Misspecification of the covariance matrix in the linear mixed model: A monte carlo simulation

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Brandon LeBeau

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctor of Philosophy

Michael Harwell

February, 2013

Acknowledgements

I would like to acknowledge the exemplary instruction, guidance, and mentorship of Dr. Michael Harwell. He helped to spur my appreciation, enthusiasm, and dedication to statistics and research. His supreme attention to detail, especially regarding this Doctoral dissertation, has made my work today infinitely better.

I would also like to acknowledge the inspirational support and guidance of Dr. Andrew Zieffler. Early on he pushed me to undertake more and push my comfort threshold. Without his persistance and knowledge I would not be as proficient in R, LaTeX, and computing in general as I am today.

Finally, I wish to thank Rachel for her support, encouragement, and love. This process would have been much more strenuous without her tolerance and enthusiasm.

Dedication

I would like to dedicate this Doctoral dissertation to my Grandma and Grandpa LeBeau and my Mom. They all strived to give me a leg up and worked perilously to push me to greater endeavors. Without their unwavering support I would not have attempted and achieved this degree.

Abstract

The linear mixed model has become a popular method for analyzing longitudinal and cross sectional data due to its ability to overcome many of the limitations found using classical methods such as repeated measures analysis of variance or multivariate analysis of variance. Although the linear mixed model allows for flexible modeling of clustered data, the simulation research literature is not nearly as extensive as classical methods. This current study looks to add to this literature and the statistical properties associated with the linear mixed model under longitudinal data conditions.

Historically when using the linear mixed model to analyze longitudinal data, researchers have allowed the random effects to solely account for the dependency due to repeated measurements. This dependency arises in this case, from repeated measurements on the same individual and measurements taken closer in time would be more correlated than measurements taken further apart in time. If measurements are taken close in time (i.e. every hour, daily, weekly, etc.), the random effects alone may not adequately account for the dependency due to repeated measurements. In this case serial correlation may be present and need to be modeled.

Previous simulation work exploring the effects of misspecification of serial correlation have shown that the fixed effects tend to be unbiased, however evidence of bias show up in the variance of the random components of the model. In addition, some evidence of bias was found in the standard errors of the fixed effects. These

simulation studies were done with all other model conditions being "perfect," including normally distributed random effects and larger sample size. The current simulation study looks to generalize to a wider variety of data conditions.

The current simulation study used a factorial design with four simulation conditions manipulated. These included: covariance structure, random effect distribution, number of subjects, and number of measurement occasions. Relative bias of the fixed and random components were explored descriptively and inferentially. In addition, the type I error rate was explored to examine any impact the simulation conditions had on the robustness of hypothesis testing. A second smaller study was also conducted that explicitly misspecified the random slope for time to see if serial correlation could overcome the misspecification of that random effect.

Results for the larger simulation study found no bias in the fixed effects. There was however evidence of bias in the random components of the model. The fitted and generated serial correlation structures as well as their interaction explained significant variation in the bias of the random components. The largest amounts of bias were found when the fitted structure was underspecified as independent. Type I error rates for the five fixed effects were just over 0.05, with many around 0.06. Many of the simulation conditions explained significant variation in the empirical type I error rates.

Study two again found no bias in the fixed effects. Just as in study one, bias was found in the random components and the fitted and generated serial correlation structures as well as the interaction between the two explaining significant variation in the relative bias statistics. Of most concern were the severely inflated type I error rates for the fixed effects associated with the slope terms. The average

type I error rate was on average twice what would be expected and ranged as high as 0.25. The fitted serial correlation structure and the interaction between the fitted and generated serial correlation structure explained significant variation in these terms. More specifically, when the serial correlation was underspecified as independent in conjunction with a missing random effect for time, the type I error rate can become severely inflated.

Serial correlation does not appear to bias the fixed effects, therefore if point estimates are all that are desired serial correlation does not need to be modeled. However, if estimates of the random components or inference are concerned care needs to be taken to at least include serial correlation in the model when it is found in the data. In addition, if serial correlation is present and the model is misspecified without the random effect for time serious distortions of the empirical type I error rate occur. This would lead to rejecting many more true null hypotheses which would make conclusions extremely uncertain.

Contents

List of Tables

List of Figures

Chapter 1

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) being 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. These situations commonly arise from a clustered sample where all of the students from a classroom are selected instead of randomly selecting students. For some additional examples of how these models are used in education see Bryk and Raudenbush (1987) and Raudenbush (1988).

The LMM can be a very general model that is able to accomodate many

different data conditions, including longitudinal studies where repeated measures are nested within individuals. The model is not limited to just two levels of clustering, but more levels of nesting can be added. For example, a three level longitudinal study with repeated measures nested within individuals nested within schools could be fit in a LMM framework. Note, the LMM is commonly depicted as a linear model with no explicit mention of levels. However, in order to avoid confusion (both for me and for the reader), the LMM in this paper will often be spoken about with respect to levels. The focus of this paper will be on longitudinal studies, although cross sectional models follow a similar framework.

The LMM has also been gaining popularity in other areas besides education including: sociology, medicine/biostatistics, statistics, and economics for many of the same reasons it has become popular in education, although the name for the model often changes. Some of these names include: two-stage random effects model, multilevel linear model, empirical Bayes' model, random regression coefficients, or contextual models (Blalock, 1984; Cnaan, Laird, & Slasor, 1997).

1.1 History

1.1.1 Classic approaches for longitudinal data

Univariate repeated measures analysis of variances (RM-ANOVA) has been a popular analysis option for analyzing repeated measures data (see Kutner, Nachtsheim, Neter, and Li (2005) and Tabachnick, Fidell, and Osterlind (2001) for more details on RM-ANOVA). RM-ANOVA has the advantage of being a rather simple approach, compared to other models for repeated measures data, bearing many

simularities to traditional analysis of variance (ANOVA). RM-ANOVA partitions the variance to within subject and between subject variation and allows multiple within and between subject variables.

The biggest drawback of this procedure is how it handles the dependency due to repeated measures, assuming sphericity for the test statistic to follow an F distribution (Kutner et al., 2005; Tabachnick et al., 2001). Sphericity assumes that the variance of the difference between two treatment means is a constant (i.e. $\sigma^2(\bar{Y}_{.j} - \bar{Y}_{.j'}) = c, j \neq j'$) where c is a constant. A related but more restrictive term, compound symmetry, assumes that the correlation between time points is equal regardless of the lag between measurements (Kutner et al., 2005). For repeated measures data, this assumption is rarely justifiable especially when time is the within subject variable. In these situtations it is commonly assumed that measurements closer in time are more related than measurements separated further in time. In addition to the sphericity assumption, the RM-ANOVA only allows for balanced designs (i.e. all subjects measured at the same time), discrete covariates (i.e. binary or categorical variables), and no missing data (Fitzmaurice, Laird, & Ware, 2004).

Due to the unrealistic sphericity assumption many researchers started to use the multivariate analysis of variance (MANOVA) or the related technique profile analysis (see Tabachnick et al. (2001) for more details on MANOVA and profile analysis). For these two procedures, the within subject variable is treated as multiple dependent variables (for example, one dependent variable for each measurement occasion). Instead of the sphericity assumption found with RM-ANOVA, MANOVA uses an unstructured covariance matrix (Núñez-Antón & Zimmerman,

2000). If there are many time points, this can lead to many covariances needed to be estimated, which would likely reduce the efficiency of estimation. Even though these approaches do not make the assumption of sphericity, only balanced longitudinal designs with complete data are allowed (Fitzmaurice et al., 2004). Due to many of these limitations, a new method has gained popularity to analyze longitudinal data, the LMM.

1.1.2 Modern approaches for longitudinal data

The linear mixed model was formulated in the early 1970's (see Lindley and Smith (1972)), but it wasn't until the late 1970's that the approach had an estimation algorithm to allow for the estimation of the model (Hartley & Rao, 1967; Dempster, Laird, Rubin, et al., 1977; Dempster, Rubin, & Tsutakawa, 1981; Harville, 1977). The model has to use an iterative procedure in order to estimate the variance terms and is a computationally difficult task. After methods for estimating these models were established, there was an influx of talk about these models during the early to mid 1980's throughout the literature for statistics, biostatistics, sociology, and education (Laird & Ware, 1982; Aitkin & Longford, 1986; Blalock, 1984; Goldstein, 1986; De Leeuw & Kreft, 1986; Longford, 1987).

Even though these models gained popularity in many areas throughout the 1980's, the increase in computing power over the past couple decades have made these models more computationally feasible. As a result, the LMM has been adopted by applied researchers for longitudinal or clustered data analysis. In response to the interest and use of the LMM, numerous simulation studies have

explored violations to model assumptions and its effect on estimation and inference, however Ferron, Hess, et al. (2004) commented that the LMM literature is not near as exhaustive as the multiple regression literature. Therefore, this study looks to add to the literature on possible consequences of model assumption violations.

1.2 Research Problem

In longitudinal studies, the repeated measures for the same person are expected to be more similar due to the fact that the same person is being measured multiple times on the same measurement scale and the repeated measures are being measured closely in time (Littell, Henry, & Ammerman, 1998). A unique aspect of the LMM is the ability to account for the dependency due to repeated measures in a few different ways. This dependency can be accounted for by the specification of 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, & Rasbash, 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.

Simulation studies are commonly used to look at assumption violations/model misspecification issues due to the considerable amount of control they offer the researcher compared to analytical studies. In a simulation study, researchers explicitly specify and alter the conditions of interest which results in a factorial type experimental design. In an analytic study, the researcher does not know the "true" parameter values therefore it is difficult to assess if model one is better than model two as the "true" model is never known. Simulation studies, however, allow the researcher to compare parameter estimates to the parameter set by the researcher in the study design. Therefore, differences found in the simulation study can only be a function of the manipulated conditions.

Chapter 2

Literature Review

2.1 The Model

2.1.1 Two-level models

Some basic notation needed to define the models. First, let i represent the level 1 units and let j represent the level 2 units. There are n_2 level 2 clusters and within each level 2 cluster there are n_{1j} units at level 1. As an example, suppose a researcher is interested in whether a math curriculum is effective in improving grades over time. The researchers collect data twice a year over a 4 year period on a standardized math achievement test. Here the level 1 units are the repeated measures and the level 2 units are the individuals (students). The same indexing can be used for cross sectional data where the level 1 units could be students and the level 2 units could be classrooms.

The similar feature in both of these situations is that the level 1 and level 2 units are correlated in some fashion. It is obvious that repeated measures taken

from the same individual will be related (someone scoring high will also tend to score high on subsequent measures). However, the same dependency is not as obvious for cross sectional data, but the same argument can be applied. One would expect students coming from the same classroom to score more similar than students from different classrooms. Classrooms characteristics such as: same teacher, same classmates, same textbook, etc. account for the reasons why you would expect students within the same classroom to score more similar than students from across classrooms.

This dependency is commonly accounted for by the random effects at level 2. These are additional residual terms that reside at level 2. The inclusion of these random effects in the model have the effect of imposing structure on the covariance of the repeated measures (Fitzmaurice et al., 2004). When the covariance between the repeated measures has been appropriately modeled, correct standard errors are obtained which leads to valid inferences regarding the regression parameters (Fitzmaurice et al., 2004).

A basic linear mixed model can be written as follows:

$$Y_{ij} = X_{ij}\beta + Z_{ij}b_j + e_{ij} \tag{2.1}$$

In this model, the Y_{ij} is the response variable for the i^{th} level 1 unit nested within the j^{th} level 2 unit. For example, the Y_{ij} may be the test score for the i^{th} student in the j^{th} classroom. Next is the X_{ij} which is a n_i by 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. Examples of covariates from the school example above could be gender at level 1 or classroom socio-economic status (SES) by aggregating student SES over all the students in each classroom. The β in the model is a p by 1 vector representing the fixed effects. Next is the Z_{ij} which is the design matrix for the level 2 random effects. This term is commonly formed from a subset of the columns of X_{ij}. The 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 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 multilevel form as follows:
Level 1:

$$Y_{ij} = Z_{ij}\beta_j + e_{ij} \tag{2.2}$$

Level 2:

$$\beta_j = A_j\beta + b_j \tag{2.3}$$

Within this formulation it is easier to see the structure of the linear mixed model. Here in the level 1 equation, the Y_{ij} and e_{ij} are identical to the above model. However, in this formulation the Z_{ij} are the level 1 covariates (no level 2 covariates are included) and the β_j now have a J subscript which allow them to vary at level 2. At level 2, the β_j are now the response variable(s). The A_j are

the level 2 covariates (no level 1 covariates here). The rest is the same as before. The β are the fixed effects and the b_j are the level 2 residuals or random effects. When you combine the equations, the $Z_{ij}A_j$ combine to form the X_{ij} from above.

Fitzmaurice et al. (2004) point out that there is an advantage to specifying the model in hierarchical form as it is easier to see which covariates are acting at which level. However, they also show a disadvantage in that when you combine the hierarchical models into a single equation, the formulation leads to an overly rigid structure where all of the level 1 covariates must have the accompanying random effects as well. Therefore, in the rest of the paper, the linear mixed model will be in the form of equation (2.1) however to aid in discussion the levels of the model will still be mentioned.

The same model can also be expressed in matrix form:

$$\mathbf{Y}_j = \mathbf{X}_j\beta + \mathbf{Z}_j\mathbf{b}_j + \mathbf{e}_j \tag{2.4}$$

Here the terms are identical as to equation (2.1) above just the i subscript is dropped.

2.1.2 Three-Level Models

The extension to a three-level model is natural from the two-level model shown in equation (2.1). To index the three level structure, i represents the level 1 units, j represents the level 2 units, and now k represents the level 3 units. Each of the k clusters has n_{2k} level 2 clusters in it, and each level 2 cluster has n_{1jk} level 1 units within it. Extending our example from above, suppose we again have repeated measures nested within individuals (students), an obvious third level in

education is the school level. Differences at the school level may help to predict the differences in math achievement over time. The nested structure would then be: repeated measures nested within individuals (students) nested within schools. A similar structure would follow when applying a three level model to cross sectional data where the students are at level 1, classrooms would be at level 2, a third level 3 could be schools, neighborhoods, cities, states, etc.

The three-level model is as follows:

$$Y_{ijk} = X_{ijk}\beta + Z^{(3)}_{ijk}b^{(3)}_{k} + Z^{(2)}_{ijk}b^{(2)}_{jk} + e_{ijk} \tag{2.5}$$

Many of the terms are the same from the two-level model, except now there is the possibility of a third level. Y_{ijk} is still the response variable. The X_{ijk} term represents the covariates in the model, also known as the design matrix. The only difference in this term is that now there can be covariates at level 1, leve1 2, and level 3 as well as variables that are aggregated across the level 1 or level 2 units. The β are again the fixed effects associated with the covariates. The main difference in the model is the addition of another Z matrix representing the design matrix of the random effects. For example, $Z^{(3)}_{ijk}$ is the design matrix for the level 3 random effects and $Z^{(2)}_{ijk}$ is the design matrix for the level 2 random effects. Both of these terms are composed of the relevant columns of X_{ijk}. The $b^{(2)}_{jk}$ and $b^{(3)}_{k}$ are the level 2 and level 3 random effects respectively. Finally, the e_{ijk} are the level 1 variance components just like in the two-level model. Extending the model beyond three levels is straightforward and is not explored further.

2.1.3 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 b_j are independent across level 2 units, normally distributed (multivariate normal when more than one random effect is in the model), each has a mean of 0, and a covariance matrix $\mathbf{G}$. A common shorthand is to say that the b_j are independent and identically distributed (*iid*) which means that each b_j are independent from one another and all come from the same probability distribution (i.e. normal distribution). This can be succintly written as: $b_j \sim iid\ N(0, \mathbf{G})$. Similarly, for a three-level model the $b_{jk}^{(2)}$ and $b_k^{(3)}$ are: $b_{jk}^{(2)} \sim iid\ N(0, \mathbf{G}^{(2)})$ and $b_k^{(3)} \sim iid\ N(0, \mathbf{G}^{(3)})$ for level two and level three respectively.

2. Every e_{ij} and e_{ijk} are independent and follows a normal distribution with mean 0 and variance σ^2 for every level 1 unit within level two or level three respectively. This can be summed up as: $e_{ij} \sim iid\ N(0, \sigma^2)$ and $e_{ijk} \sim iid\ N(0, \sigma^2)$ for level two and level three models respectively.

3. The e_{ij} and e_{ijk} are independent of the random effects at level 2 or level 3 where applicable.

4. For three level models, the $b_{jk}^{(2)}$ and $b_k^{(3)}$ are usually treated to be independent (i.e. $cor(b_{jk}^{(2)}, b_k^{(3)}) = 0$).

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. There are models using a similar framework as the LMM for response variables that are discrete (e.g. dichotomous or count) called generalized linear mixed models (GLMM). The interested reader is directed to Diggle (2002), Fitzmaurice et al. (2004), and Molenberghs and Verbeke (2005) for more information on the GLMM.

Violations to these assumptions typically impact estimates of the standard errors (and in turn inferences) more so than the fixed effect estimates. Estimates of the variance components are also affected when there are violations to model assumptions. This will be explored more in depth in the coming sections.

2.1.4 Missing Data

In longitudinal studies, missing data is more of a rule than the exception (Fitzmaurice et al., 2004). In general, it is more difficult to follow up with participants as the number of measurement occasions increases or the time frame between observations is very large, as a result, there tends to be more missing data at the later time points. This has the consequence of making the design unbalanced. For example, some participants may have four observations where as others may have six. Traditionally this was a problem with classic methods such as RM-ANOVA and MANOVA, but unbalanced data is not a problem for the LMM as the number of units at level 1 does not have to be the same for every level 2 unit.

Another interesting aspect of longitudinal studies is that you can have participants that miss an observation, but then come back for subsequent observations (Laird & Ware, 1982). This further complicates the missing data structure for longitudinal studies. With many possible reasons for missing data in longitudinal studies, it needs to be addressed and thought about by researchers before data analysis.

Another implication of missing data is in the validity of the inferences. Rubin (1976) devised what is known as missing data mechanisms or missingness mechanisms that treat missing data from a probability standpoint. Under this modern missing data framework, missing data is treated as a set of random variables, say R. Adopting the general notation from Schafer and Graham (2002), let Y_{com} represent the complete data which can be broken down as follows: $Y_{com} = (Y_{obs}, Y_{miss})$. Y_{obs} is the observed data on Y and Y_{miss} is the missing data on Y. Data is said to be missing at random (MAR) when the following is true:

$$Pr(R|Y_{com}) = Pr(R|Y_{obs}) \tag{2.6}$$

Equation (2.6) states that the missing data depends (conditional) on the observed data, not on the missing data. In other words, probability of missing can be a function of another observed variable, but not as a function of the missing data itself (Schafer & Graham, 2002).

A special case of MAR is missing completely at random (MCAR). When the

missing data is MCAR, the missing data does not depend on observed and unobserved parts of Y. Stating this in probability terms:

$$Pr(R|Y_{com}) = Pr(R) \tag{2.7}$$

This is similar to taking a random number generator to the data to exclude cases, there are no patterns in the missing data structure, it is completely random.

If the data does not follow equation (2.6) or (2.7), the data is said to be not missing at random (NMAR). Under this missing data mechanism, the data is missing as a function of Y. NMAR is a very severe missing data mechanism and is nonignorable (Rubin, 1976).

The LMM assumes that the missing data is MCAR or MAR for valid inferences. When the data is MCAR the missingness is not a function of any observed or unobserved aspect of the response variable; which means that a completely random process accounts for the missing data. This has the effect that inferences are unbiased without any further action needed. Under MAR, valid inferences can be obtained from the LMM, however the covariance matrix must have been properly modeled. This means correctly specifying the model and including relevant variables in the model (Fitzmaurice et al., 2004). When the missing data is NMAR, the researcher is put in a tough spot as nearly all longitudinal analysis methods are not valid. The missing data is truly nonignorable meaning that from the data at hand there is little the researcher can do. The researcher must use outside data sources to track down the missingness and figure out why it is missing (Fitzmaurice et al., 2004). Verbeke and Molenberghs (2000) offers a good discussion of missing data for longitudinal models including options for NMAR.

2.2 Estimation

Using the repeated measures example from above, the observations cannot be treated as independent. Also, a consequence of this dependency among the response variable is that the variance matrix is used in the estimation of the fixed effects by using generalized least squares. The problem is in estimating the variance components (which are not known) of the LMM which proved to be computationally difficult until the late 1960's to early 1970's (Jiang, 2007). Hartley and Rao (1967) were the first to establish maximum likelihood for the LMM, in addition Harville (1977) examined maximum likelihood and restricted maximum likelihood for the LMM. Both types of maximum likelihood will be discussed below.

2.2.1 Full Maximum Likelihood

The problem is finding the correct β that maximizes the log likelihood function. As discussed above, since the observations are correlated, the variance matrix needs to be considered in order to estimate β accurately. The variance matrix of the repeated measures is the difficult part to estimate and involves using an itervative procedure such as maximum likelihood (ML).

The response variable is assumed to be multivariate normal and is completely specified by the mean vector and variance matrix. The mean response vector for the model is

$$E(Y_j) = X_j\beta \tag{2.8}$$

and the variance matrix

$$Var(Y_j) = \Sigma_j = \Sigma_j(\theta) = Z_j G Z_j^{\mathsf{T}} + \sigma_e^2 I_{n1j} \tag{2.9}$$

where θ represents the variance parameters. The variance of the response variable can be partioned to the effects due to the random effects, $\mathbf{Z}_j\mathbf{G}\mathbf{Z}_j^{\mathsf{T}}$ and the level 1 error $\sigma_e^2\mathbf{I}_{n1j}$. Here, $\mathbf{Z}_j$ is the design matrix for the random effects, $\mathbf{G}$ is the variance components of the random effects, σ_e^2 is the level 1 variance, and $\mathbf{I}_{n1j}$ is an identity matrix of size n_{1j}.

The probability density function (pdf) for a multivariate normal distribution is as follows:

$$f(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp - \frac{1}{2}(y - \mathbf{X}\beta)^{\mathsf{T}}\Sigma^{-1}(y - \mathbf{X}\beta). \tag{2.10}$$

Following from the pdf for the multivariate normal distribution, the log likelihood function that needs to be maximized is the following:

$$l(\beta, \theta) = c - \frac{1}{2}\sum_{j=1}^{N} log|\Sigma_j^{-1}| - \frac{1}{2}\sum_{j=1}^{N}(y_j - X_j\beta)^{\mathsf{T}}\Sigma_j^{-1}(y_j - X_j\beta) \tag{2.11}$$

Where c is a constant and Σ_i is the variance matrix of the repeated measures defined in equation (2.9). Generally, there are no analytic techiniques to maximize this equation so iterative techniques must be used (Molenberghs & Verbeke, 2005). Examining the likelihood of equation (2.11) shows that the third term is negative, so maximizing the β (fixed effects) is equivalent to minimizing the last term. As a result, the estimator of β is generally known as the generalized least squares

estimator (sometimes called the maximum likelihood estimator) of β (Molenberghs & Verbeke, 2005):

$$\hat{\beta} = \sum_{j=1}^{N} (X_j^{\mathsf{T}} \Sigma_j^{-1} X_j)^{-1} \sum_{j=1}^{N} (X_j^{\mathsf{T}} \Sigma_j^{-1} X_j) \tag{2.12}$$

The only problem with equation (2.12) for the $\hat{\beta}$ is the fact that the Σ_j matrix is unknown. Therefore, this has to be estimated with the data as well. Here the log likelihood shown in equation (2.11) needs to be maximized with respect to theta. This is done by taking the derivative of equation (2.11) with respect to theta, set it to zero, and find the solution (Fitzmaurice et al., 2004; Jiang, 2007). Equation (2.11) is generally non-linear with no easy closed form expression (Fitzmaurice et al., 2004), therefore a computer algorithm iterates to find a solution and is said to "converge" when little change to the parameter estimates between subsequent iterations. Once this has been done, the estimate of the variance components that make up the Σ_j matrix, now $\hat{\Sigma}_j$ or $\Sigma_j(\hat{\theta})$ are used in the equation (2.12) and the fixed effect estimates are obtained.

The estimation process can be summed up into four steps:

1. Starting values for the fixed effects are found which are commonly found by ordinary least squares or are specified by the researcher.

2. The starting values for the fixed effects are plugged into equation (2.11) and maximized with respect to theta to obtain an estimate for the Σ_j matrix.

3. The $\hat{\Sigma}_j$ is plugged into equation (2.12) to obtain estimates for the β.

4. Steps 2 and 3 are repeated until the algorithm converges, which is obtained

when there is little change in parameter estimates between two subsequent iterations.

Some properties of the generalized least squares estimator for β are (Molenberghs & Verbeke, 2005; Fitzmaurice et al., 2004):

- $\hat{\beta}$ is normally distributed.
- The mean is $E(\hat{\beta} = \beta$
- The variance matrix is $Var(\hat{\beta}) = \sum_{j=1}^{N} (X_j^{\mathsf{T}} \Sigma_j^{-1} X_j)^{-1}$.

ML is an asymptotic method, which means that in large samples these properties for $\hat{\beta}$ hold.

2.2.2 Restricted Maximum Likelihood

The maximum likelihood estimates (MLE) of the variance components have been shown to be biased for small samples (Fitzmaurice et al., 2004; Jiang, 2007). Additionally, the MLE may not improve even as sample size increases when "the number fixed effects is proportional to the sample size" (Jiang, 2007). When the variance components are estimated, the fixed effects act as nuisance paramters. Restricted (or residual) maximum likelihood (REML) looks to fix these problems by applying a transformation to the data that eliminates the fixed effects from the log liklihood and use the transformed data to estimate the variance components.

When REML is used, a slightly different log likelihood is maximized

$$l_R(\theta) = c - \frac{1}{2}\sum_{j=1}^{N} log|\Sigma_j^{-1}| - \frac{1}{2}\sum_{j=1}^{N}(y_j - X_j\beta)^{\mathsf{T}}\Sigma_j^{-1}(y_j - X_j\beta) - \frac{1}{2}log|\sum_{j=1}^{N} X_j^{\mathsf{T}}\Sigma_j^{-1}X_j|. \tag{2.13}$$

Here the R subscript denotes that this log likelihood is the restricted version. The additional term in the log likelihood can be shown to be the covariance of $\hat{\beta}$. Therefore, the REML estimate makes a correction that is similar to the correction to the denominator of $\hat{\sigma^2}$ (Fitzmaurice et al., 2004).

Once the REML estimate for the variance components (i.e. $\Sigma_j(\hat{\theta})$) are obtained, this estimate is placed into equation (2.12) to obtain the fixed effects. A similar iteration procedure outlined for the maximum likelihood estimator is used for the REML estimator except the new log likelihood is used (equation (2.13)). In general, the REML estimate for Σ_j will be less biased than the ML counterpart (Fitzmaurice et al., 2004). Under reasonable conditions, the REML estimator is consistent and asymptotically normal just like the MLE (Jiang, 2007).

2.3 Inference

Inferences can be made on both the fixed effects and the random effects in the LMM. Based on the properties of the generalized least squares estimator for $\hat{\beta}$ given above, this can be used to construct confidence intervals for hypothesis tests. To calculate a 95% confidence interval about some fixed effect, β_t, one would take that parameter estimate, $\hat{\beta}_t$, plus and minus 1.96 times the standard

error of $\hat{\beta}_t$ (Fitzmaurice et al., 2004). If different confidence limits were desired, say 99%, only the multiple of the standard error would need to be changed (for example, 2.58 instead of 1.96).

Similarly, hypothesis testing can be done with the fixed effects. Suppose the following null hypothesis was established: $H_0 : \beta_t = 0$ versus the alternative $H_A : \beta_t \neq 0$ This test statistic can be calculated from a Wald statistic:

$$Z = \frac{\hat{\beta}_t}{\sqrt{\hat{Var}(\hat{\beta}_t)}}, \tag{2.14}$$

which follows the general formula as test statistic, $\hat{\beta}_t$, over a standard error, $\sqrt{\hat{Var}(\hat{\beta}_t)}$. The test statistic is compared to a standard normal distribution (Fitzmaurice et al., 2004).

An alternative to the Wald test for fixed effects is the likelihood ratio test. The likelihood ratio test compares the log likelihood of two models where one is a subset of another. As an example, suppose a researcher was interested in testing the following null hypothesis: $H_0 : \beta_2 = 0$. Using the likelihood ratio approach, the researcher would fit a "full" model with β_2 in the model and another "reduced" model that does not contain the term of interest, in this case β_2. A statistic can be calculated from twice the difference in the log likelihoods:

$$G^2 = 2(\hat{l}_{full} - \hat{l}_{red}), \tag{2.15}$$

where $\hat{l}_{full}$ corresponds to the log likelihood from the full model and $\hat{l}_{red}$ is the log likelihood from the reduced model. This statistic is compared to a chi-square distribution with the difference in the number of parameters between the full and

reduced model as the degrees of freedom (Fitzmaurice et al., 2004; Raudenbush & Bryk, 2002). In general this test has more desirable properties than the Wald tests and is recommended by Fitzmaurice et al. (2004).

One caveat with the likelihood ratio test occurs when testing fixed effects, where REML cannot be used when estimating the parameters. Instead FML must be used because when REML is used for estimation, an additional term was added to the likelihood being maximized (Fitzmaurice et al., 2004; Raudenbush & Bryk, 2002). This additional term in the likelihood function when REML is used will be different depending on which model is fit and nested models result in two different transformations of the data to get rid of the fixed effects from the likelihood function (Fitzmaurice et al., 2004).

In addition to the fixed effects, the LMM also estimates variance components of the random effects. These variance components can be of interest and inferential tests can be created about them. Typically these tests will be constructed by way of the likelihood ratio test where now the terms that are dropped are the random effects (or the covariance between two random effects). For example, a researcher could test the following hypothesis: $H_0 : G_{22} = G_{12} = 0$. Here the terms G_{22} and G_{12} represent the variance of the random effects for the slope and the covariance between the random effects for the slope and intercept respectively. This hypothesis could be conducted by fitting a full model with the random effect for slope in the model (as well as the covariance between random slopes and random intercepts) and a reduced model with the random effect for the slope removed from the model. Then one would use the same approach as above for the fixed effects (equation (2.15)).

As a note, the likelihood ratio test for the variance components can be computed when REML is used as long as the fixed effects remain the same throughout the estimation process. When the fixed effects remain constant the same transformed responses are used in the likelihood function during the estimation process (Fitzmaurice et al., 2004; Pinheiro & Bates, 2000).

2.3.1 Degrees of Freedom

Notice that up until now the degrees of freedom for the test statistics have not been discussed (except for the likelihood ratio test). This is because there is no one agreed upon solution to the problem of degrees of freedom. When the sample is large, this tends to not be a big issue as the associated degrees of freedom would be large and therefore the associated t and F distributions would be closely approximated by the standard normal and chi-square distributions respectively. But when the sample is smaller, the p-values and confidence intervals may be too small as the standard normal and chi-square distributions would have smaller tails than the t and F respectively (Fitzmaurice et al., 2004). Under small sample conditions, approximations for the degrees of freedom must be used. There are many methods possible, but the method developed by Satterthwaite (1941) is well known. Another method by Kenward and Roger (1997) is also another possible method that is similar to Satterthwaite's approach when the number of terms being tested is one.

As a result of the approximations in the degrees of freedom, the likelihood ratio approach considered above is recommended. When Wald tests or confidence intervals are constructed and the sample is "large" it is generally safe to use

standard normal and chi-square reference distributions. However, if the sample is small, the approximations mentioned above, or others, would be useful to look into.

2.4 Violation of Model Assumptions

2.4.1 Normality of Residuals

The normality of the residuals is a major assumption of the LMM and is needed for unbiased estimates and for correct inferences (Ferron, Hess, et al., 2004). The assumption of normality of the residuals is needed at both levels, but more attention in the literature has been paid to the residuals at the second level (the random effects).

Ferron, Hess, et al. (2004) commented that normality of the residuals at level one have implications for valid inferences at both levels of the model, although no studies were cited. The lack of research on violations of the level one residuals could be due to the strong literature on multiple regression showing estimation and inference to be largely robust to departures of normality of the level one residuals (Boneau, 1960; Gayen, 1949) applies to LMMs. A recent study has looked into violations of the level one residuals with the LMM, and did not find any evidence that the level one distribution significantly influenced paramter estimation (LeBeau, 2012).

More research has looked at violations of normality on the second level residuals (i.e., the random effects). However, many of the simulation studies using the LMM do not simulate data conditions representative of typical longitudinal data

(i.e., few observations per cluster). Two studies have included in their simulation a cluster size of five (Maas & Hox, 2004a, 2004b) and another study included cluster sizes of four and eight (LeBeau, 2012) which would be more representative of longitudinal data.

Simulation studies with longitudinal data have found little evidence of parameter bias in the fixed or random effects but 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. When the random effect distribution was uniform, there was no evidence of biased standard errors for the variance of the random effects (Maas & Hox, 2004b). There was no evidence of bias in the variance of the level one residuals.

Contrary to the results from the previous two studies, LeBeau (2012) found parameter bias in the fixed effects when the random effect distribution was chi-square with one degree of freedom, although this may have been due to the small amount of variability in the one simulated situation. The parameters most effected were the within-individual intercept and slope. There was also evidence of severe parameter bias (overestimation) for the variance of the random effects when the random effect distribution was not normally distributed (LeBeau, 2012).

Maas and Hox (2004a) and Maas and Hox (2004b) also studied whether the biased standard errors could be corrected by using robust standard errors (also called sandwich estimators; see (Huber, 1967; White, 1982)). For a cluster size of five and non-normal level two residuals, the robust standard errors improve

but do not completely overcome the underestimation of the standard errors for the level two variance components (Maas & Hox, 2004a, 2004b). However, this comes at the cost of overcorrecting the standard errors for the level one variance components. There was little effect on the standard errors for the fixed effects, which did not provide any evidence of bias.

2.4.2 Model Misspecification and Heterogeneity of Variance

The level 1 and level 2 residuals are usually assumed to have a constant variance. This allows one variance to be estimated instead of many which helps keep the model simple and allows the researcher to get by with less data if needed (i.e. less parameters to estimate). In addition, Ferron, Hess, et al. (2004) state that in order to get the correct standard errors the variance structure must be accurately specified. However, this assumption has not been extensively studied with the LMM. One such study by Darandari (2003) explored conditions when the assumption of heterogeneity of variance at level 2 was violated. He found that the estimates are robust to moderate and severe violations to the assumption of homogeneity of variance. The standard errors were smaller than expected when the sample size was very small at the group level.

Kasim and Raudenbush (1998) studied the violation to the assumption of the homogeneity of variance at level 1. They found no bias for the fixed effects and standard errors when the variances for each group were not homogenous. However, Raudenbush and Bryk (2002) say that if the level 1 variances vary as a function of a covariate that there may be more serious consequences (i.e. estimation bias,

incorrect inferences). If the variances are not the same for each group, this can be modeled by way of a heterogeneous variance matrix structure (Wolfinger, 1996). The LMM framework allows the researcher flexbility in modeling the variance structure which helps to improve the chances of correctly specifying the model.

Model misspecification is generally thought of as when a covariate is not included in the model when it should be (i.e. accounts for variation in the response variable). Darandari (2003) found that when a covariate that has a moderate effect size was omitted from the model resulted in positively biased variance estimates. Maeda (2007) also studied the effects of model misspecification when a level 1 covariate was omitted from the model when it should be in the model. Major biases were found in the parameter estimates and the inferences made from these parameters. These results reiterate the fact that model misspecification is an important consideration for the LMM (and other statistical techniques) and careful thought needs to be made to correctly specify the model, including covariates and variance structures.

2.4.3 Sample Size for Estimation

Sample size considerations for the LMM is an important consideration when planning a study. This is especially true since maximum likelihood are asymptotic tests and require large sample sizes for proper estimation (Maas & Hox, 2004a). Due to these reasons, there has been work done to see how many subjects one needs for accurate estimation. Typically, the highest level sample size is of most concern as there are fewer numbers at this level (Maas & Hox, 2004a). For example, in a cross sectional study, there may be an average of 30 students per classroom and

50 classrooms are sampled. Here, the higher level sample size (classroom level) is the smaller of the two as there are about 30 students per classroom, so the total level 1 sample size would be, $30 * 50 = 1500$. Therefore, there are 1500 students in the sample which is a large sample, but the question remains, is the 50 groups enough to accurately estimate the between classroom effects?

This issue is commonly exacerbated for longitudinal studies as the level 1 sample size tends to be small (i.e. few observations per subject); where 10 observations per subject is considered large (Snijders & Bosker, 1993). Therefore, after attrition is considered in the study, there may be problems with the level 1 sample size. 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, 1999, 2004a, 2004b; Browne & Draper, 2000). Additionally, the standard errors for the fixed effects are generally estimated accurately with at least 30 groups (Maas & Hox, 1999, 2004a).

However, the standard errors for the variance terms are not estimated as accurately under certain conditions. Maas and Hox (1999) found that the standard errors for the second level variance were too small when the level 2 sample size was less than 100. Another study done by Maas and Hox (2004a) found that a level 2 sample size of at least 50 was needed to estimate the standard errors of the second level variances. If the second level sample size is even smaller, then Maas and Hox (2004a) found that the standard errors for the second level variances are

estimated to be about 15% too small.

Based on these findings, Maas and Hox (2004a) came up with the following very general guidelines when planning the number of groups needed for a study. They state that if the researcher is only interested in the estimation of the fixed effects, only 10 groups are needed, if the researcher is interested in contextual effects, 30 groups are needed, but if the researcher wants correct standard errors for the estimates, at least 50 groups are needed. This should be taken very generally, as there may be other problems that cause problems with the estimation of the standard errors as discussed above (non-normality of the error variances, heterogeneity of variance, model misspecification, etc.). These additional problems would probably make the estimation of the standard errors more volatile and biased where additional groups would be needed or other actions would need to be taken to get accurate standard errors (e.g. transformations).

Finally, Afshartous and de Leeuw (2005) found that for prediction purposes the level 1 sample size was more important than the level 2 sample size. More level 1 units helped to increase the precision of the fixed effects aiding in accurate predictions. Additionally, Browne and Draper (2000) simulated unbalanced designs and found that the in general the imbalance of the design has less of an effect on estimation than the number of level 2 groups.

There have been some attempts by researchers to derive formulas to conduct a priori sample size calculations. These calculations are useful for reasearchers in the planning of studies so that they have enough subjects to estimate the parameters with precision to detect the desired effect. The interested reader is directed to Snijders and Bosker (1993) and Spybrook et al. (2011) for more information on

sample size formulas for cross sectional data designs and to Hedeker, Gibbons, and Waternaux (1999) for sample size formulas for longitudinal studies with attrition. Additionally, Hedeker et al. (1999) has formulas for different variance structures for the level 1 residuals, (e.g. first order autoregressive).

2.5 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

$$Var(Y_j) = \Sigma_j = \Sigma_j(\theta) = Z_j G Z_j^{\mathsf{T}} + \sigma_e^2 I_{n1j}, \tag{2.16}$$

which is also shown in equation (2.9) above. As can be seen from the above equation, the variance is composed of two portions, $Z_j G Z_j^{\mathsf{T}}$ is the portion of the variance that is accounted for by the random effects and the $\sigma_e^2 I_{n1j}$ 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 has 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:

$$\begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{pmatrix} \tag{2.17}$$

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, time^2, 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):

$$\begin{pmatrix} \sigma_e^2 + g_{11} & g_{11} & g_{11} & g_{11} \\ g_{11} & \sigma_e^2 + g_{11} & g_{11} & g_{11} \\ g_{11} & g_{11} & \sigma_e^2 + g_{11} & g_{11} \\ g_{11} & g_{11} & g_{11} & \sigma_e^2 + g_{11} \end{pmatrix} \tag{2.18}$$

Here σ_e^2 represents the error variance and g_{11} represents the variance of the random intercepts. As can be seen from matrix (2.18) 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.

One such example would include assuming the level one structure follows a first order autoregressive structure with a random intercept. This would yield the following covariance structure:

$$\begin{pmatrix} \sigma_e^2 + g_{11} & \rho + g_{11} & \rho^2 + g_{11} & \rho^3 + g_{11} \\ \rho + g_{11} & \sigma_e^2 + g_{11} & \rho + g_{11} & \rho^2 + g_{11} \\ \rho^2 + g_{11} & \rho + g_{11} & \sigma_e^2 + g_{11} & \rho + g_{11} \\ \rho^3 + g_{11} & \rho^2 + g_{11} & \rho + g_{11} & \sigma_e^2 + g_{11} \end{pmatrix} \tag{2.19}$$

Here the σ_e^2 and g_{11} terms are the same as above, but the new term, ρ, represents the average autocorrelation. Since ρ takes on a value between zero and one, raising this term to a power of two, three, or higher makes ρ smaller. This ultimately has the effect of decreasing correlation as the length of time between observations increases a common trait for longitudinal data.

With the inclusion of more complicated error terms, it can be helpful to include additional notation for the level one residual to separate measurement error and serial correlation denoted as $e_j = e_{(1)j} + e_{(2)j}$. Here $e_{(1)j}$ represents measurement 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 measurement error separately in the variance of the response

variable leads to the following expression:

$$Var(Y_j) = \Sigma_j = \Sigma_j(\theta) = Z_j G Z_j^{\mathsf{T}} + \sigma_e^2 I_{n1j} + \tau^2 H_j \tag{2.20}$$

Different from Equation (2.9) above, serial correlation is explicitly shown as $\tau^2 H_j$ where H_j is an n_j x n_j matrix where the (j,k)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 regardless of the type of data the model is being used for (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. The question must be asked, after removing the variation due to the random effects are the level one residuals independent from one another within an individual (Browne & Goldstein, 2010)? In other words, conditional on the level two random effects, is it safe to assume that the level one 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.

2.5.1 Time Series Models

Time series models are used when data is measured over time on the same phenomenon often with the desire to forecast (predict) ahead in time. As an example, suppose the amount of office paper in a school or company was recorded daily. A time series model could be fit to this data to attempt to forecast when the school/company would need to order more office paper. As one can imagine, these models are used frequently in economics where unemployment, economic growth, etc. are able to be modeled.

The defining feature of time series data (just like longitudinal data) is the correlated nature of the data (i.e. each observation is not independent from one another). Using the office paper example from above, it would be expected that the amount of office paper recorded on a day, would be very strongly related to how much office paper was recorded on the previous day. This dependency needs to be modeled in order to ensure appropriate model fit and for accurate forcasts to be made (Bowerman, O'Connell, & Koehler, 2005; Box & Jenkins, 1976; Shumway & Stoffer, 2000). There are a few different ways to account for this dependency in the data with time series models. Three ways will be discussed in more detail below.

The work by Box and Jenkins (1976) defined a set of models that can be used for time series data that are stationary. A stationary time series is one where the statistical properties are constant throughout the entire time series (Bowerman et al., 2005; Shumway & Stoffer, 2000). A nonstationary process is one where the statistical properties are not constant throughout the time series. For example, perhaps the mean tends to increase over time. When a process is nonstationary, a

transformation is usually taken to make the process stationary, sometimes called differencing (Bowerman et al., 2005; Shumway & Stoffer, 2000). The focus on the models below are for processes that are already stationary.

Using the office paper example from above where the current amount of office paper may be related to previous office paper amounts could be modeled by a autoregressive process. An autoregressive (AR) process is one where the current value can be explained by p past values (Bowerman et al., 2005; Shumway & Stoffer, 2000). More formally, an $AR(p)$ process is:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + w_t \tag{2.21}$$

where x_t is the value at time t, $\phi_1, \phi_2, \ldots, \phi_p$ are unknown constants, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$ are values that are 1, 2, and p points prior to the current value, and w_t is an error term (referred to as a random shock in the Box and Jenkins (1976) terminology).

A common autoregressive process used in the social sciences is the first order autoregressive process, commonly denoted as AR(1), would be:

$$x_t = \phi_1 x_{t-1} + w_t \tag{2.22}$$

where the terms are identical to equation (2.21) above. ϕ_1 represents a correlation between x_t and x_{t-1} or more simply represents the correlation between two time points separated by a lag of one. The theoretical autocorrelations for the AR(1) model would be: $\rho_k = (\phi_1)^k$ for $k \geq 1$ where k represents the time lag between measurements. This has the effect of decreasing correlations as the time lag increases. More complicated structures can be made by using a higher order AR

model (e.g. AR(2)).

Another type of time series model is the moving average (MA) model. This model says that the current value is based on a linear function of previous error terms (random shocks) (Bowerman et al., 2005; Shumway & Stoffer, 2000). More formally, a $MA(q)$ process is:

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q} \tag{2.23}$$

where x_t is the value at time t, $\theta_1, \theta_2, \ldots, \theta_q$ are unknown constants, w_t is the current error term, $w_{t-1}, w_{t-2}, \ldots, w_{t-q}$ are the prior error terms with lag 1, 2, $\ldots$, q.

The simplest MA model, the first order moving average model (MA(1)), is defined as:

$$x_t = w_t + \theta_1 w_{t-1} \tag{2.24}$$

where the terms are identical to equation (2.23) above. The theoretical autocorrelations for the MA(1) model are $\rho_1 = \frac{-\theta_1}{(1+\theta_1^2)}$ and $\rho_k = 0$ for $k > 1$. This means that all observations with a lag greater than one have no correlation and all observations with time lag of one have the correlation defined above, this structure is sometimes refered to as a toeplitz or banded structure. As the order of the MA model increases the more correlations are defined. For example, for a MA(3) model, all lags greater than three are zero and the first three lags would have nonzero correlations.

Lastly, the autoregressive and moving average processes can be combined to form the mixed autoregressive moving average (ARMA) model (Bowerman et al.,

2005; Shumway & Stoffer, 2000). An $ARMA(p, q)$ can be represented as:

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \ldots + \theta_q w_{t-q} \tag{2.25}$$

where here the terms are identical to those defined in the AR or MA process above. This model allows the current value to be based on prior values and prior error terms (random shocks).

There is also a model called an integrated ARMA model (ARIMA) which generalizes the ARMA model to include differencing (Shumway & Stoffer, 2000). The ARIMA model allows for a transformation to turn a nonstationary process to a stationary process. This model will not be discussed in further detail here and the interested reader is directed to Box and Jenkins (1976) and Shumway and Stoffer (2000).

Selecting an appropriate model for the time series data involves assessing the consistency of the sample autocorrelation function (SAC) and the sample partial autocorrelation function (SPAC) to the theoretical autocorrelation function (TAC) and theoretical partial autocorrelation function (TPAC) respectively (Bowerman et al., 2005). For example, if an AR(1) model appears to have TAC and TPAC that is consistent with the SAC and the SPAC, this model would be chosen as a starting point. Later, in model diagnostics, the adequacy of model fit would be identified and an alternative model could be fitted if needed.

2.5.2 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). There has also been some work done recently to use a data-driven technique to model the within-subject covariance matrix in LMM for longitudinal data (Pan & MacKenzie, 2003, 2006, 2007). More specifically, they use a regression approach to model the within-subject covariance matrix using polynomials of time. However, only recently have simulation studies started exploring the impact of misspecification of the level one residual structure (Ferron, Dailey, & Yi, 2002; Kwok et al., 2007; Murphy & Pituch, 2009). Other simulation studies have explored this notion using a latent growth curve approach (Sivo & Willson, 2000; Sivo, Fan, & Witta, 2005; Grimm & Widaman, 2010). There has also been a recent article discussing a bayesian framework for removing the independence assumption at both level one and level two. (Browne & Goldstein, 2010). This has the ability to represent three levels of clustering in a two level model and also flexibly model the covariance structure of the level one residuals.

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). These three terms allow for an easier way to talk about how the structure is misspecified.

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, Dailey, & Yi, 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, Dailey, & Yi, 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, Dailey, & Yi, 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, Dailey, & Yi, 2002; Kwok et al., 2007; Verbeke & Molenberghs, 2000).

2.5.3 Selecting a Covariance Structure

In most cases when researchers fit a LMM, they are interested in doing more than just looking at the fixed effect estimates. Therefore, some care and thought should be taken to selecting the covariance structure, however there are few methods that have been researched to determine their effectiveness in selecting the correct covariance structure. The few studies that have explored methods of selecting the correct covariance structure have found it difficult to empirically select the correct structure (Ferron, Dailey, & Yi, 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, Dailey, and Yi (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%. Ferron, Dailey, and Yi (2002) found that more time points, more participants, and higher levels of autocorrelation improved correct identification. In contrast to Ferron, Dailey, and Yi (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 should be modeled.

Verbeke and Molenberghs (2000) outline an informal check for the presence of serial correlation. In their technique, they define a set of transformed residuals that remove the variability attributable to the random effects. If these transformed residuals do not follow an approximately normal distribution, it may indicate that serial correlation is missing from the model.

Another option to explore the need of a serial correlation component of the LMM is the variogram which is related to the covariance function. If the response variable follows a stationary process, the variogram and covariance function are equivalent (Diggle, 2002; Sithole & Jones, 2002). The variogram has the added advantages of being well-defined for a few non-stationary processes and is easier to estimate from data with irregular time points (Diggle, 2002; Sithole & Jones, 2002). From a variogram, one can estimate the measurement error variance, how large the variation in the random effects is, and the contribution of the serial correlation component (Sithole & Jones, 2002).

Table 2.1 summarize the assumption violations and the implications to the fixed and random effects. As can be seen from the table, almost all of the problems in estimation center around the random effects or standard errors. As a result if parameter estimation of the fixed effects is the primary purpose less care needs to be taken to monitor assumption violations. However, if inferences or estimation of the variance components is desired then more care needs to be taken to ensure that model assumptions are met or realize the implications when talking about the results.

Table 2.1: Summarizing implications when model assumptions have been violated

Violation	Implication FE	Severity FE	Implication SE FE	Severity SE FE	Implication RE	Severity RE
Small within cluster size	Can lead to poor predictive power	Medium**	Likely larger standard errors	Low***	May make estimation difficult, especially with many random effects	Medium***
Few Clusters	May make estimation difficult	Medium**	Larger standard errors	Low***	Estimation difficulties of random effects, and standard errors may be biased	Medium***
Non-normal Residuals	No bias	None*	No bias	None*	No bias	None*
Non-normal RE	No bias	None**	No bias	None**	Some evidence of overestimation	Medium**
Model Misspecification	Biased parameter estimates	Large*	Never specifically explored	Unknown*	Never studied, but may cause bias in other random components	Unknown*
US or GS Covariance Structure	Little to no bias	Low**	Within individual intercept and slope overestimated	High**	Overestimation of random effects. Residuals underestimated as a result	High**
OS Covarariance Structure	Little to no bias	Low*	No bias found	None*	Produced smallest estimates of random effects, however still biased	High*

Note: FE is fixed effects, SE is standard error, RE is random effects, US is underspecified, GS is generally misspecified, OS is overspecified, * = contraditory/minimal evidence, ** = moderate evidence, and *** = substantial evidence.

Chapter 3

Methodology

For all statistical models, parameter estimates, inferences, and conclusions are only valid if the statistical assumptions for the model are adequately met or evidence exists of their robustness to assumption violations. Important assumptions for the linear mixed model (LMM) are 1) the random effects are independent and identically (normally) distributed (iid) with mean zero and covariance matrix G and 2) the residuals are independent and normally distributed with mean zero and common variance σ^2 with no correlation between measurement occasions (sometimes called an independence correlation structure). Put more simply, the common model assumptions are: $b_j \sim iid\, N(0, \mathbf{G})$ and $e_{ij} \sim iid\, N(0, \sigma^2)$. In addition, random sampling at the cluster level is also assumed.

Under certain conditions, especially when the correlation between measurements decreases slower than expected or with measurement occasions close in time (i.e. daily or weekly measurements) the independence assumption of the residuals may not be tenable. In these cases it may be beneficial to model a more

complex correlation structure, such as an autoregressive or moving average process commonly found in the time series literature (see Bowerman et al. (2005), Box and Jenkins (1976), and Shumway and Stoffer (2000) for more details).

These topics were explored further in the current simulation study. The first study explored conditions where the data are not "perfect," such as when the random effects come from a distribution other than a normal distribution. The second study explored the ability for serial correlation to overcome model misspecification of the random components of the model. More specifically, if a random effect for time underlies the data but is erroneously not modeled. This form of misspecification could occur if convergence problems arise during estimation and the random slope is dropped to aid in convergence. Both of these studies look to inform applied researchers as to which situations are most troublesome when fitting a LMM.

3.0.4 Research Questions

The following research questions was explored in the current simulation study:

1. To what extent are parameter estimates (fixed and random effects) biased when the covariance structure is incorrectly modeled?

2. Does the relationship in question 1 generalize to real world data conditions such as non-normal random effect distributions?

3. To what extent can a more complicated covariance structure overcome a missing random effect for time?

3.1 Study One

3.1.1 Research Design

A factorial research design was used for the computer simulation study. Previous simulation work (Ferron, Dailey, & Yi, 2002; Kwok et al., 2007; Murphy & Pituch, 2009) has 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 fit (i.e. ID, AR(1), MA(1), MA(2), ARMA(1,1)) resulting in a total of $30,000 * 5 = 150,000$ models.

3.1.2 Sample

Population parameters were generated from 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).

3.1.3 The Model

The basic model used in the simulation study can be seen in equation (3.1) shown below.

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 time_{ij} + \beta_2 diff_{ij} + \beta_3 ACT_{1j} + \beta_4 ACT_{1j} time_{ij} \\ + b_{0j} + b_{1j} time_{ij} + e_{(1)ij} + e_{(2)ij} \end{aligned} \tag{3.1}$$

In this equation, let i represent repeated measurements and j represent individuals. The fixed effects are represented by $\beta_0, \beta_1, \beta_2, \beta_3$, 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, ACT_{1j} 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}, $e_{(1)ij}$, 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 equation (3.1), where the $e_{(2)ij}$ and the distribution of the random components were the primary differences between the simulated conditions.

Table 3.1 shows the population values used to generate the data according to equation (3.1). Table 3.1 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. Dividing

by a very small number may artificially inflate the bias statistic. Potentially more problematic is the relatively small variance of the random slope for time. Here the parameter value is close to the boundary of zero as variances cannot be negative. This may introduce some estimation difficulties later.

Table 3.1: Population 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
Cor RE	0
Var e_{ij}	0.549
ϕ_1	0.45
θ_1	0.50
θ_2	0.30
Var $diff_{ij}$	1.250
Var ACT_{1j}	4.905

3.1.4 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:

$$Rel.Bias = \frac{\hat{\theta} - \theta}{\theta} \tag{3.2}$$

where $\hat{\theta}$ is the parameter estimate (i.e. β_k or $Var(b_{lj})$) 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:

$$Z = \frac{\hat{\beta} - \beta}{\hat{SE}} \tag{3.3}$$

where $\hat{\beta}$ is the parameter estimate, β is the simulated paramater value shown in Table 3.1, and $\hat{SE}$ 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 siulation 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:

$$\begin{aligned} Y_{ijklmn} = \mu &+ \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(l)} + \alpha_{D(m)} + \alpha_{E(n)} \\ &+ \alpha_{AB(jk)} + \alpha_{AC(jl)} + \alpha_{AD(jm)} + \alpha_{AE(jn)} + \alpha_{BC(kl)} \\ &+ \alpha_{BD(km)} + \alpha_{BE(kn)} + \alpha_{CD(lm)} + \alpha_{CE(ln)} + \alpha_{DE(mn)} \\ &+ \alpha_{ABC(jkl)} + \alpha_{ABD(jkm)} + \alpha_{ABE(jkn)} + \alpha_{ACD(jlm)} + \alpha_{ACE(jmn)} \\ &+ \alpha_{ADE(jmn)} + \alpha_{BCD(klm)} + \alpha_{BCE(kln)} + \alpha_{BDE(kmn)} + \alpha_{CDE(lmn)} \\ &+ \alpha_{ABCD(jklm)} + \alpha_{ABCE(jkln)} + \alpha_{ACDE(jlmn)} + \alpha_{BCDE(klmn)} \\ &+ \alpha_{ABCDE(jklmn)} + e_{ijklmn}. \end{aligned} \tag{3.4}$$

The above equation represents a factorial UANOVA that fits all possible interactions. In Equation (3.4), the α represent cell means, μ is the grand mean, the first

set of subcripts A, B, C, D, E represent the five simulation conditions, the subscripts in parentheses j, k, l, m, n index the factor categories, and lastly i depicts the observation number.

The model for the empirical type I error rates is simplified compared to equation (3.4) because there was only one observation per cell. As a result the higher order interaction terms were pooled into the error term. The model took the following form:

$$\begin{aligned} Y_{ijklmn} = \mu &+ \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(l)} + \alpha_{D(m)} + \alpha_{E(n)} \\ &+ \alpha_{AB(jk)} + \alpha_{AC(jl)} + \alpha_{AD(jm)} + \alpha_{AE(jn)} + \alpha_{BC(kl)} \\ &+ \alpha_{BD(km)} + \alpha_{BE(kn)} + \alpha_{CD(lm)} + \alpha_{CE(ln)} + \alpha_{DE(mn)} \\ &+ \alpha_{ABC(jkl)} + \alpha_{ABD(jkm)} + \alpha_{ABE(jkn)} + \alpha_{ACD(jlm)} + \alpha_{ACE(jmn)} \\ &+ \alpha_{ADE(jmn)} + \alpha_{BCD(klm)} + \alpha_{BCE(kln)} + \alpha_{BDE(kmn)} \\ &+ \alpha_{CDE(lmn)} + e_{ijklmn}. \end{aligned} \tag{3.5}$$

The above model contains all terms up to three-way interactions and pools the four-way and five-way interactions 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:

$$\eta^2 = \frac{SS_{trt}}{SS_{total}}. \tag{3.6}$$

In the above equation, SS_{trt} is the amount of variation attributable to the treatment of interest (e.g. covariance structure) and SS_{total} is the total sum of squares or the total amount of variation in the dependent variable. This ratio also has the nice interpretation as the percent of variation in the dependent variable explained by the independent variable of interest. The factorial design in this case also allows the interpretation of the η^2 metric as the unique percent of variation explained by each independent 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 a cpu to be used simultaneously improving the data simulation speed. Model fitting was done with the `nlme` package within `R` (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2012). Lastly, in order to check if the simulated data follows the data conditions specified by the design, the sample autocorrelation function was plotted to see if the values approximately follow the theoretical autocorrelation function for each distribution.

In addition, the empirical skewness and kurtosis of the simulated random effect distribution was computed to check for accurate random effect simulation.

3.2 Study Two

For study two, no additional data were generated. The same data generated in study one were used and another set of models were fitted to the 30 conditions when the number of subjects were the smallest (i.e. 25 subjects). All five covariance structures were again fitted, but this time without the random slope included in the model. This set of conditions were analyzed separately from the previous conditions. This study hopes to address the ability to fit a LMM under small sample size conditions when the addition of another random component (i.e. a random slope) leads to convergence problems. When convergence problems are found, many researchers would fix a random effect to be zero which simplifies the likelihood function and may alleviate convergence issues.

The model fit to the relative bias statistics was identical to equation (3.4) except that the terms with an E subscript were dropped as the number of individuals was constant and does not need to be modeled.

3.3 Simulated Data Checking

A monte carlo simulation is only helpful if the data generated closely follows the simulated data conditions. Specifically for this simulation, ensuring that the random effect distribution, residual distribution, and serial correlation structures closely follow theoretical values improves the conclusions that can be drawn from

this study.

3.3.1 Residual Distributions

To check the adequacy of the simulation random effect and residual distributions, skewness and kurtosis values were calculated for each replication. In the long run, the skewness and kurtosis values should center around the mean skewness and kurtosis values for each theoretical distribution shown in Table 3.2. The empirical values should also not depend on the serial correlation model assumed to underly the data as well. To explore this, the skewness and kurtosis values were plotted for each serial correlation structure.

Table 3.2: Theoretical skewness and kurtosis values for the normal, chi-square (1), and Laplace distributions.

Distribution	Skewness	Kurtosis
Normal	0	0
Chi-Square (1)	$\sqrt{8}$	12
Laplace	0	3

Density plots depicting the empirical skewness values for the random effects can be seen in Figure 3.1 to Figure 3.5. Investigating the density plots reveals that they are very similar across the five serial correlation structures suggesting that the generated serial correlation structure does not have an impact on the skewness of the random effect distribution. This was expected given that these two conditions should be independent of one another. In addition, there does not appear to be any noticeable differences between the two random effects suggesting that the skewness value does not depend on the random effect being simulated. Lastly, the skewness

values are centered around zero when the random effects were simulated from a normal distribution and Laplace distribution, which follows the theorectical values shown in Table 3.2. Not surprisingly, the Laplace distributed random effects have a larger variance of skewness values than the normal distributed random effects due to the larger theoretical kurtosis. When the random effects are simulated from a chi-square distribution with one degree of freedom, the density is centered just under 2. This is smaller than the theoretical value, however is still different from the other two distributions with little overlap compared to the other two distributions.

Density plots depicting the empirical kurtosis values for the random effects can be seen in Figure 3.6 to Figure 3.10. There was a similar trend for the empirical kurtosis values compared to the empirical skewness density plots. Specifically, the kurtosis values do not appear to depend on the serial correlation structure or the random effects themselves, suggesting that the simulation of the distribution of the random effects are not related to these conditions. The empirical kurtosis values when a normal distribution was simulated centers around zero with very little variation. The Laplace and chi-square distributions are shifted slightly above zero reflecting heavier tails compared to a normal distribution, although these values are not as large as the theoretical values shown in Table 3.2. Even though, these values are different from the theoretical values, they do follow the same pattern as the theoretical values shown in Table 3.2 with normal, Laplace, and chi-square distributions reflecting increasing average empirical kurtosis values.

Looking at the density plots of the empirical skewness and kurtosis values together does not suggest any significant problems with the generation of the

different distributions for the random effects. The simulated random effect distribution does not seem to be influenced by the other simulation conditions and the empirical patterns mimic the theoretical patterns. Also, even though the empirical values do not exactly replicate the theoretical values, they do seem to be different from one another and provide varying data conditions that may represent real-world data.

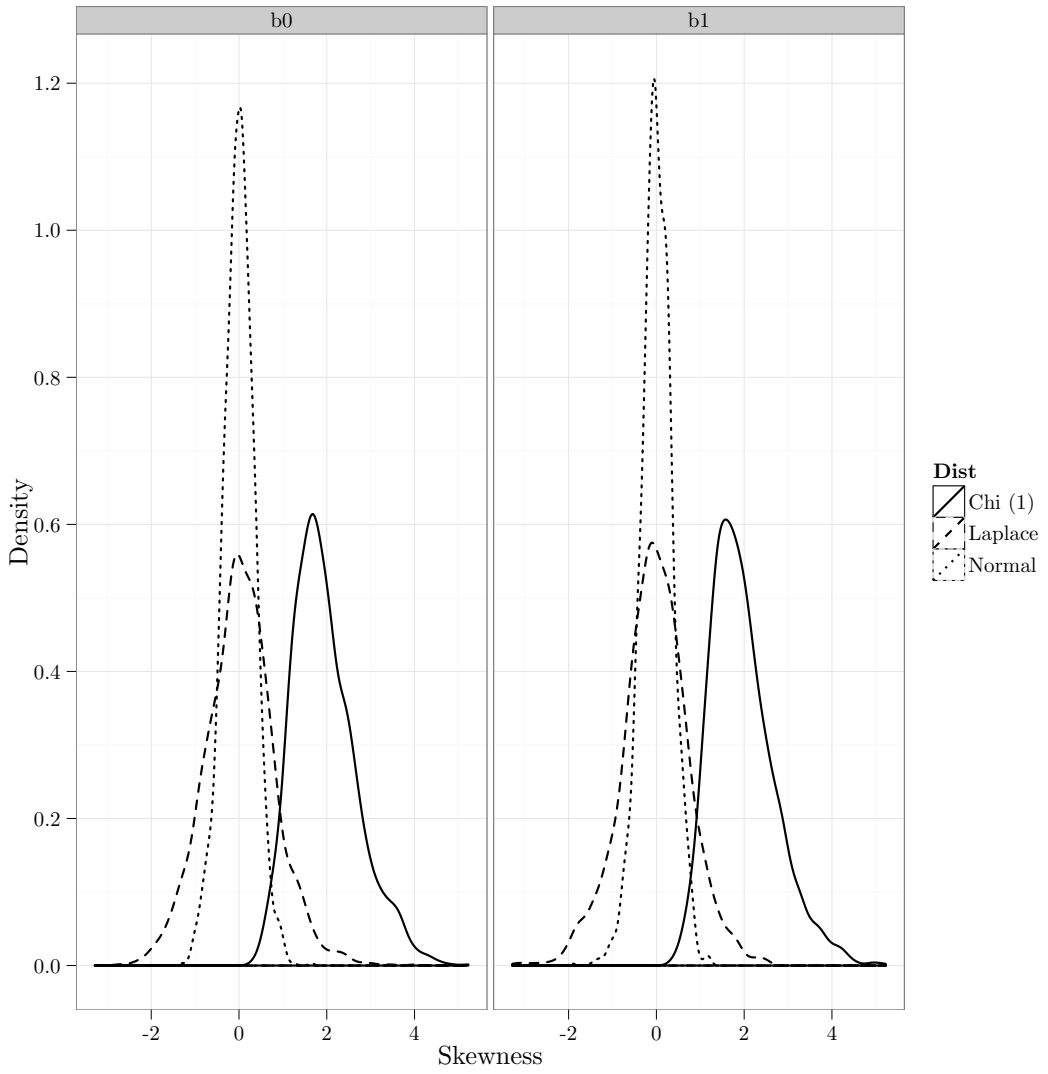


Figure 3.1: Skewness values of the random effects for the independent serial correlation structures

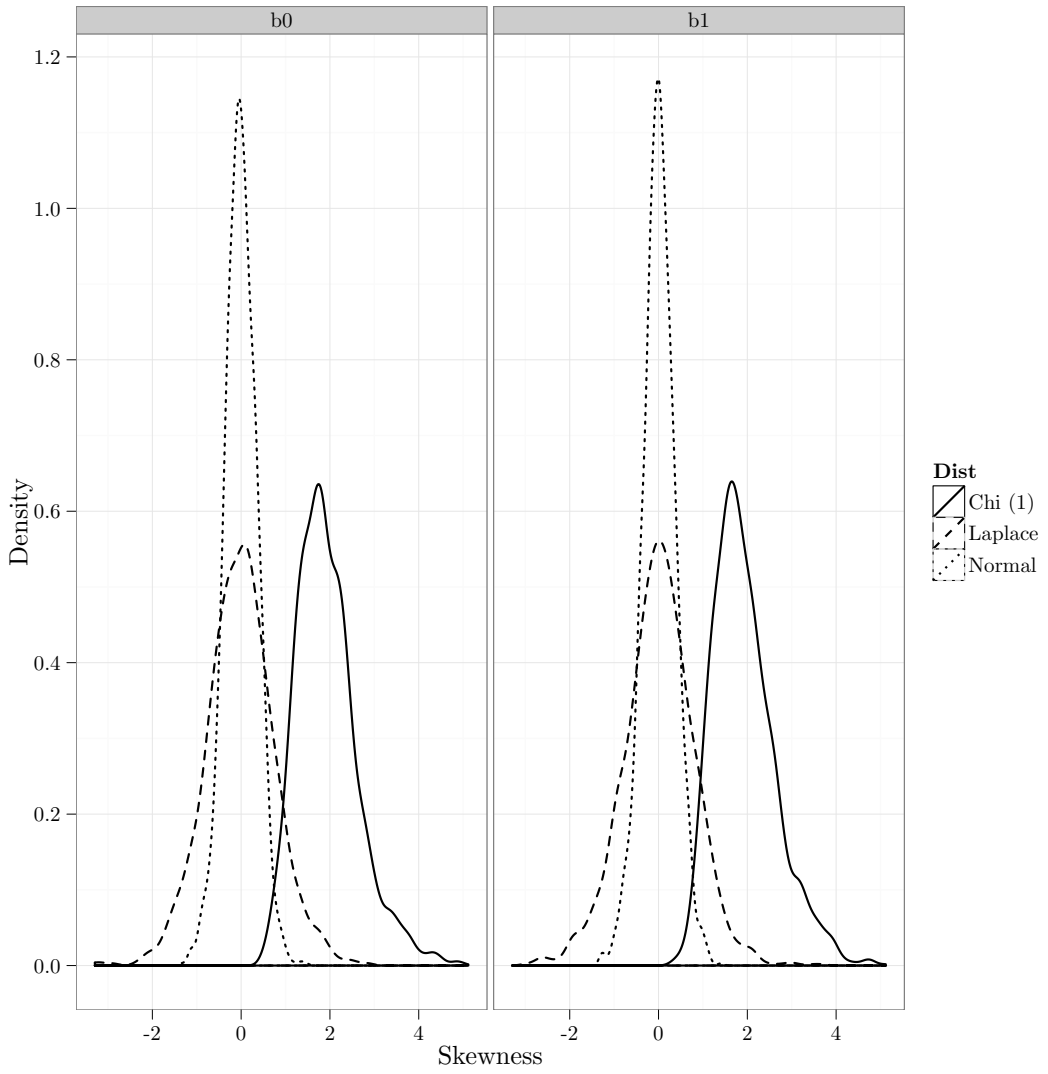


Figure 3.2: Skewness values of the random effects for the AR(1) serial correlation structures

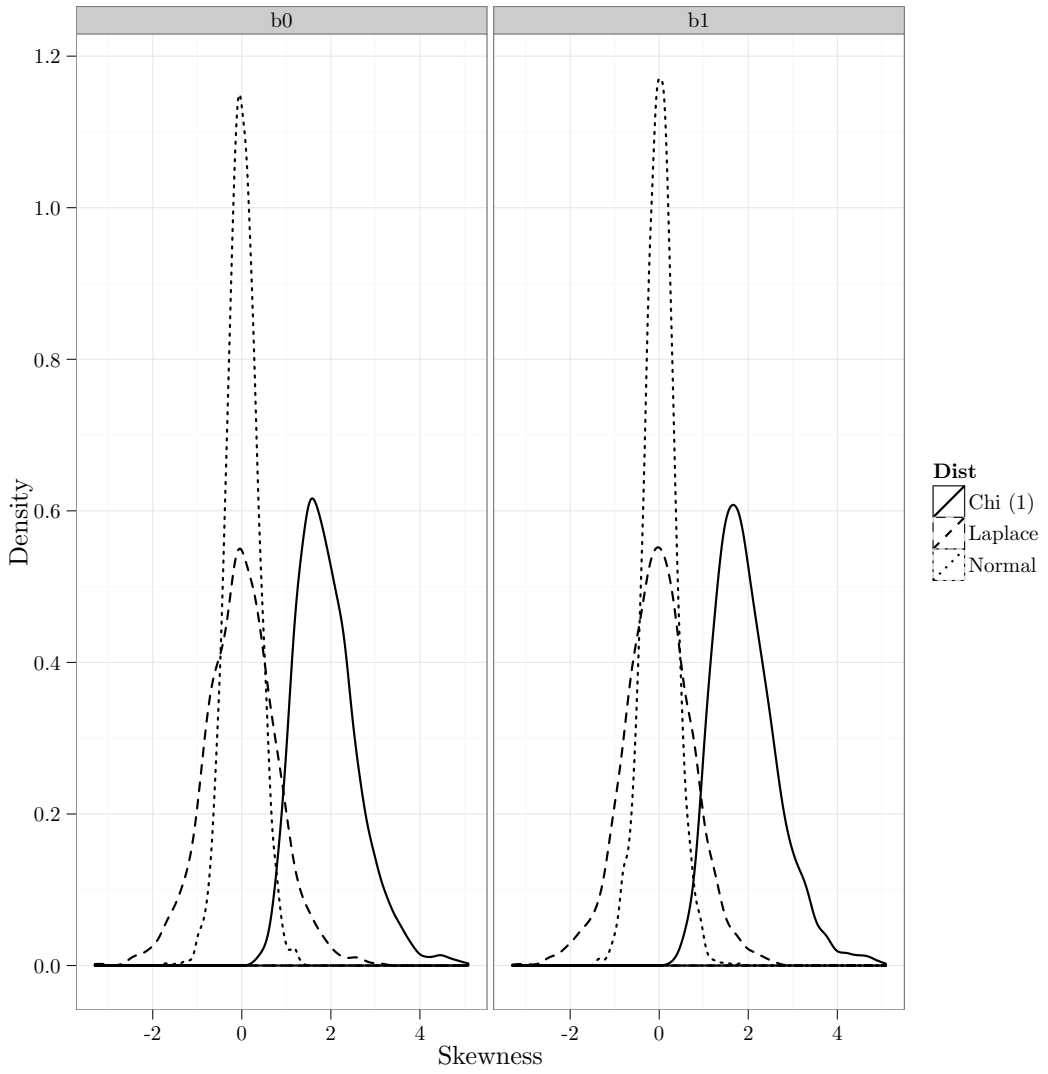


Figure 3.3: Skewness values of the random effects for the MA(1) serial correlation structures

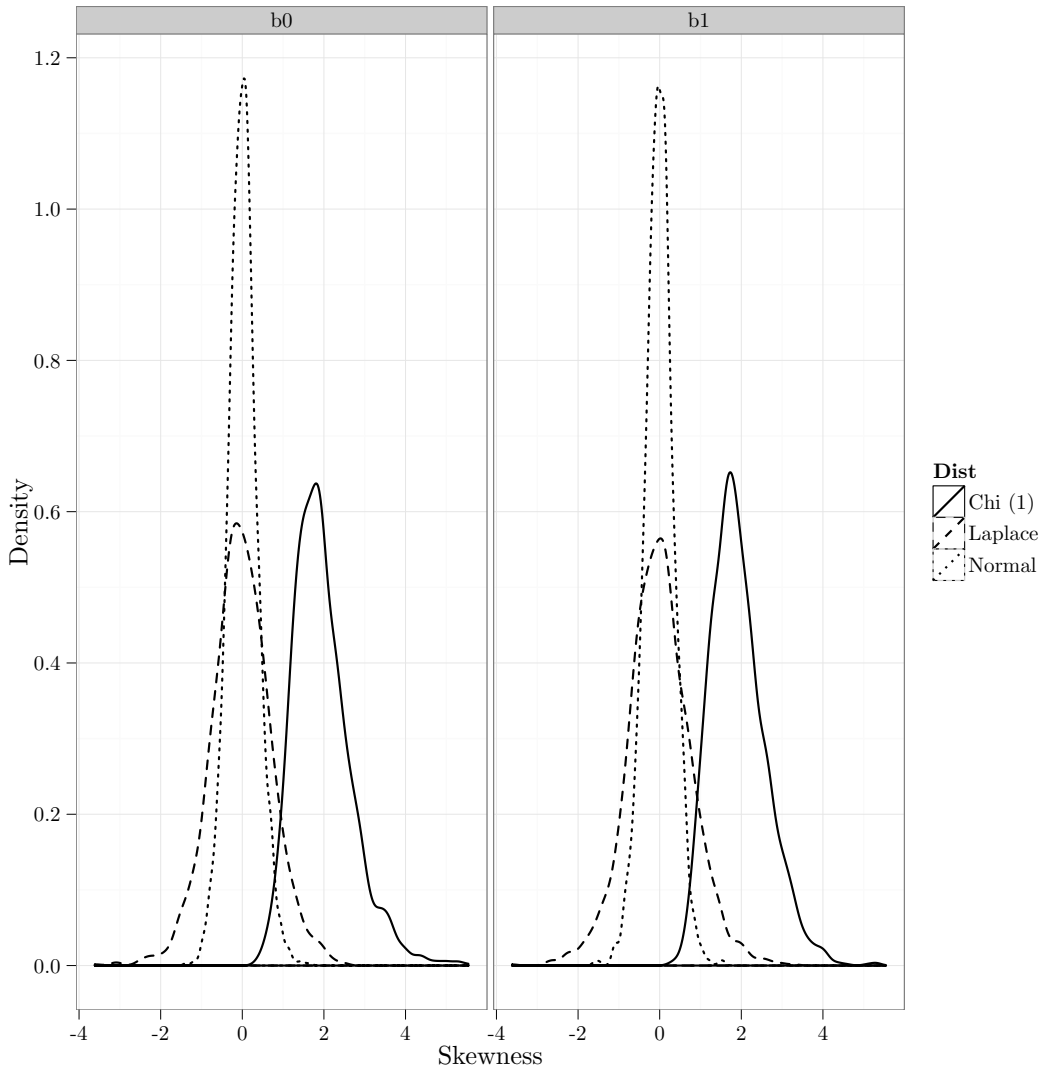


Figure 3.4: Skewness values of the random effects for the MA(2) serial correlation structures

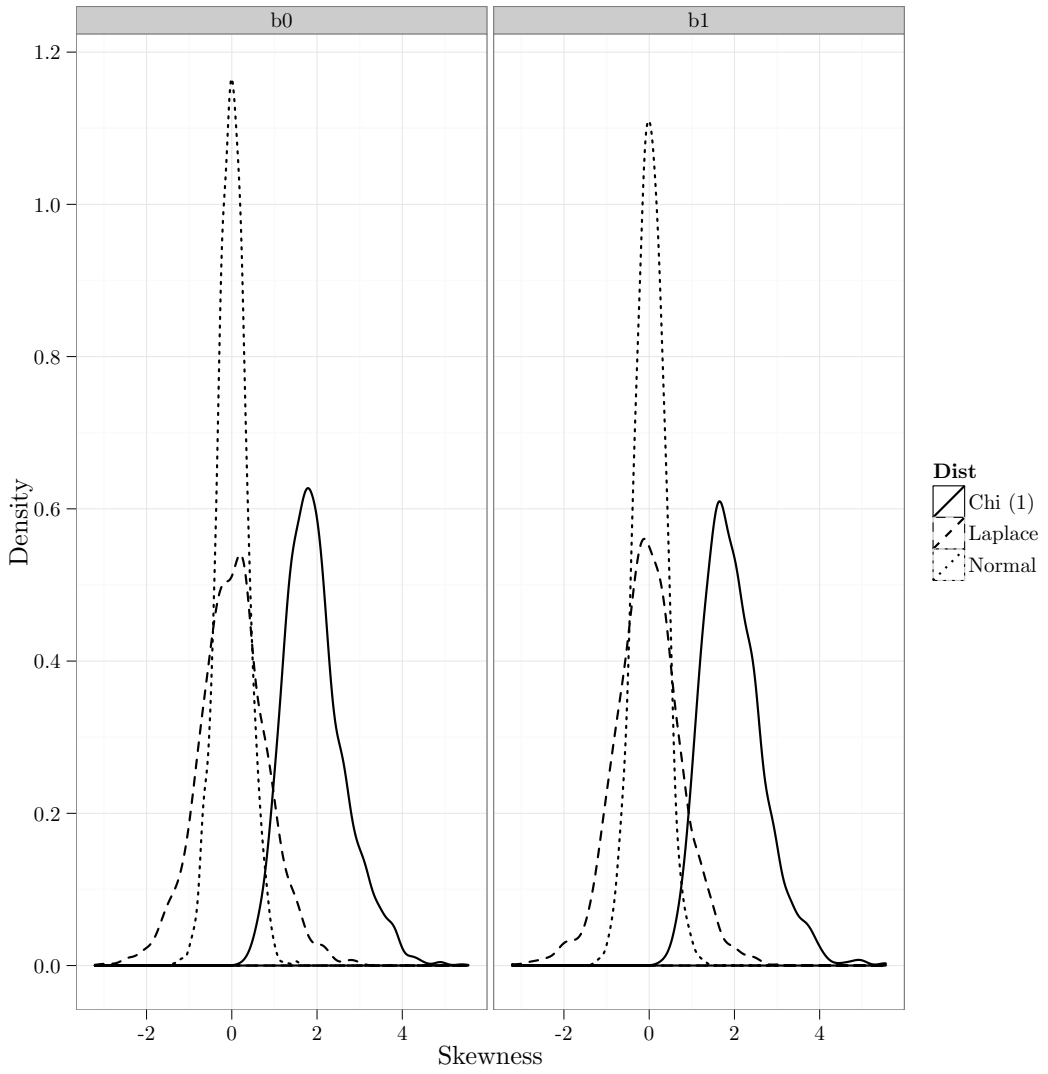


Figure 3.5: Skewness values of the random effects for the ARMA(1,1) serial correlation structures

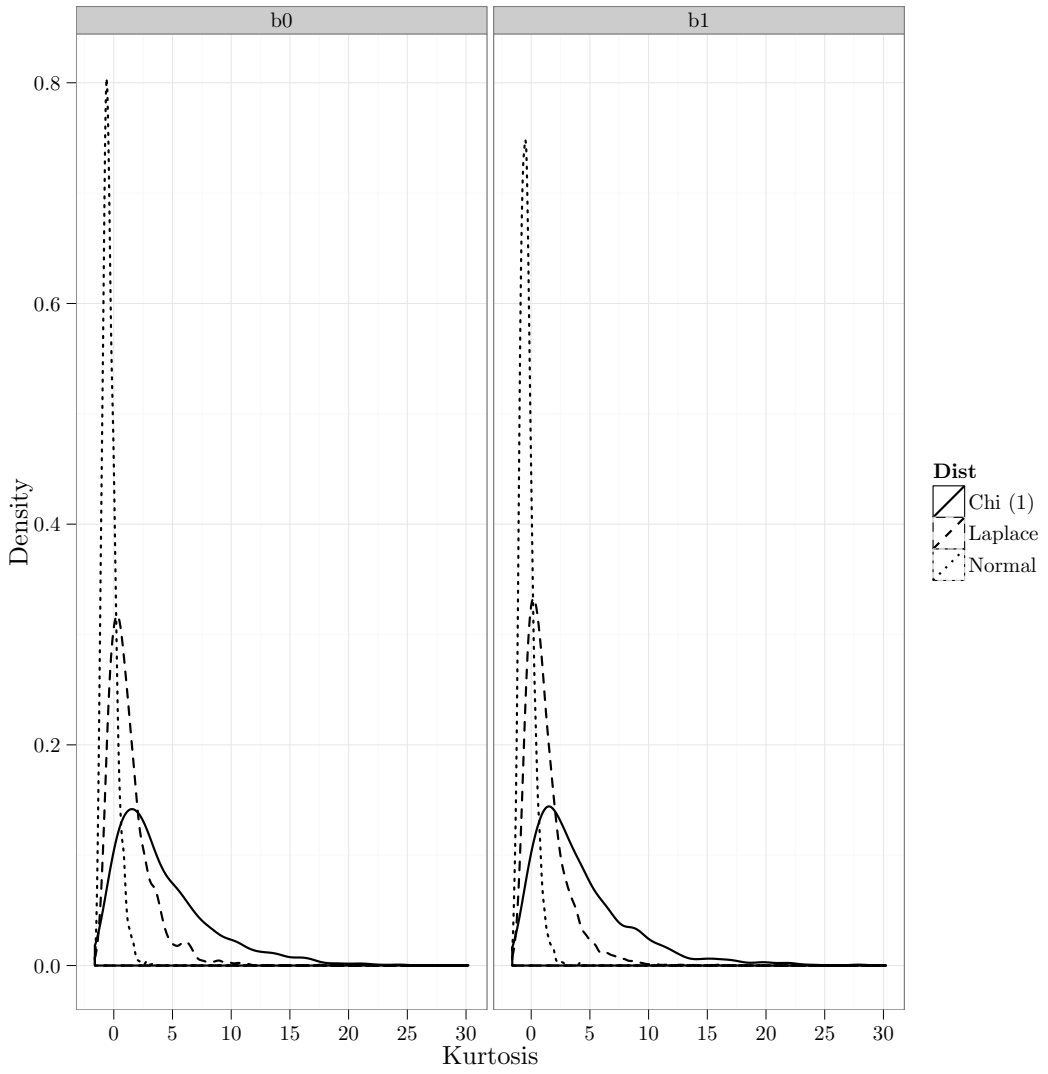


Figure 3.6: Kurtosis values of the random effects for the independent serial correlation structures

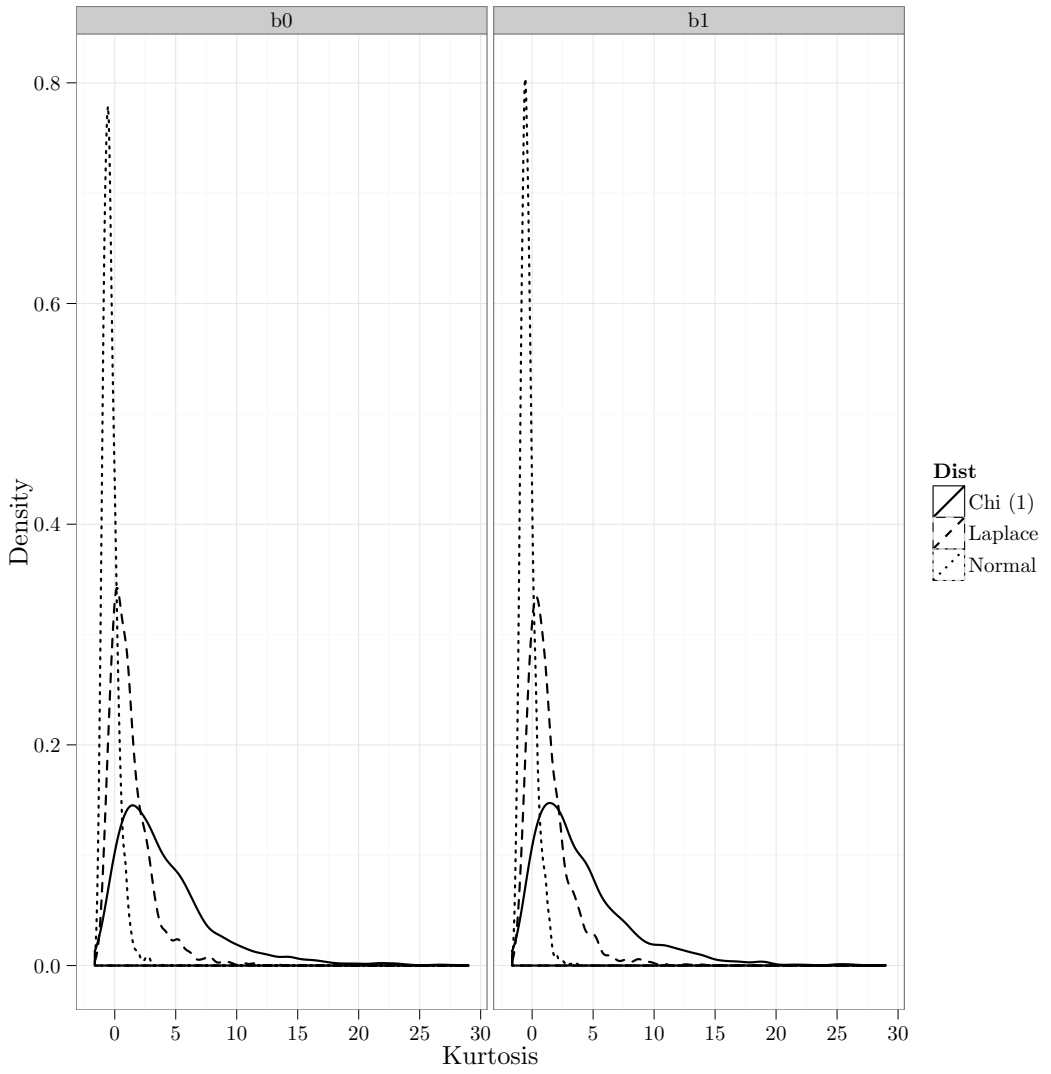


Figure 3.7: Kurtosis values of the random effects for the AR(1) serial correlation structures

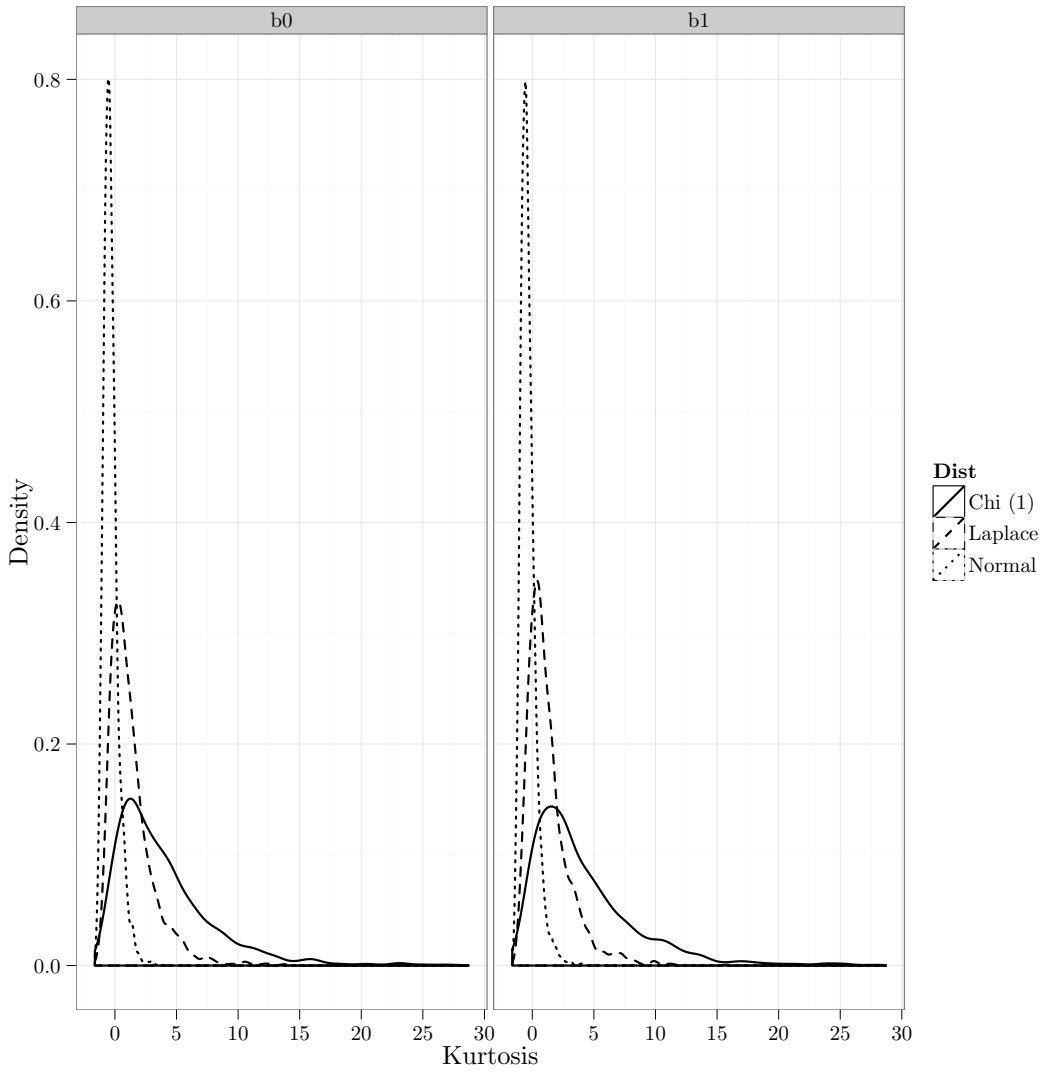


Figure 3.8: Kurtosis values of the random effects for the MA(1) serial correlation structures

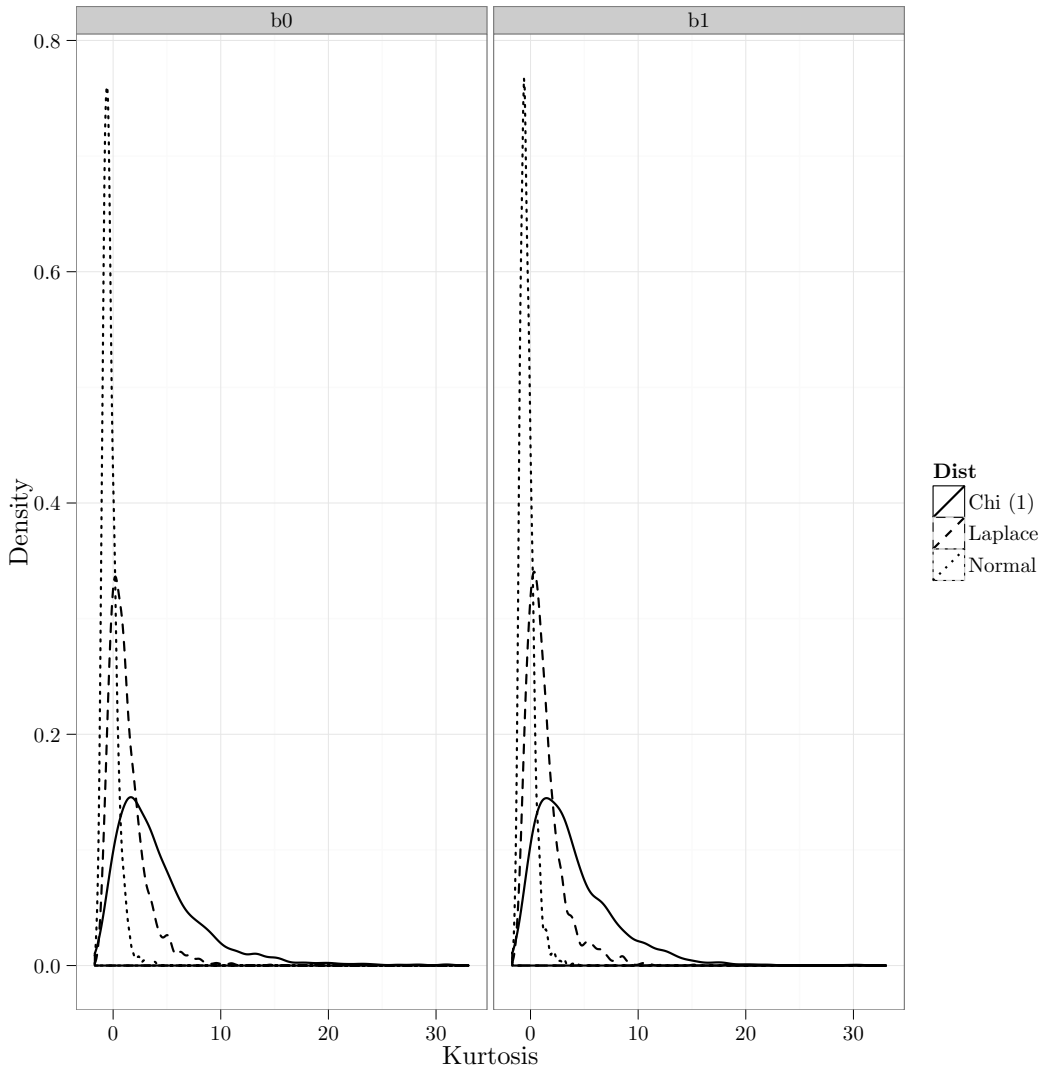


Figure 3.9: Kurtosis values of the random effects for the MA(2) serial correlation structures

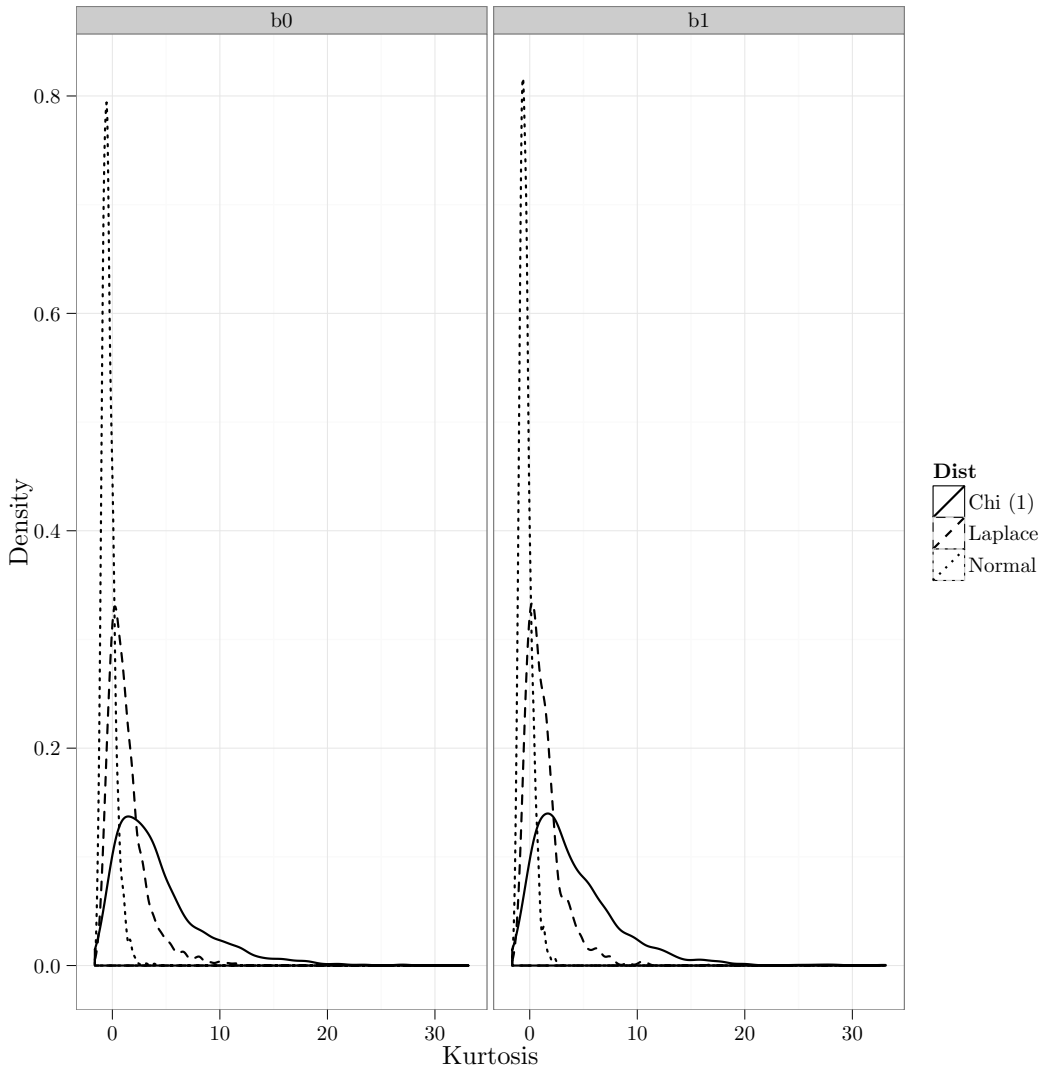


Figure 3.10: Kurtosis values of the random effects for the ARMA(1,1) serial correlation structures

3.3.2 Serial Correlation

To check how accurately the serial correlation portion of the model was simulated, the average autocorrelation for each time lag was calculated for every replication. These sample values were compared to theoretical autocorrelation values for the specific serial correlation model. The theoretical autocorrelation values for each serial correlation model can be seen in Table 3.3.

Table 3.3: Theoretical autocorrelation values for independent, AR(1), MA(1), MA(2), and ARMA(1,1) serial correlation models.

Model	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7
Independent	0	0	0	0	0	0	0
AR(1)	.450	.203	.091	.041	.018	.008	.004
MA(1)	.400	0	0	0	0	0	0
MA(2)	.485	.224	0	0	0	0	0
ARMA(1,1)	.685	.308	.139	.062	.028	.013	.006

Note: ϕ_1 for AR functions is .45, θ_1 for MA functions is .5, θ_2 for MA functions is .3.

Ensuring the simulated autocorrelations were similar to the theoretical values proved much more difficult than simulating the random effect distributions. During the initial planning of the simulation data conditions, four and eight observations within each subject was chosen. A density plot of the empirical autocorrelations with four observations when an AR(1) serial correlation was simulated is shown in Figure 3.11. The different lines reflect the density of the three possible lags when the time series length is four. The pattern diverged substantially from the theoretical values with large amounts of variation. In addition, the density of each lag was not normally distributed and for a lag of one the density was closer

to a uniform distribution. This same pattern in the density of the autocorrelations at each lag was reflected in each serial correlation structure simulated. In an attempt to replicate the results, simulating the error terms were conducted within the statistic program STATA, where the same trend was found.

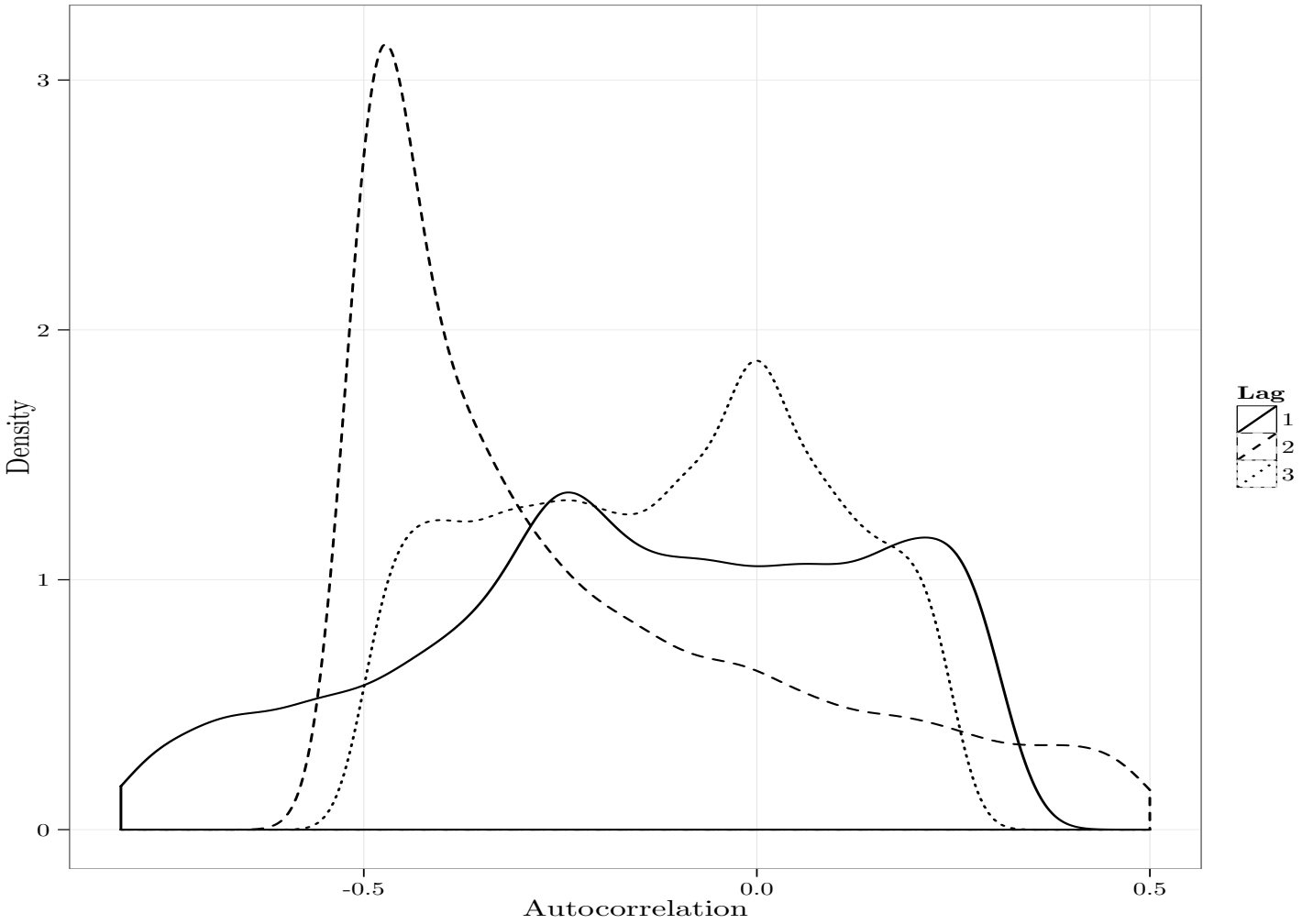


Figure 3.11: Density plots of autocorrelation for six observations

Such a small time series proved to have too much uncertainty in the precision of the autocorrelation functions. This can be seen in the equation for the standard error of the autocorrelation function:

$$SE_{r_k} = \sqrt{\frac{1 + 2\sum_{l=0}^{k-1} r_l^2}{N}} \tag{3.7}$$

where k represents the current lag, l represents the previous lag, and N is the series length. The standard error is the smallest at lag 1 where the formula simplifies to $1/\sqrt{N}$. Therefore, with a series length of four, the standard error for the first lag would be: $1/\sqrt{4} = .5$ and the standard error of all other lags would be greater than this. This shows the amount of uncertainty in the autocorrelation function where every possible value for a correlation would be a plausible value. As a result, the smallest series length was increased from four observations to six in the final simulation design.

With six observations, the standard error for the first lag would be $1/\sqrt{6} \approx .408$. This is still a large standard error, but significantly smaller than .5. Also, as can be seen below in Figure 3.12 depicting the density plot of the autocorrelations at the first three lags, the empirical autocorrelations behaved much better (i.e. approximately normally distributed). The different serial correlation structures also start to diverge with six observations. The independent serial correlation structure was centered at the same value just below zero for all of the lags which was close to the theoretical value. In contrast, the other serial correlation structures have an average autocorrelation at lag one greater than zero with the ARMA(1,1)

structure having the largest average value. Although none of the average autocorrelations are as high as the theoretical values shown in Table 3.3, they do start to diverge and is a reflection of the uncertainty in the autocorrelation function with such a small time series length.

Figure 3.13 shows the density plot for the first three lags when the time series is eight. The autocorrelation is again better behaved resembling a normal distribution with less variation than the time series of six. The average autocorrelation in Figure 3.13 for the first lag also tends to be closer to the theoretical value compared to Figure 3.12. This likely reflects the additional precision with a longer time series.

Although the average autocorrelation values are smaller than what would be expected theoretically, they are different from one another and roughly follow the theoretical trend for each structure. For example, the ARMA(1,1) structure theoretically has a structure that declines more slowly than the other structures and the density plots seem to reflect this as well. As a result, the simulation was continued, although the amount of uncertainty in the autocorrelation is a limitation of the current study.

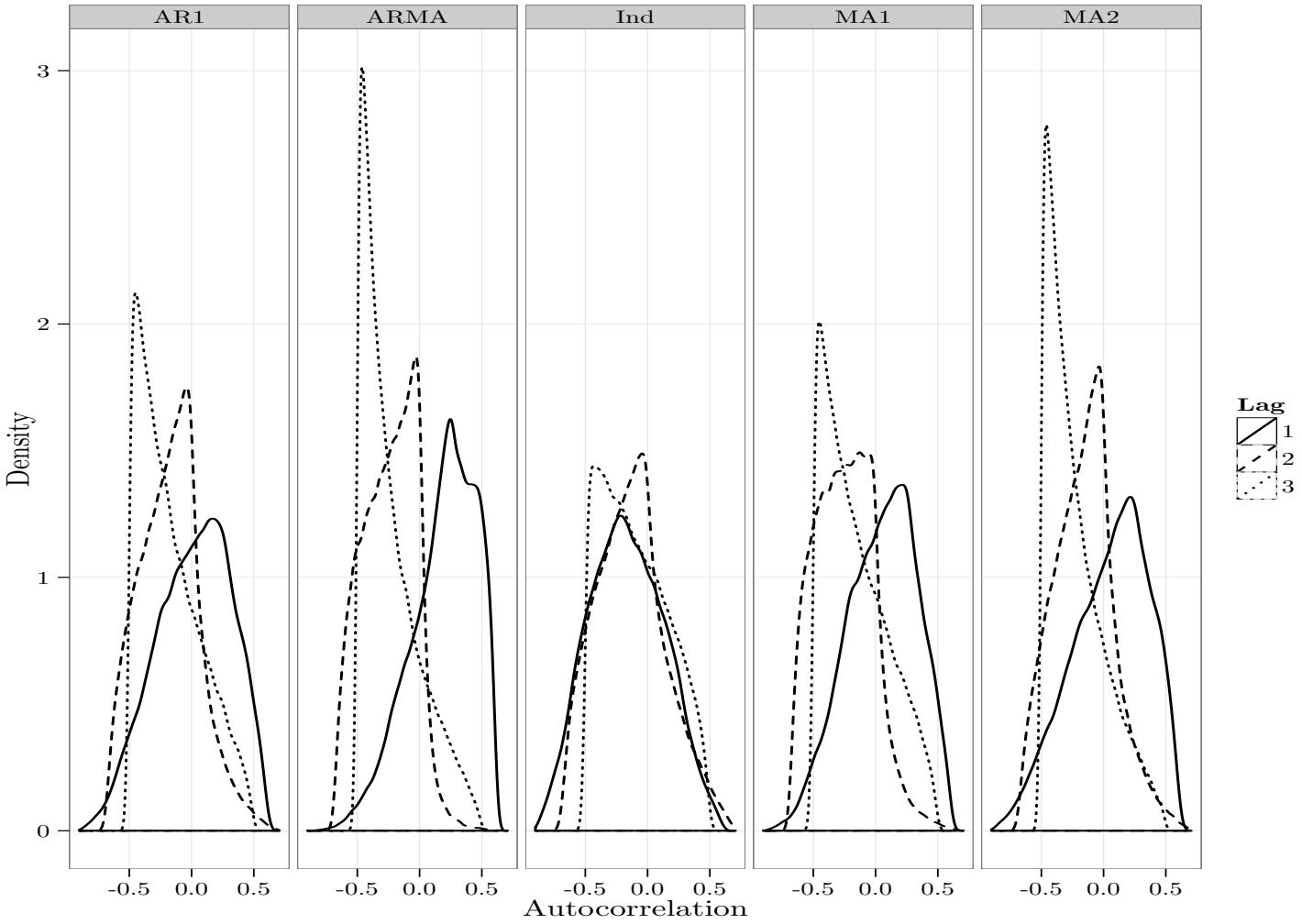


Figure 3.12: Density plots of autocorrelation for six observations

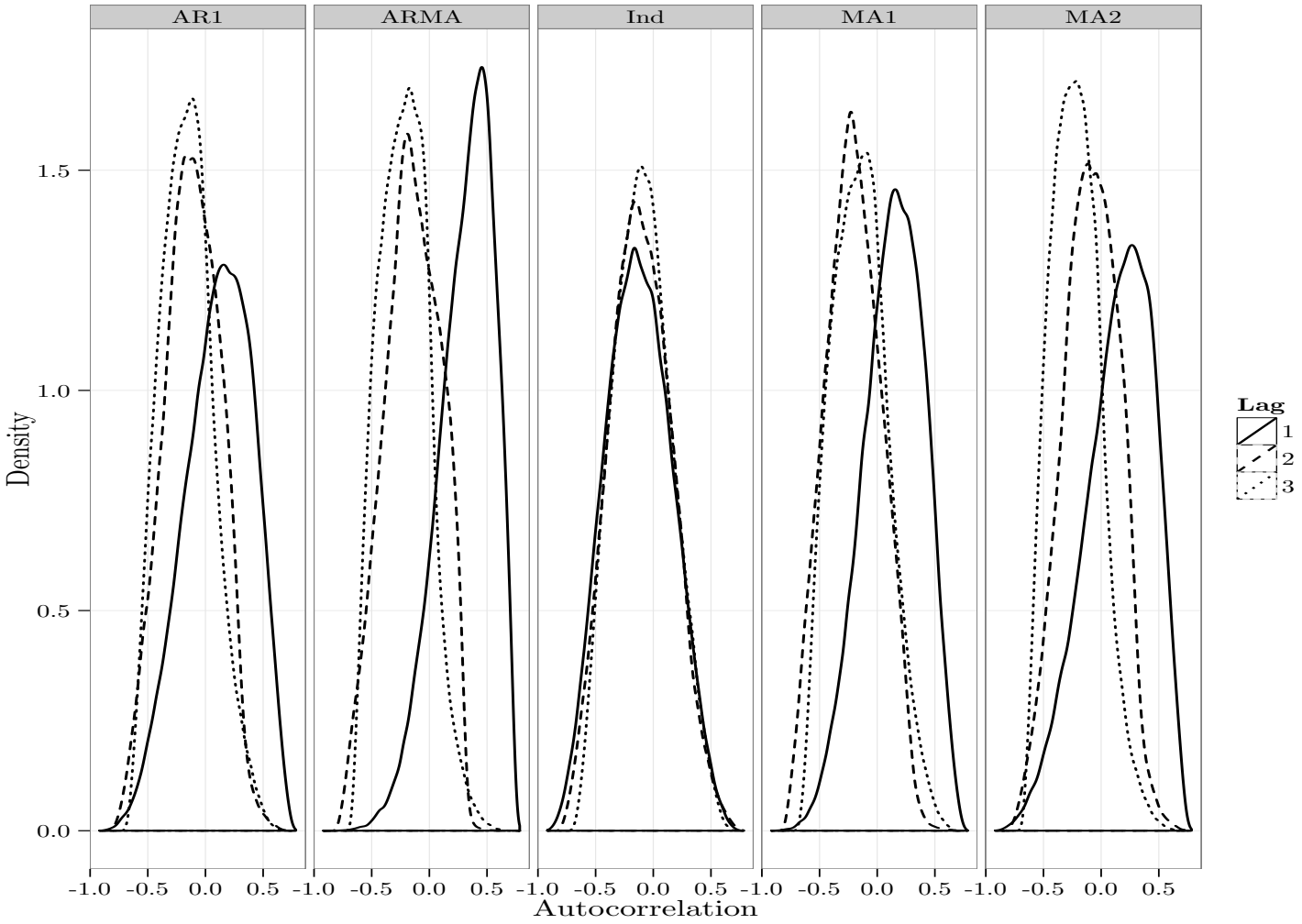


Figure 3.13: Density plots of autocorrelation for eight observations

Chapter 4

Results

4.1 Study One

4.1.1 Convergence

The convergence rates for study one can be seen in Table 4.1. 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 42.2% to a high of 98.4%. Low convergence rates tended to occur when the serial correlation structure was overspecified (e.g. ARMA(1,1) structure fit to an AR(1) structure) or when a generally misspecified serial correlation structure was fitted (e.g. AR(1) structure fit 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 indpendent structure had the best convergence rate, which was not surprising as no additional terms need to be estimated with an independent structure.

Table 4.1: Convergence rates by generated serial correlation structure and fitted serial correlation structure.

Gen SC	Fit SC	Study 1 Conv %	Study 2 Conv %
Ind	Ind	72.48	100.00
Ind	AR1	68.38	100.00
Ind	MA1	71.02	100.00
Ind	MA2	67.23	99.77
Ind	ARMA	65.10	93.17
AR1	Ind	93.88	100.00
AR1	AR1	64.88	100.00
AR1	MA1	81.37	100.00
AR1	MA2	70.78	99.93
AR1	ARMA	60.45	98.73
MA1	Ind	92.23	100.00
MA1	AR1	55.12	99.97
MA1	MA1	69.15	100.00
MA1	MA2	65.93	99.90
MA1	ARMA	63.68	99.80
MA2	Ind	95.62	100.00
MA2	AR1	61.98	100.00
MA2	MA1	84.50	100.00
MA2	MA2	68.83	99.97
MA2	ARMA	54.88	99.13
ARMA	Ind	98.37	100.00
ARMA	AR1	42.17	100.00
ARMA	MA1	88.02	100.00
ARMA	MA2	72.90	99.80
ARMA	ARMA	63.60	100.00

Note: Gen is generated, SC is serial correlation,

Fit is fitted, Conv is convergence

4.1.2 Relative Bias

Summary statistics for the relative bias of the fixed effects can be seen in Table 4.2. This 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 much more variation compared to the other three fixed effects. These terms also have at least a few very large relative bias statistics shown by the large minimum and maximum values in Table 4.2. Therefore, on average the relative bias was kept under control for all of the fixed effects, but can become a problem for the slope terms (specifically β_1). This variation will be explored further through inferential modeling.

Table 4.2: 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 is variance, Med is median, Min is minimum, Max is maximum

Summary statistics for the absolute bias, (i.e. $\hat{\beta}_i - \beta_i$ where i represents the parameter of interest), of the fixed effects are shown in Table 4.3. This table provides further evidence that all of the terms were unbiased (average and median bias statistics near zero). Looking at the minimum and maximum absolute bias statistics in this table reveals that the scale of the paramter value artificially inflated the relative bias statistics, especially for the two slope terms β_1 and β_4. The range and variance of the absolute bias statistics for β_1 and β_4 were significantly

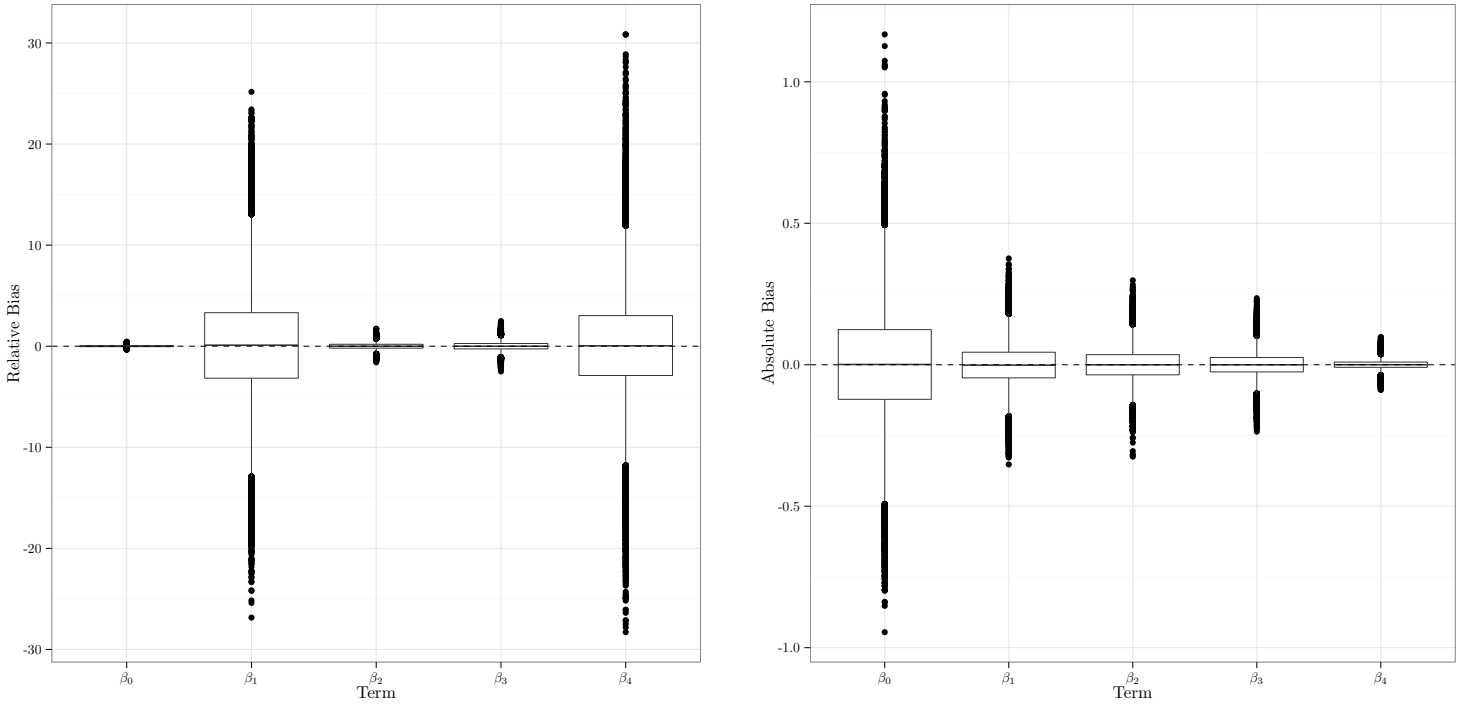
less compared to the relative bias statistics.

Table 4.3: Summary statistics for absolute bias of fixed effects

Term	Mean	Var	Med	Min	Max
β_0	0.0013	0.0376	0.0009	-0.9451	1.1676
β_1	-0.0008	0.0052	-0.0014	-0.3523	0.3758
β_2	-0.0002	0.0032	-0.0002	-0.3247	0.2983
β_3	-0.0002	0.0017	-0.0002	-0.2362	0.2351
β_4	0.0002	0.0002	0.0001	-0.0889	0.0970

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

To better explore the variability shown in Table 4.2, box plots of the fixed effects are shown in Figure 4.1, where Figure 4.1a depicts the relative bias statistics and Figure 4.1b shows the absolute bias. Looking at Figure 4.1a reveals the large amount of variability in β_1 and β_4. The interquartile range, depicted by the box in the box plot, was much larger for these terms indicating that the relative bias has a wider range of plausible values. In addition, β_1 has many potential outlying values shown by the dots where some relative bias statistics were greater than 100 in absolute value. However, Figure 4.1b shows the absolute bias has much less variability for these two terms suggesting that the scale of the parameter values influenced these large relative bias statistics. Looking back at Table 3.1 confirms that the parameter values for β_1 and β_4 were -0.014 and 0.0031 respectively. Dividing by such a small number could inflate relatively small absolute bias statistics when converting to relative bias. For example, an absolute bias statistics of 1.4 for β_1 would produce a relative bias statistic of -100 when dividing by the -0.014 parameter value.



(a) Relative bias

(b) Absolute bias

Figure 4.1: Box plots showing the bias of the fixed effects

The variation in the relative bias for the parameters were explored using ANOVA and effect sizes (i.e. η^2) were used instead of p-values due to the large sample size and strong statistical power. No four or five-way interactions had $\hat{\eta}^2$ greater than .001 and were dropped from all of 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 4.4 for all five fixed effect parameters and the variance of the random components (i.e. random effects and within cluster residuals). The values in bold in the table were $\hat{\eta}^2$ statistics that were larger than .001.

Looking at Table 4.4, there were no large $\hat{\eta}^2$ statistics for any of the five fixed effects, meaning that we cannot explain any variation in the relative bias for the fixed effects with the simulation conditions. 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 4.2. Exploring the averages shows that the 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 (.061 and .058 for β_1 and β_4 respectively), the relative bias statistic was quite small and would likely not seriously distort any findings.

Table 4.4: Eta-squared statistics for the relative bias models up to three-way interactions.

Variable	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$	η^2 Var b0	η^2 Var b1	η^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

Continued on next page

Variable	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$	η^2 b0	η^2 b1	η^2 Res
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 $> .001$, N is cluster sample size, p is within cluster sample size, Gen is generated, RE Dist is random effect distribution, SC is serial correlation, Fit is fitted, ":" represents interaction

Summary statistics for the relative bias of the random components can be seen in Table 4.5. 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 was not surprising that the minimum relative bias was small (approximately -1) compared to the maximum relative bias (approximately 10, 35, and 7 for variance of intercept, slope, and within cluster residuals respectively). For comparison, summary statistics for the absolute bias are shown in Table 4.6. The biggest difference between Table 4.5 and Table 4.6 was the reduced variation in the bias statistics and much smaller range of bias values. In addition, the absolute bias in the variance of the slope and residuals appear much smaller than the relative bias statistics, again likely due to the scale of the parameter values (e.g. parameter for variance of slope was 0.015).

Table 4.5: Summary statistics for relative bias of random components

Term	Mean	Var	Med	Min	Max
Var b0	0.4012	0.6942	0.2904	-1.0000	10.0186
Var b1	1.9116	9.2561	1.1211	-1.0000	35.4700
Var Res	0.1222	0.2645	-0.0151	-0.7943	6.6436

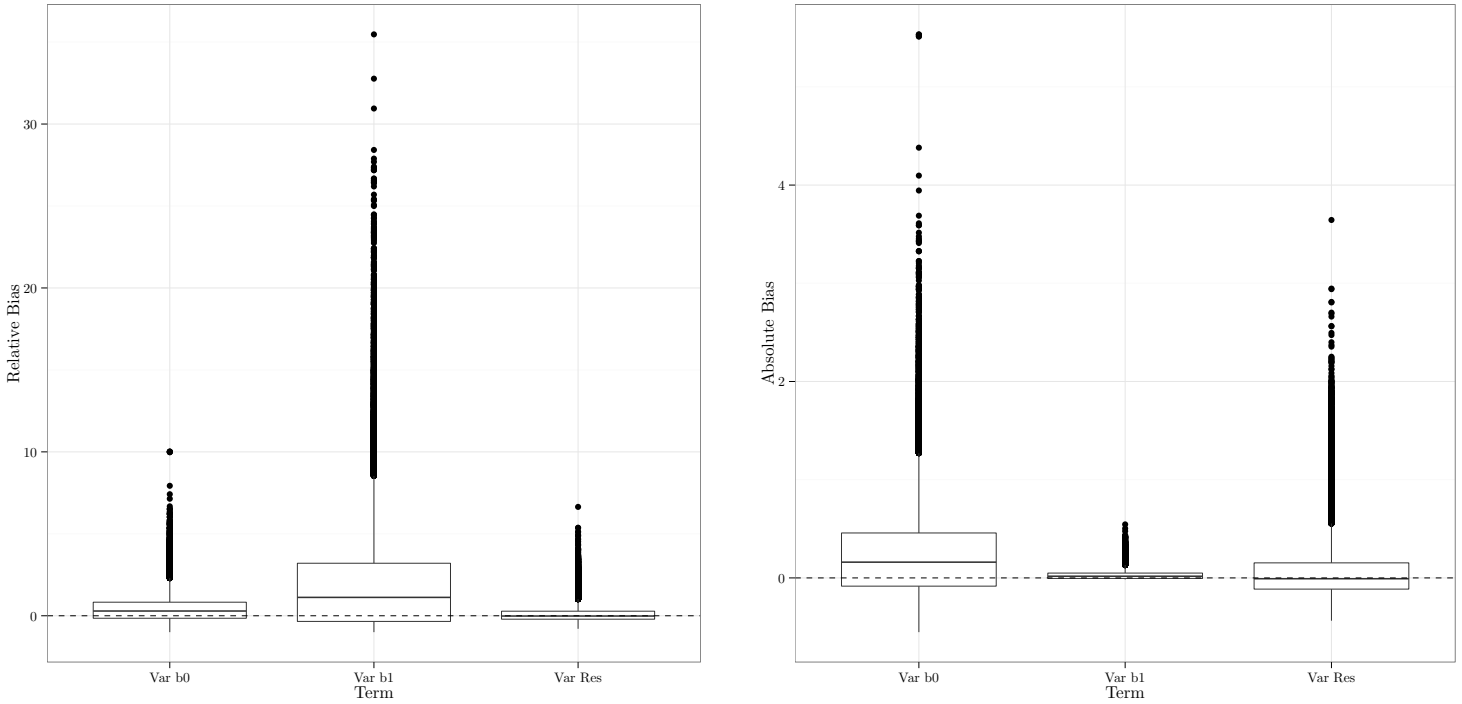
Note: Var is variance, Med is median,
Min is minimum, Max is maximum

Table 4.6: Summary statistics for absolute bias of random components

Term	Mean	Var	Med	Min	Max
Var b0	0.2216	0.2119	0.1604	-0.5524	5.5344
Var b1	0.0294	0.0022	0.0172	-0.0154	0.5449
Var Res	0.0670	0.0796	-0.0083	-0.4358	3.6445

Note: Var is variance, Med is median,
Min is miniumum, Max is maximum

Lastly, Figure 4.2 shows box plots of the relative and absolute bias for the random components. These plots further corroborate the observations made from the tables, especially that the variation in the absolute bias statistic was much smaller compared to the relative bias statistic, most notably for the variance of the slope and within cluster residuals. The variation was also smaller for the variance of the intercept, but the difference was not nearly as large (note the difference in scales between the two plots). In both of the figures, the box of the box plot contains the desired bias value of 0 suggesting this was a plausible value, but there was much more data above the zero line. The variation shown in the table and box plots will attempt to be explained through inferential modeling.



(a) Relative bias

(b) Absolute bias

Figure 4.2: Box plots showing the bias of the random components

The variation in the relative bias statistics for the random components of the model (i.e. random effects and residuals) were explored with an ANOVA and the $\hat{\eta}^2$ can be seen in the last three columns of Table 4.4. These columns reveal that there were variables that explain variation in the relative bias of the random components (i.e. $\hat{\eta}^2 > 0.001$). These effects include the cluster sample size, the within cluster sample size, the generated serial correlation structure, the fitted serial correlation structure, and the interaction between the generated and fitted serial correlation structure. No second or third order interactions had effect sizes greater than .001. An additional interaction between within cluster sample size and the generated serial correlation structure explained significant variation in the relative bias of the variance of the random slope for time.

The average relative bias for each random component by the generated serial correlation structure and fitted serial correlation structure can be seen in Table 4.7. Exploring the table shows that the relative bias was much greater for the variance of the slope compared to the other random components. This may be due to the small variance of the slope making it more difficult for the estimation algorithm to obtain an estimate for the variance of the random slopes. As discussed above, the small parameter value for the variance of the slope makes the denominator small for the relative bias statistic, which may artifcially inflate the relative bias statistic. In general, the random effects tend to be overestimated (positive relative bias statistics) whereas the residuals were commonly underestimated (negative relative bias statistics). This was not surprising as the overestimation of the random effects needs to be taken into account somewhere in the model, therefore the residuals were smaller than expected.

Table 4.7: Relative bias for the variance of the random effects

Gen SC	Fit SC	RB Var b0	RB Var b1	RB Res
Ind	Ind	0.090	0.625	-0.029
Ind	AR1	0.101	0.714	-0.038
Ind	MA1	0.130	0.771	-0.061
Ind	MA2	0.133	0.747	-0.065
Ind	ARMA	0.134	0.800	-0.068
AR1	Ind	0.657	2.664	-0.269
AR1	AR1	0.211	1.151	0.125
AR1	MA1	0.427	1.610	-0.084
AR1	MA2	0.347	1.370	-0.007
AR1	ARMA	0.251	1.233	0.085
MA1	Ind	0.536	2.837	-0.153
MA1	AR1	-0.170	0.489	0.484
MA1	MA1	0.138	1.048	0.184
MA1	MA2	0.205	1.197	0.124
MA1	ARMA	0.102	0.901	0.214
MA2	Ind	0.767	3.427	-0.261
MA2	AR1	0.131	1.369	0.303
MA2	MA1	0.519	2.231	-0.069
MA2	MA2	0.238	1.307	0.198
MA2	ARMA	0.099	1.252	0.330
ARMA	Ind	1.662	6.813	-0.218
ARMA	AR1	-0.304	0.951	1.535
ARMA	MA1	1.020	3.479	0.210
ARMA	MA2	0.675	2.330	1.000
ARMA	ARMA	0.355	1.644	0.858

Note: Gen is generated, SC is serial correlation, Fit is fitted,
RB is relative bias, p is within cluster sample size,
Var is variance, b0 is random intercept, b1 is random slope,
Res is residual

Looking at Figure 4.3 and Figure 4.4 shows 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 underly 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 was underspecified as independent, the variance of the intercept and slope were overspecified by over 1.5 times and at least 6 times respectively. The MA(1) fitted structure also has large relative bias statistics, but this structure was much improved compared to the independence structure, the variance of the intercept was about twice as large and the variance of the slope was about four times too large.

The AR(1) and ARMA(1,1) fitted structures tend to do the best to limit bias in the estimates of the variance of the random effects compared to the other structures. This may suggest that the moving average component does not aid in correcting for the presence of serial correlation in longitudinal data. Lastly, even when the correct structure was modeled there was still evidence of overestimation of the variance of the random effects and in many cases the correct structure fitted structure does not produce the smallest average relative bias statistics.

Lastly, Figure 4.5 shows that the variance of the residuals tend to be underestimated when an underspecified independence structure was fit, however this underestimation was not as large as the overspecification found in the random effects. The one exception to this was when the underlying structure was ARMA(1,1), which tends to produce an average relative bias statistics for the residuals that were comparable to the average relative bias for the variance of the intercept. Except for the systematic underestimation when an independence structure was

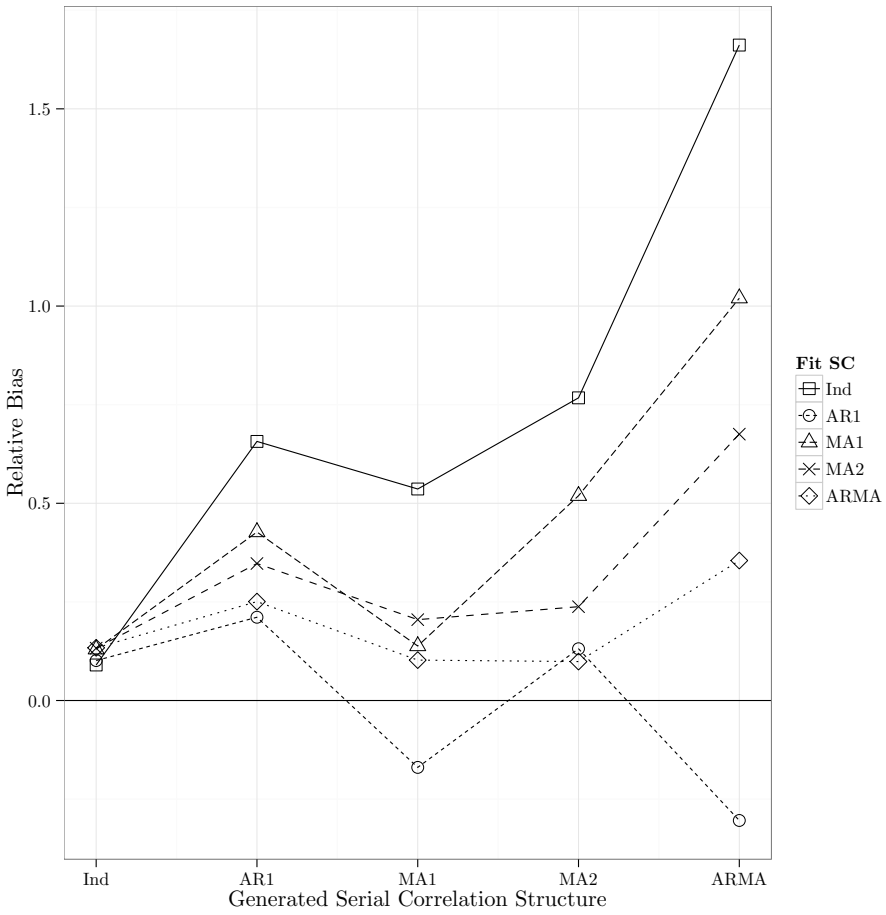


Figure 4.3: Relative bias of variance b0 by generated serial correlation structure and fitted correlation structure.

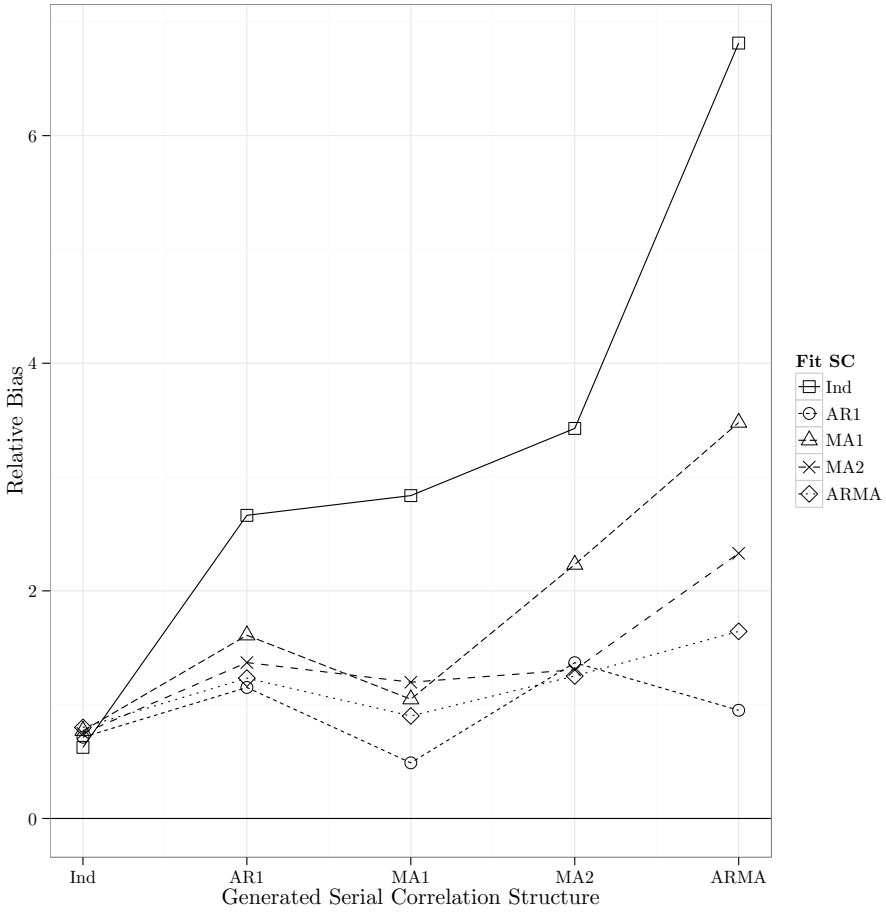


Figure 4.4: Relative bias of variance b1 by generated serial correlation structure and fitted correlation structure.

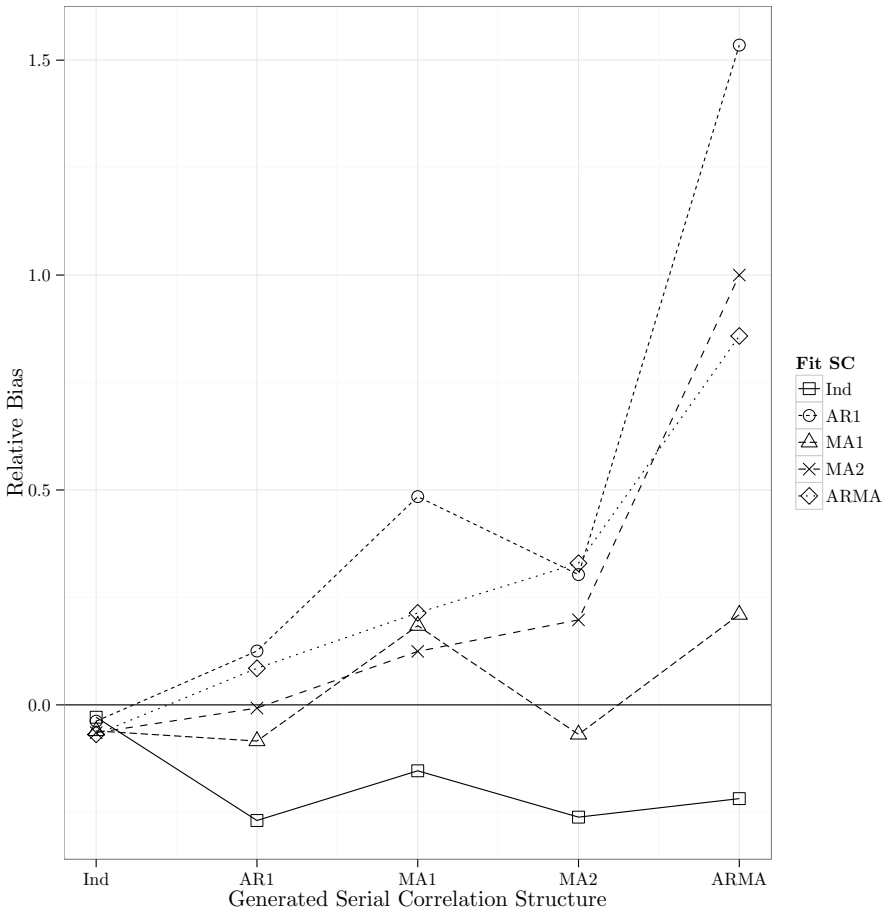


Figure 4.5: Relative bias of residuals by generated serial correlation structure and fitted correlation structure.

fitted when serial correlation was present, the average relative bias tended to be positive suggesting that all of the random components were overestimated when serial correlation was present and correctly modeled.

4.1.3 Type I Error Rate

Even though there was no evidence of bias in the fixed effects under any of the simulated data conditions, since the random components showed evidence of bias; the standard errors of the fixed effects may not be accurate causing 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). Summary statistics for the empirical type I error rate by the parameter can be seen in Table 4.8. From the table, one can see the mean and median were slightly above .05 with the largest being β_2 at .067. There also was variation in the empirical type I error rates shown by the range in values which range from about .02 to .1 for all the fixed effects.

Table 4.8: Summary statistics for empirical type I error rates

Term	Mean	Med	Min	Max
β_0	0.0565	0.0556	0.0269	0.0897
β_1	0.0657	0.0635	0.0308	0.1193
β_2	0.0670	0.0667	0.0230	0.1104
β_3	0.0579	0.0574	0.0149	0.0899
β_4	0.0661	0.0655	0.0211	0.1101

Note: Med is Median, Min is minimum, Max is maximum

Box plots can be seen in Figure 4.6 that 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 and β_3 have median type I error rates around 0.06, whereas β_1,

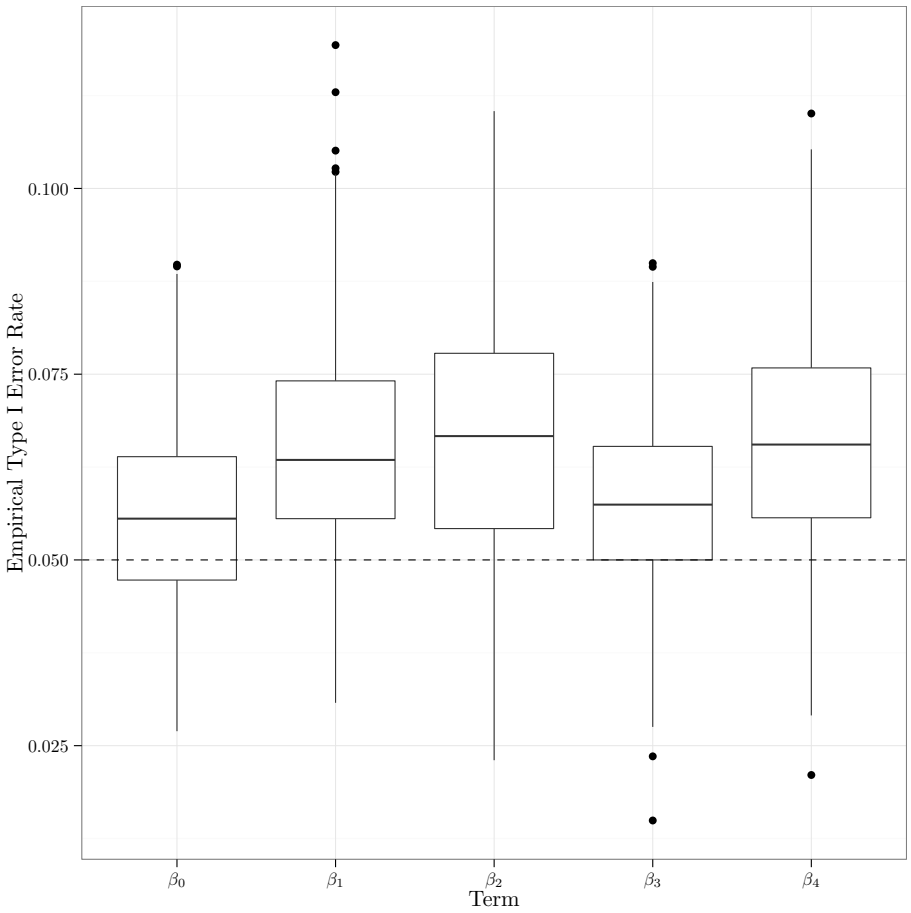


Figure 4.6: Box plot of type I error rates by parameter

β_2, and β_4 have 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 with an ANOVA.

The empirical type I error rates were modeled using an ANOVA similar to the relative bias statistics above, however in this model the average type I error rate for each simulation condition served as the dependent variable. Table 4.9 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.

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 these terms.

Table 4.9: Eta-squared statistics for the type I error rate models up to three-way interactions.

Variables	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$
Gen SC	**0.0416**	**0.0518**	**0.0338**	**0.0196**	**0.0857**
Fit SC	0.0086	**0.0145**	**0.1579**	0.0049	**0.0137**
RE Dist	**0.1133**	**0.0119**	**0.0617**	**0.0282**	**0.0286**
N	**0.0108**	**0.1111**	**0.1014**	**0.0150**	**0.0866**
p	0.0037	0.0005	**0.0152**	0.0000	0.0065
Gen SC:Fit SC	**0.0151**	**0.0412**	**0.0351**	**0.0180**	**0.0712**
Gen SC:RE Dist	**0.0339**	**0.0525**	**0.0354**	**0.0814**	**0.0820**
Gen SC:N	**0.0300**	0.0090	**0.0352**	**0.1305**	**0.0755**
Gen SC:p	**0.0468**	**0.0306**	0.0027	**0.0581**	**0.0356**
Fit SC:RE Dist	0.0060	0.0038	0.0043	**0.0117**	0.0035
Fit SC:N	0.0037	0.0030	0.0079	0.0017	0.0024
Fit SC:p	0.0030	0.0025	0.0034	**0.0131**	0.0088
RE Dist:N	**0.0160**	**0.0147**	**0.0631**	0.0072	0.0066
RE Dist:p	**0.0102**	**0.0188**	0.0096	0.0075	**0.0638**
N:p	**0.0476**	**0.0385**	**0.0240**	**0.0129**	0.0038
Gen SC:Fit SC:RE Dist	**0.0309**	**0.0205**	**0.0254**	**0.0355**	**0.0228**
Gen SC:Fit SC:N	**0.0128**	**0.0109**	**0.0103**	**0.0191**	**0.0111**

Continued on next page

Variables	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$
Gen SC:Fit SC:p	0.0067	**0.0193**	**0.0147**	**0.0254**	**0.0103**
Gen SC:RE Dist:N	**0.0397**	**0.0713**	**0.0523**	**0.0747**	**0.0380**
Gen SC:RE Dist:p	**0.1112**	**0.0989**	**0.0792**	**0.0969**	**0.0961**
Gen SC:N:p	**0.1475**	**0.0156**	**0.0601**	**0.0269**	**0.0338**
Fit SC:RE Dist:N	0.0070	0.0084	**0.0132**	**0.0188**	0.0084
Fit SC:RE Dist:p	0.0023	0.0038	**0.0152**	0.0099	**0.0107**
Fit SC:N:p	0.0010	0.0021	**0.0115**	0.0012	0.0005
RE Dist:N:p	**0.0196**	0.0051	0.0047	**0.0218**	**0.0338**

Note: Bold numbers are $> .01$, N is cluster sample size, Gen is generated, RE Dist is random effect distribution, p is within cluster sample size, SC is serial correlation, Fit is fitted, ":" represents an interaction

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 4.7. 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 MA(2) and ARMA(1,1) structures having the smallest amount of variation. Interestingly, there does not appear to be any discernable pattern in the average type I error rate by the differing random effect distributions.

Lastly, the scale of the y-axis should be taken into account. Although there was variability in the average type I error rates, this variability ranges from just over 0.04 to about 0.07 with an even smaller range when the cluster size was 50. Using the moderate criteria for robustness by Algina, Blair, and Coombs (1995) type I error rates outside the lower bound $.7\alpha = .035$ or the upper bound $1.3\alpha = .065$ were deemed to not have sufficient type I error control. A handful of the conditions fall outside the upper bound, most notably when the generated serial correlation

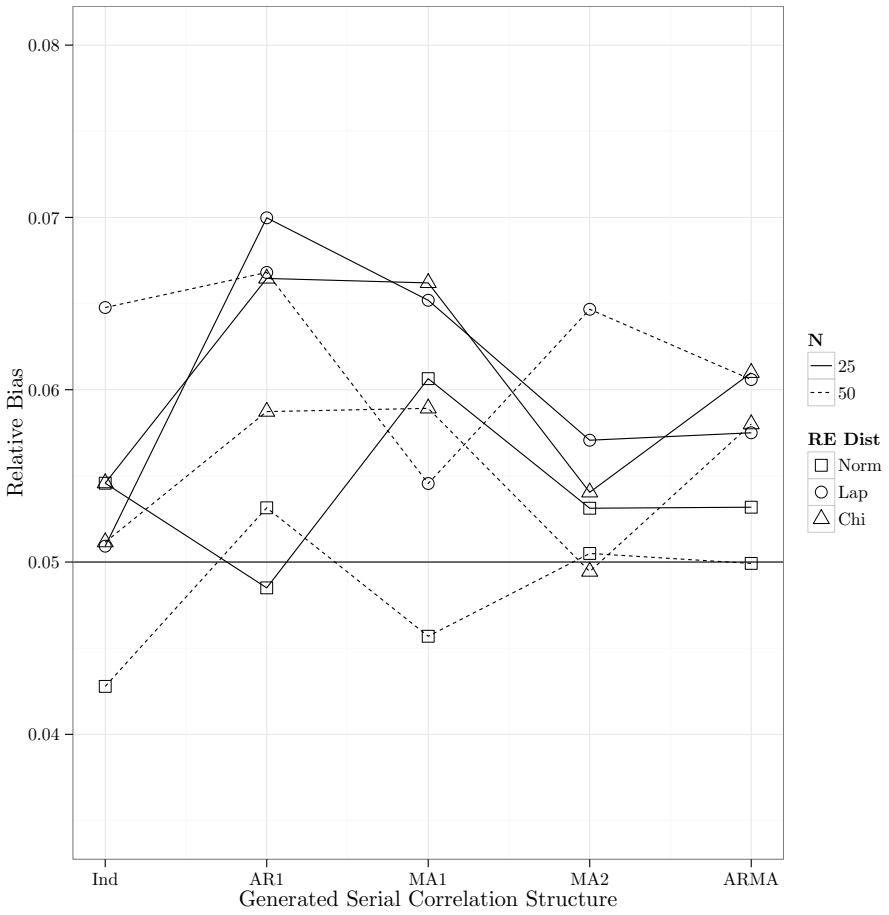


Figure 4.7: Mean type I error rate for β_0 by generated serial correlation structure, random effect disribution, and cluster sample size

structure was AR(1), cluster sizes of 25, and random effect distribution were chi-square(1) or Laplace.

Patterns for the empirical type I error rates were similar for the other parameters (i.e. $\beta_1, \ldots, \beta_4$) and were 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 were 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 4.7.

The empirical type I error rates for all simulation conditions can be seen in Table 4.10 through Table 4.14. There is a table for each of the five generated serial correlation structures, independent, AR(1), MA(1), MA(2), and ARMA(1,1), and were broken down by the other simulation conditions. As a result, the analysis conducted for the type I error rates could be replicated by pulling the data from this document.

Table 4.10: Type I error rates for independence generated serial correlation structure.

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	25	6	0.046	0.043	0.040	0.058	0.058
Ind	Norm	25	8	0.056	0.053	0.059	0.056	0.077
Ind	Norm	50	6	0.038	0.053	0.056	0.081	0.046
Ind	Norm	50	8	0.054	0.059	0.039	0.073	0.054
Ind	Lap	25	6	0.055	0.058	0.067	0.044	0.029
Ind	Lap	25	8	0.038	0.088	0.053	0.047	0.053
Ind	Lap	50	6	0.056	0.051	0.053	0.069	0.032
Ind	Lap	50	8	0.069	0.069	0.067	0.050	0.074
Ind	Chi	25	6	0.047	0.074	0.057	0.043	0.080
Ind	Chi	25	8	0.063	0.060	0.075	0.043	0.066
Ind	Chi	50	6	0.044	0.052	0.050	0.024	0.052
Ind	Chi	50	8	0.051	0.054	0.039	0.062	0.057
AR1	Norm	25	6	0.042	0.042	0.061	0.067	0.061
AR1	Norm	25	8	0.064	0.064	0.070	0.055	0.073
AR1	Norm	50	6	0.031	0.046	0.054	0.080	0.054
AR1	Norm	50	8	0.049	0.057	0.041	0.067	0.054
AR1	Lap	25	6	0.065	0.049	0.086	0.043	0.034
AR1	Lap	25	8	0.048	0.102	0.070	0.041	0.054
AR1	Lap	50	6	0.063	0.049	0.085	0.052	0.038
AR1	Lap	50	8	0.064	0.064	0.066	0.048	0.074

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
AR1	Chi	25	6	0.047	0.093	0.083	0.056	0.093
AR1	Chi	25	8	0.063	0.060	0.084	0.048	0.069
AR1	Chi	50	6	0.045	0.062	0.051	0.031	0.057
AR1	Chi	50	8	0.061	0.061	0.042	0.061	0.058
MA1	Norm	25	6	0.037	0.058	0.058	0.062	0.068
MA1	Norm	25	8	0.070	0.067	0.067	0.055	0.076
MA1	Norm	50	6	0.036	0.047	0.055	0.071	0.047
MA1	Norm	50	8	0.050	0.060	0.040	0.065	0.052
MA1	Lap	25	6	0.056	0.062	0.077	0.047	0.047
MA1	Lap	25	8	0.042	0.096	0.069	0.054	0.060
MA1	Lap	50	6	0.062	0.051	0.075	0.056	0.038
MA1	Lap	50	8	0.066	0.066	0.068	0.048	0.073
MA1	Chi	25	6	0.042	0.090	0.081	0.065	0.103
MA1	Chi	25	8	0.058	0.061	0.086	0.049	0.072
MA1	Chi	50	6	0.047	0.063	0.047	0.033	0.063
MA1	Chi	50	8	0.060	0.057	0.041	0.060	0.060
MA2	Norm	25	6	0.040	0.070	0.054	0.054	0.081
MA2	Norm	25	8	0.078	0.072	0.087	0.075	0.075
MA2	Norm	50	6	0.032	0.052	0.058	0.086	0.061
MA2	Norm	50	8	0.049	0.074	0.044	0.068	0.071
MA2	Lap	25	6	0.059	0.059	0.093	0.056	0.059
MA2	Lap	25	8	0.048	0.105	0.076	0.054	0.057
MA2	Lap	50	6	0.074	0.063	0.074	0.054	0.049
MA2	Lap	50	8	0.062	0.071	0.079	0.046	0.082
MA2	Chi	25	6	0.056	0.119	0.077	0.063	0.105
MA2	Chi	25	8	0.055	0.069	0.095	0.057	0.075
MA2	Chi	50	6	0.046	0.055	0.058	0.032	0.067
MA2	Chi	50	8	0.056	0.056	0.053	0.044	0.064
ARMA	Norm	25	6	0.038	0.057	0.057	0.054	0.080
ARMA	Norm	25	8	0.071	0.062	0.062	0.068	0.078
ARMA	Norm	50	6	0.034	0.050	0.056	0.081	0.062
ARMA	Norm	50	8	0.051	0.057	0.051	0.059	0.071
ARMA	Lap	25	6	0.060	0.050	0.098	0.047	0.060
ARMA	Lap	25	8	0.038	0.102	0.073	0.061	0.054
ARMA	Lap	50	6	0.061	0.055	0.078	0.055	0.041

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Lap	50	8	0.070	0.065	0.073	0.053	0.067
ARMA	Chi	25	6	0.046	0.099	0.085	0.049	0.095
ARMA	Chi	25	8	0.066	0.063	0.084	0.057	0.075
ARMA	Chi	50	6	0.044	0.062	0.062	0.029	0.083
ARMA	Chi	50	8	0.056	0.062	0.056	0.047	0.062

Note: Fit SC is fitted serial correlation, RE Dist is random effect distribution, N is cluster sample size, p is within cluster sample size, T1E is type I error, Norm is normal, Lap is Laplace, Chi, is chi-square(1)

Table 4.11: Type I error rates for AR (1) generated serial correlation structure.

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	25	6	0.052	0.090	0.068	0.061	0.056
Ind	Norm	25	8	0.055	0.070	0.037	0.076	0.090
Ind	Norm	50	6	0.046	0.071	0.061	0.055	0.069
Ind	Norm	50	8	0.061	0.073	0.043	0.045	0.073
Ind	Lap	25	6	0.068	0.066	0.053	0.070	0.090
Ind	Lap	25	8	0.072	0.068	0.050	0.034	0.070
Ind	Lap	50	6	0.074	0.084	0.062	0.043	0.051
Ind	Lap	50	8	0.069	0.071	0.059	0.061	0.063
Ind	Chi	25	6	0.061	0.084	0.050	0.050	0.093
Ind	Chi	25	8	0.088	0.082	0.071	0.066	0.071
Ind	Chi	50	6	0.059	0.063	0.051	0.059	0.090
Ind	Chi	50	8	0.057	0.065	0.063	0.059	0.065
AR1	Norm	25	6	0.052	0.091	0.049	0.062	0.039
AR1	Norm	25	8	0.033	0.057	0.051	0.082	0.085
AR1	Norm	50	6	0.038	0.055	0.064	0.047	0.050
AR1	Norm	50	8	0.054	0.076	0.051	0.056	0.065
AR1	Lap	25	6	0.054	0.061	0.067	0.083	0.074
AR1	Lap	25	8	0.090	0.061	0.054	0.035	0.054
AR1	Lap	50	6	0.064	0.076	0.067	0.055	0.070
AR1	Lap	50	8	0.055	0.043	0.065	0.055	0.055
AR1	Chi	25	6	0.055	0.071	0.055	0.049	0.087

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
AR1	Chi	25	8	0.066	0.082	0.082	0.057	0.066
AR1	Chi	50	6	0.047	0.050	0.038	0.047	0.088
AR1	Chi	50	8	0.057	0.051	0.036	0.063	0.054
MA1	Norm	25	6	0.054	0.095	0.079	0.062	0.038
MA1	Norm	25	8	0.042	0.063	0.058	0.090	0.077
MA1	Norm	50	6	0.048	0.055	0.065	0.048	0.053
MA1	Norm	50	8	0.061	0.083	0.046	0.050	0.066
MA1	Lap	25	6	0.051	0.069	0.062	0.077	0.077
MA1	Lap	25	8	0.083	0.061	0.067	0.037	0.053
MA1	Lap	50	6	0.068	0.075	0.066	0.045	0.057
MA1	Lap	50	8	0.061	0.063	0.072	0.048	0.052
MA1	Chi	25	6	0.043	0.062	0.076	0.046	0.084
MA1	Chi	25	8	0.082	0.084	0.082	0.066	0.069
MA1	Chi	50	6	0.053	0.062	0.045	0.055	0.084
MA1	Chi	50	8	0.055	0.041	0.050	0.062	0.048
MA2	Norm	25	6	0.060	0.103	0.091	0.057	0.045
MA2	Norm	25	8	0.040	0.072	0.063	0.086	0.077
MA2	Norm	50	6	0.051	0.053	0.072	0.048	0.067
MA2	Norm	50	8	0.062	0.071	0.054	0.054	0.059
MA2	Lap	25	6	0.062	0.071	0.062	0.084	0.071
MA2	Lap	25	8	0.083	0.059	0.071	0.045	0.050
MA2	Lap	50	6	0.080	0.086	0.072	0.033	0.061
MA2	Lap	50	8	0.060	0.063	0.063	0.052	0.050
MA2	Chi	25	6	0.058	0.058	0.077	0.064	0.070
MA2	Chi	25	8	0.077	0.098	0.098	0.068	0.071
MA2	Chi	50	6	0.073	0.062	0.056	0.056	0.076
MA2	Chi	50	8	0.061	0.059	0.051	0.069	0.059
ARMA	Norm	25	6	0.050	0.113	0.080	0.053	0.047
ARMA	Norm	25	8	0.043	0.060	0.060	0.080	0.080
ARMA	Norm	50	6	0.052	0.055	0.075	0.052	0.071
ARMA	Norm	50	8	0.055	0.086	0.052	0.043	0.061
ARMA	Lap	25	6	0.055	0.074	0.077	0.084	0.074
ARMA	Lap	25	8	0.083	0.055	0.073	0.035	0.055
ARMA	Lap	50	6	0.068	0.081	0.078	0.059	0.065
ARMA	Lap	50	8	0.064	0.067	0.064	0.054	0.061

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Chi	25	6	0.049	0.077	0.077	0.038	0.084
ARMA	Chi	25	8	0.075	0.099	0.086	0.062	0.068
ARMA	Chi	50	6	0.069	0.059	0.048	0.059	0.083
ARMA	Chi	50	8	0.059	0.056	0.046	0.065	0.049

Note: Fit SC is fitted serial correlation, RE Dist is random effect distribution, N is cluster sample size, p is within cluster sample size, T1E is type I error, Norm is normal, Lap is Laplace, Chi, is chi-square(1)

Table 4.12: Type I error rates for MA (1) generated serial correlation structure

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	25	6	0.085	0.069	0.046	0.071	0.081
Ind	Norm	25	8	0.046	0.065	0.039	0.065	0.084
Ind	Norm	50	6	0.049	0.076	0.074	0.053	0.074
Ind	Norm	50	8	0.051	0.060	0.045	0.055	0.076
Ind	Lap	25	6	0.060	0.095	0.074	0.047	0.072
Ind	Lap	25	8	0.065	0.088	0.083	0.065	0.059
Ind	Lap	50	6	0.048	0.058	0.050	0.058	0.062
Ind	Lap	50	8	0.067	0.055	0.065	0.055	0.055
Ind	Chi	25	6	0.082	0.089	0.080	0.068	0.068
Ind	Chi	25	8	0.059	0.062	0.057	0.041	0.062
Ind	Chi	50	6	0.045	0.053	0.068	0.049	0.086
Ind	Chi	50	8	0.072	0.063	0.063	0.057	0.067
AR1	Norm	25	6	0.079	0.047	0.039	0.079	0.071
AR1	Norm	25	8	0.037	0.057	0.054	0.054	0.047
AR1	Norm	50	6	0.045	0.056	0.059	0.052	0.056
AR1	Norm	50	8	0.036	0.043	0.069	0.040	0.058
AR1	Lap	25	6	0.074	0.082	0.086	0.058	0.074
AR1	Lap	25	8	0.056	0.078	0.063	0.041	0.033
AR1	Lap	50	6	0.031	0.057	0.065	0.061	0.061
AR1	Lap	50	8	0.073	0.049	0.059	0.045	0.056
AR1	Chi	25	6	0.066	0.089	0.070	0.085	0.047
AR1	Chi	25	8	0.053	0.039	0.042	0.046	0.021

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
AR1	Chi	50	6	0.038	0.056	0.052	0.045	0.066
AR1	Chi	50	8	0.083	0.052	0.049	0.056	0.056
MA1	Norm	25	6	0.073	0.051	0.063	0.057	0.051
MA1	Norm	25	8	0.054	0.057	0.060	0.060	0.066
MA1	Norm	50	6	0.040	0.051	0.073	0.045	0.054
MA1	Norm	50	8	0.042	0.045	0.089	0.052	0.076
MA1	Lap	25	6	0.071	0.071	0.089	0.058	0.068
MA1	Lap	25	8	0.047	0.070	0.079	0.065	0.041
MA1	Lap	50	6	0.035	0.052	0.072	0.061	0.064
MA1	Lap	50	8	0.077	0.066	0.082	0.055	0.061
MA1	Chi	25	6	0.090	0.074	0.102	0.065	0.056
MA1	Chi	25	8	0.055	0.043	0.061	0.028	0.040
MA1	Chi	50	6	0.044	0.044	0.072	0.058	0.069
MA1	Chi	50	8	0.083	0.064	0.075	0.053	0.047
MA2	Norm	25	6	0.069	0.076	0.066	0.080	0.080
MA2	Norm	25	8	0.051	0.057	0.081	0.045	0.054
MA2	Norm	50	6	0.052	0.063	0.061	0.061	0.058
MA2	Norm	50	8	0.037	0.042	0.087	0.051	0.073
MA2	Lap	25	6	0.078	0.085	0.098	0.064	0.081
MA2	Lap	25	8	0.059	0.081	0.100	0.078	0.050
MA2	Lap	50	6	0.031	0.074	0.083	0.061	0.067
MA2	Lap	50	8	0.067	0.057	0.105	0.054	0.059
MA2	Chi	25	6	0.069	0.086	0.109	0.069	0.059
MA2	Chi	25	8	0.050	0.033	0.066	0.040	0.053
MA2	Chi	50	6	0.039	0.058	0.055	0.061	0.069
MA2	Chi	50	8	0.077	0.066	0.068	0.060	0.040
ARMA	Norm	25	6	0.063	0.053	0.067	0.063	0.067
ARMA	Norm	25	8	0.049	0.061	0.067	0.058	0.064
ARMA	Norm	50	6	0.055	0.058	0.071	0.052	0.052
ARMA	Norm	50	8	0.045	0.051	0.090	0.051	0.078
ARMA	Lap	25	6	0.084	0.071	0.096	0.061	0.074
ARMA	Lap	25	8	0.062	0.082	0.082	0.069	0.036
ARMA	Lap	50	6	0.032	0.063	0.079	0.060	0.069
ARMA	Lap	50	8	0.073	0.057	0.096	0.051	0.062
ARMA	Chi	25	6	0.087	0.087	0.110	0.074	0.064

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Chi	25	8	0.046	0.040	0.066	0.043	0.046
ARMA	Chi	50	6	0.027	0.042	0.051	0.057	0.066
ARMA	Chi	50	8	0.082	0.058	0.076	0.064	0.052

Note: Fit SC is fitted serial correlation, RE Dist is random effect distribution, N is cluster sample size, p is within cluster sample size, T1E is type I error, Norm is normal, Lap is Laplace, Chi, is chi-square(1)

Table 4.13: Type I error rates for MA (2) generated serial correlation structure

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	25	6	0.065	0.089	0.073	0.045	0.080
Ind	Norm	25	8	0.050	0.069	0.078	0.058	0.080
Ind	Norm	50	6	0.047	0.069	0.051	0.061	0.088
Ind	Norm	50	8	0.056	0.056	0.046	0.074	0.062
Ind	Lap	25	6	0.087	0.081	0.065	0.063	0.076
Ind	Lap	25	8	0.050	0.065	0.071	0.058	0.086
Ind	Lap	50	6	0.062	0.052	0.062	0.050	0.060
Ind	Lap	50	8	0.067	0.059	0.063	0.077	0.077
Ind	Chi	25	6	0.053	0.066	0.077	0.057	0.075
Ind	Chi	25	8	0.063	0.069	0.054	0.048	0.072
Ind	Chi	50	6	0.046	0.060	0.040	0.054	0.064
Ind	Chi	50	8	0.040	0.067	0.046	0.073	0.073
AR1	Norm	25	6	0.061	0.064	0.074	0.054	0.087
AR1	Norm	25	8	0.048	0.061	0.082	0.071	0.085
AR1	Norm	50	6	0.046	0.065	0.046	0.065	0.092
AR1	Norm	50	8	0.059	0.050	0.065	0.065	0.062
AR1	Lap	25	6	0.069	0.090	0.069	0.045	0.083
AR1	Lap	25	8	0.033	0.069	0.042	0.046	0.075
AR1	Lap	50	6	0.064	0.038	0.070	0.054	0.058
AR1	Lap	50	8	0.064	0.067	0.061	0.089	0.064
AR1	Chi	25	6	0.055	0.059	0.107	0.059	0.085
AR1	Chi	25	8	0.040	0.050	0.047	0.050	0.060
AR1	Chi	50	6	0.054	0.042	0.048	0.060	0.054

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
AR1	Chi	50	8	0.049	0.071	0.040	0.074	0.071
MA1	Norm	25	6	0.057	0.062	0.080	0.050	0.087
MA1	Norm	25	8	0.047	0.065	0.087	0.062	0.094
MA1	Norm	50	6	0.049	0.068	0.042	0.065	0.072
MA1	Norm	50	8	0.051	0.055	0.048	0.075	0.068
MA1	Lap	25	6	0.068	0.075	0.075	0.060	0.078
MA1	Lap	25	8	0.038	0.067	0.059	0.056	0.067
MA1	Lap	50	6	0.066	0.048	0.073	0.059	0.064
MA1	Lap	50	8	0.075	0.066	0.049	0.071	0.073
MA1	Chi	25	6	0.047	0.057	0.098	0.054	0.070
MA1	Chi	25	8	0.065	0.057	0.065	0.044	0.060
MA1	Chi	50	6	0.051	0.062	0.053	0.055	0.053
MA1	Chi	50	8	0.050	0.061	0.057	0.070	0.066
MA2	Norm	25	6	0.055	0.076	0.085	0.043	0.073
MA2	Norm	25	8	0.042	0.072	0.093	0.063	0.090
MA2	Norm	50	6	0.042	0.051	0.051	0.071	0.082
MA2	Norm	50	8	0.056	0.051	0.072	0.075	0.059
MA2	Lap	25	6	0.079	0.085	0.091	0.058	0.076
MA2	Lap	25	8	0.047	0.059	0.080	0.062	0.071
MA2	Lap	50	6	0.055	0.040	0.101	0.049	0.063
MA2	Lap	50	8	0.062	0.067	0.073	0.086	0.073
MA2	Chi	25	6	0.061	0.061	0.098	0.055	0.110
MA2	Chi	25	8	0.051	0.057	0.081	0.045	0.063
MA2	Chi	50	6	0.059	0.053	0.053	0.056	0.065
MA2	Chi	50	8	0.043	0.068	0.060	0.068	0.062
ARMA	Norm	25	6	0.058	0.065	0.076	0.061	0.104
ARMA	Norm	25	8	0.046	0.061	0.095	0.065	0.087
ARMA	Norm	50	6	0.041	0.049	0.049	0.064	0.094
ARMA	Norm	50	8	0.054	0.058	0.065	0.075	0.065
ARMA	Lap	25	6	0.054	0.072	0.086	0.058	0.086
ARMA	Lap	25	8	0.033	0.071	0.059	0.056	0.082
ARMA	Lap	50	6	0.070	0.048	0.085	0.040	0.048
ARMA	Lap	50	8	0.059	0.066	0.066	0.087	0.077
ARMA	Chi	25	6	0.052	0.063	0.101	0.052	0.093
ARMA	Chi	25	8	0.047	0.058	0.069	0.047	0.058

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Chi	50	6	0.058	0.031	0.046	0.058	0.065
ARMA	Chi	50	8	0.053	0.070	0.049	0.070	0.070

Note: Fit SC is fitted serial correlation, RE Dist is random effect distribution, N is cluster sample size, p is within cluster sample size, T1E is type I error, Norm is normal, Lap is Laplace, Chi, is chi-square(1)

Table 4.14: Type I error rates for ARMA (1,1) generated serial correlation structure

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	25	6	0.054	0.087	0.076	0.072	0.076
Ind	Norm	25	8	0.057	0.051	0.071	0.079	0.063
Ind	Norm	50	6	0.048	0.064	0.056	0.044	0.064
Ind	Norm	50	8	0.058	0.088	0.062	0.052	0.050
Ind	Lap	25	6	0.061	0.079	0.069	0.071	0.075
Ind	Lap	25	8	0.054	0.078	0.074	0.085	0.093
Ind	Lap	50	6	0.054	0.040	0.064	0.058	0.048
Ind	Lap	50	8	0.060	0.084	0.064	0.054	0.058
Ind	Chi	25	6	0.062	0.075	0.060	0.066	0.079
Ind	Chi	25	8	0.054	0.094	0.050	0.060	0.081
Ind	Chi	50	6	0.062	0.064	0.074	0.070	0.056
Ind	Chi	50	8	0.054	0.064	0.042	0.054	0.070
AR1	Norm	25	6	0.028	0.097	0.078	0.074	0.051
AR1	Norm	25	8	0.036	0.059	0.036	0.050	0.100
AR1	Norm	50	6	0.040	0.065	0.040	0.050	0.080
AR1	Norm	50	8	0.050	0.078	0.078	0.032	0.050
AR1	Lap	25	6	0.077	0.081	0.091	0.057	0.091
AR1	Lap	25	8	0.047	0.047	0.037	0.051	0.061
AR1	Lap	50	6	0.055	0.046	0.037	0.068	0.064
AR1	Lap	50	8	0.060	0.080	0.085	0.015	0.070
AR1	Chi	25	6	0.062	0.062	0.071	0.071	0.090
AR1	Chi	25	8	0.069	0.092	0.023	0.069	0.083
AR1	Chi	50	6	0.083	0.099	0.050	0.066	0.061
AR1	Chi	50	8	0.040	0.044	0.049	0.044	0.058

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
MA1	Norm	25	6	0.052	0.069	0.089	0.074	0.072
MA1	Norm	25	8	0.063	0.038	0.061	0.073	0.082
MA1	Norm	50	6	0.049	0.064	0.051	0.041	0.051
MA1	Norm	50	8	0.062	0.088	0.066	0.058	0.056
MA1	Lap	25	6	0.076	0.076	0.071	0.066	0.058
MA1	Lap	25	8	0.051	0.065	0.072	0.076	0.076
MA1	Lap	50	6	0.056	0.054	0.064	0.051	0.056
MA1	Lap	50	8	0.057	0.088	0.095	0.053	0.059
MA1	Chi	25	6	0.062	0.072	0.080	0.057	0.075
MA1	Chi	25	8	0.058	0.090	0.041	0.061	0.075
MA1	Chi	50	6	0.063	0.066	0.057	0.059	0.049
MA1	Chi	50	8	0.051	0.053	0.038	0.053	0.053
MA2	Norm	25	6	0.057	0.078	0.093	0.054	0.078
MA2	Norm	25	8	0.068	0.045	0.077	0.057	0.080
MA2	Norm	50	6	0.040	0.066	0.063	0.032	0.047
MA2	Norm	50	8	0.057	0.074	0.079	0.054	0.059
MA2	Lap	25	6	0.066	0.075	0.097	0.069	0.060
MA2	Lap	25	8	0.039	0.064	0.069	0.069	0.067
MA2	Lap	50	6	0.058	0.050	0.077	0.055	0.052
MA2	Lap	50	8	0.064	0.086	0.100	0.050	0.060
MA2	Chi	25	6	0.060	0.063	0.110	0.066	0.063
MA2	Chi	25	8	0.062	0.071	0.076	0.059	0.082
MA2	Chi	50	6	0.061	0.050	0.081	0.069	0.056
MA2	Chi	50	8	0.055	0.050	0.060	0.046	0.031
ARMA	Norm	25	6	0.040	0.077	0.077	0.060	0.070
ARMA	Norm	25	8	0.055	0.055	0.071	0.058	0.094
ARMA	Norm	50	6	0.042	0.069	0.060	0.042	0.039
ARMA	Norm	50	8	0.042	0.074	0.086	0.065	0.059
ARMA	Lap	25	6	0.066	0.093	0.086	0.080	0.060
ARMA	Lap	25	8	0.044	0.066	0.069	0.069	0.075
ARMA	Lap	50	6	0.063	0.044	0.067	0.060	0.054
ARMA	Lap	50	8	0.083	0.092	0.101	0.035	0.047
ARMA	Chi	25	6	0.058	0.065	0.076	0.058	0.082
ARMA	Chi	25	8	0.068	0.081	0.068	0.061	0.074
ARMA	Chi	50	6	0.068	0.071	0.059	0.062	0.056

Continued on next page

Fit SC	RE Dist	N	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Chi	50	8	0.050	0.041	0.067	0.047	0.035

Note: Fit SC is fitted serial correlation, RE Dist is random effect distribution, N is cluster sample size, p is within cluster sample size, T1E is type I error, Norm is normal, Lap is Laplace, Chi, is chi-square(1)

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 (which violates the assumption of homogeneity of variance). This transformation took the following form:

$$\hat{p'}_k = \begin{cases} 2 * \sqrt{sin^{-1}\hat{p}_k} & \text{for } 0 < \hat{p'}_k < 1 \\ 2 * \sqrt{sin^{-1}(\frac{1}{4}R_k)} & \text{for } \hat{p'}_k = 0 \\ 3.14 - 2 * \sqrt{sin^{-1}(\frac{1}{4}R_k)} & \text{for } \hat{p'}_k = 1 \end{cases} \tag{4.1}$$

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, the same 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 model were very similar to the model

left in the original proportion metric with the same variables identified as having a significant effect size. Since the results were similar, interpretations that were made above in the original proportion metric were the same regardless of the scale of measurement which aids in the robustness of results. As a result, interpretations were kept in the proportion metric from above.

4.2 Study Two

4.2.1 Convergence

Convergence rates for study two were very good and can be seen in the third column of Table 4.1. Almost all of the models converged, the only exception to this rule was when an overspecified ARMA(1,1) structure and a MA(2) structure were fit. For example, the smallest convergence rate was 93% when an ARMA(1,1) structure was fit to an independence structure. The better convergence rate in this study was due to the missing random effect for time that was not modeled. Estimation of random effects was much more difficult than estimating other parameters and the parameter value for the variance of the slope was close to zero (0.015) which would make estimation more difficult. Just as in study one, only the models that converged were included in the results below.

4.2.2 Relative Bias

Summary statistics for the relative bias of the fixed effects can be seen in Table 4.15 and were similar to those in study one. On average the fixed effects have small relative bias statistics, however the variation and range of the slope terms (β_1 and

β_4) were large compared to the other terms.

Summary statistics for the absolute bias statistics for the fixed effects can be seen in Table 4.19. Similar to study one, the absolute bias statistics still have small average bias statistics, but with much smaller variation. This again suggests that the small parameter values for the slope terms may be artificially inflating the relative bias statistics.

Table 4.15: Summary statistics for relative bias of fixed effects

Term	Mean	Var	Med	Min	Max
β_0	-0.0005	0.0073	-0.0008	-0.3765	0.4710
β_1	0.0057	36.6005	0.0391	-27.1576	24.1300
β_2	0.0015	0.1260	0.0019	-1.6969	1.8122
β_3	-0.0038	0.2599	-0.0060	-2.5423	2.4612
β_4	0.1144	35.0059	0.1065	-34.2909	37.4144

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

Box plots showing the distribution of possible values for the relative and absolute bias of the fixed effects can be seen in Figure 4.8. These plots further show the reduced variation in the absolute bias compared to the relative bias statistics. The plots also were all centered closely around the dashed line of 0 indicating no bias. This suggests that the terms have negligible amounts of bias.

Table 4.16: Summary statistics for absolute bias of fixed effects

Term	Mean	Var	Med	Min	Max
β_0	-0.0013	0.0512	-0.0020	-0.9936	1.2432
β_1	-0.0001	0.0072	-0.0005	-0.3378	0.3801

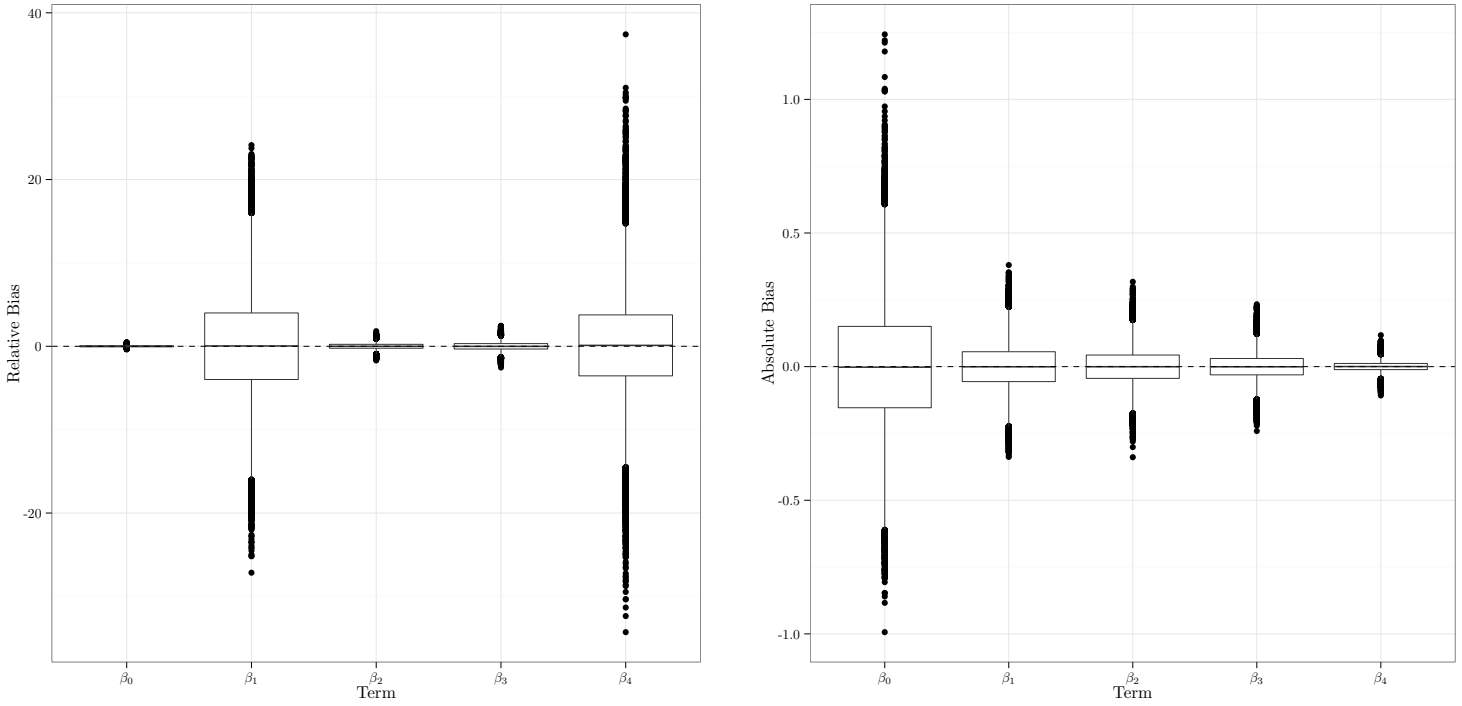
continued on next page

Term	Mean	Var	Med	Min	Max
β_2	-0.0003	0.0044	-0.0003	-0.3390	0.3174
β_3	-0.0004	0.0023	-0.0006	-0.2410	0.2333
β_4	0.0004	0.0003	0.0003	-0.1078	0.1176

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

The relative bias for the fixed and random components were explored using ANOVA. All interactions were explored initially, but no third or fourth order interactions had an $\hat{\eta}^2 > 0.001$ and were dropped from the final model. The results for all the ANOVA models for the five fixed effects and two random components are in Table 4.17. $\hat{\eta}^2$ statistics that are in bold were larger than .001 and deemed to explain a significant proportion of variation in the relative bias statistics.

Exploring Table 4.17 reveals that none of the $\hat{\eta}^2$ for the fixed effects were greater than 0.001, suggesting that on average the simulation conditions can not explain variation in the relative bias of the fixed effects. The means found in Table 4.15 can adequately summarize the average relative bias for each parameter. Just as in the first study, the terms for the slope (i.e. β_1 and β_4) tend to have larger relative bias statistics compared to the other terms (0.006 and 0.114 respectively). The average relative bias statistic for β_4 was larger, however this was likely due to the small parameter value. The average absolute bias statistic was $< .001$ providing evidence that the term was unbiased.



(a) Relative bias

(b) Absolute bias

Figure 4.8: Box plots showing the bias of the random components

Table 4.17: Eta-squared statistics for the relative bias models up to three-way interactions.

Variable	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$	η^2 Var b0	η^2 Var Res
p	0.0002	0.0001	0.0001	0.0000	0.0000	0.0000	0.0008
RE Dist	0.0000	0.0001	0.0001	0.0000	0.0001	0.0002	0.0000
Gen SC	0.0006	0.0012	0.0007	0.0006	0.0002	**0.0172**	**0.1861**
Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	**0.1388**	**0.2097**
p:RE Dist	0.0002	0.0000	0.0004	0.0003	0.0001	0.0001	0.0001
p:Gen SC	0.0005	0.0003	0.0003	0.0004	0.0001	0.0002	0.0002
p:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
RE Dist:Gen SC	0.0003	0.0008	0.0003	0.0005	0.0007	0.0004	0.0002
RE Dist:Fit SC	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	**0.0592**	**0.0922**

Note: Bold numbers are $> .001$, N is cluster sample size, Gen is generated, Fit is fitted, RE Dist is random effect distribution, p is within cluster sample size, SC is serial correlation, ":" represents interaction

Summary statistics for the relative bias and absolute bias for the random components are presented in Table 4.18 and Table 4.19 respectively. These tables again show a similar trend as the previous tables, there was more variation in the relative bias metric compared to the absolute bias metric. The variance components tended to be overestimated which was not surprising as variances have a lower bound of 0 and the parameters were small, therefore there was a constraint on how small the algorithm will estimate the variance.

Table 4.17 also shows the relative bias for the random components in the last two columns. These two columns reveal that the generated and fitted serial correlation structures were significant along with the interaction between the two when the relative bias of the random components were modeled.

Table 4.18: Summary statistics for relative bias of random components

Term	Mean	Var	Med	Min	Max
Var b0	0.0532	0.6651	-0.0314	-1.0000	14.3336
Var Res	0.4672	0.4616	0.2703	-0.6215	7.0030

Note: Var is variance, Med is median,
Min is miniumum, Max is maximum

Table 4.19: Summary statistics for absolute bias of random components

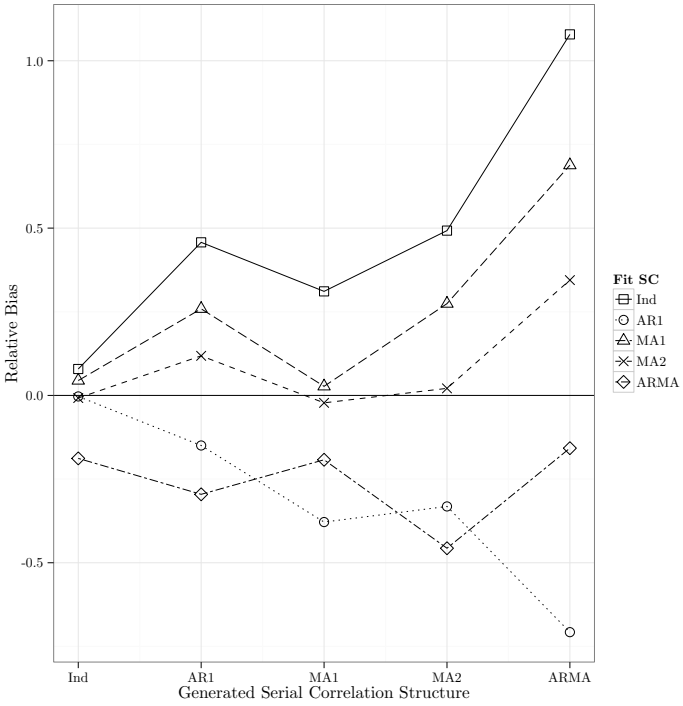
Term	Mean	Var	Med	Min	Max
Var b0	0.0294	0.2030	-0.0173	-0.5524	7.9180
Var Res	0.2563	0.1389	0.1483	-0.3409	3.8417

Note: Var is variance, Med is median,
Min is miniumum, Max is maximum

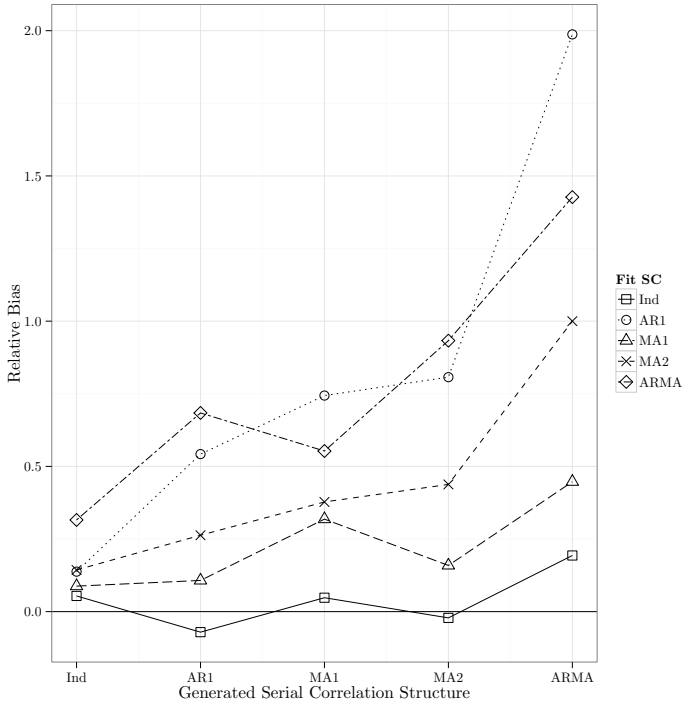
The average relative bias for the variance of the intercepts and the variance of the residuals by the generated and fitted serial correlation structures can be seen in Figure 4.9. Figure 4.9a shows there was less variability and bias when the generated serial correlation structure was either independent or MA(1) and the largest variability and bias when the generated structure was ARMA(1,1). When the fitted serial correlation structure was AR(1) or ARMA(1,1) the variance of the intercepts tends to be underestimated. In contrast, when the fitted structures were independent or MA(1) the variance of the intercepts tends to be overestimated. The MA(2) fitted serial correlation structure on average tends to show the least amount the bias compared to the other structures.

Finally, Figure 4.9b shows the average relative bias was overestimated in almost

all of the conditions. The average relative bias tends to be smallest when the fitted serial correlation structure was independent or MA(1) and tends to be largest when the fitted structure was MA(2) or ARMA(1,1). In contrast, when the generated serial correlation structure was ARMA(1,1) the average relative bias tends to be the largest (especially for ARMA(1,1), AR(1) and MA(2) fitted structures) and most variable. The other three generated serial correlation structures, AR(1), MA(1), and MA(2) show similar levels of average relative bias and variability. This was much different compared to the first study where the AR(1) and ARMA(1,1) tended to produce the smallest bias in the random components.



(a) Relative bias of variance b0



(b) Relative bias of residual

Figure 4.9: Relative bias of random components

4.2.3 Type I Error Rate

Just as in study one, the fixed effects did not show evidence of bias, however the random components did show evidence of bias. In addition, this study purposefully failed to model the random effect for time. As a result of bias in the random components and missing a random effect, the empirical type I error rates may differ from the expected $\alpha = .05$.

Summary statistics for the empirical type I error rate are shown in Table 4.20. From the table, three of the five fixed effects have average type I error rates close to .05, β_0, β_2, and β_3, whereas the other two terms, β_1 and β_4, have type I error rates about twice what would be expected. β_1 and β_4 also have a wide range of empirical type I error rates, ranging from a minimum of .05 to a maximum around .27.

Table 4.20: Summary statistics for empirical type I error rates

Term	Mean	Med	Min	Max
β_0	0.0572	0.0580	0.0240	0.0920
β_1	0.1167	0.1037	0.0520	0.2740
β_2	0.0647	0.0641	0.0320	0.1002
β_3	0.0581	0.0560	0.0320	0.0920
β_4	0.1133	0.0996	0.0460	0.2720

Note: Med is Median, Min is minimum, Max is maximum

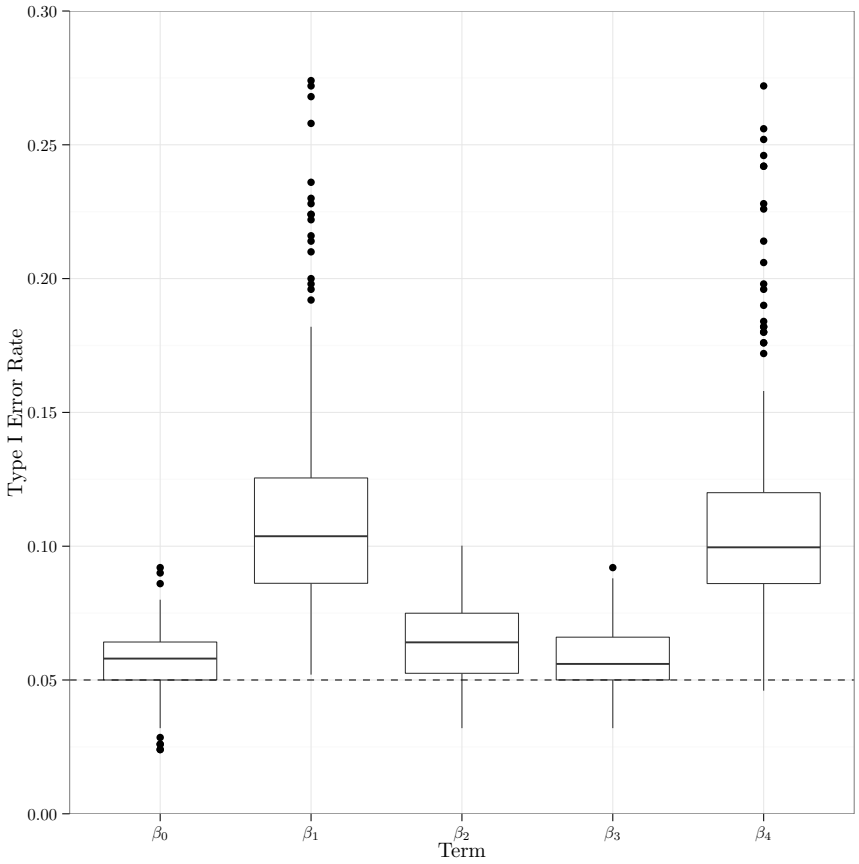


Figure 4.10: Box plots showing the distribution of type I error rates for each parameter

Figure 4.10 also displays the inflated type I error rate for the two parameters β_1 and β_4 and the increased variation for these two terms. From the figure, the median type I error rate for these two terms was around 0.1 which was about twice what would be expected based on theory. These two parameters were associated with the random effect that was deliberately not modeled, the random effect for time. The other three parameters were close to the $\alpha = 0.05$ value, but still elevated slightly.

Due to the variation found in the average empirical type I error rates, these

were modeled to attempt to explain variation with the simulation conditions and the resulting effect sizes are shown in Table 4.21. Many of the variables explained significant variation (larger than 0.01) in the empirical type I error rates and are marked in bold. There were large differences in the $\hat{\eta}^2$ statistics across the parameters. For example, for β_1 and β_4 the fitted serial correlation structure explained approximately 60% of the variation in the average empirical type I error rate whereas the other terms did not have any predictors explain this much variation in the empirical type I error rates with one term. Other covariates explaining significant variation in the empirical type I error rate for β_0, β_2, and β_3 were the two-way interaction between generated serial correlation and random effect distribution as well as the three-way interaction between the generated serial correlation, random effect distribution, and within cluster sample size. Lastly, for β_1 and β_4 the interaction between the generated and fitted serial correlation was also highly significant explaining about 20% of the variation in the empirical type I error rates. The significant interactions will be explored more fully graphically below.

Table 4.21: Eta-squared statistics for the type I error rate models up to three-way interactions.

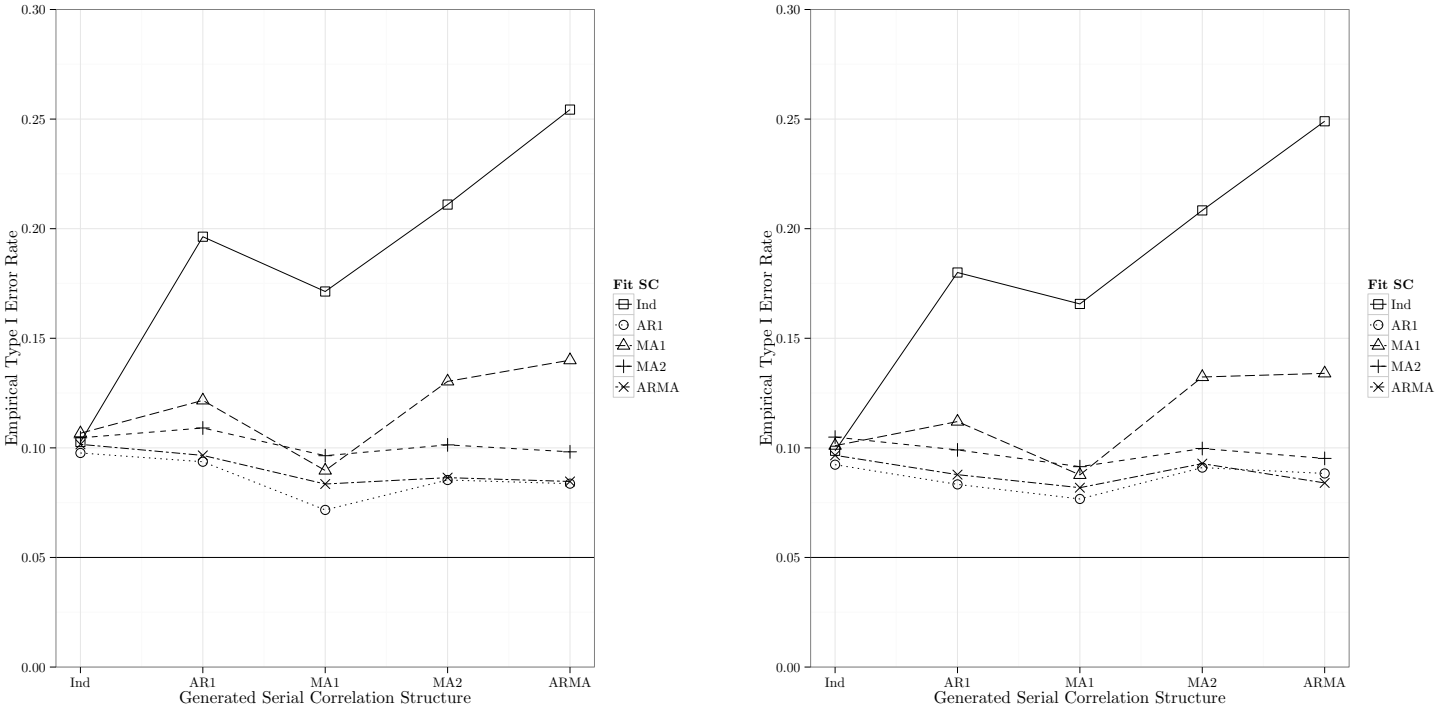
Variables	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$
Gen SC	**0.1154**	**0.0659**	**0.0243**	**0.1323**	**0.0779**
Fit SC	**0.2273**	**0.6183**	**0.2468**	**0.2485**	**0.5988**
RE Dist	**0.1153**	0.0019	**0.0524**	**0.0474**	0.0037
p	0.0066	**0.0102**	**0.0374**	**0.0320**	0.0012
Gen SC:Fit SC	**0.0709**	**0.2058**	**0.0669**	**0.0934**	**0.2065**
Gen SC:RE Dist	**0.0779**	**0.0227**	**0.2768**	**0.1094**	**0.0291**
Gen SC:p	**0.1483**	0.0058	**0.0368**	**0.0185**	0.0042

Continued on next page

Variables	$\eta^2\beta_0$	$\eta^2\beta_1$	$\eta^2\beta_2$	$\eta^2\beta_3$	$\eta^2\beta_4$
Fit SC:RE Dist	0.0017	0.0014	0.0069	0.0059	0.0013
Fit SC:p	0.0015	0.0070	0.0064	0.0023	0.0045
RE Dist:p	**0.1045**	0.0010	0.0022	**0.0213**	0.0092
Gen SC:Fit SC:RE Dist	**0.0179**	0.0040	**0.0281**	**0.0246**	0.0083
Gen SC:Fit SC:p	**0.0135**	0.0024	0.0086	0.0053	0.0019
Gen SC:RE Dist:p	**0.0766**	**0.0465**	**0.1696**	**0.2248**	**0.0467**

Note: Bold numbers are $> .01$, N is cluster sample size, Gen is generated, RE Dist is random effect distribution, p is within cluster sample size, SC is serial correlation, Fit is fitted, ":" represents interaction

The average empirical type I error rate by the generated and fitted serial correlation structures for β_1 and β_4 is shown in Figure 4.11. This figure reveals that the average empirical type I error rate was extremely inflated when the serial correlation structure was underspecified with an independence structure. This inflation was the worst when the generated serial correlation structure was ARMA(1,1) which has an empirical type I error rate about .25 for both β_1 and β_4. Also from the figure, the AR(1) fitted structure tends to have the smallest average type I error rate across the generated serial correlation structures, but was still larger than .07 for all generated serial correlation structures. Most of the variation in the average type I error rates come from the independent and MA(1) structures being severely inflated under most generated structures. However, the type I error rate was still inflated compared to what was expected at around 0.1 instead of 0.05. Lastly, the MA(1) generated serial correlation structure tended to have the smallest empirical type I error rate compared to other generated structures.



(a) β_1 (b) β_4

Figure 4.11: Average empirical type I error rates by generated and fitted serial correlation structures

The average empirical type I error rates by the generated serial correlation structure, random effect distribution, and within cluster sample size for β_0, β_2, and β_3 can be seen in Figure 4.12 through Figure 4.14 respectively. First thing to notice about Figure 4.12 through Figure 4.14 is that the range of empirical type I error rates is much smaller than the Figure 4.11 suggesting that these parameters do a much better job of controlling the empirical type I error rate. In general from the three figures, the larger within cluster sample size improves the average empirical type I error rate. Exceptions to this occur in Figure 4.13 when the generated serial correlation structure was MA(2) and the random effect distribution was chi-square (1).

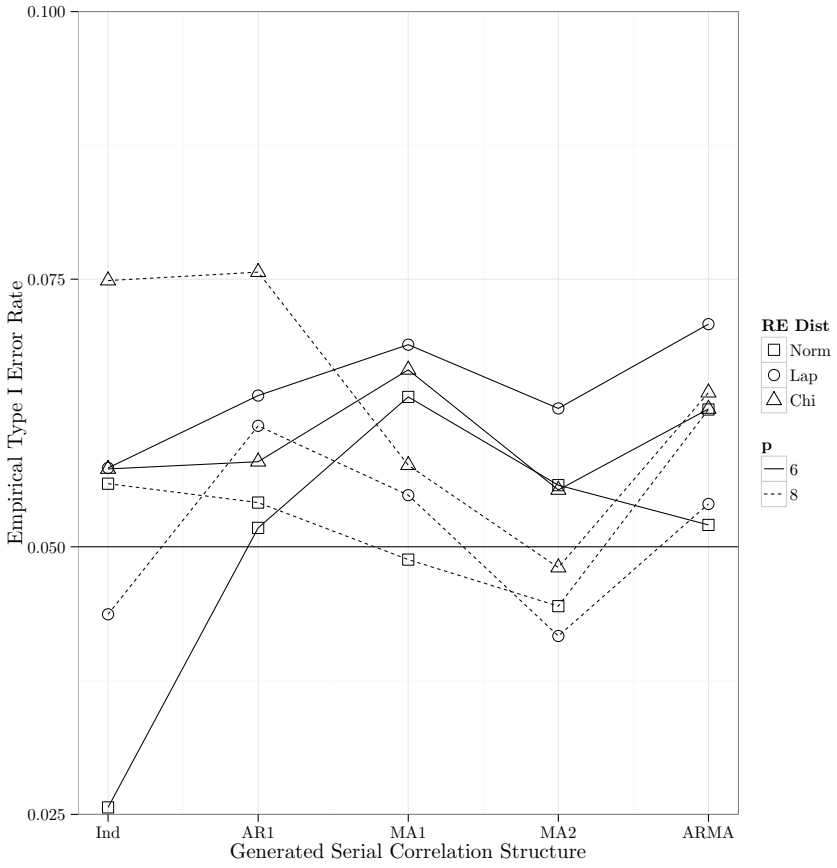


Figure 4.12: Average empirical type I error rates for β_0 by generated serial correlation structure, random effect distribution, and within cluster sample size

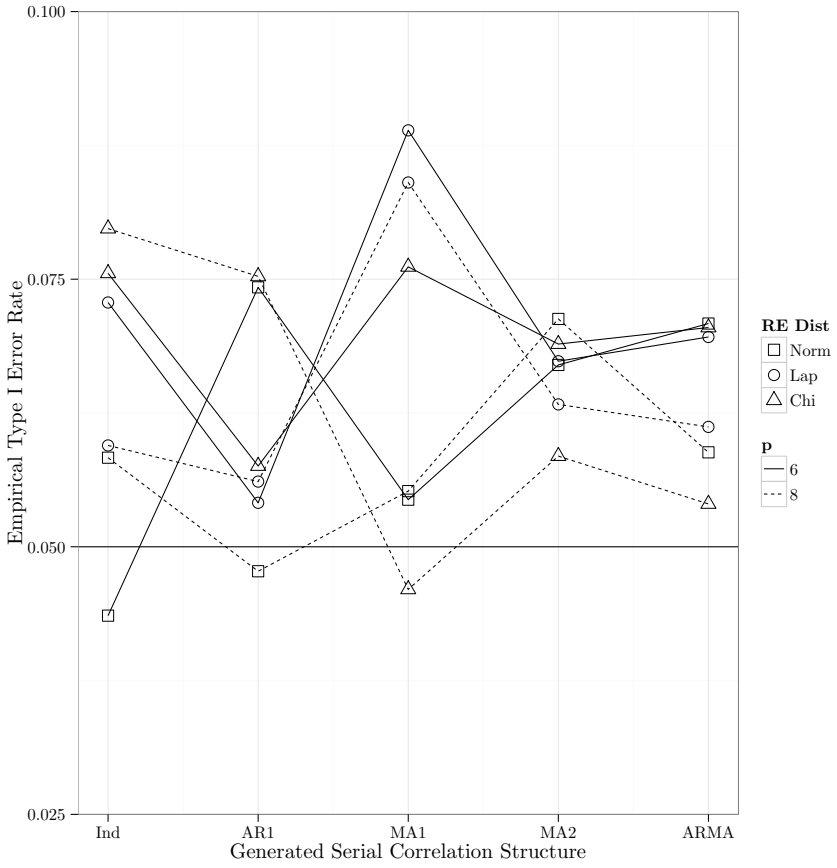


Figure 4.13: Average empirical type I error rates for β_2 by generated serial correlation structure, random effect distribution, and within cluster sample size

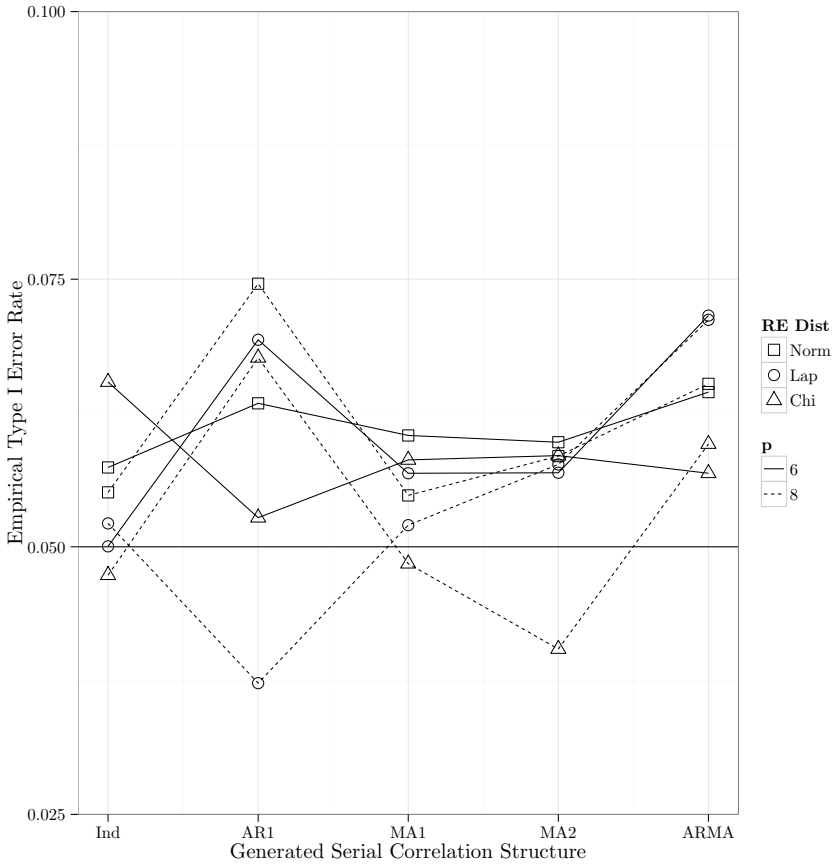


Figure 4.14: Average empirical type I error rates for β_3 by generated serial correlation structure, random effect distribution, and within cluster sample size

The empirical type I error rates for all simulated data conditions in study two are shown in Table 4.22 through Table 4.26. Each table depicts a different generated serial correlation structure.

Table 4.22: Type I error rates for independence generated serial correlation structure.

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	6	0.024	0.076	0.036	0.066	0.092
Ind	Norm	8	0.060	0.124	0.050	0.056	0.120
Ind	Lap	6	0.062	0.080	0.056	0.050	0.046
Ind	Lap	8	0.046	0.126	0.052	0.054	0.090
Ind	Chi	6	0.058	0.106	0.066	0.058	0.124
Ind	Chi	8	0.076	0.104	0.072	0.046	0.120
AR1	Norm	6	0.024	0.088	0.040	0.056	0.102
AR1	Norm	8	0.056	0.106	0.058	0.054	0.104
AR1	Lap	6	0.058	0.076	0.070	0.050	0.058
AR1	Lap	8	0.040	0.118	0.058	0.052	0.090
AR1	Chi	6	0.056	0.100	0.076	0.066	0.110
AR1	Chi	8	0.074	0.098	0.076	0.046	0.090
MA1	Norm	6	0.026	0.092	0.046	0.058	0.104
MA1	Norm	8	0.058	0.114	0.058	0.054	0.116
MA1	Lap	6	0.056	0.086	0.072	0.052	0.064
MA1	Lap	8	0.042	0.126	0.058	0.052	0.096
MA1	Chi	6	0.060	0.120	0.074	0.066	0.120
MA1	Chi	8	0.074	0.102	0.076	0.048	0.106
MA2	Norm	6	0.026	0.090	0.048	0.054	0.108
MA2	Norm	8	0.052	0.106	0.062	0.056	0.110
MA2	Lap	6	0.054	0.082	0.082	0.052	0.078
MA2	Lap	8	0.044	0.128	0.064	0.048	0.092
MA2	Chi	6	0.058	0.117	0.087	0.066	0.131
MA2	Chi	8	0.072	0.105	0.089	0.048	0.111
ARMA	Norm	6	0.029	0.103	0.048	0.053	0.107

Continued on next page

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
ARMA	Norm	8	0.053	0.104	0.064	0.055	0.111
ARMA	Lap	6	0.057	0.078	0.085	0.046	0.065
ARMA	Lap	8	0.047	0.119	0.066	0.055	0.085
ARMA	Chi	6	0.054	0.108	0.075	0.071	0.112
ARMA	Chi	8	0.078	0.097	0.086	0.049	0.099

Note: Fit is fitted, SC is serial correlation, T1E is type I error,
RE Dist is random effect distribution, p is within cluster sample size,
Norm is normal, Lap is Laplace, Chi, is chi-square (1)

Table 4.23: Type I error rates for AR (1) generated serial correlation structure.

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	6	0.066	0.192	0.058	0.070	0.134
Ind	Norm	8	0.070	0.214	0.044	0.088	0.198
Ind	Lap	6	0.076	0.172	0.046	0.076	0.190
Ind	Lap	8	0.066	0.196	0.062	0.042	0.196
Ind	Chi	6	0.068	0.182	0.050	0.070	0.182
Ind	Chi	8	0.092	0.222	0.062	0.080	0.180
AR1	Norm	6	0.046	0.102	0.066	0.060	0.058
AR1	Norm	8	0.050	0.096	0.048	0.068	0.094
AR1	Lap	6	0.056	0.082	0.050	0.066	0.086
AR1	Lap	8	0.060	0.094	0.048	0.032	0.072
AR1	Chi	6	0.054	0.092	0.050	0.046	0.106
AR1	Chi	8	0.068	0.096	0.074	0.062	0.084
MA1	Norm	6	0.054	0.124	0.082	0.064	0.070
MA1	Norm	8	0.054	0.138	0.048	0.078	0.134
MA1	Lap	6	0.070	0.102	0.058	0.068	0.124
MA1	Lap	8	0.064	0.126	0.054	0.040	0.112
MA1	Chi	6	0.058	0.118	0.060	0.052	0.126
MA1	Chi	8	0.078	0.122	0.074	0.070	0.106
MA2	Norm	6	0.046	0.120	0.080	0.062	0.084
MA2	Norm	8	0.046	0.112	0.052	0.070	0.104
MA2	Lap	6	0.062	0.096	0.060	0.070	0.108

Continued on next page

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
MA2	Lap	8	0.058	0.106	0.058	0.040	0.092
MA2	Chi	6	0.056	0.108	0.068	0.048	0.112
MA2	Chi	8	0.074	0.112	0.082	0.064	0.094
ARMA	Norm	6	0.047	0.105	0.085	0.061	0.065
ARMA	Norm	8	0.051	0.103	0.047	0.069	0.093
ARMA	Lap	6	0.057	0.081	0.057	0.067	0.095
ARMA	Lap	8	0.058	0.089	0.058	0.032	0.089
ARMA	Chi	6	0.054	0.095	0.060	0.047	0.099
ARMA	Chi	8	0.066	0.106	0.084	0.062	0.086

Note: Fit is fitted, SC is serial correlation, T1E is type I error,
RE Dist is random effect distribution, p is within cluster sample size,
Norm is normal, Lap is Laplace, Chi, is chi-square (1)

Table 4.24: Type I error rates for MA (1) generated serial correlation structure

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	6	0.080	0.160	0.042	0.074	0.180
Ind	Norm	8	0.050	0.198	0.042	0.070	0.172
Ind	Lap	6	0.080	0.158	0.082	0.066	0.182
Ind	Lap	8	0.064	0.200	0.072	0.062	0.148
Ind	Chi	6	0.078	0.146	0.072	0.072	0.136
Ind	Chi	8	0.064	0.166	0.046	0.060	0.176
AR1	Norm	6	0.058	0.056	0.048	0.062	0.086
AR1	Norm	8	0.044	0.084	0.040	0.048	0.076
AR1	Lap	6	0.064	0.080	0.078	0.050	0.106
AR1	Lap	8	0.048	0.074	0.076	0.050	0.056
AR1	Chi	6	0.058	0.078	0.064	0.052	0.060
AR1	Chi	8	0.052	0.058	0.040	0.044	0.076
MA1	Norm	6	0.064	0.070	0.052	0.050	0.088
MA1	Norm	8	0.050	0.108	0.060	0.050	0.084
MA1	Lap	6	0.066	0.082	0.092	0.060	0.106
MA1	Lap	8	0.050	0.100	0.084	0.050	0.074
MA1	Chi	6	0.064	0.100	0.080	0.052	0.076

Continued on next page

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
MA1	Chi	8	0.050	0.078	0.048	0.048	0.098
MA2	Norm	6	0.058	0.084	0.068	0.058	0.096
MA2	Norm	8	0.052	0.122	0.066	0.052	0.092
MA2	Lap	6	0.068	0.098	0.100	0.054	0.114
MA2	Lap	8	0.058	0.104	0.100	0.050	0.070
MA2	Chi	6	0.068	0.096	0.084	0.060	0.074
MA2	Chi	8	0.066	0.074	0.050	0.044	0.102
ARMA	Norm	6	0.060	0.066	0.062	0.058	0.084
ARMA	Norm	8	0.048	0.100	0.068	0.054	0.082
ARMA	Lap	6	0.066	0.084	0.092	0.054	0.102
ARMA	Lap	8	0.054	0.094	0.088	0.048	0.056
ARMA	Chi	6	0.064	0.088	0.080	0.054	0.076
ARMA	Chi	8	0.056	0.068	0.046	0.046	0.090

Note: Fit is fitted, SC is serial correlation, T1E is type I error,
RE Dist is random effect distribution, p is within cluster sample size,
Norm is normal, Lap is Laplace, Chi, is chi-square (1)

Table 4.25: Type I error rates for MA (2) generated serial correlation structure

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	6	0.064	0.210	0.056	0.068	0.206
Ind	Norm	8	0.060	0.228	0.062	0.074	0.242
Ind	Lap	6	0.080	0.236	0.066	0.066	0.184
Ind	Lap	8	0.060	0.216	0.054	0.068	0.228
Ind	Chi	6	0.066	0.152	0.046	0.070	0.214
Ind	Chi	8	0.064	0.224	0.048	0.052	0.176
AR1	Norm	6	0.052	0.092	0.068	0.058	0.078
AR1	Norm	8	0.038	0.094	0.070	0.054	0.098
AR1	Lap	6	0.050	0.094	0.060	0.050	0.094
AR1	Lap	8	0.032	0.088	0.058	0.054	0.094
AR1	Chi	6	0.052	0.062	0.066	0.054	0.096
AR1	Chi	8	0.036	0.082	0.048	0.036	0.086
MA1	Norm	6	0.056	0.138	0.070	0.060	0.116

Continued on next page

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
MA1	Norm	8	0.046	0.142	0.070	0.064	0.158
MA1	Lap	6	0.062	0.140	0.068	0.058	0.118
MA1	Lap	8	0.044	0.138	0.062	0.058	0.152
MA1	Chi	6	0.056	0.090	0.072	0.060	0.136
MA1	Chi	8	0.054	0.134	0.058	0.040	0.114
MA2	Norm	6	0.052	0.106	0.070	0.058	0.090
MA2	Norm	8	0.042	0.116	0.082	0.050	0.112
MA2	Lap	6	0.064	0.118	0.076	0.056	0.096
MA2	Lap	8	0.040	0.104	0.078	0.058	0.104
MA2	Chi	6	0.050	0.068	0.088	0.054	0.110
MA2	Chi	8	0.048	0.096	0.078	0.038	0.086
ARMA	Norm	6	0.055	0.103	0.071	0.055	0.087
ARMA	Norm	8	0.036	0.084	0.072	0.050	0.096
ARMA	Lap	6	0.059	0.099	0.067	0.055	0.089
ARMA	Lap	8	0.032	0.085	0.065	0.050	0.093
ARMA	Chi	6	0.053	0.063	0.073	0.055	0.103
ARMA	Chi	8	0.038	0.085	0.060	0.036	0.089

Note: Fit is fitted, SC is serial correlation, T1E is type I error,
RE Dist is random effect distribution, p is within cluster sample size,
Norm is normal, Lap is Laplace, Chi, is chi-square (1)

Table 4.26: Type I error rates for ARMA (1,1) generated serial correlation structure

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
Ind	Norm	6	0.072	0.274	0.062	0.072	0.272
Ind	Norm	8	0.076	0.224	0.060	0.088	0.252
Ind	Lap	6	0.090	0.258	0.058	0.088	0.226
Ind	Lap	8	0.070	0.272	0.070	0.092	0.256
Ind	Chi	6	0.080	0.230	0.072	0.080	0.246
Ind	Chi	8	0.086	0.268	0.052	0.078	0.242
AR1	Norm	6	0.038	0.106	0.050	0.064	0.086
AR1	Norm	8	0.040	0.052	0.046	0.054	0.086
AR1	Lap	6	0.058	0.086	0.062	0.056	0.086

Continued on next page

Fit SC	RE Dist	p	T1E β_0	T1E β_1	T1E β_2	T1E β_3	T1E β_4
AR1	Lap	8	0.048	0.072	0.040	0.060	0.090
AR1	Chi	6	0.052	0.082	0.058	0.046	0.100
AR1	Chi	8	0.056	0.104	0.032	0.056	0.082
MA1	Norm	6	0.060	0.138	0.062	0.072	0.136
MA1	Norm	8	0.066	0.114	0.048	0.070	0.140
MA1	Lap	6	0.074	0.148	0.064	0.072	0.120
MA1	Lap	8	0.056	0.148	0.052	0.078	0.150
MA1	Chi	6	0.068	0.124	0.068	0.064	0.132
MA1	Chi	8	0.062	0.168	0.046	0.062	0.126
MA2	Norm	6	0.046	0.102	0.090	0.056	0.100
MA2	Norm	8	0.072	0.076	0.070	0.060	0.102
MA2	Lap	6	0.068	0.108	0.088	0.072	0.088
MA2	Lap	8	0.050	0.094	0.078	0.066	0.106
MA2	Chi	6	0.060	0.087	0.080	0.048	0.091
MA2	Chi	8	0.060	0.122	0.080	0.050	0.084
ARMA	Norm	6	0.044	0.086	0.090	0.058	0.096
ARMA	Norm	8	0.060	0.060	0.070	0.054	0.092
ARMA	Lap	6	0.064	0.098	0.076	0.070	0.074
ARMA	Lap	8	0.046	0.074	0.066	0.060	0.088
ARMA	Chi	6	0.054	0.082	0.074	0.046	0.086
ARMA	Chi	8	0.058	0.108	0.060	0.052	0.068

Note: Fit is fitted, SC is serial correlation, T1E is type I error,
RE Dist is random effect distribution, p is within cluster sample size,
Norm is normal, Lap is Laplace, Chi, is chi-square (1)

Sensitivity Analysis

Just as in study one, an arcsine transformation was done on the empirical type I error rates that were analyzed above. The results for the transformed empirical type I error rates revealed very similar $\hat{\eta}^2$ suggesting that the results were robust to the choice of metric. As a result, interpretations will be made in the original metric.

Chapter 5

Conclusions

Statistical models such as the LMM rely on model assumptions to make estimation easier and computationally feasible. A common assumption made by researchers is that conditional on the random effects the within cluster residuals are indpendent within each cluster. For example, the residual at measurement occasion one is independent from the residual at measurement occasion three (and independent from all other residuals within that same cluster) after removing the contribution of the random effects. Although this assumption corresponds with the simplest residual structure, there are situations where this may not adequately model the dependency due to repeated measures, such as measurement occasions very close in time (hours, days, weeks, etc.) In these situations a more complicated within cluster residual structure may be needed, like an AR(1) structure.

Another assumption the LMM makes is normally distributed random components of the model, including the random effects. Micceri (1989) found that distributions of many widely used dependent variables were in fact not normally distributed and often severely departed from normal and symmetric distributions.

Understanding how the LMM performs when the distribution underlying the random effects is not normally distributed will help better understand the robustness of the LMM.

The current monte carlo study explored the implications for the LMM when model assumptions have not been adequately met. More specifically when serial correlation is not adequately modeled and the distribution of the random effects are not normally distributed, what implications occur to the estimation of the fixed effects, variance components, and standard errors of the fixed effects. 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). More specifically, the three-way interaction between the generated serial correlation structure, the fitted serial correlation structure, and the within cluster sample size explained significant variation in the relative bias of the random components for study one. For study two, the two-way interaction between the generated and fitted serial correlation structures explained significant variation in the relative bias of the random components.

These results show that the generated serial correlation and fitted serial correlation do matter when estimation of the variance components is concerned. 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. Specifically, the terms tended to be overestimated for every fitted serial correlation structure regardless of the generated serial correlation structure underlying the data.

In general, the random effects tend to be overestimated which was not surprising as the parameters simulated for these terms were all small making it more difficult for the estimation algorithm to underestimate these terms with a hard boundary of 0 (can not have variances smaller than 0). The one exception to the overestimation occured in study two where the AR(1) and ARMA(1,1) fitted serial correlation structures produced an underestimated variance of the intercept for all generated structures. This underestimation likely occured due to the missing random effect for time where the variation in the slope term was subsumed into the other random component. When the fitted structure was indepdent and underspecified (i.e. AR(1)) large relative bias statistics were found. This suggests that if there is evidence of serial correlation it should be modeled.

The AR(1) fitted structure tended to produce the smallest relative bias statistics in study one and the MA(1) structure in study two for the random effects across all generated serial correlation structures. This is somewhat comforting as

these are relatively simple structures with only one additional parameter needed to be estimated from the data, however differing results as to which may be best (the AR(1) structure had some of the largest average relative bias statistics in study two compared to the smallest in study one). Lastly, more measurement occasions tended to reduce the amount of relative bias for the random effects. As a result, collecting a longer time series can help reduce bias in the random effects. Longer time series may also allow for better estimates and better ability to detect serial correlation when present. Further exploration of the benefits of adding observations would be a topic to explore further.

The bias in the within cluster residuals were different than the random effects. In general, the amount of bias tended to be less and less variable compared to the random effects, which replicates previous simulation work (Kwok et al., 2007; Murphy & Pituch, 2009). The one exception is when the generated serial correlation structure was ARMA(1,1) which resulted in significantly greater variability in the relative bias statistics. Interestingly, the simple independence serial correlation structure tended to produce relative bias statistics close to 0 for most of the generated serial correlation structures. This is good news for researchers as the simplest serial correlation structure tends to produce within cluster residuals that are not severely biased, however is bad news when compared to the results of the random effects where an independent structure produced much higher relative bias statistics when a more complicated structure underlied the data. Lastly, adding more measurement occasions tended to increase the relative bias in the within cluster residuals. This may have occured due to the random effects being more accurate with more measurement occasions, resulting in more variation

being sent to the residuals.

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, 2012). 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 for study one were all close to the specified α level, however there was evidence that the type I error rates were inflated for the terms representing the missing random effect in study two. This suggests that problems occur when attempting to overcome a missing random effect for time with an overspecified covariance structure. The two terms modeling the slope had average type I error rates that were about twice what was expected by theory. For the two inflated terms, underspecifying the serial correlation structure by assuming the serial correlation structure was independent but in reality a more complicated structure underlied the data severely inflated the empirical type I error rates. The other fitted serial correlation structures did a better job of holding the empirical type I error rates in check, although they were still inflated compared to 0.05. Unfortunately for researchers, no distinct pattern emerged that limited the inflation

of the empirical type I error rates back to nominal levels.

5.1 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 or model misspecification of a random effect. 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, Dailey, & Yi, 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 underly 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.

Even though fitting an AR(1) or ARMA(1,1) structure to the data did improve the estimation of the random effects, the convergence rate did suffer when estimating these additional terms (see Table 4.1). As a result there may be a trade-off between getting better estimates of the random effects and model convergence

rate. One thing that needs to be studied further is a larger population value for the variance of the random slope. The current study had a value very close to zero which may have decreased the ability of the algorithm to converge. Evidence of this can be seen from the convergence rates of study two with a convergence rate very close to 100% for all of the simulation conditions. The only difference between the two studies is the random effect for the slope was not modeled in study two, which shows the significant difficulty in estimating the cluster specific slope effects.

Lastly, if the researcher is interested in inference about the fixed effects care needs to be taken that all of the random effects that underly the data are modeled. If all the random effects are modeled, then the empirical type I error rate is close to the specified α level set by the researcher (e.g. $\alpha = 0.05$). However, if a random effect is not modeled, the type I error rate for the fixed effects associated with that random effect can have significantly inflated type I error rates. For example, if the random effect for time is not modeled, the fixed effects for time have inflated type I error rates. Ignorning serial correlation when a random effect is not correctly modeled makes the type I error rate inflate even further.

Unfortunately, there is no a priori test to directly test for the presence of random effects or serial correlation in the data, but exploratory data analysis can go a long way in detecting the presence of a random effect. A spaghetti plot that shows the trajectories of each individual can be a good way to see if there is variation in the trajectories over time (i.e. do the slopes differ or does there appear to be a common slope among the individuals). 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 BIC to see if modeling the serial correlation improves model fit. Unfortunately, these methods have not been very reliable in selecting the correct structure (Ferron, Dailey, & Yi, 2002; Keselman et al., 1998).

5.2 Future Work

Future work exploring reasons for the poor convergence rate of the models in study one by increasing the variances of the random components would be helpful. Increasing the variance of the random components may also have an impact on the empirical type I error rate when a random effect is not modeled. Relatedly, increasing the proportion of the overall variation attributable to the random effect not being modeled may also be related to the type I error rate of the terms associated with that random effect.

Detecting serial correlation when present in the data is another area of work that needs to be explored more fully. 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. Better ways or rules of thumb to use when looking for serial correlation 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, allowing a subject to re-enter the study after missing a measurement occasion is also common in longitudinal data.

Additional work will look to relax the assumption that random effects are uncorrelated across clusters, extending the work done by Browne and Goldstein (2010) in a bayesian framework. 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 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

Afshartous, D., & de Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics*, *30*, 109 –139.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, *149*, 1–43.

Algina, J., Blair, R., & Coombs, W. (1995). A maximum test for scale: type i error rates and power. *Journal of Educational and Behavioral Statistics*, *20*(1), 27–39.

Blalock, H. (1984). Contextual-effects models: theoretical and methodological issues. *Annual Review of Sociology*, *10*, 353–372.

Boneau, C. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*, 49–64.

Bowerman, B., O'Connell, R., & Koehler, A. (2005). *Forecasting, time series, and regression: an applied approach*. South-Western Pub.

Box, G., & Jenkins, G. (1976). *Time series analysis: forecasting and control*. Prentice Hall PTR.

Browne, W., & Goldstein, H. (2010). Mcmc sampling for a multilevel model with nonindependent residuals within and between cluster units. *Journal of Educational and Behavioral Statistics*, *35*(4), 453.

Browne, W., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Bryk, A., & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158.

Chi, E., & Reinsel, G. (1989). Models for longitudinal data with random effects and ar (1) errors. *Journal of the American Statistical Association*, 452–459.

Cnaan, A., Laird, N., & Slasor, P. (1997). Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*, 2349–2380.

Darandari, E. (2003). *Robustness of Hierarchical Linear Model Parameter Estimates Under Violations of Second-Level Residual Homoskedasticity and Independence Assumptions*. (Doctoral dissertation).

De Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, *11*, 57–85.

Dempster, A. et al., Laird, N. et al., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Dempster, A., Rubin, D., & Tsutakawa, R. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 341–353.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press, USA.

Diggle, P. (1988). An approach to the analysis of repeated measurements. *Biometrics*, *44*(4), 959–971.

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.

Ferron, J., Hess, M., Hogarty, K., Dedrick, R., Kromrey, J., Lang, T., ... San Diego, A. (2004). Hierarchical linear modeling: A review of methodological issues and applications. In *Annual meeting of american educational research association, san diego*.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied longitudinal analysis*. Wiley-IEEE.

Gayen, A. (1949). The distribution of Student's t in random samples of any size drawn from non-normal universes. *Biometrika*, *36*, 353.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (2010). *Multilevel statistical models*. 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.

Grimm, K., & Widaman, K. (2010). Residual structures in latent growth curve modeling. *Structural Equation Modeling*, *17*(3), 424–442.

Hartley, H., & Rao, J. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*, 93–108.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–338.

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.

Hedeker, D., Gibbons, R., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, *24*, 70–93.

Huber, P. (1967). Maximum likelihood estimation of misspecified models. *Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability*, *1*, 221–223.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Verlag.

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*, 93–116.

Kenward, M., & Roger, J. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997.

Keselman, H., Algina, J., Kowalchuk, 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.

Keselman, H., Algina, J., Kowalchuk, 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.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill New York.

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.

Laird, N., & Ware, J. (1982). Random-effects models for longitudinal dalairdta. *Biometrics*, *38*, 963–974.

LeBeau, B. (2012). Impact of non–normal level one and two residuals on the linear mixed model. *Unpublished*.

L'ecuyer, P., Simard, R., Chen, E., & Kelton, W. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 1073–1075.

Lindley, D., & Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*, 1–41.

Littell, R., Henry, P., & Ammerman, C. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*, *76*, 1216.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*, 817–827.

Maas, C., & Hox, J. (1999). Sample sizes for multilevel modeling. *Am J Public Health*, *89*, 1181–1186.

Maas, C., & Hox, J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Maas, C., & Hox, J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427–440.

Maeda, Y. (2007). *Monte Carlo Evidence Regarding the Effects of Violating Assumed Conditions of Two-Level Hierarchical Models for Cross-Sectional Data*. (Doctoral dissertation).

Marascuilo, L., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Brooks/Cole Publishing Company Monterey^ eCalifornia California.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

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.

Núñez-Antón, V., & Zimmerman, D. (2000). Modeling nonstationary longitudinal data. *Biometrics*, *56*(3), 699–705.

Pan, J., & MacKenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, *90*(1), 239.

Pan, J., & MacKenzie, G. (2006). Regression models for covariance structures in longitudinal studies. *Statistical Modelling*, *6*(1), 43.

Pan, J., & MacKenzie, G. (2007). Modelling conditional covariance in the linear mixed model. *Statistical Modelling*, *7*(1), 49.

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer Verlag.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team. (2012). *Nlme: linear and nonlinear mixed effects models*. R package version 3.1-103.

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*, 274–308.

Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, *13*, 85–116.

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

R Development Core Team. (2010). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from R Foundation for Statistical Computing: http://www.R-project.org

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581.

Satterthwaite, F. (1941). Synthesis of variance. *Psychometrika*, *6*, 309–316.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*, 147–177.

Shumway, R., & Stoffer, D. (2000). *Time series analysis and its applications*. Springer Verlag.

Sithole, J., & Jones, P. (2002). Repeated measures models for prescribing change. *Statistics in medicine*, *21*(4), 571–587.

Sivo, S., & Willson, V. (2000). Modeling causal error structures in longitudinal panel data: a monte carlo study. *Structural Equation Modeling: A multidisciplinary journal*, *7*(2), 174–205.

Sivo, S., Fan, X., & Witta, L. (2005). The biasing effects of unmodeled arma time series processes on latent growth curve model estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 215–231.

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

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., Raudenbush, S., & TO, A. (2011). Optimal design plus empirical evidence: documentation for the fffdfffdfffdoptimal designfffdfffdfffd software.

Tabachnick, B., Fidell, L., & Osterlind, S. (2001). *Using multivariate statistics*. Allyn and Bacon Boston.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer Verlag.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Wolfinger, R. (1996). Heterogeneous variance: covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*, 205–230.