UNIVERSITY OF
KWAZULU-NATAL

INYUVESI
YAKWAZULU-NATALI

UNIVERSITY OF KWAZULU-NATAL (PMB)

DISSERTATION

STATISTICS

A REVIEW OF METHODS FOR MODELLING BOTH GAUSSIAN AND NON-GAUSSIAN LONGITUDINAL DATA WITH APPLICATION

SIPHAMANDLA SIBIYA

A technical report submitted
in partial fulfillment of the requirements for the degree of

Master of Science
in
Statistics

September 2015

Declaration

The research work is the original done by the author Siphamandla Sibiya (208500476) and it is not a duplicate source of the research work done by authors. All the references that were used to refer to are duly acknowledged.

Author: Mr. Siphamandla Sibiya   Date

Supervisor: Dr. S. Ramroop   Date

Co-Supervisor: Prof. H.G. Mwambi   Date

School of Mathematics, Statistics and Computer Science

Department of Statistics and Biometry

University of KwaZulu-Natal

Pietermaritzburg

Dedication

To My Family and Friends

Acknowledgement

The endless thanks goes to God Almighty for all the blessings He has showered onto me, which has enabled me to write this last note in my research work. During the period of my research, as in the rest of my life, I have been blessed by Almighty with some extraordinary people who have spun a web of support around me. Words can never be enough in expressing how grateful I am to those incredible people in my life who made this thesis possible. I would like an attempt to thank them for making my time during my research in the Institute a period I will treasure.

First and foremost I want to thank my supervisor Dr. Shaun Ramroop and my co-supervisor Prof. Henry Mwambi. It has been an honor to be one of their MSc students. They have taught me, both consciously and unconsciously, how good experimental statistics is done. I appreciate all their contributions of time, ideas, and funding to make my MSc experience productive and stimulating. The joy and enthusiasm they have for the research was contagious and motivational for me, even during tough times in the MSc pursuit.

Lastly, I would like to thank my parents, Mr. Jimson Sibiya and Ms. Linah, A Mabika, my sisters, brother, and my two beautiful nieces Mabee and Putsu, and relatives for their unlimited love, encouragement and support throughout my life. For my parents who raised me with a love of science and supported me in all my pursuits. For the presence of my brother Sandile Soh Gumede here at UKZN for two of my years here. And most of all for my loving, supportive, encouraging, and patient fiance Peggy Dlamini whose faithful support during the final stages of this MSc. is so appreciated. Thank you.

Abstract

The study of longitudinal data plays an integral role in medicine, epidemiology, social science, biomedical and health sciences research where repeated measurements are obtained over time for each individual. Generally, the interest is in the dependence of the outcome variable on the covariates. The analysis of the data from longitudinal studies requires special techniques, which take into account the fact that the repeated measurements within one individual are correlated. In review of this work, we explore modern developments in the area of linear and nonlinear generalized mixed-effects regression models and various alternatives including generalized estimating equations for analysis of longitudinal data and correspondence analysis (CA). Methods are described for continuous and normally distributed as well as categorical variables. We apply this theory to the analysis of complete longitudinal data from National Institute of Environment Health Sciences (NIEHS) focusing on the relationship between blood lead levels (P_bB) and some associated covariates. The results show that Placebo-treated children had a gradual (occuring) decrease in blood lead level. Succimer-treated children had an abrupt (unexpected) drop in blood lead level, followed by rebound. The average mean blood lead level of the succimer-treated children after initiation of treatment was 19.14 μg/dL lower than that of placebo-treated children. After randomization, blood lead levels had fallen by similar amounts in both chelated and placebo children, despite the immediate drops in the chelated group; there was no association between change in blood lead level and change in cognitive test score. Blood lead levels continued to fall.

Keywords: Longitudinal data, Linear mixed models, generalized linear models, random effects, correlated data, correspondence analysis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

A longitudinal study refers to an investigation where participant outcomes and possible treatments or exposures are collected at multiple follow-up times. This is also often referred to as repeated measurements data (Davis, 2002; Fitzmaurice et al., 2004). Longitudinal data analysis is defined as the analysis of the data resulting from the observations of subjects which are repeatedly measured over a series of time-points (Verbeke and Molenberghs, 2000; Diggle et al., 2002). The purpose of conducting longitudinal study is to study the change of a response over time.

Analyzing longitudinal data involves characterizing the time trends within subjects and variability or heterogeneity between subjects in the data. The data will always comprise the response, the time covariate among other covariates and the indicator of the subject on which the measurement has been made. If other covariates are recorded, for example whether the subject is in a treatment group or the control group, we may wish to relate the within- and between-subject trends to such covariates (Verbeke and Molenberghs, 2000; Diggle et al., 2002). Between- subject effects are those whose values change only from subject to subject and remain the same for all observations on a single subject, for example the treatment (i.e drug or placebo) and gender effects (i.e male or female). Within-subject effects are those whose values may differ from measurement to measurement, for example the level of CD4 count from one time period to the next one in a clinical study (Hedeker and Gibbons, 2006). According to West et al. (2007), a random factor is a variable with levels that can be thought of as being randomly sampled from a population of levels being studied and a fixed factor is a variable for which the investigator has included all levels that are of interest in the study. Random effects are random values associated with the levels of a random factor whilst fixed effects describe the relationships between the dependent variable and predictor variables (i.e., fixed factors or continuous covariates) for an entire population of units of analysis (West et al., 2007).

The study of random effect models with normal or non-normal distribution for the random effects has application in many areas of medical science or clinical studies (Verbeke and Molenberghs, 2000; West et al., 2007). Repeated measurements are very frequent in almost all scientific fields where statistical models are used. We give a few examples: in agriculture - crop yields in different fields over different years; in biology - growth curves i.e sequential measurements of size taken at regular intervals on the same height, for example of longitudinal data found in Potthoff and Roy (1964); in education - student progress under various learning conditions; in insurance - evolving relationships between premiums and claims for different firms; and in medicine - successive periods of illness and recovery under different treatment regimes.

According to Diggle et al. (2002), the main advantage of longitudinal studies over cross sectional studies is that longitudinal studies can separate aging from cohort effects Hedeker and Gibbons (2006) give an example where it can separate aging effects (i.e. changes over time within individuals) from cohort effects (i.e. differences between subjects at baseline). Longitudinal studies can distinguish changes over time within individuals from differences among individuals in their baseline levels or covariates while the cross sectional studies cannot do so. The changes within individuals over time is known as age effects while the difference among people in their baseline covariates is known as cohort. In the cross sectional data only a single response is available for each experimental unit. We now consider some advantages and challenges associated with longitudinal studies.

1.1.1 Types of Longitudinal Studies

Firstly, panel studies allow the researcher to find out why changes in the population are occurring, since they use the same sample of people every time, for example with a household panel survey, when individuals are followed and observed within their household and information is collected. Secondly, time series analysis, where a single variable is measured at different time points, for example quarterly for several years. Thirdly, repeated cross sections, which is the most common type of study in longitudinal survey studies. This involves whole survey with the same variable measured repeatedly at different time points and cohort data sets, where individuals are followed over time and a certain event of interest occurs or until the end of study. Finally, an event history data sets, which is also known as survival data analysis.

1.1.2 Advantages of Longitudinal Studies

The advantages are taken from the ideas of Hedeker and Gibbons (2006). We give some advantages of longitudinal study analysis:

1. The study provides information about individual change.
2. With same number and pattern of individuals in the study, more efficient estimators than cross-sectional design can be obtained.
3. The study economizes on subjects thus it reduces on costs
4. The between-subject variation is excluded from error.
5. The study can separate changes over time within individuals from difference between subjects at baseline or separates aging effects from cohort effects.
6. The subjects can serve as their own control in that the outcome variable can be measured under both control and experimental conditions for each subject.

1.1.3 Challenges of Longitudinal Studies

A longitudinal study just like any study, has challenges. According to Hedeker and Gibbons (2006), there are several challenges under longitudinal study analysis. These include:

1. Time-varying covariates: here the designs offer the opportunity to associate changes in exposure with changes in the response of interest however the direction of causality can be complicated by feedback between the outcome and the exposure. Examples of such covariates are age, smoking status, cumulative exposure to some risk factor and treatment in certain studies.
2. Participant follow-up: there is a risk of bias due to incomplete follow-up or drop-out of the study participants. Sometimes this is called attrition.
3. Analysis of correlated data: statistical analysis requires methods that can properly account for the intra-subject correlation of response measurements. The inferences such as statistical tests or confidence intervals can be detrimental or invalid if the correlation is ignored.

According to Davis (2002), there are two main difficulties in the analysis of data from repeated measures studies. First, the analysis is complicated by the dependence among repeated observations made on the same experimental unit. Second, the investigator often cannot control the circumstances for obtaining measurements, so that the data may be unbalanced or partially incomplete. For example, in a longitudinal study, the response from a subject may be missing at one or more of the time points due to factors that are related or unrelated to the outcome of interest. In our study of regression type models for longitudinal data, we focused on situations where the response was continuous and reasonably assumed to be normally distributed, and also the model relating mean response to time and possibly other covariates is linear in parameters that characterize the relationship. This means in one way or the other the analysis of longitudinal data strategies deal with missing data if there are any.

1.2 Analysis for Longitudinal Data

Currently researchers agree (Hedeker and Gibbons, 2006; Verbeke and Molenberghs, 2000; Diggle et al., 2002; Fitzmaurice et al., 2004) that there are several different features of longitudinal studies that must be considered when selecting an appropriate longitudinal analysis.

Response Variable

The response variable might be continuous and assumed to be normal or non-normal distributed and might be categorical where the response variable might be ordinal, dichotomous, nominal or counts. The mixed-effects linear model is popularly used for continuous and normally distributed response variables. But if the response is discrete and therefore does not have a normal distribution (e.g., a count), then one can consider a mixed-effects model like Poisson regression model. Mixed effects models can also be considered for the categorical responses, such as binary (yes or no), ordinal (e.g., sad, neutral, happy), or nominal (republican, democrat, independent). Non-linear mixed effect models can also be used for non-linear changes over time such as in the case of growth data, pharmarkokinetics, and disease dynamics models.

Number of subjects N

In the analysis based on a large sample of unbalanced longitudinal data, we consider the models like generalized mixed-effects regression models which is suitable for analysis and may not be suitable for small number of subjects N (e.g., $N < 50$).

Number of observations per subjects n_i

i $n_i = 2$ for all subjects

The data can be analyzed using changing score analysis or ANCOVA in the methods for cross-sectional data.

ii $n_i = n$ for all subjects

If the data is assumed to be balanced, then ANOVA or MANOVA for repeated measures or mixed effects models can be used in the analysis.

iii n_i varies

In the case where n_i varies from subject to subject, then methods like generalized mixed-effects regression models are required in the analysis. The reason for GLMMs is not variable n_i. The model can be used both if $n_i = n$ and variable n_i.

Number and Types of Covariates

i Single sample

In this case, the interest is in characterizing the rate of change in the population over time. Here, we can use a random-effects regression model in the analysis, since the parameters are treated as random effects and allowed to vary from subject to subject.

ii Multiple samples

The model consists of one or more categorical covariates that contrast the various treatment conditions in the design. Here, categorical covariates can also be included in a model characterizing the rate of change, e.g. gender, to see if the rate of change differs between males and females.

iii Regression

The analysis may have both continuous and categorical covariates, such as age, sex, and race.

iv Time-varying covariates

When the covariates take on time-specific values i.e., time-varying covariates, then longitudinal data methods are best suited to handle such covariates.

Type of Variance-Covariance Structure -$V(y_i)$

In this case we are dealing with different model specifications that lead to homogeneous or heterogeneous variances and homogeneous or heterogeneous covariances of the repeated measurements over time. In modeling the variance-covariance structure of the data, residual autocorrelation among the responses plays an important role in the analysis. The statistical model will be used for handling the analysis.

1.2.1 Methodology for Analyzing Longitudinal Data

In this section we review the methodology for the analysis of repeated measurements. According to the current literature including Verbeke and Molenberghs, (2000); Diggle et al., (2002); West et al., (2007) and others, the general approaches for analyzing longitudinal data include:

1. Analysis of variance (ANOVA) for repeated measures
2. Multivariate analysis of variance (MANOVA) for repeated measures
3. Mixed effects regression models
4. Covariance pattern models
5. Generalized Linear models (GLM)
   a) Generalized estimating equation (GEE)
   b) Generalized linear mixed model (GLMM)
   c) Hierarchical generalized linear model (HGLM)
6. Growth curve and Latent variable models
7. Multilevel models

Linear (and nonlinear) mixed effects models are useful for analyzing longitudinal data, providing a simple and effective way to incorporate within-subject and between-subject variation and the correlation structure of longitudinal data (Nguyen et al., 2008). The advantage of this methodology is that it accommodates the complexities of longitudinal data sets (Fitzmaurice et al., 2004). The inclusion of random effects in models for longitudinal data helps to capture the inherent correlation in the data. The use of linear mixed model methodology for the analysis of repeated measurements is becoming increasingly common due to the development of widely available software.

1.2.2 Statistical Software

The purpose of the current study is to review and summarize the random effect models for normal non-normal distributions for the analysis of longitudinal data and also understand the implementation of such models using the SAS software. The software used in the current study will be SAS (Version 9.3). Most of the analyses are carried out using PROC MIXED in SAS supplemented by PROC GENMOD, PROC LOGISTIC, PROC NLMIXED and PROC GLIMMIX in SAS for fitting non-normal data.

1.3 Objective of the Study

The primary object of the study is to review and apply advanced statistical methods to analyse correlated Gaussian and non-Gaussian longitudinal data using appropriate models for the outcomes or outcome variable. The specific objectives of the study were:

1. To review statistical methods used in the analysis of the longitudinal Gaussian non-Gaussian data.

2. To understand how to fit the model and interpret the parameter estimates, especially in terms of odds and odd ratios.

3. To understand the relative merits and demerits of the statistical software used in analyzing longitudinal data.

Chapter 2

Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) approach is used to summarize the main characteristics of the data in the form of graphs and tabulations without using a statistical model. For this data, analysis is done with a complete data set in SAS version 9.3.

2.1 Data Description

The current data was a published data set which was reported in treatment of lead-exposed children (TLC) by Rhoads (2000). In the current study we followed the subsample (N=100) of data on Blood Lead Levels from the Treatment of Lead-Exposed Children (TLC) Trial. The Treatment of Lead-Exposed Children (TLC) trial was a placebo-controlled, randomized study of succimer (a chelating agent) in children with blood lead levels of 20-44 micrograms/dL. These data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or placebo. The data is a clinical trial where the outcome of interest is continuous. The TLC clinical trial compared the effect of lead chelation with succimer to place therapy. The model of the data includes both categorical and continuous covariates.

2.2 Descriptive Statistics

The summary statistics for the number of children in the study at each time point in both groups is given in Table 2.1 and Table 2.2 where we look at measures of central tendency. In the statistical table the average lead level for children in the active treatment is 19.139 and that for children in the actual number s placebo group is 24.662. Another measure is the median which divides a data into half when the data set is ordered in descending or ascending order and is used to explain the skewness of the distribution of the data. In Table 2.1 and Table 2.2 the skewness statistic in both groups are positive implying right skewed. In both groups the kurtosis is not equal to zero and it is positive therefore the degree and direction of asymmetry of the distribution is non-normal and it is positively skewed.

There are several different measures of variation that we look at, such as the range, interquartile range, variance and standard deviation. In Table 2.1 and Table 2.2, the variance and standard deviation shows that in active group the observations are more spread out than in placebo since the standard deviation is higher than the one in placebo group and the spread of variables is much high in active than in the placebo group the range in active group is high. On each scale, the statistics show that there is a strong variation between group and treatment. The spread of the data is much higher in the active group since the interquartile range is greater than in placebo group. We see that the range of active group is 2 times greater as that of placebo group. Table 2.2 shows that the interquartile ranges for the active group is approximately 2 times as that of placebo group. This is based on the middle 50% of the data and ignores the extremes at either end of the data.

The result in the Table 2.3 indicates that there was statistical significant association between the location parameter. Based on the test statistic, the $p-value < 0.001$ for each test.

Figure 2.1 shows the estimated means and estimated individual trends. The plot seems to indicate individual to individual variability in both the intercept and slope which should be investigated at the modeling analysis stage.

Figure 2.2 shows a boxplot of lead levels across time for the active group that has a significant effect on blood level with respect to both location and variation. The plot shows a decreasing in the mean lead levels between baseline and week 1 then increasing starts again between weeks 1 and 6. Figure 2.3 shows a decrease in the mean lead levels between baseline and week 6.

Figure 2.4 show that the evolution of scatter plots with increasing blood lead levels from the diagonal suggests that the correlation remains more or less constant. There seem to be outliers in the data. In addition the histograms at each week of measurement show a fairly consistent non-normal distribution.

Table 2.1: Analysis of Blood lead levels measured grouped by Treatment

Treatment = P (Placebo)			
Moments			
N	200	Sum Weights	200
Mean	24.662	Sum Observations	4932.400
Std Deviation	5.526	Variance	30.539
Skewness	0.708	Kurtosis	0.357
Uncorrected SS	127720.220	Corrected SS	6077.371
Coeff Variation	22.407	Std Error Mean	0.390
Basic Statistical Measures			
Location		Variability	
Mean	24.662	Std Deviation	5.526
Median	23.900	Variance	30.539
Mode	21.100	Range	29.800
		Interquartile Range	7.250

Table 2.2: Analysis of Blood lead levels grouped by Treatment

Treatment = A (Active)			
Moments			
N	196	Sum Weights	196
Mean	19.139	Sum Observations	3751.300
Std Deviation	9.135	Variance	83.459
Skewness	0.635	Kurtosis	2.002
Uncorrected SS	88071.710	Corrected SS	16274.508
Coeff Variation	47.732	Std Error Mean	0.652
Basic Statistical Measures			
Location		Variability	
Mean	19.139	Std Deviation	9.135
Median	9.135	Variance	83.459
Mode	0.635	Range	61.100
		Interquartile Range	12.250

Table 2.3: Test for location

Test	Statistic		P-value	
Student's t	t	63.112	$Pr >= \|t\|$	$< .0001$
Sign	M	100	$Pr >= \|M\|$	$< .0001$
Signed Rank	S	10050	$Pr >= \|S\|$	$< .0001$

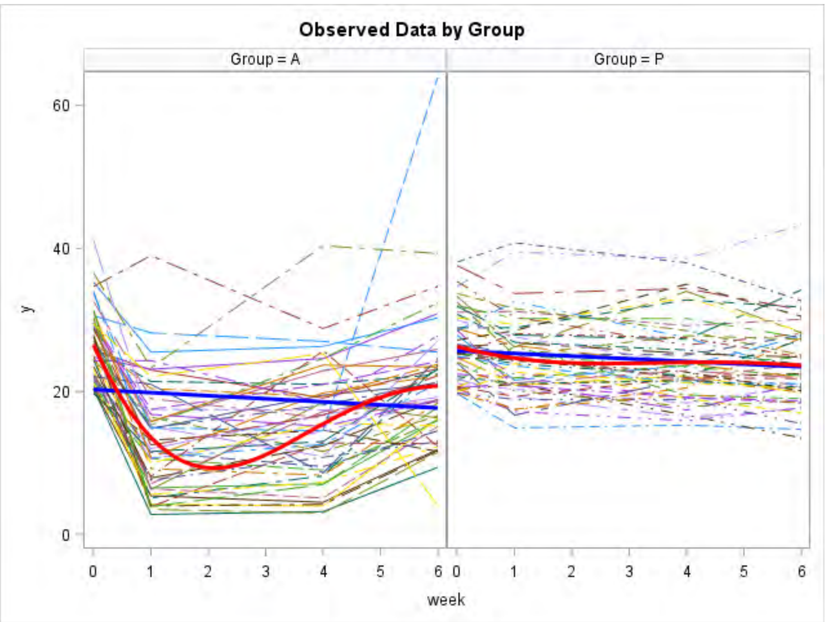

Figure 2.1: TCL data: Estimated means (solid line) and estimated individual trends (dotted line)

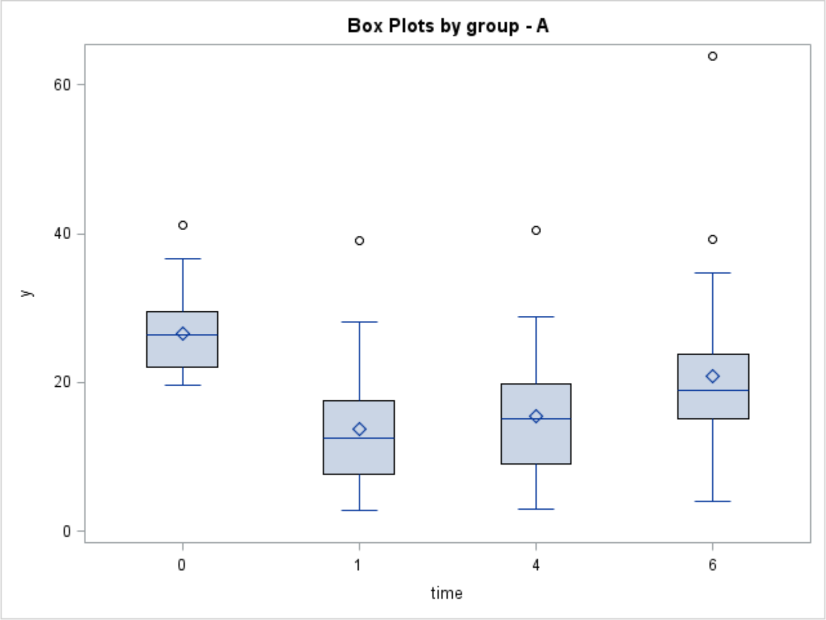

Figure 2.2: Box Plot for Blood Lead Level Measurements Across Weeks.

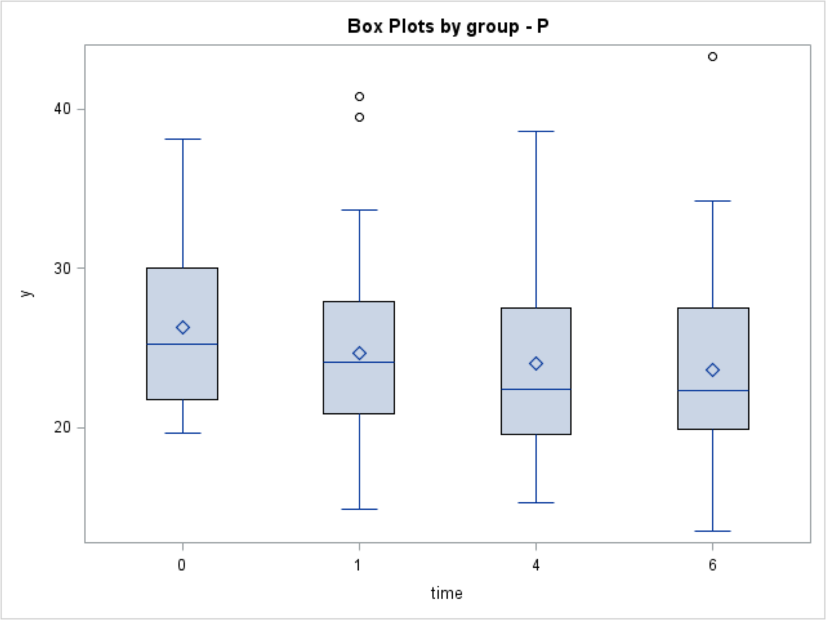

Figure 2.3: Box Plot for Blood Lead Level Measurements Across Weeks.

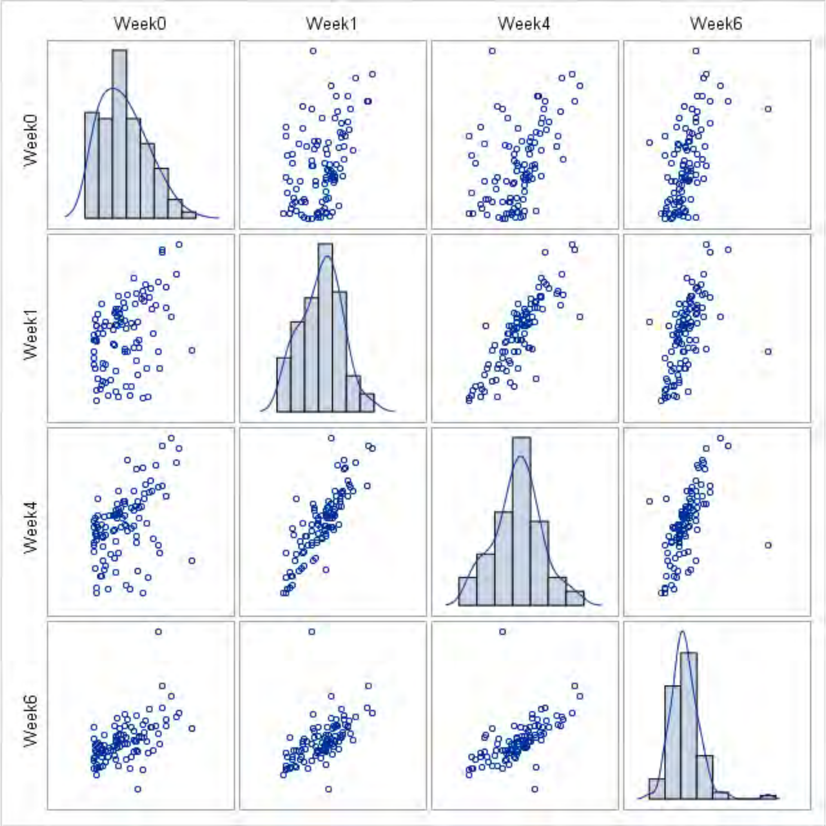

Figure 2.4: Blood Lead Level Level Measurements Across Weeks

2.3 Summary

In analysis of repeated observations we first carry out an exploratory analysis by creating tables and graphs to see if there is a change, and doing a simple linear regression analysis. Graphs of the average response over time are very helpful. And we can tell if the trends between two groups are the same, if there are between-subject and within-subject effects, if the change in the response is linear or not, and if the variance increases as the study progresses. From the graphs one can see that there is significant within and between subject effects, the variance is increasing, and there is a nonlinear positive change within the average response. The graphs also showed that the mean response is not the same within the week groups.

Chapter 3

Theory of Linear Mixed Effect Models (LMMs)

3.1 Introduction

In this chapter we present the theory behind linear mixed models. This is done because the model choice in our study involves repeated measures over time. Our main objective for this study is to find the relationship between blood lead level measurements across week and selected covariates. This chapter describes the statistical models used in the subsequent chapters. It also describes how the parameters were estimated. According to West et al. (2007), linear mixed model (LMM) is a parametric linear model for clustered, longitudinal, or repeated-measures data that provides the relationships between a continuous dependent variable and various predictor variables. The LMM includes both fixed-effect parameters associated with continuous or categorical covariates and random effects associated with random factors. The linear combination of fixed and random effects gives the linear mixed model. Fixed-effect parameters describe the relationships of the covariates to the dependent variable for an entire population and random effects describe the heterogeneity between subjects and clusters. Consequently, random effects are directly used in modeling the random variation in the dependent variable at different levels of the data. According to Howell (2010) mixed model plays an important role in statistical analysis and gives some advantages over traditional analyzes. We consider some advantages of mixed model delineated by Howell (2010) and Duchateu et al. (1998).

1. The mixed model does not assume the covariance structures in the model but it shows flexible specification of the covariance structure among repeated measure and it allows

the model to select it own set of covariance.

2. It also allows the prediction of random effects of interest by best linear unbiased prediction.

3. Linear mixed model can be extended to higher level models that implies repeated observations within individuals within cluster (Littell, 2006).

4. Mixed models can also be extended as Generalized mixed models to non-Normal outcomes.

3.2 Model Description

3.2.1 Linear Model

According to Davis (2002), we let $\mathbf{y} = (y_1, \ldots, y_n)'$ be an $n \times 1$ vector of independent observations. The linear model is given by

$$y = X\beta + \varepsilon \tag{3.1}$$

where β is a $p \times 1$ vector of unknown parameters, $\mathbf{X}$ is an $n \times p$ model matrix and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an $n \times 1$ vector of independent errors. The components of an error (ε) are assumed to be normally distributed with mean 0 and constant variance σ^2. This model can be extended to linear mixed model by including random effects which will be shown in the next section.

3.2.2 Linear Mixed Model

Linear mixed model procedures extend the general linear model so that the data is allowed to display correlation and non constant variability. Linear mixed effects models contain both fixed and random effects. They allow the analysis of between-subject and within-subject sources of variation in the longitudinal or clustered data (Fitzmaurice et al., 2004). The LMM for longitudinal data that was introduced by Harville (1977, 1976) and Laird and Ware (1982) can be written as

$$y_i = X_i\beta + Z_i u_i + \varepsilon_i \tag{3.2}$$

with $i = 1 \ldots N$ individuals. The notation Y_i is an $N \times 1$ response vector for the i^{th} individual, such that Y_{ij} denotes the $j^{th}(j = 1, 2, \ldots, n)$ observation made at time t_{ij} for the individual, X_i is the model matrix for the fixed effects for observations in individual i, Z_i is the model matrix for the random effects for observations in individual i, U_i is the random effect coefficient vector for individual i, β is the fixed effect coefficient, N is the number of subjects and ε_i is the error for observations in individual i (Laird and Ware, 1982). The random components of the model are vectors u and ε are assumed to be normally distributed with mean 0 and variance G and R_i respectively. The precise distributional assumptions about the random effects and errors are: $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})' \sim N(0, R_i)$, where R_i is the covariance matrix of error vector ε_i for observations in individual i and $U_i \sim N(0, G)$, where G denotes the covariance matrix of random effects U_i. It is also assumed that $U_1, \ldots, U_n, \varepsilon_1, \ldots, \varepsilon_n$ are independent or uncorrelated (Verbeke and Molenberghs, 2000). The elements in G and R_i are known as variance components and can be written as:

$$\begin{bmatrix} U_i \\ \varepsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R_i \end{bmatrix}\right) \tag{3.3}$$

According to Verbeke and Molenberghs (2000), if we assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independent then, it follows that $R_i = \sigma^2 I$ where I is the identity matrix. An important distinction in the linear mixed effects model is between the conditional and marginal means Y_i . The conditional or subject-specific mean of Y_i is given by

$$E(Y_i|u_i) = X_i\beta + Z_i u_i$$

and the conditional covariance of Y_i given u_i is

$$\begin{aligned} Var(Y_i|u_i) &= Var(\varepsilon_i) \\ &= R_i \end{aligned} \tag{3.4}$$

Thus

$$Y_i|u_i \sim N(X_i\beta + Z_i u_i, R_i)$$

while the marginal mean of Y_i when averaged over the distribution of random effects u_i is

$$\begin{aligned} E(Y_i) &= E(E(Y_i|u_i)) \\ &= E(X_i\beta + Z_i u_i) \\ &= X_i\beta + Z_i E(u_i) \\ &= X_i\beta \end{aligned} \tag{3.5}$$

and the marginal covariance of Y_i averaged over the distribution of u_i is

$$\begin{aligned} Var(y_i) = V_i &= E[Var(Y_i|u_i)] + Var[E(Y_i|u_i)] \\ &= E(R_i) + Var(X_i\beta + Z_i u_i) \\ &= R_i + Z_i G Z_i^{'} \\ &= Z_i G Z_i^{'} + R_i \end{aligned} \tag{3.6}$$

Thus

$$y_i \sim N(X_i\beta, Z_iGZ_i^{'} + R_i)$$

It can be shown that the observations y_i and random effects u_i have joint multivariate normal distribution

$$\begin{pmatrix} y_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} ZGZ^{'} + R_i & ZG \\ GZ^{'} & G \end{pmatrix}\right] \tag{3.7}$$

From these results, it can be shown that the mean of the posterior distribution of u_i given y_i yields the formula for the empirical Bayes estimates of the random effects. Now, the mean of the posterior distribution of u_i given y_i is

$$\begin{aligned} \hat{u}_i &= [Z_i^{'}(\sigma^2 I_{ni})^{-1}Z_i + G^{-1}]^{-1}Z_i^{'}(\sigma^2 I_{ni})^{-1}(y_i - X_i\beta) \\ &= (Z_i^{'}Z_i + \sigma^2\Sigma_u^{-1})^{-1}Z_i^{'}(y_i - X_i\beta) \end{aligned} \tag{3.8}$$

with the variance-covariance matrix as

$$\Sigma_{u|y_i} = [Z_i^{'}(\sigma^2 I_{ni})^{-1}Z_i + G^{-1}]^{-1} \tag{3.9}$$

It should be noted that intrinsically the marginal model allows negative variance components provided V_i is positive semi-definite while in the conditional model negative components does not make sense. The previous studies have shown that Gaussian theory estimation procedures for linear mixed models which consists of the maximum likelihood (ML) and the restricted maximum likelihood (REML) are some of the methods that can be used to deal with such challenges.

3.3 Estimation Methods

In this section we address the estimation problem for fixed effects, random effects and variance components in linear mixed models. According to Jiang (2007) the most often used methods of estimation in Gaussian mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML). In our case we discuss these two estimation methods including inference on population parameters.

3.3.1 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation was introduced by Fisher (1925). The aim of this method is to construct an estimator for an unknown parameters by maximizing the likelihood. This approach arises when we consider the estimation of β and α simultaneously by maximizing the joint likelihood distributed given by

$$L_{ML}(\theta, y) = -\frac{1}{2}\left\{\log|V| + (y - X\hat{\beta})^{-1}V^{-1}(y - X\hat{\beta})\right\} \tag{3.10}$$

the MLE σ^2 is equal

$$\hat{\sigma}^2_{ML} = \sum_{i=1}^{N} \frac{(y_i - X_i\hat{\beta})'(y_i - X_i\hat{\beta})}{N} \tag{3.11}$$

and it can be shown that $\hat{\sigma}^2_{ML}$ is a biased downward for σ^2 because

$$E(\hat{\sigma}^2_{ML} - \sigma^2) = -\frac{\sigma^2}{N}, \quad \text{where } N, \sigma^2 > 0$$

Thus $\hat{\sigma}^2_{ML}$ is an underestimate of σ^2.

3.3.2 Estimation of Fixed Effects Parameters

Let β denote the vector of fixed effects coefficients and α denote the vector of all variance components in G and R_i. It then follows that the variance covariance matrix V_i of Y_i is α dependent. Thus we can let $\theta = (\beta', \alpha')$ denote the vector of all parameters in the marginal model. According to Verbeke and Molenberghs (2000) the estimates for α_{ML} and β_{ML} can be obtained from maximizing $L_{ML}(\theta)$ with respect to θ that is with respect to α and β respectively. The marginal likelihood function is written as:

$$L_{ML}(\theta) = \prod_{i=1}^{n}\left\{(2\Pi)^{-\frac{n_i}{2}}|V_i(\alpha)|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y - X_i\beta)'V_i^{-1}(\alpha)(y - X_i\beta)\right)\right\} \tag{3.12}$$

given the above assumptions hold. Where $V_i(\alpha)$ is the matrix of variance components and if α is known, the MLE of β is given by

$$\hat{\beta}(\alpha) = \left(\sum X_i' W_i X_i\right)^{-1}\left(\sum X_i' W_i y_i\right) \tag{3.13}$$

where $W_i = V_i^{-1}(\alpha)$. According to Fitzmaurice et al. (2004), the estimator of β that minimizes this expression is known as generalized least squares (GLS) estimator of β denoted by $\hat{\beta}$. It can be shown that

$$E(\hat{\beta}) = \beta$$

provided $E(y_i)$ is correctly modeled and

$$\begin{aligned} Var(\hat{\beta}) &= \left(\sum X_i' W_i X_i\right)^{-1}\left(\sum X_i' W_i var(y_i) \sum W_i X_i\right)\left(\sum X_i' W_i X_i\right)^{-1} \\ &= \left(\sum X_i' W_i X_i\right)^{-1} \end{aligned} \tag{3.14}$$

provided $Var(y_i)$ is truly given by V_i. When α is not known but if an estimate $\hat{\alpha}$ is available we can set $\hat{V}_i = \hat{W}^{-1}$ and estimate β by using the expression with W_i replaced with by $\hat{W}_i$ (Verbeke and Molenberghs, 2000).

3.3.3 Estimation of Variance Components under MLE

Fisher (1925) introduced the general method for estimating variance components, or partitioning random variation into components from different sources, for balanced data. Hartley and Rao (1967) showed that the estimates of variance components could be obtained by using maximum likelihood methods, (Searle et al., 1992) using the equations resulting from the matrix representation of a mixed model. However, the estimates of the variance components were biased downward as previous stated because this method assumes that fixed effects are known and not estimated from data (West et al., 2007). Consider a mixed linear model for one trait, represented by (3.2). The Least Squared equations are

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix} \tag{3.15}$$

Absorbing the fixed effects reduces the equations to

$$Z'MZ\hat{u} = Z'My$$

with

$$M = I - X'(X^{'}X)^{-1}X'$$

when the inverse of $(X'X)$ does not exist, a generalized inverse can be used in its place (Meyer, 1989). By (Henderson, 1953) method 3 of ”fitting constants”, the estimates of variance components are then:

$$\hat{\sigma}_e^2 = \frac{(y'y - \hat{u}'Z'y - \hat{b}'X'y)}{N - r(X) - r(Z) + 1}$$

$$\hat{\sigma}_u^2 = \frac{(uZMy - (r(Z) - i)\hat{\sigma}_e^2)}{tr(Z'MZ)}$$

with $r(X)$ and $r(Z)$ denoting the column rank of X and Z, respectively, N the number of observations, and tr the trace operator. In this method any covariances between levels of u are ignored. An extension of method 3 to account for relationships between u has been considered by Sorensen and Kennedy (1984).

According to Searle et al. (1992), For estimating the variance covariance components that make up the element of V . When u and ε are taken as having zero covariance that are from $V = ZG^{'}Z + R$ and with V non singular , $VV^{-1} = I$ and supposing that V is a square matrix having elements that are not functionally related.

$$\frac{\partial V^{-1}}{\partial \theta} = -V^{-1}\left(\frac{\partial V}{\partial \theta}\right)V^{-1} \tag{3.16}$$

and

$$\frac{\partial}{\partial \theta}\log|V| = tr\left(V^{-1}\frac{\partial V}{\partial \theta}\right) \tag{3.17}$$

where elements of V are considered as function of θ. Using this result, we arrange the variance covariance components that occur in V as a vector $\theta_{h=1}^{v}$ where v is the total number of different components. Then to find the variance components estimates, we maximize the equation given by

$$\ell_{\theta_h} = \frac{\partial \ell}{\partial \theta_h} - \frac{1}{2}tr\left(V^{-1}\frac{\partial v}{\partial \theta}\right) + \frac{1}{2}(Y - X\beta)^{'}V^{-1}(Y - X\beta) \tag{3.18}$$

and by equating to zero

$$tr\left[\hat{V}^{-1}\left(\frac{\partial v}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\right] = (Y - X\hat{\beta})^{'}\hat{V}^{-1}\left(\frac{\partial v}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)(Y - X\hat{\beta}) \tag{3.19}$$

where $\left(\frac{\partial v}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)$ is $\left(\frac{\partial v}{\partial \theta_h}\right)$ written with $\hat{\theta}$ in place of θ with $X\hat{\beta} = X(X^{'}\hat{V}^{-1}X)^{-}X^{'}V^{-1}$ and we define P as

$$P = V^{-1} - V^{-1}X(X^{'}\hat{V}^{-1}X)^{-}X^{'}V^{-1}$$

that is

$$\hat{V}^{-1}(Y - X\hat{\beta}) = \hat{P}y$$

and we get the ML estimation equation as

$$tr\left[\hat{V}^{-1}\left(\frac{\partial v}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\right] = y^{'}\hat{P}\left(\frac{\partial v}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.20}$$

To find the estimates of variance components, we consider the derivatives of ML equation in terms of G and R. We distinguish θ_g and θ_r as elements of θ occurs in $Var(u) = G$ and $Var(\varepsilon) = R$, respectively. Then

$$\frac{\partial V}{\partial \theta_g} = Z\left(\frac{\partial G}{\partial \theta_g}\right)Z^{'} \tag{3.21}$$

and

$$\frac{\partial V}{\partial \theta_r} = \frac{\partial G}{\partial \theta_r} \tag{3.22}$$

Hence the ML equations become

$$tr\left[\hat{V}^{-1}\left(\frac{\partial G}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\right] = y^{'}\hat{P}\left(\frac{\partial G}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.23}$$

for each parameter θ_g of G, and

$$tr\left[\hat{V}^{-1}\frac{\partial R}{\partial \theta_h|_{\theta=\hat{\theta}}}\right] = y^{'}\hat{P}\left(\frac{\partial R}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.24}$$

for each parameter θ_r of R.

An alternative form of the maximum likelihood method known as REML estimation is frequently used to eliminate the bias in the ML estimates of the covariance parameters. We discuss REML estimation in next subsection.

3.3.4 Background on Restricted Maximum Likelihood (REML) Estimation

According to Searle et al. (1992), REML was first developed by Anderson and Bancroft (1952) and thereafter Russell and Bradley (1958) extend some ideas for balanced data. Patterson and Thompson (1971, 1974) were the first to introduce restricted maximum likelihood (REML) estimation as a method of estimating variance components without assuming that fixed effects are known in a general linear model with unbalanced data. It was developed because maximum likelihood estimation of the variance components does not account for the loss of degrees of freedom used in estimating the fixed parameters. The REML method includes an adjustment for degree of freedom used in estimating effects from the general linear mixed model. Verbyla and Cullis (1990) applied REML in a longitudinal data analysis by modeling variance components in the model.

Restricted Maximum Likelihood (REML) Estimation

Applying maximum likelihood to the linearly transformed response data vector is known as restricted maximum likelihood (REML). The transformation is done in such a way that the fixed effects is not contained in the vector of linearly transformed data. It was developed to avoid the biased variance components estimates that are produced by ML estimation. REML estimation applies maximum likelihood estimation technique to the likelihood function associated with a set of error contrasts rather than to that associated with original observations. This accounts for the loss of degrees of freedom resulting from estimation of the fixed effects and gives less biased estimates of the variance components. An error contrast is a linear combination $a^{'}y$ of the elements of y such that $E(a^{'}y) = 0$ for any β i.e if $a^{'}x = 0^{'}_{p}$. If we let

$$S = I_n - P_x \tag{3.25}$$

where

$$P_x = X(X^{'}X)^{-1}X^{'}$$

is the orthogonal projection matrix onto the column space of X. The expected value of S_y is

$$\begin{aligned} E(S_y) &= (I_n - P_x)X\beta \\ &= X\beta - X\beta \\ &= 0_n \end{aligned} \tag{3.26}$$

Each element of S_y is an error contrast. S is $n \times n$ and its rank is $n \times p$. The set of error contrasts contains at most $n-p$ linearly independent error contrasts, where the error contrasts $a^{'}_{1}y \cdots a^{'}_{k}y$ are linearly independent if the vector of $a_1 \cdots a_k$ are linearly independent. Let A be an $n \times (n-p)$ matrix such that $A^{'}A = I_{n-p}$ and $AA^{'} = n - Px$. We can show that $w = A^{'}y$ is a vector of n-p linearly independent error contrast. The REML approach applies maximum likelihood estimation techniques to $w = A^{'}y$ rather than to y. Under assumed model

$$y \sim N(X\beta, ZDZ^{'} + R) \tag{3.27}$$

Then

$$w \sim N(0_{n-p}, AVA^{'}) \tag{3.28}$$

The log-likelihood function $L_{REML}(\theta, y)$ associated with any vector of n-p linearly independent error contrasts is

$$L_{REML}(\theta, y) = -\frac{1}{2}\left\{\log|V| + \log|X^{'}V^{-1}X| + (y - X\hat{\beta})^{-1}V^{-1}(y - X\hat{\beta})\right\} \tag{3.29}$$

where

$$\hat{\beta} = (X^{'}V^{-1}X)^{-1}X^{'}V^{-1}y$$

In comparison the log-likelihood function for y is

$$L_{ML}(\theta, y) = -\frac{1}{2}\left\{\log|V| + (y - X\hat{\beta})^{-1}V^{-1}(y - X\hat{\beta})\right\} \tag{3.30}$$

The only difference is that $L_{REML}(\theta, y)$ has the additional term

$$-\frac{1}{2}\log|X^{'}V^{-1}X|$$

The estimator $\hat{\theta}$ is an REML estimator of θ if $L_{REML}(\theta, y)$ attains its maximum value at $\theta = \hat{\theta}$

$$\sigma^{2\,\hat{}}_{REML} = \frac{(y - X\hat{\beta})^{-1}(y - X\hat{\beta})}{n - p}$$

In general, the problem of obtaining an REML estimate of θ requires iterative methods of maximizing the nonlinear function $L_{REML}(\theta, y)$ subject to the constraint $\theta \varepsilon \Omega$, where Ω is the set of θ values for which $var(y)$ is positive-definite. An algorithm such as a Newton-Raphson and the method of Fisher scoring can be used.

3.3.5 Estimation of Fixed Effects Parameters under REML Estimation

The approach to inference is based on estimators obtained from maximizing the marginal likelihood function

$$L(\theta) = \prod_{i=1}^{n} \left\{ (2\Pi)^{-\frac{n_i}{2}} |V_i|^{-\frac{1}{2}} \times \exp(-\frac{1}{2}(y - X_i\beta)' V_i^{-1}(y - X_i\beta)) \right\} \tag{3.31}$$

with respect to θ, where $\theta = (\beta; \alpha)$. When REML estimation is used, we obtained the generalized least squares estimation of β which is given by

$$\hat{\beta} = \left[\sum_{i=1}^{n} X_i' \hat{V}_i X_i \right]^{-1} \sum_{i=1}^{n} \left(X_i' \hat{V}_i y_i \right) \tag{3.32}$$

where $\hat{V}_i$ is the REML estimates of V_i

3.3.6 Estimation of Variance Components under REML Estimation

Meyer (1989) states available methods used to get REML estimates, which can be divided in the following groups:

1. Methods using first derivatives of the likelihood function.

2. Methods using first and second derivatives of the likelihood function.

3. Derivative free methods.

For models with more random factors it is more difficult to find the maximum and it is also more difficult to construct derivatives. In categories 1 and 2, the derivatives can be calculated exact but in most methods approximations are used.

REML using derivatives

Methods use both first and second derivatives. The REML applications were based on the so-called Expectation-Maximization (EM) algorithm. This requires, implicitly, first derivatives of the likelihood to be evaluated. The resulting estimators then have the form of quadratics in the vector of random effects solutions, obtained by BLUP for the assumed values of variances to be estimated, which are equated to their expectations. For the mixed model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \hat{\alpha}A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix} \tag{3.33}$$

Note that $\hat{\alpha}$ is a function of the variance parameters that need to be estimated. Therefore, initially a prior (starting) value of α is used. The REML estimates of variance components using the EM algorithm can be obtained as:

$$\hat{\sigma}_e^2 = \frac{(y'y - \hat{b}'X'y - \hat{u}'Z'y)}{N - r(X)}$$

$$\hat{\sigma}_u^2 = \frac{\hat{u}'A^{-1}u + tr(A^{-1}C)\hat{\sigma}_e^2)}{q}$$

where N is the number of observations, q is the number of random genetic effect levels and C the part of the inverse of the mixed model equations that corresponds with the random effects and a denote the vector of addictive effects.

REML using the Average Information algorithm

First we discuss more formally first and second derivatives of the likelihood function. Then, the mechanism of an AI algorithm will be presented. The partial derivatives of $\ln|V|$ in equation 3.12 with respect to the variance of random effects, σ_i^2 (e.g. i = a and e) can be obtained from matrix theory (Searle, 1982). In estimating the variance from REML we make a transformation from ML to REML by making some replacements (Searle et al., 1992), namely: y by $K^{'}$y, Z by $K^{'}$Z, X by $K^{'}$X = 0, and V by $K^{'}$VK and P is replaced by $(K^{'}VK)^{-1}$ and $\hat{P} = K(K^{'}VK)^{-1}K^{'}$

With all the above replacements in the ML equation the REML equation becomes

$$tr\left[Z^{'}K(K^{'}\hat{V}K)^{-1}K^{'}Z\left(\frac{\partial G}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)\right] = y^{'}K(K^{'}\hat{V}K)^{-1}K^{'}Z\left(\frac{\partial G}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)Z^{'}K(K^{'}\hat{V}K)^{-1}K^{'}y \tag{3.34}$$

and

$$tr\left[(K^{'}\hat{V}K)^{-1}K^{'}\left(\frac{\partial R}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)K\right] = y^{'}K(K^{'}\hat{V}K)^{-1}K^{'}\left(\frac{\partial R}{\partial \theta_h|_{\theta=\hat{\theta}}}\right)K(K^{'}VK)^{-1}K^{'}y \tag{3.35}$$

which is reduced for each parameter θ_g in G to

$$tr\left[Z^{'}\hat{P}Z\frac{\partial G}{\partial\theta_h|_{\theta=\hat{\theta}}}\right]=y^{'}\hat{P}Z\left(\frac{\partial G}{\partial\theta_h|_{\theta=\hat{\theta}}}\right)Z\hat{P}y \tag{3.36}$$

for each parameter θ_r in R to

$$tr\left[\hat{P}\frac{\partial R}{\partial\theta_h|_{\theta=\hat{\theta}}}\right]=y^{'}\hat{P}\left(\frac{\partial R}{\partial\theta_h|_{\theta=\hat{\theta}}}\right)\hat{P}y \tag{3.37}$$

We note that in both ML and REML equations, there right hand side are the same.

3.3.7 Estimation (or Prediction) of Random Effects Parameters

Statistical models that include random effects are commonly used to analyze longitudinal and clustered data. These models are often used to derive predicted values of the random effects. In the prediction of random effects we look at the conditional expectations of the random effects given the observed response values, y_i, in Y_i. According to Verbeke and Molenberghs (2000) and Fitzmaurice et al. (2004), let $\hat{u}$ be the predictor for u. The best predictor for u is the conditional mean of u_i given the vector of response y_i and it is given by

$$\tilde{u}=E(u_i|y_i)$$

We consider the model shown below for each individual i :

$$y_i=X_i\beta+Z_iu_i+\varepsilon_i \tag{3.38}$$

If the $Cov(u_i,y_i)=GZ^{'}$ where y_i is the response vector and u_i vector of the individual specific parameters, then we have the joint multivariate normal distribution

$$\begin{pmatrix} y_i \\ u_i \end{pmatrix}\sim N\left[\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix},\begin{pmatrix} ZGZ^{'}+\sigma^2I_{ni} & ZG \\ GZ^{'} & G \end{pmatrix}\right] \tag{3.39}$$

Thus, the best predictor of u_i is the conditional mean of u_i given the vector of response y_i

$$\begin{aligned}\hat{U}_i=E(u_i|y_i)&=E(u_i)+GZ^{'}V^{-1}(y_i-X_i\beta)\\&=GZ^{'}V^{-1}(y_i-X_i\beta)\end{aligned} \tag{3.40}$$

since $E(u_i)=0$

If β is unknown, we use it estimate ($\hat{\beta}$) and find that

$$\begin{aligned}E(u_i|y_i)&=GZ^{'}V^{-1}(y_i-X_i\hat{\beta})\\&=GZ^{'}(V^{-1}-V^{-1}X(^{'}V^{-1}X)^{-1}X^{'}V^{-1})y\\&=GZ^{'}Py\end{aligned} \tag{3.41}$$

and the variance of $\tilde{u}$ is given by

$$Var(\tilde{u}) = GZ^{'}PZG$$

where

$$P = V^{-1} - V^{-1}X(^{'}V^{-1}X)^{-1}X^{'}V^{-1}$$

Then the distribution of $\tilde{u}$

$$\tilde{u} \sim N(0, GZ^{'}PZG)$$

This is known as the best linear unbiased predictor (BLUP). This predictor of u_i depends upon the unknown covariance among the y_i. When the unknown covariance parameters are replaced by their REML or ML estimates, the resulting

$$E(u_i|y_i) = \hat{G}Z^{'}\hat{V}^{-1}(y_i - X_i\hat{\beta})$$

is referred to as the empirical BLUP. Given the empirical BLUP $\hat{u}$, we can obtain the ι^{th} subjects predicted response profile as follows

$$\hat{y}_i = X_i\hat{\beta} + Z_i\hat{u}_i$$

3.4 Inference for the Fixed Parameters

As previously stated, the vector β of fixed effects that was introduced by Laird and Ware (1982) is estimated by

$$\hat{\beta} = \left(\sum X_i^{'}W_iX_i\right)^{-1}\sum X_i^{'}W_iy_i$$

In which the unknown vector α of variance components is replaced by its ML or REML estimates. Under the marginal model above and conditional on α and $\hat{\beta}$ follows a multivariate normal distribution with mean vector β and with variance-covariance matrix

$$\begin{aligned} Var(\hat{\beta}) &= \left(\sum X_i^{'}W_iX_i\right)^{-1}\left(\sum X_i^{'}W_ivar(y_i)\sum W_iX_i\right)\left(\sum X_i^{'}W_iX_i\right)^{-1} \\ &= \left(\sum X_i^{'}W_iX_i\right)^{-1} \end{aligned} \tag{3.42}$$

where W_i equals to V_i^{-1} and assuming that $Var(y_i) = V_i$.

3.4.1 Approximate Wald Test

According to Verbeke and Molenberghs (2000), this follows from likelihood theory under some conditions of the distribution of the ML and REML estimator $\hat{\alpha}$ can be approximated by a normal distribution with mean vector α and with covariance matrix which is given by the inverse of the fisher information matrix. Hence, the approximate standard errors for the estimates of variance components in α can be calculated from inverting minus the matrix of second-order partial derivatives of the log-likelihood function of ML or REML with respect to α. Using the asymptotic normality of the parameter estimates, approximate Wald tests and approximate Wald confidence intervals can be obtained for fixed effects

Now we consider important hypotheses about β of Equation 3.32. For each parameters β_j in β, $j = 1; \cdots ; p$, we can test the hypothesis $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$ using an approximate Wald statistics test as well as an associated confidence interval, by approximating the distribution of

$$Z = \frac{\hat{\beta}_j - \beta_j}{s.e(\hat{\beta}_j)} \tag{3.43}$$

by a standard multivariate normal distribution (Verbeke and Molenberghs, 2000). Suppose that L is a single row vector then $LCov(\hat{\beta})L^{'}$ is a single value and its square roots provides an estimates of the standard error for $L\hat{\beta}$. Thus an approximate 95% confidence interval is given by

$$L\hat{\beta} \pm 1.96\sqrt{Lcov(\hat{\beta})L'} \tag{3.44}$$

To test our estimates using Wald test, we consider the assumption that $\hat{\beta}$ is asymptomatically normal with mean β and covariance matrix, for any known matrix L, then for the hypothesis

$$H_0 : L\beta = 0 \quad \text{vs} \quad H_A : L\beta \neq 0$$

Then Wald statistic is given by

$$Z = \frac{L\hat{\beta}}{\sqrt{LCov(\hat{\beta})L'}} \tag{3.45}$$

and compare with standard normal distribution. If Z is a standard normal random variable, then Z^2 has a χ^2 distribution with 1 df. Thus, the test statistic is

$$W = (L\hat{\beta})(LCov(\hat{\beta})L^{'})(L\hat{\beta})$$

and is compared to χ^2 with 1 df. Now suppose that L has r rows, then the test is given by

$$W = (L\hat{\beta})(LCov(\hat{\beta})L^{'})(L\hat{\beta}) \tag{3.46}$$

which has a χ^2 distribution with r df (Verbeke and Molenberghs, 2000).

3.5 Inference for the Random Effects

In this section the problem of making the inference on the random effects u_i is discussed. In particular the idea of empirical Bayes and best linear unbiased predictors (BLUP) will be given attention (Verbeke and Molenberghs, 2000; Davis 2002; and Fitzmaurice et al., 2004). The concept of shrinkage estimators will be derived and the normality assumption for random effects discussed.

3.5.1 Empirical Bayes (EB) Inference

Consider the linear mixed model

$$y_i = X_i\beta + Z_iu_i + \varepsilon_i \tag{3.47}$$

where $u_i \sim N(0,G)$, $\varepsilon \sim N(0,R_i)$ and the u_i and ε_i are independent. The random effects u_i reflects how the evolution for the i^{th} subject deviates from the expected evolution $X_i\beta$. Estimation of the random effects u_i is helpful for detecting the outlying profile from the expected profile. Under hierarchical model assumptions, inference of random effects is important. The Hierarchized model can be specified as

$$Y_i|u_i \sim N(X_i\beta + Z_iu_i, R_i)$$

and $u_i \sim N(0,G)$ implying that

$$E(Y_i|u_i) = X_i\beta + Z_iu_i$$

Since u_i is a random parameters, then we consider Bayesian approaches where the prior distribution of random parameter is $u_i \sim N(0,G)$. Thus using the Bayes rule we can express the posterior distribution of the u_i given $Y_i = y_i$ as

$$f(u_i|y_i) = \frac{f(y_i|u_i)f(u_i)}{\int f(y_i|u_i)f(u_i)du_i} \tag{3.48}$$

since we know the distribution of u_i and the conditional distribution $y_{ij}|u_i$, we can show the posterior distribution of u_i is given by

$$u_i|y_{ij} \sim N(GZ_i^{'}W_i(y_i - X_i\beta), \xi_i)$$

for some matrix ξ_i after some algebraic manipulation. Thus we can use the posterior mean of u_i as an estimate of u_i that is

$$\begin{aligned} \hat{u}_i &= E(u_i|y_i) \\ &= \int f(u_i|y_i)du_i \\ &= GZ_i^{'}W_i(\alpha)(y_i - X_i\hat{\beta}) \end{aligned} \tag{3.49}$$

and the variance is given by

$$Var(\hat{u}_i) = GZ_i^{'}W_i - W_iX_i\left(\sum X_i^{'}W_iX_i\right)^{-1}X_i^{'}W_iZ_iG \tag{3.50}$$

however inference on u_i should take into account the variability in u_i, the inference for u_i is usually based on

$$\begin{aligned} Var(\hat{u}_i - u_i) &= G - Var(\hat{u}_i) \\ &= G - GZ^{'}PZG \end{aligned} \tag{3.51}$$

Thus for inference purposes once the correlated variance in the equation above is found, Wald test can be constructed to test hypothesis about $\hat{u}_i(\theta)$. Parameters in θ are replaced by their ML and REML estimates, obtained after fitting the marginal model. The estimates $\hat{u}_i = \hat{u}_i(\theta)$ is called the empirical Bayes estimates of u_i. Approximate t and F tests can be constructed in similar ways to test for fixed effects to take account for the variability introduced by replacing θ by $\hat{\theta}$.

3.5.2 Best Linear Unbiased Prediction (BLUP) of Random Effect

The resulting predictor is a best linear unbiased predictor and if β is unknown we use $\hat{\beta}$ to get

$$\begin{aligned} E(u_i|y_i) &= GZ^{'}V^{-1}(y_i - X_i\hat{\beta}) \\ &= GZ^{'}(V^{-1} - V^{-1}X(^{'}V^{-1}X)^{-1}X^{'}V^{-1})y = GZ^{'}Py \end{aligned} \tag{3.52}$$

and the variance of $\tilde{u}$ is given by

$$Var(\tilde{u}) = GZ^{'}PZG$$

where

$$P = V^{-1} - V^{-1}X(^{'}V^{-1}X)^{-1}X^{'}V^{-1}$$

The BLUPs are unbiased in a sense of population

$$E(\hat{u}_i) = E(u_i) = 0$$

but conditionally biased towards zero

$$E(\hat{u}_i|u) = GZ^{'}PZu$$

The variance of the predictor is

$$var(\hat{u}_i) = GZ^{'}PZG$$

but the variation is usually considered in terms of the prediction error variance

$$var(\hat{u}_i - u_i) = G - GZ^{'}PZG$$

which measures variation in terms of distance from the unobserved true value.

Note that the i^{th} subjects predicted response profile is as follows

$$\begin{aligned} \hat{y}_i &= X_i\hat{\beta} + Z_i\hat{u}_i \\ &= X_i\hat{\beta} + Z_iGZ^{'}V_i^{-1}(y_i - X_i\hat{\beta}) \\ &= \hat{R}_i\hat{\Sigma}_i^{-1}X_i\hat{\beta} + (I_{n_i} - \hat{R}_i\hat{\Sigma}_i^{-1})y_i \end{aligned} \tag{3.53}$$

That is, the i^{th} subject's predicted response profile is a weighted combination of the population-averaged mean response profile, $X_i\hat{\beta}$, and the i^{th} subject's observed response profile Y_i. According to Fitzmaurice et al. (2004), the subject's predicted response profile is shrunk towards the population-averaged mean response profile. The amount of shrinkage depends on the relative magnitude of R_i and Σ_i. We note that R_i characterizes the within-subject variability, while Σ_i incorporates both within subject and between-subject sources of variability. Thus, when R_i is large, and the within-subject variability is greater than the between-subject variability, more weight is given to $X_i\hat{\beta}$, the population-averaged mean response profile. When the between-subject variability is greater than the within-subject variability, more weight is given to the i^{th} subject's observed data Y_i.

3.6 Inference for the Variance Components

The mean structure is usually of primary interest in the inference, however inference for the covariance structure could be of the interest for some good reasons. Covariance modeling is useful for interpretation of the random variation in the data and it is absolutely necessary to obtain valid model-based inference for parameters in the mean structure of the model. A test for variance component helps in proving or establishing whether we do need the inclusion of random effects or not. It also important to note that in an over-parameterized model the covariance structure leads to inefficient inference and a potentially poor assessment of standard errors for estimates of the mean response profile, whereas a too restrictive specification invalidates inference about the mean structure when the assumed covariance structure does not hold (Verbeke and Molenberghs, 2000).

3.6.1 Approximate Wald Test

Using the asymptotic normality of the parameter estimates, approximate Wald tests and approximate Wald confidence intervals can be obtained similarly as for fixed effects in Section 3.4.1.

3.6.2 The Likelihood Ratio Test (LR)

In this case we are interested in comparing two nested models which are the full model and the reduced model. The reduced model is a special case of full model and the reduced model is simpler than the full model. The reduced model is nested within the full model. The LR test can be derived by comparing their maximized log-likelihood, say $\hat{\ell}_{full}$ and $\hat{\ell}_{reduced}$, where the LR test statistic is given by

$$LR = -2\ell_n\lambda_n = 2(\hat{\ell}_{full} - \hat{\ell}_{reduced}) \tag{3.54}$$

and comparing the statistical test to a χ^2 distribution with degree of freedom equal to the difference between the number of parameter in the full and reduced models. The larger the difference between the $\hat{\ell}_{full}$ and $\hat{\ell}_{reduced}$ the stronger the evidence that the reduced model is not sufficient. We note that the valid LR test can still be obtained under REML since the error contrasts U are the same in both cases, namely H_0 and H_a as long as the comparison is under the same mean structure.

3.7 Random Coefficient Models

A random coefficients model is a special case of the linear mixed model for longitudinal data (Brown and Prescott, 2006) and will be discussed below. The random effects are the covariates effect that vary among subjects. These effects are subject-specific and hence are random since each subject is randomly drawn from the population (Hedeker and Gibbons, 2006). In the analysis of random coefficients, there are three question of interest in assessing a model. First, is it a good model? Second, is a more complex model better? Finally, what contribution does individual predictors make to the model?, In order to assess models, the different model fit statistics would be examined. As an example of random coefficients models, we consider the longitudinal data used by Diggle et al. (2002) consisting of weight measurements of 48 pigs on 9 successive weeks. Pigs were identified by the variable *id*. The overall weight measurements vary from pig to pig. We treat them as random samples from a larger population and model the between-pig variability as a random effect.

3.7.1 Random Intercept Model

We follow the description of Laird and Ware (1982), Longford (1995) and Verbeke and Molenberghs (2000). The random intercepts model is a model in which intercepts are allowed to vary; the scores on the dependent variable for each individual observation are predicted by the intercept that varies across groups. This model assumes that the slopes are fixed. The model provides information about intraclass correlation coefficient which is helpful in determining whether multilevel models are required in the first place (West et al., (2007)). The random intercept model is given by:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + u_i + \varepsilon_{ij} \tag{3.55}$$

where $i = 1, 2, \ldots, N$, $u_i \sim N(0, \sigma_u^2)$ is the random subject effect, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ are within subject measurement errors. u_i and ε_{ij} are assumed to be independent of each other. Where β represents the mean changes over time in the population of interest, u_i represents the i^{th} individual deviation from the population mean intercept after the effects of covariates have been accounted for. From the equation above we can get the mean response over time for the i^{th} individual. The conditional mean of Y_{ij} given the subject specific effect u_i is is given by $E(Y_{ij}|u_i) = X_{ij}^T\beta + u_i$ and the marginal mean is given by $E(Y_{ij}) = X_{ij}^T\beta$. The marginal response variance for each response is given by

$$\begin{aligned} Var(Y_{ij}) &= Var(X_{ij}^T\beta + u_i + \varepsilon_{ij}) \\ &= Var(u_i + \varepsilon_{ij}) \\ &= \sigma_u^2 + \sigma_\varepsilon^2 \end{aligned} \tag{3.56}$$

and the marginal covariance between any two pair of responses (Y_{ij} and Y_{ik}) is given by:

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= Cov(X_{ij}^T\beta + u_i + \varepsilon_{ij}, X_{ik}^T\beta + u_i + \varepsilon_{ik}) \\ &= Cov(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ik}) \\ &= Cov(u_i, u_i) = \sigma_u^2 \end{aligned} \tag{3.57}$$

Therefore

$$\begin{aligned} Corr(Y_{ij}, Y_{ik}) &= \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{Var(Y_{ij})Var(Y_{ik})}} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \\ &= \rho \end{aligned} \tag{3.58}$$

Since the random intercept model implies a compound symmetry assumption for the variance and covariance of the longitudinal data we assume that the variance is constant over time, say σ^2 and $Corr(y_{ij}; y_{ik}) = \rho$ for all j and k, then the compound structure is given by

$$Corr(y_{ij}, y_{ik}) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

with the constraint that $\rho \geq 0$. An example of random intercept model is taken from an example in section 3.7 above, we specified random intercept term at the pig level. We thus wish to fit the model

$$Weight_{ij} = \beta_0 + \beta_1 week_{ij} + u_j + \varepsilon_{ij} \tag{3.59}$$

For all $i = 1 \cdots 9$ and $j = 1 \cdots 48$ pigs. The fixed part of the model, $\beta_0 + \beta_1 week_{ij}$, simply states that we want one overall regression line representing the population average. The random effect u_j occurs at the pig level (id).

3.7.2 Random Slopes Model

A random slopes model is a model in which slopes are allowed to vary and furthermore the slopes are different across groups. This model assumes that intercepts are fixed. The model is given by:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + u_{1i} t_{ij} + \varepsilon_{ij} \tag{3.60}$$

where $j =, \cdots, n$. Since we assume normality in model, the $Var(u_{1i}) = \sigma_1^2$ and $Cov(u_{0i}, u_{1i}) = 0$ since there is no intercepts. We extend the above example to allow for a random slope on week yields the model

$$Weight_{ij} = \beta_0 + \beta_1 week_{ij} + u_{1j} week_{ij} + \varepsilon_{ij} \tag{3.61}$$

3.7.3 Random Intercepts and Slopes Model

A model that includes both random intercepts and slopes is likely the most realistic type of model, although it is also more complex. In this model, both intercepts and slopes are allowed to vary across groups, and meanings are different in different contexts. Consider the model:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + u_{0i} + u_{1i} t_{ij} + \varepsilon_{ij} \tag{3.62}$$

or

$$Y_{ij} = X_{ij}^{'} \beta + Z_{ij}^{'} u_i + \varepsilon_{ij} \tag{3.63}$$

where $j = 1, \cdots, n$, $X_{ij}^{'} = (1 \;\; t_{ij})$, $Z_{ij}^{'} = (1 \;\; t_{ij})$, $u_{0i} \sim N(0, b_{11})$ and $u_{1i} \sim N(0, b_{22})$ are random intercept and the random slope respectively. $Cov(u_{1i}, u_{2i}) = b_{12}$ are within subject measurement errors. Note that $u_i = (u_{1i}, u_{2i})^{'}$ and $\beta = (\beta_1, \beta_2)^{'}$. Since we assume the normality in model then $u_i \sim MVN(0, \Sigma_i)$, we then consider the covariance among the components of Y_i in this linear mixed effects model with randomly varying intercepts and slope. Let a $Var(u_{0i}) = \sigma_0^2$, $Var(u_{1i}) = \sigma_1^2$ and $Cov(u_{0i}, u_{1i}) = \sigma_{01}$. The random effects u_{io} and u_{i1} are assumed to have bivariate normal distribution

$$\begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00}^2 & \sigma_{01} \\ \sigma_{01} & \sigma_{11}^2 \end{bmatrix} \right) \tag{3.64}$$

Then from the model (3.60) above, $\beta_0 + u_{io}$ is the intercept for subject i which implies u_{oi} means the deviation of the intercept of the subject i from population intercept β_o and also $\beta_1 + u_{i1}$ is the slope for the subject i therefore u_{i1} is the deviation of the slope of the subject i from the population slope β_1. According to Fitzmaurice et al. (2004), the unique elements of the (2×2) covariance matrix $G = Cov(u_i)$ we in addition assume that $R_i = Cov(\varepsilon_i) = \sigma^2 I_{ni}$. The conditional mean of Y_{ij} given the subject specific effect u_i is given by:

$$E(Y|u_i) = \beta_1 + \beta_2 t_{ij} + u_{0i} + u_{1i} t_{ij}$$

and the marginal mean and the variance of Y_{ij} is therefore given by:

$$E(Y_{ij}) = \beta_1 + \beta_2 t_{ij}$$

and

$$\begin{aligned} Var(y_{ij}) &= Var(u_{0i}) + 2t_{ij}Cov(u_{0i}, u_{1i}) + t_{ij}^2 Var(u_{1i}) + Var(\varepsilon_{ij}) \\ &= \sigma_1^2 + 2\sigma_{01} t_{ij} + \sigma_1^2 t_{ij}^2 + \sigma^2 \end{aligned} \tag{3.65}$$

respectively. Thus the variance under this model will vary over time and it is not constant over time and σ^2 is a variance within subject. The covariance can be shown that:

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik}) \\ &= \sigma_0^2 + t_{ij}\sigma_{10} + t_{ik}\sigma_{01} + t_{ij}t_{ik}\sigma_1^2 \\ &= \sigma_0^2 + (t_{ij} + t_{ik})\sigma_{01} + t_{ij}t_{ik}\sigma_1^2 \end{aligned} \tag{3.66}$$

Thus:

$$Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_0^2 + (t_{ij} + t_{ik})\sigma_{01} + t_{ij}t_{ik}\sigma_1^2}{\sqrt{Var(Y_{ij})Var(Y_{ik})}} \tag{3.67}$$

and the correlation is not constant over time. The example of random intercept and slope model from example in 3.7, by extending the example to allow for a random intercept and slope on week yields the model

$$Weight_{ij} = \beta_0 + \beta_1 week_{ij} + u_{0j} + u_{1j} week_{ij} + \varepsilon_{ij} \tag{3.68}$$

3.8 Types of Correlation/Covariance Structures

In the statistical analysis models for multivariate model for repeated measures is that the assessments for each individual are assumed to be correlated over time. The main difference between a univariate regression for independent observations and a multivariate model for repeated measures is that the results for each individual are bound to be correlated over time. Then we present some number of covariance structures that can be assumed to account for such correlation. A summary of some covariance structures that can be used are listed in the Table below.

Table 3.1: Summary of covariance structures

Structure	Description	No. of parameters	$\{i,j\}^{th}$ elements
VC	Variance components	q	$\sigma_{ij} = \sigma_k^2$
AR(1)	First order autoregressive	2	$\sigma_{ij} = \sigma^2\rho^{i-j}$
CS	Compound symmetry	2	$\sigma_{ij} = \sigma_1^2 + \sigma^2 I$
Toep	Toeplitz	m	$\sigma_{ij} = \sigma_{\|i-j\|+1}$
UN	Unstructured	m(m+1)/2	$\sigma_{ij} = \sigma_{ij}$
SP(POW)	Power spatial	2	$\sigma^2\rho^{d_{ij}}$
SP(EXP)	Exponential spatial	2	$\sigma^2\exp(-\frac{d_{ij}}{\rho})$
SP(GAU)	Gaussian spatial	2	$\sigma^2\exp(-\frac{d_{ij}^2}{\rho^2})$

Independence Structure

It assumes repeated measures are uncorrelated and the corresponding covariance structure for four observations per subject is given by:

$$\sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Autoregressive Order One (AR(1))

The problem of unequal correlation can be solved by using several approaches such as the AR(1) covariance structure. The correlation between m time units apart is ρ^m, $0 < \rho < 1$. The greater the power or distance (m), the smaller the magnitude of the covariance will be. Thus the further the measurements are apart, the lower their correlation. This covariance structure depends on two parameters $q = 2$ and the covariance for two points i.e. j and $j^{'}$ equals

$$\sigma_{jj'} = \sigma^2\rho^{|j-j^{'}|}$$

where ρ is the AR(1) parameter and σ^2 is the error variance thus the covariance structure is given by:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}$$

Compound Symmetry

This structure assumes the covariances are homogeneous. The correlation between two separate measurements is assumed to be constant no matter how far apart the measurements are. This is unrealistic in longitudinal data problem in the sense that observations closer to each other are more correlated than the ones which are further apart. It is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

Unstructured

This is the most flexible since it assumes all the variance and covariances are different. This lets the data dictate what they should be and requires the estimate of many parameters, but the more data that are used to assess the covariance structure the less data are left to estimate the parameters of linear models. The analysis that uses an UN matrix will be less powerful than analysis that uses the proper structures. It is expressed as:

$$\begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

Toeplitz

The Toeplitz structure is similar to AR(1) in that all measurements at same distance have the same correlation. But there is no assumption of exponential decay. The AR(1) is a special case of the Toeplitz and AR(1) can be estimated with single parameter and then exponentiate with the distance. The Toeplitz model has as many parameters due to distance. Toeplitz and AR(1) are reasonable choices for equally spaced observations. The covariance matrix is given by:

$$\sigma^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}$$

The Exponential Spatial Covariance Structure

This covariance structure is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & \exp(-\frac{d_{12}}{\rho}) & \cdots & \exp(-\frac{d_{1n}}{\rho}) \\ \exp(-\frac{d_{21}}{\rho}) & 1 & \cdots & \exp(-\frac{d_{2n}}{\rho}) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(-\frac{d_{n1}}{\rho}) & \exp(-\frac{d_{n2}}{\rho}) & \cdots & 1 \end{pmatrix}$$

The Power Spatial Covariance Structure

For this case, correlations decline with increasing spacing between points. It is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & \rho^{d_{i12}} & \cdots & \rho^{d_{i1(n-1)}} \\ \rho^{d_{i21}} & 1 & \cdots & \rho^{d_{i2(n-2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{d_{i(n-1)1}} & \rho^{d_{i(n-2)2}} & \cdots & 1 \end{pmatrix}$$

The Gaussian Spatial Covariance Structure

For Gaussian spatial covariance structure correlation declines with increasing distance between points. It is expressed as:

$$\sigma^2 \begin{pmatrix} 1 & \exp(-\frac{d_{12}^2}{\rho^2}) & \cdots & \exp(-\frac{d_{1n}^2}{\rho^2}) \\ \exp(-\frac{d_{21}^2}{\rho^2}) & 1 & \cdots & \exp(-\frac{d_{2n}^2}{\rho^2}) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(-\frac{d_{n1}^2}{\rho^2}) & \exp(-\frac{d_{n2}^2}{\rho^2}) & \cdots & 1 \end{pmatrix}$$

In all spatial covariance structure d_{ijk} is the distance between J^{th} and k^{th} observations within subject i and $0 < \rho < 1$. The major advantage of the spatial type structures over the AR structure which assumes equal spaced observation is that they make use of actual distance between observations which allows the modeller to be in a position to deal with unequally spaced observations within and between observations. Toeplitz and the autoregressive order allow observations that are far apart to be less strongly correlated and the correlation between two observations is a function of the separation between the observations (Verbeke and Molenberghs, 2009).

3.8.1 Model Selection

The following criteria can be used to compare the goodness-of-fit of two models. The AIC is useful for non-nested models. If two models are nested then LR is used. The Akaike's information criteria (AIC) was introduced by Akaike (1974) and Schwarz criteria (SC) (also known as Bayesian information criteria (BIC) was introduced by Schwarz et al. (1978). According to

Vittinghoff et al. (2005) these methods are used to adjust the likelihood ratio statistic $-2\log L$ which measures the deviation of the maximal possible model. The adjustment is necessary because the $-2\log L$ will always decrease as a new explanatory variable enters the model even if it is insignificant (Moeti, 2010). The AIC is given by

$$AIC = -2\log L + 2p$$

where p is the number of parameters in the model. Another criteria which adjusts the $-2\log L$ statistic for the number of parameters is SC or BIC and is given by

$$SC = -2\log L + p\log(n)$$

where p is as explained above and n is the overall sample size. According to Al-Marshadi (2011) there are two more model selection criteria that will be considered in the study. These are bias-correlated Akaike's Information Criterion (AICC) by Hurvich and Tsai (1989) and is given by

$$AICC = -2\ell + 2p(\log n + 1)$$

and Hannan and Quinn Information Criterion (HQIC) by Hannan and Quinn (1979) and is given by

$$HQIC = -2\ell + 2p\log\log n$$

In our study we are interested to compare the four information criteria in terms of their ability to identify the true structure model order with and without the help of other approaches. The smaller the value of the criteria, the better the goodness-of-fit of the model (Caley and Hone, 2002; Anderson et al., 1994).

3.9 Checking Model Assumptions (Diagnostics)

According to West et al. (2007) it is important to carry out model diagnostics to check whether distributional assumptions for the residuals are satisfied and whether the fit of the model is sensitive to unusual observations. Model diagnostics should be part of the model building process throughout the analysis of a clustered or longitudinal data set (West et al., 2007). According to Zewotir and Galpin (2005) we propose and investigate a number of diagnostics for variance components ratios, fixed effects parameters, prediction of the response variable and of random effects and the likelihood function. In this case, we focus on the definitions of a selected set of terms related to residual and influence diagnostics in LMMs.

3.9.1 Residual Diagnostics

A residual is the difference between an observed quantity and its estimated or predicted value. In the context of standard linear model is used to decide whether a given set of residuals plotted against predicted values presents a random pattern or not. The residual versus fitted plots are used to verify model assumptions and to detect outliers and potentially influential observations. According to West et al (2007) and Schabenberger (2005) residual should be assessed for normality, constant variance and outliers. In the LMMs we consider marginal residual (r_m), conditional residual (r_c) and their studentized version as described in the following subsections.

Marginal and Conditional Residuals

A marginal residual is the difference between the observed data and the estimated marginal mean. The name marginal comes from the fact that $X_i^{'}\hat{\beta}$ is the estimated marginal mean of y_i. The equation is given by

$$r_{m_i} = y_i - X_i^{'}\hat{\beta} \tag{3.69}$$

A conditional residual is the difference between the observed value and the conditional predicted value of the dependent variable. The equation for the vector of conditional residuals for a given individual i in a two-level longitudinal data set is written as follows

$$r_{c_i} = \hat{\varepsilon}_i = y_i - X_i^{'}\hat{\beta} - Z_i^{'}\hat{u}_i \tag{3.70}$$

The name conditional residual comes from the fact that $X_i^{'}\hat{\beta} + Z_i^{'}\hat{u}_i$ is the conditional mean of y_i. Residuals are used to examine model assumptions and to detect outliers and potentially influential data point. According to Schabenberger (2005) and West et al. (2007), the raw residuals r_{m_i} and r_{c_i} are usually not well suited for verifying these purposes. Even if the true model residuals or errors are uncorrelated and have equal variance, the residuals will tend to be correlated and their variance will differ. To account for the unequal variance of the residuals, various studentizations are applied.

Standard and Studentized Residuals

According to Schabenberger (2005) and West et al. (2007) a random variable is said to be standardized if the difference from its mean is scaled by standard deviation. Unfortunately, the true standard deviations are rarely known in practice, so scaling is done by using estimated standard deviations instead. The residual have mean zero but their variance is unknown as it depends on the true values of θ. Standardization is thus not possible in practice. The method of scaling residuals to divide them by the estimated standard deviation of the dependent variable is resulting to Person residuals. If the estimate is independent of the i^{th} observation , the process is called external studentizations because this is accomplished by excluding the i^{th} observation when computing the estimate of its standard error. If the observation contributes to the standard error computation then residual is said to be internally studentized (West et al., 2007). Rather than divide each individual residual by the variance of an observation, we can also consider the vector of residuals and the estimated variance $V(\hat{\theta})$. Let $\hat{C}$ denote a matrix such that $\hat{C}\hat{C}' = V(\hat{\theta})$. Then the scaled residual $r_c = \hat{C}^{-1} r_m$ have zero mean and are approximately uncorrelated. They are not exactly uncorrelated because $\hat{C}$ is an estimated matrix and V is not the variance of r_m. Scaled residuals can be useful to diagnose to ascertain whether the covariance structure of the mixed model has been specified correctly. Table 3.2 summarizes the available residuals.

Table 3.2: Summarizes available residuals

Types of Residual	Marginal	Conditional
Raw	$r_{m_i} = y_i - X_i'\hat{\beta}$	$r_{c_i} = \hat{\varepsilon}_i = y_i - X_i'\hat{\beta} - Z_i'\hat{u}_i$
Studentized	$r_{m_i}^{student} = \frac{r_{m_i}}{\sqrt{v\hat{a}r[r_{m_i}]}}$	$r_{c_i}^{student} = \frac{r_{c_i}}{\sqrt{v\hat{a}r[r_{c_i}]}}$
Pearson	$r_{m_i}^{person} = \frac{r_{m_i}}{\sqrt{v\hat{a}r[Y_i]}}$	$r_{c_i}^{person} = \frac{r_{c_i}}{\sqrt{v\hat{a}r[Y_i \mid u]}}$
Scaled	$\hat{C}^{-1} r_m$	

3.9.2 Influence Diagnostics

Influence diagnostics are formal techniques that allow one to identify observations that heavily influence estimates of the parameters in either β or θ. The idea of influence diagnostics for a given observation is to quantify the effect of omission of those observation from the data on the results of the analysis of the entire data set. The key to the implementations of influence diagnostics in the mixed procedure is the attempt to quantifying influence where possible by drawing on the basic definitions of the various statistics in the classical linear model. The basic procedure for quantifying influence is:

1. Fit the model to the data and obtain estimates of all parameters

2. Remove one or more data points from the analysis and compute updated estimates of model parameters

3. Based on full- and reduced-data estimates, contrast quantities on interest to determine how the absence of the observations changes the analysis.

Overall Influence

According to Schabenberger (2005) an overall influence statistic measures the change in the objective function being minimized. In linear mixed models fitted by maximum likelihood (ML) and restricted maximum likelihood (REML), an overall influence measure is the likelihood distance which is also referred to as the likelihood displacement. The idea is to compute the full data parameter estimates $\hat{\psi}$ and estimates based on the reduced data $\hat{\psi}_u$. The likelihood and restricted distances are obtained as

$$LD_{(u)} = 2\{\ell(\hat{\psi}) - \ell(\hat{\psi}_{(u)})\}$$

$$RLD_{(u)} = 2\{\ell_R(\hat{\psi}) - \ell_R(\hat{\psi}_{(u)})\}$$

The likelihood distance gives the amount by which the log-likelihood of the full data changes if one was to evaluate it at the reduced parameter model.

Change In Parameter Estimates

According to Schabenberger (2005) and West et al. (2007), the main difference between the Cook's distance and the MDFFITS statistic is that the MDFFITS statistic uses an externalized estimate of the variance of the parameter estimates which is based on recalculated covariance estimates using the reduced data while Cook's distance does not. The theory of Cook's distance was first introduced by Cook (1977). For the fixed effects, the two statistics are

$$D(\beta) = (\hat{\beta} - \hat{\beta}_{(u)})' v\hat{a}r(\hat{\beta})^{-1}(\hat{\beta} - \hat{\beta}_{(u)})/rank(X)$$

$$MDFFITS = (\hat{\beta} - \hat{\beta}_{(u)})' v\hat{a}r(\widehat{\beta_{(u)}})^{-1}(\hat{\beta} - \hat{\beta}_{(u)})/rank(X)$$

If the covariance parameters are updated during the influence analysis, similar statistics can be computed for $\hat{\theta}$. However, the D(θ) and MDFFITS(θ) statistics does not involve division by a matrix rank.

Change In Precision of Estimates

The effect on the precision of estimates is separate from effect on the point estimates. If the influence on the precision of the estimates is large, the MIXED procedure computes functions of the trace and determinants of the variance matrices based on the full data and the reduced data estimates

$$COVTRACE(\beta) = |trace(v\hat{a}r[\hat{\beta}]^{-1} v\hat{a}r[\beta_{(u)}]) - rank(X)|$$

$$COVRATIO(\beta) = \frac{det_{ns}(v\hat{a}r[\beta_{(u)}])}{det_{ns}(v\hat{a}r[\beta])}$$

where det$_{ns}$(M) denotes the determinant of nonsingular part of matrix M. If the influence analysis updates the covariance parameters, MIXED procedure computes similar statistics for θ :

$$COVTRACE(\theta) = |trace(v\hat{a}r[\hat{\theta}]^{-1} v\hat{a}r[\theta_{(u)}]) - q|$$

$$COVRATIO(\theta) = \frac{det_{ns}(v\hat{a}r[\theta_{(u)}])}{det_{ns}(v\hat{a}r[\theta])}$$

where q denotes the rank of $var(\hat{\theta})$. The variance matrix that is used in the computation of COVTRACE and COVRATIO for covariance parameters is obtained from the inverse Hessian matrix.

Change In Precision of Estimates

The PRESS residual is the difference between the observed value and the predicted (marginal) mean, where the predicted value is obtained without the observations in equation. The equation is given by

$$\hat{\varepsilon}_i(u) = y_i - X_i^{'} \hat{\beta}_{(u)}$$

If we compute the influence of individual observations using PROC MIXED in SAS the procedure gives these PRESS residual. When removing sets of observations, the MIXED procedure computes the PRESS statistics. This statistic is the sum of the squared PRESS residuals in a deletion set

$$PRESS_{(u)} = \sum_{i \varepsilon u} \hat{\varepsilon}_i(u)$$

The PRESS of observations of fitted values can be measured by the DFFITS statistic. A DFFITS measures the change in predicted values due to removal of a single data point. When this change is standardized by the externally estimated standard error of the predicted value in the full data and we obtain the DFFITS statistic is obtain as:

$$DFFITS = (\hat{y}_i - \hat{y}_{i(u)})/ese(\hat{y}_i)$$

Table 3.3: The table below summarizes the available influences in LMMs (Source: West et al. 2007, p. 45)

Summary of influence Diagnostics for LMMs				
Group	Name	Parameters of interest	Formula	Description
Overall influence	Likelihood distance/ displacement	ψ	$LD_{(u)} = 2\{\ell(\hat{\psi}) - \ell(\hat{\psi}_{(u)})\}$	Change in ML log-likelihood for all data with ψ estimated for all data vs. reduced data
	Restricted likelihood distance/ displacement	ψ	$RLD_{(u)} = 2\{\ell_R(\hat{\psi}) - \ell_R(\hat{\psi}_{(u)})\}$	Change in REML log-likelihood for all data with ψ estimated for all data vs. reduced data
Change in parameter estimates	Cook's D	β	$D(\beta) = (\hat{\beta} - \hat{\beta}_u)' v\hat{a}r[\hat{\beta}]^{-1}(\hat{\beta} - \hat{\beta}_u)/rank(X)$	Scaled change in entire estimated β vector
		θ	$D(\beta) = (\hat{\theta} - \hat{\theta}_u)' v\hat{a}r[\hat{\theta}]^{-1}(\hat{\theta} - \hat{\theta}_u)$	Scaled change in entire estimated β vector
	Multivariate DFFITS statistic	β	$\text{MDFFITS}(\beta) = (\hat{\beta} - \hat{\beta}_{(u)})' v\hat{a}r[\hat{\beta_{(u)}}]^{-1}(\hat{\beta} - \hat{\beta}_{(u)})/rank(X)$	Scaled change in entire estimated β vector using externalized estimates of var($\hat{\beta}$)
		θ	$\text{MDFFITS}(\beta) = (\hat{\theta} - \hat{\theta}_u)' v\hat{a}r[\hat{\theta}]^{-1}(\hat{\theta} - \hat{\theta}_u)$	Scaled change in entire estimated β vector using externalized estimates of var($\hat{\theta}$)
Change in precision of parameter estimates	Trace of covariance matrix	β	$\text{COVTRACE}(\beta) = \lvert trace(v\hat{a}r[\hat{\beta}]^{-1} v\hat{a}r[\hat{\beta}_{(u)}]) - rank(X)\rvert$	Change in precision of estimated β vector, based on trace var($\hat{\beta}$)
		θ	$\text{COVTRACE}(\beta) = \lvert trace(v\hat{a}r[\hat{\theta}]^{-1} v\hat{a}r[\hat{\theta}_{(u)}]) - q\rvert$	Change in precision of estimated θ vector, based on trace of var($\hat{\theta}$)
	Covariance ratio	β	$\text{COVRATIO}(\beta) = \frac{det_{ns}(v\hat{a}r[\hat{\beta}_{(u)}])}{det_{ns}(v\hat{a}r[\hat{\beta}])}$	Change in precision of estimated β vector, based on determinant var($\hat{\beta}$)
		θ	$\text{COVRATIO}(\theta) = \frac{det_{ns}(v\hat{a}r[\hat{\theta}_{(u)}])}{det_{ns}(v\hat{a}r[\hat{\theta}])}$	Change in precision of estimated θ vector, based on determinant of var($\hat{\theta}$)
Effect on predicted value	Sum of square PRESS residual	N/A	$PRESS_{(u)} = \sum_{i\varepsilon u}(y_i - x_i'\hat{\beta}_{(u)})$	Sum of PRESS residual calculated by deleting observations in U

3.10 Application of the Linear Mixed Models

We are going to apply the method of linear mixed models to the Treatment of Lead-Exposed Children (TLC) data which is described in Chapter 2. The covariates used in the analysis are treatment and week. The application of the linear mixed models was done under PROC MIXED in SAS 9.3. SAS PROC MIXED is a very powerful procedure for a wide variety of statistical analysis, including repeated measures analysis of variance. PROC MIXED uses the maximum likelihood (ML) or restricted/residual maximum likelihood (REML) method. PROC MIXED defines random effects as truly random. The MIXED procedure fits a variety of mixed linear models to data and allows us to use these fitted models to make statistical inferences about the data. The mixed linear model provides us with the flexibility of modeling the variances and covariances as well as not only the means of our data. Once we fit a model of the data, we use the model to draw statistical inferences via both the fixed-effects and covariance parameters. PROC MIXED computes several different statistics suitable for generating hypothesis tests and confidence intervals. The validity of these statistics depends upon the estimates and variance-covariance model we select, so it is important to choose the model carefully. The output from PROC MIXED helps us assess the model and compare it with others to check the best fitting model. Using PROC MIXED, we fit a model for blood lead level measured in micro-grams/dL over time which includes subject-specific intercepts and slope as random effects and allows both the mean intercept and mean slope (fixed effects) to differ by group. The statistically mixed model can be stated as

$$y_{ij} = \beta_0 + \beta_1 group_i + \beta_2 time_{ij} + \beta_3 group_i * time_{ij} + u_{1i} + u_{2i} time_{ij} + \varepsilon_{ij} \tag{3.71}$$

where y_{ij} is the blood lead level for subject i at time j, $time_{ij}$ is the time of measurement of y_{ij} takes values $0,1,4,6$. $\beta_0, \beta_1, \beta_2, \beta_3$ are the fixed effects parameters: intercept, group main effect, time main effect and interaction between group and time. u_{1i} is the random effect of intercept for subject i and u_{2i} is the random effect of slope for subject i. We assume that $(u_{1i}, u_{2i})' \sim N(0, G)$ where G is the variance-covariance matrix with $Var(u_{1i}) = G_{11}$, $Var(u_{2i}) = G_{21}$, $Cov(u_{1i}, u_{2i}) = G_{12} = G_{21}$ elements. We can write the model in matrix notation as

$$y = X\beta + Zu + \varepsilon \tag{3.72}$$

where y in our case is the 4×1 vector of repeated measurements , β is the 2×1 vector of fixed effects and X is the associated 4×2 full column rank, u is the 2×1 vector of random effects, Z is the associated 4×2 design matrix and ε is the 4×1 vector of residual.

The commands to fit the model in SAS code for marginal model for linear mixed model

```
/* Marginal models for all structures under both methods ML and REML*/
proc mixed data=tlc method=reml covtest asycov;
class id group;
model y = group time group*time / s influence(iter=5 effect=id est);
repeated / type=un subject=id r rcorr;
run;
```

The CLASS statement names the classification variables to be used in the analysis. The MODEL statement is required and it specifies the response (dependent) variable versus the explanatory (independent) variables. The REPEATED statement in PROCMIXED is used to specify covariance structures for repeated measurements on subjects. The OPTION method in the PROC MIXED statement specifies the estimation method. The restricted maximum likelihood is obtained with the option method=REML and the maximum likelihood estimator is obtained with the option method=MLE. The statement TYPE specifies the covariance structure of R.

Table 3.4 below is from the fitted marginal model under both ML and REML estimates under different structures. The best model was the one with the smallest AIC from the selection criteria in both methods. From Table 3.4 below the best model can be found.

Table 3.4: Model selection criteria with different structures under ML and REML

Model	Cov Parameters	Maximum Likelihood (ML)			Restricted Maximum Likelihood (REML)		
Structure	No.	AIC	AICC	BIC	AIC	AICC	BIC
UN	10	2556.5	2557.6	2592.8	2551.0	2551.6	2577.0
Toep	4	2634.8	2635.1	2655.5	2628.0	2628.1	2638.3
CS	2	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
AR(1)	2	2667.4	2667.6	2682.9	2659.3	2659.3	2664.4
VC	1	2724.2	2724.4	2737.2	2717.4	2717.4	2720.0
Spatial Exp	2	2724.1	2724.4	2739.7	2716.2	2716.2	2721.4
Spatial Gau	2	2724.2	2724.4	2737.2	2717.4	2717.4	2720.0
Spatial Pow	2	2726.0	2726.2	2741.6	2719.3	2719.4	2724.5

The best fit model is the one which has the smallest AIC value for the selection criteria under both methods i.e ML and REML in the table. In Table 3.4 the best fitting model is one with unstructured (UN) structure with AIC = 2556.5 and AIC = 2551.0 under both ML and REML respectively.

Table 3.5: Estimates of covariance parameters under MLE

Cov Parm	Subject	Estimates	Std Error	Z Value	$Pr > \|t\|$
UN(1,1)	id	27.8491	5.1217	5.44	$< .0001$
UN(2,1)	id	6.3200	7.1322	0.89	0.3756
UN(2,2)	id	92.9266	18.6132	4.99	$< .0001$
UN(3,1)	id	11.5757	5.3765	2.15	0.0313
UN(3,2)	id	67.3710	15.6973	4.29	$< .0001$
UN(3,3)	id	67.6960	13.9068	4.87	$< .0001$
UN(4,1)	id	21.8049	5.1205	4.26	$< .0001$
UN(4,2)	id	29.5423	12.1716	2.43	0.0152
UN(4,3)	id	30.5275	9.2265	3.31	0.0009
UN(4,4)	id	58.0263	8.2492	7.03	$< .0001$

Table 3.5 shows the preceding table lists the 10 estimated covariance parameters in order. The parameter estimates are labeled according to their location in the block in covariance parameter column and all of these estimates are associated with individual (children) as a subject effects. The standard error column approximates standard errors of the covariance parameters obtained from the inverse Hessian matrix. The standard errors leads to approximate Wald Z-statistics which are compared with the standard normal distribution. The results of the tests indicate that all parameters are significantly different from 0 except that cov parm UN(2,1) with p-value = 0.3756 is not significantly different from zero.

Table 3.6: Estimates of Fixed Effects under MLE

Effect	Group	Estimates	Std Error	DF	t Value	$Pr > \|t\|$
Intercept		26.0321	0.6873	97	37.88	$< .0001$
Group	A	-1.8837	0.9769	97	-1.93	0.0568
Week		-0.3669	0.1207	97	-3.04	0.0030
Group × Week	A	-0.1830	0.1715	97	-1.07	0.2885

Table 3.6 shows the preceding table lists the solution vector for the fixed effects. The estimate of the placebo treatments' intercept is 26.03, while that for active is 26.03-1.88 = 24.15. Similarly, the estimate for the placebo treatments' slope is -0.37 while that for the active is -0.37-0.18= -0.55. Thus the placebo group starting point is larger than that for the active group treatment, but their blood lead level growth rate is about three times that of the placebo.

Table 3.7: Tests of Fixed Effects under LME

Effect	Num DF	Den DF	F Value	$Pr > F$
Group	1	97	3.72	0.0568
Week	1	97	28.58	$< .0001$
Group × Week	1	97	1.14	0.2885

Table 3.7 ”Tests of Fixed Effects” displays Type III tests for all fixed effects. In this case, the $Group \times Week$ test reveals difference between the slopes that is statistically not significant at the 5% level, the p-value (0.2934) is the same as the p-value in the ”$Group \times WeekA$” row in the ”Solution for Fixed Effects” table, and that F-statistic (1.12) is the square of the $t-$statistic (-1.06). Similar connections are evident among the other rows in these two tables. The **Week** test is the one for an overall blood lead level curve accounting for possible heterogeneous slopes and it is highly significant. Finally, the **Group** row tests the null hypothesis of a common intercept and this hypothesis cannot be rejected from these data since is not significantly different from zero and p-value $(0.0593) > 0.05$.

Table 3.8: Estimates of covariance parameters under REML

Cov Parm	Subject	Estimates	Std Error	Z Value	$Pr > \|t\|$
UN(1,1)	id	28.3333	5.2183	5.43	$< .0001$
UN(2,1)	id	6.8046	7.2304	0.94	0.3756
UN(2,2)	id	93.4266	18.7663	4.98	$< .0001$
UN(3,1)	id	12.0421	5.4776	2.20	0.0313
UN(3,2)	id	67.8974	15.8520	4.28	$< .0001$
UN(3,3)	id	68.3841	14.0668	4.86	$< .0001$
UN(4,1)	id	22.2573	5.2365	4.25	$< .0001$
UN(4,2)	id	30.0842	12.3221	2.44	0.0152
UN(4,3)	id	31.3216	9.3864	3.34	0.0009
UN(4,4)	id	58.9868	8.4417	6.99	$< .0001$

Table 3.8 shows a similar results interpretation to Table 3.5 except that the estimates values are much higher than estimates under ML estimates.

Table 3.9: Estimates of Fixed Effects under REML

Effect	Group	Estimates	Std Error	DF	t Value	$Pr > \|t\|$
Intercept		26.0322	0.6943	97	37.49	$< .0001$
Group	A	-1.8830	0.9869	97	37.49	0.0593
Week		-0.3669	0.1219	97	-3.01	0.0033
Group × Week	A	-0.1831	0.1733	97	-1.06	0.2934

Table 3.9 shows the preceding table lists the solution vector for the fixed effects. The estimate of the placebo treatments' intercept is 26.03, while that for active is 26.03-1.88 = 24.15. Similarly, the estimate for the placebo treatments' slope is -0.37 while that for the active is -0.37-0.18= -0.55. Thus the placebo group starting point is larger than that for the active group treatment, but their blood lead level rate is about three times of the placebo.

Table 3.10: Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	$Pr > F$
Group	1	97	3.64	0.0593
Week	1	97	28.01	$< .0001$
Group × Week	1	97	1.12	0.2934

Table 3.10 shows similar results interpretation to Table 3.7 except that the estimates values are approximately the same under ML and REML estimates.

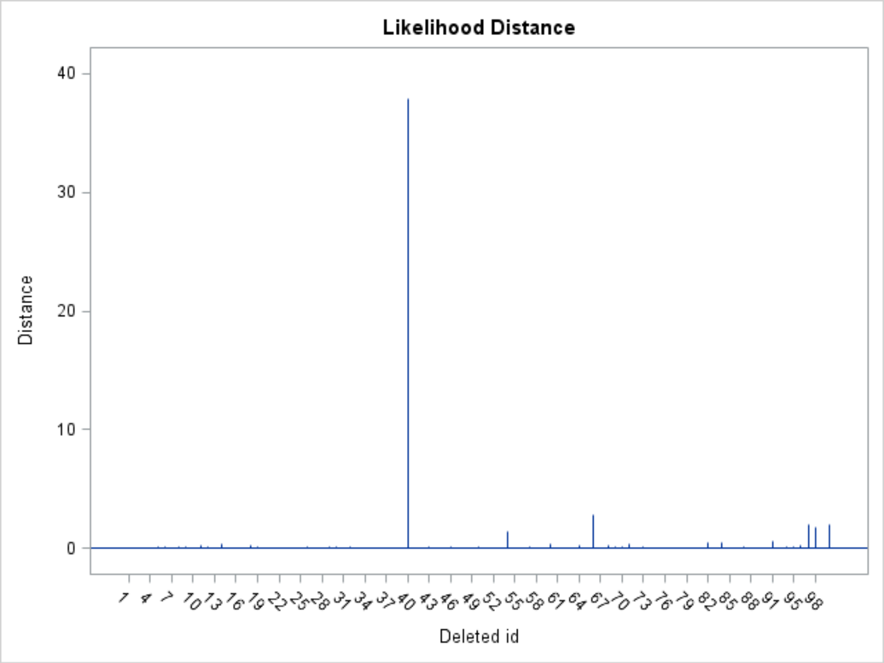

Figure 3.1: Restricted Likelihood Distance

As judged by the restricted likelihood distance, subject 40 clearly has most influence on the overall analysis followed by subject 67.

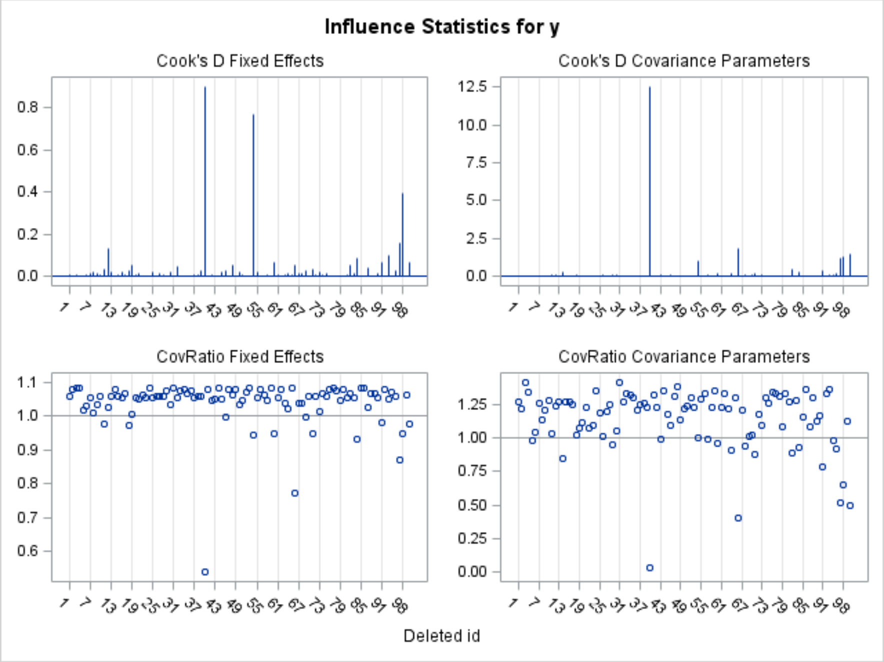

Figure 3.2: Influence Diagnostics

Figure 3.2 displays Cook's D and CovRatio statistics for the fixed effects and covariance parameters. The subject 40 has a considerable effect on the estimates of variances and covariances. This subject also affects the precision of the covariance parameter estimates more than any other subject; CovRatio is near 0. The observation who exerts the greatest influence on fixed effect is subject 40; this subject affects the variance-covariance matrix of the fixed effects more than other subjects and has small CovRatio.

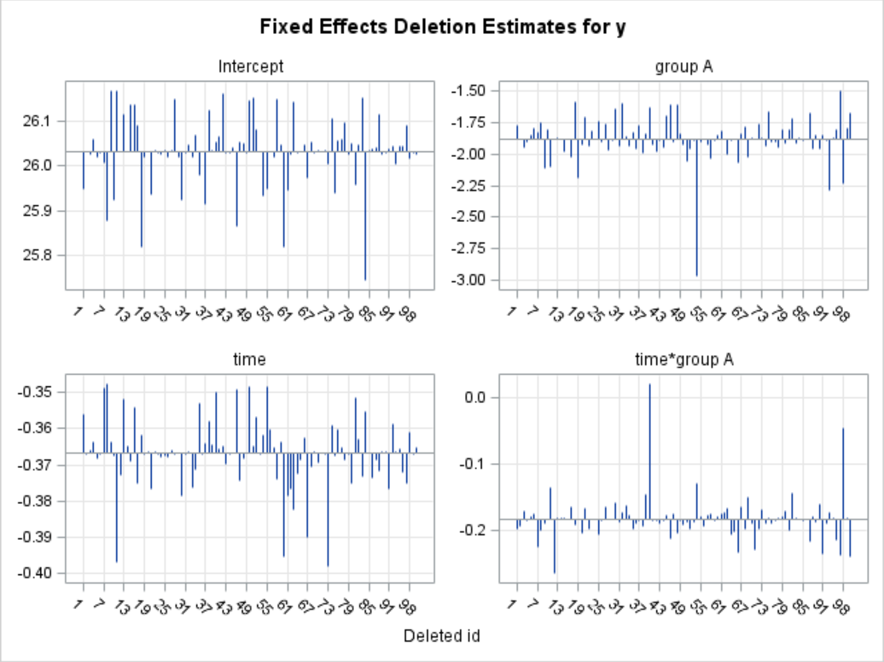

Figure 3.3: Fixed Effects Delete Estimates

Figure 3.3 shows the graphs on the left hand side of the panel represent the intercept and slope estimate for placebo; the graphs on the right hand side represent the difference in intercept and slope between active and placebo. The difference in these parameters between active and placebo is altered or improved by the removal of any child. Subject 40 changes fixed effects substantially or appreciably.

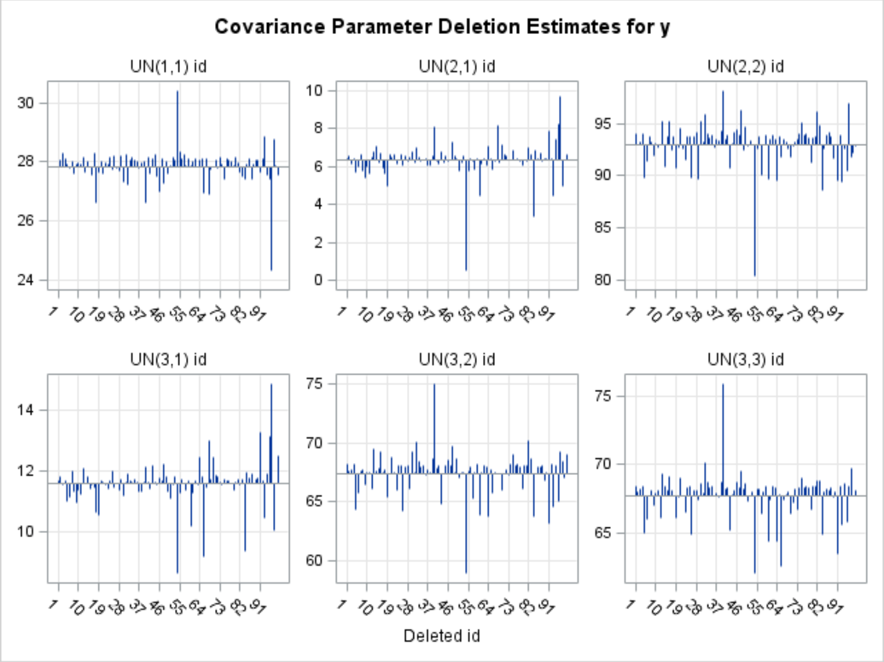

Figure 3.4: Covariance Parameter Delete Estimates

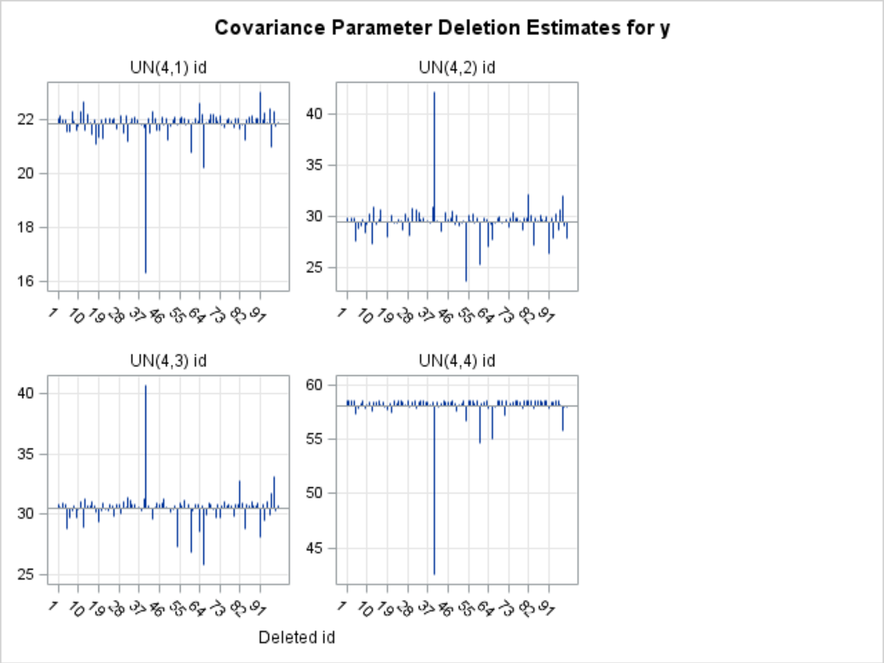

Figure 3.5: Covariance Parameter Delete Estimates

The covariance parameter deletion estimates in Output show several important features.

Subject 54 has great impact on the six covariance parameters. Removing this child from the analysis increases the variance of the random intercept and slope. The repeated measurements of child display an up and down behavior.

3.10.1 Application of the Random coefficient model

The commands to fit the model in SAS code for random intercept model are

```
/* To fit a random intercept model for all structures under both methods ML and REML*/
proc mixed data=tlc method=ml;
class id group;
model y = group time group*time / s;
random intercept / type=un sub=id g;
run;
```

The RANDOM statement defines the random effects constituting the *u* vector in the mixed model. The purpose of the RANDOM statement is to define the Z matrix of the mixed model. The statement TYPE specifies the covariance structure of G.

Table 3.11: Model selection criteria with different structures under ML and REML

Model	Cov Parameters	Maximum Likelihood (ML)			Restricted Maximum Likelihood (REML)		
Structure	No.	AIC	AICC	BIC	AIC	AICC	BIC
UN	2	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
Toep	2	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
CS	3	2681.6	2681.8	2697.1	2674.2	2674.2	2679.4
AR(1)	3	2681.6	2681.8	2697.1	2674.2	2674.2	2679.4
VC	3	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
Spatial Exp	2	2657.3	2657.6	2675.5	2650.0	2650.0	2657.7
Spatial Gau	3	2697.9	2698.2	2716.1	2690.5	2690.6	2698.3
Spatial Pow	3	2681.6	2681.8	2697.1	2679.4	2674.2	2674.2

The best fit model is the one which has the smallest AIC value for the selection criteria under both methods i.e ML and REML in the table. In Table 3.11 the best fitting model is one with toeplitz (toep), unstructured (UN) and variance component (VC) structure with AIC = 2556.5 and AIC = 2551.0 under both ML and REML respectively. In this case we will fit model with UN or toep because it gives us same results.

Table 3.12: Estimates of Parameter Estimates under REML

Cov Parm	Subject	Estimates	Standard error	Z value	$Pr > Z$
Variance	id	21.7622	4.3470	5.01	$< .0001$
Residual		33.7271	2.7677	12.19	$< .0001$

Covariance parameter estimates

Table 3.12 shows the covariance parameter estimates. This can be thought of as the correlation between the observations within the group. σ^2_{week} represents the variability between children: $\sigma^2_{week} = 21.7622$ and is highly significant different from zero.
$\sigma^2_{residual}$ represents the variability within children: $\sigma^2_{residual} = 33.7271$ and is also highly significant.

In our case, the observations from different children are independent and observations from the same child are correlated. The total variability, correcting for group differences is decomposed as within-cluster variability and between cluster variability is equal to

$$\sigma^2 = \sigma^2_{week} + \sigma^2_{residual} = 21.7622 + 33.7271 = 55.4895$$

The overall correlation between repeated measurements is given by

$$ICC = \frac{\sigma^2_{week}}{\sigma^2_{week}+\sigma^2_{residual}} = \frac{21.7622}{21.7622+33.7271} = 0.392$$

The between-child variability accounts for 39% of all variability. The week factor explains 39% of the total variability after the correction for group, or indicating that 39% of the variation in the data is accounted for by allowing the intercept and the slope to vary across individuals.

Table 3.13: Estimates of Fixed Effects under LME

Effect	Group	Estimates	Std Error	DF	t Value	$Pr > \|t\|$
Intercept		25.6854	0.9100	97	28.23	$< .0001$
Group	A	-5.3184	1.2935	295	-4.11	$< .0001$
Week		-0.3721	0.1722	295	-2.15	0.0315
Group × Week	A	-0.07391	0.2448	295	-0.30	0.7629

Table 3.13 show the preceding table lists the solution vector for the fixed effects. The estimate of the placebo treatments' intercept is 26.03, while that for active is 25.69-5.32 = 20.37.

Similarly, the estimate for the placebo treatments' slope is -0.37 while that for the active is -0.37-0.07= -0.44. Thus the placebo group starting point is larger than that for the active group treatment, but their blood lead level rate is approximately the same as that of the placebo.

Table 3.14: Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	$Pr > F$
Group	1	295	16.91	$< .0001$
Week	1	295	11.17	0.0009
Group × Week	1	295	0.09	0.7629

Table 3.14 shows the result of "Type 3 Tests of Fixed Effects" for all fixed effects. There is a strong evidence of a relationship between the baseline covariates group placebo and the subsequent responses. The p-value is $< .0001$ which is highly significant. The week effect is highly significant and concludes that there is evidence since the p-value is 0.0009. There is no evidence of $Group \times week$ interaction and the p-value is 0.7629; that shows that the effects are not significantly different from zero. Their changes in the response variable treatment over time are the same for all drug treatments.

Table 3.15: Estimates of Covariance Parameter under REML

Cov Parm	Subject	Estimates	Standard error	Z value	$Pr > Z$
Variance	id	22.3276	4.4799	4.98	$< .0001$
Residual		33.9557	2.7959	12.14	$< .0001$

Covariance parameter estimates

Table 3.15 presents the covariance parameter estimates. These are estimates for random effects portion of the model. In this case, we find that the estimated value of $\sigma^2_{week} = 22.3276$ the week variance component that represents the variability between children and $\sigma^2_{residual}$ represents the variability within children: $\sigma^2_{residual} = 33.9557$. The variance component for children is highly significant between children variation and the residual variance is also significant at 5% level of significance. The total variation is equal to

$$\sigma^2 = \sigma^2_{week} + \sigma^2_{residual} = 22.3276 + 33.9557 = 56.2838$$

The observations from the same child are correlated; the overall correlation between repeated measurements is given by

$$ICC = \frac{\sigma^2_{week}}{\sigma^2_{week}+\sigma^2_{residual}} = \frac{22.3276}{22.3276+33.9557} = 0.397 \approx 40\%$$

The between-child variability accounts for 40% of all variability. The week factor explains 40% of the total variability after the correction for group.

Table 3.16: Estimates of Fixed Effects under REML

Effect	Group	Estimates	Std Error	DF	t Value	$Pr > \|t\|$
Intercept		25.6854	0.9176	97	27.99	$< .0001$
Group	A	-5.3184	1.3044	295	-4.08	$< .0001$
Week		-0.3721	0.1728	295	-2.15	0.0321
Group × Week	A	-0.07391	0.2456	295	-0.30	0.7637

Table 3.16 show the preceding table lists the solution vector for the fixed effects. The estimate of the placebo treatments' intercept is 25.69, while that for active is 25.69-5.32 = 20.37. Similarly, the estimate for the placebo treatments' slope is -0.37 while that for the active is -0.37-0.07= -0.44. Thus the placebo group starting point is larger than that for the active group treatment, but their blood lead level rate is about the same as that of the placebo.

Table 3.17: Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	$Pr > F$
Group	1	295	16.63	$< .0001$
Week	1	295	11.10	0.0010
Group × Week	1	295	0.09	0.7637

The results in Table 3.17 show that there is strong evidence of a relationship between the baseline covariates group placebo and the subsequent responses. The p-value is $< .0001$ which is highly significant. The week effect is highly significant and concludes that there is evidence since the p-value is 0.0010. There is no evidence of $Group \times week$ interaction and the p-value is 0.2934 that shows that the effect is not significantly different from zero. Their changes in the response variable treatment over time are the same for all drug treatments.

3.10.2 Application of the Random Intercept and Slope Model

The commands to fit the model in SAS code for random intercept and random slope model are */* To fit a random intercept and slope model for all structures under both methods ML and REML*/*
proc mixed data=tlc method=ml;

```
class id group;
model y = group time group*time / s;
random intercept time / type=un sub=id g gc v vcorr;
run;
```

Table 3.18: Model selection criteria with different structures under ML and REML

Model	Cov Parameters	Maximum Likelihood (ML)			Restricted Maximum Likelihood (REML)		
Structure	No.	AIC	AICC	BIC	AIC	AICC	BIC
UN	2	2643.4	2643.7	2661.5	2636.3	2636.3	2644.1
Toep	2	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
CS	3	2681.6	2681.8	2697.1	2674.2	2674.2	2679.4
AR(1)	3	2681.6	2681.8	2697.1	2674.2	2674.2	2679.4
VC	3	2655.3	2655.6	2670.9	2648.0	2648.0	2653.2
Spatial Exp	3	2657.3	2657.6	2675.5	2650.0	2650.0	2657.7
Spatial Gau	3	2697.9	2698.2	2716.1	2690.5	2690.6	2698.3
Spatial Pow	3	2681.6	2681.8	2697.1	2679.4	2674.2	2674.2

Table 3.19: Estimated Correlation Matrix

Row	Col1	Col2	Col3	Col4
1	1.0000	0.1323	0.2736	0.5444
2	0.1323	1.0000	0.8495	0.4053
3	0.2736	0.8495	1.0000	0.4932
4	0.5444	0.4053	0.4932	1.0000

Table 3.20: Estimates of Fixed Effects

Effect	Group	Estimates	Std Error	DF	t Value	$Pr > \|t\|$
Intercept		25.6854	0.7375	97	34.83	$< .0001$
Group	A	-5.3184	1.0482	198	-5.07	$< .0001$
Week		-0.3721	0.1728	97	-2.15	0.0337
Group × Week	A	-0.07391	0.2456	198	-0.30	0.7638

Table 3.20 shows the preceding table lists the solution vector for the fixed effects. The estimate of the placebo treatments' intercept is 25.69, while that for active is 25.69-5.32 = 20.37. Similarly, the estimate for the placebo treatments' slope is -0.37 while that for the active is -0.37-0.07= -0.44. Thus the placebo group starting point is larger than that for the active group treatment, but their blood lead level rate is about the same as that of the placebo.

Table 3.21: Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	$Pr > F$
Group	1	198	25.74	$< .0001$
Week	1	97	11.10	0.0012
Group × Week	1	198	0.09	0.7638

The results from the Type III F statistic corresponding to Group × Week in output indicates there is no evidence to believe that the slopes are equal (p-value = 0.7638); therefore, we assume that a common slope model is sufficient to describe the data. In the test of null hypothesis, we found that there is no evidence that the slopes are equal to zero (p-value = 0.7638). The slope effects are not significantly different from zero.

3.11 Summary

The proposed methodology copes with the difficulty of the analysis of longitudinal data. Thereby, we dealt with theoretical and computational aspects which are substantially challenging in the linear regression setting. In the analysis of linear mixed effect models, we assume the continuous response because LMM is not suitable for modelling a binary response. We fitted LMM using SAS procedure MIXED and we used model selection criteria to choose best model with the best covariance structure. We fitted different models such as linear mixed model and as well for random effects structure (i.e. random intercept, and random intercept and slope model). After comparing models, we choose linear mixed model with unstructured covariance structure as our best fitted model since it has a small AIC which estimates the quality of each model relative to each of the other models. We found that linear mixed effects models are flexible methods for modelling continuous longitudinal data and the major advantage of linear mixed model is that it accommodates the complexities of typical longitudinal data sets and can provide information on individual approach as well as population approach and can handle missing data. If the longitudinal response is discrete, then we have more than one way to extend generalized linear models to longitudinal setting; this is discussed in the following chapter.

Chapter 4

Models for Discrete Correlated Longitudinal Data

The data set is a longitudinal study measuring the prevalence of the Respiratory Syncytial Virus (RSV) in children. The study of the data is obtained when a response is measured repeatedly on a set of units. The data set is a part of the study carried out by the Kenyan Medical Research Institute and the Welcome Trust in Kilifi, Kenya. The data set presents a form of completeness which has to be properly accounted for in order to carry out an appropriate analysis of this data which leads to correct conclusions. The data set has 320 individuals (children) that were recruited in the study and all were measured and recorded. For each child, the following information was collected with the variable names in brackets. These were: the number of visits (visit); the time between visits (dt); the sampling types, active sampling if the field worker went to visit the child and passive if the child was brought to the clinic to be sampled (actpass); the age in months of the child at the visit (age); the response variable of whether the child is infected or uninfected (rsv); and the prevalence of the virus in the blood (prev) which is a continuous variable. The aim of the study was to understand the dynamics of the disease in children. The paper by Mwambi et al. (2011) gave an approach to estimate the force of infection for the disease. However this is not the focus of the current work.

Table 4.1: Table of variables

Variable	Levels and coding
id	$1, \ldots, 320$
rsv	1=uninfected, 2=infected
dt	$0, \ldots, 181$
visit	$1, \ldots, 144$
actpass	1=active sampling, 2=passive sampling
age	$0, \ldots, 12$ months
prev	a continuous variable ranging from min=0.0 to max=0.047516

Table 4.1 shows that the measured status of a rsv (infected or uninfected) in an individual is a binary response outcome of non-Gaussian variable. The generalized linear model in a longitudinal setting seems the best option to deal with such a data set. In this work we focus on the data derived from repeated measurements from the same individual. Thus the assumption of independent measurements cannot be used because observations within an individual are correlated or dependent. The chapter discusses the theory of generalized linear models (GLMs) to accommodate responses that follow non-Gaussian data distributions by first introducing components of a GLM and the exponential family. We also consider logistic regression which is a special case of GLM. Since the standard of GLM assumes that the observations are uncorrelated then the extensions have been developed to allow for correlation between observations, as occurs in longitudinal studies and clustered designs. The case of correlated non-Gaussian longitudinal data is then discussed using generalized estimating equations (GEEs) as in Liang and Zeger (1986). GEEs applies to marginal models for longitudinal binary data. An important aspect of this approach is the specification of a working correlation structure. The working correlation structure represents the correlation believed to be present among responses within subjects and as such is incorporated into the random component of the model (Lalonde et al., 2013). The application of this model is done under different correlation structures to RSV data using the SAS software and the results are discussed.

4.1 Generalized Linear Models (GLMs)

Generalized linear models (GLMs) are an extension of classical linear models to model non-normal response variables. The generalized linear model (GLM) was first introduced by Nelder and Wedderburn (1972). Generalized linear models extend classical regression analysis for independent normally distributed random variables with constant variance to other types of the response variables. These models are suitable when the response variable is non-normal distributed along with exploratory variables that are categorical (Davis, 2002). The application of generalized linear models was further extended by introducing quasilikelihood by Wedderburn (1974). The GLM approach is most often used in common analysis because provides a theoretical framework for many commonly used statistical models and simplifies the tool of those different models in statistical software, since essentially the same algorithm can be used for estimation, inference and assessing the model sufficiently for all generalized linear models (Venables and Ripley, 2002). The parameters are estimated by using the maximum likelihood methods (McCullagh and Nelder, 1989). As stated previously, GLMs extend the range of application of linear statistical models by accommodating response variables with non-normal distributions such as the Poisson distribution, the Binomial distribution, the Bernoulli distribution, the Gamma distribution and other distributions. If the response variable is assumed to be nonlinear then the link function which is one of the components of the GLM is used as the response variable.

4.2 The Exponential Family

In this section we describe the generalized linear model as formulated by Nelder and Webberburn (1972) and discuss estimation of the parameters and tests of hypotheses. We assume that the observations come from a distribution in the exponential family with probability density function

$$f(y_i) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \tag{4.1}$$

where θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. According to other researchers and authors, the function $a_i(\phi)$ has the form

$$a_i(\phi) = \frac{\phi}{w} \tag{4.2}$$

where w is a known prior weight, usually 1. The parameters θ_i and ϕ are essentially location and scale parameters. The function $b(\theta_i)$ is called the cumulant function which is helpful in generating the mean and variance. We show below how $b(\theta_i)$ is used to find mean and variance.

The mean and variance of Y can be derived by using this equation

$$\int f(y;\theta,\phi)dy = 1 \tag{4.3}$$

By taking the first and second derivative with respect to θ from both sides of above equation, This leads us to the two equations of the form

$$\int (y - b'(\theta_i))f(y;\theta,\phi)dy = 0 \tag{4.4}$$

and

$$\int \left[\frac{1}{\phi}(y - b'(\theta_i))^2 - b''(\theta_i)\right] f(y;\theta,\phi)dy = 0 \tag{4.5}$$

By solving this equation for $\mu = E(y)$ and $Var(y) = E[(y-\mu)^2]$ respectively, we get the solutions

$$E(y) = b'(\theta_i) \text{ and } Var(y) = \phi V(\mu), \text{ where } V(\mu) = b''(\theta_i)$$

We note that in general mean and variance are dependent since

$$\begin{aligned} Var(y) &= \phi b''[b'^{-1}(\mu_i)] \\ &= \phi V(\mu) \end{aligned} \tag{4.6}$$

The function $V(\mu)$ is called the variance function. The function $b'^{-1}(.)$ which express θ as a function of μ is the link function and $b'(.)$ is the inverse link function. There are several distributions which belong to this structure and for classification purposes we briefly relate the above formulation to the Normal, Poisson, Bernoulli and Binomial distributions which all fall under the exponential family of distributions.

For example, suppose that y is normally distributed with mean μ and variance σ^2. Then distribution is given by

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{(2\Pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &= \exp\left\{\log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\right)\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\right\} \\ &= \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right\} \end{aligned} \tag{4.7}$$

where $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$ and $C(y,\phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$

Therefore, the mean and variance is given by

$$E(y) = b'(\theta) = \mu$$

and

$$Var(y) = \phi Var(\mu) = \sigma^2$$

which is independent on μ. The variance function is $V(\mu) = 1$, and the dispersion parameter is $\phi = \sigma^2$.

For a Poisson distribution with mean μ. Then distribution is given by

$$\begin{aligned} f(y,\mu) &= \frac{\exp(-\mu)\mu^y}{y!} \\ &= \exp\{y\log\mu - \mu - \log y!\} \end{aligned} \tag{4.8}$$

where $\theta = \log\mu$, $b(\theta) = \exp(\theta)$, $\phi = 1$ and $C(y,\phi) = \log y!$
The mean and variance is given by

$$E(y) = b'(\theta) = \mu$$

and

$$Var(y) = \phi Var(\mu) = \mu$$

which depends on μ. In this case, the variance function is $V(\mu) = \mu$ and the dispersion parameter is $\phi = 1$.

For Bernoulli distribution with mean π. Then distribution is given by

$$\begin{aligned} f(y,\mu) &= \pi^y(1-\pi)^{1-y} \\ &= \exp\{y\log\pi + (1-y)\log(1-\pi)\} \\ &= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} \end{aligned} \tag{4.9}$$

which is also in the form of an exponential family where $\theta = \log\frac{\pi}{(1-\pi)}$, $\phi = 1$, and
$b(\theta) = -\log(1-\pi) = \log(1+\exp(\theta))$, since $\pi = \frac{\exp(\theta)}{(1+\exp(\theta))}$.
Then the mean and variance is given by:

$$E(y) = b'(\theta) = \pi$$

and

$$Var(y) = \phi Var(\mu) = \pi(1-\pi)$$

which is dependent on μ. In this case, the variance function is $V(\mu) = \pi(1-\pi)$ and the dispersion parameter is $\phi = 1$.

For binomial distribution with parameters n and π (i.e $y \sim Bin(n,\pi)$). In this case,

$$\begin{aligned} f(y;\mu,\sigma^2) &= \binom{n}{y}\pi^y(1-\pi)^{n-y} \\ &= \exp\left\{\log\left(\binom{n}{y}\pi^y(1-\pi)^{n-y}\right)\right\} \\ &= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right)+n\log(1-\pi)+\log\binom{n}{y}\right\} \end{aligned} \tag{4.10}$$

$\theta = \log(\frac{\pi}{1-\pi})$, $\phi = 1$ and $b(\theta) = n\log(1+\exp(\theta))$. Therefore

$$E(Y) = b'(\theta) = n\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right) = n\pi$$

and

$$Var(Y) = \phi b''(\theta) = n\left(\frac{\exp(\theta)}{(1+\exp(\theta))^2}\right) = n\pi(1-\pi)$$

In this case, the variance function is $V(\mu) = n\pi(1-\pi)$ and the dispersion parameter is $\phi = 1$.

4.2.1 Components of a GLM

The generalized linear regression model is characterized by the following features:

- A **random component:** this component identifies the response, y_i and assumes a distribution that follows the exponential family. It is given by:

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\} \tag{4.11}$$

- **Systematic component:** this specifies the explanatory or predictor variables. The covariates x_i is combined linearly with the coefficients to form the linear predictor

$$\eta = X\beta \tag{4.12}$$

  where the i^{th} row of X is given by $x_i = (1, x_{i1}, \ldots, x_{ip})^{'}$ with x_{ij}, $i = 1\cdots n$ equal to the value of j^{th} explanatory variable or predictor $j = 1\cdots p$ and the $\beta = (\beta_0, \ldots, \beta_1, \beta_p)^{'}$ is a regression coefficient.

- **Link component:** this specifies the relationship between the mean of the random and systematic components: the linear predictor $X_i\beta = \eta_i$ is a function of the mean parameter μ_i via a link function, $g(\mu_i)$.

$$\begin{aligned}\eta_i &= g(\mu_i) \\ &= x_i^{'}\beta\end{aligned} \tag{4.13}$$

According to Davis (2002) and McCulloch and Neuhaus (2001), the $g(\mu_i)$ must be monotonic and a differentiable function such that

$$\eta_i = g(\mu_i)$$

Thus

$$g(\mu_i) = \sum_j \beta_j x_{ij} \ \ , \ i = 1, \ldots, N$$

relating the linear predictor to the mean response follows

$$\mu = g^{-1}(\eta_i) = E(y)$$

We model a function of the mean as a combination of linear predictors. This function g(.) is monotone which means as the systematic part gets larger, μ gets larger too and again when the systematic part gets smaller, μ gets smaller too. The relationship between E(y) and the systematic part can be nonlinear. Table 4.2 shows the summary of canonical link.

Table 4.2: Summary of canonical link

Distribution	Natural Parameter	Canonical link
Normal	μ	Identity
Poisson	$\log(\mu)$	Log
Bernoulli	$\log(\frac{\mu}{1-\mu})$	Logit
Gamma	$\frac{1}{\mu}$	Inverse
Binomial	$\log(\frac{\mu}{1-\mu})$	Logit

We demonstrate examples of the GLM family now. Example 4.2.1.1: Let Y be normal distributed with mean μ. Then the density function of the normal distribution is

$$f(y;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

The above density function follows the exponential family and it can be written as

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &= \exp\left\{\log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\right)\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\right\} \\ &= \exp\left\{\frac{y\mu-\frac{\mu^2}{2}}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}\right\} \end{aligned} \tag{4.14}$$

where $\theta=\mu$, $b(\theta)=\frac{\theta^2}{2}$, $\phi=\sigma^2$ and $C(y,\phi)=-\frac{1}{2}\log(2\pi\sigma^2)-\frac{y^2}{2\sigma^2}$. The canonical link is the density.

Example 4.2.1.2: Let Y be Poisson distributed with mean μ. Then the density function of the Poisson distribution is

$$\begin{aligned} f(y,\mu) &= \frac{\exp(-\mu)\mu^y}{y!} \\ &= \exp\{y\log\mu-\mu-\log y!\} \end{aligned} \tag{4.15}$$

where $\theta=\log(\mu)$, $b(\theta)=\exp(\theta)$, $\phi=1$ and $C(y,\phi)=\log y!$
Then the canonical link is log link.

Example 4.2.1.3: Let Y be Bernoulli distributed with mean π. Then the density function of the Bernoulli distribution is

$$\begin{aligned} f(y,\mu) &= \pi^y(1-\pi)^{1-y} \\ &= \exp\{y\log\pi+(1-y)\log(1-\pi)\} \\ &= \exp\left\{y\log\frac{\pi}{(1-\pi)}+\log(1-\pi)\right\} \end{aligned} \tag{4.16}$$

which is also in the form of an exponential family where $\theta=\log(\frac{\pi}{1-\pi})$, $\phi=1$, $C(y,\phi)=0$ and $b(\theta)=-\log(1-\pi)=\log(1+\exp(\theta))$, since $\pi=\frac{\exp(\theta)}{1+\exp(\theta)}$, and the canonical link is logit link.

4.2.2 Estimation of the Model Parameters

The concept of generalized linear models (GLMs) unifies different approaches to explaining variation in data in terms of a linear combination of covariates (Agresti, 2002). The GLM model which consists of a single response variable and the predictor variable is a member of the exponential family distribution. Generalized linear modeling transforms the relationship between the linear predictor and the mean response, such that a nonlinear relationship can be modeled as a linear. This admits a model specification allowing for continuous or discrete responses and allows a description of the variance as a function of the mean response. The GLM family members are linearized by mean of a link function and are fitted using ML technique with the help of iterative algorithms. A single algorithm can be used to estimate the parameters of an exponential family GLM using maximum likelihood. The likelihood $L(\beta,\phi)$ is given by

$$L(\beta,\phi)=\prod_{i=1}^{N} f(y_i;\beta,\phi)=\prod_{i=1}^{N}\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi}+C(y_i,\phi)\right\} \tag{4.17}$$

The estimation of the parameters in β is done by maximizing the log-likelihood defined as

$$\ell(\beta,\phi)=\sum_{i=1}^{N}\log f(y_i;\beta,\phi)=\sum_{i=1}^{N}\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi}+C(y_i,\phi)\right\} \tag{4.18}$$

which assumes independent exponential family responses y_i. To find the MLE of β_j, we differentiate $\ell(\beta,\phi)$ with respect to β_j which is given us

$$\frac{\partial\ell}{\partial\beta_j}=\left(\frac{\partial\ell}{\partial\theta}\right)\left(\frac{\partial\theta}{\partial\mu}\right)\left(\frac{\partial\mu}{\partial\eta}\right)\left(\frac{\partial\mu}{\partial\beta_j}\right) \tag{4.19}$$

The first factor in equation (4.21) is

$$\frac{\partial\ell}{\partial\theta}=\frac{y-b'(\theta)}{\phi},\ \text{ where }\ \mu=b'(\theta), \tag{4.20}$$

The second factor is

$$\frac{\partial\theta}{\partial\mu}=\frac{1}{b''(\theta)},\ \text{ where }\ V(\mu)=b''(\theta) \tag{4.21}$$

The third factor $\frac{\partial\mu}{\partial\eta}$ will depend on the link function, where $\eta=X^{'}\beta$. The last factor is

$$\frac{\partial\mu}{\partial\beta_j}=x_{ij} \tag{4.22}$$

where x_{ij} is the j^{th} element of the covariate vector $x_i=x$ for the i^{th} observation. Then by substituting equation (4.20, 4.21 and 4.22) into equation 4.19 and equating to zero, we have

$$\frac{\partial\ell}{\partial\beta}=\sum_{i=1}^{N}\frac{(y_i-\mu_i)}{V(\mu)}\left(\frac{\mu_i}{\partial\beta}\right)x_{ij}=0 \tag{4.23}$$

Then the MLE of the parameter vector β is obtained based on solving the estimating equations above. The estimation of β depends on the density function only through the mean and variance function $V(\mu_i)$. The score equations can be solved by using the iterative algorithm such as Newton-Raphson, Fisher scoring and re-weighted least square (RWLS). The ML estimation for β is carried out via Newton-Raphson,

$$\beta^{(t+1)} = \beta^{(t)} + \left(\ell''(\beta^{(t)})\right)^{-1} \ell'\left(\beta^{(t)}\right), \tag{4.24}$$

where ℓ is the log-likelihood function for the entire sample $y_i, \ldots, y_N$. We let ℓ, ℓ' and ℓ'' denote the contribution of the observations y_i to the log-likelihood and its derivatives. The Fisher scoring iterative equation is given by

$$\beta^{(t+1)} = \beta^{(t)} + \left[-E(\ell''(\beta^{(t)}))\right]^{-1} \ell'\left(\beta^{(t)}\right), \tag{4.25}$$

where the expected Hessian matrix becomes

$$-E(\ell''(\beta^{(t)})) = H^{(t)} = X^{'}WX \text{ and } W = Diag\left\{Var(y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2\right\}^{-1}$$

An iterative of Fisher scoring is then

$$\beta^{(t+1)} = \beta^{(t)} + [X^{'}WX]^{-1}X^{'}A(y-\mu) \tag{4.26}$$

where W is a diagonal matrix with main diagonal elements and $A = W\left(\frac{\partial \eta}{\partial \mu}\right)$, then we note that A and W are related. Note that $\frac{\partial \eta}{\partial \mu} = Diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$. This lead us to Reweighted least square iterative equation which is given by

$$\beta^{(t+1)} = [X^{'}WX]^{-1}X^{'}Wz, \tag{4.27}$$

where

$$\begin{aligned} z &= \eta + \left(\frac{\partial \eta}{\partial \mu}\right)(y-\mu) \\ &= (z_1, \ldots, z_N)^{'}, \end{aligned} \tag{4.28}$$

where $z_i = \eta_i + \left(\frac{\partial \eta_i}{\partial \mu_i}\right)(y_i - \mu_i)$ and is called the adjusted dependent variate or we can call it a linearized form of link function g evaluated at y.

Fisher scoring method is similar to the Newton-Raphson method but the difference is that Fisher scoring uses the expected value of matrix called expected information while the Newton-Raphson method uses the matrix itself or the observed information.

4.2.3 Simplication for canonical links

According to Agresti (2002), by using results of the canonical link of the likelihood equations with the

$$\eta_i = \sum \beta_j x_{ij} \tag{4.29}$$

for this model

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial b'(\theta)}{\partial \theta_i} = b''(\theta) \tag{4.30}$$

Then recall that $b''(\theta)$ is the variance function

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi} \frac{1}{V_i} V_i \, x_{ij} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi} \tag{4.31}$$

with the canonical link the second derivatives of the log-likelihood have components

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = -\frac{x_{ij}}{\phi} \left(\frac{\partial \mu_i}{\partial \beta_k} \right) \tag{4.32}$$

This does not depend on the observation y_i for $i = 1 \cdots n$, so

$$\frac{\partial^2 \ell}{\partial \beta^2} = E \left(\frac{\partial^2 \ell}{\partial \beta^2} \right) \tag{4.33}$$

Therefore under the canonical link $H = -j$ and the Newton-Raphson and Fisher scoring algorithms are identical since ϕ is constant for all observations in the likelihood estimating equation are

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij}, \quad \text{for} \quad j = 1, \ldots, p \tag{4.34}$$

These equations equate the sufficient statistics for the model parameters to their expected values (Agresti, 2002).

4.3 Inference of Parameter Estimates

Our primary interest is to test general hypothesis about the vector of parameters β

$$H_o : L\beta = 0 \ \text{ vs } \ H_a : L\beta \neq 0$$

since $\hat{\beta}$ is the MLE of β, it follows that $L\hat{\beta}$ is the MLE of $L\beta$. Therefore

$$L\hat{\beta} \sim N(L\beta, \ LVar(\beta)L')$$

where L is a known constants of dimension say $m \times p$. The hypotheses on single or groups of parameters can be tested in different ways in GLMs. There are three commonly used statistics for inference which is such as Wald test, Likelihood ratio test and Score test.

4.3.1 Wald Test

The Wald test statistic is commonly used to test the significance about the regression coefficients for each independent variable. The test hypothesis is given by

$$H_o : L\beta = 0 \text{ vs } H_a : L\beta \neq 0$$

and the Wald test statistics is given by

$$W = (L\hat{\beta} - L\beta)^{'}[LVar(\hat{\beta})L']^{-1}(L\hat{\beta} - L\beta)$$

and under H_o is asymptotically distributed as χ^2 with $d.f$ equal to rank(L). Note that the

$$Var(\hat{\beta}) = (\textstyle\sum_{i=1}^{N} X_i^{'} W_i X_i)^{-1}$$

where $W_i = diag(w_i)$ and $w_i = \left\{var(\mu_i)[g'(\mu_i)]^2\right\}^{-1}$

4.3.2 Likelihood Ratio Test

The likelihood ratio test is a widely used procedure for testing hypothesis involving nested models. This is a test for two nested models. There is a full fitted model and a reduced model that omits some variables. It reject the null hypothesis when the maximum likelihood under null hypothesis is significantly smaller than the likelihood under alternative hypothesis.

$$LRT = -2\ln\left[\frac{L(reduced\ model)}{L(full\ model)}\right] \sim \chi^2_{f-r} \tag{4.35}$$

The full model has f variables and the subset has R variables and so the test value is compared against a χ^2 with $f - r$ degrees of freedom.

4.3.3 Score Test

According to Liang (1999) we assume that the $Y_i^{'}$s are independent and represent a sample from the population, the likelihood function for β and φ is simply proportional to

$$\begin{aligned} L(\beta,\phi) &\propto \prod_{i=1}^{N} f(y_i;\beta,\phi) \\ &= \prod_{i=1}^{N} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + C(y_i,\phi)\right\} \end{aligned} \tag{4.36}$$

where $\eta_i = g(\mu_i) = X_i^{'}\beta$

The log-likelihood function is

$$\ell(\beta,\phi) = \sum_{i=1}^{N} \log f(y_i;\beta,\phi) = \sum_{i=1}^{N} \left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + C(y_i,\phi)\right\} \tag{4.37}$$

The score function is given by

$$U = \frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi}, \text{ where } \mu = b'(\theta_i),$$

and the score vector is given by

$$\begin{aligned} U &= \frac{\partial \ell}{\partial \theta} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi} \frac{\partial \theta_i}{\partial \beta} \\ &= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi} \frac{G'(\eta_i)}{V(\mu_i)} x_i \end{aligned} \tag{4.38}$$

and the information matrix is given by

$$\begin{aligned} I &= Var(U) \\ &= E(UU') \\ &= \sum_{i=1}^{n} \frac{G'(\eta_i)^2}{V(\mu_i)} x_i x_i^{'} \end{aligned} \tag{4.39}$$

Then the score vector is distributed as

$$U \sim MVN_p(0, I)$$

so that

$$Q = UI^{-1}U' \sim \chi^2(p)$$

4.4 Goodness of Fit in GLM

4.4.1 Deviance

Here we want to estimate Y by $\hat{\mu}$ and we expect that in n data points we can estimates n parameters. The deviance function is a very useful method for comparing the two models when one model has parameters that are a subset of another. The deviance is equal to twice the difference between log-likelihood for reduced and full models or fitted and saturated models. The theoretical definition of deviance is given by

$$D = 2\{\ell(y, y) - \ell(\hat{\mu}, y)\} \tag{4.40}$$

This has an asymptotical χ^2 distribution with degrees of freedom $N - p$, where p is the number of parameters in the reduced models. In the hypothesis testing, a test using the deviance is equivalent to a likelihood ratio test. Let us find out the deviance of the difference distributional equations. Let $\hat{\mu}_i$ denote the MLE of μ_i under the model of interest and let $\tilde{\mu}_i = y_i$ denote the MLE under the saturated model (McCullagh and Nelder, 1989). From the first principles of exponential family, the examples are:

Example 4.4.1.1: Deviance for a Binomial model

$$\begin{aligned} D(y,\hat{u}) &= 2\sum\left[y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right) - y_i \log\left(\frac{\hat{\mu}_i}{n_i}\right) - (n_i - y_i)\log\left(\frac{n_i - \hat{\mu}_i}{n_i}\right)\right] \\ &= 2\sum\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right] \end{aligned} \tag{4.41}$$

Example 4.4.1.2: Deviance for a Poisson model

$$\begin{aligned} D(y,\hat{u}) &= 2\sum\left[y_i \log(y_i) - y_i - \log(y_i!) - y_i \log(\hat{\mu}_i) + \hat{\mu}_i + \log(y_i!)\right] \\ &= 2\sum\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right] \end{aligned} \tag{4.42}$$

Example 4.4.1.3: Deviance for a Normal model

$$\begin{aligned} D(y,\hat{u}) &= 2\sum\left\{y_i(y_i - \hat{\mu}_i) - \frac{1}{2}y_i^2 + \frac{1}{2}\hat{\mu}_i^2\right\} \\ &= 2\sum\left\{\frac{1}{2}y_i^2 - y_i\hat{\mu}_i^2 + \frac{1}{2}\hat{\mu}_i^2\right\} \\ &= \sum(y_i - \hat{\mu}_i)^2 \end{aligned} \tag{4.43}$$

Example 4.4.1.4: Deviance for a Bernoulli model

$$\begin{aligned} D(y,\hat{u}) &= 2\sum\left[y_i \log(y_i) + (1 - y_i)\log(1 - y_i) - y_i \log(\hat{\mu}_i) - (1 - y_i)\log(1 - \hat{\mu}_i)\right] \\ &= 2\sum\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i)\log\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right)\right] \end{aligned} \tag{4.44}$$

4.5 Estimation of the Scale Parameter

When we assumed that the scale parameter is unknown, an estimate is obtained by using the following methods:

- **The deviance method**

  $\hat{\phi} = \frac{D}{N-p}$
  where N is the number of sample cases (number of rows in the data set we are modeling) and p is number of parameters.

- **Pearson χ^2**

  For example

  - Normal: $\chi^2 = SSE = (y_i - \hat{\mu})^2/Var(\hat{\mu})$
  - Poisson: $\chi^2 = (y_i - \hat{\mu})^2/\hat{\mu}$
  - Binomial: $\chi^2 = (y_i - \hat{\mu})^2/[\hat{\mu}(1-\hat{\mu})]$

  If the model is correct then
  $$\hat{\phi} = \frac{\chi^2}{N-p}$$
  In the following distributions we give ϕ :

  - Normal: $SSE/(N-p) \approx \sigma^2 = \phi$
  - Poisson: $\chi^2/(N-p) \approx 1 = \phi$
  - Binomial: $\chi^2/(N-p) \approx 1 = \phi$

  If the model holds, then $\chi^2/(N-p)$ is a consistent estimate for σ^2 in the asymptotical sequence $N \to \infty$ for fixed n_i's.

- **The method of moments or MLE**
  In this section, the method of moments agrees with MLE where the estimate is

  $\hat{\phi} = \frac{Var(y_i)}{Var(\hat{\mu})} = \frac{\sum(y_i-\hat{\mu})^2}{(n-p)Var(\hat{\mu})}$

and p is the number of parameters estimated.

4.6 Distribution of the Scaled Deviance

The likelihood ratio criterion (LRC) compare two models in the exponential family and has the form

$$-2\log\lambda = 2\sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{\phi}$$

Then we can write the likelihood ratio criterion as follows

$$-2\log\lambda = \frac{D(y,\hat{\mu})}{\phi}$$

Note that $D(y,\hat{\mu})$ does not depend on unknown parameters and it is called deviance. From the equation above, we note that LRC is the deviance divided by the scale parameter ϕ and we can call this the scaled deviance.

4.7 Theory of Generalized Estimating Equations (GEEs)

4.7.1 Introduction

Correlated data are very common in the longitudinal data setting. The most common applications include longitudinal and clustered data. Generalized estimating equations (GEEs) are a suitable and general approach to analyse in these kinds of correlated data. GEE was introduced by Liang and Zeger (1986) as a method of estimation regression model parameters dealing with correlated data. The GEE method, an extension of the quasi-likelihood (QL) approach, is being increasingly used to analyze longitudinal and other clustered data, especially when the outcome measure of interest is discrete (i.e binary or count data) rather than continuous (Hanley et al., 2003). This approach generalized the estimation method of quasi-likelihood of Wedderburn (1974) to the correlated data (Yan et al., 2007). An alternative generalization was proposed by Lee and Nelder (2001). The GEE approach focuses on models for the mean of the correlated observations within clusters without fully specifying the joint distribution of the observations (Yan et al., 2007).

4.7.2 Advantages of GEE

The GEE, like any other model, has some advantages. The main advantage of GEE lies in the unbiased estimation of population-averaged regression coefficients dispate possible misspecification of the correlation structure. According to Lipsitz et al. (1994) GEE provides some benefits over other models:

- Accounts for within-subject/within-clustered correlation.
- Allows for missing data.
- It has a range of correlation structures.
- Allows for time-varying covariates.

- Allows for infrequently or irregularly-timed or mistimed measurements.

GEEs have consistent and asymptotically normal solutions even with misspecification of the correlation structure. It avoids the need for multivariate distribution by only assuming a functional form for the marginal distribution at each time point (i.e y_{ij}). The covariance structure is treated as a nuisance. GEE cases are assumed to be dependent within subjects and independent between subjects. The correlation matrix that represents the within-subject dependencies is estimated as part of the model.

4.8 Assumption of GEEs

- The responses are $Y_1, Y_2, ..., Y_n$ are correlated or clustered, i.e., cases are NOT independent.
- It uses quasi-likelihood estimation rather than maximum likelihood estimation (MLE) or ordinary least squares (OLS) to estimate the parameters.
- The homogeneity of variance does NOT need to be satisfied
- Covariates can be the power terms or some other nonlinear transformations of the original independent variables and it can have interaction terms.
- Errors are correlated.
- It has a covariance specification
- Missing data in Weighted GEE (for handling missing at random (MAR) dropouts) and GEE missing complete at random (MCAR) dropouts)

4.9 Specification Needed for GEEs

According to Liang and Zeger (1986), Guo (2011) and Yan et al. (2007) the Generalized Estimating Equations procedure extends the generalized linear model to allow for analysis of repeated measurements or other correlated observations, such as clustered data and the setting is as follows: one each of $i = 1, \ldots, N$ subjects, there are made n_i measurements $y_i = (y_{1i}, \ldots, y_{in_i})$. The independent is assumed to be the measurements on the different subjects. Measurements on the same subject are allowed to be correlated. The model specification of a GEE requires three elements. The model formulation is similar to that of a GLM but full specification of the joint distribution is not required and thus no likelihood function

$$g(\mu_i) = x_i^T \beta \tag{4.45}$$

Systematic part: This relates the expectation $E(y_{ij}) = \mu_{ij}$ to the linear predictor via the link function.

$$g(\mu_{ij}) = \eta_{ij} = x_i^T \beta \tag{4.46}$$

Random part: This specifies how the variance $Var(y_{ij})$ is related to the mean $E(y_{ij})$ by specifying a variance function $Var(\mu_{ij})$ such that $Var(y_{ij}) = \phi Var(\mu_{ij})$

The correlation part: This is the part which differentiates the GEE model from the GLM. We need to apply a correlation structure for observation on the same unit. This is done by specifying a working correlation matrix.

According to Owusu-Darko et al. (2014) correlated data are modeled using the same link function and linear predictor setup (systematic component) as the independence case. The random component is described by the same variance functions as in the independence case but the covariance structure of the correlated measurements must also be modeled. Let the vector of measurements on the i^{th} subject be $Y_i = [Y_{i1}, Y_{i2}, \cdots, Y_{in_i}]'$ with corresponding vector of means $\mu_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{in_i}]'$ and let V_i be an estimate of the covariance matrix of Y_{ij}. The unknown regression coefficient vector β is of primary interest. The GEE for estimating β is an extension of the independence estimating equation to correlated data and is given by

$$\sum_i D_i' V_i^{-1}(y_i - \mu_i) = 0 \tag{4.47}$$

where $D_i = D_i(\beta) = \frac{\partial \mu_i}{\partial \beta}$ and V_i is the working covariance matrix of Y_i. V_i can be expressed in terms of a working correlation matrix $R(\alpha)$

$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}} / \phi \tag{4.48}$$

where A_i is a diagonal matrix with element $Var(y_{ij} = V(\mu_{ij})$ specified as function of the means μ_{ij}, α is some unknown parameters. The parameter α can be estimated through method of moments or another set of estimating equations (Prentice, 1988) . Let D and V denote conforming matrices constructed for D_i and V_i defined above. Then according to Liang and Zeger (1986) the asymptotical covariance for the p covariates is given by a $p \times p$ matrix

$$V_\beta = \lim_{n \to \infty} n(D^T V^{-1} D)^{-1} D^T V^{-1} (y - \mu)(y - \mu)^T V^{-1} D (D^T V^{-1} D)^{-1} \tag{4.49}$$

where $\hat{\mu} = g^{-1}\hat{\eta}$ with $\hat{\eta}$ being $\hat{\eta}$ evaluated at convergence.

4.9.1 Iteratively Reweighted Least Squares Algorithm

According to Liang and Zeger (1986) the solution is obtained by alternating between estimation of ϕ, β and α using method of moments estimators for ϕ and α. Thus in summary the IRWLS proceeds as follows:

- Step 1: Assuming $R = I$ and $\phi = 1$, provide initial estimate of β with GLM algorithm.
- Step 2: Estimate ϕ and α
- Step 3: Use an updated ϕ and α to estimate β
- Return to step 2 and 3 until convergence.

According to Liang and Zeger (1986) at a given iteration the correlation parameters α and scale parameter ϕ can be estimated from the current Person residuals defined by

$$\hat{r}_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{Var(\mu_{ij})}}$$

where μ_{ij} depends upon the current value β. We can estimate ϕ by

$$\hat{\phi} = \sum_{i=1}^{k} \sum_{j=i}^{n_i} \frac{\hat{r}_{ij}}{N - p} \tag{4.50}$$

where $N = \sum n_i$. This is the longitudinal analogue of the familiar Person statistic (McCullagh, 1983; Liang and Zeger, 1986). Given $\hat{\phi}$, a method of moments estimator for the parameter α is

$$\hat{\alpha} = \hat{\phi} \sum_{i=1}^{N} \sum_{j>j'}^{N} \frac{\hat{r}_{ij}\hat{r}_{ij'}}{\sum_{i=1}^{N} \frac{1}{2} n_i(n_i - 1) - p} \tag{4.51}$$

4.9.2 Newton-Iteration

To solve the system of equation using the Newton-iteration method, there are some steps which must be followed. Then the fitting algorithm becomes:

1. Compute an initial estimate of β from a GLM (i.e by assuming independence)
2. Compute an estimate $R(\alpha)$ of the working correlation on the basis of the current Person residuals and the current estimate of β
3. Compute an estimate of the variance as

$$V_i = \phi A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$$

4. Compute an updated estimate of β based on the Newton-step

$$\beta_{j+1} = \beta_j + \left[\sum_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right]^{-1} \left[\sum_i \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (y_i - \mu_i(\beta))\right] \tag{4.52}$$

This should iterate through step $2-4$ until convergence. Note that ϕ need not be estimated until the last iteration. The GEE estimate $\hat{\beta}$ of β is often very similar to the estimates obtained if observations were treated as being independent. In other words, the estimate $\hat{\beta}$ for GEE is often very similar to the estimates obtained by fitting a QL-method to the data Yan et al. (2007)

4.10 Estimation

Parameter Estimation: The controls in this group allow you to specify estimation methods and to provide initial values for the parameter estimates.

4.10.1 Estimation of Regression Coefficients $\hat{\beta}$

We estimate β by solving the generalized estimating equations (GEE). The GEE estimator of β is the solution of 4.47 where $\hat{\alpha}$ is a consistent estimator α and $D_i = \frac{\partial \mu_i}{\partial \beta}$ (Hedeker and Gibbons, 2006). For example normal case, $\mu_i = X_i\beta$, $D_i = X_i$ and $Var(\hat{\alpha}) = \phi R(\hat{\alpha})_i$ which yields for solving $\hat{\beta}$

$$\sum_i X_i' [R_i(\hat{\alpha}) X_i]^{-1} (y_i - X_i\beta) = 0 \tag{4.53}$$

Therefore

$$\hat{\beta} = \left[\sum_i X_i' [R(\hat{\alpha})_i]^{-1} X_i\right]^{-1} \left[\sum_i X_i' [R(\hat{\alpha})_i]^{-1} y_i\right] \tag{4.54}$$

This is similar to weighted least-squares (WLS) estimator and is more generally used because solution only depends on the mean and variance of y, these are quasi-likelihood estimates (Hardin and Hilbe, 2003). According to Liang and Zeger (1986) this is asymptotically normality if the variance function $V(\mu)$ is incorrectly specified and the working correlation matrix R is not the true correlation matrix.

4.10.2 Estimating the Covariance of $\hat{\beta}$

In order to perform hypothesis tests and construct confidence intervals, we are interested to obtain standard errors associated with the estimated regression coefficients. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\beta})$. The GEE provides two versions of these:

1. **Robust or "empirical" or sandwich estimator**
   The estimator
   $$V(\hat{\beta}) = M_0^{-1} M_1 M_0^{-1} \tag{4.55}$$
   where
   $$M_0 = \textstyle\sum_i \frac{\partial \hat{\mu}_i'}{\partial \beta} \hat{V}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta}$$
   and
   $$M_1 = \textstyle\sum_i \frac{\partial \hat{\mu}_i'}{\partial \beta} \hat{V}_i^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' \hat{V}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta}$$
   is called the empirical or robust estimator of the covariance matrix of $\hat{\beta}$. According to Owusu-Darko et al. (2014) it has the characteristic of being a consistent estimator of the covariance matrix of $\hat{\beta}$ even if the working correlation matrix is misspecified. We notice that if $\hat{V}_i = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$ then the two estimator equations are equal. This occurs only if the true correlation structure is correctly modeled. In most cases, the robust or "sandwich" estimator provides a consistent estimator of $V(\hat{\beta})$ even if the working correlation structure $R_i(\alpha)$ is not the true correlation of y_i (Hedeker and Gibbons, 2006 and Owusu-Darko et al., 2014).

2. **Naive or "model-based" estimator**
   This is the "GEE-version" of the inverse of the Fisher information often used in GLMs as an estimator of the covariance estimates of the maximum likelihood estimator (MLE) of $\hat{\beta}$. This is given by
   $$V(\hat{\beta}) = M_0^{-1} \tag{4.56}$$
   For $D_i = X_i$ then $V(\hat{\beta})$ becomes
   $$V(\hat{\beta}) = \left[\sum_{i=1}^{N} X_i \hat{V}_i^{-1} X_i \right]^{-1} \tag{4.57}$$
   Here $Cov(\hat{\beta})_m$ is a consistent estimator of the covariance matrix of $\hat{\beta}$ if the mean model and the working correlation matrix are correctly specified.

4.11 Working Correlation Matrix

In GEE modeling one has to specify the working correlation matrix in estimating the covariance of the estimates. According to Owusu-Darko et al. (2014) the specification of the working correlation matrix accounts for the form of the within-subject correlation of responses on dependent variables. One of the aims of this work is to find out whether using different working correlation matrices for estimation would substantially affect the estimates and standard errors with respect to model-based and empirical-based estimator. The working correlation $R(\alpha)$ matrix is

$$V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}} / \phi \tag{4.58}$$

where $R(\alpha)$ is a $n_i \times n_i$ matrix of correlation coefficients, number is between -1 and 1. The parameter α is a tunable parameter on which $R(\alpha)$ depends. The $R(\alpha)$ is assumed to depend on a set of parameters α. The over dispersion parameter ϕ is assumed to be known. If it is unknown it is included as a parameter to be estimated from the data using some methods as mentioned above. The α is estimated in terms of Person residuals given by

$$\hat{e_{ij}} = \frac{y_{ij} - \hat{\mu_{ij}}}{\sqrt{V(\hat{\mu_{ij}})}} \tag{4.59}$$

subject to the structure of the assumed correlation. There are several specific choices of the form of the working matrix $R(\alpha)$ to model the correlations of the individual responses.

Independent (R_{IN})

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{j=k} \\ 0 & j \neq k \end{cases} \qquad R_{IN} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Exchangeable (R_{EX})

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{j=k} \\ \alpha & j \neq k \end{cases} \qquad R_{EXCH} = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}$$

Autoregressive (R_{AR-1})

$$Corr(y_{ij}, y_{ik}) = \alpha^k \quad \text{for} \quad t = 0, 1, \ldots, n_i - j \qquad\qquad R_{AR-1} = \begin{pmatrix} 1 & \alpha & \cdots & \alpha^{n-1} \\ \alpha & 1 & \cdots & \alpha^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{n-1} & \alpha^{n-2} & \cdots & 1 \end{pmatrix}$$

Toeplitz (R_{TOEP})

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{j=k} \\ \alpha & j = 1, 2, \ldots, n_i - t \end{cases} \qquad\qquad R_{TOEP} = \begin{pmatrix} 1 & \alpha_1 & \cdots & \alpha_{n-1} \\ \alpha_1 & 1 & \cdots & \alpha_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha\rho_{n-1} & \alpha_{n-2} & \cdots & 1 \end{pmatrix}$$

Unstructured (R_{UN})

$$Corr(y_{ij}, y_{ik}) = \begin{cases} 1 & \text{j=k} \\ \alpha_{jk} & j \neq k \end{cases} \qquad\qquad R_{UN} = \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & 1 & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & 1 \end{pmatrix}$$

4.12 Inference

Now our interest is to test the hypothesis concerning the elements of β. We consider the hypothesis of the form

$$H_0 : L\beta = d \tag{4.60}$$

where L is a $r \times p$ matrix of constants imposing ℓ linearly independent constants on the elements of β and d is a $r \times 1$ vector of constants. Since $\hat{\beta}$ is asymptotically normal, then with the large sample approximation

$$L\hat{\beta} \sim N(L\hat{\beta}, L\hat{V}_{\beta}L^T) \tag{4.61}$$

the Wald χ^2 test statistics is given by

$$\chi^2 = (L\hat{\beta} - d)(L\hat{V}_{\beta}L^T)^{-1}(L\hat{\beta} - d) \tag{4.62}$$

that has an asymptotic χ^2_{ℓ} distribution if H_0 is true.

4.13 Model Fit Analysis under the QIC Statistics

According to Hanley et al. (2003), the Akaike information criterion (AIC) is a goodness-of-fit measure for likelihood based models. It is defined as

$$AIC = -2L + 2p \tag{4.63}$$

where L is the log likelihood and p is the number of parameters in the model. Since GEE is a non-likelihood based model, we do not have a likelihood function in context. However we may have quasi-likelihood. We propose to replace the likelihood by the quasi-likelihood Q under the working independent model (Pan, 2001). An extension of AIC is QIC, the quasi-likelihood under the independent model information criteria. This measure is more appropriate for GEE which is a quasi-likelihood method. Like the AIC, the smaller the QIC the better. The QIC is defined as

$$QIC(R) = -2Q(g^{-1}(x\beta_R)) + 2trace(A_I^{-1}V_{MS,R}) \tag{4.64}$$

where $-2Q(g^{-1}(x\beta_R))$ is the value of the quasi-likelihood calculated with the proposed correlation structure R and $g^{-1}(x\beta_R) = \hat{\mu}$ where g^{-1} is the inverse link function for the model, a logit for this model. We define A_I as the variance matrix under the independent model and we define $V_{SM,R}$ as the sandwich estimate of variance under the hypothesized correlation structure R.

4.14 Application of GEEs

The aim of this section is to analyze RSV data using application of Generalized Estimating Equations (GEE) models under various working correlation assumptions. From our analysis, we check the test of model-based and empirical-based standard error estimates on coefficients estimation and parameter estimates of the study based on our GEE assumption model. The application was carried out using SAS PROC GENMOD which is an implanted procedure in SAS suited to fitting both GLM and their extensions to GEEs allowing for the specification of the quasi-likelihood and the correlation structure for correlated data.

4.14.1 Statistical Model

The statistical defining of data is refer to the paper by Mwambi et al. (2011)

$$\text{rsv}_{ij} = \begin{cases} 1, & \text{uninfected for subject } i \\ 2, & \text{otherwise} \end{cases} \quad \text{and}$$

$$\text{actpass}_{i1} = \begin{cases} 1, & \text{active sampling for subject } i \\ 2, & \text{otherwise} \end{cases}$$

From the Table 4.3 the statistical model can be written as:

$$\log(E(y_{ij})) = \beta_0 + \beta_1 actpass_{i1} + \beta_2 age_{ij} + \beta_3 dt_{ij} \tag{4.65}$$

where β_1 is the main effect of an *actpass*, β_2 is the main effect of an *age* effect, β_3 is the main effect of *dt* effect and β_4 is the main effect of *visit* effect.

The commands to fit the model in SAS code for GEE model are

```
proc genmod data=spha descending;
class id;
model rsvpos=actpass age dt /dist=bin link=logit;
Repeated sub=id/type=cs covb corrw modelse;
run;
```

The GENMOD procedure can fit models to correlated responses by the GEE method and it uses maximum likelihood methods. The options in the REPEATED statement specify the correlation structure as well as convergence criteria. The CORR option is probably the most important option in the REPEATED statement. It is used to specify the "working correlation matrix" and the SUBJECT option identifies the cluster. MODELSE statement displays an analysis of parameter estimates table using model-based standard errors. The "Analysis of Parameter Estimates" table based on empirical standard errors is displayed by using default in SAS.

4.14.2 Hypothesis Tests

We can test for $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ in (4.65) to investigate whether there is a significant association between the pattern of change of the responses and actpass.

4.14.3 Output in SAS and Interpretation

Table 4.3: Estimated Coefficients, Standard Errors and P-Values : GEE Models

Model	INDEPENDENCE		EXCHANGEABLE		AR(1)		TOEPLITZ	
Parameter	Estimates	Std Error	Estimates	Std Error	Estimates	Std Error	Estimates	Std Error
Intercept	-1.8674^*	0.575 (0.585)	-1.8146^*	0.581 (0.581)	-1.9215^*	0.565 (0.595)	-1.9214^*	0.565 (0.594)
actpass	-2.2382^*	0.454 (0.544)	-2.2308^*	0.438 (0.548)	-2.2115^*	0.445 (0.544)	-2.2129^*	0.445 (0.544)
age	0.3889	0.174 (0.206)	0.3486	0.188 (0.213)	0.3935	0.174 (0.212)	0.3936	0.174 (0.211)
dt	-0.0111	0.029 (0.028)	-0.0098	0.029 (0.028)	-0.0101	0.029 (0.028)	-0.0100	0.029 (0.028)

Note: * Shows a parameter estimate that has significant effects at 5% level of significance.

(.) Shows model-based standard error estimates (std err).

Table 4.3 gives the analyzes of GEE parameter estimation for the main effect based on our four assumptions (independent, autocorrelation, Toeplitz and exchangeable) with their respective model-based and empirical-based standard error estimates. It could be concluded from the table that the parameter estimates for the variables (actpass, age and dt) are approximately the same for both empirical and model-based parameter estimates for all the assumptions excluding exchangeable. However, the standard errors for the robust and naive cases are marginally different. This may indicate that the true correlation structure for the GEE is not correctly modeled using the independence, autocorrelation and Toeplitz model assumption.

The parameter estimation for actpass is highly significant but has the highest standard errors values for both empirical and model-based estimations. The estimation of the model age parameter was seen to be statistically significant at $\alpha = 0.05$ significance level for all the assumptions under empirical-based estimates but was statistically insignificant under model-based estimates. The parameter estimation for dt was statistically not significant at $\alpha = 0.05$ significance level for all assumptions in both model estimates.

A vital observation of the standard errors for model-based and empirical-based estimation is marginally different and relatively small for all our assumptions except the exchangeable GEE model. The parameter estimates for empirical and model based in the GEE exchangeable model are the same. The standard error estimate for empirical and model-based are approximately the same. The parameter intercept is significant. The estimated standard error estimates for robust or sandwich estimators of the model for all parameters estimates are equal. We notice that if

$$\hat{V}_i = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)'$$

then the model-based and empirical-based standard errors estimates are equal. This occurs only if the true correlation structure is correctly modeled. In our case, comparing the analysis of exchangeable GEE model with the other working correlation assumptions discussed above, we choose the exchangeable GEE model as the best fit for our analysis.

4.14.4 Summary

Generalized estimating equations (GEEs) are a suitable and general approach to analysis in these kinds of correlated data. The GEE method is an extension of the quasi-likelihood (QL) approach and is being increasingly used to analyze longitudinal and other clustered data, especially when the outcome measure of interest is discrete (i.e. binary or count data) rather than continuous. The GEE method is known to provide consistent regression parameter estimates regardless of the choice of working correction structure, provided $\sqrt{n}$ consistent nuisance parameters are used. However, it is essential to use the appropriate working correlation structure in small samples, since it improves the statistical efficiency of $\hat{\beta}$. The GEE works best if the number of observations per subject is small and the number of subjects is large and also works best if it is in longitudinal that the measurements are taken at the same time for all subjects. GEE is not good for high unbalanced data sets. It allows for a flexible relationship to be modeled between the response and any model covariates. It also allows for correlation within the counts, which is very important when considering variance regression. Coefficient $\beta_{\hat{GEE}}$ is asymptotically correct if the underlying regression mean is model is correct. Suppose the mean is correctly specified and, the variance and correlation structure are incorrect but GEE model still provides consistent estimates of the parameters, and even if assumed correlation model is incorrect the consistent estimates of the valid standard errors can be obtained with the sandwich covariance estimator. That is the main advantage of GEE models. The GEE is an easy methodology to use, it runs well in SAS and produces population estimates directly. In SAS, it uses PROC GENMOD procedure to fit model. PROC GENMOD is a powerful tool to conduct GLM as well as the extension to GEE where correlated outcome data were taken into account.

Chapter 5

Generalized Linear Mixed Models (GLMMs)

5.1 Introduction

According to Pan and Lin (2005), Generalized linear mixed models (GLMMs) as proposed by Breslow and Clayton (1993) are obtained from the extension of generalized linear models (GLMs) that was proposed by McCullagh and Nelder (1989) by including random effects into the linear predictors and include the well-known linear mixed models (LMMs) for normal responses (Laird and Ware, 1982) as a special case. The applications are useful in various disciplines such as the analysis of clustered data including longitudinal data or repeated measures. According to Feddag and Mesbah (2006), GLMMs are extension of GLMs that accommodate correlated and over-dispersed data by adding random effects to the linear predictor. These models are useful when the interest of the analyst lies in the individual response profiles rather than the marginal mean $E(y_{ij})$. These models are useful for modeling the dependence among response variables existing in longitudinal or repeated measures studies, for accommodating overdispersion among binomial or Poisson responses, and for producing shrinkage estimators in multiparameter problems, such as the construction of maps of small area disease rates (Breslow and Clayton, 1993). The estimates of the parameters are obtained by maximum likelihood or restricted maximum likelihood. Breslow and Clayton (1993) used the approximations in the penalized quasi-likelihood (PQL) and the marginal quasi-likelihood (MQL) approach to find the regression parameter estimates and the variance components. The other methods are Markov Chain Monte Carlo (MCMC), Bayesian approach, and Maximum stimulated likelihood (MLS). In our case we will focus on Laplace approximation which is a combination of multivariate Taylor series expansion and Laplace approximation that is fast, computationally accurate and gives a likelihood for likelihood ratio tests.

5.2 Generalized Linear Mixed Models

The generalized linear mixed models combine the linear mixed models described in Chapter 3 with the generalized linear models introduced in Chapter 4. According to Gbur et al. (2012), the GLMMs is an extension of generalized linear model which includes the random effects into the linear predictor. In general, let Y be response variable whose conditional distribution given the random effects belongs to the exponential family or can be written as quasi-likelihood (Gbur et al., 2012). Given a vector u_i of random effect for cluster i, it is assumed that all responses Y_{ij} are independent with density function that is given by

$$f(y_{ij}|u,\theta_{ij},\phi) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij},\phi)\right\} \tag{5.1}$$

in which θ_{ij} is now modeled as

$$\theta_{ij} = x_{ij}^{'}\beta + Z_{ij}^{'}u_i \tag{5.2}$$

and u_i is assumed that

$$u_i \sim N(0,D)$$

where θ_i and ϕ are parameters and $b(\theta_i)$ and $c(y_i,\phi)$ are known functions. Let $x_1,\ldots,x_p$ be the set of explanatory variable for fixed effects and $u_1,\ldots,u_q$ be the set of random effects. The linear predictor of the model for the observation given the random effects is given by

$$\begin{aligned}\eta_{ij} &= g\left(E[Y_j|u_1,\ldots,u_q]\right)\\ &= \beta_0 + \sum_{i=1}^{p}\beta_i x_{ij} + \sum_{k=1}^{q} Z_{kj}u_k, \qquad j=1,\ldots,n\end{aligned} \tag{5.3}$$

where β_0 is the overall mean, β_i is the i^{th} fixed effects coefficient, x_{ij} is the i^{th} fixed effects explanatory variable on the j^{th} observations, Z_{kj} is the binary indicator variable for the effect of the k^{th} random effects, u_k on the j^{th} observation and $g(.)$ is the link function relating the condition mean of the response to the predictor. In matrix form, the linear predictor can be written as

$$\begin{aligned}\eta &= g\left(E[Y|u]\right)\\ &= X\beta + Zu\end{aligned} \tag{5.4}$$

where Y is the $n \times 1$ vector of response, X is the $n \times (p+1)$ fixed effects design matrix, β is the $(p+1) \times 1$ vector of fixed effects coefficients, Z is the $n \times q$ design matrix for random effects and u is the $q \times 1$ vector of random effects.

The expectation of the GLMMs are

$$E\left(Y|u\right) = g^{-1}(X\beta + Zu) = g^{-1}(\eta) \tag{5.5}$$

The conditional variance is given by

$$V\left(Y|u\right) = R = \phi V^{\frac{1}{2}} P V^{\frac{1}{2}} \tag{5.6}$$

where P is a working correlation matrix, $V^{\frac{1}{2}}$ is a diagonal matrix with the square root of the variance function on diagonal and ϕ is a scale parameter. The relationship between the linear predictor and the vector of observations is modeled as

$$(Y|u) \sim (g^{-1}(\eta), R)$$

This means that the conditional of y given u has mean $g^{-1}(\eta)$ and variance R. According to Gbur et al. (2012), if P is the identity matrix, then the R is an $n \times n$ covariance matrix. The random effects u are assumed to be a multivariate normally distributed mean zero and variance G i.e $u \sim MVN(0, G)$

$$E(u) = 0 \qquad \text{and} \qquad Var(u) = G$$

5.3 Likelihood Function of GLMMs

According to Verbeke and Molenberghs (2005) and Schelldorfer et al. (2012), we let $f_{ij}(y_{ij}|u_i, \beta, \phi))$ denote the conditional density function of Y_{ij} and u_i and the marginal distribution of Y_{ij} is given by

$$f_i(y_i, \beta, u_i, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|u_i, \beta, \phi)) f(u_i, G) du_i \tag{5.7}$$

where $f(u_i, G)$ is the density of $u \sim (0, G)$ distribution. According to Moudud (2009) the distribution of random effects depends on unknown parameter is given by

$$
\begin{aligned}
L(\beta, G, \phi, y) &= \prod_{i=1}^{n} f(y_i|G, \beta, \phi) du \\
&= \prod_{i=1}^{n} \int \prod_{j=1}^{n_i} f(y_{ij}|G, \beta, \phi) f(u_i, G) du_i
\end{aligned}
\tag{5.8}
$$

The likelihood function of GLMMs is from the idea of marginal likelihood function from LMMs which is an integral of the joint density function. The likelihood function of a GLMM is given by the following expansion

$$
\begin{aligned}
L(\beta, \theta, \phi) &= \int \prod_{i=1}^{n} \left[\exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + C(y_i, \phi) \right\} \right] \frac{1}{(2\pi)^{\frac{q}{2}}} \exp\left\{ -\frac{1}{2} ||u||_2^2 \right\} du \\
&= \frac{1}{(2\pi)^{\frac{q}{2}}} \int \left\{ \exp\left(\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} + C(y_i, \phi) \right) - \frac{1}{2} ||u||_2^2 \right\} du
\end{aligned}
\tag{5.9}
$$

where $||u||_2^2 = (u - \hat{u})' Q(\hat{u})(u - \hat{u})$. According to Verbeke and Molenberghs (2005) and Jiang (2007) the integral of equation (5.8) cannot be worked out analytically under non-normal linear mixed model but numerical approximations are required.

5.4 Estimation and Inference in GLMMs

The estimation of the parameters is obtained by maximum likelihood or restricted maximum likelihood (REML). Breslow and Clayton (1993) used the approximations in the marginal quasi-likelihood (MQL) approach to find the regression parameter estimates and the variance components are estimated by REML or the profiled maximum likelihood. They also used the penalized quasi-likelihood (PQL); this is based on first-order Taylor expansions around the maximum of current estimates of the random effects via the first-order Laplace approximation of the integrals. These approaches produce biased estimates for both the regression and variance components parameters (Feddag and Mesbah, 2006). Breslow and Lin (1995) provide the correction factor for the estimates of the variance component derived from the second-order Laplace approximations and extend this bias correction to the GLMMs with multivariate random effects. The several approximation methods that were carried out by former studies such as Guass-Hermite quadrature, Laplace approximation, Penalized quasi-likelihood and Marginal quasi-likelihood are used to find the estimates.

5.5 Laplace Approximation

The Laplace's method is an alternative approach of the approximation of integral. Sun (2011) and Breslow and Lin (1995) used the fifth-order Laplace approximation to estimate random effects with a single random effect per cluster and Raudenbush et al. (2000) extend this idea to high order approximation and multiple dependent random effects per cluster. The integrals in $L(\beta, D, \phi)$ can be written in the form

$$I = \int e^{Q(b)} du \tag{5.10}$$

and the second-order Taylor expansion of $Q(b)$ about $\hat{b}$ is given by

$$Q(b) \approx Q(\hat{b}) + \frac{1}{2}(b - \hat{b})' Q''(\hat{b})(b - \hat{b}) \tag{5.11}$$

where the first-order term of the Taylor expansion disappears since the expansion is done about $\hat{b}$ and $Q''(\hat{b}) = -\ell''(\beta, D, \phi)|_{b=\hat{b}}$ is a Hessian of the log-likelihood evaluated at $\hat{b}$. Then by using the approximation in the Laplace approximation, the quadratic term leads us to

$$I \approx (2\pi)^{\frac{q}{2}} |Q''(\hat{b})|^{-\frac{1}{2}} e^{Q(\hat{b})} \tag{5.12}$$

and the marginal log-likelihood becomes

$$\begin{aligned} \ell(\theta, y) &= \log \int \exp\left(\ell(\theta, \hat{b}, y) - \frac{1}{2}(b - \hat{b})' Q''(\hat{b})(b - \hat{b})\right) db \\ &= \ell(\theta, \hat{b}, y) - \frac{1}{2} \log \left| \frac{Q''(\hat{b})}{2\pi} \right| \end{aligned} \tag{5.13}$$

This gives good approximation if we have many repeated measures per subject.

5.6 Approximation of the Data

According to Verbeke and Molenberghs (2005), we re-write GLMM as

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \varepsilon_{ij} \\ &= h(x'_{ij}\hat{\beta} + z'_{ij}\hat{b}_i) + \varepsilon_{ij} \end{aligned} \tag{5.14}$$

where $h(.)$ is the inverse link function and the error terms have the variance equal to

$$V(Y_{ij}|b_i) = \phi v(\mu_{ij})$$

and where $v(.)$ is usually the variance function of mean in an exponential family in generalized linear model. The several methods of the tools for the approximation of the data that was proposed by Breslow and Clayton (1993) and other current studies will be discussed in these section. The two commonly used methods that will be discussed are penalized quasi-likelihood and marginal quasi-likelihood

5.6.1 Penalized Quasi-Likelihood (PQL)

The penalized quasi-likelihood (QPL) is used for the estimation of the parameters under maximum likelihood (ML) or restricted maximum likelihood (REML). According to Breslow and Clayton (1993), the PQL approximation is one of the approaches used to find the the regression parameter estimates and the variance components are estimated by the REML or the profiled maximum likelihood. PQL is based on Taylor expansions around the maximum of current estimates of the random effects via the Laplace approximation of the integrals. In our case the linear Taylor expansion is about β and $\hat{b}_i$ and the model is given by

$$\begin{aligned} Y_{ij} &\approx h(x_{ij}^{'}\hat{\beta}+z_{ij}^{'}\hat{b}_i)+h'(x_{ij}^{'}\hat{\beta}+z_{ij}^{'}\hat{b}_i)x_{ij}^{'}(\beta-\hat{\beta})+h'(x_{ij}^{'}\hat{\beta}+z_{ij}^{'}\hat{b}_i)z_{ij}^{'}(b_i-\hat{b}_i)+\varepsilon_{ij} \\ &= \hat{\mu}_{ij}+V(\hat{\mu}_{ij})x_{ij}^{'}(\beta-\hat{\beta})+V(\hat{\mu}_{ij})z_{ij}^{'}(b_i-\hat{b}_i)+\varepsilon_{ij} \end{aligned} \tag{5.15}$$

In vector notation we re-write the expansion as

$$Y_i = \hat{\mu}_i + V_i x_i^{'}(\beta-\hat{\beta})+V_i z_i^{'}(b_i-\hat{b}_i)+\varepsilon_i \tag{5.16}$$

and by re-ordering terms gives us

$$\begin{aligned} Y_i^* &\equiv +V_i^{-1}(Y_i-\hat{\mu}_i+)x_i\hat{\beta}+z_i\hat{b}_i \\ &\approx x_i\hat{\beta}+z_i\hat{b}_i+\varepsilon_i^* \end{aligned} \tag{5.17}$$

which is an approximate LMM with pseudo responses Y_i^* and where $\varepsilon_i^* = \hat{V}^{-1}\varepsilon_i$. According to Verbeke and Molenberghs (2005) model fitting proceeds by iterating between updating the pseudo responses and fitting the model approximate model in the equation Y^* similar to a linear mixed model until convergence. This approach produces biased estimates for both the regression and variance components parameters (Feddag and Mesbah, 2006). Breslow and Lin (1995) provide the correction factor for the estimates of the variance component derived from the second-order Laplace approximation and extend this bias correction to the GLMM with multivariate random effects.

5.6.2 Marginal Quasi-Likelihood (MQL)

The marginal quasi-likelihood (MQL) approach is also used to find the regression parameter estimates and the variance components are estimated by REML or the profiled maximum likelihood (Breslow and Clayton, 1993). This approach also uses the Laplace approximation but the MQL linear Taylor expansion via around β and $\hat{b}_i = 0$ and the model is given by

$$\begin{aligned} Y_{ij} &\approx h(x_{ij}'\hat{\beta}) + h'(x_{ij}'\hat{\beta})x_{ij}'(\beta - \hat{\beta}) + h'(x_{ij}'\hat{\beta})z_{ij}'(b_i - \hat{b}_i) + \varepsilon_{ij} \\ &= \hat{\mu}_{ij} + V(\hat{\mu}_{ij})x_{ij}'(\beta - \hat{\beta}) + V(\hat{\mu}_{ij})z_{ij}'(b_i) + \varepsilon_{ij} \end{aligned} \tag{5.18}$$

In vector notation we re-write the expansion as

$$Y_i \approx \hat{\mu}_i + V_i x_i'(\beta - \hat{\beta}) + V_i z_i' b_i + \varepsilon_i \tag{5.19}$$

and by re-ordering terms gives us

$$\begin{aligned} Y_i^* &\equiv +V_i^{-1}(Y_i - \hat{\mu}_i +)x_i\hat{\beta} \\ &\approx x_i\hat{\beta} + z_i\hat{b}_i + \varepsilon_i^* \end{aligned} \tag{5.20}$$

According to Verbeke and Molenberghs (2005) this is similar to PQL; the model fitting proceeds by iterating between updating the pseudo responses and fitting the model approximate model in the equation Y^* similar to a linear mixed model until convergence.

5.6.3 Marginal Quasi-Likelihood (MQL) versus Penalized Quasi-Likelihood (PQL)

According to Molenberghs and Verbeke (2005) the differences and the relationship between MQL and PQL are as follows:

- With higher-order Taylor expansion, both approximations approach produce improvement estimates
- With increasing or large number (n_i) of measurements per subject
    1. MQL provides biased estimates, while
    2. PQL provides the consistent estimates values.
- With few repeated measurements per cluster both produced bad results for binary outcomes.
- MQL only performs reasonably well if random effects variance is very small.

5.7 Approximation of the Integral

According to Verbeke and Molenberghs (2005), the likelihood contribution of every subject is of the form

$$\int f(x)\phi(x)dx \tag{5.21}$$

where $\phi(x)$ is the density of the multivariate normal distribution. According to McCulloch (1997) and Sun (2011), the Gaussian quadrature methods replace the integral by a weighted sum and the integration is in the form

$$\int_{-\infty}^{\infty} f(x)\exp(-x^2)dx \tag{5.22}$$

which is approximately

$$\sum_{i=1}^{m} w_i f(x_i) \tag{5.23}$$

where m is the order of the approximation, the nodes x_i are solutions to the m^{th} order Hermite polynomial and w_i are well chosen weights. According to Verbeke and Molenberghs (2005), the approximation will be more accurate if we have higher m. In case of Gaussian quadrature the nodes and weights are fixed independent $f(x)\phi(x)$ and in case of adaptive Gaussian quadrature the nodes and weights are adapted to the support of $f(x)\phi(x)$ (Verbeke and Molenberghs, 2005).

5.8 Inference

We showed that GLMMs can be estimated by fitting the pseudo-data by linear mixed models. Thus we can use the same techniques of inference discussed in the previous chapter to find the inference of GLMMs since we will be fitting pseudo data for GLMMs using the same estimation methods used for linear mixed models.

5.9 Application to the Treatment of Lead-Exposed Children (TLC) Data

The background for this data is given in Chapter 2. We analyze the TLC data using generalized linear mixed models, because the data are frequency counts. Interest lies in testing whether the intensities of occurrences are significantly different between the two treatment groups. The covariates used in the analysis are treatment and week the same as in Chapter 3. The GLIMMIX procedure fits statistical models to data with correlations or non-constant variability and where the response is not necessarily normally distributed. These models are known as generalized linear mixed models (GLMM). This model is like linear mixed models, as it assumes the random effects to be normally distributed. The GLIMMIX procedure select the distribution of the response variable conditional on normally distributed random effects and the data can have any distribution in the exponential family. The exponential family contains many of the fundamental discrete and continuous distributions. For example, the binary, binomial, Poisson, and negative binomial distributions are discrete members of this family while the normal, beta, gamma, and chi-square distributions are examples of the continuous distributions in this family. The GLIMMIX procedure allows us to specify a generalized linear mixed model and to perform estimation and inference in such models. PROC GLIMMIX will be used in SAS 9.3 to fit the model. The syntax is similar to that of the MIXED procedure and includes CLASS, MODEL and RANDOM statements. In the case of fitting GLMMs, we compare three types of fitted models of GLMM, namely the random intercept model, random slope model and random intercept and slope model so that we will understand how to fit GLMMs. The review of SAS for random effects modeling focuses on the GLIMMIX. Table 5.1 shows the GLMM results when we consider the random intercept using the PQL and MQL methods.

The commands to fit the model in SAS code for penalized quasi-likelihood (PQL) for GLMM with random effect.

```
proc glimmix data=tlc method=RSPL;
class id group;
model y = group time group*time / dist=poisson link=log solution;
random intercept / type=un sub=id;
run;
```

The MODEL statement is required in every model. The option method in the PROC GLIMMIX statement specifies the estimation method. The PQL is obtained with the option method=RSPL and the MQL is obtained with the option method=RMPL. The default is method=RSPL.

Both Table 5.1 and 5.2 result for fixed effects parameter estimates for fitting random intercept model in both MQL and PQL methods shows that group and time effects were significantly different at 5% significance level. The results also show that the interaction effects between group and time were not significantly different at 5% significance level. Table 5.3 shows the result of the random intercept model for the estimated variance component and residual variance estimate. The result for fitting the random intercept model using both PQL and MQL methods with unstructured covariance structure indicates the two variance components namely the random intercept and the measurement error variances are estimated as 0.04416 and 1.7687 with standard errors given by 0.009454 and 0.1455 respectively in the PQL approximation while under MQL approximation the estimates are 0.0475 and 1.7661 with standard errors given by 0.01005 and 0.1461 respectively. We note that the random intercept variance component is underestimated under PQL approximation compared MQL approximation.

Table 5.1: Parameter estimates, standard errors and p-values for fixed effects under random intercept model

	PQL			MQL		
Parameter	Estimates	SE	P-value	Estimates	SE	P-value
Intercept	3.2360	0.041	< .0001	3.2463	0.041	< .0001
group	-0.2413	0.061	< .0001	-0.2315	0.061	0.0002
time	-0.01515	0.008	0.0584	-0.01515	0.008	0.0582
time*group	-0.00830	0.012	0.4949	-0.00830	0.012	0.4946

Table 5.2: Tests of Fixed Effects for the random intercept model

	PQL		MQL	
Effect	F-value	P-value	F-value	P-value
group	15.92	< .0001	14.20	0.0002
time	10.10	0.0016	10.11	0.0016
time*group	0.47	0.4949	0.47	0.4946

Table 5.3: Estimates variance component for random intercept model under PQL and MQL

		PQL		MQL	
Structure	Cov Parm	Estimates	Std Error	Estimates	Std Error
	UN(1,1)	0.04416	0.009454	0.04757	0.01005
	Residual (VC)	1.7687	0.1455	1.7661	0.1461

In the case of random slope model, similar analysis was done. Table 5.4 shows the results of the analysis of fixed effects under random slope models using the unstructured covariance structure and shows the solution of fixed effects that was a parameter estimates measuring the covariates effects. Table 5.4 shows the result for fixed effects parameter estimates for fitting random slope model in both MQL and PQL methods and indicates that the group effect was significantly different at 5% significance level. The results also show that the time and interaction effects between group and time were not significant at 5% level of significance level. Table 5.5 shows the Type III analysis for fixed effects. The result for Type III analysis for the random slope model for both MQL and PQL methods result shows that group and time effects were significantly different at 5% significance level. The results also show that the interaction effect between group and time was not significantly different at 5% level of significance level. Table 5.6 shows the results of an estimation of variance of components namely random slope model and measurement error. The two components of variance of random slope and measurement error variance were estimated as 0.002315 and 2.0121 with standard errors given by 0.000610 and 0.1649 respectively under the PQL approximation. In the MQL approximation of the estimates were 0.002543 and 2.0955 with corresponding standard errors given by 0.000683 and 0.1742. Also in the random slope model we note that the variance component is underestimated under PQL approximation as compared to the MQL approximation.

Table 5.4: Parameter estimates, standard errors and p-values for fixed effects under random slope model

	PQL			MQL		
Parameter	Estimates	SE	P-value	Estimates	SE	P-value
Intercept	3.2474	0.03034	$< .0001$	3.2463	0.03097	$< .0001$
group	-0.2298	0.04594	$< .0001$	-0.2315	0.04690	$< .0001$
time	-0.01737	0.01091	0.1146	-0.01515	0.01123	0.1806
time*group	-0.01164	0.01622	0.4737	-0.00830	0.01666	0.6187

Table 5.5: Tests of Fixed Effects for the random slope model

	PQL		MQL	
Effect	F-value	P-value	F-value	P-value
group	25.03	$< .0001$	24.38	$< .0001$
time	8.18	0.0052	5.37	0.0226
time*group	0.51	0.4737	0.25	0.6187

Table 5.6: Estimates of variance component for random slope model under PQL and MQL

		PQL		MQL	
Structure	Cov Parm	Estimates	Std Error	Estimates	Std Error
	UN(1,1)	0.002315	0.000610	0.002543	0.000683
	Residual (VC)	2.0121	0.1649	2.0955	0.1742

In the final analysis of the TLC data was made to model both random intercept and slope. The model was fitted without the interaction between group and time under MQL methods because the analyses were not converged when this interaction were included in the model. The convergence was only possible under different covariance structure for variance components and MQL approximation when the interaction between group and time was not included in the model. The results show that when we fit both the random intercept and slope in the same model, we find that the PQL method does perform better than the MQL under some number of covariance structures. Table 5.7 and Table 5.8 show the results for the fixed effects analysis that allow both a random intercept and slope also show that the group and time effects were significantly different at 5% significance level under both PQL and MQL methods. The interaction was not significant at 5% significance level under PQL and under MQL methods does not appear on the result since the model does not converge.

Table 5.7: Parameter estimates, standard errors and p-values for fixed effects under random intercept and slope model

	PQL			MQL		
Parameter	Estimates	SE	P-value	Estimates	SE	P-value
Intercept	3.2458	0.03090	$< .0001$	3.2441	0.03061	$< .0001$
group	-0.2407	0.04627	$< .0001$	-0.2297	0.04601	$< .0001$
time	-0.01967	0.007989	0.0156	0.01873	0.005973	0.0023
time*group	-0.01060	0.01215	0.3841	.	.	.

Table 5.8: Test of fixed effects for the random intercept and slope model

	PQL		MQL	
Effect	F-value	P-value	F-value	P-value
group	27.07	$< .0001$	24.92	$< .0001$
time	16.90	$< .0001$	9.83	0.0023
time*group	0.76	0.3841	.	.

Table 5.9: Estimates of variance component for random slope model under PQL and MQL

		PQL		MQL	
	Cov Parm	Estimates	Std Error	Estimates	Std Error
Structure	UN(1,1)	0.007326	0.008382	0.007284	0.008399
	UN(2,1)	0.008570	0.001704	0.01030	0.002076
	UN(2,2)	2.57E-19	.	5.87E-19	.
	Residual (VC)	1.7732	0.1463	1.7425	0.1438

From the unstructured covariance structure specification for both PQL and MQL methods respectively, we note that the elements of the matrix $D = Var(u_i)$ is estimated as

$$\hat{D} = \begin{pmatrix} 0.007326 & 0.008570 \\ 0.008570 & 2.57E-19 \end{pmatrix} \quad \text{and} \quad \hat{D} = \begin{pmatrix} 0.007284 & 0.01030 \\ 0.01030 & 5.87E-19 \end{pmatrix}$$

which indicates a positive correlation between random intercept and slope. From the table we find that the within-subject error variance (Residual, 1.7732 and 1.7425) is big relative to the between-subject variance of the intercepts (the UN(1,1) term, which is V ar(b1i) = G11 = 0.007326 and 0.007284) in both PQL and MQL methods respectively.

5.10 Estimation via NLMIXED

We focus on the use of PROC NLMIXED procedure to fit GLIMMIX to longitudinal data. PROC NLMIXED in SAS is a very flexible procedure for fitting non-linear mixed effects models. PROC NLMIXED directly maximizes an approximate integrated likelihood via numerical quadrature. PROC NLMIXED has an option for a number of quadrature points using during evaluation of integrals, e.g. QPOINTS=30 specifies that 30 quadrature points are to be used for each random effects.

The commands to fit the model in SAS code for GLMM with SAS code (NLMIXED) with random effect.

```
/*NLMIXED: Adaptive Gaussian Quadrature
proc nlmixed data=tlc qpoints=20 maxiter=100 technique=newrap cov ecov;
parms beta0=1.6 beta1=0.0 beta2=0.4 beta3=0.5 sigma=3.9;
eta=beta0 + beta1*time +beta2*group +beta3*time*group+b1;
mu=exp(eta);
```

```
MODEL y POISSON(mu);
random b1  NORMAL (0,sigma**2) SUBJECT=id;
run;
```

and the commands to fit the model for GLMM with Gaussian quadrature

```
/*NLMIXED: Gaussian Quadrature
proc nlmixed data=tlc qpoints=20 noad maxiter=100 technique=newrap cov ecov;
parms beta0=1.6 beta1=0.0 beta2=0.4 beta3=0.5 sigma=3.9;
eta=beta0 + beta1*time +beta2*group +beta3*time*group+b1;
mu=exp(eta);
MODEL y POISSON(mu);
random b1  NORMAL (0,sigma**2) SUBJECT=id;
run;
```

The default method in PROC NLMIXED for computing this integral is adaptive Gaussian quadrature. The NOAD option in the PROC NLMIXED statement requests nonadaptive Gaussian quadrature. The QPOINTS=$n > 0$ statement specifies the number of quadrature points to be used during evaluation of integrals. The PARMS statement identifies the unknown parameters and their starting values. The MODEL statement defines the dependent variable and its conditional distribution given the random effects. Here a Poisson conditional distribution is specified with mean μ. The RANDOM statement defines the single random effect to be b1, and specifies that it follows a normal distribution with mean 0 and variance $sigma^{**2}$.

The result for fitting the model using PROC NLMIXED Gaussian quadrature and adaptive Gaussian quadrature assuming random intercept model is given in Table 5.10.

Table 5.10: Estimates for Gaussian quadrature and Adaptive Gaussian quadrature

Gaussian quadrature									
Q-Points	Q=10			Q=30			Q=50		
Parameter	Estimates	SE	$Pr > \|t\|$	Estimates	SE	$Pr > \|t\|$	Estimates	SE	$Pr > \|t\|$
β_0	3.2279	0.03895	$< .0001$	3.2262	0.03838	$< .0001$	3.2279	0.03895	$< .0001$
β_1	-0.01515	0.005994	0.0131	-0.01515	0.005994	0.0131	-0.01515	0.005994	0.0131
β_2	-0.2481	0.05655	$< .0001$	-0.2470	0.05655	$< .0001$	-0.2481	0.05655	$< .0001$
β_3	-0.00830	0.009133	0.3657	-0.00830	0.009133	0.3657	-0.00830	0.009133	0.3657
σ	0.2295	0.02017	$< .0001$	0.2301	0.01952	$< .0001$	0.2295	0.02017	$< .0001$
-2ℓ	2745.2			2745.1			2745.2		
Adaptive Gaussian quadrature									
Q-Points	Q=10			Q=30			Q=50		
Parameter	Estimates	SE	$Pr > \|t\|$	Estimates	SE	$Pr > \|t\|$	Estimates	SE	$Pr > \|t\|$
β_0	3.2279	0.03895	$< .0001$	3.2279	0.03895	$< .0001$	3.2279	0.03895	$< .0001$
β_1	-0.01515	0.005994	0.0131	-0.01515	0.005994	0.0131	-0.01515	0.005994	0.0131
β_2	-0.2481	0.05655	$< .0001$	-0.2481	0.05655	$< .0001$	-0.2481	0.05655	$< .0001$
β_3	-0.00830	0.009133	0.3657	-0.00830	0.009133	0.3657	-0.00830	0.009133	0.3657
σ	0.2295	0.02017	$< .0001$	0.2295	0.02017	$< .0001$	0.2295	0.02017	$< .0001$
-2ℓ	2745.2			2745.2			2745.2		

The result in Table 5.10 indicates that there is no difference in parameter estimates of Gaussian quadrature and adaptive Gaussian quadrature. This takes into consideration in each log-likelihood corresponds to the maximum of the approximation to the model involving the log-likelihood corresponding to different quadrature points are not necessarily comparable. This means that difference in log-likelihood value reflects the difference in the quality of numerical approximation and thus higher log-likelihood value does not necessarily correspond to better approximation which happens when we choose a large value of QPOINTS to increase the accuracy of the numerical integration algorithm. The standard errors are approximately much closer as those obtained using GLIMMIX procedure in the random intercept in Table 5.1. The adaptive Gaussian quadrature methods give estimates much closer to the quasi-likelihood approximation under GLIMMIX than the Gaussian quadrature points. The standard errors under adaptive Gaussian quadrature and Gaussian quadrature remain the same as the number of quadrature point increases.

5.11 Summary

Generalized linear mixed model extends the concept approach represented by the linear mixed effect model. It assumes natural heterogeneity across individuals in subsets of the regression coefficients. The focus of GLMMs is on inference about individuals. The regression parameters (β) have subject-specific interpretations in terms of change in the transformed mean response for a specific individual. Generalized linear mixed models have been implemented in the SAS procedures PROC GLIMMIX and PROC NLMIXED. Both procedures approach parameter estimation as an optimization problem, which solves for an approximation of the marginal log-likelihood. PROC NLMIXED accomplishes this using an integral approximation through Gaussian, whereas PROC GLIMMIX relies on approximation of linear nixed models. PROC NLMIXED directly maximizes an approximate integrated likelihood via numerical quadrature. The present study show that the likelihood approximation may not be accurate if too few quadrature points are used. The results show that using different Q can lead to considerable differences in estimates and standard errors. For example, using non-adaptive quadrature Q=3, we found no difference in the week effect between both group ($\beta = -0.152$, SE=0.059, p-value=0.081). Using adaptive quadrature with Q=50, we found a non-significant interaction between the week and group ($\beta = -0.008$, SE=0.009, p-value=0.3657).

Chapter 6

Theory of Correspondence Analysis (CA)

6.1 Historical Background of Correspondence Analysis

According to Zhou (2008) correspondence analysis (CA) is a universally popular data analysis method. In France, CA was developed under the strong influence of Jean-Paul Benzecri and in Japan it was developed under Chikio Hayashi. The name correspondence analysis is a translation of the French *analyse des correspondances*. According to Zhou (2008) and Doey and Kurta (2011) it has had many other names, including optimal scaling, reciprocal averaging, optimal scoring, and appropriate scoring in the United States; quantification method in Japan; homogeneity analysis in the Netherlands; dual scaling in Canada; and scalogram analysis in Israel.These names are thought to stem from the fact that CA has been used to analyze different questions and has therefore been given a different name each time to answer a different question. CA is described in more detail in French by Benzecri (1973) and Lebart et al. (1977). In Japanese, the subject is described in Kobayashi (1981) and Komazawa and Hayashi (1982). In English, CA is described in Ludovic et al. (1984), Greenacre (1984), and Greenacre and Hastie (1987) just to name a few. The variation of development and application of CA ranges from various fields, such as biometry, psychometrics, linguistics, health care and science (Zhou, 2008). Therefore, CA can be considered as a very flexible method of data analysis in all situations where an exploratory or more in-depth analysis of categorical data is required.

6.2 Introduction

Mazzarol and Soutar (2008) state that correspondence analysis is an exploratory data analytic technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence in the rows and columns. The results give information that is similar in nature to those produced by Factor Analysis techniques and they allow one to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency cross-tabulation table. Correspondence analysis is a statistical technique that provides a graphical representation of cross tabulations or contingency tables. De Leeuw and Mair (2007) propose that contingency tables arise at any time it is possible to assign events into two or more different sets of categories, such as product and location for purchases in market research or symptom and treatment in medical testing. Correspondence analysis has become most popular in fields such as ecology where data is collected on the bunch of various animal species in specific sampling units and also in market research because researches in this area frequently collect categorical data due to the simplicity of this collection method (Ter Braak, 1986). Correspondence analysis was initially recommended as an inductive method for analyzing semantic data and was considered as a standard, unifying and integrated analysis framework (Murtagh, 2005). CA is used to analyze research questions across many fields (Greenacre, 1993). CA can be used as an exploratory data technique for categorical data since ecologists have been able to transform these complicated tables into straightforward graphical displays. Ecologist data is multidimensional thus making visualization of more than two dimensions difficult and CA is analyzes this form of data because of its ability to extract the most important dimensions, allowing simplification of the data matrix (Doey and Kurta, 2011). Murtagh (2005) gives two explanations that contribute to its success:

- The idea of distributional equivalence allows a table of positive values to be given a mathematical structure that compensates as far as possible for randomness in the choice of weighting and subdivision of categories.
- The great number of data analysis available for working in very different application fields is able to merge processing frameworks in a single software package.

Greenacre (1984) states that correspondence analysis has several features that distinguish it from other techniques of data analysis. An important feature of correspondence analysis is the multivariate treatment of the data through simultaneous consideration of multiple categorical variables. The nature of multivariate of correspondence analysis can unveil relationships that would not be detected in a series of pair wise comparisons of variables. Another important feature is the graphical display of row and column points in biplots which can help in detecting structural relationships among the variable categories and objects. Finally, correspondence analysis has highly flexible data requirements. The strict data requirement is a rectangular data matrix with non-negative entries. Correspondence analysis is most effective if the following conditions are satisfied:

- The data matrix is large enough, so that optical inspection or simple statistical analysis cannot unveil its structure.
- The variables are homogeneous, so that calculating the statistical distances between the rows or columns will makes sense.
- The data matrix is a rational unstructured, i.e., its structure is either unknown or poorly understood.
- Normalization procedures in CA can be used to determine whether and how similarity of the row and column variables, as well as the relationship between them, can be interpreted in terms of row and column coordinates and the origin of the plot (Nilsson, 2011).

A well-defined advantage of correspondence analysis over other methods generating joint graphical displays is that it produces two double or dual displays whose row and column geometries have similar interpretations, facilitating analysis and detection of relationships. According to Zhou (2008) and Nilsson (2011) this technique is similar to principal component analysis, but it is better suited for analyzing categorical data. Principal component analysis, on the other hand, is better suited for continuous data (Lebart et al., 1984). Another difference between the two techniques is how the data matrix is decomposed. While the total Chi-square value is decomposed in CA, total variance is decomposed in principal component analysis (Zhou, 2008). In short, CA involves mapping a Chi-square distance into a particular Euclidean distance. When these converted points are plotted, as the distance between two points get closer, the similarities between their profiles increase (De Leeuw and Mair, 2007 and Zhou, 2008). The aim is to have different but complementary analytic tools to facilitate interpretation of the data (Murtagh, 2005). The primary goal for CA is to transform a table of numerical information into a graphical display, in which each row and each column is represented as a point.

6.3 Advantages of Correspondence Analysis

Some of the advantages include:

- CA can simplify complex data from a potentially large table into a simpler display of categorical variables while presenting all of the valuable information in the data set. Correspondence analysis is useful when other statistical techniques cannot be used to analyze data because of certain assumptions that are met due to its flexible data requirements. An example from Doey and Kurta (2011) is that when a Likert scale is used to collect data and the spaces between descriptors, i.e. ″never″, ″sometimes″ and ″often″ are not necessary equivalent, then CA is a useful technique because it focuses mainly on how variables correspond to another and whether there is a significant difference between these variables.

- Another benefit of CA is that when one wishes to analyze continuous data with CA, the data can be categorized and subsequently analyzed as discrete data because CA demonstrates how variables are associated by the approximate distance of points to one another on the biplot and not simply that they are associated.

- CA reveals relationships that would not be identified using other non-multivariate statistical techniques such as performing pairwise comparisons yet CA represents data using two dual displays which are the display for the row data and display for the column data.

- CA is also good way to examine data validity and facilitates the treatment of outliers.

6.4 Assumption of Correspondence Analysis

Doey and Kurta (2011) state that when we disobey the following assumptions, we can make the conclusions drawn about the association among variables imprecise and the biplot a less variable guide for analyzing the data

- The homogeneity of variance across row and column variable must be met which assumes that the statistical properties are similar across rows and columns. For example, there must not be any empty variables

- CA assumes that the data being analyzed is discrete however the original continuous variables can be categorized into discrete variables.

- The data should be made up of more than three categories otherwise if CA is used to analyze only two or three categories this analysis is unlikely to be more informative than the original table.

- All values in the frequency table must be non-negative so that distances between the points on the biplot are always positive.

We also note that CA does not make distributional assumptions.

6.5 Notation

By using the ideas of Zhou (2008) and Nilsson (2011), we let F be a matrix of frequencies with size $(I \times J)$ and with elements (f_{ij}). F is a matrix composed of non-negative values with row and column sums which are non-zero and this matrix has rank q. The correspondence matrix P with size $(I \times J)$ is defined as a matrix where all elements in F are divided with grand total n, which is $P = \frac{1}{n}F$. Furthermore, let the vectors of row and column sums be denoted as r and c, which can also be expressed a centroids of row and column clouds in their respective spaces r and c, with row centroid defined as $c = R^{'}r$ and column centroid as $c = C^{'}c$. Let 1 be a rows vector of ones and I be an identity matrix, each of appropriate order. Denote a matrix-valued function that creates a diagonal matrix from a vector by diag(.). Define,

- $N = 1'F1$ as the sum of all elements in F;
- $P = \frac{1}{n}F$ as the matrix of relative frequencies (the correspondence matrix);
- $r = P1$ as the vector of row marginal proportions (row masses);
- $D_r = diag(r)$ as the diagonal matrix of row masses;
- $D_c = diag(c)$ as the diagonal matrix of column masses;
- $R = D_r^{-1}P^{'}$ as the row profile; and
- $C^{'} = D_c^{-1}P^{'}$ as the column profile.

The scalar n is the sum of all elements in F. The matrix P is a matrix of relative frequencies. The vector r contains row marginal proportions or row masses. The vector c contains column marginal proportions or column masses. The matrices D_r and D_c are diagonal matrices of marginals. The rows of R contain the row profiles. The elements of each row of R sum to one. Each (i,j) element of R contains the observed probability of being in column j given membership in row i. Similarly, the columns of C contain the column profiles.

6.6 Basic Concepts and Definitions

There are certain fundamental concepts in correspondence analysis which are described below:

6.6.1 Primitive Matrix

The data matrix or contingency table $N(I,J)$ is called the primitive matrix or primitive table. The elements of this matrix are n_{ij}.

6.6.2 Row-profile and Column-profile Tables

According to De Leeuw and Mair (2007), Greenacre (1984) and Beh (2004) in interpreting a contingency tables, we compute the conditional frequencies so that it can make sense to compare the actual frequencies in each cell. Each row and each column has a different number of respondents, called the base of respondents. It is necessary to reduce either the rows or columns to the same base for comparison. Let us consider a contingency table of $N(I,J)$ with I rows $(i = 1,2,\ldots,I)$ and J columns $(j = 1,2,\ldots,J)$ having frequencies n_{il}. Marginal frequencies are denoted by n_{i+} and n_{+j}

$$n_{i+} = \sum_j n_{ij} \qquad \text{and} \qquad n_{+j} = \sum_j n_{ij} \tag{6.1}$$

The total frequency is given by

$$n = \sum_j \sum_i n_{ij} \tag{6.2}$$

6.6.3 Row Profiles

Let $f_{j|i}$ denote the conditional frequencies associated to row profiles. The profile of each row i is a vector of conditional densities is given by

$$f_{j|i} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i+}}{n}} = \frac{n_{ij}}{n_{i+}} \qquad j = 1,2,\cdots,J \tag{6.3}$$

The matrix of row profile of the complete set may be denoted by $(I \times J)$ matrix **R**

6.6.4 Column Profiles

Let $f_{i|j}$ denote the conditional frequencies associated to column profiles. The profile of each column j is a vector of conditional densities given by

Table 6.1: Matrix of Row Profiles

Rows	Columns		Row Mass
	1	2 $\cdots$ j	
1	n_{11}/n_{1+}	$n_{12}/n_{1+}\cdots n_{1j}/n_{1+}$	1
2	n_{21}/n_{2+}	$n_{22}/n_{2+}\cdots n_{2j}/n_{2+}$	1
$\vdots$	$\vdots$	$\vdots$ $\vdots$	$\vdots$
I	n_{i1}/n_{i+}	$n_{i2}/n_{i+}\cdots n_{ij}/n_{i+}$	1
Columns Mass	n_{+1}/n_{++}	$n_{+2}/n_{++}\cdots n_{+j}/n_{++}$	1

$$f_{j|i} = \frac{\frac{n_{ij}}{n}}{\frac{n_{+j}}{n}} = \frac{n_{ij}}{n_{+j}} \qquad i = 1, 2, \cdots, I \tag{6.4}$$

The matrix of column profile of the complete set may be denoted by $(I \times J)$ matrix $\mathbf{C}$

Table 6.2: Matrix of Column Profiles

Rows	Columns		Row Mass
	1	2 $\cdots$ j	
1	n_{+1}/n_{+1}	$n_{12}/n_{+2}\cdots n_{1j}/n_{+j}$	n_{+1}/n_{++}
2	n_{+1}/n_{+1}	$n_{22}/n_{+2}\cdots n_{2j}/n_{+j}$	n_{+2}/n_{++}
$\vdots$	$\vdots$	$\vdots$ $\vdots$	$\vdots$
I	n_{i1}/n_{+1}	$n_{i2}/n_{+2}\cdots n_{ij}/n_{+j}$	n_{+i}/n_{++}
Columns Mass	1	1 $\cdots$ 1	1

Average row profile

$$\bar{r} = n_{+j}/N \qquad (j = 1, 2, \cdots, J) \tag{6.5}$$

Average column profile

$$\bar{c} = n_{i+}/N \qquad (i = 1, 2, \cdots, I) \tag{6.6}$$

6.6.5 Masses or Weights

Another fundamental concept in correspondence analysis is the concept of mass. This concept is also based on rows and column of the contingency table. According to Dufour (2008) and Greenacre (2002) the mass is the proportion of the whole table that is in the category represented by the row or column. Mass is the ratio of the row or column count to the total table count. The mass of the i^{th} row which can be denoted as m_{r_i} is equal to marginal frequency of the i^{th} row divided by the grand total. Therefore

$$m_{r_i} = n_{+i}/n \tag{6.7}$$

Similarly, the mass of the j^{th} column which can be denoted as m_{c_j} is equal to marginal frequency of the j^{th} row divided by the grand total. Therefore

$$m_{c_j} = n_{j+}/n \tag{6.8}$$

6.6.6 Distances

According to Greenacre (2007) and Mazzarol and Soutar (2008) a variant of Euclidean distance is called the weighted Euclidean distance which is used to measure and so that design the distances between profile points. In this case the weighting assigns to differential weighting of the dimension of the space and not to the weighting of the profiles. The distance between two rows i and i' is given by

$$d^2(i,i^{'}) = \sum_{j-1}^{J} \frac{1}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{ij}}{n_{i'+}} \right)^2 \tag{6.9}$$

and the distance between two column j and $j^{'}$ is given by

$$d^2(j,j^{'}) = \sum_{i-1}^{I} \frac{1}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij}}{n_{j'+}} \right)^2 \tag{6.10}$$

Thus the distance that is obtained is called the Chi-squared distance. This differs from the Euclidean distance in that each squared is weighted by the inverse of the frequency corresponding to each term. According to Greenacre (2007) the analysis of each squared by the expected frequency is called variance standardizing and commits for the large variance in high frequencies and the smaller variance in low frequencies. If there is no standardization achieved then the differences between larger populations would contribute to be large and thus manage the distance calculation while the differences between the smaller proportions would lead to be submerged. The weighting factors are used to equalize these differences. The fundamental reason for choosing the Chi-square distance is that it satisfies the principle of distributional equivalence Mazzarol and Soutar (2008) expressed as follows:

- If two rows i and $i^{'}$ of $N(I,J)$ are proportioned and if they are replaced by only one which is the sum column by column then the distance between columns does not change in $N(J)$.

- If two columns j and $j^{'}$ of $N(I,J)$ are proportioned and if they are replaced by only one which is the sum row by row then the distance between columns does not change in $N(I)$.

6.7 Inertia

According to Mazzocchi (2008) inertia is a measure of association between two categorical variables based on the Chi-squared statistic. In correspondence analysis the proportion of inertia interpreted by each of the dimensions can be observed as a measure of goodness-of-fit because the capability of correspondence analysis depends on the degree of association between row and column. According to De Leeuw and Mair (2007) the term inertia is taken from the term "moment of inertia" in mechanics where the physical object has a center of gravity and every particle of an object has a certain mass (m) and distance (d) from the center of gravity. Then the moment of inertia of an object is the quantity md^2 that are summed in all the particles that create the object

$$moments\ \ of\ \ inertia = \sum md^2 \tag{6.11}$$

This theory has an equivalence or correspondence in correspondence analysis. In correspondence analysis there is a set of elements of profile points with masses adding up to 1 and these points have the average profile and distance between profile points. Each profile contributes to the inertia of the whole set of elements. The inertia is the sum of squares of the singular values or the sum of the eigenvalues and is given by

$$\text{inertia} = \sum_{k=1} \alpha_k^2 = \sum_{k=1} \lambda_k^2 \tag{6.12}$$

The inertia of a profile point is computed by the following formulas:

For the i^{th} row profile

$$\text{inertia} = m_i \sum_j \frac{(r_{ij} - \bar{r}_j)^2}{\bar{r}_j} \tag{6.13}$$

where r_{ij} is the ratio of $\frac{n}{n_{i+}}$ and $\bar{r}_j$ is $\frac{n_j}{n}$ and for the j^{th} column profile

$$\text{inertia} = m_j \sum_i \frac{(c_{ij} - \bar{c}_i)^2}{\bar{c}_i} \tag{6.14}$$

The total inertia of the contingency table is given by

$$\text{Total inertia} = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{6.15}$$

which is the Chi-square statistic that is divided by n. Total inertia measure the overall association between row and column and it also equals to the sum of the eigenvalues. It corresponds to the Chi-square value divided by the number of observations (Mazzocchi, 2008).

6.7.1 Link between CA and the Chi-square Statistic

According to Dufour (2008) if we let I_t be the total inertia and χ^2 be the value of Chi-square statistic calculated from the contingency table then total inertia is given by

$$I_t = \frac{\chi^2}{n} \quad \text{where} \quad \chi^2 = \sum \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}} = \sum \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \tag{6.16}$$

where O_{ij} is the count of row i and column j of the table, E_{ij} is the value expected under the assumption of row-by-column independence, and n is the total table count.

6.7.2 Contributions of Points to Principal Point Inertias

The contributions of the row and columns points to the inertia on the k-th dimension are the inertia components:

$$\text{for row } i: \qquad \frac{r_i f_{ik}^2}{\lambda_k} = r_i \phi_{ik}^2 \tag{6.17}$$

and

$$\text{for column } j: \qquad \frac{c_j g_{jk}^2}{\lambda_k} = c_i \gamma_{jk}^2 \tag{6.18}$$

where $f_{ik} = \sqrt{\lambda_k}\phi_{ik}$ and $g_{ik} = \sqrt{\lambda_k}\gamma_{ik}$ the relationship between principal and standard coordinates that are exactly the coordinates proposed for the standard correspondence analysis biplot which shows that the squared lengths of these coordinates are the contributions to the principal axes (Mazzarol and Soutar, 2008).

6.7.3 Contributions of Principal Axes to Point Inertias

The contributions of the dimensions to the inertia of the $i-th$ and $j-$th column points

$$\text{for row } i: \qquad \frac{f_{ik}^2}{\sum_k f_{ik}^2} \qquad \text{and for column } j: \qquad \frac{g_{jk}^2}{\sum_k g_{ik}^2} \tag{6.19}$$

where the denominators are the squared χ^2-distances between the corresponding profile point and the average profile. The summary statistics of the row and column points includes mass, contribution to inertia, inertia and squared cosines which is a summary statistics for all subsection of section 6.6. The formulas to compute these statistics are given in the Table 6.3 below.

Table 6.3: Row and Column Summary Statistics

Summary Statistic	Formula
Row mass	r
Column mass	c
Row partial contribution to inertia	$D_c^{-1}sq(A)$
Row partial contribution to inertia	$D_c^{-1}sq(B)$
Column inertia	$(\frac{1}{T})D_r^{-1}sq(AD_u)1$
Column inertia	$(\frac{1}{T})D_r^{-1}sq(BD_u)1$
Row squared cosine	$diag[sq(AD_u)1]^{-1}sq(AD_u)$
Column squared cosine	$diag[sq(BD_u)1]^{-1}sq(BD_u)$

According to Zhou (2008) and Nilsson (2011) one of the most important advantages of CA is to represent the high dimensional categorical data set into a low dimensional space. The plot consists of two overlaid plots, one for row points and the other for the column points. The row and column points are row and column profile respectively, which is rescaled so that distance between profiles can be displayed as ordinary Euclidean distance, then orthogonally rotated to a principal axes orientation. Thus, once the rescaled row and column coordinates have been calculated by the formulas in Table 6.3, they can often scatter plotted in a two dimensional space. The squared singular value, which is the square of the diagonal entries of the D_u matrix, are plotted versus each dimension index in the fit plot. The residual versus the centered frequencies may also be plotted if necessary (Zhou, 2008). The centered data are calculated by the formula $P - rc^{'}$ which subtracts the expected relative frequencies from these centered frequencies by

$$(P - rc^{'}) - A^m D_u^m B^m$$

where m refers to the fact that only m of the dimensions are involved in the calculation. Note that distance between row points or distance between column points have meaning, however, distance between row and column points are not well interpretable (Zhou, 2008).

6.8 Basic Computational Algorithm

According to Nilsson (2011) a central theme in correspondence analysis is singular value decomposition (SVD) which revolves around the concept of dimension reduction of a data set. The aim of CA is to find a low dimensional approximation of the data set to represent the row and column profiles that the dimension should be $min\{I,J\}-1$. According to Zhou (2008) the required row and column coordinates generated from CA are based on the generalized SVD of the relative frequency matrix P,

$$P = AD_uB^{'}$$

where

- A is an $(n \times q)$ the eigenvectors (matrix whose columns are the left generalized singular vectors)
- D_u is a $q \times q$ diagonal matrix of generalized singular values
- B is an $m \times q$ matrix whose column are the right generalized singular vectors
- $A^{'}D_r^{-1}A = B^{'}D_r^{-1}B = I$

According to Zhou (2008) there is a insignificant part of the generalized SVD of P consisting of a singular value of 1 and associated left and right singular vectors, which is discarded before any results are displayed. The remaining left and right singular vector defines the orthogonal principal axes of the column and row points respectively. In practice, the generalized SVD is computed indirectly by performing an ordinary SVD, where the ordinary SVD of any matrix Q is given by

$$Q = UD_uV^{'}$$

under the following relation or constraint $U^{'}U = V^{'}V = I$. Therefore, the following steps can be performed in order to compute the generalized SVD of the correspondence matrix P,
Step 1: Calculate the matrix S standarized residual

$$S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}} \tag{6.20}$$

where $D_r = diag(r)$ and $D_c = diag(c)$ are diagonal matrices of row and column masses respectively, $\mathbf{P} = \{p_{ij}\}$ is a correspondence analysis, $r = \{r_i\}$ and $\mathbf{c} = \{c_j\}$ are row and column masses whose elements add up to one in each case. The row and column masses respectively is given by

$$r_i = \sum_{j=1}^{J} p_{ij} \qquad \text{and} \qquad c_j = \sum_{i=1}^{I} p_{ij} \tag{6.21}$$

and using standard matrix notation, we write row and column masses respectively as

$$\mathbf{r} = \mathbf{P1} \qquad \text{and} \qquad \mathbf{c} = P^T\mathbf{1} \tag{6.22}$$

and the correspondence analysis is given by

$$\mathbf{P} = \frac{1}{n}\mathbf{N} \tag{6.23}$$

where **N** denote the $(I \times J)$ data matrix with positive row and column sums and the elements of P is given by

$$P = \{p_{ij}\} \quad \text{where} \quad p_{ij} = \frac{n_{ij}}{n} \tag{6.24}$$

Step 2: Compute the ordinary SVD of S:

$$S = UD_uV^T \quad \text{where} \quad U^TU = V^TV = I \tag{6.25}$$

where D_α is the diagonal matrix of positive singular values in descending order $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \alpha_4 \geq \cdots$, U and V are the left and right singular vectors respectively.

Step 3: Standard coordinates A of rows and B of columns, respectively:

$$A = D_r^{-\frac{1}{2}}U \qquad \text{and} \qquad B = D_c^{-\frac{1}{2}}V \tag{6.26}$$

Step 4: Principal coordinates F of rows and F of columns, respectively:

$$F = D_r^{-\frac{1}{2}}UD_u = AD_u \qquad \text{and} \qquad G = D_c^{-\frac{1}{2}}UD_u = BD_u \tag{6.27}$$

Step 5: Principal inertias λ_k:

$$\lambda_k = \alpha_k^2, \qquad k = 1, 2, \ldots, K \text{ where} K = min\{I-1, J-1\} \tag{6.28}$$

Then $P = AD_uB^{'}$ is the generalized SVD.

6.9 Interpretation of Correspondence Analysis

Mazzarol and Soutar (2008) and Doey and Kurta (2011) state that the interpretation of the results of correspondence analysis involves the interpretation of numerical results and factor graphics flexible by CA. The former implies selection of significant axes and significant points.

6.9.1 Selection of Significant Axes

Mazzarol and Soutar (2008) state that in the selection of significant axes we consider only two types of factor axes which are first order factor axes and second order factor axes. The first order factor axes are considered on the basis of contribution to the total inertia while the second order factor axes are considered on basis of contributions to the unusualness or eccentricity which is $\cos^2 \varphi$.

First order factor axes

Let M be the number of significant axes that can decisive by any of the following rules:

- Sum of the inertia explained by the first M axes exceeds a certain edge, mostly 80% of the inertia
- Single out all the axes whose eigenvalues exceed $\frac{1}{[min(I-1,J-1)]}$

Second order factor axes

According to Mazzarol and Soutar (2008) once the first factor axes has been selected then the second order factor axes are selected as follows. We first let $M^{'}$ be the rank of a factor axis for point i of $N(I)$ or j for $N(J)$ exists such that

$$\cos^2 \varphi(i) \geq k \qquad \text{or} \qquad \cos^2 \varphi(j) \geq k \tag{6.29}$$

where k is normally equal to 0.25
Hence, the number of axes that are selected for interpretation is equal to $M + M^{'}$

6.10 Application of CA

The application of CA was applied to Respiratory Syncytial Virus data (RSV). The RSV data was described in Chapter 4. The analysis was done in SAS. The CORRESP procedure was used to perform the CA to find a low-dimensional graphical representation of the matrices of Respiratory Syncytial Virus data. ODS GRAPHICS procedure is used to plot the row and column points where the row points are row profiles which are rescaled so that the distance between profiles can be displayed as ordinary Euclidean distances and then orthogonally related to a principal axes orientation. Similar process is done to the column points. A distance between row points or column points indicates the correspondence between them. The approximate relationship between child age and prevalence of virus in blood in two-dimensional space can be represented through both quantitative and geometric ways.

The commands to fit the model in SAS code for Correspondence Analysis are

```
ods graphics on;
* Perform Simple Correspondence Analysis;
proc corresp all data=spha outc=Coor;
tables age, prev;
run;
ods graphics off;
```

Table 6.4 indicates that there is no significant association between child age and prevalence of virus in blood because of either the constantly small change of Percentage of Chi-square decomposition or the constantly increasing cumulative chi-square decomposition.

Table 6.5 shows the tables that summarize the CA for the row variable (child age). The ″Row Coordinates″ table displays the coordinates of the child age in the joint plot. The row coordinates shows that in the first dimension, all the child age except ***age7*** and ***8*** make little contribution to the total inertia. This may be caused by the low marginal frequency which can be considered as one reason that makes this points deviate from others. The ″Summary Statistic″ table displays various statistics including quality and mass of the presentation. The category with low quality is ***age1 (0.009)*** and is not well presented by the two principal co-ordinates. The categories with high quality are ***age7, 8*** and ***11***. The quality statistic is equal to the sum of the square cosines, which are displayed in Table 6.5. The square cosines are the square of the cosines of the angles between each axis and a vector from the origin to the point. Thus, points with a square cosine near 1 are located near a principal coordinate axis and so have high quality. In our analysis, ***age11*** has high quality. The ″Partial Contribution to Inertia″ table indicates how much of the total inertia is accounted for by each category in each dimension. This table corresponds to the spread of the points in the joint plot in the horizontal and vertical dimensions. In the principal coordinate, ***age10*** and ***age2*** contributes the most. In the second principal coordinate, ***age7*** contribute the most. The ***age7-9*** seems to contribute nothing in principal coordinates.

Table 6.7 shows the tables summarize the CA for the column variable (prevalence). The analysis of the column is similar. The column coordinates clearly show that all the prevalence of virus except ***prev(0.005208)*** make small contribution to the total inertia since their coordinates are all approximately zero. The marginal frequency of ***prev(0.005208)*** may also be the reason for it's deviation. The "Summary Statistics" table with quality of the representation statistic has the same interpretation as in Table 6.6. The "Partial Contributions to Inertia" table which indicates how much of the inertia is accounted for by each category in each dimension and the "Squared Cosine" table also has a similar interpretation as in Table 6.6.

Table 6.4: Inertia and Chi-Square Decomposition of CA for Child Age vs Prevalence of Virus in Blood

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	$---+^{4}----+^{8}----+^{12}----+^{16}----+^{20}---$
1.00000	1.00000	899.00	16.87	16.87	*********************
0.96854	0.93808	843.33	15.83	32.70	********************
0.93956	0.88277	793.61	14.90	47.60	*******************
0.85920	0.73822	663.66	12.46	60.06	****************
0.83061	0.68991	620.23	11.64	71.70	***************
0.73235	0.53634	482.17	9.05	80.75	***********
0.68639	0.47113	423.55	7.95	88.70	**********
0.64562	0.41683	374.73	7.03	95.73	*********
0.50306	0.25306	227.51	4.27	100.00	*****
Total	5.92634	5327.78	100.00		
Degrees of Freedom = 99					

Row Profiles										
	0.00788	0.02385	0.002513	0.003188	0.005208	0.018182	0.021523	0.025424	0.041276	0.047516
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.50000	0.00000	0.00000	0.50000	0.00000
2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.27778	0.72222	0.00000
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.10067	0.88591	0.01342	0.00000
4	0.14935	0.00000	0.00000	0.00000	0.00000	0.00000	0.79221	0.05844	0.00000	0.00000
5	0.78107	0.00000	0.00000	0.17160	0.00000	0.00000	0.04734	0.00000	0.00000	0.00000
6	0.03333	0.00000	0.00000	0.92222	0.04444	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
8	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
9	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	0.00000	0.00000	0.66102	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.33898
11	0.00000	0.3333	3 0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.66667
12	0.00000	0.90164	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.09836

Column Profiles										
	0.00788	0.02385	0.002513	0.003188	0.005208	0.018182	0.021523	0.025424	0.041276	0.047516
1	0.00000	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.01235	0.00000
2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.17544	0.96296	0.00000
3	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.10345	0.77193	0.02469	0.00000
4	0.14557	0.00000	0.00000	0.00000	0.00000	0.00000	0.84138	0.05263	0.00000	0.00000
5	0.83544	0.00000	0.00000	0.25893	0.00000	0.00000	0.05517	0.00000	0.00000	0.00000
6	0.01899	0.00000	0.00000	0.74107	0.08889	0.00000	0.00000	0.00000	0.00000	0.00000
7	0.00000	0.00000	0.00000	0.00000	0.88889	0.00000	0.00000	0.00000	0.00000	0.00000
8	0.00000	0.00000	0.00000	0.00000	0.02222	0.00000	0.00000	0.00000	0.00000	0.00000
9	0.00000	0.00000	0.18750	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
10	0.00000	0.00000	0.81250	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.31250
11	0.00000	0.25676	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.59375
12	0.00000	0.74324	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.09375

Table 6.5: Row and Column Profile of CA for Child Age vs Prevalence of Virus in Blood

Table 6.6: Child Age Results of CA for Child Age vs Prevalence of Virus in Blood

Row Coordinates		
	Dim1	Dim2
1	-0.5108	-1.3325
2	-0.5108	-1.1718
3	-0.5108	-0.9138
4	-0.5108	-0.2089
5	-0.5108	0.4379
6	-0.5108	0.9929
7	-0.5108	3.2729
8	-0.5108	3.2729
9	1.9579	0.0000
10	1.9579	0.0000
11	1.9579	0.0000
12	1.9579	0.0000

Summary Statistics for the Row Points			
	Quality	Mass	Inertia
1	0.0090	0.0022	0.0850
2	0.3146	0.1201	0.1053
3	0.3435	0.1657	0.0892
4	0.1003	0.1713	0.0878
5	0.1663	0.1880	0.0863
6	0.2123	0.1001	0.0992
7	0.5782	0.0445	0.1425
8	0.5782	0.0011	0.0036
9	0.2162	0.0100	0.0299
10	0.4357	0.0656	0.0974
11	0.5814	0.0634	0.0705
12	0.4253	0.0679	0.1032

Partial Contributions to Inertia for the Row Points		
	Dim1	Dim2
1	0.0006	0.0042
2	0.0313	0.1759
3	0.0432	0.1475
4	0.0447	0.0080
5	0.0490	0.0384
6	0.0261	0.1052
7	0.0116	0.5081
8	0.0003	0.0127
9	0.0384	0.0000
10	0.2516	0.0000
11	0.2430	0.0000
12	0.2601	0.0000

Squared Cosines for the Row Points		
1	0.0012	0.0078
2	0.0502	0.2643
3	0.0818	0.2617
4	0.0859	0.0144
5	0.0959	0.0705
6	0.0444	0.1679
7	0.0137	0.5645
8	0.0137	0.5645
9	0.2162	0.0000
10	0.4357	0.0000
11	0.5814	0.0000
12	0.4253	0.0000

Table 6.7: Prevalence Results of CA for Child Age vs Prevalence of Virus in Blood

Column Coordinates		
	Dim1	Dim2
0.00788	-0.5108	0.3658
0.02385	1.9579	0.0000
0.002513	1.9579	0.0000
0.001388	-0.5108	0.8768
0.005208	-0.5108	3.1700
0.018182	-0.5108	-1.3757
0.021523	-0.5108	-0.2541
0.025424	-0.5108	-0.9519
0.041276	-0.5108	-1.2054
0.047516	1.9579	0.0000

Summary Statistics for the Column Points			
	Quality	Mass	Inertia
0.00788	0.1390	0.1758	0.0842
0.02385	0.4686	0.0823	0.1136
0.002513	0.3049	0.0534	0.1133
0.001388	0.2126	0.1246	0.1018
0.005208	0.5966	0.0501	0.1460
0.018182	0.0048	0.0011	0.0842
0.021523	0.1013	0.1613	0.0875
0.025424	0.4070	0.1902	0.0920
0.041276	0.2524	0.0901	0.1032
0.047516	0.6205	0.0712	0.0742

Partial Contributions to Inertia for the Column Points		
	Dim1	Dim2
0.00788	0.0458	0.0251
0.02385	0.3155	0.0000
0.002513	0.2047	0.0000
0.001388	0.0325	0.1021
0.005208	0.0131	0.5362
0.018182	0.0003	0.0022
0.021523	0.0421	0.0111
0.025424	0.0496	0.1837
0.041276	0.0235	0.1395
0.047516	0.2729	0.0000

Squared Cosines for the Column Points		
0.00788	0.0919	0.0471
0.02385	0.4686	0.0000
0.002513	0.3049	0.0000
0.001388	0.0539	0.1587
0.005208	0.0151	0.5815
0.018182	0.0006	0.0042
0.021523	0.0812	0.0201
0.025424	0.0910	0.3160
0.041276	0.0384	0.2139
0.047516	0.6205	0.0000

Table 6.8: Contingency Table of Child Age vs Prevalence of Virus in Blood

Contingency Table										
	0.00788	0.02385	0.002513	0.003188	0.005208	0.018182	0.021523	0.025424	0.041276	0.047516	Sum
1	0	0	0	0	0	1	0	0	1	0	2
2	0	0	0	0	0	0	0	30	78	0	108
3	0	0	0	0	0	0	15	132	2	0	149
4	23	0	0	0	0	0	122	9	0	0	154
5	132	0	0	29	0	0	8	0	0	0	169
6	3	0	0	83	4	0	0	0	0	0	90
7	0	0	0	0	40	0	0	0	0	0	40
8	0	0	0	0	1	0	0	0	0	0	1
9	0	0	9	0	0	0	0	0	0	0	9
10	0	0	39	0	0	0	0	0	0	20	59
11	0	19	0	0	0	0	0	0	0	38	57
12	0	55	0	0	0	0	0	0	0	6	61
Sum	158	74	48	112	45	1	145	171	81	64	889

Table 6.8 provides the contingency table of the row point which stands for child age versus the column point which stands for the prevalence of virus in blood. Notice that for each cell in the right-hand column, **SUM** is the total number of one child age, regardless of the other ages and prevalence. Similarly, each cell of the bottom **SUM** is the total number of child age occurred to prevalence virus, regardless of other prevalence of virus and child ages. Therefore, it is unsuitable to compare the quantitative difference of child age or prevalence of virus simply from this table generated from one specific sample.

Table 6.9: Contribution to the Total Chi-Square Statistic of Child Age vs Prevalence of virus in Blood

Contribution to the Total Chi-Square Statistic										
	0.00788	0.02385	0.002513	0.003188	0.005208	0.018182	0.021523	0.025424	0.041276	0.047516	Sum
1	0.35	0.16	0.11	0.25	0.10	447.50	0.32	0.38	3.73	0.14	453.05
2	18.98	8.89	5.77	13.45	5.41	0.12	17.42	4.35	478.96	7.69	561.04
3	26.19	12.26	7.96	18.56	7.46	0.17	3.39	379.13	9.72	10.61	475.45
4	0.61	12.68	8.22	19.19	7.71	0.17	380.06	14.06	13.88	10.96	467.54
5	352.33	13.91	9.02	3.00	8.46	0.19	13.61	32.15	15.23	12.03	459.92
6	10.39	7.41	4.81	459.62	0.06	0.10	14.52	17.12	8.11	6.41	528.53
7	7.03	3.29	2.14	4.98	721.11	0.04	6.45	7.61	3.60	2.85	759.11
8	0.18	0.08	0.05	0.12	18.03	0.00	0.16	0.19	0.09	0.07	18.98
9	1.58	0.74	151.04	1.12	0.45	0.01	1.45	1.71	0.81	0.64	159.56
10	10.37	4.86	407.98	7.35	2.95	0.07	9.52	11.22	5.32	59.43	519.06
11	10.02	43.63	3.04	7.10	2.85	0.06	9.19	10.84	5.14	283.91	375.80
12	10.72	497.47	3.26	7.60	3.05	0.07	9.84	11.60	5.50	0.63	549.74
Sum	448.74	605.39	603.39	542.35	777.64	448.50	465.94	490.36	550.08	395.38	5327.78

Table 6.9 displays the contributions to the total Chi-square statistic for each child age and prevalence of virus. The last column summarizes the contributions for child age, where ***age7*** contributes the most, a fact clear from configuration plot. Similarly, the last row summarizes the contributions for prevalence of virus where ***prev*****(0.005208)** makes the largest contribution.

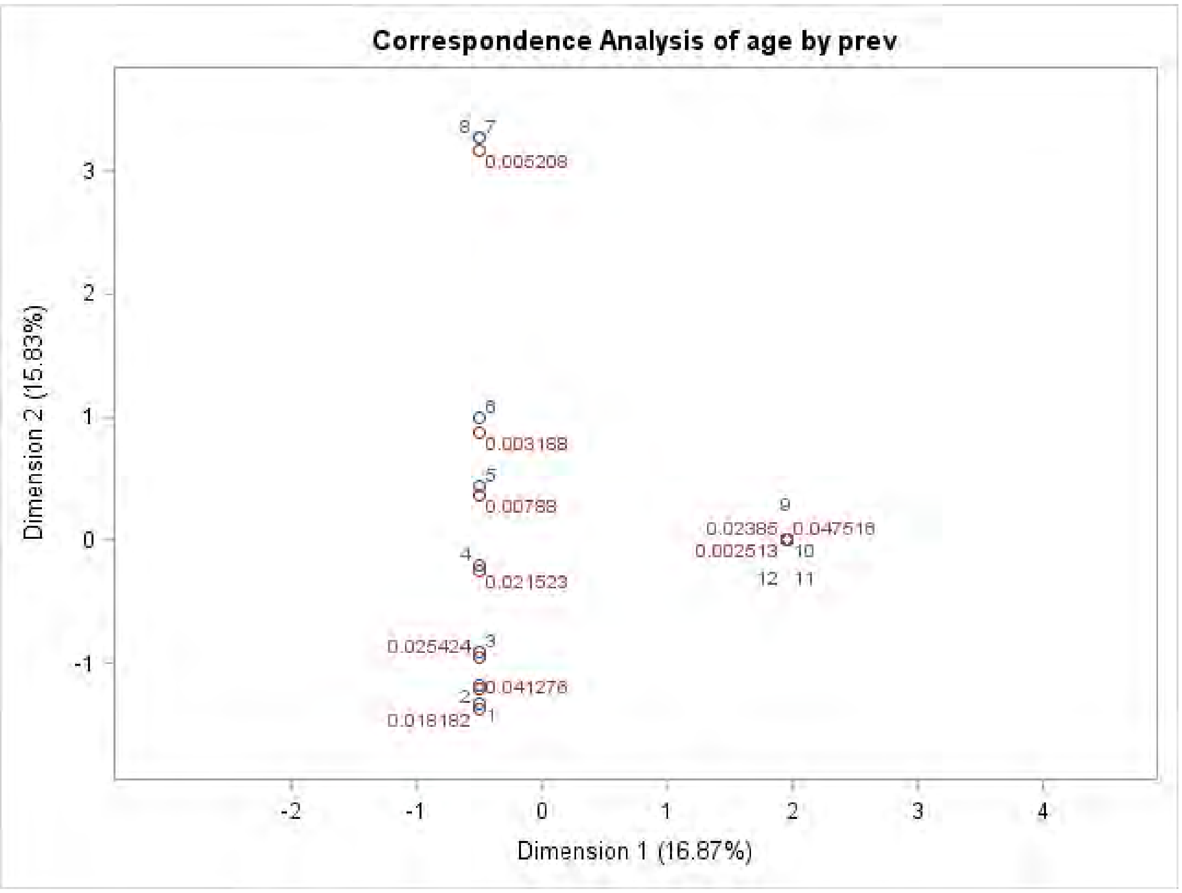

Figure 6.1: CA Plot of Child Age vs Prevalence of Virus in Blood

Fig 6.1 shows the CA Plot of Child Age vs Prevalence of Virus in Blood. To interpret the plot, we start by interpreting the row points separately from the column points. The prevalences of virus are all far from the centroid and they lie along one dimension. They make relatively large contributions to the chi-square statistic and the inertia of one dimension. The horizontal dimension seems to be largely determined by ***prev(0.00788)*** versus ***prev(0.047516)*** points. In the row points, the ***age12*** points is near the centroid and has a small coordinates on one dimension that is near zero. The horizontal dimension seems to be largely determined by ***age4*** and ***age10*** points. The two interpretations of one dimension shows the association with

age4 has ***prev(0.00788)*** virus in blood and with ***age10*** has ***prev(0.047516)*** virus in blood. This means that in young age the child has small prevalence of virus in blood. Distances between row and column points are not defined. The plot shows more children in ***age4*** than we would expect if the rows and columns were independent have ***prev(0.00788)*** and more children who are in ***age10*** than we would expect if the rows and columns were independent have ***prev(0.047516)***. The plot of correspondence analysis can be complex. This means that the data points may appear close to each other but, in fact are placed far apart on the slantwise dimension. That is, the first principal axis reflects 16.87% of the total inertia while for the second axis it is 15.83%. In total, the two-dimensional correspondence plot of Figure 6.1 graphically depicts 31.7% of the association between the variables.

6.11 Summary

In summary, according to Sourial et al. (2010) Correspondence Analysis can be a very helpful tool to display the relationships among categorical variables and generate hypotheses for future analysis, and it is easily implemented with most statistical software. Because CA explores the clustering among categorical variable responses, it can discover how responses within and between variables are related; knowledge that may not otherwise be discovered through a pairwise analysis. We believe that correspondence analysis is an underutilized technique which can play a compatible role in analyzing epidemiological data and therefore deserves greater consideration in this field.

Chapter 7

Discussion and Future Directions

Models for the analysis of longitudinal data are omnipresent these days throughout empirical research. Indeed, models and analysis techniques for longitudinal data, be it for Gaussian or non-Gaussian outcomes are arising up in biometry, epidemiology, medical statistics and survey applications. The models are attractive for the intuition behind their formulation. The inferential apparatus are well developed and methods have been implemented in standard software packages. In this work, we have represented basic methodology for Gaussian and non-Gaussian longitudinal data including the linear and generalized linear mixed model, generalized estimating equations and CA. We also placed a strong emphasis on the use of these methods in conjunction with an outcome. Furthermore, we have indicated how models for longitudinal data are playing a role in research.

Mixed models provide a general framework for the analysis of continuous and discrete repeated measurements, based on linear and non-linear models. In general, parameters in mixed models do not immediately yield population-based inferences. Mixed models specify the full distribution of Y_i. Mixed models are more sensitive to model misspecification than most models for cross-sectional data. Generalized linear models provide a framework for relating response and predictor variables by extending traditional linear model theory to nonlinear data. This is very important in many areas of epidemiologic research where outcomes are dichotomous or otherwise not normally distributed. Generalized estimating equations (GEEs) and generalized linear mixed models (GLLMs) offer a way to analyse such data with reasonable statistical efficiency.
Linear mixed effect model rely on assumptions of multivariate normality and likelihood-based inferences for both the fixed and random effects are relatively straightforward. In contrast, when the longitudinal response is discrete, we have seen that there is more than

one way to extend generalized linear mixed models to the longitudinal setting. This led to the development of marginal and conditional models for non-Gaussian longitudinal data. In our case there was no conditional models. In general, we have seen that likelihood-based approaches are somewhat more difficult to formulate in the non-Gaussian data setting than it is the case with continuous responses. This has led to various avenues of research where more tractable approximations have been developed (e.g. MQL and PQL methods) and where likelihood-based approaches have been abandoned altogether in favour of semi-parametric methods (e.g. GEE approaches).

Our review of the developments of regression models for longitudinal data has focused exclusively on extensions of generalized linear mixed models. Limitations of space have precluded a discussion of non-linear models (i.e., models where the relationship between the mean and covariates is non-linear in the regression parameters) for longitudinal data. The GEE method is semi-parametric, in that the estimating equations are derived without fully specifying the joint distribution of the vector of repeated measures. This is a very appealing feature of the GEE approach, especially for the analysis of discrete longitudinal data, because for the latter case the total number of parameters in the saturated model for the joint distribution of the vector of responses grows exponentially with the number of repeated measures. Although an appealing feature of the GEE approach is its robustness to misspecification of the within-subject association; there are settings where it can be appealing to model the covariance. The implementations of GEE in standard statistical software packages provide only very limited options for modelling the covariance. In particular, there are few choices of models for the ”working covariance” when the data are highly unbalanced and irregularly spaced in time. This is in contrast to models for continuous responses (e.g. general linear models and linear mixed-effects models), where there are a broad class of models for the covariance. Future work is needed in both the formulation and implementation of flexible models for the working covariance in GEE methods.

The present study demonstrates the application of TLC and RSV data analyses using SAS. We first describe the basic modelling framework and demonstrate how various models fit to longitudinal data via SAS. The results presented suggest that the best approach probably is by fitting linear curves for the relevant data sets as a regression of response on time. The result in linear mixed model, random intercept, and random intercept and slope model shows that the week (time) effect is significant at 5%. Clearly blood lead level is linearly related to week. The fixed parameter associated with group is not significant. This indicates that on average neither placebo nor active drug differ significantly concerning their initial blood lead level. Also the interaction parameter is not significant; suggesting that average of active drug increases faster over time than placebo blood lead level. Since the group variable was found to be insignificant it was removed from the model. It should be noted that the sample size used was relatively small. Therefore it would be of benefit to investigate this on a larger sample and increasing observation period. Increasing the observation period and sample size will probably is the best route for future studies.

The results of the model selection show that the correlation matrix resembled an unstructured (UN) in all fitted models. In the analysis of GEE, the results of the model parameter estimates were approximately the same, but the standard errors varied GEE model varies within and across the parameters. A parameter estimate for the ACTPASS effect was statistical significant at 5% level of significance. We conclude that there are differences in the responses of child depending on whether they are visited by fieldworker or they are brought to the clinic. We noticed that the effects of Age and DT did not show a significant contribution towards child response. The parameter estimate for empirical and model based on the GEE model with exchangeable covariance structure are the same. For our data the model with exchangeability covariance structure is chosen. These models perform poorly when there many observations from a handful of subjects. From our findings, we recommend that ACTPASS should be strengthened since the contrast effect of their average performance with PREV decreases over time. In the analysis of GLMM, the results from GLIMMIX and NLMIXED were similar which should give us more confidence in both. The results from GLMMs and NLMIXED models indicate that the group and age effects are significant at 5% significance level while the interaction between group and age shows non-significance differences in both procedures.

As we discussed earlier that CA is an exploratory method of data analysis which does not only quantify multivariate categorical data but yields a graphical representation of the inner structure of the data. The application of CA as a graphical method of data analysis is almost unlimited, but Lebart et al. (1984) and Zhou (2008) suggest that there are conditions that should be satisfied if CA is to be most effective. Firstly, the data matrix must be large enough that visual inspection or simple statistics analysis can reveal its structure. Secondly, the variables must be homogeneous so that it makes sense to calculate a statistical distance between rows and columns and so that distances can be interpreted meaningfully. Lastly, the data matrix must be amorphous a priori. CA is a statistical technique that is specifically designed to explore relationships within and between two or more categorical variables. It can be used to analyse binary data without distributional assumptions. CA provides a unique graphical display thatcan be used to show how the variable response categories are related. While we believe that CA is a very useful technique, its limitation is that distances between row and column points are not mathematically defined. Another limitation of CA is that all of the relevant variables are included in the analysis. If a key variable is overlooked in the design stage of the research, then the final scaling solutions is impoverished or weakened. The possible future research includes:

1. Expanding the applicability of and improving the estimation method for linear mixed models.

2. Extending the theory of correspondence analysis to multiple correspondence analyses.

3. The development of statistical technique to handle complex survey design and analysis of correlated data with measurement error in response from longitudinal survey.

4. We can also consider the penalized generalized estimating equations for analysing longitudinal data with high-dimensional covariates.

Bibliography

Agresti, A. (2002). *Categorical data analysis*, Volume 359. John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on 19*(6), 716–723.

Al-Marshadi, A. H. (2011). New procedure to improve the order selection of autoregressive time series model. *Journal of Mathematics and Statistics 7*(4), 270–274.

Anderson, D., K. Burnham, and G. White (1994). Aic model selection in overdispersed capture-recapture data. *Ecology 75*(6), 1780–1793.

Anderson, R. L. and T. A. Bancroft (1952). Statistical theory in research.

Beh, E. J. (2004). Simple correspondence analysis: a bibliographic review. *International Statistical Review 72*(2), 257–284.

Benzecri, J. (1973). *LAnalyse des Donnes*, Volume vol. 1 and vol. 2. Paris: Dunod, France.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*(1), 81–91.

Brown, H. and R. Prescott (2006). *Applied mixed models in medicine*. John Wiley & Sons.

Caley, P. and J. Hone (2002). Estimating the force of infection; mycobacterium bovis infection in feral ferrets mustela furo in new zealand. *Journal of Animal Ecology 71*(1), 44–54.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 15–18.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. Springer.

De Leeuw, J. and P. Mair (2007). Simple and canonical correspondence analysis using the r package anacor. *Department of Statistics, UCLA*.

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of longitudinal data*. Oxford University Press.

Doey, L. and J. Kurta (2011). Correspondence analysis applied to psychological research. *Tutorials in Quantitative Methods for Psychology 7*(1), 5–14.

Dufour, A. (2008). Correspondence analysis (coa or ca).

Feddag, M. L. and M. Mesbah (2006). Approximate estimation in generalized linear mixed models with applications to the rasch model. *Computers & Mathematics with Applications 51*(2), 269–278.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fitzmaurice, G. M., N. M. Laird, and J. H. Ware (2004). Applied longitudinal analysis.

Gbur, E., W. W. Stroup, K. S. McCarter, S. Durham, L. J. Y. abd Mary Christman, M. West, and M. Kramer (2012). *Analysis of generalized linear mixed models in the agricultural and natural resources sciences*. American Society of Agronomy; Soil Science Society of America; Crop Science Society of America.

Greenacre, M. (2002). The use of correspondence analysis in the exploration of health survey data. Technical report.

Greenacre, M. (2007). Correspondence analysis in practice.

Greenacre, M. and T. Hastie (1987). The geometric interpretation of correspondence analysis. *Journal of the American statistical association 82*(398), 437–447.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*.

Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics 20*(2), 251–269.

Guo, X. (2011). *Longitudinal Data Analysis in Social Science Data*. Ph. D. thesis, University of Alberta, Canada.

Hanley, J. A., A. Negassa, J. E. Forrester, et al. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American journal of epidemiology 157*(4), 364–375.

Hannan, E. J. and B. G. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 190–195.

Hardin, J. and J. Hilbe (2003). *General estimating equations*. Boca Raton, FL, Chapman & Hall.

Hartley, H. O. and J. N. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika 54*(1-2), 93–108.

Harville, D. A. (1976). Confidence intervals and sets for linear combinations of fixed and random effects. *Biometrics*, 403–407.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338.

Hedeker, D. and R. D. Gibbons (2006). *Longitudinal data analysis*, Volume 451. John Wiley & Sons.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics 9*(2), 226–252.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*(2), 297–307.

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

Kobayashi, R. (1981). *An Introduction to Quantification Theory*. Tokyo: Japan Union of Scientists and Engineers.

Komazawa, T. and C. Hayashi (1982). *Quantification theory and data processing*. Tokyo: Asakura-shoten.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 963–974.

Lebart, L., A. Morineau, and N. Tabard (1977). Techniques de la description statistique. *Dunod, Paris*, 351.

Lee, Y. and J. A. Nelder (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika 88*(4), 987–1006.

Liang, K.-Y. (1999). *Generalized Linear Models Estimating Functions and Multivariate Extensions*. Institude of Statistical Science, Academia Sinica.

Lipsitz, S. R., G. M. Fitzmaurice, E. J. Orav, and N. M. Laird (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 270–278.

Littell, R. C. (2006). *SAS for mixed models*. SAS institute.

Longford, N. T. (1995). *Random coefficient models*. Springer.

Ludovic, L., A. Morineau, and K. M. Warwick (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. *New York: John Wiley & Sons*.

Mazzarol, T. W. and G. N. Soutar (2008). Australian educational institutions' international markets: a correspondence analysis. *International Journal of Educational Management 22*(3), 229–238.

Mazzocchi, M. (2008). *Statistics for marketing and consumer research*. Sage.

McCullagh, P. and J. Nelder (1989). Generalized linear models.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association 92*(437), 162–170.

McCulloch, C. E. and J. M. Neuhaus (2001). *Generalized linear mixed models*. Wiley Online Library.

Meyer, K. (1989). Estimation of genetic parameters. *Reviewes on Molecular and Quantitative Genetic Aproaches in Honor of Alan Robertson*, 159–167.

Moeti, A. (2010). *Factors affecting the health status of the people of Lesotho.* Ph. D. thesis, School of Statistics and Actuarial Science, University of KwaZulu-Natal.

Molenberghs, G. and G. Verbeke (2005). Models for discrete longitudinal data.

Moudud, A. (2009). An efficient algorithm for the pseudo likelihood estimation of the generalized linear mixed models (glmm) with correlated random effects.

Mwambi, H., S. Ramroop, L. White, E. Okiro, D. J. Nokes, Z. Shkedy, and G. Molenberghs (2011). A frequentist approach to estimating the force of infection for a respiratory disease using repeated measurement data from a birth cohort. *Statistical methods in medical research 20*(5), 551–570.

Nelder, J. and R. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Nguyen, D. V., D. Şentürk, and R. J. Carroll (2008). Covariate-adjusted linear mixed effects model with an application to longitudinal data. *Journal of nonparametric statistics 20*(6), 459–481.

Nilsson, A. (2011). *Evaluation of scoring index with different normalization and distance measure with correspondence analysis*. Sweden: Lund University.

Owusu-Darko, I., I. K. Adu, and N. K. Frempong (2014). Application of generalized estimating equation (gee) model on students academic performance. *Applied Mathematical Sciences 8*(68), 3359–3374.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics 57*(1), 120–125.

Pan, Z. and D. Lin (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics 61*(4), 1000–1009.

Patterson, H. and R. Thompson (1974). Maximum likelihood estimation of components of variance. In *Proc. Eighth International Biochem. Conf*, pp. 197–209.

Patterson, H. D. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033–1048.

Raudenbush, S. W., M.-L. Yang, and M. Yosef (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of computational and Graphical Statistics 9*(1), 141–157.

Rhoads, M. (2000). The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *Pediatrics 110*(4), 787–791.

Russell, T. S. and R. A. Bradley (1958). One-way variances in a two-way classification. *Biometrika*, 111–129.

Schabenberger, O. (2005). Mixed model influence diagnostics. In *SUGI*, Volume 29, pp. 189–29.

Schelldorfer, J., L. Meier, and P. Bühlmann (2012). Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using 1-penalization. *Journal of Computational and Graphical Statistics 23*(2), 460–477.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

Searle, S., G. Casella, and C. MacCulloch (1992). Variance components. *Wiley series in probability and mathematical statistics*.

Searle, S. R. (1982). Matrix algebra useful for statistics. *New York 1982*.

Sorensen, D. and B. Kennedy (1984). Estimation of genetic variances from unselected and selected populations. *Journal of Animal Science 59*(5), 1213–1223.

Sourial, N., C. Wolfson, B. Zhu, J. Quail, J. Fletcher, S. Karunananthan, K. Bandeen-Roche, F. Béland, and H. Bergman (2010). Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of clinical epidemiology 63*(6), 638–646.

Sun, L. (2011). Comparison of different estimation methods for linear mixed models and generalized linear mixed models.

Ter Braak, C. J. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology 67*(5), 1167–1179.

Venables, W. N. and B. D. Ripley (2002). *Modern applied statistics with S*. Springer.

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

Verbyla, A. and B. R. Cullis (1990). Modelling in repeated measures experiments. *Applied Statistics*, 341–356.

Vittinghoff, E., D. V. Glidden, S. C. Shiboski, and C. E. McCulloch (2005). *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika 61*(3), 439–447.

West, B., K. Welch, A. Gałecki, and B. Gillespie (2007). Linear mixed models: a practical guide using statistical software.

Yan, J., S. Hojsgaard, U. Halekoh, and M. J. Yan (2007). The generalized estimating equation package. *Geepack version 1*, 0–13.

Zewotir, T. and J. S. Galpin (2005). Influence diagnostics for linear mixed models. *Journal of data science 3*(2), 153–177.

Zhou, T. (2008). *Correspondence Analysis of Elevator Malfunction Matrices*. Ph. D. thesis, University of California, LA.