

# **Multilevel Modelling of Reading Achievement in Children and Youth**

**Yifei Fang B.Sc.**

**A Dissertation**

Presented to the University of Dublin, Trinity College  
in partial fulfilment of the requirements for the degree of

**Master of Science in Computer Science (Data Science)**

Supervisor: Bahman Honari

August 2019

# Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

---

Yifei Fang

August 8, 2019

## Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

---

Yifei Fang

August 8, 2019

# Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Bahman Honari. He continuously supported and helped me in finalising the research topic and methodology particularly. With his expertise in statistics, I was able to complete the project and thesis.

Growing Up in Ireland is a really meaningful study. Thanks for the joint efforts of ESRI and Trinity College Dublin. I am grateful to all staffs who have worked on the study, especially Caoimhe O'Reilly from ESRI. I received a lot of helps from her.

I very much appreciate the useful R package of R2MLwiN developed by the Centre for Multilevel Modelling and the University of Bristol. Lastly, I would like to thank my flatmate Ying for providing me guidance in the writing.

YIFEI FANG

*University of Dublin, Trinity College  
August 2019*

# Multilevel Modelling of Reading Achievement in Children and Youth

Yifei Fang, Master of Science in Computer Science  
University of Dublin, Trinity College, 2019

Supervisor: Bahman Honari

Despite the vast studies on the longitudinal study of Growing Up in Ireland (GUI), little is known about the perspective of the advanced statistical modelling. The study investigates cognitive development and discovers influential factors using GUI data set. The predictors emerged from the literature review are household income, home literacy experiences, expectation of parents and phone ownership. In the analysis, the work involves data preprocessing, feature selection and multilevel modelling with predictors. Iterative generalised least squares algorithm is used for parameter estimation. A multilevel analysis yields the growth rate of the reading development of 9-18 years old. The assumptions of the models are held: linear relationship of variables, normal distribution of residuals and homogeneity of variance.

Two-level multilevel models are built, and the final linear growth model achieves the best score. Linear growth models who have higher intercepts tend to have steeper slopes of books coefficient and flatter slopes of phone ownership. The variance of reading achievement between students increases with more books and decreases with owning phones. So there is a strong negative influence of phone ownership while book is a positive factor. The study, as a part of GUI, starts applying cutting-edge statistical models in three waves of the child cohort. The study could be carried on with more predictors and more cross-level interactions in the future.

**Keywords:** *Multilevel Modelling, Reading Achievement, Growing Up in Ireland*

# Contents

<b>Acknowledgments</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>Chapter 1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Growing Up in Ireland . . . . .	2
1.3 Motivation . . . . .	3
1.4 Overview . . . . .	4
<b>Chapter 2 Literature Review</b>	<b>5</b>
2.1 Child Academic Outcomes . . . . .	5
2.2 Modelling Techniques . . . . .	6
2.2.1 Analysis of Variance . . . . .	6
2.2.2 Latent Curve Modelling . . . . .	7
2.2.3 Hierarchical Linear Modelling . . . . .	7
<b>Chapter 3 Methodology</b>	<b>9</b>
3.1 Mixed-effects Modelling . . . . .	9
3.1.1 Fixed Effects . . . . .	10
3.1.2 Random Effects . . . . .	10
3.2 Estimation Methodology . . . . .	11

3.2.1	Maximum Likelihood Estimation . . . . .	11
3.2.2	Iterative Generalised Least Squares . . . . .	12
3.3	Evaluation Metrics . . . . .	13
3.3.1	Deviance . . . . .	14
3.3.2	AIC and BIC . . . . .	14
3.3.3	Pseudo R Squared . . . . .	15
3.3.4	Likelihood Ratio Tests . . . . .	17
<b>Chapter 4 Data Preparation</b>		<b>19</b>
4.1	Data Preprocessing . . . . .	20
4.1.1	Data Cleaning . . . . .	20
4.1.2	Data Scaling . . . . .	21
4.1.3	Data Transformation . . . . .	22
4.2	Feature Selection . . . . .	23
4.3	Exploratory Data Analysis . . . . .	25
4.3.1	The Target Variable . . . . .	25
4.3.2	Explanatory Variables . . . . .	28
<b>Chapter 5 Implementation of Multilevel Modelling</b>		<b>31</b>
5.1	A Variance Components Model . . . . .	32
5.1.1	Estimation of Residuals . . . . .	34
5.2	A Linear Growth Curve Model . . . . .	35
5.2.1	Models with Explanatory Variables . . . . .	35
5.2.2	Random Slopes Models . . . . .	36
5.3	Model Building . . . . .	38
5.3.1	A Model with Books Predictor . . . . .	39
5.3.2	A Model with Household Predictor . . . . .	40
5.3.3	A Model with Phone Predictor . . . . .	40
5.3.4	The Final Multilevel Model . . . . .	41
<b>Chapter 6 Results and Discussion</b>		<b>43</b>
6.1	Model Diagnostics . . . . .	43
6.1.1	The Baseline Model . . . . .	43
6.1.2	The Final Model . . . . .	44

6.2	Model Inference . . . . .	47
6.2.1	Residuals Estimation . . . . .	49
6.2.2	Variance Estimation . . . . .	50
6.2.3	Examining Variation of Coefficients . . . . .	51
6.3	Model Evaluation . . . . .	52
6.4	Interpretation of Model Output . . . . .	54
6.5	Conclusion . . . . .	56
	<b>Chapter 7 Future Work</b>	<b>57</b>
	<b>Bibliography</b>	<b>59</b>
	<b>Appendices</b>	<b>61</b>



# List of Tables

4.1	The Dependent Variable . . . . .	19
4.2	Long Format of Data Structure . . . . .	23
4.3	The Potential Explanatory Variables . . . . .	25
4.4	Descriptive Statistics of Response Variables . . . . .	25
5.1	Likelihood Ratio Test of Model 2 . . . . .	39
5.2	Likelihood Ratio Test of Model 4 . . . . .	41
6.1	Model Estimation . . . . .	48
6.2	Model Evaluation . . . . .	53
6.3	Elapsed Time and Converge Iteration . . . . .	54
6.4	The Impacts of Predictors . . . . .	55

# List of Figures

4.1	The Box Plot of Reading Achievement . . . . .	22
4.2	Correlation Matrix of Logit Score in Drumcondra Reading Test . . . . .	24
4.3	Trend of 100 Observations . . . . .	26
4.4	The Distribution of the Response Variables . . . . .	27
4.5	The Frequency of Books . . . . .	28
4.6	The Frequency of Household Type . . . . .	29
4.7	The Frequency of Phone Ownership . . . . .	30
5.1	The Variance Components Model . . . . .	33
5.2	The Random Intercept Model . . . . .	36
5.3	The Random Slopes Model . . . . .	38
6.1	Standardised Residuals VS Normal Distribution . . . . .	44
6.2	The Density Estimation of Intercept Coefficient . . . . .	45
6.3	The Density Estimation of Books Coefficient . . . . .	46
6.4	The Density Estimation of Phone Coefficient . . . . .	47
6.5	Caterpillar Plots of Residuals . . . . .	49
6.6	Variance against Predictors in Model 2 and 4 . . . . .	50
6.7	Intercepts VS Slopes in Model 5 . . . . .	52

# Chapter 1

## Introduction

### 1.1 Background

Longitudinal design research has applied to many fields experiments and surveys recently, such as in medical, educational and psychological research. The focus of this research is investigating cognitive development, more precisely reading ability development in a longitudinal study of children. The study involves with continuous-time monitoring processes of children lives.

Education is a process of equipping people with knowledge, skills and so on. In schools, children and youths receive a compulsory education at around the ages of 6 to 16. During this period, cognitive, social and other skills are meant to develop to a intermediate or advanced level. A lot of money and efforts have been spent on improving reading performance and promoting an effective reading habit. However, there are many aspects and factors that hinder the development. Students often face a proximate risk of low achievement when they take a academic high school curriculum. Reasons behind the continued low achievement are vulnerable since experts can predict the independently academic resilience. It is critical to identify the highly related independent variables to the growth rate of the reading ability. The reasons why child's language skills progress or deteriorate could be figured out. The finding process is complicated. Given the fact of that defining the growth is hard and unclear, applying the sophisticated modelling mechanism helps define this development.

Multilevel modelling is an appropriate tool for modelling data from a longitudinal

study. It was first been fully applied in educational science and has been widely used by now. In statistics, multilevel models are officially called as mixed-effects models and first introduced by Ronald Fisher in 1950. As an alternative to multivariate analysis, differences of each student in growth curves can be examined. With the provision of adjusting covariates, another advantage is allowing correlation of variables.

## 1.2 Growing Up in Ireland

Growing Up in Ireland (GUI) is a national longitudinal study tracking two groups of children in contemporary Irish families. The survey investigated the lives of the same subjects and gathered information from children, parents and teachers for repeatedly over every four years. The objective of the GUI is examining the well-being of children. Most possible factors contribute to or undermine the development were investigated and put in the questionnaires. Compared with other studies conducted in the country, GUI is more significant and comprehensive. Because a wide range of topics and domains ranging from health to educational development is included. Besides child cohort spans a long period. Many latent insights contribute to policy formation and services provision towards children and families.

GUI Cohort'98 includes anonymous details on children and youth who are based in Ireland. The sample consists of three waves. 8500 nine-year-old children who were born between November 1997 and October 1998 first participated in the initial interview between September 2007 and May 2008. And researchers re-investigated and recorded the results of 13-years-old Child from 2011 to 2012. The final round data collection was taken place when they were at age 17 or 18 between 2015 and 2016. There have been previous studies using child cohort of GUI data. Minister for Health and Children gave a thorough report listing their findings after calculating the concrete ratios and the proportion of the composition of the indicator in the GUI data set. They hoped to improve citizens' understanding of children well-beings and help government propose policies to improve child's lives. These reports enclosed factors which affected child's reading achievement. Moreover, some other researchers and institutes have adopted the basic analysis methods regarding children educational development.

## 1.3 Motivation

There are some few problems in existing studies of GUI. First, researchers paid great attention to health problems or difficulties during the transition to high school. Fewer efforts were spent on studying cognitive achievement, especially on children reading ability. Reading ability enables children effectively to learn and be adaptive in a new environment. Cultivating effective reading ability is a complex cycle that requires the integration of cognition, skill, and affection.

Second, more cutting-edge data statistical tools and models need to be applied to cognitive ability research. Most studies only focused on one wave and produced a report, such as "the lives of 9 years old" [1]. They seldom covered all three waves due to the inconsistency of GUI child cohort. Since the questionnaires were designed differently and results of the wave 3 were new. Subjects can be more comprehensively analysed and studied over a longer observation period of three waves. Furthermore, more adequate data makes the research results more objective. Besides, self-correlated factors were included in the linear regression models. To continue previous researches and further analyse the GUI data set, this study aims to implement advanced statistical models and find the changes of child's reading proficiency when they get older. Because hidden information waits for a deeper exploration.

Taking the above situation into account, the research objectives are defining the improvement of children reading ability, identifying the key factors, and choosing the best model. The purpose is helping policymakers make decisions and improving children reading skills.

For statistical analysis, the following assumptions are put forward and tested in the modelling. The first one is the hierarchy of data structure. Since the models are built on the base of the nested structure. Assuring hierarchical data sample is the first requirement. The next assumption is a strong effect of time on the child's reading performance. Hence, studying the growth rate is the main focus. Also, the data population follows the Gaussian distribution. This valid assumption is the next condition of models. The final assumption is chosen explanatory factors correlated to the reading scores. The extent of influence and inner comparison of each factor could be detected after modelling.

## 1.4 Overview

Given the circumstance of Irish children, analysing the data set is the first step. After fully understanding the data, data preprocessing is followed next. Although it costs a large number of efforts and time, data cleaning is required. The next step is seeking and screening negative or positive influential factors of education development. After that, the most important step comes, building and comparing models. The last step is testing the assumptions and analysing the output of statistical modelling.

The structure of the thesis is introduced in the following paragraph. Chapter 1 Introduction states the current background of the study, a general description of GUI data set and research motivation as well as objectives. The final section 1.4 gives a brief overview of the research procedures along with a thesis structure. Chapter 2 includes a literature review of children reading outcomes and possible factors. Moreover, potential theoretical frameworks and models are first introduced. In chapter 3, concepts and formulas behind multilevel modelling are explained. It is followed by the logic of evaluation metrics and how to examine and compare models performance. How the data get prepared for models is described in chapter 4. The process is mainly divided into data preprocessing and feature selection. There are also preliminary results of data analysis before modelling. Implementing procedure of the multilevel models is in chapter 5. It starts with a variance components model and then linear growth curve models. A detailed depiction of five models is in the end. Chapter 6 describes results part where previous assumptions are tested and the results of models are analysed. A highly summary of multilevel models is in section 6.5 of conclusion.

# Chapter 2

## Literature Review

Based on related researches, studying children's education development is usually developed by a longitudinal research. Longitudinal data, also called as functional data, is a collection of repeated observations from the same subject over a long period. Due to the hierarchical data structure, multilevel modelling has been rapidly developed recently and widely used in the education field. Literature review summarises the current related researches and studies, providing potential factors related to the reading attainment and possible modelling methods.

### 2.1 Child Academic Outcomes

Experts in education area proposed many factors influencing child academic outcomes. It was raised that parental school involvement promotes child academic achievement. Parents with lower socioeconomic status potentially lack ability to involve children's schooling and have more negative experiences [2]. The cross-country analysis was conducted in studying the formation of children cognitive ability. It discovered that the family environment and parental investment were key input factors to children cognitive ability. Families with higher household income would invest more on children, which resulted the differences of cognitive development [3]. Study shows that reading development benefits from early home literacy experiences. Home literacy activities were proved to improve the understanding of basic concepts, promoting children's reading skills. The research proposed the significance of raising later literacy skills [4].

In the related researches of Growing Up in Ireland, child academic achievements are influenced by many factors. Impact of technology use on childrens development was examined by Cross-sectional analysis [5]. Ordinary Least Squares (OLS) technique was the estimation method in the two-staged sampling frame. Findings showed a negative association between mobile phone ownership and academic outcomes. While nine-year-old children owning a mobile phone is correlated with lower socioeconomic status. Other dependent factors were enclosed in the the nine-year-old cohort [6]. Children with higher achievement in reading test scores are mostly from high household income families. In addition, there are other useful factors: the child's mother highest level of education and access to educational books at home. Moreover, the research showed high influences of school and classroom aspects in the mathematics scores [1].

## 2.2 Modelling Techniques

Repeated measures design is asking one question over two or more two time points on the group of subjects. So to proceed repeated data analysis, the first step is manipulating within-subject conditions. Since one subject has more than one data point from a source. Moreover, there is a different growth trend between each individual. Traditional linear regression model based on the homogeneity of variance in data sample can not satisfy the random errors. For repeated measures analysis, there are many underlying available methods: Analysis of variance, Hierarchical Linear Modelling and Latent Curve Modelling.

### 2.2.1 Analysis of Variance

With data sample of experimental studies, analysis of variance (ANOVA) is developed for analysing and testing the differences between group means by Ronald Fisher. The analysis can be subdivided into two types. One-way ANOVA refers to one response variable with 2 levels. And two-way ANOVA can have two outcome variables and multiple levels. ANOVA analysed the contribution of variation from each group, the effects of controllable factors on the research results are determined and measured. The total sum of squares (SST) are divided into two components:  $SST = SSR + SSE$ , where explained variability is denoted by SSR and unexplained variance is SSE. F test



is used to examine the association of predictors and the response variable. It requires the balance and independence of the sample and the equality of covariates over an even spacing time. Due to the limitation of raw changed values and high requirement of a sample, ANOVA models are not appropriate in GUI data.[7]

### **2.2.2 Latent Curve Modelling**

Latent Curve Modelling, also known as latent growth mixed models, estimates the growth of the latent variables over some times. In the latent variable mixed growth model, each potential class has a different random effect. In practical applications, individuals are put into different classes and the development of trajectories is examined. It has an important significance because individuals of different groups may not only have different trajectories but may have different predictors and outcome variables. In the way, it can distinguish different potential changes. Besides, it estimates the overall probability of individuals in each class. Latent curve modelling presents the average development trajectory of each cluster and the differences between individuals in the same cluster. Meanwhile, it also suggests a possible cluster that each individual is most likely to belong to. It helps researchers identify potential changes of different types and test the relationship between different groups and variables [8].

The requirement using latent curve modelling is the existence of several different growth patterns. The different pattern corresponds to each potential class in the unobservant population. The overall development of children is assumed homogeneous in the GUI data set. Moreover, having different predictors and the response variable is complicated and unnecessary. Therefore, studying GUI data set does not need latent growth curve models.

### **2.2.3 Hierarchical Linear Modelling**

Combining the advantages of univariate mixed ANOVA model, Hierarchical Linear Modelling (HLM) offers a reasonable alternative approach. HLM is also referred to multilevel modelling. Multilevel models, as an extension of OLS regression models, are based on Maximum Likelihood estimation. The common used parameter estimation methods are Iterative generalised least squares, Restricted generalised least squares and Markov chain Monte Carlo [9]. The multilevel linear model was first

proposed by Lindley and Smith in 1972. Owing to the limitation of computational techniques, the model's parameter estimation method did not develop until the invention of Expectation-maximisation algorithm (EM). EM algorithm finds the maximum likelihood and solves the parameter estimation by an iterative process. It made the application of multilevel modelling possibly. In 1983, Strenio, Weisberg, and Bryk applied the method in sociological researches. Subsequently, Goldstein proposed IGLS in 1986. IGLS usually starts with a reasonable parameter estimate, and then the parameters are estimated step by step by iteration.

Compared with standard regression and ANOVA models, multilevel models are more applicable for nested data and repeated measures. Since it allows the dependent random errors between subjects. Multilevel models support unbalanced data and continuous input predictor variables. Under the assumption that the random errors are independent between each student. In HLM analysis, the error in the traditional regression analysis should be decomposed into two parts: one is the error caused by the difference between the first level of observations, and the other is the error caused by the variation of the second level. It provides a solution to analysing differences of individuals with a higher level. Multiple variables could be input to different levels simultaneously. They can be either imported to one level or many levels according to users' preference. Therefore, HLM is more flexible and efficient.

# Chapter 3

## Methodology

Longitudinal data are generated when measurements of subjects are repeatedly recorded more than twice. It is also called repeated measures data which has repeated observations across more than two time points. The benefit of the repetition is that researchers can study the effects of time. The multivariate model has traditionally been used in time-series data. It could be regarded that the response variable has 3 responses for each child and conducting a multivariate analysis. Nevertheless, the observations of GUI data are not independent. And the focus of the study is not only the time effect on children but also the factors affecting the reading ability. The repeated measures model is more appropriate under this circumstance. Pooling all entities and using ordinary least squares (OLS) could lead to a biased inference. The more efficient estimates can be obtained by IGLS.

### 3.1 Mixed-effects Modelling

Dealing with repeated measures, mixed-effects models allow both fixed effects and random effects, which fixes correlated errors. By taking account of other potentially relevant covariates, mixed-effects modelling provides a solution for studying several within-subjects effects [10]. The general equation 3.1 of the linear mixed-effects model is as below:

$$Y = X\alpha + Z\beta + e \tag{3.1}$$

where  $y$  represents the outcome variable,  $X$  and  $Z$  are full-rank, known as covariate matrices.  $\alpha$  is the population mean to be estimate of,  $\beta$  is a random-effects (individual-specific) coefficient, following a multivariate Gaussian distribution:  $\beta_i \sim \mathcal{N}(0, D^*)$ .  $D$ , the covariance matrix of the random effects, should be positive semi-definite.  $e$  is called error or residual. In the parameter estimation,  $e \sim \mathcal{N}(0, \sigma^2 I)$  and the variance  $\sigma^2$  should be positive. [11]

### 3.1.1 Fixed Effects

Fixed effects analysis is generalised by building multiple regression models and averaging the parameter across the individual. Variables which are treated to be fixed should change in value over time [12]. It is also required that data should range across more than two time points. In fixed-effects models, variables are thought as constant across the individual. All levels share a common effect size. It means that the group means is modelled as fixed for each grouping. In the longitudinal data analysis, fixed effects could be used to represent the subject-specific means [13].

$$\begin{aligned} y_{i_j} &= \alpha_i + \beta_1 x_{i_j} + e_{i_j} \\ e_{i_j} &\sim \mathcal{N}(0, \sigma_e^2) \end{aligned} \tag{3.2}$$

The ANOVA model is commonly expressed in equation 3.2. Intercepts  $\alpha_i$  varies among each entity. Consequently, each one of these factors is treated as the fixed effect of subject  $i$ . It contains more than one determinant of the response variable, which is correlated with the independent variables within changes over time. Thus fixed parameters allow observable differences between subjects but cannot control the changes over time.

### 3.1.2 Random Effects

Random-effects model, so-called multilevel model, is a kind of hierarchical linear model. The goal is to analyse longitudinal designs with repeated-measures regressions. It deals with the data which are drawn from a hierarchy of different groups. In random-effects models, parameters are understood as random variables. Group effects are random, which means that it accepts the differences both within and between individuals related

to that hierarchy of data structure. It provides mix-effects models with the between-individual correlation structure. In general equation 3.3,  $i$  means the group indicator and  $j$  stands for the individual.

$$y_{i_j} = \mu + u_i + w_{i_j} \quad (3.3)$$

where  $\mu$  is the grand average of the entire data sample,  $u_i$  is the group-specific random effect.  $w_{i_j}$  is the individual-specific random effect and deviates from  $j$ th group [14].  $u_i \sim \mathcal{N}(0, \sigma_\mu^2)$  and  $w_{i_j} \sim \mathcal{N}(0, \sigma^2)$ .

## 3.2 Estimation Methodology

Estimation of parameters is a process that determines the appropriate values of parameters in models with data sample. There are many techniques, for example, least-squares principle.

### 3.2.1 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE), the most widely applicable method, generates the estimation for parameters by maximising a likelihood function of  $\theta$  :

$$L(\theta) = f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta) \quad (3.4)$$

while  $x_i$  is numbering for a random sample  $X_1, \dots, X_n$ .

The ultimate goal is estimating the HLM parameters that have most likely produced for the observed data. HLM yields simultaneous estimation of fixed and random components by maximising the function 3.4. When the sample size is large, the estimation of MLE is nearly unbiased. The general procedure is

1. Take the log of likelihood function 3.4:

$$l(\theta) = \ln L(\theta) \quad (3.5)$$

2. Set the derivatives of  $l(\theta)$  to be 0 and  $\theta = \hat{\theta}$

$$\frac{\partial l(\theta)}{\partial \theta} = 0$$

3. Solve for  $\hat{\theta}$

There are two different MLE functions: Full maximum likelihood (FML) and Restricted maximum likelihood (RML). FML includes both the regression coefficients and the variance components, whilst RML only includes the variance components and estimates the regression coefficients in the following estimation step. Another difference is that FML uses less computation and saves more time. However, FML may produce biased estimates of variance components because it only focuses on the fixed parameters. While RML only compares differences of the random parameters. There is a trade-off in choosing MLE. This research focuses on estimating the covariances and the variances among the variables in the multilevel model. IGLS algorithm derived from FML is applied in this research.

### 3.2.2 Iterative Generalised Least Squares

Ordinary least squares (OLS), is also referred to linear least squares. It estimates the coefficients in a linear regression model by minimising the squares residual (the sum of differences between fitted values and observed values).

$$\min \sum_{i=1}^n \|(\hat{\alpha} + \hat{\beta}x_i) - y_i\|^2 \quad (3.6)$$

Minimising the residual requires the calculation of the first-order conditions according to  $\alpha$  and  $\beta$ . After deriving and setting it to zero,  $\hat{\beta} = \frac{\text{cov}(X, Y)}{\text{var}(X)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ .  $\mathbf{Y}$  is an  $n \times 1$  vector with  $\text{cov}(Y | X\beta) = \text{cov}(E) = V$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}, \text{cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \quad (3.7)$$

If covariance matrix  $V$  is known, which means that estimate of  $\beta$  is consistent, OLS can be used. While maximising the likelihood is a nonlinear problem, Iterative generalised least squares (IGLS) algorithm is a reasonable solution without knowledge

of the covariance matrix. IGLS method is an estimation for the normal mixed-effects model, providing a greater flexibility of the covariance matrix. For multilevel regression models, maximum likelihood estimation is an iterative process. IGLS is the method of FML under the assumption of knowing fixed part estimates. It estimates variances and covariances used fixed coefficients from the likelihood function. IGLS can only estimate the fixed effect in a general linear regression model.

An iterative procedure starts initial estimated values for the regression coefficients and ends until the process converges. The initial estimate is generated from OLS and zero value of the variance component. It is followed by estimating  $\hat{\beta}$  according to equation 3.7. It then improves the estimate of covariance matrix  $V$ . The procedure above is iteratively repeated until it converged [15]. In the first iteration, a complex iteration procedure tries improving on the starting values. Then the likelihood function is evaluated and the second iteration is performed. This procedure continues until the process converges. In other word, an iterative sequence leads to reaching a stable solution. IGLS estimation may produce biased estimates of the random parameters because it does not take into account the sampling variation of the estimates for variance components.

### 3.3 Evaluation Metrics

Model evaluation metrics are used to measure the goodness of fit given the observed data and estimated parameters. These kinds of measurement help select an optimal model that suffices as the sole model. There are many means comparing regression models' performance, for example, R-squared score ( $R^2$ ) or Mean Absolute Error (MAE).

For mixed-effects models, model fit and the explanatory power aspects should be both considered. The dominance analysis (DA) method measures model adequacy and is designed for evaluating factors' importance [11]. Under the assumption of variables following Gaussian distribution, statistical test is for examining the significance of unknown population parameters estimation. Instead of using the normal t-tests, z-tests are adopted due to lacking knowledge of the variance and the large data sample size. The process is almost identical to the chi-squared test. The only difference is that t-test estimates the standard deviate. The details procedure of chi-squared test could be found in the chapter 3.3.4.

Evaluation metrics based on the likelihood are performed for statistical hypothesis testing. To test the significance of models and predictors, the following evaluation metrics are calculated. Multilevel models are nested with each other. They all based on the initial baseline model. At convergence the estimating process, the log-likelihood value according to the formula 3.5 is calculated for judging more complex models.

Different well-justified criterion has a different perspective towards the best-fitted model. A rigorous model selection process often includes multiple evaluation metrics. The evaluation metrics below all depend on the value of likelihood. Deviance, AIC and BIC are important criteria for choosing the best predictor subsets in regression. AIC and BIC criteria can be also used in the non-nested models. These metrics show the significance of predictors and models. To compare the random slope models to random intercept models with explanatory variables, Likelihood Ratio Tests are performed for model selection criterion.

### **3.3.1 Deviance**

The deviance statistic is generated by the difference of the log-likelihoods. Between the fitted model and the baseline model, deviance is calculated using equation 3.8. It is used to indicate the hypothesis test that model changes improve the fit of the model or not. The changes could be additional model predictors or setting of random coefficients. In linear regression models evaluation, deviance finds the sum of squares of total residuals. Moreover, it depends on the sample size, the degree of freedom of the model and the goodness of fit.

$$\text{Deviance} = -2 \times \text{LogLikelihood} \quad (3.8)$$

The lower value of deviance, the more accurate the mixed-effects model is. In most cases, large values of deviance suggest poorly fitting models.

### **3.3.2 AIC and BIC**

Two well-known approaches Akaike information criterion (AIC) and Bayesian information criterion (BIC) describes an adequate fit for models. They are both penalised-likelihood criteria dealing with a trade-off between the model fit and the model simplic-



ity. They interpret the relative quality of multilevel models and identify parsimonious models. Akaike first put forward AIC criterion based on the relationship of K-L information and likelihood theory. AIC, an unbiased estimator, evaluates candidate models by considering the complexity of models. [16]

$$\text{AIC} = -2\log L(\hat{\theta}) + 2k \quad (3.9)$$

where  $k$  is the number of model parameters. A second penalty component is added compared with the deviance statistic. Another closely similar approach is Bayesian Information Criteria, also called as Schwarz information criterion. Since AIC does not penalise the number of parameters as strongly as BIC. BIC is a better asymptotic property compared with AIC because AIC tends to overfit or underfit the sample data. BIC could deal with overfitting problem when the data sample is large. Moreover, AIC will have a risk of choosing a model with too many parameters. Since AIC could not detect an increase in the likelihood by adding parameters. If the data sample is sufficient, BIC will have less risk of choosing such models.

$$\text{BIC} = -2\log L(\hat{\theta}) + k\log(n) \quad (3.10)$$

where  $k$  is the number of model parameters and  $n$  means the sample size. AIC and BIC both receive a negative contribution of the Log-likelihood and a positive contribution of the parameters. Maximising the likelihood function leads to a better inference. So lower AIC and BIC, the better the fit. It indicates that a model is more likely to be the true model with the data sample.

### 3.3.3 Pseudo R Squared

R-squared is a common use statistic measurement for evaluating OLS regression models.

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\text{var}(y_i - \hat{y}_i)}{\text{var}(y_i)} \quad (3.11)$$

In the above formula,  $\bar{y}$  is the mean of the sample data and  $\hat{y}_i$  is the predicted value. An equivalent equation written at the end of the row is more intuitive for

variance explained components. The value of  $R^2$  always lies in the range between 0 and 1 and can never be negative. In addition, the value of  $R^2$  will not change much when including explanatory variables. Since it is not possible to explain variation change in the response variable.

Applying the  $R^2$  evaluation metric in the multilevel analysis is tricky due to random slopes. The explained and unexplained variances are determined on the values of the predictors. Pseudo-R-Squared refers to the  $R^2$  statistic in multilevel models. It measures the explanatory power of a model or a predictor.

Pseudo-R-Squared can be defined in several ways. Bryk and Raudenbush first proposed it in 1992. R-Squared compares the candidate model to a baseline model by explaining each variance component.

$$R_{\text{level1}}^2 = 1 - \frac{\sigma_e^2(\text{candidate})}{\sigma_e^2(\text{null})}$$

$$R_{\text{level2}}^2 = 1 - \frac{\sigma_u^2(\text{candidate})}{\sigma_u^2(\text{null})}$$

$$R_{\text{Total}}^2 = \frac{(\sigma_e^2(\text{null}) - \sigma_e^2(\text{candidate})) + (\sigma_u^2(\text{null}) - \sigma_u^2(\text{candidate}))}{\sigma_e^2(\text{null}) + \sigma_u^2(\text{null})}$$

where candidate represents the candidate model and null refers to the baseline variance components model. In the two-level mixed-effects model, three scores of Pseudo-R-squared are computed:  $R_{\text{level1}}^2$  for level 1,  $R_{\text{level2}}^2$  for level 2 and  $R_{\text{Total}}^2$  for the total variance. The formula shows calculating Pseudo-R-squared with two variance components: the first level variance component  $\sigma_e^2$  and the second level variance component  $\sigma_u^2$ . This sort of variance-component-specific calculation is one-sided. Because it uses variance from each level instead of the total variance. Pseudo-R-Squared changes by addition of new variables in the model or taking predictors into account in the random effects.[17]

Here is another definition of R-squared using MLE based on the study of Maddala [18].

$$R_D^2 = 1 - \frac{-2\ln(L_\beta)}{-2\ln(L_0)} = 1 - \frac{\text{Deviance}(\text{candidate})}{\text{Deviance}(\text{null})}$$

In equation above, -2 was left deliberately and deviance was formed. Consequently, Deviance is denoted by D.  $L_\beta$  and  $L_0$  are the likelihood of the full and null model

independently. Basically, it is just an evaluation metric using a variation of deviance.

Accordingly, a well-rounded Pseudo-R-Squared can explain the proportional variance reduction and include the total variance. Snijders and Bosker suggested a measurement defining it in error of predicting individual mean scores in level 2 and entities mean scores in level 1.

$$\begin{aligned} R_1^2 &= 1 - \frac{\sigma_e^2(\text{candidate}) + \sigma_u^2(\text{candidate})}{\sigma_e^2(\text{null}) + \sigma_u^2(\text{null})} \\ R_2^2 &= 1 - \frac{\sigma_e^2(\text{candidate})/n + \sigma_u^2(\text{candidate})}{\sigma_e^2(\text{null})/n + \sigma_u^2(\text{null})} \end{aligned} \quad (3.12)$$

In equation 3.12,  $n$  signifies the number of individuals. In the case,  $n$  is 6165.  $\sigma_e^2$  is the variance component at level 1 and  $\sigma_u^2$  is the individual variance of scores. Unlike the  $R^2$  of the OLS regression model, it can be negative when the variance estimation is close to zero, especially when there is a negative correlation. For positive values, high pseudo  $R^2$  means a good fit with the data sample. It is opposite with the negative pseudo  $R^2$ , lower is better.

### 3.3.4 Likelihood Ratio Tests

There is another common used hypothesis testing method: Likelihood Ratio Tests (LRTs). It was proposed by Neyman-Pearson to compare two nested models. LRTs are used to test the improvement of random slope models. The assumption is that increasing the number of parameters improves the performance of models by random predictors substituting for fixed predictors. In the case, LRT is applied to test the assumption. The format of the LRT suggested by its name, is mainly consisted of the ratio of two models' maximum likelihood value.

$$LRT = -2 \ln\left(\frac{L_1(\hat{\theta})}{L_2(\hat{\theta})}\right)$$

According to equation 3.8, the formula could be reformed to the following form.

$$LRT = -2(\ln L_1(\hat{\theta}) - \ln L_2(\hat{\theta})) = \text{deviance}_1 - \text{deviance}_2 \quad (3.13)$$

$L_1$  is the likelihood of the data sample without any assumption. Strictly speaking,

it is the maximum value with all the parameters unrestricted under the maximum likelihood estimates. While  $L_2$  is the value from a model where the parameters are restricted as random rather than being fixed.

With degrees of freedom and the approximation, Chi-Square distribution test is normally followed within the next step. The right-tailed test rejects the hypothesis if  $LRT > z$ .

$$\alpha = P(LRT > z) = 1 - P(LRT < z) \approx 1 - F(z)$$

$$z = F^{-1}(1 - \alpha)$$

$z$  is a specified critical value for the chi-square test. As illustrated above, it is approximately equalled to the inverse of the CDF. Actually, the process is complicated and needs a large computation power. Hence, it is done relying on the function "lrttest" in the R environment.

# Chapter 4

## Data Preparation

Growing Up in Ireland (Child Cohort) data set is obtained from the Irish Social Science Data Archive (ISSDA). ISSDA provides a wide range of quantitative data sets in the social sciences field mainly for research and educational purposes. The longitudinal data is nested within people across 3 waves. Reading attainment is the response variable. Compared with other kinds of test scores, the reading test scores are the most consistent variables across three waves. The chosen responses from three waves measuring children’s reading ability are listed in the following table.

Table 4.1: The Dependent Variable

Waves	Variable Name	Description of Label
wave1 [19]	readings	Drumcondra Reading Test
wave2 [20]	vrls	Drumcondra Verbal Reasoning Test
wave3 [21]	CognitiveVocabularyTotal	Cognitive Test-Vocabulary Test

Two non-identical and age-appropriate tests were recorded in GUI dataset. They are the Drumcondra Primary Reading Test (DPRT) in wave 1 and 2 and Cognitive Ability Tests in wave 3. These are similar tests both indicating academic and intellectual performance in Irish primary schools. DPRT is designed for primary school students and divided into six levels with parallel forms. The cognitive test is completed by post-primary school teenagers. It included a set of tests: vocabulary test, semantic fluency test and mathematics test. In this research, the vocabulary test is the focus.

## 4.1 Data Preprocessing

To study longitudinal effects, combining the inconsistent data sets according to "ID" indicator is necessary. The merged data frame only contains individuals who participated 3 waves. It has 6165 observations and 2983 variables. Besides, it is intending to quantify the distribution of the response variable and the factors.

Since real-world data has many incomplete attributes and errors. Transforming data into a clean and clear format is the first step before fitting into models. Otherwise, the vast noise and outliers of raw data may lead to an inaccurate inference.

### 4.1.1 Data Cleaning

The method of identifying noise is plotting box plots. Due to the pattern of the response variables, items outside the normal range are considered as noise and replaced with null values. It is assumed that there are no outliers.

In the data set, 2981 (99.9%) out of 2983 variables have missing data. With respect to the dependent variable, there are 122 (1.98%), 410 (6.65%) and 28 (0.45%) missing values in Wave 1, 2 and 3 separately. Owing to the mass missingness in other variables, simple k-nearest neighbours (KNN) algorithm is not suitable for imputing the response variable. Iterative imputation is conducted in variables with over 100 missing data. As for variables with less missing values, Random imputing method is adopted. 1% of random imputing data will not cause the biased parameter estimates in the modelling process.

Multivariate imputation by chained equations (MICE) is the key algorithm in conducting iterative imputation. The first step is finding and including related variables by the correlation coefficient. There will be a detailed introduction of how to calculate it in section 4.2. Then it is followed by replacing every missing observation in each variable with its mean value. The third step is building a linear regression model for one variable. And then this entails using the newly imputed values for other variables, cycling through each of the variables and looping through five iterations [22]. As seen in Figure 4.1, the response variables are normally distributed. Random imputation is drawn from the Gaussian distribution:  $X \sim \mathcal{N}(\mu, \sigma^2)$ .

Most of the related features to predictors are categorical. Replacing missing cells with the most frequent value is the common statistical strategy. Although it may

increase the risk of biased Bayesian inference. The proportion of such replacement is small, ranging from 0.4% to 3%.

### 4.1.2 Data Scaling

The outcome variables are scores of the reading tests. However, they were sat in different levels of the test according to which class and year they were in. Comparing the reading achievement of children from different level is tricky. The DPRT is a standardised measurement developed specifically for group administration. Since there are gaps between groups of subjects, data scaling and standardisation methods are performed to ensure unbalance of the sample. Data collectors have scaled the scores based on an expected posterior calculated by two parameters of difficulty and discrimination. And then they standardised these reading score of DPRT. Therefore, it enables the comparison between children in different level.

Unlike DPRT, the result of Cognitive Ability Tests is a rank from 2 to 17. So the response variable from wave 3 is a category and is not standardised. In order to cope with the consistency of previous outcomes, the standardisation process of Cognitive test scores was proceeded by the scikit-learn package using the equation of 4.1.

$$z_i = \frac{x_i - \mu}{\sigma} \quad (4.1)$$

The distribution of standardised response variables is displayed through quartile ranges with the box plot. It sorts values based on the normal distribution and identifies the outliers. The little rectangle is generated by the 25th percentile (Q1), median value and 75th percentile (Q3).

$$IQR = Q3 - Q1. \quad (4.2)$$

Inter Quartile Range (IQR) is calculated by the difference between the upper and the lower quartile according to equation 4.2. The lower limit ( $Q1 - 1.5 * IQR$ ) is the horizontal line below the rectangle. The upper limit ( $Q3 + 1.5 * IQR$ ) is the horizontal line above the rectangle. Little dots below the lower limit or above the upper limit are considered as the outliers.

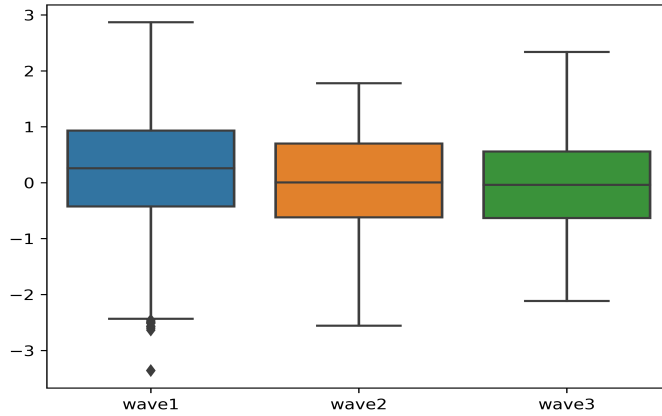


Figure 4.1: The Box Plot of Reading Achievement

The summary of central tendency is described in Figure 4.1. There are only three outliers in wave 1. The overall trend is a decrease when children were 13-year-old and an increase when they turned to 18-year-old. Individual changes will be further discussed in the following section.

### 4.1.3 Data Transformation

Data reduction is essential in dealing with a high dimensional data structure. Little information could be extracted from variables with too many missing values. In addition, more time and computation are needed to impute and recover the missing data. Hence, dimension reduction is carried out to handle thousands of features. The approach is removing variables if the missing value ratio is greater than 30%.

The desired data structure is a long format. For multilevel analysis, the input has to be hierarchical sorted data. Hence, it is needed to reshape and sort the data frame first by students ID and then by wave indicator. After transforming from a wide to a long style, the input data structure is shown in Table 4.2.



Table 4.2: Long Format of Data Structure

ID	Wave	household	phone	books	Reading
1000	1	1	0	3	-1.301237
1000	2	1	1	3	-0.395305
1000	3	1	1	3	-0.630139
3000	1	1	0	4	0.177569
3000	2	1	1	2	0.525956
3000	3	1	1	2	0.854718

## 4.2 Feature Selection

The Pearson Correlation Coefficient (PCC) is calculated to find interesting factors which may affect the reading ability. The formula 4.3 measures the strength of linear relationships between X and Y variables. The PCC is a symmetric measurement that falls between -1 and 1. A value of 0 indicates no linear correlation between two explanatory variables. Judging the correlation just by the PCC is arbitrary. Correlation matrix in Figure 4.2 is a mere reference.

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (4.3)$$

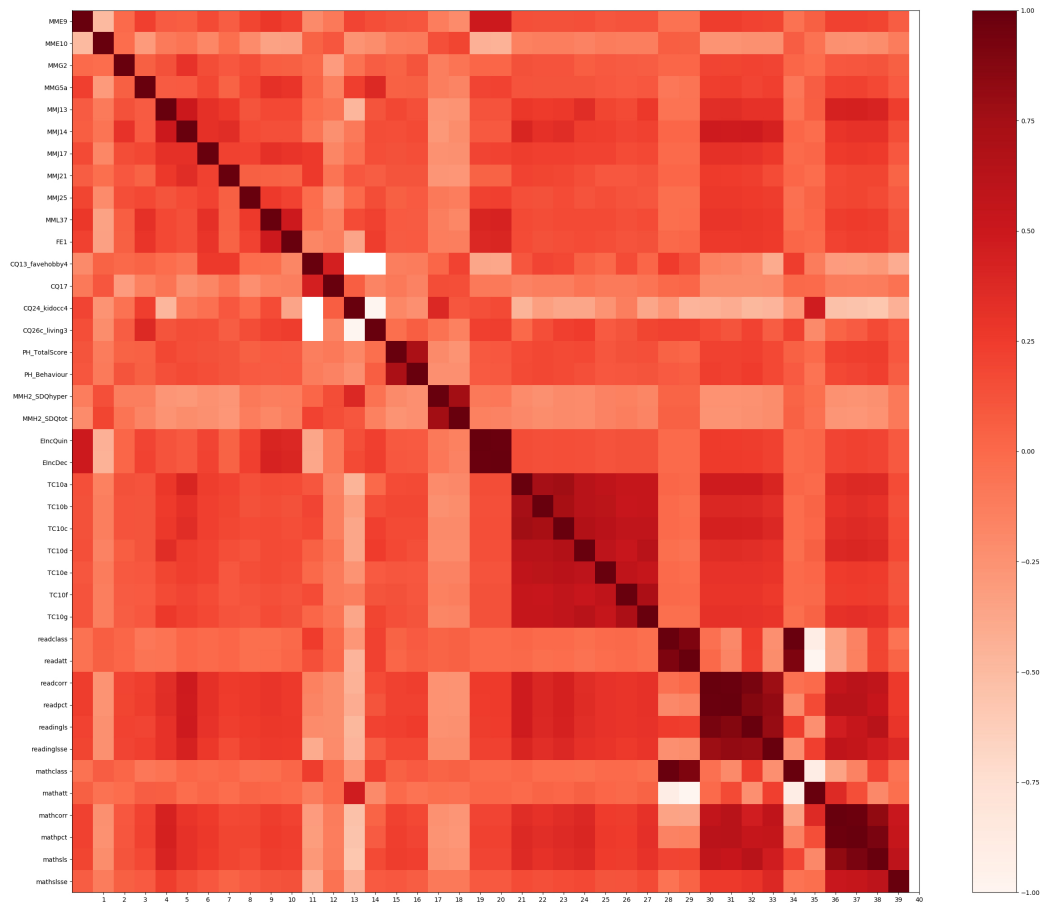


Figure 4.2: Correlation Matrix of Logit Score in Drumcondra Reading Test

Choosing independent variables requires a careful consideration from all aspects. No self-correlation inside explanatory variables is the basic principal of linear regression models. In multilevel modelling, ensuring the independence of individuals is the first requirement. Combining the results of PCC, advise of professionals from GUI study and previous literature review, the final decision of explanatory variables appears in Table 4.3. Because too many missing values in the two predictors of parents' expectation and household income, the rest three independent variables are chosen as the input of the models.

Table 4.3: The Potential Explanatory Variables

Description	Wave	Variable Name	Missingness
Access of	1	MMJ25	0 (0%)
Books	2	pc2e24	175 (2.8%)
Expectation of Parents	1	MMJ17	0 (0%)
	2	pc3c3	181 (2.9%)
	3	sc3b1	2171 (35.2%)
Household	1	Equivinc	440 (7%)
Annual	2	w2equivinc	600 (10%)
Income	3	w3equivinc	632 (10%)
Household Type	1	hhtype4	0 (0%)
	2	w2hhtype4	172 (2.8%)
	3	w3hhtype4	52 (0.8%)
Phone	1	CQ18	26 (0.4%)
Ownership	2	cq2q18	221 (3.6%)

## 4.3 Exploratory Data Analysis

To preliminarily understand the data, Exploratory Data Analysis (EDA) is frequently conducted. The most common approaches are the calculations of mean and standard deviation.

### 4.3.1 The Target Variable

However, it is not clear how the reading ability of individuals changes over time. Consequently, first 100 observation from sample data is plotted in Figure 4.3 to make assumptions. And it is assumed a linear trend of reading score across waves.

Table 4.4: Descriptive Statistics of Response Variables

Waves	Length	Mean	Standard Deviation
wave1	6165	0.2293	0.9711
wave2	6165	0.0340	0.8609
wave3	6165	-0.0013	1.0009
Total	18495	0.0873	0.9516

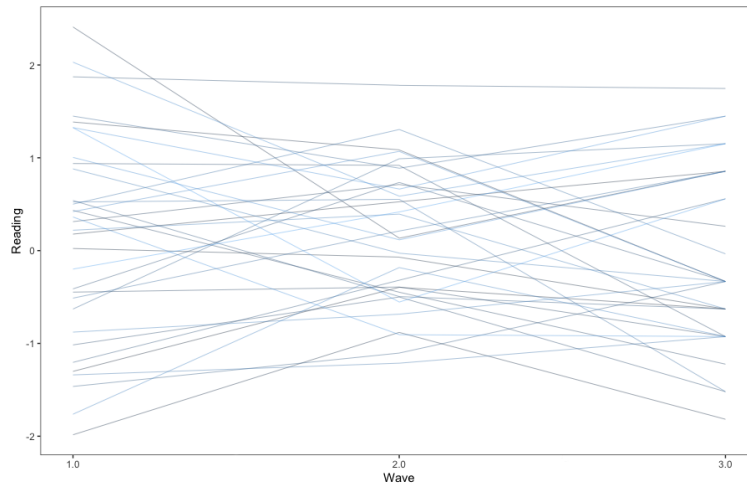


Figure 4.3: Trend of 100 Observations

According to Table 4.4, the overall mean decreases constantly. But as discussed in the 4.1.2 section, the trend goes down and up. The assumption is that reading achievement changes within the three waves. Actually, Figure 4.3 indicates a slight growth over the three time points despite a drop in wave 2. It can be assumed there is a growth when children get older. And a linear growth model based on the assumption can be built.

The distribution of independent variables from each wave is depicted in the density plots. The figures are plotted by the probability density function (PDF). This function explains the relative likelihood of retaining a given value of independent variables. As shown in Figure 4.4, they follow a left skewed distribution. The response variable in wave 3 displays the most obvious negative skew among them. This means the overall performance is negative. Because of separate tests, the pattern does not represent that the reading ability of the individual shows a declining trend over time. Therefore, it is demanded to apply the multilevel modelling method and dive deeper into a deeper analysis.

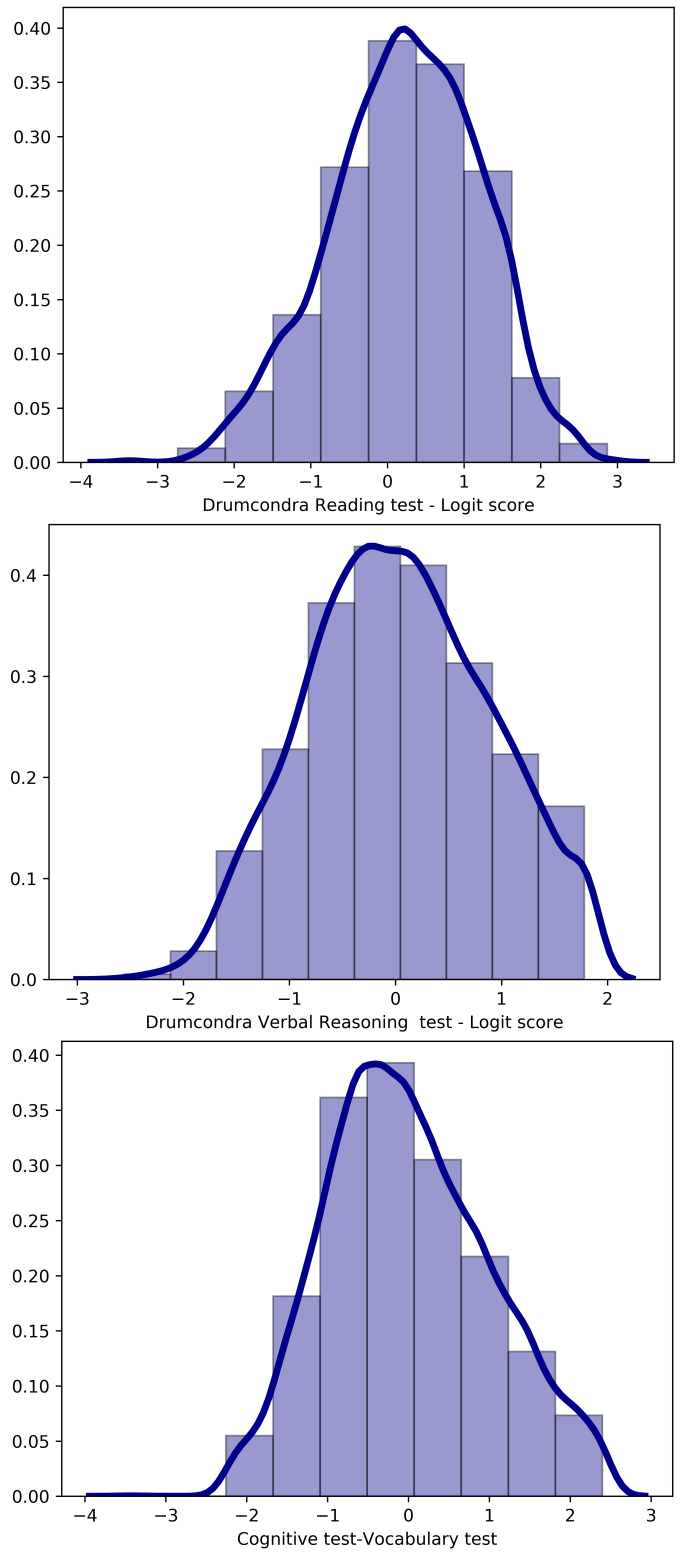


Figure 4.4: The Distribution of the Response Variables

### 4.3.2 Explanatory Variables

According to the correlation matrix, the factor of books was assumed to have a strong linear relationship with the target variable. Household type and phone ownership variables were chosen in investigating the impact of owning a mobile phone [5]. In this study, these three variables are the explanatory variables. Later on, they will be put into the models and compared with each other. Due to the inconsistency of data, it is vital to visualise the explanatory variables. Otherwise, it could cause inaccurate inference. For instance, the value of 1 in the phone ownership means a no in wave 1, but indicates a yes in wave 2. So checking and reordering these factors is essential.

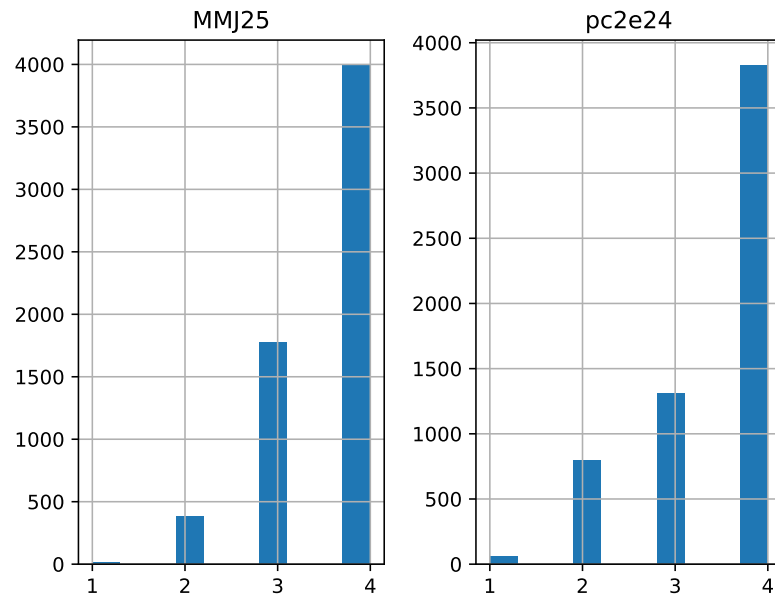


Figure 4.5: The Frequency of Books

Figure 4.5 conveys the frequency of the number of books in home. The value ranks from the first class through to the end of fourth class: zero, less than ten, ten to thirty and more than thirty books. Roughly 65% of children have access to more than 30 books in home. Compared with values in wave 1 and 2, there is a slight decrease in the number of books. It means overall children had fewer books when they got older. The general environment was friendly in developing the reading interest of children.

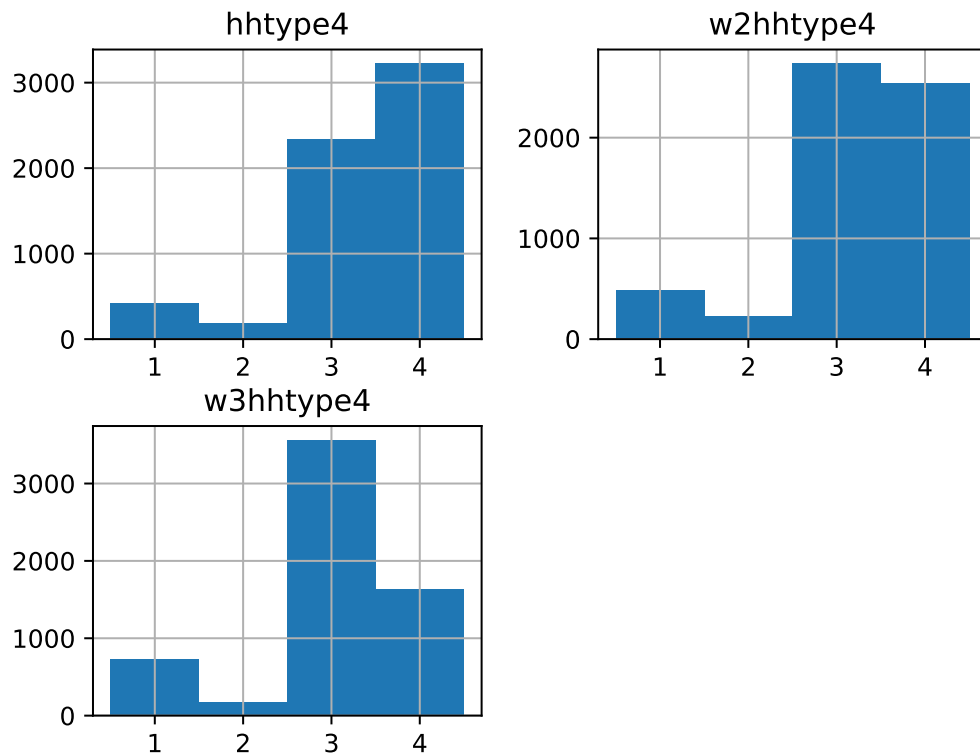


Figure 4.6: The Frequency of Household Type

After re-ranking as a difficulty of raising children, the frequency of household type is shown in Figure 4.6. The input is a categorical variable: 1 (Single with more than 3 child), 2 (Single with 1 to 2 child), 3 (Couple with more than 3 child) and 4 (couple with 1 to 2 child). Most families were a married couple and they were having more child between 2007 to 2016 in Ireland.

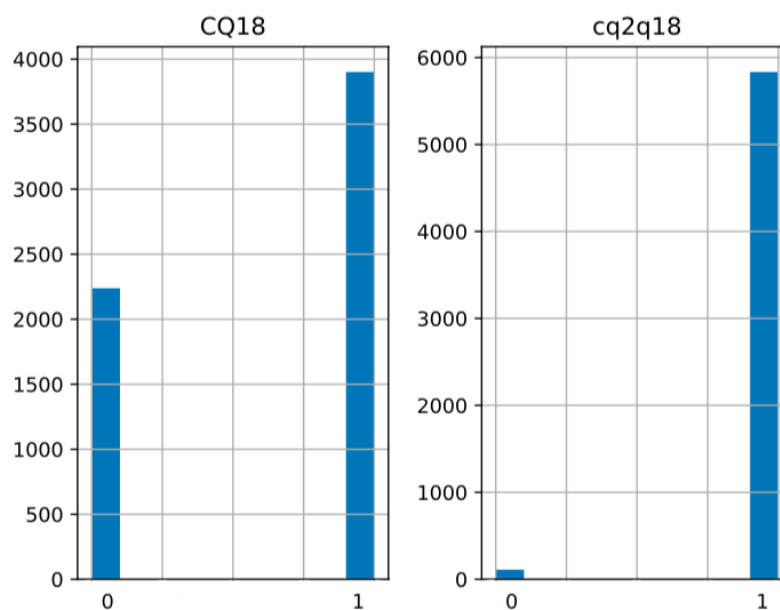


Figure 4.7: The Frequency of Phone Ownership

Phone ownership is a binary variable. It comes from the question at the first two waves "Do you have your own mobile phone?". As shown in Figure 4.7, the ratio of not owning a mobile phone and owning a mobile phone is 1:2 in 9 years old. And when children turned to 13, the proportion of owing mobile phone boosted to 99.9%. It suggests a high level of mobile phone ownership in the 13-year-old group. It could be assumed that almost every child had a phone since then. Although the phone ownership is missing for wave 3, using 1 for creating a new indicator is natural. The change of phone ownership could be described as that phone ownership of children is growing rapidly overages in Ireland.



## Chapter 5

# Implementation of Multilevel Modelling

Chapter 5 includes a guideline for employing the efficient multilevel models. For longitudinal data, a 2-level mixed-effect model is established with repetitions of waves as level 1 units and respondents as level 2 units. There are two requirements: (1) The size of data sample is large (2) Many subjects are at level 2. Otherwise, the variances could be negative. All models are fitted into a general statistical inference framework IGLS. The estimation will converge after several iterations. The general procedure is starting with a variance components model and then building more complex models with adding predictors. Then it is followed by checking the substantial fix or random coefficients. The later models are derived from the variance components model. Variance components model estimates the variation covered by the hierarchical data structure - two levels. [23]

MLwiN is used for fitting multilevel models. It is a software program developed by the University of Bristol and has become established to be the most advanced one in multilevel modelling [24]. In the study, the R package of R2MLwiN is implemented to run the MLwiN software from the R environment [25].

## 5.1 A Variance Components Model

This variance components model, considered as a null model, is a baseline model for comparing with more elaborate models. Building the model is particularly for univariate repeated measures. It is generalised according to variance between-occasion and within-person. Variance components here refer to between-individual variance and between-wave variance. A two-level model with the reading score as the response variable and constant 1 as the only explanatory variable is set up first. The overall variance of the outcome can be partitioned into components for child and wave.

$$\begin{aligned} \text{Reading}_{i_j} &= \beta_{0_{i_j}} \text{constant} \\ \beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\ u_{0_j} &\sim \mathcal{N}(0, \sigma_{u_0}^2) \\ e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2) \end{aligned} \tag{5.1}$$

where  $\text{Reading}_{i_j}$  is the reading ability score at  $i$ th measurement wave for the  $j$ th student.  $\beta_0$  is the overall mean value of students.  $u_{0_j}$  serves as the individual-level random effect while  $e_{0_{i_j}}$  corresponds to the wave-level residual error. The variance of the reading score is the sum of level 1  $\sigma_e^2$  and level 2  $\sigma_u^2$ :  $\text{var}(y|\beta_0) = \text{var}(u_{0_j} + e_{0_{i_j}}) = \sigma_u^2 + \sigma_e^2$ . In order to vividly depict the implementation, a figure corresponding to the equation is shown below.

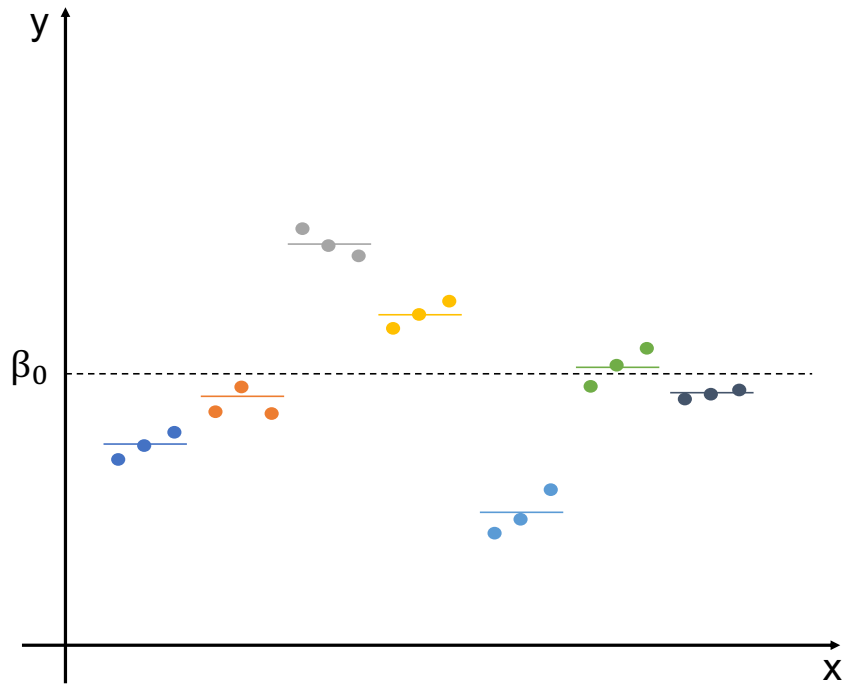


Figure 5.1: The Variance Components Model

The variance components model does not have a slope because there is no predictor included. There are seven subjects in Figure 5.1. Actually, 6165 horizontal lines representing 6165 students are created in the variance components model. Every three observations are clustered in one group (student). These three dots indicate the performance of the reading score from the same individual. The first dot means the reading score of children 1 in his or her 9 years old. The second dot is the score when children turned to 13. The final dot represents the performance of a cognitive reading test in wave 3.  $e_{0_{i_1}}$  is residual in estimating the mean value of the first student.  $u_{0_1}$  is the distance of the overall mean  $\beta_0$ . These estimated horizontal line means no growth in students' reading achievement.

Intra-level-2-unit correlation refers to the correlation between students. It can be expressed by the Intra-class Correlation Coefficient (ICC), so-called Variance Partition Coefficient (VPC), and is calculated according to equation 5.2. The role of VPC is examining the relation of two levels. It provides statisticians with the significance of the hierarchical data structure. If the VPC is greater than the general threshold of 0.05,

the nested data could not be ignored and it is necessary to build multilevel models. Otherwise, it is supposed to be a model with only one level. [26]

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \quad (5.2)$$

### 5.1.1 Estimation of Residuals

Two-level mixed model 5.1 has two kind of residuals at two levels. The parameter estimates of  $u_{0_j}$  and  $e_{0_j}$  are along the following equation 5.4. Residuals have two general interpretations. The first function is acting as parameter values of the distribution explaining the variation among the level 2 units. In addition, it could provide an efficient estimation of the fixed coefficients. Another role is as individual estimates for each level.

$u_{0_j}$  is in a second level and  $e_{0_j}$  is in a first level of variance components model. It is known the general equation is 5.3

$$u_{\hat{0}_j} = E(u_{0_j} | Y, \hat{\beta}_0, \hat{\Omega}) \quad (5.3)$$

In the variance components model:

$$\begin{aligned} u_{\hat{0}_j} &= \frac{n_j \sigma_u^2}{n_j \sigma_u^2 + \sigma_e^2} \tilde{y}_j \\ e_{\tilde{0}_j} &= \tilde{y}_{i_j} - u_{\hat{0}_j} \\ \tilde{y}_j &= \frac{\tilde{y}_{i_j}}{n_j} \end{aligned} \quad (5.4)$$

In order to MLE normally distributed values, it derives correlations between subjects and times within the reading scores. IGLS is conducted the random part under the condition of knowing the regression coefficients. Assuming that the residuals follow the normal distribution, IGLS is a suitable method for estimating the values of the residuals.

## 5.2 A Linear Growth Curve Model

Based on the variance components model, linear growth curve model increases the complexity by including fixed explanatory variables and extending them at the student level. The process is generally fitting a 2-level continuous response model with predictors. Variation in the individual level provides a separate curve for each child. The between waves within individual variation is estimated in the wave level. Since the gap between waves is a period of four years, which is quite large. It is considered that the residuals in the wave level are independent and not correlated. Estimation of coefficients in the wave level involves only a few parameters. As time is eternal and will not be affected by other factors.

First, the model includes each predictor in the fixed part. Then a random-effect model is extended based on the previous model considering factors as random effects. It is by adding factor, for example, phone ownership in level two. It is worth noting that VPC is no longer fixed because the random slopes lead to heteroscedasticity. These two models are compared by carrying out likelihood ratio tests on 2 degrees of freedom. If there is distinguished variation between the individual in their linear growth rates, the likelihood drops significantly. And the best one is chosen for factors comparison and the final model construction. Based on the previous results of four models, the final model will contain factors in the fixed or in the random part.

### 5.2.1 Models with Explanatory Variables

The variance components model is added with explanatory variables of books, household type and phone ownership. Therefore, three new random intercept models are built. It is considered that the factors are fixed effects.

$$\begin{aligned} \text{Reading}_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_1 X_{i_j} \\ \beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\ u_{0_j} &\sim \mathcal{N}(0, \sigma_{u_0}^2) \\ e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2) \end{aligned} \tag{5.5}$$

Combining the baseline model and linear regression model, predictors are fitted with the fixed slope coefficient  $\beta_1$  in equation 5.5. Variance and residuals in level one and

level two fluctuate correspondingly. The detailed change will be further discussed chapter 6.

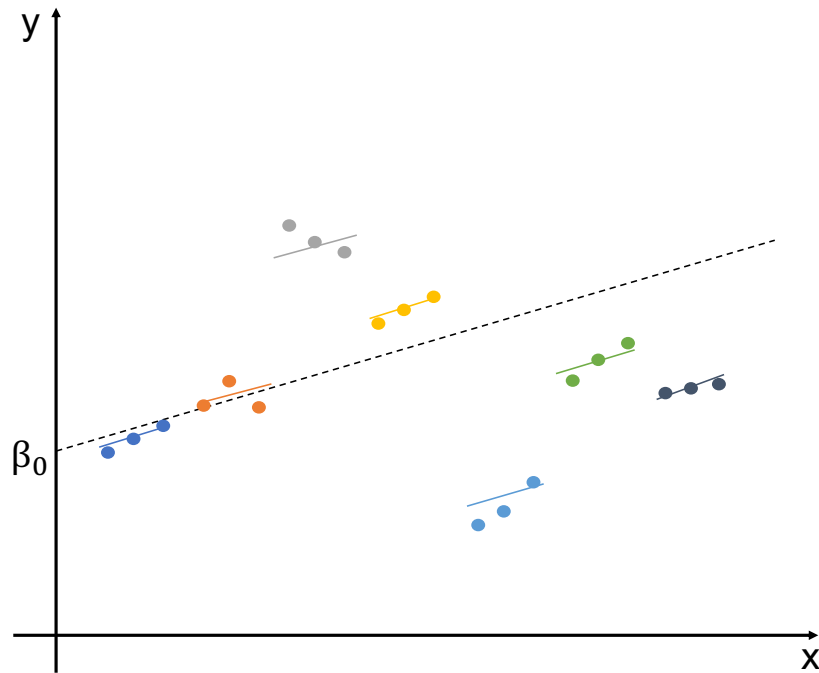


Figure 5.2: The Random Intercept Model

These three models could be regarded as random intercept models with a single slope. Variation in the level 2 contributes to the different intercepts. Since the slope is fixed and intercepts are random. Compared with the 5.1, 6165 horizontal lines change to parallel lines. They share the same slope accounted for the predictor. However, when estimating the response variable, the error inside of each subject is considerably high.

### 5.2.2 Random Slopes Models

Rather than factors to be fixed, coefficients of slopes are assumed to be random. It accords with the assumption that the linear growth rate varies among student. If the result of the likelihood ratio test between candidate and baseline model is significant, the factor dose vary from student to student. So the coefficient of this predictor will

be set random. Else the model stays in the previous stage - random intercept model with fixed predictors.

$$\begin{aligned}
\text{Reading}_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_{1_j} X_{i_j} \\
\beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\
\beta_{1_j} &= \beta_1 + u_{1_j} \\
\begin{bmatrix} u_{0_j} \\ u_{1_j} \end{bmatrix} &\sim \mathcal{N}(0, \Omega_u); \Omega_u = \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{0_1}}^2 \\ \sigma_{u_{0_1}}^2 & \sigma_{u_1}^2 \end{bmatrix} \\
e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2)
\end{aligned} \tag{5.6}$$

In equation 5.6, it is worthy noting that a new sub-equation is appended for the random coefficient  $\beta_{1_j}$  and the coefficient  $\beta_{1_j}$  now has a subscript j. It suggests that the slope coefficient varies between students. Simultaneously, the dimension of the covariance matrix in wave 1 grows from  $1 \times 1$  to  $2 \times 2$ .  $\beta_1$  is the mean growth rate deviates according to  $\sigma_{u_1}$ .  $u_{1_j}$  refers to the random effect residual of X factor.  $\sigma_{u_1}^2$  indicates the variance of the slope  $\beta_{1_j}$ . While  $\sigma_{u_{0_1}}^2$  is the covariance of intercept and the independent variable.

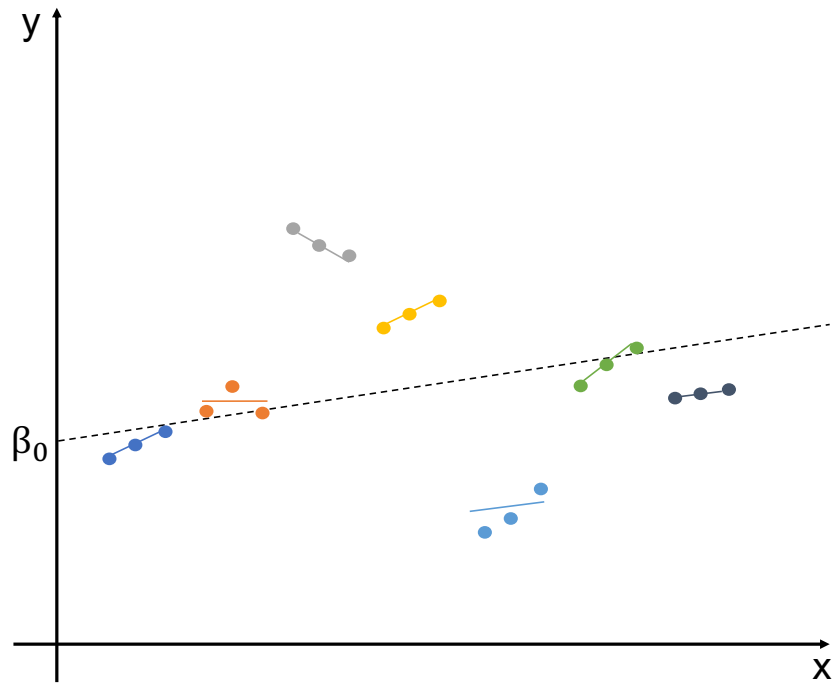


Figure 5.3: The Random Slopes Model

In Figure 5.3, the slopes are interpreted as the growth rate of each subject. Since there is a variation of the predictor between individual. The slopes are random and calculated by the covariance matrix  $\Omega_u$ . On the basis of Figure 5.2, the general dashed line changes. In the meantime, other small 6165 lines vary from each other in both intercepts and slopes.

### 5.3 Model Building

There are three models for each factor for comparing their impact on the reading scores. The general process is including the number of books, household type and phone ownership to the baseline model separately. When each set of explanatory variables added to the model, the model improves at accurately predicting the response variable for each case. The building models process is separated into two steps generally. The step is building a random intercept model 5.5 with a factor based on the variance components model 5.2. It is followed by checking the p-value of the predictor. If the p-



value is significant, building a random slope model 5.6 based on the model built in step 1 is the next step. Otherwise, it is unnecessary to proceed the next step and include the factor in the final model building. Linear growth rate varies between individuals around its mean value. Rather than the slope is fixed, the slope coefficient of each significant predictor is random.

### 5.3.1 A Model with Books Predictor

A random intercept model is built first according to equation 5.5. The overall mean value of observations is -0.6117 within the confidence interval of [-0.6833, -0.5402]. And the value of fixed slop is 0.1981 within the confidence interval of [0.1786, 0.2176]. When checking the significance of the slope and intercept parameter, the p-values calculated from the Z test are both smaller than 0.05. The null hypothesis is that the predicted and original values are equal. It turns out that the null hypothesis should be rejected due to the small p-value. So predictor books is an important predictor.

To test the growth rate changing from student to student, the slope of books factor is set to be random. In other word, number of books is considered as a random effect based on the previous random intercept model. The likelihood ratio test is conducted between the random intercept model and random slope model:

Table 5.1: Likelihood Ratio Test of Model 2

LogLikelihood	DF	Chi-Squared	P-Value
-22096			
-22061	2	69.441	***

DF: Degrees of Freedom

Significant: \*\*\* (0), \*\* (0.01), \* (0.05)

Non-Significant: "." (0.1), " " (1)

A likelihood rises from -22096 to -22061 as shown in Table 5.1. It is concluded a significance according to a Chi-squared distribution on 2 degrees of freedom. There is a significant variation of the linear growth rate on the individual level. With the factor

of books included as the random effect, the equation of the final model 2 is as below:

$$\begin{aligned}
Reading_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_{1_j} \text{books}_{i_j} \\
\beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\
\beta_{1_j} &= \beta_1 + u_{1_j} \\
\begin{bmatrix} u_{0_j} \\ u_{1_j} \end{bmatrix} &\sim \mathcal{N}(0, \Omega_u); \Omega_u = \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{0_1}}^2 \\ \sigma_{u_{0_1}}^2 & \sigma_{u_1}^2 \end{bmatrix} \\
e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2)
\end{aligned} \tag{5.7}$$

### 5.3.2 A Model with Household Predictor

Basing on the variance components model, the random intercept model is constructed with the factor household type. After 4 iterations, the estimate for the intercept is 0.0631. The p-value of the intercept parameter is 0.00825. While the p-value of the household type is 0.2609. Since one p-value is greater than 0.05, the null hypothesis can not be rejected. It is concluded that the household type factor is not significant in predicting the reading score.

From the result above, changing the model 3 to the random slope model is unnecessary. Since further improvement dose not make any difference. Model 3 should include the household type factor in the fixed part. The equation of model 3 is as below:

$$\begin{aligned}
Reading_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_1 \text{household}_{i_j} \\
\beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\
u_{0_j} &\sim \mathcal{N}(0, \sigma_{u_0}^2) \\
e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2)
\end{aligned} \tag{5.8}$$

### 5.3.3 A Model with Phone Predictor

The general process of building a linear growth model with phone factor is similar to the previous models. The first step is building a model with phone predictor based on the variance components and linear regression model. Then changing the coefficient of the slop to be random is the next step. In the random intercept model, the intercept is 0.2406 in the confidence interval between 0.2079 and 0.2734. The estimate of the slope

is -0.1756 within the confidence interval of [-0.205, -0.146]. The result of the t-test is that p-values close to the zero. Such small p-values show strong evidence against the null hypothesis. Adding the predictor of phone ownership to the baseline model improves the performance.

Table 5.2: Likelihood Ratio Test of Model 4

LogLikelihood	DF	Chi-Squared	P-Value
-22218			
-22207	2	21.504	***

DF: Degrees of Freedom

Significant: \*\*\* (0), \*\* (0.01), \* (0.05)

Non-Significant: "." (0.1), " " (1)

Similarly, the likelihood decreases after including phone predictor as a random explanatory variable. Hence, the growth ratio indeed varies between children and the slope should not be fixed. The structure of the fourth model with phone ownership is the same as the second model. The only difference with model 2 is the explanatory variable, using phone ownership instead of predictor books.

$$\begin{aligned}
 Reading_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_{1_j} phone_{i_j} \\
 \beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\
 \beta_{1_j} &= \beta_1 + u_{1_j} \\
 \begin{bmatrix} u_{0_j} \\ u_{1_j} \end{bmatrix} &\sim \mathcal{N}(0, \Omega_u); \Omega_u = \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{0_1}}^2 \\ \sigma_{u_{0_1}}^2 & \sigma_{u_1}^2 \end{bmatrix} \\
 e_{0_j} &\sim \mathcal{N}(0, \sigma_{e_0}^2)
 \end{aligned} \tag{5.9}$$

### 5.3.4 The Final Multilevel Model

A final model combines the results of previous models and is considered as a robust multilevel model. In this model, it enables studying and comparing the effects of possible predictors with the growth rate. Model 5 dose not consider the household variable because this predictor is not significant. The final model only includes the

books as well as phone ownership variables and considers them as the random effects.

$$\begin{aligned}
Reading_{i_j} &= \beta_{0_{i_j}} \text{constant} + \beta_{1_j} \text{books}_{i_j} + \beta_{3_j} \text{phone}_{i_j} \\
\beta_{0_{i_j}} &= \beta_0 + u_{0_j} + e_{0_{i_j}} \\
\beta_{1_j} &= \beta_1 + u_{1_j} \\
\beta_{3_j} &= \beta_3 + u_{3_j}
\end{aligned} \tag{5.10}$$

$$\begin{bmatrix} u_{0_j} \\ u_{1_j} \\ u_{3_j} \end{bmatrix} \sim \mathcal{N}(0, \Omega_u); \Omega_u = \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{0_1}}^2 & \sigma_{u_{0_3}}^2 \\ \sigma_{u_{0_1}}^2 & \sigma_{u_1}^2 & \sigma_{u_{1_3}}^2 \\ \sigma_{u_{0_3}}^2 & \sigma_{u_{1_3}}^2 & \sigma_{u_3}^2 \end{bmatrix}$$

$$e_{0_j} \sim \mathcal{N}(0, \sigma_{e_0}^2)$$

As seen in the above equation, three factors are composed of the final model.  $\beta_{0_{i_j}}$  is generated from the overall mean of students and residuals variation in student and wave level. The overall average is predicted by the fixed parameter  $\beta_0$ .  $\beta_{1_j}$  and  $\beta_{3_j}$  are random coefficients and vary amid the student level. They are controlled by the variance and covariance in only in the student level but not in the wave level. The dimension of covariance matrix  $\Omega_u$  is now a  $3 \times 3$  because two random factors are included. There is a new interaction between predictor books and phone.

# Chapter 6

## Results and Discussion

### 6.1 Model Diagnostics

Model diagnostics aim at testing models assumptions and investigating systematic misfit of the data sample. Only on the condition of valid assumptions, the statistical analysis is credible and a reasonable interpretation could proceed.

#### 6.1.1 The Baseline Model

In the baseline model, VPC in reading scores between individual within waves is 0.586 calculated by equation 5.2 :  $0.5302/(0.5302 + 0.3753)$ . It means that attainment of 58.6% variance can be attributed to differences of each student among three waves. Such a large value of VPC shows a high degree of resemblance between-subject level and a strong dependence in waves level. In the repeated measures data model, most of the variation is at the second level, which is the subject level. In consequence, the VPC at level 2 is quite high.

Due to the VPC value in the variance components model, the assumption of the nested and hierarchical data structure is true. Especially, there is a variation between individuals among waves. The z score of the intercept is 8.47 and p-value is  $2.49 \times 10^{-17}$ , indicating the significance of the intercept parameters. These estimated parameters are the general results that new models can be built upon. To conclude, the assumption of hierarchy is tested to be valid according to the variation between individuals. Variance components model is a baseline to evaluate the performances of other models.

## 6.1.2 The Final Model

The assumption before applying linear growth curve models is normality of the residuals. In order to examine the normally distributed random effects, various visualisation explains the estimated residuals at both levels. The validation of the final model is tested and illustrated by some of the possible methods.

Normal probability plots are produced to check the assumption: residuals at both level 1 and level 2 follow the Gaussian distributions. If the lines look fairly linear, the assumption is valid.

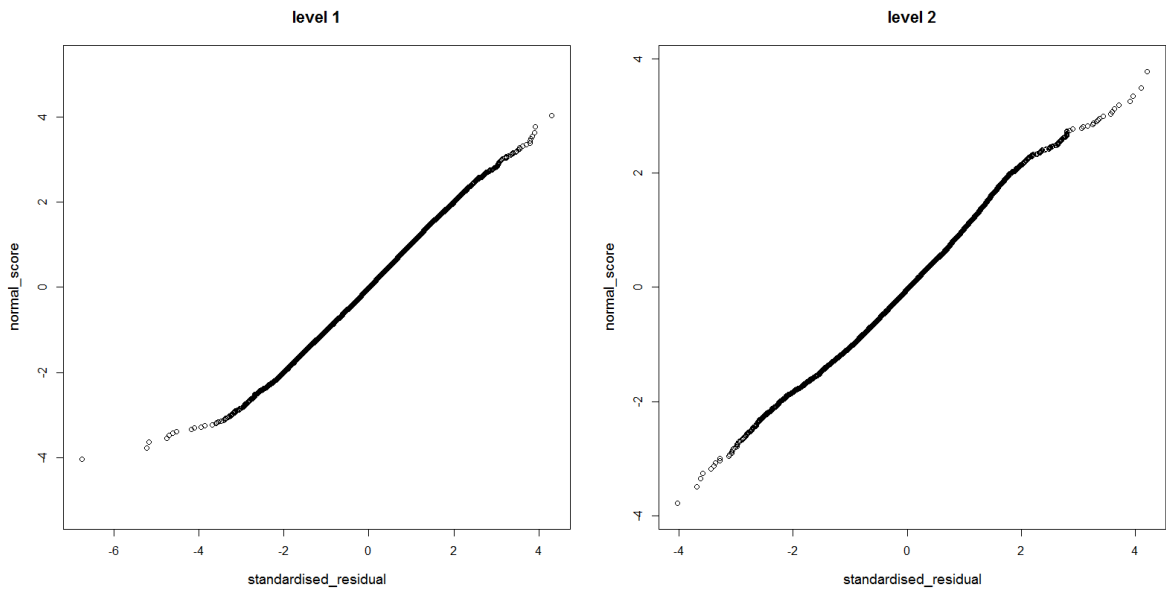


Figure 6.1: Standardised Residuals VS Normal Distribution

As shown in Figure 6.1, ranked residuals are scattered against the interrelated points on a Gaussian distribution curve. The x-axis is the residuals after the standardisation. The y-axis is the values generated from the normal distribution. The left-hand side plot depicts the normality of residuals in the wave level (level 1). The right-hand side plot shows the residuals of the individual level (level 2). In level 1, the most of the line is linear except for the minor tails part. However, the tails of the distribution are partially heavier in level 2 because the upper right is longer than the usual. Further proof of normality at level 2 need to be conducted. After all the dependent variable has been normalised. The normality of residuals at level 1 is reasonable.

Multivariate normality of the random effects could be examined through the following density plots. Density kernel estimation enables smoothing out the noise by a Gaussian kernel smoother. The plots will show smoother distributions more generally.

$$bw = 0.9n^{-\frac{1}{5}} \min(sd(x), \frac{IQR(x)}{1.34}) \quad (6.1)$$

The bandwidth in the density plot is calculated by equation 6.1. Standard deviation is denoted as *sd*. IQR is interquartile range function, calculated by equation 4.2.

Moreover, the density plot is drawn to further check the normality of residual at level 2. IQR of the intercept estimation is 0.4985 when the upper quartile is 0.2641 and the lower quartile is -0.2344.

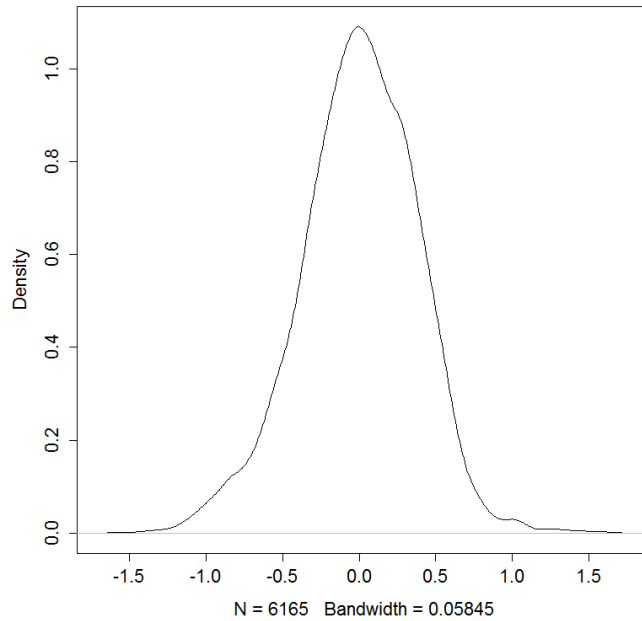


Figure 6.2: The Density Estimation of Intercept Coefficient

Figure 6.2 shows the distribution of the second level residuals. The right tail of the density curve is not smooth between the interval of 1 and 1.7. Generally speaking, the random effect from waves and students indicators is uniformly normal distributed with the mean value of 0 and over a continuous interval of -1.5235 and 1.5915. It corresponds to Figure 6.1. Overall speaking, the rough part is minor and residuals are considered following normal distribution.

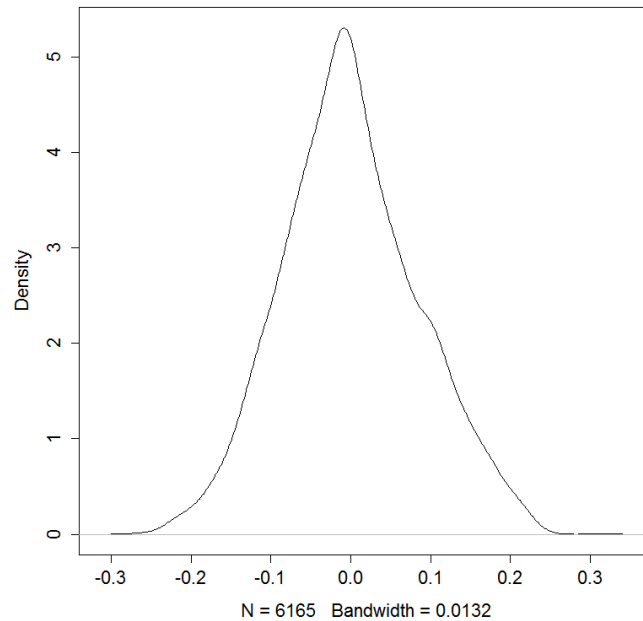


Figure 6.3: The Density Estimation of Books Coefficient

The distribution of the slope coefficient for books predictor is shown in Figure 6.3. ICC is 0.1126 calculated by the first quartile of -0.05739 and the third quartile of 0.5519. The bandwidth is 0.013. Since the books predictor has four values, the density curve is loosely distributed between the range of -0.2737 and 0.3132. The curve is clearly not a skewed distribution. The mean of books coefficient estimation is zero and located at the centre of the plot. Hence, this density curve could be considered symmetric.

In Figure 6.4, the bandwidth calculated by equation 6.1 is 0.006. Since the difference between the upper and lower quartile is 0.054 (=IQR). Compared with the 0.013 of the books estimation, the bandwidth is smaller. On the other hand, the estimated phone coefficient has a narrow distribution between the interval of -0.3542 and 0.5836. Despite the small bandwidth and the narrow peak, the mean value corresponds to the peak identically. The density curve is roughly symmetrical.

All in all, residuals at both levels are normally distributed on the strength of linearity in Figure 6.1. The estimation of slopes for predictors is tenable. According to the above three density plots, the density curves can be seen as symmetrical. The parameters have a zero mean and finite constant variance. Derived from it, the residuals are normally distributed. The previous assumption could hold. It is possible that



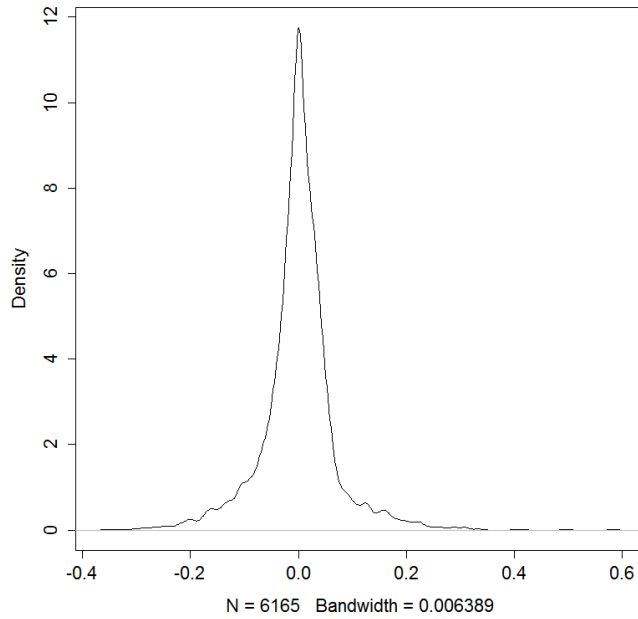


Figure 6.4: The Density Estimation of Phone Coefficient

the true distribution of random effects dose not follow Gaussian distribution while the estimates appear to follow the distribution. Nevertheless, the distribution of explained variance is considered normal.

## 6.2 Model Inference

As from the model diagnostics, the assumption of normality is valid. The model inference, confidence intervals of parameters and model predictions are reasonable. The estimation of coefficients is in Table 6.1. Model 1 is the variance components model. While model 2 and 4 have random slopes of predictor books and phone separately. Model 3 includes the household type as the slope is fixed. The final model (model 5) include both predictors books and phone as random effects.

Z value is calculated and then compared with a critical value. If it is greater than the critical value, the null hypothesis is rejected. A confidence interval (CI) is an estimate of parameters' interval, computed from the z test statistics of the observed data. The true value of the estimated parameter may or may not lies in the interval. There is a 95% probability associates with the reliability of the IGLS estimation procedure.

Table 6.1: Model Estimation

	Multilevel Models				
	Model 1	Model 2	Model 3	Model 4	Model 5
<b>Fixed Part</b>	<b>Coefficient, Standard Error and CI</b>				
Intercept	0.0873	-0.6816	0.0631	0.2354	-0.531
SD	(0.0103)	(0.0348)	(0.0239)	(0.0174)	(0.039)
CI	[0.07, 0.11]	[-0.75, -0.61]	[0.02, 0.11]	[0.2, 0.27]	[-0.61, -0.45]
P-Value	***	***	**	***	***
books		0.2162			0.2115
SE		(0.0101)			(0.01)
CI		[0.2, 0.24]			[0.19, 0.23]
P-Value		***			***
household			0.0104		
SE			(0.0093)		
CI			[-0.01, 0.03]		
P-Value			#		
phone				-0.1704	-0.1544
SE				(0.0159)	(0.016)
CI				[-0.2, -0.14]	[-0.19, -0.12]
P-Value				***	***
<b>Random Part</b>	<b>Coefficient (Standard Error)</b>				
	<b>Student Level</b>				
Variance	0.5302	0.5208	0.5291	0.5996	0.6537
	(0.0119)	(0.0864)	(0.0119)	(0.0271)	(0.1197)
books×Intercept		-0.0774			-0.0762
		(0.0269)			(0.0303)
books		0.0397			0.034
		(0.0087)			(0.0087)
phone×Intercept				-0.0642	-0.1356
				(0.0185)	(0.0407)
phone				0.0856	0.0825
				(0.0197)	(0.0203)
books×phone					0.0187
					(0.0104)
	<b>Wave Level</b>				
Variance	0.3753	0.3703	0.3756	0.3581	0.3561
	(0.0048)	(0.0050)	(0.0048)	(0.0051)	(0.0052)

Significance: \*\*\* (0), \*\* (0.01), \* (0.05), "." (0.1), "#" (1)

## 6.2.1 Residuals Estimation

Residuals at level 2 represent individual departures from the mean value. Given the large number of 6165 subjects in the data sample, it is difficult to plot every student. Caterpillar plots are used to see the residuals from the records of the first 100 students. The residuals are from the second level and re-ranked according to the estimated mean values. In the baseline model 1, the standard deviation is the same throughout the all data sample since there is no predictor. By looking at the confidence intervals of the other three plots, about 50% of residuals overlap 0. It means that students vary from the average at the 5% level, which is significant. In model 2, 60% of residuals at the centre are near zero. 30% of residuals at the lower and upper end of model 4 plot where the confidence intervals do not lap over the zero.

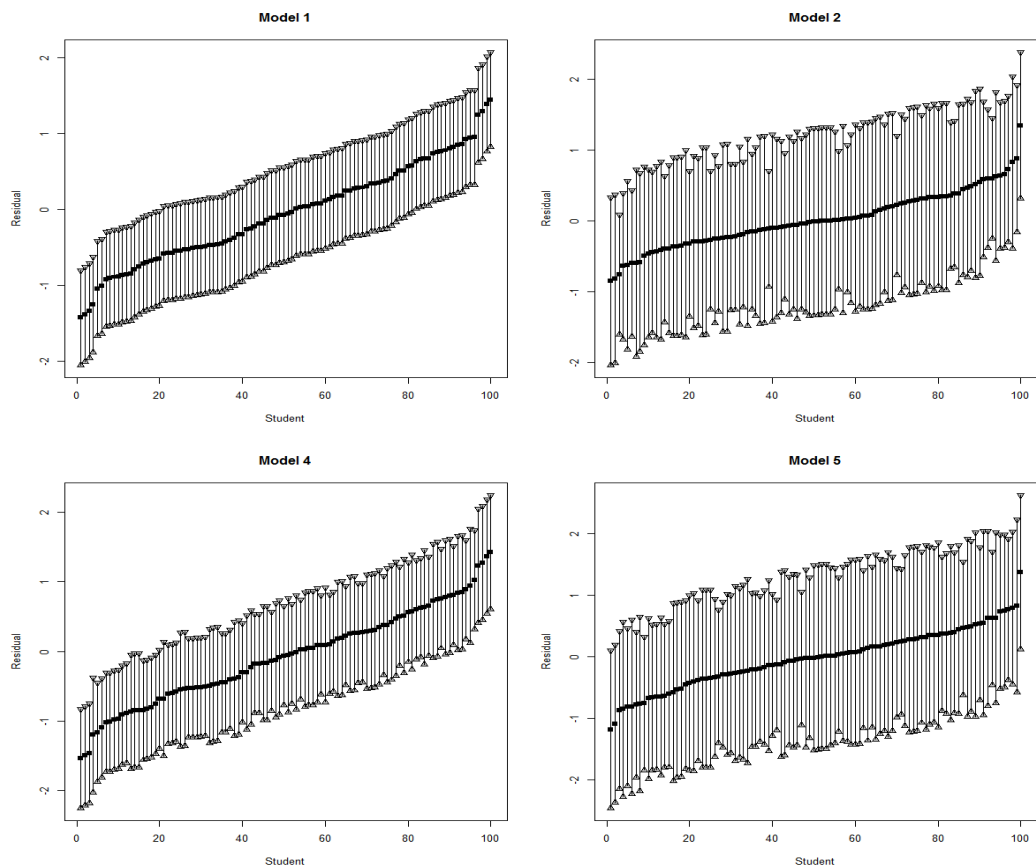


Figure 6.5: Caterpillar Plots of Residuals

## 6.2.2 Variance Estimation

By calculating and comparing variance components at each level, the changes of variation are straightforward. The output of the variance components model only provides a baseline and estimates the contribution of random effects at each level. The model itself could not make any prediction. From Table 6.1, the grand mean of reading scores is 0.0873. Additionally, the variation of reading scores is 0.5302 between individuals and variation among each wave is 0.3753. While adding explanatory variables, the overall variance at the level 1 dose not change much. The focus is examining changes in the second level (student level).

When adding books predictor, the variance between students depends on number of books. The variance of level 2 reduces by 0.01, which means that books factor is individual-varying. The variance in student-level rises from 0.5 to 0.6 when including phone predictor. So as to the model 4, the phone ownership varies between individuals. In the final model, the variance of 0.65 at level 2 is gained on the baseline model with both the two predictors. While the standard error of 0.12 is the highest compared with other models. The overall variance produced by each predictor is shown by the variance plots. The plots include the variance changes in model 2 and Model 4. It is to further test that a constant coefficient of variation is whether partitioned between values of predictor.

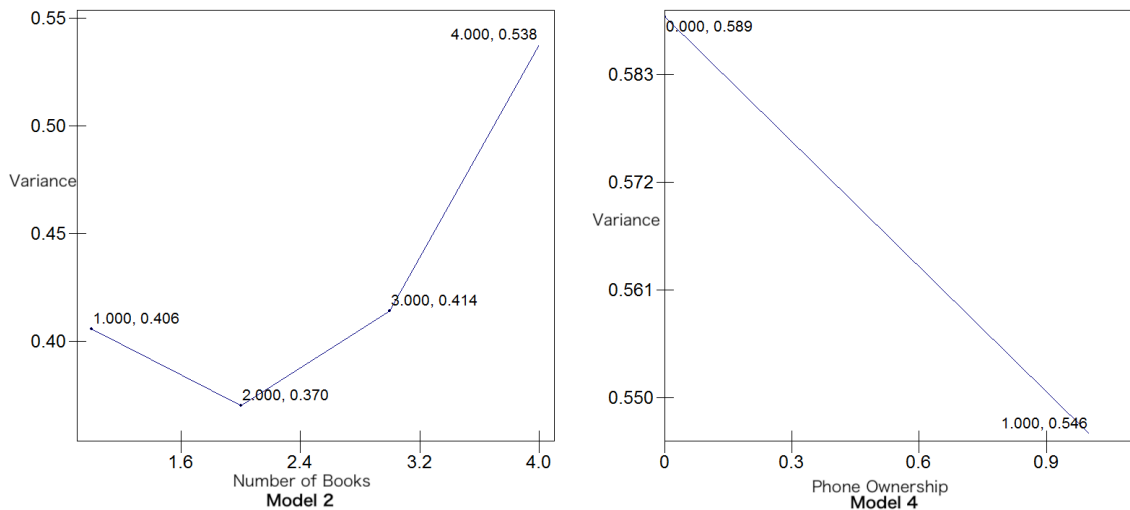


Figure 6.6: Variance against Predictors in Model 2 and 4

In model 2, variance at level 2 (student level) of reading scores first dropped from 0.406 to 0.37. Then it raised back to 0.414 and finally reached to 0.538. As more books in home, variance first decreased when the value was 2 and then increased at the value of 4. Overall speaking, the variance between students increases with adding books.

The right-hand side plot in Figure 6.6 shows the difference of variance at level 2 when including phone ownership predictor. The variance is 0.589 when the value of phone ownership is 0. And it declined to 0.546 at the value of 1. The overall variance decreases with owing mobile phones.

### 6.2.3 Examining Variation of Coefficients

There are only two random slope coefficients in the previous four models. The estimate of  $\beta_1$  is the coefficient of predictors. In model 2,  $\beta_1$  is close to the fixed coefficient (0.198) of the model with a single slope. Anyhow, the estimated standard deviation of the slope is 0.1992 by taking the square root of the variation 0.0397. Students have the mean growth rate of 0.2162 with a standard deviation of 0.199. From the z test, the growth rate has 95% coverage interval of 0.2 to 0.24. In model 4, 95% of the estimation of phone slopes is normally distributed between -0.19 and -0.12 from the z test. The slopes differ from the mean value of -0.17 with a variance estimated as 0.0856 (standard error 0.29). Similarly,  $\beta_1$  of model 4 is near to -0.175 obtained from the single slope model.

The intercepts in the fixed part diverse between models. In Model 2, the estimated mean value is 0.6816 (SE 0.0348) and their variance is 0.5208. Furthermore, there is a negative covariance between the intercept and slope coefficient. Since the estimated covariance is -0.0774 (SE 0.0269). It is interpreted as individuals with higher intercepts prone to have milder slopes. The negative correlation is -0.54 calculated by the following equation

$$Correlation = \frac{\sigma_{u_0}^2}{\sqrt{\sigma_{u_1}^2 \times \sigma_{u_0}^2}}$$

In Model 4, the intercept estimation is 0.2354 and the standard error is 0.0174. This negative correlation between slopes and intercepts is  $-0.0642/\sqrt{0.08560.5996} = -0.28$ . The negative covariance will possibly encompass steeper slopes when lines have lower intercepts. The detailed pattern could be further explained in the prediction plots.

The covariance matrix of the two models are described as below:

$$\Omega_u(\text{Model2}) = \begin{bmatrix} 0.5208 & -0.0774 \\ -0.0774 & 0.0397 \end{bmatrix} \quad \Omega_u(\text{Model4}) = \begin{bmatrix} 0.5996 & -0.0642 \\ -0.0642 & 0.0856 \end{bmatrix}$$

$$\Omega_u(\text{Model5}) = \begin{bmatrix} 0.6537 & -0.0762 & -0.1356 \\ -0.0762 & 0.034 & 0.0187 \\ -0.1356 & 0.0187 & 0.0825 \end{bmatrix}$$

In the final model (Model 5), the covariance matrix is  $\Omega_u(\text{Model5})$  listed above. The scatter plots reveal the relationship of the intercepts and slopes.

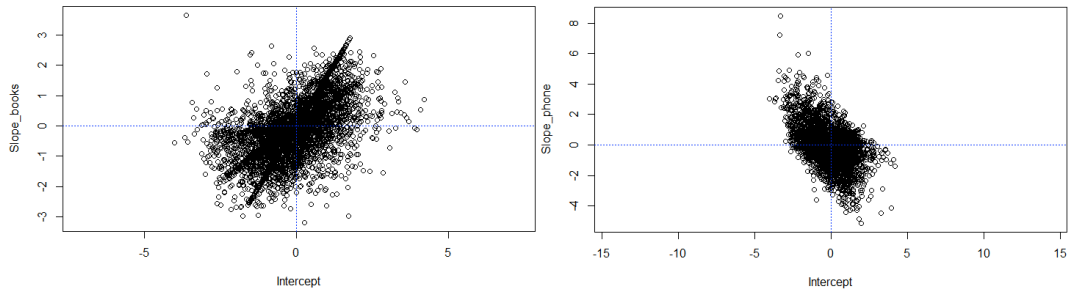


Figure 6.7: Intercepts VS Slopes in Model 5

In the above Cartesian coordinate, x-axis is the intercept while y-axis is the slope. Intercepts in level 2 are positively correlated with the slopes of books predictor while are negatively correlated with the slopes of the phone predictor. The slopes coefficient of books predictor are steeper when the intercepts are greater. It is completely opposite in the slopes of phone ownership, smaller intercepts lead to the steeper slopes.

## 6.3 Model Evaluation

Model evaluation aims at judging and comparing complex models. Degrees of freedom limits by the number of parameters, which is estimated to be varying in the final calculation. The basis of model evaluation is the Log-likelihood statistic. Other metrics, for example, deviance, AIC and BIC, are generated based on the Log-likelihood to compare the performance of models.

Table 6.2: Model Evaluation

	<b>DF</b>	<b>LogLikelihood</b>	<b>Deviance</b>	<b>AIC</b>	<b>BIC</b>	$ \mathbf{R}_1^2 $	$ \mathbf{R}_2^2 $
Model 1	3	-22285.3	44570.6	44576.6	44600.1		
Model 2	6	-22060.9	44121.9	44133.9	44180.8	0.0159	0.0178
Model 3	6	-22278.8	44557.6	44577.3	44608.6	0.0009	0.002
Model 4	6	-22207.2	44414.4	44426.4	44473.4	0.0459	0.1109
Model 5	10	-21995.8	43991.5	44011.5	44089.8	0.1151	0.2329

DF: Degrees of Freedom

The baseline log-likelihood value is -22285.3. Four elaborate models increase the log-likelihood more or less. It is since that when MLE function maximises the likelihood values, the value grows with increasing sample size. The higher value of log-likelihood or lower value of deviance dose not represent a better model fit in the scenario of the model overfit.

On the other hand, AIC and BIC evaluation penalise the likelihood function by the number of parameters. These two evaluations are calculated by equation 3.9 and 3.10. Lower values uncover a better model fit. BIC is higher than AIC normally because BIC also considers the sample size. Evaluation scores of models are smaller than the baseline model except for model 3. AIC and BIC of model 3 increases by 0.7 and 8.5 independently. It is included that the AIC remains the same and BIC rises when adding the explanatory variable of household type.

Due to the negative nature of pseudo  $R^2$ , it is considered that taking absolute values after the calculation of equation 3.12. The model 3 has the lowest  $R^2$  at both levels. It can be inferred that model 3 performs badly in predicting the response variable. In model 2, the performance of model fit is similar at level 1 and 2. However, the phone ownership predictor leads to a higher  $R_2^2$  notably.

The most significant improvement appears in model 5. The deviance declines by 579. AIC and BIC are 565 and 510 separately, which are smaller than the baseline model outstandingly. It has the highest  $R^2$  value, which is 0.1 roughly greater than other single predictor models.

As for the estimate process, model 3 takes less time since there is no additional predictor variance. In Table 6.3, the elapsed time triples in the final model compared with model 3. Model 2 converges after 6 iterations. And it takes 15.63 seconds, which

Table 6.3: Elapsed Time and Converge Iteration

	<b>Model 1</b>	<b>Model 2</b>	<b>Model 3</b>	<b>Model 4</b>	<b>Model 5</b>
Time	13.65s	15.63s	8.89s	13.63s	22.53s
Iteration	3	6	4	5	8

is the longest time and highest iterations excluding the final model. For the reason that there is a high variation in the values of books predictor.

Undoubtedly, improving the performance need to cost more time and computation. The more spent time could be ignored with such improvement of model 5. The best model of Model 5 converges after 8 iterations and used 22.53 seconds.

## 6.4 Interpretation of Model Output

The estimate the growth rate of books is 0.2115 in model 5, which indicates that reading scores improves when more books children can have access to in home. The fixed parameter for phone is -0.1544, specifying that the response variable reduces when subjects owe a phone. Comparing the two slope parameters, books have a more positive influence in improving children reading ability than the negative effect of phones. Since the estimated growth rate is 0.01 in model 4. Household type merely affects the reading achievement.

In order to study the impacts of each value in the predictor, binary or categorical variables corresponding to predictors are added in the variance components model. The predicted mean values of reading scores are generated from fixed slopes models and recorded in Table 6.4.

The average reading score of the students is 0.0873 shown in Table 6.1. After analysing the influence of each variable, it can be seen from Table 6.4 that subjects have no access to books have the lowest reading score, only -0.433. The average score of students without mobile phones was the highest, reaching 0.241. When subdividing each predictor, the first one is the respondent ownership of books. The results show that the student's reading score is positively correlated with the number of books in home. The improvement of reading scores is approximately 0.2 between each category in the books predictor. As mentioned in chapter 4, the majority of students have more



Table 6.4: The Impacts of Predictors

Predictor	Value	Label	Reading Scores
Number of Books	1	0 books	-0.433
	2	< 10 books	-0.221
	3	10 – 30 books	-0.012
	4	> 30 books	0.179
Household Type	1	Single (> 3 child)	-0.023
	2	Single (1 – 2 child)	-0.058
	3	Couple (> 3 child)	0.153
	4	Couple (1 – 2 child)	0.067
Phone Ownership	0	No	0.241
	1	Yes	0.065

than 30 books in home. They obtain a higher reading score generally, reading score is nearly 0.18. The second predictor is the family household type. It is surprisingly found that more children from a family may lead to a better reading score. Moreover, students from a single family probably have a poorer reading achievement compared with those belonging to couple household type. The children with the best reading scores are from the family type - parents with more than 3 children. Subjects in a single family with 1 to 2 child do -0.058, and they are 0.211 points lower than subjects in a couple family with more than 3 children. Finally, having mobile phones produces a negative effect on the reading achievement in subjects. Children or young people without phones do 0.241 of a standard deviation 0.13, which is remarkably 0.176 points better than those who have phones.

## 6.5 Conclusion

In conclusion, the model includes the random effects of phone and books has a desirable model fit with the data sample. In section 6.1 of model diagnostics, the following models' assumptions are valid: (1) The data structure is hierarchical between individuals among waves. (2) Residuals at both level 1 and level 2 follow a normal distribution. (3) The estimation of parameters is acceptable. In the final model, linear growth lines who have higher intercepts tend to have steeper slopes of books coefficient and flatter slopes of phone ownership. Therefore, students rapidly grow their reading skills when more books in home. On the contrary, the growth rate declines when children own mobile phones.

The model output proves that the household type is not a good predictor, in contrast to the result in the study of Dempsey [5]. Because the household type predictor only has a slight effect on the reading attainment. Whereas books and phone ownership have a relatively strong effect. The predictor of phone improves the model fit better than the predictor of books. Phone ownership is a negative effect on the response variable. It is shown that reading scores drop 0.18 points with owning mobile phones for most people. While there is a positive relationship between books and reading ability, about 0.2 points of improvement when approximately ten more books in home. Students who have fewer siblings probably have a poorer reading achievement. Besides, children from a couple family may have a better reading achievement.

# Chapter 7

## Future Work

Many potential adaptations and improvements in many aspects could be realised for the future. There can be a deeper analysis of the modelling mechanism, data sampling and estimating process.

It may exist a higher level of a data hierarchy. For instance, the GUI data set has a variation of reading scores caused by other variables. A three-level mixed-effects model could be built. In the case, the other variable is added at level 3. Wave is still the level 1 and level 2 remains within the individual. A more complicated model could have a better estimate and performance with the data sample. Due to the limitation of data and lack of time, the two-level hierarchical model is built. The estimation method only includes a deterministic procedure of IGLS. The estimation and iterations are the same after every run. The certainty may lead to biased estimates. More advanced stochastic estimation could be applied, such as simulation method of Markov Chain Monte Carlo (MCMC). This kind of method has an advantage of more accurate and less biased estimation.

More evaluation metrics can be taken into consideration, such as the Comparative Fit Index (CFI) proposed by Bentler in 1990 [27]. R Squared described in the section 3.3.3 compares models' performances based on residuals variance. Other R scores can be used to check the fitness functions.

There is still a limitation in the successive child cohort of 9-year-old, 13-year-old and 17-year-old students. Collecting information at the three time points likely cause a big time gap. More compact periods can produce a better result when studying the

changes. If more waves are covered in the child cohort, a new proposal of modelling mechanism is the latent growth curve model. And the assumption is not limited to linear growth. The trend could be polynomial. Besides, there is a strong inconsistency between each wave. Adjusting and reforming these variables is time-consuming. Moreover, only few complete factors are gathered through all three waves. Three predictors are cleaned, transformed and fed to the models. Other interesting factors can be preprocessed and included when time permits. For example, household income is an intended predictor. However, there are a lot of missing observations and abandoning it is the best choice. The research can be expanded by adding other achievements, ranging from math scores to IQ scores. Instead of investigating reading achievement, a more comprehensive assessment may be generated.

# Bibliography

- [1] J. Williams, S. Greene, E. Doyle, E. Harris, R. Layte, S. McCoy, C. McCrory, A. Murray, E. Nixon, T. O’Dowd, *et al.*, “Growing up in ireland national longitudinal study of children. the lives of 9 year olds,” 2009.
- [2] N. E. Hill and L. C. Taylor, “Parental school involvement and children’s academic achievement: Pragmatics and issues,” *Current directions in psychological science*, vol. 13, no. 4, pp. 161–164, 2004.
- [3] L. C. Fernald, A. Weber, E. Galasso, and L. Ratsifandrihamanana, “Socioeconomic gradients and child development in a very low income population: evidence from madagascar,” *Developmental science*, vol. 14, no. 4, pp. 832–847, 2011.
- [4] B. A. Levy, Z. Gong, S. Hessels, M. A. Evans, and D. Jared, “Understanding print: Early reading development and the contributions of home literacy experiences,” *Journal of Experimental Child Psychology*, vol. 93, no. 1, pp. 63–93, 2006.
- [5] S. Dempsey, S. Lyons, and S. McCoy, “Later is better: mobile phone ownership and child academic development, evidence from a longitudinal study,” *Economics of Innovation and New Technology*, pp. 1–18, 2018.
- [6] S. McCoy, A. Quail, and E. Smyth, “Growing up in ireland—influences on 9-year-olds learning: Home, school and community,” *Dublin: The Stationery Office*, 2012.
- [7] A. Cuevas, M. Febrero, and R. Fraiman, “An anova test for functional data,” *Computational statistics & data analysis*, vol. 47, no. 1, pp. 111–122, 2004.
- [8] B. O. Muthén, “Latent variable modeling in heterogeneous populations,” *Psychometrika*, vol. 54, no. 4, pp. 557–585, 1989.

- [9] R. Stephen and B. Anthony, *Hierarchical linear models*. Sage Publications, Thousand Oaks, CA, 2002.
- [10] J. Bruin, “newtest: command to compute new test @ONLINE.” <https://stats.idre.ucla.edu/stata/ado/analysis/>, Feb. 2011.
- [11] G. Verbeke and E. Lesaffre, “A linear mixed-effects model with heterogeneity in the random-effects population,” *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 217–221, 1996.
- [12] P. D. Allison, *Fixed effects regression models*, vol. 160. SAGE publications, 2009.
- [13] Wikipedia contributors, “Fixed effects model — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Fixed\\_effects\\_model&oldid=906386938](https://en.wikipedia.org/w/index.php?title=Fixed_effects_model&oldid=906386938), 2019. [Online; accessed 18-July-2019].
- [14] Wikipedia contributors, “Random effects model — Wikipedia, the free encyclopedia.” [https://en.wikipedia.org/w/index.php?title=Random\\_effects\\_model&oldid=902542656](https://en.wikipedia.org/w/index.php?title=Random_effects_model&oldid=902542656), 2019. [Online; accessed 18-July-2019].
- [15] H. Goldstein, “Multilevel mixed linear model analysis using iterative generalized least squares,” *Biometrika*, vol. 73, no. 1, pp. 43–56, 1986.
- [16] H. Bozdogan, “Model selection and akaike’s information criterion (aic): The general theory and its analytical extensions,” *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.
- [17] W. Luo and R. Azen, “Determining predictor importance in hierarchical linear models using dominance analysis,” *Journal of Educational and Behavioral Statistics*, vol. 38, no. 1, pp. 3–31, 2013.
- [18] S. Nakagawa and H. Schielzeth, “A general and simple method for obtaining  $r^2$  from generalized linear mixed-effects models,” *Methods in ecology and evolution*, vol. 4, no. 2, pp. 133–142, 2013.
- [19] The Economic and Social Research Institute (ESRI), “Growing up in ireland child cohort wave 1 - 9 years.” <http://www.ucd.ie/issda/data/GUChild/>

- GUChildWave1, 2008. [dataset]. Version 1. Irish Social Science Data Archive. SN: 0020-01.
- [20] The Economic and Social Research Institute (ESRI), “Growing up in ireland child cohort wave 2 - 13 years.” <http://www.ucd.ie/issda/data/GUChild/GUChildWave2>, 2012. [dataset]. Version 1. Irish Social Science Data Archive. SN: 0020-02.
- [21] The Economic and Social Research Institute (ESRI), “Growing up in ireland child cohort wave 3 - 17/18 years.” <http://www.ucd.ie/issda/data/GUChild/GUChildWave3>, 2016. [dataset]. Version 2. Irish Social Science Data Archive. SN: 0020-03.
- [22] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?,” *International journal of methods in psychiatric research*, vol. 20, no. 1, pp. 40–49, 2011.
- [23] J. Rasbash, W. Browne, H. Goldstein, M. Yang, I. Plewis, M. Healy, G. Woodhouse, D. Draper, I. Langford, and T. Lewis, “A users guide to mlwin,” *London: Institute of Education*, vol. 286, 2000.
- [24] R. J. B. W. H. M. Charlton, C. and B. Cameron, “Mlwin version 3.04,” *Centre for multilevel modelling, University of Bristol*, 2019.
- [25] Z. Zhang, R. M. Parker, C. M. Charlton, G. Leckie, W. J. Browne, *et al.*, “R2mlwin: A package to run mlwin from within r,” *Journal of Statistical Software*, vol. 72, no. 10, pp. 1–43, 2016.
- [26] H. Goldstein, *Multilevel statistical models*, vol. 922. John Wiley & Sons, 2011.
- [27] E. Ryu and S. G. West, “Level-specific evaluation of model fit in multilevel structural equation modeling,” *Structural Equation Modeling*, vol. 16, no. 4, pp. 583–601, 2009.

# Appendix

## Abbreviations

**AIC** Akaike Information Criterion

**ANOVA** Analysis of Variance

**BIC** Bayesian Information Criterion

**CI** Confidence Interval

**DF** Degrees of Freedom

**GUI** Growing Up in Ireland

**HLM** Hierarchical Linear Modelling

**IGLS** Iterative Generalised Least Squares

**LRT** Likelihood Ratio Test

**MCMC** Markov chain Monte Carlo

**MLE** Maximum Likelihood Estimation

**OLS** Ordinary Least Squares

**PCC** Pearson Correlation Coefficient

**SE** Standard Error

**VPC** Variance Partition Coefficient