ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my advisor, Professor Jeffrey D. Long, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research and scholarship. Without his guidance and persistent help this dissertation would not have been possible.

I would like to thank my committee members, Professors Michael R. Harwell, Mark L. Davison, and Melanie M. Wall, who inspired me greatly to work in this dissertation. Their willingness to motivate me contributed tremendously to my dissertation.

In addition, a thank you to Professor Joan B. Garfield who introduced me statistics education, and whose enthusiasm for teaching had lasting effect on my teaching.
DEDICATION

This dissertation would be incomplete without a mention of the support given me by my wife, So Young Park, and my son, Hyun Suk Ryoo, to whom this dissertation is dedicated.
ABSTRACT

Model building or model selection with linear mixed models (LMM) is complicated by the presence of both fixed effects and random effects. The fixed effects structure and random effects structure are co-dependent, so selection of one influences the other.

Most presentations of LMM in psychology and education are based on a multi-level or hierarchical approach in which the variance-covariance matrix of the random effects is assumed to be positive definite with non-zero values for the variances. When the number of fixed effects and random effects is not known, the predominant approach to model building is a step-up procedure in which one starts with a limited model (e.g., few fixed and random intercepts) and then additional fixed effects and random effects are added based on statistical tests.

A procedure that has received less attention in psychology and education is top-down model building. In the top-down procedure, the initial model has a single random intercept but is loaded with fixed effects (also known as an ”over-elaborate” model). Based on the over-elaborate fixed effects model, the need for additional random effects is determined. Once the number of random effects is selected, the fixed effects are tested to see if any can be omitted from the model.

There has been little if any examination of the ability of these procedures to identify a true population model (i.e., identifying a model that generated the data). The purpose of this dissertation is to examine the performance of the various model building procedures for exploratory longitudinal data analysis. Exploratory refers to the situation in which the correct number of fixed effects and random effects is unknown before the analysis.
# Contents

Acknowledgements ................................... i

Dedication ............................................. ii

Abstract ............................................. iii

List of Tables .......................................... vii

List of Figures ......................................... ix

1 Introduction ......................................... 1

1.1 Chicago Longitudinal Study .......................... 7

1.2 Literature Review ..................................... 9

1.2.1 Model building procedure ......................... 9

1.2.2 Variable selection ................................ 13

2 Methods ............................................... 15

2.1 Linear Mixed Effects Model .......................... 15

2.1.1 Statistical Models for Longitudinal Data .......... 15

2.1.2 Formulation of LMM ............................... 18

2.1.3 Parameter Space ................................... 20

2.1.4 Estimation of Parameters ........................... 20
2.2 Model Selection ................................................ 21
  2.2.1 Tools for Model Selection .............................. 23
  2.2.2 Step Up Approach .................................. 28
  2.2.3 Top Down Approach ................................ 37
  2.2.4 Subset Approach ................................... 41

3 Data Sets ....................................................... 42
  3.1 Mathematics and Reading Achievement Scores ............ 42
  3.2 Model Building for the CLS Data sets ..................... 50
    3.2.1 Step 1 - Fitting Fixed Effects ....................... 50
    3.2.2 Step 2 - Adding Random Effects ..................... 60
  3.3 Parameter estimates ..................................... 62

4 Methods and Results ......................................... 65
  4.1 Design of the Simulation ................................ 66
  4.2 Classification Criteria ................................ 70
    4.2.1 Similarity ........................................ 71
    4.2.2 Total effect ...................................... 81
  4.3 Results of Similarity ................................... 84
  4.4 Results for Total Effect ................................ 86

5 Findings and Conclusions .................................... 90
  5.1 Sample Size ............................................. 91
  5.2 Model Building Approaches ............................. 94
  5.3 Total Effects ........................................... 96
  5.4 True Model Selection .................................. 98
  5.5 Limitations ............................................. 99
  5.6 Conclusion .............................................. 102
## List of Tables

1.1 Hypothesis test in model selection on PSID data .......................... 3
1.2 Model comparison between the linear and the cubic model ............... 5
2.1 Formulas for information criteria ........................................ 27
3.1 Missing data on both Mathematics and Reading .......................... 43
3.2 Correlation among static predictors ....................................... 44
3.3 LRT results for CLS mathematics data - time transformations .......... 52
3.4 LRT results for CLS reading data - time transformation ................. 53
3.5 LRT results for CLS mathematics data - static predictors ............... 53
3.6 LRT results for CLS reading data - static predictors ..................... 54
3.7 Parameter estimates for Mathematics for the the model of Equation (3.5) . 55
3.8 Parameter estimates for Reading for the model of Equation (3.6) ........ 56
3.9 Test result for interaction terms in Mathematics .......................... 56
3.10 Test result for interaction terms in Reading ............................... 57
3.11 Parameter estimates for mathematics with \( q = 3 \) ....................... 58
3.12 Parameter estimates for reading with \( q = 2 \) ............................ 59
3.13 Test result for random effects terms in Mathematics ..................... 61
3.14 Test result for random effects terms in Reading ......................... 61
3.15 Parameter estimates for Mathematics .................................... 63
3.16 Variance components for Mathematics ................................... 63
3.17 Parameter estimates for Reading ............................... 64
3.18 Variance components for Reading ............................... 64

4.1 Ratio of sample size and the number of parameters ................. 67
4.2 Index for selection on main effects ................................. 68
4.3 Total effects for Mathematics ........................................ 82
4.4 Total effects for Reading .............................................. 83
4.5 Proportions of selected approximate models in mathematics .... 85
4.6 Proportions of selected approximate models in reading ........... 85
4.7 Proportion of selected interaction models for mathematics - Step Up . . . . 86
4.8 Proportion of selected interaction models for mathematics - Top Down ... 87
4.9 Proportion of selected interaction models for reading - Step Up ........... 88
4.10 Proportion of selected interaction models for reading - Top Down ........... 88

5.1 Rate of selection of an approximate model according to sample sizes ... 92
5.2 Rate of selection of an approximate model according to model building approaches 94
5.3 Selection of time transformations for Mathematics- Sample size of 300 . . . 95
5.4 Selection of time transformations for Reading- Sample size of 300 ...... 96
5.5 Selection of the interaction effects according to the total effect ........... 97
5.6 Rate of selecting the true Model ........................................ 99
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Inspections of changes in annual income over time on the PSID data</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Changes in annual income conditioned on gender, education and age</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>Missing data on both Mathematics and Reading (Proportion)</td>
<td>44</td>
</tr>
<tr>
<td>3.3</td>
<td>Mean growth curves conditioned on gender</td>
<td>47</td>
</tr>
<tr>
<td>3.4</td>
<td>Mean growth curves conditioned on CPC program</td>
<td>48</td>
</tr>
<tr>
<td>3.5</td>
<td>Mean growth curves conditioned on magnet school attendance</td>
<td>48</td>
</tr>
<tr>
<td>3.6</td>
<td>Mean growth curves conditioned on risk index</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison between the approximate model and the true model for Mathematics</td>
<td>75</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison between the approximate model and the true model for Mathematics</td>
<td>76</td>
</tr>
<tr>
<td>4.7</td>
<td>Comparison between the approximate model and the true model for Reading</td>
<td>79</td>
</tr>
<tr>
<td>4.8</td>
<td>Comparison between the approximate model and the true model for Reading</td>
<td>80</td>
</tr>
<tr>
<td>4.9</td>
<td>Total Effect</td>
<td>83</td>
</tr>
<tr>
<td>4.10</td>
<td>Proportion of selected approximate models</td>
<td>85</td>
</tr>
<tr>
<td>4.11</td>
<td>Proportion of selected interaction models - Mathematics</td>
<td>87</td>
</tr>
<tr>
<td>4.12</td>
<td>Proportion of selected interaction models - Reading</td>
<td>89</td>
</tr>
<tr>
<td>5.2</td>
<td>Proportion of selected interaction models</td>
<td>98</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Longitudinal data have become one of the popular data types in applied empirical analysis in the social and behavioral sciences. For example, change in student performance over time is a common research topic in education and the data collected consist of repeated measurements of the same cohort of subjects over time. In response to this popularity, tools for statistical analysis of longitudinal data have been developed and made available to applied researchers. Applied researchers decide how these tools are used in data analysis, including decision rules for model selection. The subject of this paper is the decision rules used for model selection in longitudinal data analysis with emphasis on exploratory analysis.

Exploratory analysis occurs when the researcher has little or no preconceived idea regarding the models to fit to the sample data. Compared with statistical inference such as parameter estimation, exploratory analysis is often ad hoc and subjective in nature (Tukey, 1977 (53); Diggle, et. al., 2002 (10); Verbeke and Molenberghs, 2000 (54)). Perhaps one reason for the subjectivity is the variety of model components that must be specified. These components include time transformations, e.g., polynomial transformations, covariance (correlation) structure for the correlated observation over time, and the structure of random error.

As an illustration of the above issues, consider the Panel Study of Income Dynamics
PSID, begun in 1968, which is a longitudinal study of a representative sample of U.S. individuals (men, women, and children) and the family units in which they reside. It emphasizes the dynamic aspects of economic and demographic behavior, but its content is broad, including sociological and psychological measures. The PSID sample consists of two independent samples: a cross-sectional national sample and a national sample of low-income families. The cross-sectional sample, collected by the Survey Research Center, was an equal probability sample of households from the 48 contiguous states and was designated to yield about 3,000 completed interviews. The second sample, collected by the Bureau of the Census for the Office Economic Opportunity, selected about 2,000 low-income families with heads under the age of sixty from the Survey of Economic Opportunity respondents (Hill, 1992 (23)).

The data were analyzed by Faraway (2006, (13)), who selected a random sample subset of this data, consisting of 85 heads of household who were aged 25-39 in 1968 and had complete data for at least 11 of the years between 1968 and 1990. The variables included were annual income, gender, years of education and age in 1968. Here we focus on the analysis of annual income. As suggested in the literature on model selection (for example, Diggle, et al. (2002, (10))), applied researchers usually first look at the mean growth changes over time and/or look at the non-parametric smoothing known as locally weighted scatterplot smoothing (LOWESS). These two visual inspections are depicted in Figure 1.1 that shows the mean growth (left) and the LOWESS curve (right) of annual income over time.

Based on the figures, there may be different views among applied researchers regarding the selection on the fitted model. Some may consider a linear model for the income data as was considered by Faraway (Chapter 9, 2006 (13)). Others may consider a 3rd order polynomial model due to the fluctuations in the early 80s and mid 80s. In other words, the model selection based on the visual inspection is subjective and ad hoc.

Table 1.1 shows the statistical comparison of the linear and cubic models. By both the
Figure 1.1: Inspections of changes in annual income over time on the PSID data

(a) Mean growth curve

(b) LOWESS curve

likelihood ratio test (LRT) and the information criteria (AIC and BIC), the cubic model is preferred. This result based on the statistical comparison is more objective than the one based on the visual inspection.

Table 1.1: Hypothesis test in model selection on PSID data

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear model</td>
<td>4</td>
<td>35039.08</td>
<td>35060.74</td>
<td>-17515.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cubic model</td>
<td>6</td>
<td>35024.63</td>
<td>35057.12</td>
<td>-17506.32</td>
<td>18.45</td>
<td>2</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

In a real data analysis, we do not know the true model and need to construct an appropriate model representing the change over time based on sample data. One desirable feature of such model selection is consistency, meaning that the same model is selected by different researchers. To obtain the best fitted model for data, many researchers rely on both the visual inspection and the statistical comparison in exploratory analysis. In my point
of view, the visual inspection is unnecessary in the model selection if applied researchers fit the polynomial transformations to sample data. In other words, the statistical comparison among polynomial transformations provides a consistent result. Also an acceptable predicted curve is a result showing the change over time.

Before discussing the statistical comparison based on the LRT test or the information criteria, I would like to consider the time transformations that are fitted to sample data. To build a model for longitudinal data, many functions of time may be applied, for instance, polynomial, trigonometric, exponential, logarithmic transformations, and rational functions, and their combinations. Among those listed, polynomial functions are most well known and commonly used among applied researchers in the social and behavioral sciences. For this reason, this paper will focus on model building using polynomial functions. An advantage of examining polynomials is that they have a natural nesting structure meaning the LRT can be used for model comparison as was illustrated in Table 1.1.

In terms of selecting predictors in building the fitted model to sample data, applied researcher should consider two aspects such as which predictors should be included in the fitted model and what type of interaction between time variables and predictors should be considered. In exploratory analysis, applied researchers often perform an analysis that involves comparing models with different predictors.

Consider additional examples from the PSID data. Figure 1.2 shows how the means in annual income change over time by considering different groups in gender, education and age, separately. Figure 1.2 (a) indicates that the male group starts at a higher mean of income in 1968 and the rate of increment over time is also higher than the female group. The difference in change between groups is apparent. Figure 1.2 (b) indicates that both groups start with a similar mean of income in 1968, but the higher education group’s income increases faster than that of the lower education group. Figure 1.3 (c) indicates that both groups have a similar trend in changes over time though the older group has a slightly higher means of income than the younger group. Based on the figures, applied researchers
may include not only main effects for gender and education but also interactions between time and the predictors. Others may have different ideas in variable selections due to subjectivity.

For example, Faraway (2006, (13)) analyzed the PSID data by considering additional visual inspections such as individual growth curves. As a result, he selected the fitted model for the PSID data as the linear model, including all three predictors as main effects. And there was no interactions between time and the predictors in the fitted model. On the other hand, I chose the cubic model including gender and education predictors with up to a cubic random effects by applying the LRT with a significant level $\alpha = 0.05$. The model building process that was applied for this analysis will be discussed in detail in Chapter 2.

Instead of comparing model parameters, I compare two models in terms of the statistical comparison by looking at the LRT and the information criteria. The result is summarized in Table 1.2. The two models are both very different from each other in terms of the number of parameters. They are also statistically different when the LRT is applied. By both the LRT and the information criteria, the cubic model is preferred.

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model</td>
<td>10</td>
<td>34832.10</td>
<td>34886.25</td>
<td>-17406.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cubic Model</td>
<td>21</td>
<td>34298.67</td>
<td>34412.39</td>
<td>-17128.34</td>
<td>555.43</td>
<td>11</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The model selection introduced so far is very similar to what we discussed in the traditional linear model. In other words, it has been commonly discussed in the linear traditional model that applied researchers select the fitted polynomial model, the predictors, and their effects. In this paper, I will consider the linear mixed effects model (LMM) introduced by Laird and Ware (1982, (30)). LMM models subjects variation, which is different from the traditional linear model. More precisely, we construct an additional variance-covariance structure for the variation between subjects, which is called random effects. The formation
Figure 1.2: Changes in annual income conditioned on gender, education and age

(a) Gender

(b) Education

(c) Age
of the LMM will be discussed in detail in Chapter 2.

Prior to the advent of the LMM, we had used a repeated measures analysis of variance (RM-ANOVA) fitted to longitudinal data, including mixed effects. Compared with RM-ANOVA, LMM has the advantage of accommodating missing data. In addition, LMM is more flexible in correlation structure of observations over time (Fitzmaurice et al., 2009 (14)). Data used in this study have missing observations. To avoid the complexity coming from the missing and correlation structure, I consider the correlation structure as “unstructured”. This will be discussed in greater detail in Chapter 2.

In this paper, I mainly focus on the model selection procedure by applying the statistical comparison such as the LRT and the information criteria, to identify the correct model for longitudinal data through the simulation study. Before discussing further technical issues, let me introduce the main data set in this study.

1.1 Chicago Longitudinal Study

The Chicago Longitudinal Study (CLS) is a federally-funded investigation of the effects of an early and extensive childhood intervention in central-city Chicago called the Child-Parent Center (CPC) Program. The study began in 1986 to investigate the effects of government-funded kindergarten programs for 1,539 children in the Chicago Public Schools.

The study is in its 23rd year of operation. Besides investigating the short- and long-term effects of early childhood intervention, the study traces the scholastic and social development of participating children and the contributions of family and school practices to children’s behavior. The CPC program is a center-based early intervention that provides educational and family support services to economically disadvantaged children from preschool to third grade. It is funded by Title I of the landmark Elementary and Secondary Education Act of 1965. This CPC program is the second oldest (after Head Start) federally funded preschool program in the U.S. and is the oldest extended early childhood intervention. It
has operated in the Chicago Public Schools since 1967.

The major rationale of the CPC program is that the foundation for school success is facilitated by the presence of a stable and enriched learning environment during the entire early childhood period and when parents are active participants in their children’s education.

The Chicago Longitudinal Study has four main objectives:

1. To evaluate comprehensively the impact of the CPC program on child and family development.

2. To identify and better understand the pathways (child, family, and school-related) through which the effects of program participation are manifested, and more generally, through which scholastic and behavioral development proceeds.

3. To document and describe children’s patterns of school and social competence over time, including their school achievement, academic progress, and expectations for the future.

4. To determine the effects of family, school, neighborhood, and child-specific factors and practices on social competence broadly defined, especially those that can be altered to promote positive development and to prevent problematic outcomes.

The data I will consider are the mathematics and reading test scores over time, from kindergarten to 9th grade. The fitted models used to make inferences about objective 3 are included in Chapter 3.

In addition to the effect of the CPC program, the effects of three other variables, gender, risk factor, and magnet school attendance, will be investigated in this paper. Risk factor and magnet school attendance variable were measured, based on student, parent, and teacher surveys. The details on data sets will be discussed in Chapter 3. Rather than focus on issues of parameter interpretation, this paper will emphasize model building with the CLS data.
1.2 Literature Review

In a longitudinal data analysis, there are a number of issues regarding model fitting. Here listed two issues among them that are closely related to my research topics in this paper. The one is that the researcher must select a structure for correlated dependent variables over time, define sources of error such as subject-specific error and/or measurement error, and select a model for changes of the means over time. The model for mean change can involve various time transformations such as polynomials and other power transformations (Long and Ryoo, 2010 (32)). The other is that variable selection is still an issue as in a regression modelling (Harrell, 2001 (19)). As indicated by Harrell, variable selection is used when the analyst is faced with a series of potential predictors but does not have the necessary subject matter knowledge to enable to prespecify the important variables to include in the model.

In spite of a number of issues regarding model fitting, LMM parameter estimation has been developed and implemented in existing statistical software packages such as R, SAS, SPSS, etc. Examples of various statistical software packages can be found in West et. al. (2007 (58)). However, the accuracy in parameter estimation depends on how accurate the model is. As I narrowed down interests to reviewing the literature on model selection including variable selection, I focused on which topics have been developed in model selection.

1.2.1 Model building procedure

Model selection in LMM has received less attention than parameter estimation. Even though many statistical procedures have been proposed, the number of studies discussing LMM model selection are very limited. Jiang et. al. (2008, (26)) pointed out that model selection in LMM has never been seriously addressed in the literature. There are many reasons for the under-development of model selection methods. The reasons include the complexity of growth patterns, the variety of covariance structure in error space, etc.
I have been interested in developing the model selection methods that provide a consistent model for a given data. To minimize the variability of models selected among applied researchers, I have thought that model selection based on hypothesis tests such as the LRT should be used as a primary tool. In other words, it would be better to minimize the subjectivity during the use of visual inspections, in order to obtain a consistent model. In this paper I focus on polynomial models and model comparison using the LRT. The models to be compared differ in the order of polynomial, the number of static (time invariant) predictors, and the number of random effects. In spite of the difficulty in model selection of LMM for longitudinal data, LMM is probably the most widely used method for analyzing longitudinal data. Though RM-ANOVA and RM-MANOVA have been traditionally used in analysis, these methods are most appropriate for analyzing experimental data. Such designs seems to be the exception in much behavioral and social science research. In addition, RM-ANOVA and RM-MANOVA have restrictive assumptions regarding the parameter spaces that will be discussed in detail in Chapter 2.

Due to the abundance of non-experimental designs and missing data, many applied researchers in education and psychology consider using LMM for longitudinal data analysis. However, the model selection procedures used by researchers are arguably ad hoc and inconsistent. For example, some researchers lean heavily on visual inspection and the testing of a few number of closely related models. In some situations this can be a good model selection strategy but in other situations it might not.

Other researchers apply alternative tools to find the best fitting models among a number of candidates. For example, model selection criteria such as AIC, BIC, a conditional AIC (Vaida and Blanchard, 2005 (56)) the frequentist’s Bayes factor (Datta and Lahiri, 2001 (11)), and Generalized Information Criteria (Jiang and Rao, 2003 (25)), and likelihood ratio test (LRT) have been used. In this paper, I investigated model selection approaches based on the LRT.

LMM model selection is more demanding than traditional regression because three
components of the model must be selected: Fixed effects, random effects, and random
error. In this paper I do not consider the selection of random error. Instead, the simplest
error structure is considered in all models. Model selection procedures in LMM are usually
based on both visual inspection and a hypothesis test such as the LRT (Pinheiro and Bates
West, Welch and Galecki (2007, (58)). Pinheiro and Bates (2000, (39)), for example,
suggested that applied researchers investigate characteristics of each part of the LMM by
looking at summarized graphs such as mean growth curves, individual plots, residual plots,
etc. This visual inspection allows applied researchers to screen out candidates from the
fitted model. Then they compared candidates by applying the LRT.

However, as seen with the PSID data, the visual inspection may result in different mod-
els according to researchers. To obtain the consistent model for a given data, we had better
depend less on the visual inspection but rely more on the use of the LRT or information
criteria. In other words, if researchers set the same significant level, the LRT provides a
unique model for a given data. If researchers apply the same information criterion in an
analysis, the fitted model will be the same.

A potential problem in the LRT, however, is that tests associated with different LMM
components have different characteristics. For example, the LRT for variance components
of random effects has the boundary value problem in testing that results in a relatively
complicated sampling distribution of the statistic. In other words, if the testing parameter
is on the boundary of parameter space, the LRT does not have a $\chi^2$ distribution. At least four
methods have been proposed as remedies for the variance components testing problem. One
is to use the mixture distribution with the LRT (Self and Liang (1987, (47)); Stram and Lee
(1994, (51))). Another is to use the score test (ST; Silvapulle and Silvapulle (1995, (49));
Verbeke and Molenberghs (2003, (55)). Another is to use the parametric bootstrap method
(Pinheiro and Bates (2000, (39)); Faraway (2006, (13))). Finally, Bayesian methods have
also been suggested (Carlin and Louis, 2009 (3)). In spite of recent advances, each method
also has its own limitation. For example, the method of using the mixture distribution is based on approximation theory, which is inappropriate for small sample size analysis.

Unfortunately, the model building strategies have also received less attention. In the social and behavioral science literature, the model building strategies can be summarized in four approaches:

1. Step Up (Hox, 2002 (24); Raudenbush and Bryk, 2002 (43))
2. Top Down (Diggle et. al., 2002 (10); Verbeke and Molenberghs, 2002 (54))
3. Subset (Shang, J., and Cavanaugh, J.E. 2008 (48); Gurka, 2006 (18))
4. Inside-Out (Pinheiro and Bates, 2000 (39))

The first two approaches require that models compared should be nested, whereas the last two do not. The details will be discussed in Chapter 2. In this paper, I study sampling characteristics of the first two approaches in terms of model selection. A simulation study using the LRT with the approaches is discussed in Chapter 4.

Because LMM for longitudinal can be applied in a variety of conditions, to this end, the simulation study is confined to growth curves that consist of polynomial functions and time invariant covariates (i.e., static predictors) of change over time. In addition, methods will be examined that are perhaps most relevant to exploratory analysis.

Two methods of model selection are examined in the simulation study: Step Up and Top Down. After discussing background information on the LMM for longitudinal data in Chapter 2, I introduce the three model selection approaches in Chapter 3. The simulation study based on the CLS data set is discussed in Chapter 4. The performance of each method in terms of model selection is discussed in Chapter 5.
1.2.2 Variable selection

In addition to model building procedure, it is also a special case of model selection that we select static predictors among candidates in a data set. I want to discuss this procedure as variable selection. The problem of variable selection is one of the most pervasive model selection problems. In the process of variable selection, stepwise variable selection, backward elimination, and forward selection have been commonly used. To the contrary, Copas and Long (1991, (7)) stated one of the most serious problems with stepwise modelling eloquently when they said, “The choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so a predictor is more likely to be included if its regression coefficients is over-estimated than if its regression coefficient is underestimated.” Derksen and Keselman (1992, (8)) found that the final model usually contained less than half of the actual number of authentic predictors. There are many reasons for using methods such as full-model fits or data reduction, instead of using any stepwise variable selection algorithm.

Nevertheless, if stepwise selection must be used, a global test of no regression should be made before proceeding, simultaneously testing all candidate predictors and having degrees of freedom equal to the number of candidate variables plus any nonlinear or interaction terms if necessary (Harrell, (19)). If this global test is not significant, selection of individually significant predictors is usually not warranted. In this paper, it is hard to be implemented in the systematic procedure of model selection that we apply the full-model fits or data reduction. Thus, the stepwise selection of significant candidate was applied. The LRT results were used as a stopping criterion. In practice, recommended stopping rules are the LRT, AIC, and Mallows’ $C_p$ (Harrell, (19)).

Even though forward stepwise variable selection was used in this paper and is the most commonly used method, the elimination method is preferred when collinearity is present (Mantel, 1970 (33)). The Elimination method using Wald statistics becomes extremely
efficient when the method of Lawless and Singhal (1978, (31)) is used. For a given data set, bootstrapping can help decide between using full and reduced models. Bootstrapping can be done on the whole model and compared with bootstrapped estimates of predictive accuracy based on stepwise variable selection for each resample (Efron and Tibshirani, 1993 (12)). Sauerbrei and Schumacher (1992, (45)) developed the bootstrap method given by Chen and George (1985, (4)) combining stepwise variable selection. However, a number of the following drawbacks were pointed at Harrell (2001, (19)). First, the choice of an $\alpha$ cutoff for determining whether a variable is retained in a given bootstrap sample is arbitrary. Second, the choice of a cutoff for the proportion of bootstrap samples for which a variable is retained, in order to include that variable in the final model, is somewhat arbitrary. Third, selection from among a set of correlated predictors is arbitrary, and all highly correlated predictors may have a low bootstrap selection frequency. It may be the case that none of them will be selected for the final model even though when considered individually each of them may be highly significant. Fourth, by using the bootstrap to choose variables, one must use the double bootstrap to resample the entire modelling process in order to validate the model and to derive reliable confidence intervals. This may be computationally prohibitive.

It should be mentioned that the literature on the variable selection discussed above is in the field of linear model (LM). Thus, if we consider random effects in addition to fixed effects, we do not have a guarantee that discussion in variable selection is still valid. However, I do consider the discussion above in investigation on model selection with the linear mixed effects model (LMM).
Chapter 2

Methods

In this chapter, I discuss three statistical models for longitudinal data and argue that the LMM has features that make it attractive for the analysis of longitudinal data in education and psychology. Next, I discuss the formulation of the LMM, issues regarding parameter space, and estimation. In addition, I also discuss constraints on my simulation study in this chapter.

2.1 Linear Mixed Effects Model

2.1.1 Statistical Models for Longitudinal Data

The linear mixed effects model (LMM) for longitudinal data has been widely used among applied researchers in a variety of sciences since its modern introduction by Laird and Ware (30, 1982). However, in psychology and education the most common methods for the analysis of longitudinal data have been analysis of variance (ANOVA) type models. The LMM has more advantageous properties than ANOVA type statistical methods in terms of allowance of missing data, and various options for the variance-covariance matrix of random effects. On the other hand, the LMM can be applied wherever the ANOVA type statistical methods are applied. In this section, I discuss three different statistical models for longitudinal data: the univariate repeated-measures ANOVA (RM-ANOVA), the multi-
variate repeated-measures ANOVA (RM-MANOVA), and LMM.

RM-ANOVA has a very similar structure as a randomized block design or the closely related split-plot design. For this reason, early in its develop, the ANOVA methods seemed a natural method for repeated measures. (e.g., Yates, 1935 (59); Scheffé, 1959 (46)). In the RM-ANOVA scheme, the individuals in the study are regarded as the blocks. The RM-ANOVA can be expressed as

\[ Y_{ij} = X_{ij}^{T} \beta + b_i + e_{ij}, \quad i = 1, \cdots, N; j = 1, \cdots, n, \tag{2.1} \]

where \( Y_{ij} \) is the dependent variable, \( X_{ij} \) is the vector of indicator variables for the study factors (e.g., treatment group, time, and their interaction) and \( X_{ij}^{T} \) is its transpose, \( \beta \) is a vector of regression parameters, \( b_i \sim N(0, \sigma_b^2) \), and \( e_{ij} \sim N(0, \sigma_e^2) \). In this design, the blocks or plot effects are regarded as random rather than fixed effects. The random effect, \( b_i \), represents an aggregation of all the unobserved or unmeasured factors that make individuals respond differently. The consequence of including a single, individual-specific random effect is that it induces positive correlation among the repeated measurement, albeit with the following highly restrictive “compound symmetry” structure for the covariance structure: constant variance and constant covariance. Formally, this is expressed as,

\[
\text{Var}(Y_{ij}) = \sigma_b^2 + \sigma_e^2 \\
\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_b^2.
\]

The covariance structure among the repeated measures is a symmetric matrix consisting of diagonal entries, \( \text{Var}(Y_{ij}) \), and off-diagonal entries, \( \text{Cov}(Y_{ij}) \). Until Henderson (1963, (22)) developed a related approach for unbalanced data, it was limited to only balanced and complete data.

A related approach for the analysis of longitudinal data with an equally long history,
but requiring somewhat more advanced computations, is MANOVA. While the univariate RM-ANOVA is conceptualized as a model for a single response variable, allowing for positive correlation among the repeated measures on the same individual via the inclusion of a random subject effect, MANOVA is a model for multivariable responses. As originally developed, MANOVA was intended for the simultaneous analysis of a single measure of a multivariate vector of substantively distinct response variables. In contrast, while longitudinal data are multivariate, the vector of responses are commensurate, being repeated measures of the same response variable over time. However, there was a common feature between data analyzed by MANOVA and longitudinal data, which is that they are correlated. This led to the development of a very specific variant of MANOVA, known as RM-MANOVA.

Both ANOVA-based approaches have shortcomings that limit their usefulness in real-date applications. For RM-ANOVA, the constraint on the correlation among repeated measures is somewhat unappealing for longitudinal data, where the correlations are expected to decay with increasing separation in time. And the assumption of constant variance across time is often unrealistic. Finally, the repeated measures ANOVA model was developed for the analysis of data from designed experiments, where the repeated measures are obtained at a set of occasions common to all individuals, the covariates are discrete factors, and the data are complete. As a result, early implementations of the repeated-measures ANOVA could not be readily applied to longitudinal data that were irregularly spaced or incomplete, or when it was of interest to include quantitative covariates in the analysis.

For RM-MANOVA, there are at least two practical consequences of the constraint that the MANOVA formulation forced the within-subject covariates to be the same for all individuals. First, RM-MANOVA cannot be used when the design is unbalanced over time. Second, the RM-MANOVA did not allow for general missing-data patterns to arise. Thus, while ANOVA methods can provide a reasonable basis for a longitudinal analysis in cases where the study design is very simple, they have many shortcomings that have limited
their usefulness in real-data applications. In educational and psychological data, there is considerable variation among individuals in both the number and timing of measurements. The resulting data are highly unbalanced and not readily amenable to ANOVA methods developed for balanced designs.

Since the 1960s and 1970s, researchers began to use a two-stage model to overcome some of the limitations of the ANOVA methods (Laird and Ware 1982, (30)). In this formulation, the probability distribution for the multiple measurements has the same form for each individual, but the parameters of that distribution vary over individuals. The distribution of these parameters, or random effects, in the population constitutes the second stage of the model. Such two-stage models have several desirable features. There is no requirement for balance in the data. This design allows explicit modelling and analysis of between- and within-individual variation. Often, the individual parameters have a natural interpretation which is relevant to the goals of the study, and their estimates can be used for exploratory analysis.

2.1.2 Formulation of LMM

In the early 1980s, Laird and Ware (1982, (30)) proposed a flexible class of linear mixed-effects models (LMM) for longitudinal data that expanded a general class of mixed models introduced by Harville (1977, (20)). LMM can handle the complications of mistimed and incomplete measurements in a very natural way.

The linear mixed-effects model (LMM) has the form (Laird and Ware, 1982 (30))

\[ y_i = X_i \beta + Z_i b_i + e_i, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i \]  

(2.2)

where \( y_i \) is the \( n_i \times 1 \) vector of observations for the \( i \) subject, \( X_i \) is an \( n_i \times p \) design matrix of independent variables for the fixed effects, \( Z_i \) is an \( n_i \times q \) design matrix of independent variables for the random effects, \( \beta \) is a \( p \times 1 \) vector of fixed effects parameters, the
are independent \( q \times 1 \) vectors of random effects with \( N(0, D) \) distribution, and the \( e_i \) are independent \( n_i \times 1 \) vectors of random errors with \( N(0, \sigma^2 I_i) \) distributions. The \( b_i \) are independent of the \( e_i \). The total number of observations is \( \sum_{i=1}^{N} n_i \).

The role of the three parts of the LMM can be explained as follows. The fixed effects are the population average coefficients for time variables and other predictors, which models the variations of mean growth change in data for the population. In contrast, the random effects account for the heterogeneity among the subjects by allowing differences from the overall average. Finally, the error accounts for the variation unexplained by the fixed and random effects.

The random effects and error are assumed to be independent, and their parameter spaces should be considered in model selection. However, these two parts represent unobserved variations and there are various options for the structure, which prevents me from constructing procedure of model building in an objective way. Thus, for simplicity, I only consider the random effects part as an “unstructured” variance-covariance matrix and restrict the error space as one-dimensional by setting \( e_i \sim N(0, \sigma^2 I) \). In addition, I consider well-formulated models in the selection of variables (Peixoto, 1987 (40)) such that if a variable \( X^n \) is included in a LMM, the model is hierarchically well formulated when all terms of order less than \( n \) are also included in the model. Similarly, if an interaction term \( X^n_1X^n_2 \) is included in the model, all low-order terms \( X^i_1X^j_2, 0 \leq i \leq n, 0 \leq j \leq m \) must also remain in the model, even if they are not statistically significant (Morrell, et. al., 1997 (37)).

This single level formula above (2.2) can be extended to a multilevel formula (Pinheiro and Bates, 2000 (39), Chapters 2). For example, we can consider nested levels of random effects. This multilevel extension structure has been applied in education data. For example, the Junior School Project data were collected from primary schools in inner London. The data are described in detail in Mortimore, Sammons, Stoll, Lewis, and Ecob (1988, (38)). In these data, there is a multilevel structure such as three years performance scores nested within students nested within class. The analysis can be found at Faraway (2006,
(13)). In this paper, I, however, do not consider such multilevel structure in model building but single-level structure proposed by Laird and Ware (1982, (30)).

2.1.3 Parameter Space

The parametric space for the LMM can vary in terms of the structure of fixed effects, random effects and random errors. In the setting of equation (2.2), the parameter space is

$$\mathbb{R}_p \times \mathbb{R}_+^q \times \mathbb{R}_+^{(q-1)/2} \times \mathbb{R}_+$$

(2.3)

where $\mathbb{R}$ is the real number space and $\mathbb{R}_+$ is the nonnegative real number space.

In the model selection procedures, only nested models will be tested. When the fixed effects between nested models are compared, the other parameters will be fixed, which indicates the difference of the number of parameters in two nested models is 1. When the random effects between nested models are compared, the other parameters will be fixed. However, the difference of the number of parameters between two nested models is $q + 1$ for the case that Equation (2.3) is the base model. That is, the full model has the following parameter space:

$$\mathbb{R}_p^{\times} \times \mathbb{R}_+^{q+1} \times \mathbb{R}_+^{(q+1)/2} \times \mathbb{R}_+$$

(2.4)

This will be explained in greater details later in this section with model equations.

2.1.4 Estimation of Parameters

Let us assume $\sigma^2$ and the random effects covariance matrix $D$ to be known. Then the estimator of the fixed-effects parameter vector $\beta$ is the generalized least squares estimator
(Laird and Ware, 1982)

\[
\hat{\beta} = \left( \sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i^T V_i^{-1} y_i,
\]

(2.5)

where \( V_i = Z_i D Z_i^T + \sigma^2 I_i, i = 1, \cdots, N \). The covariance matrix of these estimates is given by

\[
\text{cov}(\hat{\beta}) = \left( \sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1}.
\]

(2.6)

When \( \sigma^2 \) and \( D \) are not known, but estimates of \( \sigma^2 \) and the random effects covariance matrix \( D \) are available, then we can estimate \( \hat{\beta} \) by using the expression (2.5) with replacing those estimates in \( \sigma^2 \) and \( D \).

The variance components \( \sigma^2 \) and \( D \) are estimated either using maximum likelihood (ML) or restricted maximum likelihood (REML). The marginal log-likelihood for computing maximum likelihood estimates is given by Laird, Lange, and Stram, 1987 as

\[
l_{ML}(\beta, \theta) = -\frac{\sum_{i=1}^{N} \ln(2\pi)}{2} - \frac{\sum_{i=1}^{N} \ln |V_i|}{2} - \frac{\sum_{i=1}^{N} (y_i - X_i \beta)^T V_i^{-1} (y_i - X_i \beta)}{2},
\]

(2.7)

where the vector \( \theta \) contains the unique elements of \( \sigma^2 \) and \( D \). To compute REML estimates of the variance components, the log-likelihood becomes

\[
l_{REML}(\beta, \theta) = l_{ML} + \frac{p \ln(2\pi)}{2} + \frac{1}{2} \left( \ln |\sum_{i=1}^{N} X_i^T X_i| - \ln |\sum_{i=1}^{N} X_i^T V_i^{-1} X_i| \right)
\]

(2.8)

### 2.2 Model Selection

There is no single strategy in model building for longitudinal data that is applicable to every situation. In the literature on model building, four common strategies have been
discussed. Model building is sometimes thought of as an iterative procedure that requires a series of model comparison tests and visual investigations based on an appropriate mean growth curve for the observed data (Pinheiro and Bates, 2000 (39)). In some areas, there are accepted models for particular situations, such as versions of nonlinear growth models (see Grasela and Donn, 1985 (16), for example).

In contrast to estimation, there has been relatively little research on the effectiveness of different model building strategies with LMM. Among the strategies that have been discussed in the applied literature are


2. Top down approach (Verbeke and Molenberghs (2000) (54), Chapter 9; Diggle (1988) (9); Diggle, Heagerty, Liang, and Zeger (2002) (10), Chapters 3, 4, and 5)


4. Inside-Out approach (Pinheiro and Bates, 2000 (39), Chapter 4).

In this paper, I will study the performance of the first two approaches. Though the subset search method is often used in model selection, the subset search is a computer-intensive method. Due to the computer-intensive, it is not possible to implement the subset search in my program that compares all nested models and provides the best fitted model. Thus, it is excluded in this paper. The Inside-Out approach is not considered in this paper even though that approach does make sense. The reason to be excluded in this paper is that it depends on the visual investigation of fixed and random effects. The approach is summarized very simply here: the approach starts with individual fits by group, using plots of the individual coefficients to decide the random-effects structure, and finally fits a mixed-effects model to the complete data. For details see Pinheiro and Bates (2000 (39), Chapter 4).
Before examining both the step up and the top down approaches in detail, I first review the tools for model comparison in this section such as likelihood ratio test and information criteria. In the end of this section, I also describe the two approaches: step up and top down, in detail.

2.2.1 Tools for Model Selection

In this section, I describe two different types of tools that can be used to compare models: the likelihood ratio test (LRT) and information criteria (IC). The former can be used when comparing nested models whereas the latter can be used to examine nested and non-nested models.

Likelihood Ratio Test

The LRT statistic for testing \( H_0 : \theta \in \Theta_0 \) versus \( H_1 : \theta \in \Theta_C \) is defined as,

\[
\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)},
\]

where \( L(\theta|x) \) is a likelihood function for a given data \( x \). A LRT is any test that has a rejection region of the form \( \{ x : \lambda(x) \leq c \} \), where \( c \) is any number satisfying \( 0 \leq c \leq 1 \).

Suppose that \( \hat{\theta} \), a maximum likelihood estimate (MLE) of \( \theta \), exists on the entire parameter space \( \Theta \) and that \( \hat{\theta}_0 \) is a MLE on \( \Theta_0 \). The LRT statistic is

\[
\lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}.
\]

\( c \) is specified since it depends on the class of tests. Under suitable regularity conditions such that the first and second derivatives of \( L(\theta|x) \) are continuous on \( \Theta \), if \( H_0 \) is true, then the distribution of \( -2\log \lambda \) is asymptotically that of \( \chi^2 \) with dimension \( \dim(\Theta) - \dim(\Theta_0) \) degrees of freedom (see Wilks, 1938 (57)).
The strong assumptions that first and second derivatives of $L$ with respect to $\theta_j \in \Theta$ are continuous functions almost everywhere in a certain region of the $\theta$-space for almost all possible samples $\Theta_0$ had been investigated by Chernoff (1954 (5)). He provided a representation of the asymptotic distribution of $-2 \ln \lambda_n$. Consider that $\Theta_0$ is an $r$-dimensional hyperplane of $n$-dimensional $\Theta$. Let $\theta_0$ lie on $\Theta_0$ and also be an interior point of $\Theta$. Then the distribution of $-2 \ln \lambda_n$ is that of $\chi^2_{n-r}$.

Self and Liang (1987, (47)) developed the results of Chernoff (Chernoff, 1954 (5)) by introducing the cone, $C$, defined as a set of points such that if $x \in C$ then $a(x - \theta_0) + \theta_0 \in C$, where $a$ is any real, nonnegative number.

**Theorem 2.2.1**

(Theorem 3 in Self and Liang, 1987 (47)). Let $\theta_0$ be a boundary point of both $\Theta_0$ and $\Theta^C_0$ and also be an interior point of $\Theta = \Theta_0 \cup \Theta^C_0$. Then the asymptotic distribution of $-2 \ln \lambda_N$ is the same as that of

$$\sup_{\theta \in C_\Theta - \theta_0} \left[ -(Z - \theta)^T I(\theta_0) (Z - \theta) \right] - \sup_{\theta \in C\Theta_0 - \theta_0} \left[ -(Z - \theta)^T I(\theta_0) (Z - \theta) \right] \quad (2.9)$$

where $Z$ has a multivariate Gaussian distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$.

It is convenient to rewrite (2.9) as

$$\inf_{\theta \in \tilde{C}_0} ||\tilde{Z} - \theta||^2 - \inf_{\theta \in \tilde{C}} ||\tilde{Z} - \theta||^2$$

where

\[
\tilde{C} = \left\{ \tilde{\theta} : \tilde{\theta} = \Lambda \frac{1}{2} P^T \theta \mbox{ for all } \theta \in C_\Theta - \theta_0 \right\} \\
\tilde{C}_0 = \left\{ \tilde{\theta} : \tilde{\theta} = \Lambda \frac{1}{2} P^T \theta \mbox{ for all } \theta \in C\Theta_0 - \theta_0 \right\}
\]
where \( \tilde{Z} \) has a multivariate Gaussian distribution with mean 0 and identity covariance matrix and \( PAP^T \) represents the spectral decomposition of \( I(\theta_0) \).

Self and Liang (1987 (47)) proved the existence of a consistent MLE, the large sample distribution of that estimator, and the large sample distribution of LRT statistics for nine cases. Their nine cases were reviewed by Stram and Lee (1994 (51), 1995 (52)) who focused on LMMs. They assumed that the true value of the parameter \( \sigma^2 \) for measurement error is non-zero and there are no additional constraints being imposed on the parameter estimates \( \beta \) for the fixed effects. In other words, \( \sigma^2 \) and \( \beta \) lie in the interior of the admissible region for these parameters. In Stram and Lee’s paper (1994 (51)), they restricted their descriptions of the geometry of \( \Theta_0 \) and \( \Theta_0^C \) to deal only with \( D \) that is the variance-covariance matrix of the random effects in LMMs. The following properties are the summary of the results on the asymptotic behavior of likelihood ratio tests for non-zero variance components of the random effects in Stram and Lee (1994 (51), 1995 (52)).

Case 1  Testing \( D = 0 \) versus \( D = (d_{11}) \). This corresponds to Case 5 of Self and Liang (1987) (47) and the asymptotic distribution of \( -2 \ln \lambda_N \) is a 50:50 mixture of \( \chi^2_0 \) and \( \chi^2_1 \).

Case 2  Testing \( D = \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix} \) against \( D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix} \) positive semidefinite. The large-sample distribution of \( -2 \ln \lambda_N \) is a 50:50 mixture of \( \chi^2_2 \) and \( \chi^2_1 \).

Case 3  Testing \( D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix} \) with \( q \times q \) positive-definite matrix \( D_{11} \), against \( D = \begin{pmatrix} D_{11} & d_{12} \\ d_{12}^T & d_{22} \end{pmatrix} \) \((q+1) \times (q+1)\) positive-semidefinite matrix. The large-sample distribution of \( -2 \ln \lambda_N \) is a 50:50 mixture of \( \chi^2_{q+1} \) and \( \chi^2_q \).

Case 4  Testing \( D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix} \) with \( q \times q \) positive-definite matrix \( D_{11} \), against \( D = \begin{pmatrix} D_{11} & D_{12} \\ D_{12}^T & D_{22} \end{pmatrix} \) with \( D_{22} \) diagonal matrix of order \( k \times k \) and \( D_{12} \) a matrix making \( D \) be at least positive-semidefinite. The large-sample distribution of \( -2 \ln \lambda_N \) is a mixture of \( \chi^2 \).
random variables with degrees of freedom $q_k, q_k - 1, q_k - 2, \ldots, (q - 1)k$, where the mixing probability for the $\chi^2_j$ component is $\binom{k}{j} 2^{-k}$ where $j = 0, \ldots, k$.

Let $l_{ML0}$ be the marginal log-likelihood from the maximum likelihood (ML) estimation computed under the null model and $l_{ML1}$ be the marginal log-likelihood from the maximum likelihood estimation computed under the alternative model. Then the ML log-likelihood ratio test statistic is defined as follows:

$$LRT_{ML} = 2(l_{ML1} - l_{ML0}).$$  \hfill (2.10)

Similarly, using the REML log-likelihood, an alternative test statistic is defined as follows:

$$LRT_{REML} = 2(l_{REML1} - l_{REML0}).$$  \hfill (2.11)

Even though it has been known that the test based on (2.10) and (2.11) is conservative under the distribution of the test statistic as a single $\chi^2$ (see Pinheiro and Bates, 2000 (39)), I do apply a single $\chi^2$ distribution due to the lack of accurate mixture distributions in the model comparison purpose (See Stram and Lee, 1994 (51) and 1995 (52)). Another reason that I apply a single $\chi^2$ distribution is that most applied researchers do not use the mixture distribution. Also it is known that LRT is relatively accurate when the number of fixed effects tested is small, the sample size is large, and the random effects do not vary between the full and reduced model (Morrell, 1998 (36)).

**Information Criteria**

As an alternative to the LRT, information criteria have also been used to compare models. This is especially the case when comparing mixed models that are non-nested, such as models with different covariance structures or power transformations of the predictors (see e.g., Long and Ryoo, 2010 (32)). Information criteria (IC) such as Akaike’s Information
Criterion (AIC), and its variants (e.g., AICC and CAIC), and Schwartz’s Bayesian Information Criterion (BIC) (see formulae in table 2.1 below), and many other variations, are often used for these purposes. In general, these information criteria are functions of the calculated likelihood for a given model with penalty term based on the number of parameters in the model, and possibly the sample size. The use of these criteria is strictly subjective; no formal inference based on their values can be made. However, these information criteria can be used to determine substantial differences in fit. Comparison of the values of the criteria for a set of models simply indicates the superior fitting model, with the most common definitions having the smallest value for the best fitting model (see Table 2.1).

When discussing model selection criteria, two important large-sample concepts are efficiency and consistency. Efficient criteria choose the best model of finite dimension when the “true model” (which is known) is of infinite dimension. In contrast, consistent criteria choose the correct model with probability approaching 1 when a true model of finite dimension is assumed to exist and is included in the candidate models fit to the sample data. Selection criteria usually fall into one of the two categories. For instance, the AIC and AICC are efficient criteria, while the BIC and CAIC are considered to be consistent criteria (Gurka, 2006 (18)). Table 2.1 shows the formulas for the IC measures. As indicated in the footnote of Table 2.1, the $l$ can be either $l_{ML}$ or $l_{REML}$, but in this paper I only consider $l_{ML}$.

Table 2.1: Formulas for information criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Formula (Smaller is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>$-2l + 2s$</td>
</tr>
<tr>
<td>AICC</td>
<td>$-2l + 2s \left( \frac{N^<em>}{N^</em> - s - 1} \right)$</td>
</tr>
<tr>
<td>CAIC</td>
<td>$-2l + s(\ln N^* + 1)$</td>
</tr>
<tr>
<td>BIC</td>
<td>$-2l + s(\ln N^*)$</td>
</tr>
</tbody>
</table>

Note: Here, $l$ is either $l_{ML}$ or $l_{REML}$, $s$ refers to the number of parameters of the model, and $N^*$ is a function of the number of observations, that is, $\sum_{i=1}^{N} n_i$ where $N$ is the number of subjects and $n_i$ is the number of observations within $i$ subject.
2.2.2 Step Up Approach

The step up approach is one method of model building for longitudinal data (Pinheiro and Bates (2000) (39); Raudenbush and Bryk (2002) (43), Chapters 5 and 6; Hox (2002) (24), Chapter 5). It is a common practice that applied researchers search for the best fitting model starting from the simplest model, such as a random-intercepts model, proceeding to more complex models until the selected model is not significantly different from the more complex model. There are three different parts of LMMs to which this step up approach can be applied: fixed effects, random effects and measurement error.

Furthermore, if longitudinal data include static predictors, fitting fixed effects can be considered as three different steps such as choosing time transformation, selecting main effects from static predictors, and selecting interaction effects between the time transformation and the main effects. The procedure of fitting fixed effects can be explained as follows. First, we find the best fitting random-intercepts model by varying the degree of polynomial of the time predictor in the fixed effects structure. Second, we test the significance of each static predictor that appears as a single or main effect in the model. Third, we test interaction terms between the time transformations and the static predictor in fixed effects structure. After fitting the fixed effects, we finally investigate the variance covariance structure of the random effects. The size of variance-covariance matrix will be limited to the highest degree of time predictor selected in the fixed effects.

Before describing the steps in detail, I consider constraints in my program that provides the best fitted polynomial model by applying the LRT. All steps described above were implemented in my program. But I constrain the degree of interactions between time transformations and main predictor, the degree of random effects, and the measurement error. To better understand these constraints, let us assume that we fit a $6^{th}$ order polynomial model as fixed-effects and select four main effects. Now, we consider interaction effects between time transformations and main effects. Theoretically, we may consider an
interaction effect between the 6\textsuperscript{th} order term. However, we merely interpret the effects. In practice, we do not consider such high order interaction. To avoid this, I constrain the highest possible interaction term as 3\textsuperscript{rd} order. To discuss the constraint of random-effects, we recall the Equation 2.2.

\[ y_i = X_i \beta + Z_i b_i + e_i, \quad i = 1, \ldots, N, \ j = 1, \ldots, n_i \]

In the formulation, the column space of \( Z_i \) is a subset of the column space of \( X_i \), which tells us that the highest degree of random effects is theoretically the same as the degree of time transformations. If we assume that we fit a 6\textsuperscript{th} order polynomial, we can have \( b_{i0} \) to \( b_{i6} \) that provide up to 28 parameters in the variance covariance structure. We do not want to have too many parameters. Thus in my program I constrain the highest degree of random-effects as 3\textsuperscript{rd} order.

In this paper, I restricted the measurement error part in the simplest form as \( R_i = \sigma_i^2 I_i \). More complex structures for the random error portion are possible but not considered in this paper (see Pinheiro and Bates, 2000 (39)). The step up approach can be described using the fitting a model in each step below.

**Step 1: Intercept-only Model**

The simplest LMM that we will consider is the intercept-only model,

\[ y_{ij} = \beta_0 + b_{0i} + e_{ij}, \quad (2.12) \]

where \( b_{0i} \sim N(0,d_{11}) \) and \( e_{ij} \sim N(0,\sigma^2) \). \( \beta_0 \) is the fixed effect that is a constant over time, \( b_{0i} \) is the random effect representing between-subjects variation, \( e_{ij} \) is the error.

The model above is different from the means-only model that does not include the random effects part \( b_{0i} \). With longitudinal data, repeated measures are clustered within sub-
jects. To model such between-subjects variability, the random effects part \( b_{0i} \) is basically added. That is, \( d_{11} > 0 \).

Equation (2.12) can be rewritten as a hierarchical linear model (HLM) (see Raudenbush and Bryk, 2002 (43)):

\[
\begin{align*}
  y_{ij} & = \beta_{0i} + e_{ij} \\
  \beta_{0i} & = \beta_0 + b_{0i}.
\end{align*}
\]

This intercept-only model becomes the base for the model comparison in the step up approach.

**Step 2: Adding Fixed Effects**

As described in the introduction of the step up approach, I consider three steps to model the fixed effects: One is to fit a growth curve model by time transformations, another is to select main effects among static predictors, and the other is to fit the interactions between time transformations and the main effects.

**Step 2a: Adding Time Transformation in Fixed Effects.** To fit time transformations in fixed effects, I consider polynomial models and mainly focus on finding the degree of the polynomial. The following process was implemented in my program. First, we compare the base model (Equation (2.12)) with a linear random intercepts model

\[
y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + b_{0i} + e_{ij}.
\]  

(2.13)

That is, we test which model is better fitted for data by applying the LRT with the significance level \( \alpha = 0.05 \). If the linear random intercepts model (Equation (2.13)) is not significantly different from the base model, we select the time transformations of the base model for fixed effects. Otherwise, we compare the linear random intercepts model with a
quadratic random intercepts model whose degree of time transformation is one more than that of the linear random intercepts model.

This comparison of \((p - 1)^{th}\) order random intercepts model with \(p^{th}\) order random intercepts model continues unless the LRT result is not statistically significant any more. To avoid the saturated model, this comparison will be stopped when \(p\) becomes the number of time points less than 1.

When increasing the degree of polynomial to be compared, we consider the well-formulated model. That is, if a variable \(X^n\) is included in a LMM, all terms of order less than \(n\) are also included in the model. Similarly, if an interaction term \(X^i_1X^j_2\) is included in the model, all lower-order terms \(X^i_1X^j_2\) for \(0 \leq i \leq n\) and \(0 \leq j \leq m\) must be also remain in the model even if they are statistically not significant, for the model to remain well formulated (see Morrell, Pearson, and Brant, 1997 (37)).

Griepentrog et al. (1982, (17)) show that the standardized statistics used to test that the parameters are zero will be the same for the untransformed and transformed parameters only for the highest order of polynomial terms and interactions. And Peixoto (1990, (41)) shows that the estimation space of a linear regression model will be the same for the untransformed and transformed models only if they well formulated. Morrell et al. (1997, (37)) extends the results in the LMMs. In education and psychology, it is of interest to make an inference from the best fitted model, i.e., to predict a growth based on the best fitted model. Thus, applied researchers want to make invariant estimates for not only the higher order term but also the lower terms. In this paper, I consider the well-formulated models in model building procedure.

Assuming that the maximum order in time transformations is \(p\), I formulate the resulting model as

\[
y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + b_0 + e_{ij} \tag{2.14}
\]
where $0 \leq p \leq n - 1$. Regardless of the significance of $\beta_k, k = 0, \cdots, p - 1$, the model includes all lower order polynomial terms, to be a well-formulated model.

**Step 2b: Adding Static Predictors in Fixed Effects.** In the literature on variable selection, it is often discussed that the static predictors are selected either by the confirmatory assessment by researchers or by the exploratory assessment (See Kutner, Nachtsheim, Neter, and Li, 2004 (27), Chapter 9; Steinley and Brusco, 2008 (50), Section 2). In the confirmatory assessment, the static predictors were chosen on the basis of prior knowledge and should be retained for comparison with earlier studies even if some of the static predictors turn out not to lead to any error variance reduction in the study at hand. In exploratory assessment, the number of static predictors that remain after initial screening typically is still large. Hence, some applied researchers usually will wish to reduce the number of static predictors to be used in the final model. In this paper, we consider the exploratory assessment. To identify static predictors that affect the response variable, we apply the LRT with a significance level, $\alpha = 0.5$.

As a second step of fitting the fixed effects part, static predictors are individually tested and added if the LRT result is significant in comparison between two nested models such as

$$y_{ij} = \sum_{k=0}^{p} \beta_k^{i} + b_0 + e_{ij}$$  \hspace{1cm} (2.15)$$

$$y_{ij} = \sum_{k=0}^{p} \beta_k^{i} + x_{1i} + b_0 + e_{ij}$$  \hspace{1cm} (2.16)$$

where $x_{1i}$ is a static predictor. The null hypothesis $H_0$ is that two models are not different from each other, that is, the coefficient of $x_1$ is not different from zero. By applying the likelihood ratio test ($df = 1$), it will be determined whether the variable $x_1$ is added or not.

In this step, $K$ statistical tests are run to determine whether each predictor is statistically significant or not, where $K$ is the number of static predictors ($x_1, \cdots, x_K$). The resulting
model is

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} + b_{0i} + e_{ij} \]  

(2.17)

where \( q \) is the number of static predictors that are statistically significant with significant level as \( \alpha = 0.05 \). Since we compare two models (2.15) and (2.16), there is always one random effect in this stage of the selection.

**Step 2c: Adding Interactions between Time Transformation and Static Predictors.**

Interactions in the fixed effects structure considered in this step produce up to \( p \times q \) additional terms where \( p \) is the highest degree of time transformations and \( q \) is the number of static predictors selected in step 2b. In this paper, I do not consider any dynamic predictors that vary over time but only consider static predictors that do not vary over time. To better understand model equations compared in this step, I re-write Equation (2.17) as a HLM model that is the reduced model and I also write an equation as a HLM model that is the full model in the LRT as follows.

\[
\begin{align*}
y_{ij} &= \sum_{k=0}^{p} \beta_k t_{ij}^k + e_{ij} \\
\beta_{0i} &= \beta_0 + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} + b_{0i} \\
\beta_{1i} &= \beta_1 \\
\beta_{2i} &= \beta_2 \\
&\quad \vdots \\
\beta_{pi} &= \beta_p
\end{align*}
\]
and

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + e_{ij} \]

\[ \beta_{0i} = \beta_0 + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} + b_{0i} \]

\[ \beta_{1i} = \beta_1 + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} \]

\[ \beta_{2i} = \beta_2 \]

\[ \vdots \]

\[ \beta_{pi} = \beta_p. \]

As we can see the model equations above, we test if the full model including an interaction effect between the time transformation and the static predictors selected in step 2b is significantly different from the reduced model. That is, if there is a set of static predictors as in step 2b, we test if the interaction between the time transformation and the set is significant. The HLM model equations can be written in the LMMs as follows.

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_k \beta_{1+p+lq} + b_{0i} + e_{ij} \]

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_k x_{k-p} + \sum_{l=1}^{r+1} \sum_{k=1}^{q} t_{ij}^l x_k \beta_{1+p+lq} + b_{0i} + e_{ij} \]

by the likelihood ratio test with \( df = q \). In addition, we also add all lower interaction terms for the model to be well-formulated.

Assuming that \( r \) is the highest degree polynomial that was significantly different from lower order interactions but not significantly different from the higher order interaction, we
obtain the resulting model as

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{r+p} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_k \beta_{1+p+l-q} + b_0 + e_{ij}, \]  

(2.18)

which is the best fitted model among random-intercept models.

**Step 3: Adding Random Effects**

In parameter space, random effects are independent of fixed effects (see Equations (2.3) and (2.4)). Since random effects were structured to investigate between-subject variations, the space of random effects parameter is a subspace of error space. However, when we construct a structure of random effects, we use the relationship between design matrices of fixed-effects and random effects. Based on the notations in Equation (2.2), we consider \( Z_i \) generated by the subset of columns in \( X_i \). That is, the column space of \( Z_i \) is the sub-space of the column space of \( X_i \). In this paper, I do not consider \( Z_i \) as any column subspace of \( X_i \) but consider \( Z_i \) as a well-formulated column subspace of \( X_i \). That is, similar to step 2, the lower order terms will always be in the model. That is, if the random quadratic term is added in the model, the constant and linear random effects terms are also included regardless of their significance. Thus, the resulting model is well-formulated (Morrell, Pearson, and Brant, 1997 (37)).

As many researchers point out, if we test a variance component such as \( \sigma^2 \), the null hypothesis tested is on the boundary of the parameter space (Self and Liang, 1987 (47); Stram and Lee, 1994 (51), 1995 (52); Verbeke and Molenberghs, 2003 (55); Molenberghs and Verbeke, 2007 (35)). Even though using a single \( \chi^2 \) test for LRT statistic is more conservative, i.e., yields \( p \)-values that tend to be too large, than a mixture of \( \chi^2 \)s as a distribution for test, I use this conservative LRT test with a single \( \chi^2 \) distribution because the LRT result is pretty accurate for fixed effects and many applied researchers in education and psychology still use the LRT.
In this step, the following two models are compared,

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t^k_{ij} + \sum_{k=p+1}^{r+p} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t^l_{ij} x_k \beta_{1+p+l} + b_{0i} + \sum_{k=0}^{s} b_{kit_{ij}} + e_{ij} \]

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t^k_{ij} + \sum_{k=p+1}^{r+p} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t^l_{ij} x_k \beta_{1+p+l} + b_{0i} + \sum_{k=0}^{s} b_{kit_{ij}} + e_{ij}. \]

However, the variance-covariance structures are of the following forms

\[
\begin{pmatrix}
\text{Var}(b_0, b_0) & \text{Cov}(b_0, b_1) & \cdots & \text{Cov}(b_0, b_s) \\
\text{Cov}(b_1, b_0) & \text{Var}(b_1, b_1) & \cdots & \text{Cov}(b_1, b_s) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(b_s, b_0) & \text{Cov}(b_s, b_1) & \cdots & \text{Var}(b_s, b_s)
\end{pmatrix}, \quad (2.19)
\]

\[
\begin{pmatrix}
\text{Var}(b_0, b_0) & \text{Cov}(b_0, b_1) & \cdots & \text{Cov}(b_0, b_s) & \text{Cov}(b_0, b_{s+1}) \\
\text{Cov}(b_1, b_0) & \text{Var}(b_1, b_1) & \cdots & \text{Cov}(b_1, b_s) & \text{Cov}(b_1, b_{s+1}) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\text{Cov}(b_s, b_0) & \text{Cov}(b_s, b_1) & \cdots & \text{Var}(b_s, b_s) & \text{Cov}(b_s, b_{s+1}) \\
\text{Cov}(b_{s+1}, b_0) & \text{Cov}(b_{s+1}, b_1) & \cdots & \text{Cov}(b_{s+1}, b_s) & \text{Var}(b_{s+1}, b_{s+1})
\end{pmatrix}. \quad (2.20)
\]

Since the variance-covariance matrix is symmetry, we do have the degree of freedom of the hypothesis testing as \( s + 1 \).

Assuming that \( s \) is the highest degree polynomial in random effects that was significantly different from lower order random effects model but not significantly different from the higher order random effects model, we obtain the resulting model as

\[ y_{ij} = \sum_{k=0}^{p} \beta_k t^k_{ij} + \sum_{k=p+1}^{r+p} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t^l_{ij} x_k \beta_{1+p+l} + b_{0i} + \sum_{k=0}^{s} b_{kit_{ij}} + e_{ij} \]  \( (2.21) \)
where $0 \leq s \leq p$. In Equation (2.20) above, $s + 1$ random components that is the last row of Equation (2.20) are added on Equation (2.21). Comparing with Equation (2.18), we can count $\frac{(s+1)(s+2)}{2} - 1$ addition random components in Equation (2.21).

### 2.2.3 Top Down Approach

The top down approach is another approach to model building for longitudinal data (Verbeke and Molenberghs (2000) (54), chapter 9; Diggle (1988) (9); Diggle, Heagerty, Liang, and Zeger (2002) (10), chapters 3, 4, and 5). In contrast to the step up approach, the top down approach deliberately favors overparameterized models for fixed effects, because an underparametrized model (for example, fitting linear rather than quadratic time trends) will tend to induce spurious autocorrelation structure into the residuals (Diggle (1988) (9)). In addition, when the data are from a designed experiment in which the only relevant explanatory variables are the treatment labels, it is a sensible strategy to use a ”saturated model” for the mean structure. This incorporates a separate parameter for the mean response at each time point within each treatment group (Verbeke and Molenberghs (2000) (54)). However, this overparameterization may not allow random error and so, may cause too restrictive a specification on the covariance structure that invalidates inferences when the assumed structure does not hold (Altham (1984) (1)).

### Step 1: Saturated Random Intercept Model

The saturated model for the mean structure is the polynomial model passing through all means over time. For example, if there are 9 time points, the polynomial model passing through all 9 means can be constructed by a polynomial of order 8. In practice, this should be avoided, since it does neither allow for random error nor is parsimonious. However, it is definitely the most complex model for time transformations in fixed effects. According
to the time points $n$, the saturated random-intercept model can be fitted as follows:

$$y_{ij} = \sum_{k=0}^{p_1} \beta_k t_{ij}^k + b_{0i} + e_{ij}$$  \hspace{1cm} (2.22)

where $p_1 = n - 1$ assuming balance on time (i.e., $n_i = n$) and $b_{0i}$ is a random effects. This is base model in the model comparisons.

**Step 2: Fixed Effects**

In this step we construct the fixed effects part of the model with polynomials. We select the highest order terms that are statistically significant with all lower order terms included regardless the significance.

**Step 2a: Trim Time Transformations in Fixed Effects.** From the saturated model, we trim off the non-significant highest order term from time transformations in fixed effects. Applying the LRT, we compare the following two nested models

$$y_{ij} = \sum_{k=0}^{p_2} \beta_k t_{ij}^k + b_{0i} + e_{ij}$$

$$y_{ij} = \sum_{k=0}^{p_2+1} \beta_k t_{ij}^k + b_{0i} + e_{ij}$$

where $0 \leq p_2 \leq p_1 - 1$

This step will stop either if the model having $p_2$ degree of the time transformation is statistically significantly different from the model having $p_2 + 1$ degree of the time transformation or if there was no statistically significant results until $p_2 = 0$. Assuming that the selected order for the time transformation is $p$, I formulate the resulting model as

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + b_{0i} + e_{ij}$$  \hspace{1cm} (2.23)

where $0 \leq p \leq n - 1$. Regardless of the significance of $\beta_k, k = 0, \ldots, p - 1$, the model in-
cludes all lower order polynomial terms, to be a well-formulated model (Morrell, Pearson, and Brant, 1997 (37)).

**Step 2b: Adding Static Predictors.** Similar to step 2b of the step up approach, I test each static predictor by applying the LRT with the significance level $\alpha = 0.05$ on the following two nested models.

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + b_{0i} + e_{ij},$$

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + x_{1i} + b_{0i} + e_{ij},$$

where $x_{1i}$ is a static predictor.

Assuming that $q$ static predictors were significant, we obtain the following model

$$y_{ij} = \sum_{k=0}^{p} \beta_{k_{ij}}^k + \sum_{k=p+1}^{p+q} \beta_{k_{ij}} x_{k_{ij}} - p + b_{0i} + e_{ij},$$

(2.24)

**Step 2c: Adding Interactions between Time Transformations and Static Predictors.**

Similar to step 2c in the step up approach, we test two nested models,

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_{k_{ij}} x_{k_{ij}} - p + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_{k} \beta_{1+p+l q} + b_{0i} + e_{ij}$$

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_{k_{ij}} x_{k_{ij}} - p + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_{k} \beta_{1+p+l q} + b_{0i} + e_{ij},$$

with the LRT using $df = q$. In addition, we also add all lower interaction terms for the model to be well-formulated.

Assuming that $r$ is the highest degree polynomial that was significantly different from lower order interactions but not significantly different from the higher order interaction, we
obtain the resulting model as

\[ y_{ij} = \sum_{k=0}^{p} \beta_{k} t_{ij}^k + \sum_{k=p+1}^{p+q} \beta_{k} x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} \beta_{1+p+l} t_{ij}^l x_{k} + b_{0i} + e_{ij}. \]  \hspace{1cm} (2.25)

### Step 3: Adding Random Effects

Similar to step 3 in the step up approach, we test the following two models

\[ y_{ij} = \sum_{k=0}^{p} \beta_{k} t_{ij}^k + \sum_{k=p+1}^{p+r} \beta_{k} x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_{k} \beta_{1+p+l} + b_{0i} + \sum_{k=1}^{s} b_{ki} t_{ij}^k + e_{ij}, \]

\[ y_{ij} = \sum_{k=0}^{p} \beta_{k} t_{ij}^k + \sum_{k=p+1}^{p+r} \beta_{k} x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij}^l x_{k} \beta_{1+p+l} + b_{0i} + \sum_{k=1}^{s+1} b_{ki} t_{ij}^k + e_{ij}. \]

However, these two models above cannot show what variance-covariance structures were compared. We can write two variance-covariance structures for both model as follows.

\[
\begin{pmatrix}
\text{Var}(b_0, b_0) & \text{Cov}(b_0, b_1) & \cdots & \text{Cov}(b_0, b_s) \\
\text{Cov}(b_1, b_0) & \text{Var}(b_1, b_1) & \cdots & \text{Cov}(b_1, b_s) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(b_s, b_0) & \text{Cov}(b_s, b_1) & \cdots & \text{Var}(b_s, b_s)
\end{pmatrix}, \hspace{1cm} (2.26)
\]

\[
\begin{pmatrix}
\text{Var}(b_0, b_0) & \text{Cov}(b_0, b_1) & \cdots & \text{Cov}(b_0, b_s) & \text{Cov}(b_0, b_{s+1}) \\
\text{Cov}(b_1, b_0) & \text{Var}(b_1, b_1) & \cdots & \text{Cov}(b_1, b_s) & \text{Cov}(b_1, b_{s+1}) \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\text{Cov}(b_s, b_0) & \text{Cov}(b_s, b_1) & \cdots & \text{Var}(b_s, b_s) & \text{Cov}(b_s, b_{s+1}) \\
\text{Cov}(b_{s+1}, b_0) & \text{Cov}(b_{s+1}, b_1) & \cdots & \text{Cov}(b_{s+1}, b_s) & \text{Var}(b_{s+1}, b_{s+1})
\end{pmatrix}. \hspace{1cm} (2.27)
\]

Since the variance-covariance matrix is symmetric, we do have the degrees of freedom of the hypothesis testing as \( s + 1 \).
Assuming that $s$ is the highest degree polynomial in random effects that was significantly different from lower order random effects model but not significantly different from the higher order random effects model, we obtain the resulting model as

$$y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{k=p+1}^{r+p} \beta_k x_{k-p} + \sum_{l=1}^{r} \sum_{k=1}^{q} t_{ij} l x_k \beta_{1+p+l} q + b_0 + \sum_{k=0}^{s} b_k t_{ij}^k + e_{ij}$$ \hspace{1cm} (2.28)

where $0 \leq s \leq p$.

In summary of the step up and the top down approaches, they are different only on step 1 and 2a. The rest of model building procedure is identical. However, the steps 2b, 2c, and 3 are conditioned on the result of steps 1 and 2a, which distinguish from each other.

### 2.2.4 Subset Approach

Another approach that can be used to select a LMM is the subset approach. The main difference between the subset approach and the step up and top down approaches is that information criteria are used to compare models rather than the LRT. In addition, researchers can compare non-nested models, which is an advantage over the step up and top down approaches that requires nested models to be compared.

However, there are also the disadvantages such as difficulty in determining subset of candidates and a computer intensive approach. Simulation study in this paper does not include subset approach as serious as the other approaches. I will not discuss the result but explain how the study was conducted. Firstly, I selected the best fitted time transformation, say $p$, among random intercepts models similar to step 2a. Secondly, I constructed subsets including all possible candidates having the degrees of time transformations as $p-1$, $p$, and $p+1$. Thirdly, I obtained all information criteria for all candidates in the subset. Finally, I selected the best model having the smallest information criteria. Due to the computer-intensive, I exclude this approach in this paper.
Chapter 3

Data Sets

3.1 Mathematics and Reading Achievement Scores

For the simulation study to be described in Chapter 4, a real data set was adopted as the population. In the data, two response variables, reading and mathematics scores, were considered separately. That is, the model building procedures for two response variables were considered separately.

The participants were from the Chicago longitudinal study (CLS) which consisted of $N = 1,539$ African American (93%) and Hispanic (7%) children who grew up in high-poverty neighborhoods in central-city Chicago and graduated from government-funded kindergarten programs in 1985-1986. The same cohort of students took standardized mathematics and reading achievement tests with vertically equated scale scores each year from kindergarten to $9^{th}$ grade of the Iowa Test of Basic Skills (ITBS).

In the data analysis, data with $N = 1,531$ were considered due to the missingness in the explanatory variable, gender. Missing data for both mathematics and reading are given in Table 3.1, respectively. Since students whose scores were missed more likely have lower score or no score in the previous grade and the missingness is not related to any static predictor, the missing data mechanism is missing at random (MAR) (Fitzmaurice et. al., 2004 (15)). After drastic increment (11% of $N$), missing data are fairly linear growth, varying
Table 3.1: Missing data on both Mathematics and Reading

<table>
<thead>
<tr>
<th></th>
<th>Mathematics Missing data</th>
<th>Proportion</th>
<th>Reading Missing data</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>8.00</td>
<td>0.01</td>
<td>8.00</td>
<td>0.01</td>
</tr>
<tr>
<td>1st Gr</td>
<td>181.00</td>
<td>0.12</td>
<td>189.00</td>
<td>0.12</td>
</tr>
<tr>
<td>2nd Gr</td>
<td>228.00</td>
<td>0.15</td>
<td>218.00</td>
<td>0.14</td>
</tr>
<tr>
<td>3rd Gr</td>
<td>250.00</td>
<td>0.16</td>
<td>251.00</td>
<td>0.16</td>
</tr>
<tr>
<td>4th Gr</td>
<td>315.00</td>
<td>0.20</td>
<td>306.00</td>
<td>0.20</td>
</tr>
<tr>
<td>5th Gr</td>
<td>310.00</td>
<td>0.20</td>
<td>305.00</td>
<td>0.20</td>
</tr>
<tr>
<td>6th Gr</td>
<td>341.00</td>
<td>0.22</td>
<td>336.00</td>
<td>0.22</td>
</tr>
<tr>
<td>7th Gr</td>
<td>410.00</td>
<td>0.27</td>
<td>407.00</td>
<td>0.26</td>
</tr>
<tr>
<td>8th Gr</td>
<td>473.00</td>
<td>0.31</td>
<td>455.00</td>
<td>0.30</td>
</tr>
<tr>
<td>9th Gr</td>
<td>498.00</td>
<td>0.32</td>
<td>497.00</td>
<td>0.32</td>
</tr>
</tbody>
</table>

from 2% to 5% increments on both mathematics and reading. The missing proportions on both mathematics and reading are fairly similar. The plot for missing data is given in Figure 3.1.

In each population considered for the simulation study, the response variable is either mathematics or reading test scores and the static predictors are gender, participation in a child-parent center program (cpc), participation in a magnet school (magnet), and a risk index variable. The gender, cpc and magnet variables are binary. For the gender variable, the female group was considered as the reference group and was coded as 0. For the cpc variable, the cpc participant group was considered as the reference group and was coded as 0. For the magnet variable, the magnet school participant group was considered as the reference group and was coded as 0. For the magnet variable, the magnet school participant group was considered as the reference group and was coded as 0. The continuous risk index variable was based on student, parent and teacher surveys and ranged from 0 to 7.75. The cpc and gender variables were measured at kindergarten. The risk variable was based on several measurements at different time points with scores combined to yield a single overall score indicating each student’s risk level. The magnet variable indexed magnet school attendance from 4th grade to 8th grade, and was considered as a static predictor characterizing the educational back-
ground of each student. The risk and magnet variables were also used as static predictors along with gender and cpc.

Table 3.2 shows the correlation matrix among the static predictors. There is no strong relationship between any two predictors. Thus, there is no co-linearity problems among the predictors.

Table 3.2: Correlation among static predictors

<table>
<thead>
<tr>
<th></th>
<th>gender</th>
<th>cpc</th>
<th>magnet</th>
<th>risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>1.000</td>
<td>0.042</td>
<td>0.093</td>
<td>0.028</td>
</tr>
<tr>
<td>cpc</td>
<td>0.042</td>
<td>1.000</td>
<td>0.145</td>
<td>-0.045</td>
</tr>
<tr>
<td>magnet</td>
<td>0.093</td>
<td>0.145</td>
<td>1.000</td>
<td>-0.163</td>
</tr>
<tr>
<td>risk</td>
<td>0.028</td>
<td>-0.045</td>
<td>-0.163</td>
<td>1.000</td>
</tr>
</tbody>
</table>

In the simulation study, the CLS data set will constitute the population and a LMM developed from the data will constitute the true model (or population model). Then sam-
Figure 3.2: Mean growth curves in math and reading data sets

![Mean growth curves in math and reading data sets](image)

Samples will be generated according to the true model and the step up and top down methods previously considered in the Chapter 2 will be performed for each sample.

One way to investigate the trend of change over time is to look at the mean growth curves. Figure 3.2 shows the mean growth changes for both mathematics and reading. Since the test is vertically equated, we are able to compare test scores over grade, which allows us to investigate the growth of students’ development in each subject. The fluctuations shown in the figure are interesting. The slower rate of change at 3rd grade may indicate that the concepts introduced at this grade may be more difficult than other grades. Rather than offer substantive analysis for the growth trajectories, the focus of this paper is on building statistical models for the growth trends.

The figure also indicates the mathematics and reading scores curves are very similar. A visual inspection suggests that a 2nd or 4th order polynomial models might be appropri-
ate based on counting the fluctuations of the curves over time. These visual inspections represent mean changes over time, which can be modelled within the fixed effects part of the LMM. However, these polynomial models for the fixed effects of the LMM may be different from the best fitting models, since there are interactions among the fixed effects, the random effects and the error parts of the LMM.

In addition to looking at the mean growth curves, we need to investigate the variance-covariance structure of the LMM. Many approaches for this have been suggested, for example, looking at a plot of the ordinary least square residual profiles versus time (Verbeke and Molenberghs (2000), (54)), or an "inside-out" model building approach (Pinheiro and Bates (2000), (39)). Though a number of approaches based on visual inspection are available, the resulting model selected may change from researcher to researcher.

In this section, I focus on describing the CLS data before fitting LMMs. In addition to the time transformations, there are four static predictors that will be investigated. To investigate discrepancies and/or interactions between groups that are classified by static predictors, I draw the mean growth curves conditioned on the four static predictors. Based on the figures listed below, it can be observed how the static predictors are related to the response variables. For example, if the growth curves for two levels of a static predictor are parallel over time, that indicates constant effects over time. If two lines cross, interaction effects between the time transformation and the static predictor are necessary.

Figure 3.3 shows two mean growth curves conditioned on gender for each response variable. Both mathematics and reading scores differentiate as grade increases. For mathematics, the female group has a higher mean than the male group for each grade. The difference between the means of the two groups gets larger as grade increases. From the 3rd grade, the difference increases. Based on the graph, it is reasonable to consider a main effect and an interaction effect with the time transformations for gender. For reading, the difference between the means of the two gender groups gets larger as grade increase. This difference increases from kindergarten onward. Based on the graph, we expect a main
effect and an interaction effect with the time transformations for gender.

Figure 3.4 shows the two mean growth curves conditioned on cpc for each response variable. Both mathematics and reading scores show close to a constant difference between the two cpc groups. For mathematics, the cpc group has a higher mean than the non-cpc group for each grade. Based on the graph, it is reasonable to consider a main effect but not an interaction effect for cpc. For reading, the difference between the means of the two cpc groups is also near constant. Based on the graph, we expect a main effect but not an interaction effect with the time transformations for cpc.

Figure 3.5 shows the two mean growth curves conditioned on the magnet variable for each response variable. For mathematics, the magnet group has a higher mean than the non-magnet group at all grades. The difference between the means of the two groups gets larger as grade increases. Based on the graph, it is reasonable to consider a main effect and an interaction effect with the time transformations for magnet. For reading, the difference between the means of the two magnet groups is constant. Based on the graph, we expect a
Figure 3.4: Mean growth curves conditioned on CPC program

(a) Mathematics

(b) Reading

Figure 3.5: Mean growth curves conditioned on magnet school attendance

(a) Mathematics

(b) Reading
main effect but not an interaction effect with the time transformations for magnet.

The risk variable is continuous ranging from 0 to 7.75. In order to investigate the effect of the risk variable, it was divided into 2 discrete levels. Figure 3.7 shows the two groups created based on a median split. Both mathematics and reading scores differentiate as grade increases. For mathematics, the lower risk group has higher a mean than the high risk group for all grades. The difference between the means of the two groups gets larger as grade increases. Based on the graph, it is reasonable to consider a main effect and an interaction effect with the time transformations for risk. For reading, the difference between the means of the two risk groups gets larger as grade increases. Based on the graph, we expect a main effect and an interaction effect with the time transformations for risk.
3.2 Model Building for the CLS Data sets

Based on the investigation of the data sets in the last section, I will construct two LMMs for the CLS data using the step up approach that was discussed in Chapter 2. These best fitting models may not be true models for the CLS data but will provide true models for the simulation study. The parameter estimates from the models fitted to both data sets will be used in the generation of the sample data for the simulation study. That is, the parameter estimates for the best fitting model are going to be the population parameters.

The goal of this study is to find the most efficient way either to select the true model or to select an adequate approximation to the true model. By “adequate approximation” I mean a model that includes the most substantial effects.

3.2.1 Step 1 - Fitting Fixed Effects

Fitting Time Transformations

In Step 1, I determine the fixed effects for the time transformations. Results for both mathematics and reading are summarized in Tables 3.3 and 3.4, respectively. Using a significance level of $\alpha = 0.05$, a $6^{th}$ order polynomial and $2^{nd}$ order polynomial were selected for mathematics and reading, respectively. The $6^{th}$ order polynomial model for mathematics data was a little more complex than that in Figure 3.2 whereas the $2^{nd}$ order polynomial for reading data was slightly simpler than that in Figure 3.2. In Tables 3.3 and 3.4, ‘Df’ indicates the number of parameters in each model and ‘Chi Df’ indicates the degree of freedom in the LRT. Since the Df increases by 1, the Chi Df is 1 in all model comparisons. As indicated in the tables, the $p$-values clearly indicate the plausibility of the $6^{th}$ degree polynomial for mathematics and the $2^{nd}$ degree polynomial for reading, though a case for a higher order model with reading can certainly be made.

In Tables 3.3 and 3.4, the fixed effects were determined by limiting the random effects to random intercepts and the error part to $\sigma^2 I_i$. Each row in Tables 3.3 and 3.4 indicates the
result of the following model comparison by assuming balance on time, i.e., \( n_i = n \). In the model comparison, Equation (3.1) is the base model fitted with the \( n^{th} \) order polynomial while Equation (3.2) is the full model fitted with the \((n + 1)^{th}\) order polynomial model.

\[
y_{ij} = \sum_{k=0}^{n} \beta_k t_{ij}^k + b_0i + e_{ij} \quad \text{(3.1)}
\]

\[
y_{ij} = \sum_{k=0}^{n+1} \beta_k t_{ij}^k + b_0i + e_{ij} \quad \text{(3.2)}
\]

where \( n = 0, \ldots, 8 \) is the degree of time transformation in the fixed effects portion. For mathematics (Table 3.3), the 6\(^{th}\) order polynomial is not significantly different from the 7\(^{th}\) order polynomial model. Thus, the 6\(^{th}\) order polynomial is selected as the highest order of the time transformations. For reading (Table 3.4), the 2\(^{nd}\) order polynomial is not significantly different from the 3\(^{rd}\) order polynomial model. Thus, the 2\(^{nd}\) order polynomial is selected as the highest order of the time transformations.

**Fitting Main Effects**

In this step, I select for the true model the significant static predictors among the gender, cpc, magnet, and risk index variables. The results are summarized in Tables 3.5 and 3.6 for each response variable. The LRT results are based on the following model comparisons,

\[
\text{Reduced model} \quad y_{ij} = \sum_{k=0}^{n} \beta_k t_{ij}^k + b_0i + e_{ij}, \quad \text{(3.3)}
\]

\[
\text{Full model} \quad y_{ij} = \sum_{k=0}^{n} \beta_k t_{ij}^k + x_{li} + b_0i + e_{ij}, \quad \text{(3.4)}
\]

where \( n \) is 6 for mathematics and 2 for reading and \( x_{li} \) for \( l = 1, 2, 3, 4 \) are statistic predictors where \( l = 1 \) is for the gender variable, \( l = 2 \) for the cpc variable, \( l = 3 \) for the magnet variable, and \( l = 4 \) for the risk index variable. In Tables 3.5 and 3.6, the reduced model was compared with the model including only one static predictor at a time.
Table 3.3: LRT results for CLS mathematics data - time transformations

<table>
<thead>
<tr>
<th>n</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>122510.00</td>
<td>122532.27</td>
<td>-61252.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>95274.03</td>
<td>95303.72</td>
<td>-47633.01</td>
<td>27237.97</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>93703.40</td>
<td>93740.52</td>
<td>-46846.70</td>
<td>1572.63</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>93608.50</td>
<td>93653.04</td>
<td>-46798.25</td>
<td>96.90</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>93334.73</td>
<td>93386.69</td>
<td>-46660.36</td>
<td>275.78</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>93253.13</td>
<td>93312.52</td>
<td>-46618.57</td>
<td>83.59</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>93167.55</td>
<td>93234.36</td>
<td>-46574.78</td>
<td>87.58</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>93168.30</td>
<td>93242.53</td>
<td>-46574.15</td>
<td>1.25</td>
<td>1</td>
<td>0.2634</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>93165.44</td>
<td>93247.10</td>
<td>-46571.72</td>
<td>4.86</td>
<td>1</td>
<td>0.0275</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>93164.89</td>
<td>93253.96</td>
<td>-46570.44</td>
<td>2.56</td>
<td>1</td>
<td>0.1098</td>
</tr>
</tbody>
</table>

1 n denotes the highest order of the time transformations of the model
2 Df denotes the number of parameter that was considered in fitting the model
3 AIC = Akaike’s Information Criterion
4 BIC = Bayesian Information Criterion
5 logLik = Log-likelihood
6 Chisq = χ² test statistic
7 Chi Df = Degree of freedom of the LRT
8 Pr(>Chisq) = p-value

All of the p-value are small and close to 0, which means that all the static predictors have statistically significant main effects in the random intercept models. In Tables 3.5 and 3.6 ‘Chisq’ indicates the χ² statistics in the LRT. χ² statistics are different over the static predictors. For mathematics, the gender variable, the cpc variable, the magnet variable, and the risk variable have χ² values of 33.47, 53.28, 81.70, and 105.93, respectively. For reading, the gender variable, the cpc variable, the magnet variable, and the risk variable have χ² values of 57.02, 57.71, 94.97, and 127.49, respectively. The model equations for both mathematics and reading can be written as, respectively,

\[
y_{ij} = \sum_{k=0}^{6} \beta_k t_{ij}^k + \beta_7 \cdot x_{1i} + \beta_8 \cdot x_{2i} + \beta_9 \cdot x_{3i} + \beta_{10} \cdot x_{4i} + b_{0i} + e_{ij}, \quad (3.5)
\]

\[
y_{ij} = \sum_{k=0}^{2} \beta_k t_{ij}^k + \beta_3 \cdot x_{1i} + \beta_4 \cdot x_{2i} + \beta_5 \cdot x_{3i} + \beta_6 \cdot x_{4i} + b_{0i} + e_{ij}, \quad (3.6)
\]

52
Table 3.4: LRT results for CLS reading data - time transformation

<table>
<thead>
<tr>
<th>n</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>122508.82</td>
<td>122531.10</td>
<td>-61251.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>98525.42</td>
<td>98555.13</td>
<td>-49258.71</td>
<td>23985.40</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>98364.07</td>
<td>98401.20</td>
<td>-49177.04</td>
<td>163.35</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>98363.56</td>
<td>98408.12</td>
<td>-49175.78</td>
<td>2.51</td>
<td>1</td>
<td>0.1128</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>97951.92</td>
<td>98003.91</td>
<td>-48968.96</td>
<td>413.63</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>97883.24</td>
<td>97942.66</td>
<td>-48933.62</td>
<td>70.68</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>97779.04</td>
<td>97845.88</td>
<td>-48880.52</td>
<td>106.20</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>97754.51</td>
<td>97828.77</td>
<td>-48867.26</td>
<td>26.53</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>97756.28</td>
<td>97837.97</td>
<td>-48867.14</td>
<td>0.23</td>
<td>1</td>
<td>0.6281</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>97734.67</td>
<td>97823.79</td>
<td>-48855.33</td>
<td>23.61</td>
<td>1</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 3.5: LRT results for CLS mathematics data - static predictors

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced model</td>
<td>9</td>
<td>93167.55</td>
<td>93234.36</td>
<td>-46574.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$l = 1$ (gender)</td>
<td>10</td>
<td>93136.08</td>
<td>93210.32</td>
<td>-46558.04</td>
<td>33.47</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 2$ (cpc)</td>
<td>10</td>
<td>93116.27</td>
<td>93190.50</td>
<td>-46548.14</td>
<td>53.28</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 3$ (magnet)</td>
<td>10</td>
<td>93087.86</td>
<td>93162.09</td>
<td>-46533.93</td>
<td>81.70</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 4$ (risk)</td>
<td>10</td>
<td>93063.62</td>
<td>93137.85</td>
<td>-46521.81</td>
<td>105.93</td>
<td>1</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

where $x_{1i}$, $x_{2i}$, $x_{3i}$ and $x_{4i}$ indicate the gender variable, the cpc variable, the magnet variable, and the risk variable, respectively.

Let us consider the relation between these $\chi^2$ statistics and the effect sizes for each main effect. If Equations (3.5) and (3.6) are the best fitting models for mathematics and reading, then it is expected that the $\chi^2$ values will be consistent with the $t$ values computed in the parameter estimates in Tables 3.7 and 3.8. The results for the $t$ values are consistent with the $\chi^2$ statistics obtained for the LRTs except for the one in which gender predicts reading. The conception of effect size will be discussed in detail in Chapter 4.

In Tables 3.7 and 3.8, all parameter estimates are negative, which was expected based on Figures 3.4 to 3.7. Since the reference group for the gender variable is female and
Table 3.6: LRT results for CLS reading data - static predictors

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced model</td>
<td>5</td>
<td>98364.07</td>
<td>98401.20</td>
<td>-49177.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$l = 1$ (gender)</td>
<td>6</td>
<td>98309.05</td>
<td>98353.61</td>
<td>-49148.52</td>
<td>57.02</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 2$ (cpc)</td>
<td>6</td>
<td>98308.36</td>
<td>98352.92</td>
<td>-49148.18</td>
<td>57.71</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 3$ (magnet)</td>
<td>6</td>
<td>98271.10</td>
<td>98315.66</td>
<td>-49129.55</td>
<td>94.97</td>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>$l = 4$ (risk)</td>
<td>6</td>
<td>98238.58</td>
<td>98283.14</td>
<td>-49113.29</td>
<td>127.49</td>
<td>1</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

the mean growth curve for the female group has a higher mean than the male group at all grades in Figure 3.4, the parameter estimates were expected to be negative. For the cpc variable, the reference group is the cpc participant group and the mean growth curve for the cpc participant group has a higher mean than the non-participant group at all grades, which indicates that the parameter estimates were expected to be negative. For the magnet variable, the reference group is the magnet school participant group and the reference group has a higher mean at all grades. Thus, it was expected that the parameter estimates for the magnet variable would be negative. For the risk variable, a high value for a student indicates that the student is at risk, which tells us that the risk variable might be negatively related with the response, which can also be seen in Figure 3.7.

Fitting Interaction Effects between Time Transformations and Static Predictors

As the last step in fitting the fixed effects, I add interaction terms between the time transformations and static predictors in the random intercepts model. When adding the interaction terms between time transformations and static predictors, we do not consider the higher degree time terms than selected in the previous steps. For example, the highest interaction terms for mathematics will be $x_{ij} \cdot t_i^6$ because the degree of the time transformations is 6. Since I only consider well-formulated models, the model always includes all lower interaction terms. In practice, applied researchers often avoid selecting the cubic or higher order
interaction terms between time transformations and static predictors. In fitting the models, I limited the degree of interaction terms with main effects to 3\textsuperscript{rd} order interaction terms. This limitation is applied to the mathematics data but is not to the reading data, since the highest order in the time transformation is a 2\textsuperscript{nd} order polynomial.

The LRT was used to compare the two nested models of Equations (3.7) and (3.8). I do not consider the interaction terms according to each static predictor but the interaction terms between time transformations and the set of predictors selected at the previous step. The parentheses in Equations (3.7) and (3.8) indicate the set of static predictors. Therefore, the difference in parameters between the full model (Equation (3.8)) and the reduced model (Equation (3.7)) is 4.

The results are summarized in Tables (3.9) and (3.10). There are much larger $\chi^2$ values (188.99 in mathematics and 157.66 in reading) when we compare the model with no interaction terms ($q = 0$) with the model with linear interaction terms ($q = 1$), which tells us that it is most significant to add linear interaction terms with static predictor for both math

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>74.43079</td>
<td>1.27080</td>
<td>58.57022</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>9.70969</td>
<td>1.06546</td>
<td>9.11317</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>11.11536</td>
<td>1.28170</td>
<td>8.67235</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-6.07583</td>
<td>0.58547</td>
<td>-10.37763</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.29244</td>
<td>0.12348</td>
<td>10.46663</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.12170</td>
<td>0.00422</td>
<td>9.36097</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.00422</td>
<td>0.00045</td>
<td>9.36097</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-3.51686</td>
<td>0.63530</td>
<td>-5.53573</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-4.06201</td>
<td>0.66435</td>
<td>-6.11423</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-6.99092</td>
<td>1.08559</td>
<td>-6.43975</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.98574</td>
<td>0.20699</td>
<td>-9.59345</td>
</tr>
</tbody>
</table>

Table 3.7: Parameter estimates for Mathematics for the the model of Equation (3.5)
Table 3.8: Parameter estimates for Reading for the model of Equation (3.6)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>82.59708</td>
<td>1.37232</td>
<td>60.18781</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>11.70412</td>
<td>0.12366</td>
<td>94.64725</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.17365</td>
<td>0.01356</td>
<td>-12.80943</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-5.19612</td>
<td>0.68931</td>
<td>-7.53812</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-4.59088</td>
<td>0.72105</td>
<td>-6.36689</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-8.14033</td>
<td>1.17554</td>
<td>-6.92476</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-2.42209</td>
<td>0.22447</td>
<td>-10.79028</td>
</tr>
</tbody>
</table>

Reduced model \( y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{m=0}^{q} \left( \sum_{l=0}^{4} \beta_{p+l+m+1} \cdot x_{li} \right) t_{ij}^m + b_0 + e_{ij}, \) (3.7)

Full model \( y_{ij} = \sum_{k=0}^{p} \beta_k t_{ij}^k + \sum_{m=0}^{q+1} \left( \sum_{l=0}^{4} \beta_{p+l+m+1} \cdot x_{li} \right) t_{ij}^m + b_0 + e_{ij} \) (3.8)

where \( q = 0,1,2. \)

Table 3.9: Test result for interaction terms in Mathematics

<table>
<thead>
<tr>
<th>( q )</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>92941.49</td>
<td>93037.99</td>
<td>-46457.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>92760.50</td>
<td>92886.69</td>
<td>-46363.25</td>
<td>188.99</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>92741.02</td>
<td>92896.91</td>
<td>-46349.51</td>
<td>27.48</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>92722.85</td>
<td>92908.43</td>
<td>-46336.43</td>
<td>26.17</td>
<td>4</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

\(^1 q\) indicates the highest order of interaction terms between time transformations and the static predictors selected in the previous steps.

Since all model comparisons result in statistical significance, we consider adding the cubic interaction terms in mathematics and the quadratic interaction terms in reading. The parameter estimates for these selected models for both mathematics and reading are summarized in Tables 3.11 and 3.12. The model equations for mathematics and reading are as
Table 3.10: Test result for interaction terms in Reading

<table>
<thead>
<tr>
<th>q</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>98077.91</td>
<td>98144.75</td>
<td>-49029.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td>97928.25</td>
<td>98024.79</td>
<td>-48951.13</td>
<td>157.66</td>
<td>4</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>97920.52</td>
<td>98046.77</td>
<td>-48943.26</td>
<td>15.73</td>
<td>4</td>
<td>0.0034</td>
</tr>
</tbody>
</table>

follows,

\[
y_{ij} = \sum_{k=0}^{6} \beta_k t_{ij}^k + \sum_{m=0}^{3} \left( \sum_{l=0}^{4} \beta_{6+l+m+1} \cdot x_{li} \right) t_{ij}^m + b_0 + e_{ij}, \quad (3.9)
\]

\[
y_{ij} = \sum_{k=0}^{2} \beta_k t_{ij}^k + \sum_{m=0}^{2} \left( \sum_{l=0}^{4} \beta_{2+l+m+1} \cdot x_{li} \right) t_{ij}^m + b_0 + e_{ij}, \quad (3.10)
\]

where \( q = 0, 1, 2 \).
Table 3.11: Parameter estimates for mathematics with $q = 3$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time transformation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>69.21437</td>
<td>1.46145</td>
<td>47.36023</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>8.74826</td>
<td>1.34422</td>
<td>6.50808</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>12.03539</td>
<td>1.28792</td>
<td>9.34484</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-6.15658</td>
<td>0.57926</td>
<td>-10.62829</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.29361</td>
<td>0.12212</td>
<td>10.59262</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.12193</td>
<td>0.01201</td>
<td>-10.14946</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.00423</td>
<td>0.00045</td>
<td>9.49523</td>
</tr>
<tr>
<td>Main effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-2.26790</td>
<td>0.73522</td>
<td>-3.08466</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-5.44686</td>
<td>0.76850</td>
<td>-7.08769</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-3.55945</td>
<td>1.26424</td>
<td>-2.81549</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.50552</td>
<td>0.23996</td>
<td>-6.27400</td>
</tr>
<tr>
<td>Linear interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.42823</td>
<td>0.44499</td>
<td>0.96233</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>1.18162</td>
<td>0.46750</td>
<td>2.52755</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-0.48534</td>
<td>0.74198</td>
<td>-0.65412</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>0.18251</td>
<td>0.14447</td>
<td>1.26326</td>
</tr>
<tr>
<td>Quadratic interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{15}$</td>
<td>-0.28300</td>
<td>0.12186</td>
<td>-2.32241</td>
</tr>
<tr>
<td>$\beta_{16}$</td>
<td>-0.21756</td>
<td>0.12822</td>
<td>-1.69677</td>
</tr>
<tr>
<td>$\beta_{17}$</td>
<td>-0.29233</td>
<td>0.20220</td>
<td>-1.44578</td>
</tr>
<tr>
<td>$\beta_{18}$</td>
<td>-0.10652</td>
<td>0.03964</td>
<td>-2.68728</td>
</tr>
<tr>
<td>Cubic interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{19}$</td>
<td>0.02299</td>
<td>0.00903</td>
<td>2.54740</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>0.01175</td>
<td>0.00951</td>
<td>1.23579</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.03437</td>
<td>0.01496</td>
<td>2.29688</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>0.00812</td>
<td>0.00294</td>
<td>2.75999</td>
</tr>
</tbody>
</table>
Table 3.12: Parameter estimates for reading with $q = 2$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time transformation</td>
<td>$\beta_0$ 76.02779</td>
<td>1.57040</td>
<td>48.41289</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$ 13.96250</td>
<td>0.47845</td>
<td>29.18272</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$ -0.27533</td>
<td>0.05255</td>
<td>-5.23979</td>
</tr>
<tr>
<td>Main effect</td>
<td>$\beta_3$ -2.41923</td>
<td>0.79323</td>
<td>-3.04984</td>
</tr>
<tr>
<td></td>
<td>$\beta_4$ -5.69423</td>
<td>0.82969</td>
<td>-6.86306</td>
</tr>
<tr>
<td></td>
<td>$\beta_5$ -6.51715</td>
<td>1.36112</td>
<td>-4.78808</td>
</tr>
<tr>
<td></td>
<td>$\beta_6$ -1.43415</td>
<td>0.25865</td>
<td>-5.54480</td>
</tr>
<tr>
<td>Linear interaction</td>
<td>$\beta_7$ -1.09465</td>
<td>0.24719</td>
<td>-4.42848</td>
</tr>
<tr>
<td></td>
<td>$\beta_8$ 0.72696</td>
<td>0.25995</td>
<td>2.79655</td>
</tr>
<tr>
<td></td>
<td>$\beta_9$ -0.52419</td>
<td>0.41236</td>
<td>-1.27118</td>
</tr>
<tr>
<td></td>
<td>$\beta_{10}$ -0.35966</td>
<td>0.08044</td>
<td>-4.47103</td>
</tr>
<tr>
<td>Quadratic interaction</td>
<td>$\beta_{11}$ 0.06282</td>
<td>0.02711</td>
<td>2.31742</td>
</tr>
<tr>
<td></td>
<td>$\beta_{12}$ -0.07344</td>
<td>0.02852</td>
<td>-2.57495</td>
</tr>
<tr>
<td></td>
<td>$\beta_{13}$ 0.02558</td>
<td>0.04511</td>
<td>0.56698</td>
</tr>
<tr>
<td></td>
<td>$\beta_{14}$ 0.01771</td>
<td>0.00883</td>
<td>2.00591</td>
</tr>
</tbody>
</table>
3.2.2 Step 2 - Adding Random Effects

The goal of Step 2 is to add random effects terms. As discussed in Chapter 2, the parameters for random effects are not \( b_i \)'s in Equation (2.21) but the variances and covariances in Equation (2.19). However, the variance-covariance structure is defined by \( b_i \)'s if we consider the “unstructured” variance-covariance structure of the LMM. This is the only type of structure considered in this paper. Since the column space of \( Z_i \) in Equation (2.2) is a subspace of the column space of \( X_i \) in Equation (2.2), the number of \( b_i \)'s are bounded by the degree of time transformations selected in the fixed effects part.

As mentioned previously, I only consider well-formulated models. Thus, if \( b_q \) is selected, then \( b_0, \ldots, b_{q-1} \) are in the model. In this paper, I limit the number of \( b_i \)'s to 4, since applied researchers rarely consider the cubic or higher degree random effects. Thus, all possible cases are

\[
\begin{align*}
\text{random – intercept} & \quad \text{including } b_0, \\
\text{linear} & \quad \text{including } b_0, b_1, \\
\text{quadratic} & \quad \text{including } b_0, b_1, b_2, \\
\text{cubic} & \quad \text{including } b_0, b_1, b_2, b_3.
\end{align*}
\]

The LRT was used to compare the following two nested models (Equations (3.11) and (3.12)). In the full model (Equation (3.12)), there is one additional \( b_{r+1} \) term. As you can see in Equation (2.19), the full model, then, includes \( r + 1 \) more parameters than the reduced model. Thus, the degrees of freedom in the LRT increase by 2, 3, and 4 in Tables
3.13 and 3.14, in which the results are shown.

Reduced model \( y_{ij} = \sum_{k=0}^{p} \beta_{kt} \cdot t_{ij} + \sum_{m=0}^{q} \left( \sum_{l=0}^{4} \beta_{p+l+m+1} \cdot x_{li} \right) \cdot t_{ij}^m + \sum_{n=0}^{r} b_{nit} + e_{ij} \), \hspace{1cm} (3.11)

Full model \( y_{ij} = \sum_{k=0}^{p} \beta_{kt} \cdot t_{ij} + \sum_{m=0}^{q} \left( \sum_{l=0}^{4} \beta_{p+l+m+1} \cdot x_{li} \right) \cdot t_{ij}^m + \sum_{n=0}^{r+1} b_{nit} + e_{ij} \), \hspace{1cm} (3.12)

where \( r = 0, 1, 2 \).

The results of the LRTs are summarized in Tables 3.13 and 3.14. All test results show statistical significance. Thus, I select cubic random effects for mathematics and quadratic random effects for reading. The best fitting models are given in Equations (3.13) and (3.14).

Table 3.13: Test result for random effects terms in Mathematics

<table>
<thead>
<tr>
<th>( r )</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>25</td>
<td>92722.85</td>
<td>92908.43</td>
<td>-46336.43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>27</td>
<td>90892.06</td>
<td>91092.48</td>
<td>-45419.03</td>
<td>1834.79</td>
<td>2</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>90502.17</td>
<td>90724.86</td>
<td>-45221.08</td>
<td>395.89</td>
<td>3</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>90266.61</td>
<td>90518.99</td>
<td>-45099.30</td>
<td>243.56</td>
<td>4</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

\( ^1 r \) indicates the highest degree of random effects

Table 3.14: Test result for random effects terms in Reading

<table>
<thead>
<tr>
<th>( r = 0 )</th>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>97920.52</td>
<td>98046.77</td>
<td>-48943.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( r = 1 )</td>
<td>19</td>
<td>96158.52</td>
<td>96299.63</td>
<td>-48060.26</td>
<td>1765.99</td>
<td>2</td>
<td>0.0000</td>
</tr>
<tr>
<td>( r = 2 )</td>
<td>22</td>
<td>96056.73</td>
<td>96220.11</td>
<td>-48006.36</td>
<td>107.80</td>
<td>3</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
3.3 Parameter estimates

The best fitting models for both the mathematics and reading data sets can be written as

\[
\text{Math}_{ij} = \sum_{k=0}^{6} \beta_{kj} t_{ij}^k + \sum_{m=0}^{3} \left( \sum_{l=0}^{4} \beta_{6+l+m+1} \cdot x_{li} \right) t_{ij}^m + \sum_{n=0}^{3} b_{nj} t_{ij}^n + e_{ij},
\]

(3.13)

\[
\text{Read}_{ij} = \sum_{k=0}^{2} \beta_{kj} t_{ij}^k + \sum_{m=0}^{2} \left( \sum_{l=0}^{4} \beta_{2+l+m+1} \cdot x_{li} \right) t_{ij}^m + \sum_{n=0}^{2} b_{nj} t_{ij}^n + e_{ij},
\]

(3.14)

where \( x_{li} \) has \( l = 1, \cdots, 4 \) indicating the gender, cpc, magnet and risk variables, respectively. The parameter estimates are listed in Tables 3.15 and 3.16 for mathematics and reading, respectively. The results are slightly different from those in Tables 3.11 and 3.12 due to the random effects parts. These parameter estimates will be used in simulating response data from the data distribution corresponding to the fitted model objects. Generating data will be discussed in detail in Chapter 4.
Table 3.15: Parameter estimates for Mathematics

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time transformation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>69.12425</td>
<td>1.39317</td>
<td>49.61640</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>8.86521</td>
<td>1.23223</td>
<td>7.19446</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>12.16875</td>
<td>0.99866</td>
<td>12.18506</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-6.27614</td>
<td>0.44366</td>
<td>-14.14622</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.32717</td>
<td>0.09353</td>
<td>14.18911</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.12592</td>
<td>0.00920</td>
<td>-13.68062</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.00441</td>
<td>0.00034</td>
<td>12.89624</td>
</tr>
<tr>
<td>Main effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-2.29623</td>
<td>0.70118</td>
<td>-3.27478</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>-5.37690</td>
<td>0.73290</td>
<td>-7.33644</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-3.55000</td>
<td>1.20603</td>
<td>-2.94355</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-1.48237</td>
<td>0.22883</td>
<td>-6.47791</td>
</tr>
<tr>
<td>Linear interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{11}$</td>
<td>0.57645</td>
<td>0.48513</td>
<td>1.18825</td>
</tr>
<tr>
<td>$\beta_{12}$</td>
<td>1.20882</td>
<td>0.50874</td>
<td>2.37612</td>
</tr>
<tr>
<td>$\beta_{13}$</td>
<td>-0.66940</td>
<td>0.81746</td>
<td>-0.81888</td>
</tr>
<tr>
<td>$\beta_{14}$</td>
<td>0.15003</td>
<td>0.15772</td>
<td>0.95121</td>
</tr>
<tr>
<td>Quadratic interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{15}$</td>
<td>-0.33505</td>
<td>0.12413</td>
<td>-2.69915</td>
</tr>
<tr>
<td>$\beta_{16}$</td>
<td>-0.23867</td>
<td>0.13041</td>
<td>-1.83015</td>
</tr>
<tr>
<td>$\beta_{17}$</td>
<td>-0.20451</td>
<td>0.20795</td>
<td>-0.98347</td>
</tr>
<tr>
<td>$\beta_{18}$</td>
<td>-0.10042</td>
<td>0.04048</td>
<td>-2.48093</td>
</tr>
<tr>
<td>Cubic interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_{19}$</td>
<td>0.02694</td>
<td>0.00909</td>
<td>2.96482</td>
</tr>
<tr>
<td>$\beta_{20}$</td>
<td>0.01362</td>
<td>0.00956</td>
<td>1.42453</td>
</tr>
<tr>
<td>$\beta_{21}$</td>
<td>0.02588</td>
<td>0.01518</td>
<td>1.70437</td>
</tr>
<tr>
<td>$\beta_{22}$</td>
<td>0.00755</td>
<td>0.00297</td>
<td>2.54517</td>
</tr>
</tbody>
</table>

Table 3.16: Variance components for Mathematics

<table>
<thead>
<tr>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>148.42</td>
<td>-47.16</td>
<td>11.08</td>
<td>-0.72</td>
<td>43.53</td>
</tr>
<tr>
<td>-47.16</td>
<td>42.96</td>
<td>-9.22</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>11.08</td>
<td>-9.22</td>
<td>2.29</td>
<td>-0.16</td>
<td></td>
</tr>
<tr>
<td>-0.72</td>
<td>0.59</td>
<td>-0.16</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td></td>
<td></td>
<td></td>
<td>43.53</td>
</tr>
</tbody>
</table>
Table 3.17: Parameter estimates for Reading

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time transformation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>76.16987</td>
<td>1.22501</td>
<td>62.17882</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>13.79263</td>
<td>0.50905</td>
<td>27.09468</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.24725</td>
<td>0.05465</td>
<td>-4.52383</td>
</tr>
<tr>
<td><strong>Main effect</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>-2.34559</td>
<td>0.61929</td>
<td>-3.78757</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>-5.64832</td>
<td>0.64806</td>
<td>-8.71571</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>-6.62972</td>
<td>1.06168</td>
<td>-6.24457</td>
</tr>
<tr>
<td>( \beta_6 )</td>
<td>-1.44055</td>
<td>0.20189</td>
<td>-7.13544</td>
</tr>
<tr>
<td><strong>Linear interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_7 )</td>
<td>-1.13019</td>
<td>0.26311</td>
<td>-4.29556</td>
</tr>
<tr>
<td>( \beta_8 )</td>
<td>0.73864</td>
<td>0.27659</td>
<td>2.67053</td>
</tr>
<tr>
<td>( \beta_9 )</td>
<td>-0.40419</td>
<td>0.43890</td>
<td>-0.92092</td>
</tr>
<tr>
<td>( \beta_{10} )</td>
<td>-0.34738</td>
<td>0.08546</td>
<td>-4.06467</td>
</tr>
<tr>
<td><strong>Quadratic interaction</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \beta_{11} )</td>
<td>0.06292</td>
<td>0.02825</td>
<td>2.22700</td>
</tr>
<tr>
<td>( \beta_{12} )</td>
<td>-0.07980</td>
<td>0.02973</td>
<td>-2.68378</td>
</tr>
<tr>
<td>( \beta_{13} )</td>
<td>0.01044</td>
<td>0.04686</td>
<td>0.22281</td>
</tr>
<tr>
<td>( \beta_{14} )</td>
<td>0.01447</td>
<td>0.00918</td>
<td>1.57612</td>
</tr>
</tbody>
</table>

Table 3.18: Variance components for Reading

<table>
<thead>
<tr>
<th></th>
<th>( b_0 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( \sigma^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0 )</td>
<td>88.91</td>
<td>2.18</td>
<td>0.03</td>
<td>83.67</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>2.18</td>
<td>8.14</td>
<td>-0.61</td>
<td></td>
</tr>
<tr>
<td>( b_2 )</td>
<td>0.03</td>
<td>-0.61</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 4

Methods and Results

To show the efficiency of model building between the step up and the top down approaches, I generated data sets by simulating responses, mathematics and reading test scores, from the distribution corresponding to the best fitting models. In other words, the distribution of the random effects and error was multivariate normal and values based on this distribution were generated.

I first simulated response data under the best fitting models for both mathematics and reading (see Equations (3.13) and (3.14)), by using the parameter estimates of the best fitting models. In this generating process, all variables except the response variables were fixed. That is, each subject had the same predictor values as in the original data except for the response data. As a result, the generated data included the same missing portion as the original CLS data, and each subject had the same static predictors as in the original CLS data and the number of measurements for each subject were the same as the original CLS data. This step was performed by using the `simulate()` function in the `stats` package implemented in R.

The parameter estimates obtained from the best fitting models, Equations (3.13) and (3.14), were used as population parameters for generating the sample data. Thus, from now on, the parameter estimates obtained in Chapter 3 will be considered and used as population parameters. In order to investigate the efficiency of model building with the generated data
sets, I sampled the simulation subjects from the 1531 subjects in the generated response data with different sample sizes, 100, 300 and 500. Then in each replication I fitted the LMMs to the sample data sets.

The results that are discussed in this chapter are based on the fitted models on these sample data sets. The design of the simulation is discussed in detail in the Section 4.1, the criteria of measuring efficiency is discussed in Section 4.2, and the results are described in Section 4.3.

4.1 Design of the Simulation

Based on the population parameters listed in Tables 3.15 to 3.18, 2000 sets of 1531 response data were generated. Then I sampled 100, 300 and 500 simulation subjects from each generated data set of 1531. When a subject was sampled, their entire vector of repeated measures was selected. The sample sizes, 100, 300, and 500 out of 1531 subjects constitute 6.5%, 19.6%, and 32.7%, respectively of the total initial simulated subjects. Results based on these three sample sizes can be used to show the efficiency of model selection in terms of sample size. In this simulation, all variables except the response variables were fixed when the data sets were generated. In other words, the design matrix for an individual was the same as in the true model.

Both the ratios of number of parameters to subject, and number of parameters to total possible data points listed in Table 4.1. The reason that I mention “possible” data points instead of exact data points is that the data points vary in terms of subjects who were selected in each sample. For example, if a student is selected and has only 5 data points in the original data, he/she has only 5 data points in the generated data, whereas if a student is selected and has a full 10 data points in the original data, he/she also has a full 10 data points in the generated data.

The same design was applied to both the mathematics and reading data. Two-thousand
Table 4.1: Ratio of sample size and the number of parameters

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Subjects</th>
<th>The number of parameter Mathematics (=34)</th>
<th>The number of parameter Reading (=22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
<td>2.94</td>
<td>4.55</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>8.82</td>
<td>13.64</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>14.71</td>
<td>22.73</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>29.41</td>
<td>45.45</td>
</tr>
<tr>
<td>3000</td>
<td></td>
<td>88.24</td>
<td>136.36</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>147.06</td>
<td>227.27</td>
</tr>
</tbody>
</table>

data sets were fitted with the LMMs whose time transformations are polynomials. As seen in Equations (3.13) and (3.14), the CLS mathematics data had a $6^{th}$ order polynomial model while the CLS reading data had a $2^{nd}$ order polynomial model. In this way, I investigated which model building approach was best in terms of selection of polynomials.

In the following subsection, I discuss how I recorded the results in the simulation study. Briefly, for each method (step up, top down), I recorded the highest order terms in the fixed effects and the random effects portions of the models.

**Nature of the Results**

All results of fitting the LMMs to the 2000 data sets were recorded separately in terms of fixed effects and random effects. More precisely, the result of the fixed effects part for each data set was recorded with the three different parts: The time transformations, the main effects of the static predictors, and the interactions between the time transformations and the main effects.

Since there were 10 time points from kindergarten to $9^{th}$ grade in the CLS data, the degree of time transformations ranged from 0 to 9. Based on the LRT test results, the degree of time transformations for each sampled data was recorded. In this paper, the degree is denoted as the number following $P$. For example, the $6^{th}$ order polynomial model
is denoted as $P_6$.

Among four static predictors, gender, cpc, magnet and risk index, I assigned the subindex for each combination of main effects selected, which is listed in Table 4.2. The numbers range from 0 to 4 indicating how many static predictors were selected. The combination and the number of static predictors selected are denoted as the numbers following $M$. For instance, consider $M_{102}$. The subindex 10 means that cpc and risk index were selected in the fitted model and 2 means that two of four static variables were selected.

Table 4.2: Index for selection on main effects

<table>
<thead>
<tr>
<th>Subindex</th>
<th>Static predictor</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gender</td>
<td>cpc</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The degree of interaction notation varies from 0 to 3. Based on the selection of main effects, the interaction part was considered as the interaction between the set of main effects selected and the time transformations. For example, if the gender and the risk predictors were selected as main effects, the interaction between \{gender, risk\} and grade (or interac-
tions with $grade^2$, $grade^3$) were considered in the upcoming model comparisons. I denote the interaction term with $I$ whose number following $I$ denotes the degree of interaction effects. Therefore, $I2$ indicates, for example, that the interaction between the quadratic time transformation and main effects was selected.

The degree of the random effects part was determined by the number of $b_i$s and it varies from 0 to 3. As illustrated in Equation (2.19), the degree directly tells us how many variables in the variance-covariance structure there are. For example, when the degree is 2, this means that we have $b_{i0}$, $b_{i1}$ and $b_{i2}$, and the parameters $Var(b_{i0})$, $Var(b_{i1})$, $Var(b_{i2})$, $Cov(b_{i0}, b_{i1})$, $Cov(b_{i0}, b_{i2})$ and $Cov(b_{i1}, b_{i2})$. Thus, 1, 3, 6, and 10 parameters were fitted according to the random-intercept, linear, quadratic, and cubic random effects models, respectively. I denote the random effects with $R$. For example, $R2$ indicates that the quadratic random effects was the highest order term selected.

In summary, I use the following notation to describe the best fitted model for each replication, which is illustrated with an example. Suppose the best fitting model had a $6^{th}$ order polynomial with cubic interactions between the time transformations and all four of the static predictors and the cubic random effects part. The resulting model is denoted as

$$\overset{P6}{\text{Time transformation}} \overset{M_{164}}{\text{Main effects}} \overset{I3}{\text{Interaction}} \overset{R3}{\text{Random effects}}.$$ (4.1)

Since I consider only well-formulated models, the lower order terms for each part were always included in the fitted model. That is, the model (4.1) always included a constant, the linear, quadratic, cubic, $4^{th}$ order, and $5^{th}$ order polynomial terms in the time transformations, the linear and the quadratic interactions, and the linear and the quadratic random effects terms.
4.2 Classification Criteria

To investigate the efficiency of the model building approaches, I will examine which approach most frequently identifies the true model. Identification means that the structure of the best fitting model is the same structure as the true model with the same polynomial transformations, the same main effects and interactions, and the same random effects. In this simulation, the identification did not work very well due to additional error that is not generally counted in a simulation study. Since I fitted a LMM to each simulated sample data set, there is additional error introduced by sampling data from each generated data set. In other words, there are two errors in this simulation. The first error happens when I generate 1531 response data and the second happens when I sample 100, 300, or 500 simulated subjects from the 1531 generated response data. That is, the second error causes additional variability and such sampling error negatively affects the selection rate. To compensate for this dispersion caused by such sampling error, I define an identification by finding an “approximate” model to the true model in this paper. The similarity between an approximate model and the true model will be defined and discussed in the following subsection. Thus, results for investigation on the efficiency of the step up and the top down will be based on the selection of approximate models.

In addition, I investigate the impact of the effect sizes of static predictors on the selection of the interaction effects between time transformations and static predictors. In the model building procedure, the set of main effects selected was considered in constructing the interaction effects with the time transformations. That is, I tested all static predictors individually when I selected the main effects. Then, the set of selected static predictors was considered when I constructed the interaction effects between the set of static predictors and the time transformations. I do not examine which model building approach better identifies the structures of main effects and interaction effects in the true model but consider a “total effect” size by quantifying effects sizes of all main effects selected and their
interaction by considering the Euclidean distance. In the following two subsections, I will discuss what is meant by “similarity” and “total effect” regarding the model selection in detail.

### 4.2.1 Similarity

Similarity to the true model is defined as follows: If the best fitting model for each simulated sample data set has the same structure in the time transformations and main effects as the true model, the best fitting model is called an approximate model to the true model. For example, if a fitted model includes the 6th order time transformation and all four main effects for the CLS math data, the model is approximate to the true model.

As another illustration, I consider $P_{6M_{16}4I0R0}$ and $P_{6M_{16}4I3R3}$ to be approximate models to the true model $P_{6M_{16}4I3R3}$ for the Mathematics CLS data, since both models have the 6th order polynomial transformation and include the four main effects that are the same as the true model. According to the number of parameters, the first model, $P_{6M_{16}4I0R0}$, includes all 13 parameters consisting of 11 for the fixed effects, 1 for the random effects, and 1 for the error part. The second model, $P_{6M_{16}4I3R3}$, includes 34 parameters consisting of 23 fixed effects, 10 random effects, and 1 random error parameter. The difference in parameters between the two models is 21, which is the greatest number between the approximate models and the true model in this simulation study.

In contrast to the difference in the number of parameters between approximate models and the true models, their predicted curves are similar. To show the graphical similarity, I consider the approximate model for mathematics and reading whose number of parameters is the smallest among approximate models. The approximate model for the mathematics data is $P_{6M_{16}4I0R0}$ while the approximate model for the reading data is $P_{2M_{16}4I0R0}$. To obtain the fitted curves for the approximate model and the true model, I fitted them to mathematics and reading data. In Figures 4.1 to 4.16, I obtained three different curves in
the same axes such as the true model, the approximate model having the smallest number of parameters among approximate models, and the mean growth curve conditioned on the statics predictors for both mathematics and reading. Since the gender, cpc, and magnet variables are binary, I consider eight different groups that are classified by gender, cpc, and magnet. For example, in Figure 4.1, we can see the similarity between the true model and the smallest model conditioned on female, cpc attendance, and magnet school. On the other hand, since the risk variable is continuous, there is no such group as the gender, cpc, and magnet variables. Instead of considering any group in the risk variable, I used the mean value based on the entire 1531 subjects’ in Figures 4.1 to 4.16. Among the eight different figures for mathematics and reading, there is slight graphical discrepancy between the approximate model and the true model in two figures for mathematics (Figures 4.1 and 4.3(a)) and in one figure for reading (Figure 4.5). In Section 4.3, I will discuss how efficient the step up and the top down performed in model selection.

Mathematics

Figure 4.1 shows the approximate model curve, the true model curve, and the mean growth curve for the female, cpc attendance and magnet school attendance group fixing risk at its mean value. The curve for the approximate model and the true model are alike. Both the approximate model and the true model are higher than mean growth curve over all time points. On the other hand, the approximate model curve is slightly higher from kindergarten to 3rd grade, whereas the true model curve is slightly higher from 5th grade to 8th grade. The curves are identical at 4th and 9th grades. This crossing is due to the presence of interaction terms in the true model.
Figure 4.1: Comparison between the approximate model and the true model for Mathematics for female, cpc, magnetic and mean risk group.
Figure 4.2: Comparison between the approximate model and the true model for Mathematics

Figure 4.2 shows the three curves for the female, cpc attendance, and non-magnet school attendance group. The approximate model curve and the true model curve are slightly higher than the mean growth curve. But, the approximate model and the true model are almost identical. Even though I looked at eight different figures according to groups classified by the static predictors, six figures indicate that the approximate model curve is almost identical to the true model curve (Figures 4.2, 4.3(b), 4.4). In two figures (Figures 4.1 and 4.3(a)), there are slight discrepancies but the curves are fairly close.

Figure 4.3 (a) shows the three curves for the female, non-cpc attendance, and magnet school attendance group. Both the approximate model and the true model are higher than the mean growth curve. Similar to Figure 4.1, the approximate model is higher from kindergarten to 3rd grade, whereas the true model is higher from 3rd grade to 9th grade.
Figure 4.3: Comparison between the approximate model and the true model for Mathematics

(a) Female, non-CPC, Magnet

(b) Female, non-CPC, non-Magnet

Both figures 4.1 and 4.3(a) occur when groups attending magnet school are considered. Figure 4.3 (b) shows the three curves for the female, non-cpc attendance, and non-magnet school attendance group. Both the approximate model and the true model are lower than the mean growth curve. However, the approximate model and the true model are almost identical, graphically.

Figures 4.1 to 4.3 include the three curves for female. Figure 4.4 includes four different figures of the three curves for male. In all the graphs of Figure 4.4, the approximate model curve and the true model are nearly identical.
Figure 4.4: Comparison between the approximate model and the true model for Mathematics

Mathematics for male, cpc, magnetic and mean risk group

Mathematics for male, cpc, non-magnetic and mean risk group

Mathematics for male, non-cpc, magnetic and mean risk group

Mathematics for male, non-cpc, non-magnetic and mean risk group
Figure 4.5: Comparison between the approximate model and the true model for Reading

Figure 4.5 shows the most different approximate model curve, the true model curve, and the mean growth curve for the female, cpc attendance and magnet school attendance group, with risk fixed at its mean value. Both the approximate model and the true model are higher than the mean growth curve over the 10 time points. On the other hand, the approximate model is slightly higher from kindergarten to 2nd grade, whereas the true model is slightly higher from 5th grade to 8th grade.

Figure 4.6 shows the three curves for the female, cpc attendance, and non-magnet school attendance group. The approximate model curve and the true model curve are slightly higher than the mean growth curve. But, the approximate model and the true model are almost identical. Even though I looked at eight different figures according to
Figure 4.6: Comparison between the approximate model and the true model for Reading

For female, cpc, non-magnetic and mean risk group grades, Figure 4.6, 4.7, and 4.8 illustrate that the approximate model curve is very close to the true model curve. In one figure (Figure 4.5), there were slight discrepancies, but the curves were fairly close.

Figure 4.7 (a) and (b) show the three curves for the female, non-cpc attendance, and magnet school attendance group. Both the approximate model curve and the true model curve are almost identical, graphically. Since the true and approximate models are of degree 2 polynomials, they look very different from the mean growth curve.

Figures 4.5 to 4.7 show the three curves for female. Figure 4.8 shows four different figures of the three curves for male. In all figures in Figure 4.8, the approximate model curve and the true model curve are almost identical.
Figure 4.7: Comparison between the approximate model and the true model for Reading

(a) Female, non-CPC, Magnet

(b) Female, non-CPC, non-Magnet
Figure 4.8: Comparison between the approximate model and the true model for Reading

Reading for male, cpc, magnetic and mean risk group

Reading for male, cpc, non-magnetic and mean risk group

Reading for male, non-cpc, magnetic and mean risk group

Reading for male, non-cpc, non-magnetic and mean risk group
4.2.2 Total effect

In addition to the graphical similarity criterion, I consider a statistical criterion to measure the effect sizes of the main effects and the interaction effects in each fitted model. To be consistent with the notation of the model selection procedure, I define the “total effect” (TE) for a model $M$ as

$$\text{TE}_M = \sqrt{\sum_{i=1}^{n} (\text{Effect size})^2_i} ;$$

where $n$ varies from 0 to 16 in this simulation. Since the highest degree of interaction terms was capped at 3rd order, we have up to four main effects and twelve interaction effects between time transformations and main effects. Therefore, $n$ varies from 0 to 16. Since the larger effect size for a variable indicates a stronger relationship between the variable and the response, the larger TE also indicates a stronger relationship between the variables and the response. That is, TE is a larger-is-better measure. TE does not measure the difference between parameter estimates and true parameter but measure how the fitted model is related to effect sizes in model selection.

I summarize the results of approximate models for both mathematics and reading in Tables 4.3 and 4.4. The differences over the structures of random effects are depicted in Figure 4.9. In Tables 4.3 and 4.4, each column indicates the degree of the interaction effects in the fitted model. For example, ‘Linear’ means that the fitted model includes up to linear interaction terms between the time transformations and the main effects. On the other hand, each row in Tables 4.3 and 4.4 indicates the degree of the random effects. For example, ‘Quad’ means that the fitting model includes two random effects.

In the simulation study, it is desirable to have the model with the larger total effect selected more often. For the purpose of investigation into the efficiency of model selection approaches with the LMM, it is no doubt that the higher rate of selecting the true model indicates the more efficient method. On the other hand, as we saw with the model building
procedure introduced in Chapter 2, the order of fitting the effects is, from first to last, the
time transformations, the main effects, the interaction effects between the time transforma-
tions and the main effects, and the random. Thus, it is possible that some models will not
be considered due to the order of fitting. For example, if two main effects of four static
predictors are selected, then the interactions between more than two main effects and the
time transformations will not be considered. In a similar fashion, if an interaction effect
between the static predictors and the time transformations is larger than in the true model,
then the model including the interaction effects can be selected more often. In both the
step up and the top down approaches, interaction terms were added before random effects
terms. We can expect that the most frequently identified interaction terms are the linear
ones for both mathematics and reading. The results will be discussed in detail in Section
4.3. In the following section, I summarize the results in terms of the two criteria, graphical
similarity and statistical total effects.

Table 4.3: Total effects for Mathematics

<table>
<thead>
<tr>
<th>Degree of random effects</th>
<th>Degree of interaction effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main</td>
</tr>
<tr>
<td>Sum of TEs</td>
<td>52.023</td>
</tr>
</tbody>
</table>
Table 4.4: Total effects for Reading

<table>
<thead>
<tr>
<th>Degree of random effects</th>
<th>Degree of interaction effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main</td>
</tr>
<tr>
<td>Linear</td>
<td>15.192</td>
</tr>
<tr>
<td>Quad</td>
<td>15.218</td>
</tr>
<tr>
<td>Sum of TEs</td>
<td>46.562</td>
</tr>
</tbody>
</table>

Figure 4.9: Total Effect

(a) Mathematics

(b) Reading
4.3 Results of Similarity

Since the true models are 6th order polynomials and 2nd order polynomials for mathematics and reading, respectively, we may expect that the top down approach will perform better for mathematics than the step up approach, and the step up approach is expected to perform better for reading than the top down approach. However, the results shows that the step up approach performed better than the top down approach for all models. In addition, I can see that the discrepancy of performance between the step up and the top down is larger when the true model is a relatively low order polynomial model. That is, the proportion of selecting approximate models in reading is greater than in mathematics. The results are summarized in Tables 4.5 and 4.6 and in Figure 4.10.

As the sample size increases, the proportion of selecting approximate models also increases. The increment from 100 to 300 simulation subjects is much greater than that from 300 to 500 simulation subjects on both mathematics and reading. For the 100 sample size, the performance of both the step up and top down methods are very poor in terms of selecting a good approximating model for both response variables. However, for samples consisting of more than 300 subjects, in more than 70% of the simulations, a good approximate model is selected with both the step up and top down approaches (see Figure 4.10).

Figure 4.10 also shows that the step up approach performed better than the top down approach regardless the sample size. For the step up approach with sample size of 300, the approximate model was selected in 73.1% and 38% of the samples for mathematics and reading, respectively. These proportions were greater than those of the top down approach (71.5% and 73.0%, respectively). For the step up approach with sample size of 500, the approximate models were selected in 92.5% and 96.2% of the samples for mathematics and reading, respectively. These proportions were much greater than those of the top down approach (88.4% and 84.2%, respectively).
Table 4.5: Proportions of selected approximate models in mathematics

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Step Up</th>
<th>Top Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.036</td>
<td>0.077</td>
</tr>
<tr>
<td>300</td>
<td>0.731</td>
<td>0.715</td>
</tr>
<tr>
<td>500</td>
<td>0.925</td>
<td>0.884</td>
</tr>
</tbody>
</table>

Table 4.6: Proportions of selected approximate models in reading

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Step Up</th>
<th>Top Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.143</td>
<td>0.122</td>
</tr>
<tr>
<td>300</td>
<td>0.830</td>
<td>0.730</td>
</tr>
<tr>
<td>500</td>
<td>0.962</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Figure 4.10: Proportion of selected approximate models

(a) Mathematics

(b) Reading
### 4.4 Results for Total Effect

Based on the true models for both mathematics and reading, I summarize the total effects in Tables 4.3 and 4.4. According to the criterion that the model with higher total effects is selected more often, we expected the models having a linear interaction term between the time transformations and the main effects to be more frequently selected (see Tables 4.3 and 4.4). The simulation results are summarized in Tables 4.7 to 4.10. Each column indicates the degree of interaction terms, and each row indicates the sample size of the sampled data in Tables 4.7 to 4.10.

For mathematics, the fitted models having a linear interaction term were the most frequently selected regardless of the sample size or the model building approach. This result tells that if a model has the highest total effects, then the model is the most frequently selected but not the same as the true model. One the other hand, the result also indicates that the proportion of selecting models having a linear interaction term decreased as the sample size increases. Whereas the proportion of selecting models having a cubic interaction term increased as the sample size increases. The model having only main effects was rarely selected, even though the total effect is not close to zero. The result for the model having only main effects seems to be odd but it can be explained by looking at the model building procedure. Since the model having a linear interaction term has highest TE, the model having only main effects is rarely selected in the model comparison with the model having the linear interaction term whose TE is the highest (see Table 4.3).

Table 4.7: Proportion of selected interaction models for mathematics - Step Up

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Degree of interaction effects</th>
<th>Main</th>
<th>Linear</th>
<th>Quad</th>
<th>Cubic</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td></td>
<td>0.110</td>
<td>0.671</td>
<td>0.205</td>
<td>0.014</td>
</tr>
<tr>
<td>300</td>
<td></td>
<td>0.004</td>
<td>0.577</td>
<td>0.270</td>
<td>0.149</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.000</td>
<td>0.395</td>
<td>0.234</td>
<td>0.372</td>
</tr>
</tbody>
</table>
Table 4.8: Proportion of selected interaction models for mathematics - Top Down

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Degree of interaction effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main</td>
</tr>
<tr>
<td>100</td>
<td>0.065</td>
</tr>
<tr>
<td>300</td>
<td>0.004</td>
</tr>
<tr>
<td>500</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Figure 4.11: Proportion of selected interaction models - Mathematics

(a) Step Up
(b) Top Down
For reading, the fitted models having a linear interaction term were the most frequently selected regardless of the sample or the model building approach. This result tells that if a model has the highest total effects, then the model is the most frequently selected but not the same as the true model. The result also indicates that the proportion of selecting models having a linear interaction term decreased as the sample size increased. In contrast, the proportion of selecting models having a quadratic interaction term increased as the sample size increased. The model having only main effects was rarely selected, even though the total effect is not close to zero. The result for the model having only main effects seems to be odd but it can be explained by looking at the model building procedure. Since the model having a linear interaction term has highest TE, the model having only main effects is rarely selected in the model comparison with the model having the linear interaction term whose TE is the highest (see Table 4.4).

Table 4.9: Proportion of selected interaction models for reading - Step Up

<table>
<thead>
<tr>
<th></th>
<th>Main</th>
<th>Linear</th>
<th>Quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.084</td>
<td>0.798</td>
<td>0.118</td>
</tr>
<tr>
<td>300</td>
<td>0.010</td>
<td>0.720</td>
<td>0.270</td>
</tr>
<tr>
<td>500</td>
<td>0.001</td>
<td>0.572</td>
<td>0.427</td>
</tr>
</tbody>
</table>

Table 4.10: Proportion of selected interaction models for reading - Top Down

<table>
<thead>
<tr>
<th></th>
<th>Main</th>
<th>Linear</th>
<th>Quad</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.098</td>
<td>0.776</td>
<td>0.127</td>
</tr>
<tr>
<td>300</td>
<td>0.008</td>
<td>0.720</td>
<td>0.273</td>
</tr>
<tr>
<td>500</td>
<td>0.001</td>
<td>0.571</td>
<td>0.428</td>
</tr>
</tbody>
</table>
Figure 4.12: Proportion of selected interaction models - Reading

(a) Step Up

(b) Top Down
Chapter 5

Findings and Conclusions

To investigate the efficiency of the model building approaches in the LMM, I conducted the simulation study based on the longitudinal data of the Chicago longitudinal study (CLS). The CLS data consisted of 1531 students’ mathematics and reading test scores from kindergarten to 9th grade. The CLS data also included the static predictors, gender, cpc program attendance, magnet school attendance, and risk index for each student. The response variables were fitted with the LMMs, and the model equations for mathematics and reading can be written as below, respectively,

\[ Math_{ij} = \sum_{k=0}^{6} \beta_{ik}t_{ij}^k + \sum_{m=0}^{3} \left( \sum_{l=0}^{4} \beta_{6+l+m+1} \cdot x_{li} \right) t_{ij}^m + \sum_{n=0}^{3} b_{ni} t_{ij}^n + e_{ij}, \] (5.1)

\[ Read_{ij} = \sum_{k=0}^{2} \beta_{ik}t_{ij}^k + \sum_{m=0}^{2} \left( \sum_{l=0}^{4} \beta_{2+l+m+1} \cdot x_{li} \right) t_{ij}^m + \sum_{n=0}^{2} b_{ni} t_{ij}^n + e_{ij}, \] (5.2)

where \( x_{li} \) has \( l = 1, \cdots, 4 \) indicating the gender, cpc, magnet and risk variables, respectively.

In this paper, two model building approaches, the step up and the top down, were considered. In the simulation, 2000 data sets were generated from the distribution of the true model for the CLS data whose parameters were obtained from fitting the models of Equations (5.1) and (5.2). For each of the 2000 generated data sets, 1531 simulated subject repeated measures data was generated. From the 1531 simulated subjects I sampled 100, 300, and 500 individuals at random.
A complication of the data generation is that it introduces additional sampling variation in the study. To compensate for this additional error, I considered the selection of an approximate model, which is similar to but not exactly equal to the true model. The simulation focused on the efficiency of the methods for selecting the true model and the approximate model, though selection rates for the former were extremely low (see Chapter 4). In order to investigate the relationship between the population effect sizes and the rate of selection, I focused on the total effects as indexed by Euclidean distance (see Chapter 4). By applying the LRT in the model comparison in this simulation study, I found that the step up approach performed better in the selection of an approximate model than the selection of the true model. In addition, by looking at the total effect of the population parameters in the interaction effects, I found that the model having the largest total effect is the most frequently selected in the simulation study. In this Chapter, I consider under what conditions researchers might apply these findings in their own analyses. Also, I discuss limitations of my simulation study and future directions.

5.1 Sample Size

The goal of the simulation study was to help suggest ways that applied researchers might select models in exploratory longitudinal data analysis. An important issue in this regard is sample size. We first consider what constitutes an adequate sample size in the longitudinal data analysis to select an approximate model.

The true model for the mathematics data included 34 parameters. The rate of selecting an approximate model was relatively high, 73.1%, when each sampled data set from each generated data set included 8.8 times more simulated subjects (sample size of 300 simulated subjects) than the number of parameters in the true model. If a data set included 14.7 times more simulated subjects (sample size of 500 simulated subjects) than the number of parameters in the true model, the rate of selecting an approximate model was 92.5%.
In contrast to the mathematics data, the true model for the reading data had 22 parameters. The rate of selecting an approximate model was also relatively high, 83.0\%, when each sampled data set included 13.6 times more simulated subjects (sample size of 300 simulated subjects) than the number of parameters in the true model. If a data set included 22.7 times more simulated subjects (sample size of 500 simulated subjects) than the number of parameters in the true model, the rate of selecting an approximate model was 96.2\%.

If we increase the ratio between the sample size and the number of parameters, the rate of selection of an approximate model becomes higher for the mathematics and reading data. The results are summarized in Table 5.1. The first column indicates the ratio between the simulated subjects and the number of parameters, the second column indicates the ratio between the data points and the number of parameters, and the third column indicates the rate of selection of an approximate model.

Table 5.1: Rate of selection of an approximate model according to sample sizes

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th>Reading</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ratio 1</td>
<td>Ratio 2</td>
<td>Rate</td>
<td>Ratio 1</td>
<td>Ratio 2</td>
<td>Rate</td>
</tr>
<tr>
<td>100</td>
<td>2.94</td>
<td>29.41</td>
<td>0.036</td>
<td>4.55</td>
<td>45.45</td>
<td>0.143</td>
</tr>
<tr>
<td>300</td>
<td>8.82</td>
<td>88.24</td>
<td>0.731</td>
<td>13.64</td>
<td>136.36</td>
<td>0.830</td>
</tr>
<tr>
<td>500</td>
<td>14.71</td>
<td>147.06</td>
<td>0.925</td>
<td>22.73</td>
<td>227.27</td>
<td>0.962</td>
</tr>
</tbody>
</table>

1. Ratio 1 indicates the ratio between the simulated subjects and the number of parameters
2. Ratio 2 indicates the ratio between the data points and the number of parameters

In Figure 5.1, I consider the relationship between the ratio of sample sizes with the number of parameters and the rate of selection of an approximate model. Figure 5.1 shows that as the ratio of subjects to parameters increase, the proportion of selection of an approximate model increases. There is a substantial jump in selection proportion as the rate increases from 5 to 9. Based on these results, if an applied researcher was to select an approximate model at a relatively high rate, then about 15 times more simulated subjects than the number of parameters would be needed. On the other hand, if the ratio of subjects to
parameters is low, for example, less than 9, the rate of selection is very poor. In such cases, the researcher might have to consider simpler models or collect additional data. Note the models, especially for mathematics, were relatively complex having a large number of parameters. The results of the simulation suggest that such models should only be investigate when the ratio of subjects to parameters is sufficient, meaning greater than about 10.
5.2 Model Building Approaches

The other finding in this simulation study was that the step up approach performed better than the top down regardless of the degree of time transformations in the true models. In Chapter 3, I selected the best fitting models for both mathematics and reading with 6th and 2nd order polynomial models, respectively.

Based on the steps of the model building methods, I had an expectation that the top down approach would perform better for the mathematics data and the step up approach perform better for the reading data. However, as we saw with the results in Chapter 4, the step up approach performed better at selecting an approximate model for both the higher order and the lower order models. The results are summarized in Table 5.2.

Table 5.2: Rate of selection of an approximate model according to model building approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Mathematics</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100  300  500</td>
<td>100  300  500</td>
</tr>
<tr>
<td>Step up</td>
<td>0.036 0.731 0.925</td>
<td>0.143 0.830 0.962</td>
</tr>
<tr>
<td>Top down</td>
<td>0.077 0.715 0.884</td>
<td>0.122 0.730 0.842</td>
</tr>
</tbody>
</table>

The result for the sample size of 100 for mathematics does not support that the step up method performed better than the top down method. But both the step up and the top down performed poorly with a sample size of 100. It is probably meaningless to conclude any efficiency difference between the two approaches with a sample size of 100 due to such low rates of selection.

The results for sample sizes of 300 and 500 for both mathematics and reading show that the step up method performed better at selection of an approximate model. Thus, applied researchers may expect that the step up method performs better at selecting an approximate model when the sample size is relatively large relative to the number of parameters. In this simulation, “large” means about 15 times more subjects than parameters.
The step up approach has the advantage that it is consistent with common approaches to model building that begin with simpler models and then add complexity when warranted. In the model selection procedure, the step up approach compares two nested models from the lowest degree whereas the top down approach compares two nested models from the highest degree. In the CLS data, there are 10 time points requiring a relative high order polynomial for the mathematics data. With the top down method, we must begin with an overly elaborate model that has a higher order polynomial than we think is actually warranted. The simulation results suggest that consideration of such over elaborate models may not be needed and building from simpler models might suffice.

In Tables 5.3 and 5.4, I summarize the results of selecting time transformations when the sample size is 300 for mathematics and reading (the results were similar for 500 subjects). The tables show that the top down method selected a greater portion of overly elaborate models having higher order polynomials than the true model.

For example, consider the results for reading in Figure 5.4. For the top down approach 14.4% of the incorrect model selections occurred for models with a higher degree polynomial than the true model. In contrast, the step up approach had only 2.6% incorrect selections for overly elaborate models. Again, this suggests that building from simpler models to more complex might have advantages for the relatively complex models considered in this study.

Table 5.3: Selection of time transformations for Mathematics- Sample size of 300

<table>
<thead>
<tr>
<th>Degree of the time transformations</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step Up</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.015</td>
<td>0.002</td>
<td><strong>0.966</strong></td>
<td>0.016</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Top Down</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td><strong>0.944</strong></td>
<td>0.015</td>
<td>0.002</td>
<td>0.036</td>
</tr>
</tbody>
</table>

\footnote{1}{The degree of the true model is 6.}
Table 5.4: Selection of time transformations for Reading- Sample size of 300

<table>
<thead>
<tr>
<th>Degree of the time transformations</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step Up</td>
<td>0.000</td>
<td>0.000</td>
<td>0.974</td>
<td>0.026</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Top Down</td>
<td>0.000</td>
<td>0.000</td>
<td>0.857</td>
<td>0.024</td>
<td>0.020</td>
<td>0.021</td>
<td>0.020</td>
<td>0.016</td>
<td>0.004</td>
<td>0.037</td>
</tr>
</tbody>
</table>

1 The degree of the true model is 2.

5.3 Total Effects

The true models considered in the simulation were based on real data and had a number of static predictor effects. Because of this it was necessary to quantify the overall effect size in order to study how the sizes of the various main and interaction effects influenced the selection process. As discussed in Chapter 4, the total effect (TE) for a model $M$ is defined by

$$TE_M = \sqrt{\sum_{i=1}^{n} (Effect\ size_i)^2}.$$  (5.3)

The TE measures the degree to which the main effects and the interaction effects between the time transformations and the main effects influence the model selection process. Furthermore, the TE allows us to identify the same interaction effects as in the true model.

In this simulation, I found that the model having the highest TE was the most frequently selected regardless of the sample size and regardless of the method of the model building. For both mathematics and reading data, the linear interaction effects have the highest TE. The rates of selected interaction effects are summarized in Table 5.5, which shows that the linear interaction effects are the most frequently selected. Unfortunately, this characteristic does not tell us what the true interaction effects are. However, as sample size increased, the rate of the selection of true interaction effects increased while the rate of the selection of the model having the highest TE decreased. Thus, we found that if sample size was small, the model having the highest TE was most frequently selected. On the other hand, if sample
size was large enough, the rate of the selection of the true interaction effects was closer to that of the model having the highest TE (see Figure 5.2). For mathematics data, the true interaction effects model is cubic. The true interaction effects for reading is quadratic.

Table 5.5: Selection of the interaction effects according to the total effect

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Mathematics</th>
<th></th>
<th></th>
<th></th>
<th>Reading</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main</td>
<td>Linear</td>
<td>Quadratic</td>
<td>Cubic</td>
<td>Main</td>
<td>Linear</td>
<td>Quadratic</td>
</tr>
<tr>
<td>100</td>
<td>0.110</td>
<td><strong>0.671</strong></td>
<td>0.205</td>
<td><strong>0.014</strong></td>
<td>0.084</td>
<td><strong>0.798</strong></td>
<td><strong>0.118</strong></td>
</tr>
<tr>
<td>300</td>
<td>0.004</td>
<td><strong>0.577</strong></td>
<td>0.270</td>
<td><strong>0.149</strong></td>
<td>0.010</td>
<td><strong>0.720</strong></td>
<td><strong>0.270</strong></td>
</tr>
<tr>
<td>500</td>
<td>0.000</td>
<td><strong>0.395</strong></td>
<td>0.234</td>
<td><strong>0.372</strong></td>
<td>0.001</td>
<td><strong>0.572</strong></td>
<td><strong>0.427</strong></td>
</tr>
</tbody>
</table>

1. The linear interaction effects are the model having the highest TE for both mathematics and reading
2. The true interaction effects for mathematics is cubic
3. The true interaction effects for reading is quadratic
5.4 True Model Selection

In the simulation study, I investigated the efficiency of the model building approaches for longitudinal data in an exploratory analysis using the LRT. After defining what is meant by an approximate model, I investigated the efficiency of selecting the approximate model instead of the true model based on repeated sampling from the population. The reason that I introduced the approximate model is due to the additional error from sampling simulated subjects from each generated data set.

In Table 5.6, I summarize the rate of selecting the true model, to show how much the sampling error affects the process. For example, in the sample size of 500, the results are 22.9% for the mathematics data and 39.8% for the reading data in the step up approach. That is, the selection of the true model was relatively rare - no greater than about 40% - even for the simplest true model and the largest sample size.

I can think of two reasons for such a low rate of selecting the true models. One is
Table 5.6: Rate of selecting the true Model

<table>
<thead>
<tr>
<th>Mathematics</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Step Up</td>
</tr>
<tr>
<td>100</td>
<td>0.000</td>
</tr>
<tr>
<td>300</td>
<td>0.064</td>
</tr>
<tr>
<td>500</td>
<td>0.229</td>
</tr>
</tbody>
</table>

that when interaction effects were selected, the model having the highest total effects was the most frequently selected instead of the true model. In Figure 5.2, the highest rate of selecting the interaction effect occurs in the model that consists of the linear interaction effect and has the highest total effect. To correctly select the interaction effects, we need to examine the change of rates over sample size. The true model is the only model whose rate of selection increased as sample size increased.

The second reason is that the data sets fitted with the LMMs were sampled from each generated data with different sample sizes of 100, 300, and 500. This sampling process caused additional error. Therefore, the rate of selecting the true model was fairly low. It is unclear what the rate of selection would be if there were no simulation error.

5.5 Limitations

The results of the simulation suggest that the step up method has advantage when the goal is to select a model that approximates the true model. By “approximates” I mean a model that produces, among other things, a predicted growth curve similar to the true model. As seen, sample size plays an important role as the ratio of subjects to model parameters must be relatively large (say, 10 or greater) to produce high selection rates.

The true models of the simulation study were based on real data from the CLS. The true models were developed based on analysis of the data using methods and strategies that are
familiar to applied researchers in the social and behavioral sciences.

The true model for the mathematics might be considered relatively complex as it included a 9th order polynomial. This is probably on the extreme of what applied researchers would consider in practice. On the other hand, the quadratic polynomial for the reading data seems very consistent with what would be considered in practice.

As with any simulation study there are limitations to the design and implementation. In this section I discuss these limitations.

To investigate the efficiency of the model building approaches, I considered the step up and the top down methods in this paper. These two model building approaches are based on the model comparison using the LRT. As introduced in Section 2.2, the LRT has been widely used in model comparison and the result is accurate when the sample size is large and the testing parameter is in the interior of parameter space. In the context of LMMs, the LRT result is accurate when the LRT based on maximum likelihood estimation is used to test hypotheses about the fixed effects parameters or when the LRT based on restricted maximum likelihood estimation is used to test hypotheses about the variance-covariance parameters (Morrell, 1998 (36)).

The LRT has its own limitation in that it is not very accurate when the testing parameter is on the boundary of the parameter space. That is called the “boundary value problem”. Such a problem occurs when the LRT is applied to test a variance component, as in this case we have,

\[ H_0 : \sigma^2 = 0 \quad \text{vs.} \quad H_a : \sigma^2 > 0. \]  \hspace{1cm} (5.4)

Since the null hypothesis value is on the boundary of the parameter space, \([0, \infty)\), the LRT result will not be very accurate (see Pinheiro and Bates, 2000 (39)), such that its \(p\)-value will tend to be too large.
In the LRT applied in the model building, the general two-sided hypothesis

\[ H_0 : \sigma^2 = 0 \quad \text{vs.} \quad H_a : \sigma^2 \neq 0 \]

was used, since the LRT is asymptotically equivalent and the asymptotic null distribution is well known to be $\chi^2_1$ (Cox and Hinkey, 1990 (6)). However, when the null hypothesis is on the boundary of the parameter space as in hypotheses (5.4), the LRT may provide inaccurate results in model selection. This boundary value problem occurs when we compare two nested models in terms of random effects, since the full model in the hypothesis test includes at least one more variance component than the reduced model, see Expressions (2.19) and (2.20).

To solve the parameter boundary value problem, some have considered an extension of the LRT such as approximating the reference distribution as a mixture (see Stram and Lee, 1994 (51) and 1995 (52)). Others have considered an alternative test, such as the score test (see Verbeke and Molenberghs, 2003 (55); Molenberghs and Verbeke, 2007 (35)). Yet others have considered parametric bootstrap methods (see Faraway, 2006 (13); Pinheiro and Bates, 2000 (39)). Finally, information criteria such as the AIC and BIC have been suggested in place of the LRT. In general, these alternative methods work better than the LRT, since the LRT may provide inaccurate results in the boundary value problem. But they are all comparable and none of them appear to dominate in terms of efficiency, since all have their own limitations under certain circumstances. For example, Stram and Lee (1994 (51)) propose a 50:50 mixture of a $\chi^2$ and a mass at zero. Unfortunately, the relative proportions of these two components vary from case to case (see Faraway, 2006 (13); Pinheiro and Bates, 2000 (39)). Future research might focus on the use of the LRT for selecting fixed effects and one of the alternative methods for selecting the random effects.

With longitudinal data, we sometimes see dynamic predictors that vary over time, which can be modelled in a similar fashion to the time transformations. In this simulation study,
all four predictors in the CLS data were static predictors and did not vary over time. If we considered any dynamic predictors in this study, we would select a model for the dynamic predictors after selecting the time transformations but before selecting the main effects in the model building procedure. Future research might consider the inclusion of dynamic predictors in the models.

In this simulation study, there were constraints on the highest degree of both the interaction effects and the random effects. The highest degree was capped at the 3rd order, which was applied to the mathematics data. Since the higher order interaction effects or the higher order random effects are rarely fitted in practice, I limited the highest degree to the 3rd order in my program. However, this might be inconsistent with how applied researchers determine random effects for inclusion. If the order of polynomial is relatively low, as it was for the reading models, then there is a tendency to include a random effect for every time transformation fixed effect. With the mathematics data, the order of the polynomial was very high, so such “automatic” inclusion seems unreasonable. Still, future simulations might provide alternative scenarios for building models to be consistent with common practice.

5.6 Conclusion

The results of my simulation show that the step up approach performed better at selecting an approximate model to the true model regardless of sample size and regardless of the complexity of the true model (See Table 5.2). As the ratio of sample size to the number of parameters increased, the rate of selecting an approximate model also increased (see Table 5.1 and Figure 5.1). By applying the total effect measurement and examining the changes of rates of selecting interaction effects over sample size, we can identify the interaction effects of the true model. In other words, as sample size increases, the rate of selecting the interaction effects in the true model increases and gets closer to that in the model having
the highest total effects (see Table 5.5 and Figure 5.2).

Based on the findings in this paper, I list recommendations for applied researchers in the selection of time transformations, main effects, interaction effects, and random effects as follows. When fitting the time transformations, it is more efficient to compare the nested models from the lowest degree than to compare the nested models from the highest degree.

When selecting the main effects of static predictors I tested each individually. This might not be optimal and in practice, the nature of testing depends on the research questions. By considering a set of the static predictors selected in the previous step, we fit interaction effects between the set and the time transformations to longitudinal data. This procedure provides the best fitted model that is almost identical with the true model except for its random effects. In addition, it is desirable to have about 14.7 times more subjects than the number of parameters in the LMM, to achieve 92.5% percent of selecting an approximate model.

As for future directions, there are a number of ideas I have based on the current study and its limitations. First, I am interested in studying the alternative tools for selecting random effects previously mentioned. Since the LRT provides an inaccurate result with the boundary value problem, there is a need to develop an alternative tool when testing random effects. In Section 5.5, I listed four alternatives such as using a mixture distribution, the score test, parametric bootstrap methods, and information criteria. In future work, I play to examine the model selection with the parametric bootstrap methods, since the parametric bootstrap seems to be simple, transparent and efficient.

The parametric bootstrap methods can be explained as follows. Under the null hypothesis, we resample from our sample data in order to approximate the sampling distribution of the static in question. Data are generated under the null model, then the fit of the null and alternative models is assessed, say with the LRT. The process is repeated a large number of times and the proportion of LRTs exceeding the observed value is used to estimate the \( p \)-value. In future work I plan to investigate the efficiency of model comparison tools using
both the analytic LRT and the parametric bootstrap methods.

Finally, I would like to examine model selection with the LMMs for longitudinal data with models that include dynamic predictors. In this paper, I discussed the efficiency for longitudinal data static predictors effects only. Though the static predictors are of different types such as quantitative and categorical, not dynamic predictor was considered. In educational and psychological data, we sometimes see dynamic predictors and I think that selection of models with such predictors will become increasingly important in the year to come.
References


subset selection algorithms: Frequency of obtaining authentic and noise variables. 
*British Journal of Mathematical and Statistical Psychology*, 45, 265-282.


109


