generated serial correlation structures, independent, AR(1), MA(1), MA(2), and ARMA(1, 1) were explored in the current Monte Carlo study along with three different simulated random effect distributions, normal, chi-square (1), and Laplace.

Study results showed that the fixed effects on average were unbiased and none of the simulation conditions explained significant variation in the relative bias of the fixed effects for either of the studies. However, there was evidence of bias in the variance components and simulation conditions did explain significant variation in the average relative bias. This is similar to previous research when serial correlation was not modeled and the random components were normally distributed (Kwok et al., 2007; Murphy & Pituch, 2009).

Unfortunately, no real pattern to which fitted serial correlation is best emerged, for example overspecified or underspecified covariance structures did not consistently provide better estimates of the random components. Instead including some measure of serial correlation, when present, helps to alleviate some bias concern for the random effects. However, even correctly modeling the serial correlation structure tended to produce biased random components of the model. The AR(1) and ARMA(1, 1) tended to produce the smallest amounts of bias in the random components, however the convergence rate was impacted when these additional parameters were included in the model.

For both the fixed effects and random components, the simulated random effect distribution did not explain significant variation in the relative bias statistics. This is contrary to prior work exploring the robustness of the LMM to normality assumptions (Maas & Hox, 2004a; LeBeau, 2013). Results from this prior work found that the simulated random effect distribution did not produce bias in the fixed effects, but did introduce bias into the random effects. However, these studies did

not build explanatory models to see which study conditions explain variation in the relative bias statistics. Adding the more complicated serial correlation structures may have influenced this relationship and overpowered the influence of the non-normal random effect distribution.

This Monte Carlo study also explored the type I error rates of the five fixed effects. The fixed effects were all slightly elevated compared to the $\alpha = 0.05$ level. Increasing the sample size at both levels of the model was the best way to help limit the slight inflation found in the empirical type I error rates. Trends regarding the generated or fitted serial correlation structure and the simulated random effect distribution were not as clear.

Recommendations

Recommendations for researchers come in three different groups. First, if the researcher is only interested in the estimates of the fixed effects, then one does not need to worry about the serial correlation. The results showed that the relative bias for the fixed effects were not affected by any of the simulation conditions studied, including the generated or fitted serial correlation structures, random effect distribution, sample size considerations, or missing a random effect. These results are similar to other Monte Carlo studies with the linear mixed model (Ferron et al., 2002; Kasim & Raudenbush, 1998; Kwok et al., 2007; Maas & Hox, 2004a; Murphy & Pituch, 2009).

However, if the researcher is interested in estimates of the random effects, more care needs to be taken. In general, the random effects tend to be overestimated when serial correlation is present and ignored (i.e. an independence structure is assumed to underlie the data when this is not the case). Although still overestimated, more measurement occasions (i.e. within cluster sample size) and fitting an AR(1) or ARMA(1, 1) serial correlation structure tends to limit the overestimation of the random effects.

Lastly, if the researcher is interested in inference about the fixed effects care needs to be taken to explore whether serial correlation is present in the data. This is especially important when the number of individuals (clusters) and the number of repeated measurements are small. Although not severely inflated, it is likely that the α value specified by researchers is slightly larger in practice.

Unfortunately, there is no a priori test to directly test for the presence of serial correlation in the data. To look for serial correlation, a variogram could be used or descriptively looking at the average correlations between measurement occasions. Another tactic would be to use a procedure such as the likelihood ratio test or model

fit indices such as the AIC or SBC to see if modeling the serial correlation improves model fit. Unfortunately, these methods have not been very reliable in selecting the correct structure (Ferron et al., 2002; Keselman et al., 1998).

Future Work

Future work exploring reasons for the poor convergence rate of the models is needed. Increasing the variances of the random components to see if that aids the poor convergence rates would be helpful. Increasing the variance of the random components may also have an impact on the empirical type I error rates and would be useful to explore.

Detecting serial correlation when present in the data is another area of work that needs to be explored. Currently it is difficult to detect serial correlation from the data putting researchers in a difficult position when searching for serial correlation in their data. Procedures to use when looking for serial correlation in the data would provide guidance for researchers. Exploring additional missing data structures would also be useful. The current study used dropout as a missing data structure as this commonly occurs in longitudinal data, however it is not the only way missing data occurs. For example, having a subject to re-enter the study after missing a measurement occasion is also common in longitudinal data.

Finally, additional work that relaxes the assumption that random effects are uncorrelated across clusters, extending the work done by Browne and Goldstein (2010) in a Bayesian framework, could be a new extension of this group of models. This would give researchers the flexibility of modeling three levels of nesting through the use of a two level model. Situations where this would be most helpful would be when relatively few level three units are sampled, for example when only five schools are sampled. It would likely not be possible to model this third level of nesting with only five units, however accounting for this dependency through correlated random effects at level two may be useful and necessary if the third level of nesting accounts for a significant amount of variation.

References

Box, G. & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. Eaglewood Cliffs, NJ: Prentice Hall.

Browne, W. & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics, 15*(3), 391-420. doi: 10.1007/s001800000041

Browne, W. & Goldstein, H. (2010). MCMC sampling for multilevel model with nonindependent residuals within and between cluster units. *Journal of Educational and Behavioral Statistics, 35*(4), 453-473. doi: 10.3102/1076998609359788

Bryk, A. & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147-158. doi: 10.1037/0033-2909.101.1.147

Chi, E. & Reinsel, G. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association, 84*(406), 452-459. doi: 10.1080/01621459.1989.10478790

Diggle, P. (1988). An approach to the analysis of repeated measurements. *Biometrics, 44*(4), 959-971. doi: 10.2307/2531727

Diggle, P. (2002). *Analysis of longitudinal data*. New York, NY: Oxford University Press.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research, 37*(3), 379-403. doi: 10.1207/S15327906MBR3703_4

Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.

Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine, 13*(16), 1643-1655. doi: 10.1002/sim.4780131605

Harwell, M., Post, T., Cutler, A., Maeda, Y., Anderson, E., Norman, K., & Medhanie, A. (2009). The preparation of students from national science foundation-funded and commercially developed high school mathematics curricula for their first university mathematics course. *American Educational Research Journal, 46*(1), 203-231. doi: 10.3102/0002831208323368

Kasim, R., & Raudenbush, S. (1998). Application of gibbs sampling to nested variance components models with heterogeneous within-group variance.

Journal of Educational and Behavioral Statistics, 23(2), 93-116. doi: 10.3102/10769986023002093

Keselman, H., Algina, J., Kowalchunk, R., & Wolfinger, R. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics – Simulation and Computation, 27*(3), 591-604. doi: 10.1080/03610919808813497

Keselman, H., Algina, J., Kowalchuck, R., & Wolfinger, R. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology, 52*(1), 63-78. doi: 10.1348/000711099158964

Kwok, O., West, S., & Green, S. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. Multivariate Behavioral Research, 42(3), 557-592. doi: 10.1080/00273170701540537

LeBeau, B. (2013, April). *Impact of non-normal random components on the linear mixed model*. Poster session presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

L'Ecuyer, P., Simard, R., Chen, E., & Kelton, W. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research, 50*(6), 1073-1075. doi: 10.1287/opre.50.6.1073.358

Littell, R., Henry, P., & Ammerman, C. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science, 76*(4), 1216-1231. Retrieved from https://dl.sciencesocieties.org/publications/jas/abstracts/76/4/1216

Maas, C. & Hox, J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137. doi: 10.1046/j.0039-0402.2003.00252.x

Maas, C. & Hox, J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis, 46*(3), 427-440. doi: 10.1016/j.csda.2003.08.006

Maas, C. & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences, 1*(3), 86-92. doi: 10.1027/1614-2241.1.3.86

Marascuilo, L. & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole Pub. Co.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

Murphy, D. & Pituch, K. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education, 77*(3), 255-284. doi: 10.3200/JEXE.77.3.255-284

Núñez-Antón, V. & Zimmerman, D. (2000). Modeling nonstationary longitudinal data. *Biometrics, 56*(3), 699-705. doi: 10.1111/j.0006-341X.2000.00699.x

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2012). The nlme package: Linear and nonlinear mixed effects models (Versions 3.1-103) [Software]. Available from https://cran.r-project.org/web/packages/nlme/index.html

Post, T., Medhanie, A., Harwell, M., Norman, K., Dupuis, D., Muchlinski, T., … Monson, D. (2010). The impact of prior mathematics achievement on the relationship between high school mathematics curricula and postsecondary mathematics performance, course-taking, and persistence. *Journal for Research in Mathematics Education, 41*(3), 274-308. Available from http://www.jstor.org/stable/20720139

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics, 13*(2), 85-116. doi: 10.3102/10769986013002085

Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Snijders, T. & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics, 18*(3), 237-259. doi: 10.3102/10769986018003237

Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Wolfinger, R. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*(2), 205-230. Available from http://www.jstor.org/stable/1400366