# Tweedie Mixed Models for Unevenly Spaced Longitudinal Data

by

Nakisa Tamjidi

**Bachelor of Science (Medicinal Chemistry), University of New Brunswick, 2014**

**A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF**

**Master of Science**

In the Graduate Academic Unit of of Mathematics and Statistics

| | |
|---|---|
| Supervisor(s): | Guohua Yan, Ph.D, Statistics |
| | M. Tariqul Hasan, Ph.D, Statistics |
| Examining Board: | Nicholas Touikan, Ph.D, Mathematics, Chair |
| | Namwei Wang, Ph.D, Statistics |
| | James Watmough, Ph.D, Mathematics |

This report is accepted by the
Dean of Graduate Studies

**THE UNIVERSITY OF NEW BRUNSWICK**

**June, 2022**

# Abstract

In Longitudinal set-ups, we often need to face many instances where we must deal with successive count responses given at unevenly spaced time intervals. In these scenarios, we consider the study designs to be complete, as opposed to an evenly spaced design with non-response or missing values.

In dealing with the variably spaced longitudinal data sets, here we propose a method to estimate parameters for unevenly spaced longitudinal data by working with the flexible class of Tweedie generalized linear mixed models, with both subject-specific and time-specific random effects. This class of models are able to handle a variety of data types, including continuous, discrete and mixed data.

The proposed method is demonstrated with the analysis of patient-controlled analgesia dataset from Henderson and Shimacura (2003), with the exception of omitting the 'no response' values for the individuals from the dataset as they are neither a missing response, nor can they be qualified as a zero, and thus creating an unbalanced dataset.

To maintain the accuracy of the correlations in real studies, the duration of the time intervals is generally kept small, leading to a large value of T (total number of intervals). For inferences, the regression parameters are estimated by the scoring method and the unevenly spaced correlation parameters of random effects by using moment estimators.

# Dedication

This report is dedicated to my parents, my aunt and my siblings for their unconditional love and support.

# Acknowledgements

Foremost, I would like express my sincerest gratitude to my supervisors, Dr. Guohua Yan and Dr. Tariqul Hasan for their incredible support, patience and guidance.

I would like to give my warmest thanks to Dr. Yan and Dr. Hasan for their continued help and advice throughout the whole process and I would like to express my deepest appreciation for their kindness and encouragement.

I am also grateful to Dr. Jeffery Picka and Dr. Renjun Ma for their invaluable lessons and their support.

Finally, I would like to thank the University of New Brunswick and the Department of Mathematics and Statistics for providing such an excellent academic program and supportive environment. Many thanks to the always friendly faculty members and staff.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Longitudinal data analysis models have been gaining tremendous popularity in recent years. This is particularly apparent when it comes to dealing with clinical studies and medical assessments. Typically, the main objective of a longitudinal study is to estimate how a response variable is affected by the covariates and how it changes over time. One frequent use of longitudinal models in medical studies is to observe a particular response for subjects in pre/post-op study designs, where a single baseline health related measurement is obtained, a treatment is done, and follow-up measurements are collected over time.

In order to simplify the analysis, most longitudinal studies are designed in a balanced structure, where every subject is assessed on the same exact schedule with the same number of measurements taken. For example, we may look at the Health Care Utilization (HCU) data (Sutradhar 2003; Sutradhar 2011) where repeated numbers of yearly physician visits for four years from 180 individuals were studied as a function of various covariates such as gender, education level, chronic disease status, and age of individuals.

Similarly, many researchers such as Leppik et al. (1985); Thall and Vail (1990);

Jowaheer and Sutradhar (2002), have studied the four two weekly evenly balanced longitudinal seizure counts arising from a clinical trial of 59 epileptics, as a function of various covariates, namely an intercept, the adjuvant treatment, baseline seizure rate, the age of the person in that year, and the interaction between treatment and baseline seizure rate.

However, despite the efforts of researchers to design a perfect time structured data set, in reality, difficulties may arise with executing a completely balanced study, due to a number of reasons. Most notably, collecting data from live subjects can be hard and unpredictable due to scheduling limitations and participant dropouts. Thus, researchers often have to deal with incomplete data sets with missing values or non-responses, meaning they would not be able to validly apply their measurements to the balanced statistical models, when in fact, they are dealing with an unbalanced data set as a result of different numbers of repeated measures, or sequence of repeated measures taken at irregular times for different individuals.

However, some longitudinal studies are set up as unevenly spaced time intervals, in which they are deemed as complete datasets. For example, there are occasions where a patient may provide successive count responses at unevenly spaced time intervals as opposed to the usual equi-distant studies. These measurements are often taken with other covariate information at any given time points, meaning the responses and covariates are complete as contrasted with the afore-mentioned longitudinal data subject to non-response or missing.

We use the term unbalanced specifically to refer to any design in which the sampling units are measured at different time points. Missing or incomplete data on the other hand would be considered under a different category. For example, missing clinic visits might be a reason for lack of balanced designs, but they usually

do not tell the whole story. A general class of models for unbalanced longitudinal data had been proposed by Huang and Fitzmaurice (2005). Their class of models is based on separate specifications of the moments for the mean, standard deviation, and correlation. Many researchers, such as Diggle et al. (2002) and Molenberghs and Verbeke (2005) have worked with marginal, mixed and conditional models, including the transition models. All these models not only can be used for balanced longitudinal data, but it is also entirely possible to apply them to unbalanced data as well. Cnaan et al. (1997) illustrates the attractive quality of using general linear Mixed models for unbalanced longitudinal data since no adjustment is necessary. On the other hand, mean and covariance structures need to be specified in order to use a Gaussian marginal model. Generalized estimating equations (GEEs) which do not specify distribution and covariance structure is a marginal approach that can be seriously considered for these situations. In the transition class of models, conditional models of the outcome at the current time point are modeled as a function of the past outcomes and covariates, Diggle et al.(2005). Transition models are useful when our interest is in predicting future outcomes or when past history contains important adjustor variables. Although they are commonly used for equally spaced data (Heagerty, 2002), they can be extended to unequally spaced designs as well. Marginal models are often referred to as "population-averaged models" and mixed models as "subject-specific models".

To deal with the variably spaced longitudinal data sets, we consider working with the flexible class of Tweedie generalized linear models, with both subject-specific and time-specific random effects. An important feature of our proposed model is that we take into account of different numbers of repeated measures, as well as measures taken at uneven times for different individuals.

Hence, as an example, when looking at a model for unevenly spaced longitudinal Poisson counts with time interval of 4 (t=4 weeks as the duration of the study), if an individual only reports for three out of the four weeks (in terms of counts), the 3 count response measurements collected would be considered unevenly spaced, and the study to be complete. The 'no response' value for the individual is neither a missing response, nor can be qualified as a zero, since no probability can be assigned for a non-existing event (Oyet, 2019).

To accommodate the correlations among the repeated yearly visits in the HCU data, Sutradhar (2011) have used an auto-regressive order 1 (AR(1)) type dynamic model for Poisson counts. Similarly, Jowaheer and Sutradhar (2002) have used an AR(1) type dynamic model for negative binomial counts to model the correlations among the repeated seizure counts of an individual. These AR(1) models are very practical for two reasons. First, these models for discrete count data produce a correlation structure exhibiting decaying correlations as the time lag increases. Secondly, the correlations under such models appear to be functions of the time dependent covariates, as expected.

To maintain the accuracy of the correlations in real studies, the duration of the time intervals is generally kept small, leading to a large value of T (total number of intervals). For inferences, the regression parameters are estimated by the scoring method and the unevenly spaced correlation parameters of random effects by using moment estimators. In this paper, we discuss a Generalized Linear Mixed Model (GLMM) framework for longitudinal responses (Li, 2018). The framework is based on the class of Tweedie exponential dispersion distributions (Jørgensen, 1987) which includes as special cases the Poisson, normal gamma, inverse Gaussian, compound Poisson, gamma and so on (Ma, 1999); for this very reason, the proposed model

framework is able to handle various data types, including both continuous and discrete data.

Our proposed method of estimating parameters is less restrictive since the measurement time points do not have to be evenly spaced. Here we use a generalized AR(1) structure for analyzing the correlation. This correlation structure is known as the Markov correlation. An optimal equation to predict the random effects has been obtained based on the best linear unbiased predictor of the random effects. This approach leads to improved computational efficiency since all the formulas are explicit in each iterative step.

We demonstrate the method using an unevenly spaced patient-controlled analgesia (PCA) dataset where hospital patients were given the opportunity to control their own pain relief medication following an abdominal surgery. The purpose of this example was to compare the two dosing regimens, where the bolus is 2 mg of morphine with a lock-out period of 8 minutes for the first group of 30 patients, whereas the bolus was 1 mg of morphine with a lock-out period of 4 minutes for the second group of 35 patients.

Our plan is to consider the 'no response' value for each of the individuals as neither a missing response, nor be qualified as a zero, since no probability can be assigned for a non-existing event. Meaning each individual can have measurement times anywhere between zero and twelve, thus making the dataset an unevenly spaced longitudinal dataset.

The remaining part of this report is organized as follows. In Chapter 2 we introduce an overview of statistical models for longitudinal studies. Two general approaches for analyzing longitudinal data are presented, including the Generalized Linear Mixed Models (GLMM) and the Tweedie exponential dispersion models.

In chapter 3, we present our model for unequally spaced longitudinal data. In this chapter we discuss the model specifications, the moment structure for the response and the orthodox best linear unbiased predictors of the random effects, as well as the estimation for the parameters.

In chapter 4, the proposed method is furthermore tested with the analysis of patient-controlled analgesia (PCA) dataset. The test statistics are run using the R Core Team (2020) software and relevant results are considered and discussed.

In Chapter 5, there will be a conclusion summarizing the report and a brief discussion of our proposed model and options for directions of further research.

# Chapter 2

# Literature Review

## 2.1 An introduction to longitudinal studies

### 2.1.1 Longitudinal studies

Longitudinal studies are observational designs in which measurements for response variables are collected repeatedly over time. The defining features of the longitudinal study is that the data gathered are of the same individuals and the measurements are commensurate, meaning the same variable is being observed over a period of time, thus, allowing us to directly study the change of the response variable.

Longitudinal data analysis has been a big field of interest in statistics for some time now. For many years, researchers have been working on deriving new, sophisticated techniques for analyzing these types of studies. Longitudinal data deals with repeated observations which are often (positively) correlated. Hence, the main objective of a longitudinal study is to estimate the changes of a response variable within the subjects and characterize the factors that affected the outcome.

Longitudinal studies are commonly used in many research projects, particularly

when it comes to dealing with clinical trials and medical assessments. One of the common uses of the longitudinal models in medical studies is to observe a particular response for subjects in a pre/post-up study designs, where a single baseline health related measurement is obtained, a treatment is done, and follow-up measurements are collected.

**General Data Structure**

Let $y_{ij}$ represent the longitudinal response recorded at the $j^{th}$ response of the $i^{th}$ subject and $x_{ijk}$ denote the $k^{th}$ covariate for the $ij^{th}$ measurement, with $i$ in $\{1, \cdots N\}$, and $j = 1, \cdots n_i$ and $k = 1, \cdots p$, where N is the total number of subjects, $n_i$ is the total number of repeated measurements of the $i^{th}$ subject, and p is the number of covariates.

A longitudinal data set can be expressed by table 2.1.

Table 2.1: Layout of a longitudinal data set

| Subject (i) | Repeated measurement (j) | $y_{ij}$ | $x_{ijk}$ |
|---|---|---|---|
| 1 | 1 | $y_{11}$ | $x_{111} \cdots x_{11p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 1 | $n_1$ | $y_{1n_1}$ | $x_{1n_11} \cdots x_{1n_1p}$ |
| 2 | 1 | $y_{21}$ | $x_{211} \cdots x_{21p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 2 | $n_2$ | $y_{2n_2}$ | $x_{2n_21} \cdots x_{2n_2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| N | 1 | $y_{N1}$ | $x_{N11} \cdots x_{N1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| N | $n_N$ | $y_{Nn_N}$ | $x_{Nn_N1} \cdots x_{Nn_Np}$ |

A few key aspects of longitudinal studies are:

- they can be observational or experimental,
- they are correlational research based,

- they are often compared and contrasted with cross-sectional research,
- they require data collection over a period of time (often years or even decades), as opposed to data collection at a single point in time.

The benefit of the longitudinal study is that it allows us to look at changes of a response variable for subjects beyond a specific time interval. Because of this, these methods are particularly useful when it comes to measuring developments over long periods of time and doing research on life-long effects of specific issues. This is also the main difference between longitudinal studies and cross-sectional studies, since in cross-sectional studies, the comparisons between subjects are done at a single point in time.

Longitudinal studies can be real assets in research studies involving identical twins; specifically, in looking at cases when the twins grew up together versus when they grew up apart (nature vs. nurture). This would present an excellent opportunity since the subjects share the same genetic make-up. Researchers can track these participants through their life span and observe how certain elements would differ and change at various points in their lives and perhaps find some of the reasons as to why some developmental shifts happen and explore whether the changes in some specific variables are due to environmental factors or genetics.

As with any type of research study methods, longitudinal studies have both strengths and weaknesses. One of the main difficulties in conducting a longitudinal study comes from the very same feature that makes it unique: it requires following subjects for long periods of time (years or even decades). As a result of this, we face some problems, mainly because the studies in turn can come out to be quite expensive. Due to it being very hard to find subjects willing to participate in a long-run study, and the fact that these studies can cost a lot, usually, only small groups of

subjects are used in longitudinal studies, which later makes it difficult to apply the results to real life situations and larger populations. Another problem arises from the unknown time span of the studies: participants tend to drop out due to many number of unforeseen reasons (from moving, to dealing with family or health related issues, to passing away, to simply losing interest in continuing with the study).

In some cases, this can lead to the reduction of the strength or the effectiveness of the study and can cause an attrition bias. If the subjects remaining at the end no longer reflect the original sample group, the attrition would also threaten the validity of the study, where the result could no longer be generalized to the rest of the population.

## 2.2 General approaches for longitudinal methods

Over the years, researchers have come up with several popular approaches to deal with longitudinal data. Among those, linear mixed effects models can be considered for analyzing longitudinal data when the response variable are continuous. On the other hand, extensions of generalized linear models for longitudinal data are extremely popular when dealing with discrete (e.g. binary or a count) as well as continuous response variables. Several general approaches have been described by Diggle et al.(2002) to model longitudinal data. In Fitzmaurice et al.(2008), various methods are introduced for the analysis of such data. In this report we discuss three main extensions of generalized linear models defined according to Zeger and Liang (1986): (1) marginal models, which are also referred to as "population averaged models" for analyzing longitudinal data when the response variable is discrete or continuous; (2) generalized mixed effects models, also referred to as "subject-specific models"; and (3) transitional models . Each of these models allows time-invariant

predictors that never change (e.g., biological sex) and time-varying predictors (e.g., age) and handle irregularly timed and missing data.

### 2.2.1 Linear Mixed Effects Models

Many approaches are available for analysis of continuous longitudinal data. Another modelling approach is about a mixed-effects or random-effects model (Laird and Ware, 1982). Over the past two decades, a lot of emphasis has been put on linear mixed models.

Let $Y_i = (Y_{i1}, Y_{i2} \ldots, Y_{in})$ be a vector of correlated responses for the $i$th subject $i = 1, 2, \ldots, N$. The linear mixed effects model assumes that the longitudinal responses depend on a combination of population or fixed effects parameters, $\beta$, and random or subject-specific effects that are unique to a particular individual. The marginal mean responses (averaged over the distribution of the random effects) can be expressed in the general form

$$E(Y_i) = X_i\beta \tag{2.1}$$

The introduction of random effects induces covariance among the responses and $\mathrm{Cov}(Y_i) = \Sigma_i$ has a distinctive random effects structure. With the inclusion of random effects, the covariance among the repeated measures can be expressed as a function of time.

The linear mixed effects model can be represented in two steps: first, assume $Y_i$ has a normal distribution that depends on population specific effects, $\beta$, and individual specific effects, $b_i$, where $e_i$ is a vector of errors, and $e_i \sim N(0, R_i)$ and $Z_i$ is a matrix of covariates, linking the vector of random effects $b_i$ to $Y_i$;

$$Y_i = X_i\beta + Z_i b_i + e_i \tag{2.2}$$

second, the response for the $i^{th}$ subject and the $j^{th}$ occasion is assumed to differ from the population and $b_i \sim N(0, G)$. In other words, the response for the $i^{th}$ subject at the $j^{th}$ occasion is assumed to differ from the population mean, $X_{ij}\beta$, by a subject effect, $b_i$, and a within-subject measurement error, $e_{ij}$.

$R_i = \text{Var}(e_i)$, and describes the covariance among observations when we focus on the response profile of a specific individual. Therefore, it is the covariance of the $i^{th}$ subject's deviations from his/her mean profile $X_i\beta + Z_i b_i$. It is assumed that $R_i = \sigma^2 I$ where I is the identity matrix. In the mixed model, the vector of regression parameters, $\beta$, are the fixed effects, which are assumed to be the same for all individuals. On the other hand, the subject specific regression coefficients $b_i$, describe the mean response profile of a specific individual.

$$Y_i = X_i\beta + Z_i b_i + e_i \tag{2.3}$$

$$E(Y_i \,|b_i) = X_i\beta + Z_i b_i \tag{2.4}$$

$$E(Y_i) = X_i\beta \tag{2.5}$$

$$\text{Cov}\,(Y_i \,|b_i) = \text{Cov}\,(e_i) = R_i = \sigma^2 I \tag{2.6}$$

$$
\begin{aligned}
\text{Cov } (Y_i) &= \text{Cov } (Z_i b_i) + \text{Cov } (e_i) \\
&= Z_i \text{ Cov}(b_i)\ Z'_i + \text{Cov } (e_i) \\
&= Z_i G Z'_i + R_i \\
&= Z_i G Z'_i + \ \sigma^2 I \qquad (2.7)
\end{aligned}
$$

Thus, the introduction of random effects, $b_i$, induces a correlation (marginally) in the response $Y_i$.

### 2.2.2 Generalized Linear Models

#### 2.2.2.1 Marginal Models: Generalized Estimation Equations (GEE)

A generalized estimation equation model is designed for analyzing the regression relationship between covariates of interest and repeated responses. This approach was derived from marginal models, a class of regression models and by the extension of generalized linear models to fit longitudinal data. The term marginal here indicates that the model for the mean response at each occasion depends only on the covariates of interest and not on any random effects or previous responses. This is in contrast to mixed effects regression model, which is designed to analyze the correlation structure of the repeated responses.

**Fitting Marginal Models**

A marginal model can be applied to use the framework of generalized linear models (GLM) (edler and Wedderburn (1972). The marginal model can be broken down to three specific parts.

1. The conditional expectation or mean of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$ is

assumed to be independent across each subject and only depend on the covariates through a known function. Therefore, the correlation between $\mu_{ij}$ and the covariates are shown as $g(\mu_{ij}) = X'_{ij}\beta$.

2. The conditional variance of each $Y_{ij}$ given the covariates is assumed to depend on the mean according to $\text{Var}(Y_{ij}|X_{ij}) = \phi \text{V}(\mu_{ij})$, where $\text{V}(\mu_{ij})$ is a known variance function of the mean, $\mu_{ij}$, and $\phi$ is a scale parameter which may need to be estimated.

3. The conditional within-subject association among the vectors of repeated responses given the covariates is assumed to be a function of an additional set of association parameters, $\alpha$. Thus, the correlation between $Y_{ij}$ and $Y_{ik}$ is a function of these additional parameters and may also depend on $\mu_{ij}$ and $\mu_{ik}$.

An estimate of $\beta$ can be obtained as the solution to the following generalized estimating equations:

$$\sum_{i=1}^{n} D'_i V_i^{-1} (Y_i - \mu_i) = 0, \tag{2.8}$$

where $D_i = \partial\mu_i/\partial\beta$ and $V_i$ is a covariance matrix, i.e. $\text{V}_i \approx \text{Cov}\ (Y_i)$

$$D_i = \begin{bmatrix} \partial\mu_{i1}/\ \partial\beta_1 & \partial\mu_{i1}/\ \partial\beta_2 & \dots & \partial\mu_{i1}/\ \partial\beta_p \\ \dots & \dots & \dots & \dots \\ \partial\mu_{in_i}/\ \partial\beta_1 & \partial\mu_{in_i}/\ \partial\beta_2 & \dots & \partial\mu_{in_i}/\ \partial\beta_p \end{bmatrix}. \tag{2.9}$$

$D_i$ is a function of $\beta$ and $V_i$ is a function of both $\beta$ and $\alpha$.

We can express $V_i$ as function of the variances and correlations

$$V_i = \quad A_i^{1/2} R(\alpha) \; A_i^{1/2}, \tag{2.10}$$

where $A_i$ is a diagonal matrix with $\mathrm{Var}(\mathrm{Y}_{ij}) = \phi \, V(\mu_{ij})$ as the $j^{th}$ diagonal element,

$$\boldsymbol{A_i} = \begin{bmatrix} \phi \, \mathrm{V}(\mu_{ij}) & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \phi \, \mathrm{V}(\mu_{ip}) \end{bmatrix}, \tag{2.11}$$

and $A_i^{1/2}$ is a diagonal matrix with the standard deviations, $\sqrt{\phi V(\mu_{ij})}$ along the diagonal. Since the generalized estimating equations depend on both $\beta$ and $\alpha$, an iterative two stage procedure is required. $R(\alpha)$ is the correlation matrix $\mathrm{Cor}(Y_i)$, here a function of $\alpha$. Given a current estimates of $\alpha$ and $\phi$, an estimate of $\beta$ is obtained as the solution to the generalized estimating equations:

$$\psi(\ \beta) = \sum_{i=1}^{n} \frac{\partial \mu_i'}{\partial \beta} \, V_i^{-1} (Y_i - \mu_i(\ \beta)) \tag{2.12}$$

Given current estimate of $\beta$, estimates of $\alpha$ and $\phi$ are obtained based on the standardized residuals,

$$e_{ij} = \frac{(Y_{ij} - \ \hat{\mu}_{ij})}{\sqrt{V(\hat{\mu}_{ij})}} \tag{2.13}$$

**Properties of GEE estimates**

Assuming the estimates of $\alpha$ and $\phi$ are consistent, $\hat{\beta}$, the solution to the generalized

estimating equations has the following properties:

1. $\hat{\beta}$ is a consistent estimator of $\beta$,
2. in large samples, $\hat{\beta}$ has a multivariate normal distribution,
3. $\text{Cov}\left(\hat{\beta}\right) = B^{-1}MB^{-1}$,

where

$$B = \sum_{i=1}^{N} D'_i V_i^{-1} D_i \tag{2.14}$$

$$M = \sum_{i=1}^{N} D'_i V_i^{-1} \text{Cov}(Y_i)\ V_i^{-1} D_i \tag{2.15}$$

Note that B and M can be estimated by replacing $\alpha$, $\phi$ and $\beta$ by their estimates and by replacing $\text{Cov}(Y_i)$ by $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.

That is, we can use the empirical or so-called "sandwich" variance estimator: the components B and M can be thought of as the "bread" and "meat" of this sandwich estimator of $\text{Cov}\left(\hat{\beta}\right)$.

$$\widehat{\text{Cov}}\left(\hat{\beta}\right) = \left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i\right)\left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^{-1} \hat{V}_i^{-1} \hat{D}_i\right)\left(\sum_{i=1}^{N} \hat{D}_i' \hat{V}_i^{-1} \hat{D}_i\right)^{-1}$$

**Advantages of GEE estimators**

The GEE designs are widely used in the neurodegenerative disease literature such as clinical trials for Huntington's disease.

1. In many cases, the GEE estimator is almost as precise or efficient when compared to the MLE method. Also, GEEs are well liked for their computational simplicities.

2. The GEE estimator has a robustness property, yielding a consistent estimate of $\beta$ even if the within-subject associations among the repeated measures have been mis-specified.

3. When the model for the covariance is mis-specified, valid standard errors for $\hat{\beta}$ can be obtained using the "sandwich" estimator of Cov $\left(\hat{\beta}\right)$. Thus, although the GEE approach and the "sandwich" estimator of Cov $\left(\hat{\beta}\right)$ are more widely used in marginal models for discrete data, they can also be applied in the linear models for continuous data.

**Disadvantages of GEE estimators**

1. GEEs assume missing data are "missing completely at random" (MCAR), which is not true for neurodegenerative disease studies. Therefore, extensions to the more flexible MAR assumption have been proposed, including a weighted-estimating equations approach.

2. We cannot perform hypothesis testing on correlation parameters since these are not directly estimated.

3. Usual methods such as likelihood ratio tests cannot be used to test and compare model fits, because the focus is solely on regression parameters, not all model parameters (i.e., regression and correlations parameters).

### 2.2.3 Generalized Linear Mixed-effects Models

Generalized linear models can be extended to longitudinal data by allowing a subset of the regression coefficients to vary randomly from one individual to another. The generalized linear mixed models (GLMMs) are an extension of linear mixed models to allow response variables from different distributions. We can think of GLMMs as an extension of generalized linear models to include both fixed and random effects. The general form of the model in matrix notation is

$$y = X\beta + Zb + \varepsilon \tag{2.17}$$

Where the outcome variable, $y$, is a $N \times 1$ column vector, $X$ is a $N \times p$ matrix of the p predictor variables, $\beta$ is a $p \times 1$ column vector of the fixed-effects regression coefficients, $Z$ is the $N \times q$ design matrix for the q random effects (the random complement to the fixed $X$), b is a $q \times 1$ vector of the random effects (the random complement to the fixed $\beta$), and $\varepsilon$ is a $N \times 1$ column vector of the residuals.

$$\overbrace{\mathbf{y}}^{\text{N x 1}} = \overbrace{\underbrace{\mathbf{X}}_{\text{N x p}} \quad \underbrace{\boldsymbol{\beta}}_{\text{p x 1}}}^{\text{N x 1}} + \overbrace{\underbrace{\mathbf{Z}}_{\text{N x q}} \quad \underbrace{\boldsymbol{u}}_{\text{q x 1}}}^{\text{N x 1}} + \overbrace{\boldsymbol{\varepsilon}}^{\text{N x 1}} \tag{2.18}$$

Mixed models for longitudinal data explicitly identify individual (random effects) and population characteristics (fixed effects). These models are extremely flexible, as they can accommodate any degree of imbalance in the data. Therefore, we can work with data that do not necessarily posses the same number of observations on each subject or that the measurements are taken at the same time. In

addition, the use of random effects allows us to model the covariance structure as a continuous function of time.

The generalized linear model is actually a family of probability models that includes the normal, Bernoulli, Poisson and Gamma distributions. These models are a class of regression models that not only include the linear model, but also many of the important nonlinear models used in the biomedical research field, such as linear regression for continuous data, logistic regression for binary data, and Poisson models for counts.

The generalized linear mixed model for non-Normal responses, $Y_i$, can be written as follows: first, assume that the conditional distribution of each $Y_{ij}$ given $b_i$ belongs to the exponential family with conditional mean

$$g(E[Y_{ij}|b_i]) = X_{ij}\beta + Z_{ij}b_i \tag{2.19}$$

where $g$ is a known link function; secondly, the $b_i$ are assumed to vary independently from one individual to another and $b_i \sim N(0,\ G)$. Also note that there is an additional assumption of conditional independence, meaning, given $b_i$, the responses $Y_{i1},\ Y_{i2},\ \ldots,\ Y_{ip}$ are mutually independent.

Mixed effects models are most useful when the scientific objective is to make inferences about individuals rather than the population averages. These models focus mainly on the individuals and the influence of covariates on them. Regression parameters, $\beta$, measure the direct influence of covariates on the responses of heterogeneous individuals. We assume that the data for a single subject are independent observations from a distribution belonging to the exponential family, and the regression coefficients can vary from subject to subject, based on the random effects

distribution.

#### 2.2.3.1 Generalized Linear Mixed Model for a Continuous Response

Suppose that $Y_{ij}$ is a continuous response and the changes in the mean response over time are related to the covariates. The conditional mean, $Y_{ij}$, depends on the fixed and random effects as follows:

$$\eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i, \tag{2.20}$$

where $X'_{ij} = Z'_{ij} = (1, t_{ij})$, with

$$\begin{aligned} E(Y_{ij}\,|b_i) &= \eta_{ij} = X'_{ij}\beta + Z'_{ij}\, b_i \\ &= \eta_{ij} = \beta_1 + \beta_2\, t_{ij} + b_{1i} + b_{2i}\, t_{ij} \\ &= \eta_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})\, t_{ij} \end{aligned} \tag{2.21}$$

The conditional mean of $Y_{ij}$ is related to the linear predictor by an identity link function $\eta_{ij} = gE(Y_{ij}\,|b_i) = E(Y_{ij}\,|b_i)$. Hence, $\varphi = \sigma^2$.

The $Y_{ij}$ are independent and assumed to have a normal distribution, with $\mathrm{Var}(Y_{ij}\,|b_i) = \sigma^2$, which does not depend on the conditional mean.

#### 2.2.3.2 Generalized Linear Mixed Model for Counts

Suppose that $Y_{ij}$ is a count. Thus, the conditional mean of $Y_{ij}$ depends on the fixed and random effects as follows:

$$\eta_{ij} = {X'}_{ij}\beta + {Z'}_{ij}\, b_i, \tag{2.22}$$

where ${X'}_{ij} = {Z'}_{ij} = (1, t_{ij})$, with

$$\log E(Y_{ij}\,|b_i) = \eta_{ij} = {X'}_{ij}\beta + {Z'}_{ij}\, b_i \tag{2.23}$$

The conditional mean of $Y_{ij}$ is related to the linear predictor by a log link function. The $Y_{ij}$ are independent and assumed to have a Poisson distribution, with $\mathrm{Var}(Y_{ij}\,|b_i) = E(Y_{ij}\,|b_i)$, hence $\varphi = 1$.

#### 2.2.3.3 Generalized Linear Mixed Model for a Binary Response

Suppose that $Y_{ij}$ is a binary response taking values of 0 to 1. The conditional mean of $Y_{ij}$ depends on the fixed and random effects as follows:

$$\eta_{ij} = {X'}_{ij}\beta + {Z'}_{ij}\, b_i = {X'}_{ij}\beta + b_i, \tag{2.24}$$

where $Z_{ij} = 1$ for all $i = 1, \ldots, N$, and $j = 1, \ldots, n_i$, with

$$\log\left\{\frac{\Pr\left(Y_{ij} = 1|b_i\right)}{\Pr\left(Y_{ij} = 0|b_i\right)}\right\} = \eta_{ij} = {X'}_{ij}\beta + b_i \tag{2.25}$$

The conditional mean of $Y_{ij}$ is related to the linear predictor by a logit link function. The single random effect $b_i$ is assumed to have a univariate normal distri-

bution with zero mean. The $Y_{ij}$ are independent and have a Bernoulli distribution, with $\text{Var}(Y_{ij}|b_i) = E(Y_{ij}|b_i)\,1 - E(Y_{ij}|b_i)$, hence $\varphi = 1$.

#### 2.2.3.4 Transitional (Markov) Models

Transitional models give us the opportunity to express the joint distribution as a series of conditional distributions

$$f(Y_{i1},\ Y_{i2},\ldots,\ Y_{ip}) = f(Y_{i1})\ \ f(Y_{i1}|Y_{i2})\ldots f(Y_{ip}|Y_{i1},\ldots,\ Y_{i,p-1}) \tag{2.26}$$

This is known as a transitional model, since it represents the probability distribution at each time point as conditional on the past and provides a complete representation of the joint distribution. In transitional models the conditional distribution of each response is expressed as an explicit function of the past responses and covariates. Therefore, the correlation between the repeated responses can be said to be arising due to the past values of responses explicitly influencing the present observations. In other words, the present outcomes depend on the past values. The explicit function of past responses at the $j^{th}$ occasion can be denoted as $H_{ij} = \{Y_{i1},\ldots,\ Y_{i,j-1}\}$.

## 2.3 Covariance Structure

One of the defining features of longitudinal data is that they are correlated since measurements on the same subjects are taken repeatedly over time. When we are modelling longitudinal data, we have to pay attention to the mean response over time and the covariance among repeated measures on the same individuals. Since the vector of residuals (observed responses minus fitted responses) depends on the specification of the model for the mean, these two mentioned aspects are interrelated.

Therefore, the covariance between any pair of residuals depends on the model for the mean (depends on $\beta$). For modelling the covariance or correlation among repeated measures, we have to keep in mind that longitudinal data are not only correlated, but for the most part, they are also positively correlated. Also, the positive correlation among the repeated measures can be used to advantage in the study of change over time, since our main focus is to analyze the change in the mean response over time. All subjects share a common variance matrix and correlation matrix to characterize the average dependence among repeated observations eve though the covariance matrix may vary from subject to subject. Let $\text{Cov}(Y_i,\ Y_i)$ represent the covariance structure between the $i^{th}$ subject. Let $\text{Corr}(Y_{it}, Y_{is})$ represent the correlation between times t and s of $i^{th}$ subject.

Thus, the covariance structures that characterized the correlated data within subjects are defined as

$$Cov\{Y_i, Y_i\} = E\{((Y_i) - E(Y_i))((Y_i) - E(Y_i))^{'} \tag{2.27}$$

and its matrix can be defined as:

$$\text{Cov}(Y_i) = \begin{bmatrix} \text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \text{Cov}(Y_{i1}, Y_{i3}) & \dots & \text{Cov}(Y_{i1}, Y_{in}) \\ \text{Cov}(Y_{i2}, Y_{i1}) & \text{Var}(Y_{i2}) & \text{Cov}(Y_{i2}, Y_{i3}) & \dots & \text{Cov}(Y_{i2}, Y_{in}) \\ \dots & \dots & \dots & \dots & \dots \\ \text{Cov}(Y_{in}, Y_{i1}) & \text{Cov}(Y_{in}, Y_{i2}) & \text{Cov}(Y_{i1}, Y_{i3}) & \dots & \text{Var}(Y_{in}) \end{bmatrix} \tag{2.28}$$

We can define the correlation matrix as

$$\text{Corr}(Y_i) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1n} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{n1} & \rho_{n2} & \rho_{n3} & \dots & 1 \end{bmatrix}. \tag{2.29}$$

The matrix is symmetric, thus $\text{Corr}(Y_{is}, Y_{it}) = \rho_{st} = \rho_{ts} = \text{Corr}(Y_{it}, Y_{is})$.
There are other covariance structures and their corresponding correlation matrix that are useful when dealing with longitudinal data (Guo, 2011).

**Compound Symmetry**

Compound symmetry is among the first covariance models used for the analysis of repeated measures data. In this model, we assume that the variance, $\sigma^2$, is constant across all occasions, and $\text{Corr}\,(Y_{ij},\ Y_{ik}\ ) =\ \rho$ for all j and k.

$$\text{Cov}(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \dots & \dots & \dots & \dots & \dots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}, \tag{2.30}$$

where $\rho \geq 0$.

The compound symmetry covariance is rather simple, with only two parameters despite the number of measurement occasions. However, the downfall of this model is the fact that it does make the assumption that the correlation between any pair of measurements is the same regardless of the time interval between the measurements.

The constraint on the correlation among repeated measurements brings up some

problems since correlations for most longitudinal data, are expected to decay with increasing separation in time. In addition, the assumption of constant variances across time is unrealistic. The real-life examples of longitudinal studies have shown that variances are rarely constant over time for a set a data.

**Toeplitz**

The Toeplitz covariance pattern makes the assumption that any pair of responses that are equally separated in time have the same correlation. In addition, we also assume that the variance, $\sigma^2$, is constant across all occasions, and $corr\,(Y_{ij},\ Y_{ij+k}\ ) =\ \rho_k$ for all j and k.

$$\mathrm{Cov}(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{bmatrix}. \tag{2.31}$$

The downfall of this model is the fact that it does make the assumption that the correlation between any pair of measurements is the same regardless of the time interval between the measurements.

**Autoregressive**

The autoregressive model for the covariance makes the assumption that the variance, $\sigma^2$, is constant across all occasions, and $\mathrm{Corr}\,(Y_{ij},\ Y_{ij+k}\ ) =\ \rho^k$ for all j and k, and $\rho \geq 0$.

$$\mathrm{Cov}(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{bmatrix}. \tag{2.32}$$

The autoregressive covariance has only two parameters, regardless of the number of measurements occasions. As with Toeplitz covariance, this covariance structure is only appropriate when the measurements that are made at equal (or approximately equal) intervals of time. Note that the Toeplitz covariance has n parameters (1 variance parameter and n-1 correlation parameter). Therefore, the first-order Autoregressive model can be considered a special case of Toeplitz covariance. Both structures have constant variance, but under the AR(1) structure the correlations change over time according to the power function. The Toeplitz structure does not have this requirement.

The "first-order" autoregressive process happens when there is dependence only on the previous error. Similarly, the "second-order" model is structured when the dependence is on the two previous errors, and so on. Thus, the autoregressive covariance can be thought of as resulting from a process where the error term at the $i^{th}$ occasion is a deterministic function of the error at the previous occasion, $\rho$, $e_{i.j-1}$, meaning that the recent past predicts the present, in addition to an independent source of random error, $\omega_{ij}$.

$$e_{ij} = \rho\ e_{ij-1} + \ \omega_{ij}, \tag{2.33}$$

where $\omega_{ij} \sim N(0, \sigma^2 [1 - \rho^2])$ and the process is initiated by an error, $e_{i0}$, where $e_{i0} \sim N(0, \sigma^2)$.

$$\text{Var}(e_{ij}) = \sigma^2 \tag{2.34}$$

$$\text{Cov}(e_{ij}, e_{ik}) = \sigma^2 \rho^{|j-k|} \tag{2.35}$$

**Exponential**

If the measurement occasions are not equally spaced over time, we can rearrange the autoregressive covariance model as follows: let $t_{i1}, \ldots, t_{in}$ denote the observation times for the $i^{th}$ individual and assume that the variance, $\sigma^2$, is constant across all measurements,

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}, \tag{2.36}$$

where $\rho \geq 0$.

The correlation between any pair of repeated measures decreases exponentially with the time separations between them. The exponential covariance model can be re-written as

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij} - t_{ik}|} \\ &= \sigma^2 \exp(-\theta |t_{ij} - t_{ik}|) \end{aligned} \tag{2.37}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$.

A notable feature of this model is that it is invariant under linear transformations of the time scale. If we replace $t_{ij}$ by $(a + bt_{ij})$, meaning, if we replace time measured in weeks by time measured in days, the same form for the covariance matrix holds.

If the measurements are made repeatedly at the same occasion, such as when replicate measurements on an individual are gathered, we would assume that the correlation is one and that it would rapidly decrease to zero as the time separation between the measurements increase.

The troublesome aspect of this model comes from the assumption that the responses are measured without errors. This assumption would be unrealistic in most longitudinal studies particularly when it comes to the health science experiments. In addition, in longitudinal studies, correlation among repeated measurements that decay to zero are rarely observed. The mentioned covariance structures can all be found in Fitzmaurice et al. (2004).

## 2.4 Tweedie exponential dispersion models

An exponential dispersion model (EDM) is a family of distributions that has two parameters: a linear exponential family and an additional dispersion parameter. These models are especially important in dealing with different response types and are the response distributions for generalized linear models. In real life, some data can have both continuous as well as discrete values. Jørgensen (1997) was the one to propose the Tweedie exponential dispersion models to deal with these types of data (Lall, 2014). An exponential dispersion model is characterized by its variance function $V$, which describes the mean-variance relationship of the distribution when the dispersion is constant. If Y has an exponential dispersion model (EDM) distribution with

mean $\mu$ and dispersion parameter $\phi$, then we can show the variance function $V$ has the form

$$V(Y) = \phi V(\mu) \tag{2.38}$$

There is special interest in exponential dispersion models with power mean-variance relationship. Tweedie distributions, $Tw_p(\mu, \sigma^2)$, have an expected mean of $E(Y) = \mu$ and variance $V(Y) = \phi\mu^p$, where p is a parameter that controls the variance of the distribution Jørgensen (1987a).

By incorporating distribution-free random effects into the Tweedie models, we can work more freely and more flexibly with longitudinal data and the various correlation structures of it; more details can be found in Ma and Jørgensen (2007). A random variable Y with a Tweedie distribution $Tw_p(\mu, \sigma^2)$ is defined as follows

$$f_p(y; \mu, \sigma^2) = \begin{cases} c_p(y; \sigma^2) \exp\left\{\frac{1}{\sigma^2}\left(\frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right)\right\} & \text{if } p \neq 1, 2, \\ c_2(y; \sigma^2) \exp\left[-\frac{1}{\sigma^2}\left\{\frac{y}{\mu} + \log(\mu)\right\}\right] & \text{if } p = 2, \\ c_1(y) \exp\left[y\log(\mu) - \mu\right] & \text{if } p = 1, \end{cases}$$

where the explicit $c_p(y; \sigma^2)$ are given by Jørgensen (1987, 1997). The density does not exist for $0 < p < 1$. For $1 < p < 2$, the Tweedie distribution corresponds to the compound Poisson distribution which accounts for continuous data with exact zeroes. Some special cases of the Tweedie distribution are as follows: for p=0, the Tweedie distribution becomes the normal distribution, for p=1, we get the Poisson distribution, for p=2 we get the gamma distribution and for p=3, we get the inverse Gaussian.

# Chapter 3

# Unequally Spaced Longitudinal Data

## 3.1 Model Specification

Let the outcome variable $y_{ij}$ represent the response at the $j^{th}$ measurement occasion on the $i^{th}$ subject. Each response is taken at time $t_{ij}$. Let $i = 1, \ldots, N$ represent the number of independent subjects and $j = 1, \ldots, n_i$ represent the number of measurement occasions for each individual subject.

Let $Y = (Y_{11}, Y_{12} \ldots, Y_{1n_1}, Y_{21}, Y_{22}, \ldots, Y_{N1}, \ldots, Y_{Nn_N})'$.

We consider the subject specific random effect $U_i$ for the response of the $i^{th}$ subject and the time specific random effect $V_{ij}$ for the response from the $j^{th}$ measurement occasion of the $i^{th}$ subject. Let $W = (U', V')'$ denote the vector of the random effects where $U = (U_1, \ldots, U_i, \ldots, U_N)'$ and $V = (V'_1, \ldots, V'_i, \ldots, V'_N)'$ with $V_i = (V_{i1}, \ldots, V_{ij}, \ldots, V_{in_i})'$ respectively.

Table 3.1: Layout of a longitudinal data set

| Subject (i) | Repeated measurement (j) | $t_{ij}$ | $y_{ij}$ | $x_{ijk}$ |
|---|---|---|---|---|
| 1 | 1 | $t_{11}$ | $y_{11}$ | $x_{111} \cdots x_{11p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| 1 | $n_1$ | $t_{1n_1}$ | $y_{1n_1}$ | $x_{1n_11} \cdots x_{1n_1p}$ |
| 2 | 1 | $t_{21}$ | $y_{21}$ | $x_{211} \cdots x_{21p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| 2 | $n_2$ | $t_{2n_2}$ | $y_{2n_2}$ | $x_{2n_21} \cdots x_{2n_2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| N | 1 | $t_{N1}$ | $y_{N1}$ | $x_{N11} \cdots x_{N1p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | |
| N | $n_N$ | $t_{Nn_N}$ | $y_{Nn_N}$ | $x_{Nn_N1} \cdots x_{Nn_Np}$ |

### 3.1.1 Assumption 1

Subject-specific random effects $U_1, \ldots, U_i, \ldots, U_N$ are positive, independently and identically distributed with mean 1 and variance $\sigma^2$. That is

$$E(U_i) = 1 \quad \text{and} \quad \text{Var}(U_i) = \sigma^2.$$

### 3.1.2 Assumption 2

Given the subject-specific random effects $U$, the moment structure of the time-specific positive random effects $V$ can be expressed as

$$E(V_{ij}|U) = U_i \qquad \text{and} \qquad \text{Cov}(V_{ij}, V_{i'j'} \mid U) \;=\; \begin{cases} \tau^2 \rho_{(j,j')}\, U_i & \text{if } i = i' \\ 0 & \text{if } i \neq i' \end{cases}$$

with $\rho_{(j,j')} = 1$ for $j = j'$ .

### 3.1.3 Assumption 3

Given the random effects $W$, the components of $Y$ are conditionally independent, and the conditional distribution of $Y_{ij}$, given $W$, depends on $V_{ij}$ only, which is

$$\mathrm{Y}_{ij} \mid W \quad \sim \quad \mathrm{Tw}_p(\mu_{ij}V_{ij}, \epsilon^2 V_{ij}^{1-p})$$

where $\mu_{ij} = \exp(x'_{ij}\beta)$ with vector of covariates $x_{ij}$ and regression parameter vector $\beta$.

The conditional expectation is $E(Y_{ij} \mid W) = \mu_{ij}V_{ij}$, therefore as $E(U_i) = E(V_{ij}) = 1$, the unconditional expectation is $E(Y_{ij}) = EE(Y_{ij}|W) = \mu_{ij}E(V_{ij}) = \mu_{ij}$.

The Tweedie family can also be called the power-variance family. Meaning that if $X \sim \mathrm{Tw}_p(\mu, \sigma^2)$, then $\mathrm{Var}(X) = \sigma^2\mu^p$. Therefore, $E(Y_{ij} \mid W) = \mu_{ij}V_{ij}$ and $\mathrm{Var}(\mathrm{Y}_{ij} \mid W) = \epsilon^2 V_{ij}^{1-p}(\mu_{ij}V_{ij})^p = \epsilon^2\mu_{ij}^p V_{ij}$.

## 3.2 Moment structure

The unconditional expectation of the response $Y_{ij}$ can be expressed as

$$E(Y_{ij}) = EE(Y_{ij} \mid W) = \mu_{ij}E(U_i) = \mu_{ij} \tag{3.1}$$

The unconditional variance of the responses $Y_{ij}$ has the form

$$\begin{aligned}
\mathrm{Var}(Y_{ij}) &= E\{\mathrm{Var}(Y_{ij} \mid W)\} + \mathrm{Var}\{E(Y_{ij} \mid W)\} \\
&= \epsilon^2 \mu_{ij}^p E(V_{ij}) + \mu_{ij}^2 \mathrm{Var}(V_{ij}) \\
&= \epsilon^2 \mu_{ij}^p E(V_{ij}) + \mu_{ij}^2 \{\mathrm{Var}(E(V_{ij} \mid U)) + E(\mathrm{Var}(V_{ij} \mid U))\} \\
&= \epsilon^2 \mu_{ij}^p + \mu_{ij}^2 (\sigma^2 + \tau^2) \qquad (3.2)
\end{aligned}$$

The covariances between the response $Y_{ij}$ and $Y_{i'j'}$ is:

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = E\{\mathrm{Cov}(Y_{ij}, Y_{ij'}) \mid W)\} + \mathrm{Cov}\ E(Y_{ij} \mid W), E(Y_{i'j'} \mid W)\} \qquad (3.3)$$

$If\ i = i',\ j = j'$

$$\begin{aligned}
\mathrm{Cov}(Y_{ij}, Y_{i'j'}) &= \mathrm{Var}(Y_{ij}) \\
&= \epsilon^2 \mu_{ij}^p + \mu_{ij}^2 (\sigma^2 + \tau^2 \rho_{(j,j')}) \qquad (3.4)
\end{aligned}$$

$If\ i = i',\ j \neq j'$

$$\begin{aligned}
\mathrm{Cov}(Y_{ij}, Y_{i'j'}) &= E\{\mathrm{Cov}(Y_{ij}, Y_{ij'}) \mid W)\} + \mathrm{Cov}\{E(Y_{ij} \mid W), E(Y_{i'j'} \mid W)\} \\
&= \mu_{ij} \mu_{ij'} (\sigma^2 + \tau^2 \rho_{(j,j')}) \qquad (3.5)
\end{aligned}$$

*Otherwise*

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = 0 \qquad (3.6)$$

Therefore, the unconditional covariance structure of the response is obtained as

follows:

$$\mathrm{Cov}(Y_{ij}, Y_{i'j'}) = \begin{cases} \epsilon^2\mu_{ij}^p + \mu_{ij}^2(\sigma^2 + \tau^2\rho_{(j,j')}) & \text{if } i = i', j = j' \\ \mu_{ij}\mu_{ij'}(\sigma^2 + \tau^2\rho_{(j,j')}) & \text{if } i = i', j \neq j' \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

## 3.3 The best linear unbiased predictors of random effects

Let the inverse of the covariance matrix of $Y$ be denoted as $\mathrm{Cov}^{-1}(Y)$. The best linear unbiased predictors of random effects (BLUP) of subject-specific random effects $U$ given $Y$ can be shown as follows:

$$\begin{aligned} \hat{U} &= E(\ U) + \mathrm{Cov}(\ U,\ Y)\mathrm{Cov}^{-1}(\ Y)\{\ Y - E(Y)\} \\ \hat{U} &= E(\ U) + \mathrm{Cov}(\ U)\ B'\mathrm{Cov}^{-1}(\ Y)\{\ Y - E(Y)\} \end{aligned} \quad (3.8)$$

Where the vector $E(U) = (1, \ldots, 1, \ldots, 1)'$ is of dimension $N \times 1$. The covariance matrix $\mathrm{Cov}(\ U)$ is a $N \times N$ diagonal matrix of $\sigma^2$. The matrix B is an $Nn_i \times N$ block diagonal matrix with column vectors $B_i = (\mu_{i1}, \ldots,\ \mu_{ij}, \ldots,\ \mu_{in_i})'$, $i = 1, \ldots,\ N$ as its diagonal blocks.

We can derive the orthodox BLUP of time-specific random effects $V$ given the response $Y$ as follows:

$$\begin{aligned} \hat{V} &= E(\ V) + \mathrm{Cov}(\ V,\ Y)\mathrm{Cov}^{-1}(\ Y)\{\ Y - E(\ Y)\} \\ &= E(\ V) + \mathrm{Cov}(\ V)\mathbf{D}'\mathrm{Cov}^{-1}(\ Y)\{\ Y - E(\ Y)\} \end{aligned} \quad (3.9)$$

Where the vector $E(V) = (1, \ldots, 1, \ldots, 1)'$ is of dimension $Nn_i \times 1$. The covariance matrix Cov( $V$) is a $Nn_i \times Nn_i$ diagonal matrix with elements $\mathrm{Cov}(V_{ij}, V_{i'j'})$ which can be shown as

$$\begin{aligned} \mathrm{Cov}(V_{ij}, V_{i'j'}) &= E\{\mathrm{Cov}(V_{ij}, V_{i'j'}) \mid \boldsymbol{U})\} + \mathrm{Cov}\{E(V_{ij} \mid \boldsymbol{U}), E(V_{i'j'}) \mid \boldsymbol{U})\} \\ &= E\{\mathrm{Cov}(V_{ij}, V_{i'j'}) \mid \boldsymbol{U})\} + \mathrm{Cov}(U_i, U_{i'}) \end{aligned} \tag{3.10}$$

*If* $i = i'$

$$\begin{aligned} \mathrm{Cov}(V_{ij}, V_{ij}) &= E\{\mathrm{Cov}(V_{ij}, V_{i'j'}) \mid \boldsymbol{U})\} + \sigma^2 \\ &= E(\tau^2 \rho_{(j,j')} U_i) + \sigma^2 \\ &= \tau^2 \rho_{(j,j')} \times 1 \times E(U_i) + \sigma^2 \\ &= \tau^2 \rho_{(j,j')} + \sigma^2 \end{aligned} \tag{3.11}$$

*Otherwise*

$$\begin{aligned} \mathrm{Cov}(V_{ij}, V_{i'j'}) &= E\{\mathrm{Cov}(V_{ij}, V_{i'j'} \mid \boldsymbol{U})\} + \sigma^2 \\ &= 0 \end{aligned} \tag{3.12}$$

In summary:

$$\mathrm{Cov}(V_{ij}, V_{i'j'}) = \begin{cases} \tau^2 \rho_{(j,j')} + \sigma^2 & \text{if } i = i', j \neq j' \\ 0 & \text{otherwise} \end{cases} . \tag{3.13}$$

The matrix D is an $Nn_i \times Nn_i$ diagonal matrix of the vector $(\mu_{11}, \ldots, \mu_{ij}, \ldots, \mu_{Nn_i})\prime$.

The estimating equations for the regression parameters can be constructed by using the linear predictors of V and Y.

The BLUP provides a common method for estimating regression parameters and random effects parameters for all Tweedie mixed models. In fact, the orthodox BLUP approach specifically includes random effects and combines the "subject-specific" and the "population-averaged" inferences within a common model (Lee and Nelder, 1996).

## 3.4 Estimation of Parameters

### 3.4.1 Estimation of regression parameters

$$\psi(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \boldsymbol{X}'_{ij} \frac{\mu_{ij}^{1-p}(\boldsymbol{\beta})}{\epsilon^2} \left[ y_{ij} - \hat{V}_{ij}(\boldsymbol{\beta}) \mu_{ij}(\boldsymbol{\beta}) \right] \tag{3.14}$$

As indicated by Ma and Jørgensen (2007), the estimating equation $\psi(\boldsymbol{\beta}) = 0$ can be solved iteratively, where the value of $\boldsymbol{\beta}$ is updated by $\boldsymbol{\beta}^* = \boldsymbol{\beta} - \mathrm{S}^{-1}(\boldsymbol{\beta})\psi(\boldsymbol{\beta})$, with the explicit expression of the sensitivity matrix given by

$$\mathrm{S}(\boldsymbol{\beta}) = -\boldsymbol{X}' \mathrm{diag}\left\{E(\mathbf{Y})\right\} \mathrm{Cov}^{-1}(\mathbf{Y}) \mathrm{diag}\left\{E(\mathbf{Y})\right\} \boldsymbol{X} \tag{3.15}$$

with $E(\mathbf{Y}) = \{E(Y_{11}), \ldots, E(Y_{Nn_N})\}'$ and $\psi(\boldsymbol{\beta}) = \boldsymbol{X}^{\boldsymbol{T}} diag \boldsymbol{E}(\boldsymbol{Y}) \mathrm{Cov}^{-1}(\mathbf{Y}) \{\mathbf{Y} - E(\mathbf{Y})\}$.

### 3.4.2 Estimation of random effects parameters

The random effects parameters and correlation parameters can be estimated using adjusted Pearson estimators (Ma, 1999). A bias correction is introduced to account

for the use of the orthodox BLUP, rather than the true random effects, in estimating these parameters. Iterative expressions for these estimators were derived in Li (2018) for the case of multivariate, correlated responses, and the estimates for this univariate model are simply a special case of these and are reproduced here. The iterative equation for estimating $\sigma^2$ is

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{i=1}^{N}\left\{(\hat{U}_i - 1)^2 + \hat{\sigma}_{r-1}^2 - \hat{\sigma}_{r-1}^4\ \mu_i{}'\text{var}(\ Y_i)^{-1}\ \mu_i\right\}, \tag{3.16}$$

where $\hat{\sigma}_{r-1}^2$ is the estimate from the previous iteration.

The iterative equation for estimating $\tau^2$ and $\epsilon^2$ can be expressed as

$$\begin{aligned}\hat{\tau}_r^2 \;=\; & \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\Big\{(\hat{V}_{ij} - \hat{U}_i)^2 + \hat{\tau}_{r-1}^2 - \sigma^4\ \mu_i{}'\text{var}(\ Y_i)^{-1}\ \mu_i - \text{cov}(V_{ij},\ Y_i) \\ & \text{var}(\ Y_i)^{-1}\text{cov}(\ Y_i, V_{ij}) + 2\text{cov}(U_i,\ Y_i)\text{var}(\ Y_i)^{-1}\text{cov}(\ Y_i, V_{ij})\Big\},\end{aligned} \tag{3.17}$$

and

$$\begin{aligned}\hat{\epsilon}_r^2 \;=\; & \frac{1}{J}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{1}{\mu_{ij}^p}\Big\{(y_{ij} - \mu_{ij}\hat{V}_{ij})^2 + \mu_{ij}^2(\sigma^2 + \tau^2) \\ & -\mu_{ij}^2\text{cov}(V_{ij},\ Y_i)\text{var}(\ Y_i)^{-1}\text{cov}(\ Y_i, V_{ij})\Big\},\end{aligned} \tag{3.18}$$

respectively. $J$ is the sum of the total number of observations for all individuals. The explicit forms for the estimators of $\sigma^2$, $\tau^2$ and $\epsilon^2$ are similar to those presented in Ma and Jørgensen (2007). In next subsection we present the estimation of the correlation parameters $\rho_{(j,j')}$.

### 3.4.3 Estimation of correlation parameter under the flexible AR(1) correlation structure

For the autoregressive of order 1 (AR(1)) structure, we can set $\rho_{(j,j')} = \rho^{|j-j'|}$ for any $j \neq j'$. Under the AR(1) structure, the correlation matrix has the form

$$\mathfrak{R} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{J-1} \\ \rho & 1 & \rho & \dots & \rho^{J-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{J-1} & \rho^{J-2} & \rho^{J-3} & \dots & 1 \end{bmatrix}. \tag{3.19}$$

For the unstructured correlation structure $\rho_{(j,j')}$ is

$$\begin{aligned} \rho_{(j,j')} &= \frac{\text{cov}\left\{(V_{ij} - U_i)(V_{ij'} - U_i)\right\}}{\left[\left\{\text{var}(V_{ij} - U_i)\right\}\left\{\text{var}(V_{ij'} - U_i)\right\}\right]^{1/2}} \\ &= \frac{\text{cov}\left\{(\hat{V}_{ij} - \hat{U}_i)(\hat{V}_{ij'} - \hat{U}_i) + b_{i(j,j')}\right\}}{\left[\left\{\text{var}(\hat{V}_{ij} - \hat{U}_i) + b_{i(j,j)}\right\}\left\{\text{var}(\hat{V}_{ij'} - \hat{U}_i) + b_{i(j',j')}\right\}\right]^{1/2}} \end{aligned} \tag{3.20}$$

We can estimate the time specific random effect $V_{ij}$ as following, where the matrix for each individual is set up separately, taking into account the differing measurement occasions

$$\hat{V}_{ij} = \text{cov}(V_{ij},\ Y_i)\text{var}(\ Y_i)^{-1}\text{matrix}(Y_i - \mu_i,\ J_i, 1)$$

We estimate the subject specific random effect $U_i$ by the following:

$$\hat{U}_i \;=\; (\sigma^2\mu_{ij})\mathrm{var}(\mathrm{Y}_i)^{-1}(\mathrm{Y}_i - \mu_i)$$

To estimate $\rho$ under AR(1) structure, it would be sufficient to estimate lag 1 ($\rho^1 = \rho$) correlation only:

$$\hat{\rho} \;=\; \frac{\sum_{i=1}^{N}\sum_{j=1}^{J-1}\left\{(\hat{V}_{ij}-\hat{U}_i)(\hat{V}_{i(j+1)}-\hat{U}_i)+b_{i(j,j+1)}\right\}}{\left[\left\{\sum_{i=1}^{N}\sum_{j=1}^{J-1}(\hat{V}_{ij}-\hat{U}_i)^2+b_{i(j,j)}\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{J-1}(\hat{V}_{i(j+1)}-\hat{U}_i)^2+b_{i(j+1,j+1)}\right\}\right]^{1/2}}. \tag{3.21}$$

where $b_{i(j,j')}$ is the correction term which can be simplified as

$$\begin{aligned} b_{i(j,j')} \;=\; & \rho_{(j,j')}\tau^2 - \Big\{[\mathrm{cov}(\hat{V}_{ij},\hat{V}_{ij'}) + \sigma^2\ \mu_i{}'\mathrm{var}(\ Y_i)^{-1}\left[\sigma^2\ \mu_i\right] \\ & -\mathrm{cov}(\ Y_i, V_{ij}) - \mathrm{cov}(\ Y_i, V_{ij'})]\Big\}. \end{aligned}$$

.

# Chapter 4

# Data Analysis

## 4.1 Patient-controlled analgesia data

The patient-controlled analgesia (PCA) dataset was presented by Henderson and Shimacura (2003) where hospital patients were given the opportunity to control their own pain relief medication following an abdominal surgery. The patients were in control of releasing the drug at any given time, however after each use there would be a lock-out time where the patients had to wait in order to have access to the drug again. The whole experiment was observed for all the patients over the first 48 hours following their surgery. There were 12 measurement occasions observed for all individuals (a measurement occasion was recorded at every 4 hours). The outcome of the experiments were the numbers of times (counts) a patients requested for a dosage of the drug within each of the time periods of 4 hours.

sixty-five patients were divided into two groups. Group one contained thirty patients, where each individual was provided with a bolus of pain-relief drug of 1 mg morphine. The lock-out time for this group was set to be four minutes. Group two contained thirty-five patients, where a dose of 2 mg morphine was provided to them.

Their lock-out time was set at 8 minutes.

In this longitudinal setup, as opposed to an equi-spaced count responses, there were instances were an individual patient provided successive count responses at unevenly spaced time intervals. In this example, since the time interval is 12 (t=12 4-hours as the duration of the study), if an individual only reports for 11 out of the 12 time intervals (in terms of counts), the 11 count response measurements collected would be considered unevenly spaced, and the study to be complete. The 'no response' value for the individual is neither a missing response, nor can be qualified as a zero, since no probability can be assigned for a non-existing event. We use a Poisson Distribution, conditional of random effects $W = (U',\ V')$, which correspond to $p = 1$.

Let the outcome variable $y_{ij}$ represent the response of the $i^{th}$ subject at the $j^{th}$ measurement occasion. The observations are recorded at time $t_{ij}$ (time at the $j^{th}$ measurement occasion for the $i^{th}$ individual); therefore the $U_1, \ldots,\ U_i,\ \ldots,\ U_N$ are patient-specific random effects and $V_{i1}, \ldots,\ V_{ij},\ \ldots,\ V_{in_i}$ are time-specific random effects for the count responses of the drugs.

$$Y_{ij} \mid W \ \sim \ \mathrm{Tw}_p(\mu_{ij}V_{ij}, \epsilon^2 V_{ij}^{1-p})$$

$$y_{ij} \ \sim \ \mathrm{Poisson}(\mu_{ij})$$

The patient-specific random effects of $U_i$ and time-specific random effects of $V_{ij}$ as specified in assumptions 1 and 2 respectively are discussed in Chapter 3.

The two different regiments of medication application of our bolus drug (mor-

phine), are considered as the covariates. For the purpose of analysis, the covariate drug was coded as 1mg for patients in group 1, and 2mg for patients in group 2. The collection of regression parameters is denoted as $\beta = (\beta_0, \beta_1, \beta_2)$, where $\beta_0, \beta_1, \beta_2$ correspond to the intercept, group 1 and group 2 respectively.

Suppose that $\rho_{(j,j')}$ is the dynamic dependence parameter relating the responses $y_{ij}$ with its adjacent past response $y_{ij'}$.

The model is fit using the R Core Team (2020) software. To obtain the convergence, 1000 iterations were performed.

Figure 4.1 and Figure 4.2 help us view the marginal distribution of the response variable for the two groups in the experiment.

Figure 4.3 helps us visualize the response variable and its relationship with the covariate (dosage of drug) over time. It seems that overall, patients in group 2 had a higher number of counts across all time points than the patients in group 1. Overall, the average response seems to decrease over time for both groups.

The random effects parameters and correlation parameters can all be estimated using the before mentioned adjusted Pearson estimators (Ma, 1999). These estimators begin by assuming a method of moments approach to estimating the random effects parameters, as well as a known correlation structure. A bias correction is taken to account for the use of the orthodox BLUP, rather than the true random effects, in estimating these parameters. Iterative expressions for these estimators were derived in Li (2018) for the case of multivariate, correlated responses, and the estimates for this univariate model are simply a special case we have used here.

In such studies, the measurement occasions for each individual might differ, ranging from 1 to t. Thus, we are going to introduce a measurement index (Jidx) to our data. The unique length of the measurement index for each individual is

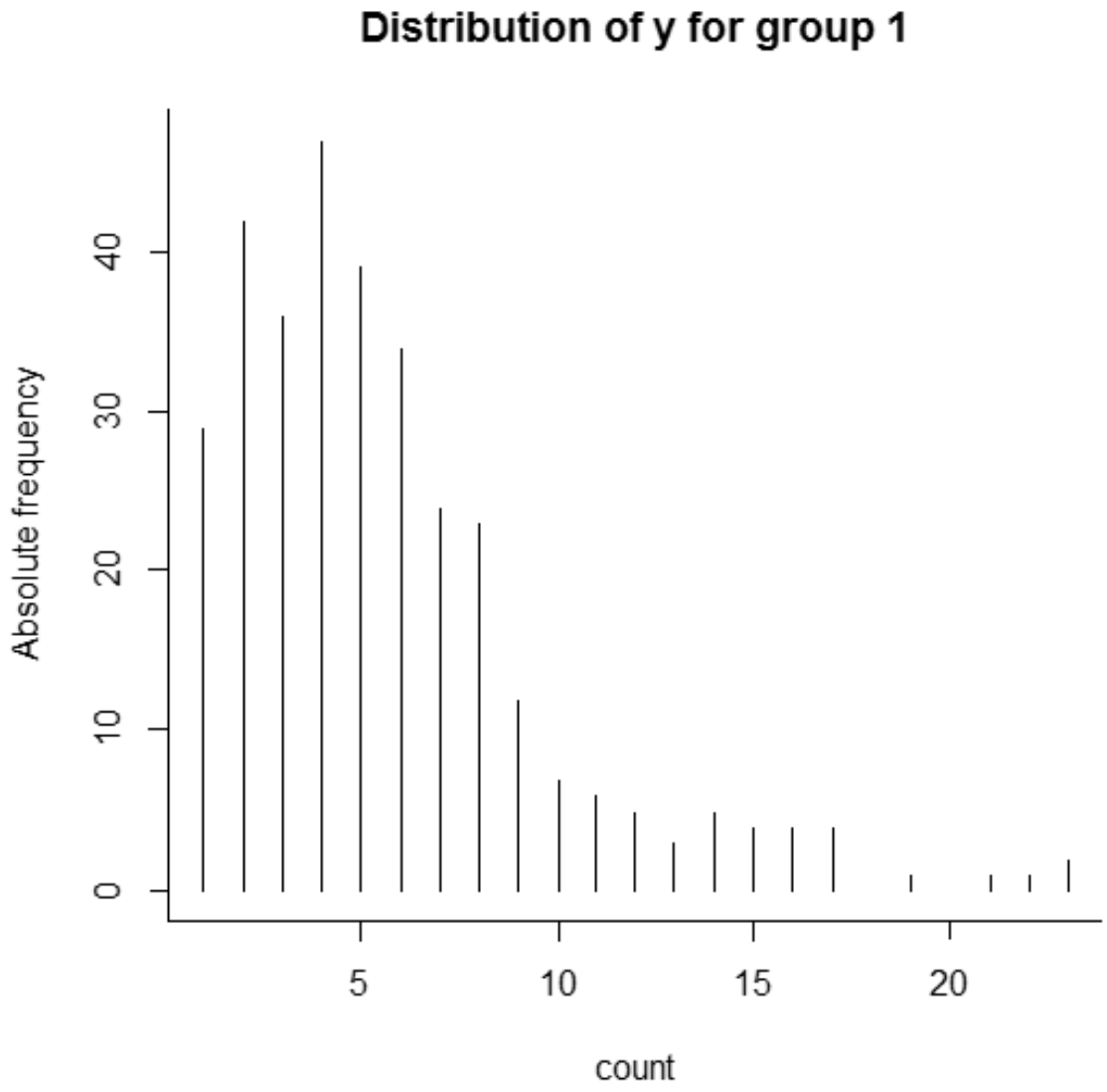


Figure 4.1: Distribution of the response variable for Group 1

specified as J[i]. Instead of setting up the correlation matrix as t by t, as we would do in an equally spaced study, we set up n correlation matrices of J[i] by J[i] for all individuals, each based on their number of measurements. Similarly, all random effects are calculated by incorporating each individual differing number of measure-

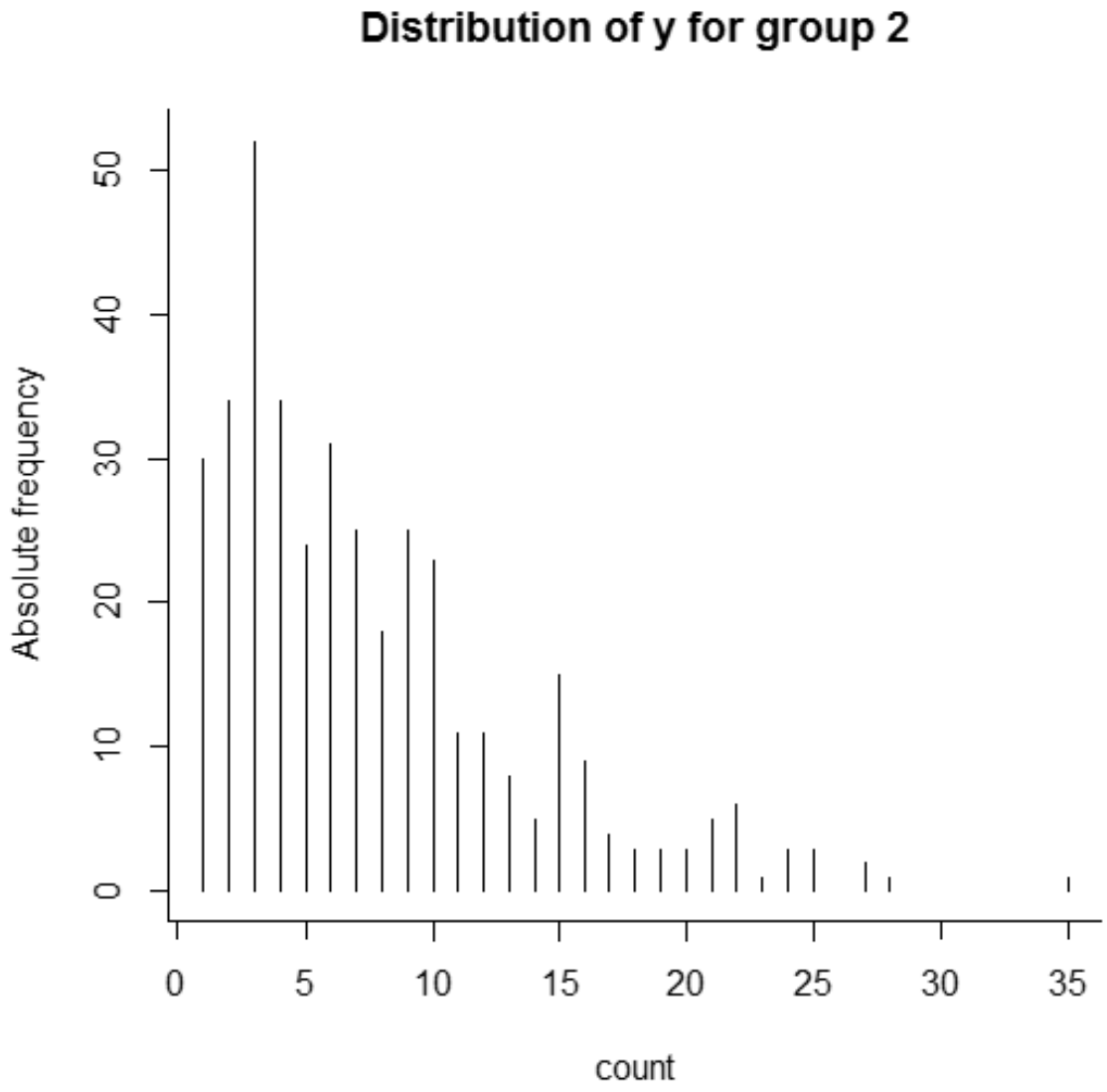


Figure 4.2: Distribution of the response variable for Group 2

ment occasions. This report will use an autoregressive structure of order 1, otherwise known as AR(1). In this structure, there is only one correlation parameter.

We first examine the equally spaced portion of the data. Then we look at the unequally spaced data, excluding the no responses.

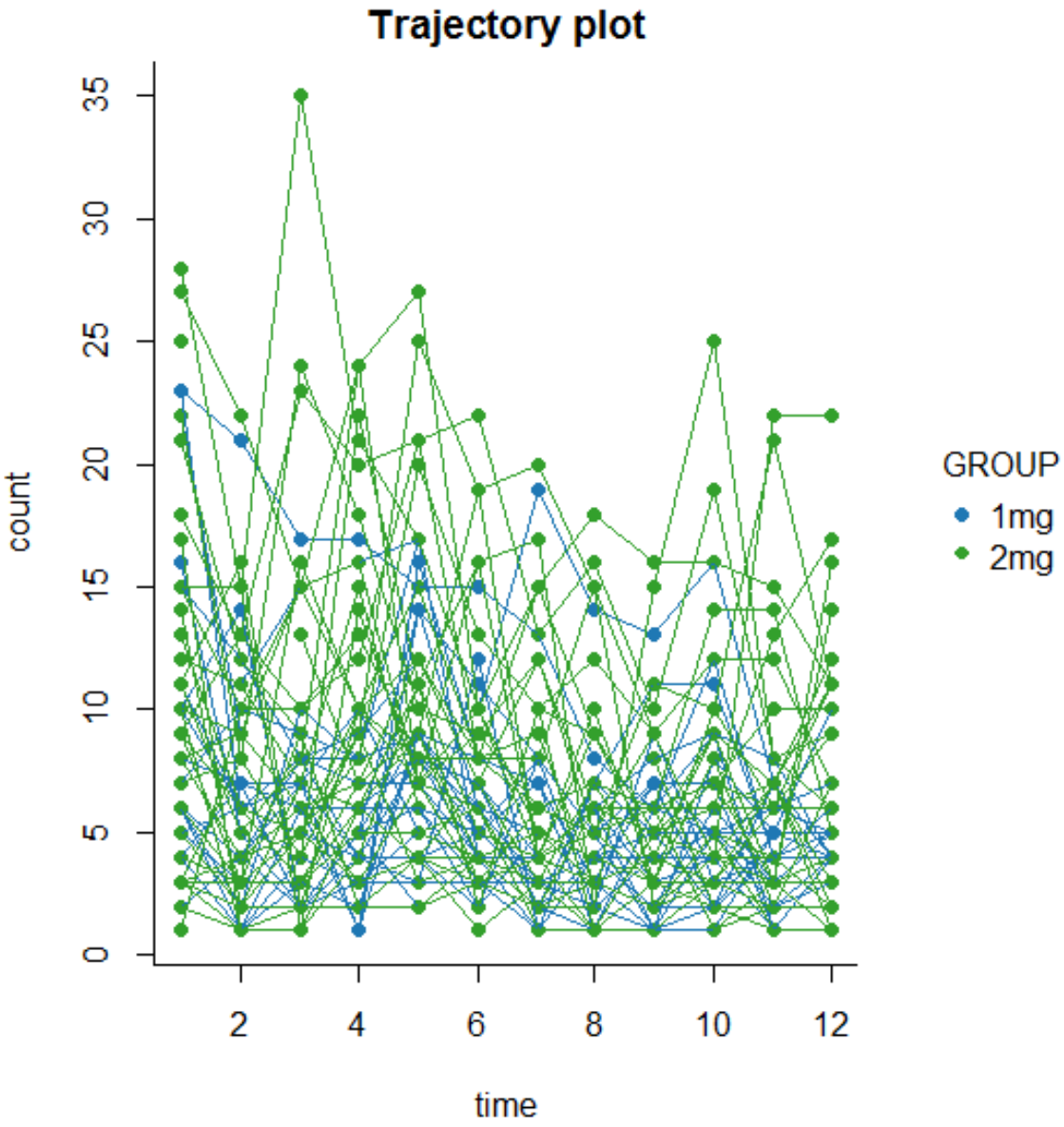


Figure 4.3: Individual trajectories of patients over time

The results are shown in the sections below.

## 4.2 Equally spaced model analysis

Ma et al.(2015) considered the similar model using a generalized linear mixed models based on the Tweedie distributions. This model makes the same assumptions and steps, but analyzes the responses as evenly spaced data where all 65 individuals would have the same number of measurement occasion. The results for the evenly-spaced response variables are listed in Table 1.

Table 4.1: Parameter estimates for the PCA data based on the equally spaced mixed effects model

| | Iteration 1 | | | Iteration 472 | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | P-value | Estimate | SE | P-value |
| Intercept | 1.0182 | 0.0499 | 0.0000 | 0.9564 | 0.0672 | 0.0000 |
| Group 1 (1 mg) | -0.0233 | 0.0309 | 0.4574 | -0.0178 | 0.0415 | 0.6679 |
| Group 2 (2 mg) | 0.1029 | 0.0018 | 0.0000 | 0.1086 | 0.0023 | 0.0000 |
| $\sigma^2$ | 0.3186 | | | 0.0026 | | |
| $\tau^2$ | 0.4274 | | | 0.014 | | |
| $\epsilon^2$ | 1.0000 | | | 1.0000 | | |
| $\rho$ | 0.5000 | | | 0.8547 | | |

In Table 4.1, the estimates of the regression and random effects parameters for the responses are listed. Looking at Table 4.1, we can see that the group effect (the two different drug dosage and administration regime) have a different effect on the response in terms of count. It seems that patients in Group 1 reported fewer times at which they required a dosage of drug per period of time than patients in group 2.

In order to determine whether the association between the response count and the covariate term in the model is statistically significant, we compare the p-value for each term to our significance level to assess the null hypothesis. The null hypothesis is defined as there not being an association between the term and the response. Usually, a significance level (denoted as $\alpha$ or alpha) of 0.05 is adequate. A significance level

of 0.05 indicates a 5 percent risk of concluding that an association exists when there is no actual association.

From Table 4.1 we can see that p-value is less than $\alpha$ , for the response of patients in Group 2. When the p-value is less than or equal to the significance level, we can conclude that there is a statistically significant association between the response variable and the term. Here a small p-value indicates that we can reject the null hypothesis, thus concluding that there is sufficient evidence of a significant association between the Group 2 drug dosage and administration regime and the response counts reported by the patients.

## 4.3 Unequally spaced model analysis

In this section, the proposed unequally spaced model is demonstrated with the analysis of patient-controlled analgesia dataset, under the assumption that the 'no response' values for the individuals from the dataset are neither a missing response, nor can they be qualified as a zeroes, thus creating an unbalanced dataset. To maintain the accuracy of the correlations, the duration of the time intervals is kept small, leading to a large value of T (total number of intervals). For inferences, the regression parameters are estimated by the scoring method and the unevenly spaced correlation parameters of random effects by using moment estimators.

From Table 4.2 we can see that P-value $\leq \alpha$, for the response of patients in Group 2. When the p-value is less than or equal to the significance level, we can conclude that there is a statistically significant association between the response variable and the term. Here the small p-value indicates that we can reject the null hypothesis and thus conclude that there is sufficient evidence of a significant association between the Group 2 drug dosage and administration regime and the

Table 4.2: Parameter estimates for the PCA data based on the unequally spaced mixed effects model

| | Iteration 1 | | | Iteration 949 | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | P-value | Estimate | SE | P-value |
| Intercept | 1.1022 | 0.0499 | 0.0000 | 1.0736 | 0.0578 | 0.0000 |
| Group 1 (1 mg) | -0.0109 | 0.0309 | 0.7251 | -0.0069 | 0.0356 | 0.8474 |
| Group 2 (2 mg) | 0.0967 | 0.0308 | 0.0000 | 0.0989 | 0.0021 | 0.0000 |
| $\sigma^2$ | 0.2419 | | | 0.0013 | | |
| $\tau^2$ | 0.3593 | | | 0.0045 | | |
| $\epsilon^2$ | 1.0000 | | | 1.0000 | | |
| $\rho$ | 0.5000 | | | 0.8900 | | |

response counts reported by the patients, similar to the equally spaced data analysis.

Table 4.2, shows the same trends as Table 4.1 regarding the group effects, however, the difference in the rates of decrease of counts between the two groups, seems to be greater when we consider the dataset as uneven, and disregard the 'no response' or zero values. This means that, group 2 reports a lower number of counts overall, and the rate of decrease as time goes by is greater compared to group 1.

Next, we look at the individual fits for 8 subjects (4 in each group). Patients 12,13,14 and 15 are in group 1, whereas, patients 43, 44, 45, and 46 are in group 2. We can see that the model in some cases overestimate or underestimate, however, overall it seems to be doing an adequate job, (see figure 4.4).

The estimates of dispersion and correlation parameters $\sigma^2$, $\tau^2$, $\epsilon^2$ and $\rho$ for the response counts are 0.0013, 0.0045, 1.0000 and 0.8900, respectively. The estimates of the variance parameters $\sigma^2$, $\tau^2$ and $\epsilon^2$ indicate that there is additional variation in the responses beyond what can be characterized by the random effect. The values of $\sigma^2$ and $\tau^2$ indicate the variation of each subject and of the responses at different measurement times, respectively. The value of $\rho$ indicates the correlation among the

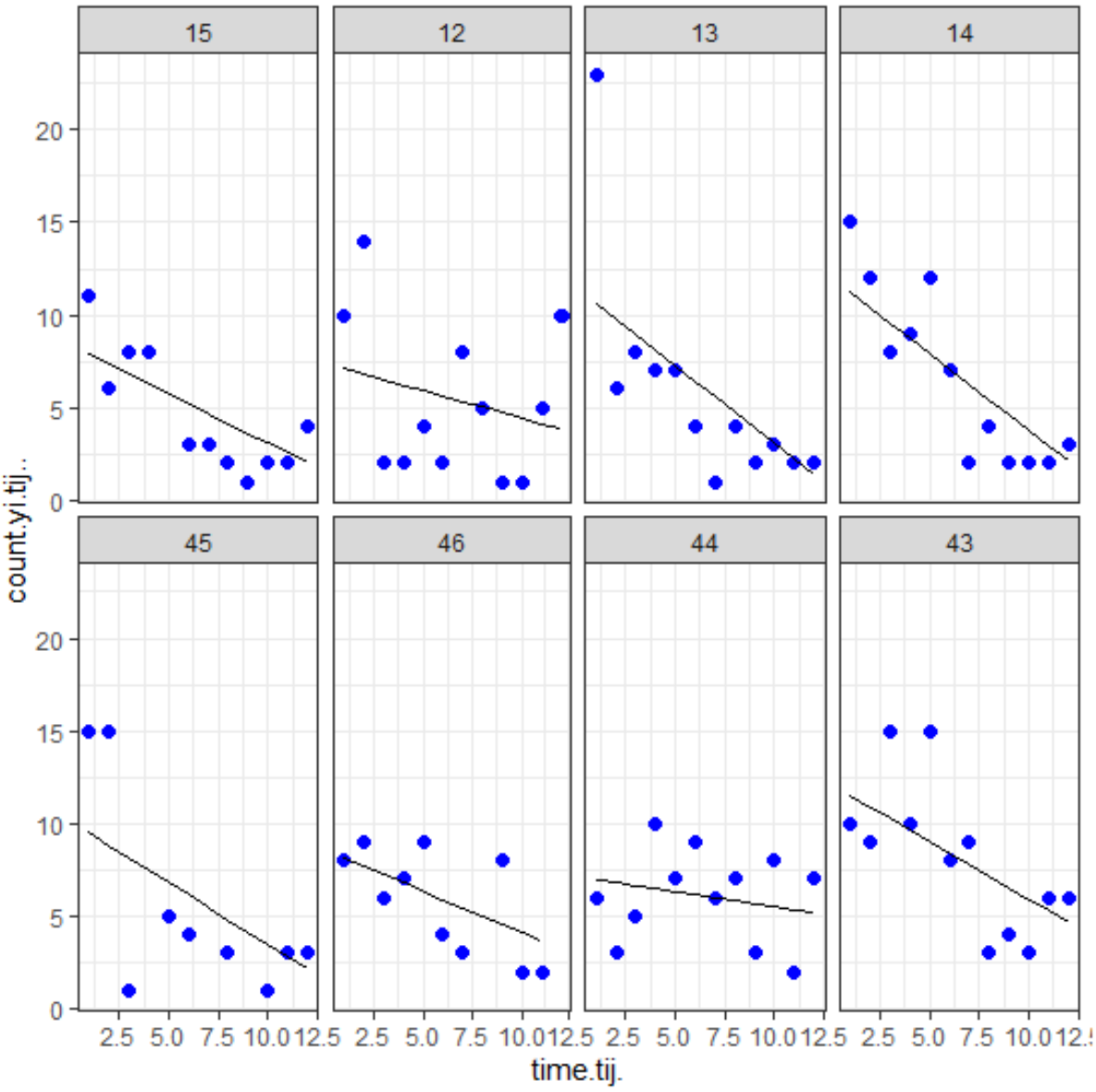


Figure 4.4: Individual fits for 8 subjects (4 in each group)

repeated measurements of each response.

Figure 4.5 is a scatter plot of the predicted subject-specific random effects of each of the 65 patients. The plot shows no obvious outliers between the patients.

From the two approaches, we can see that the regression and random effect pa-

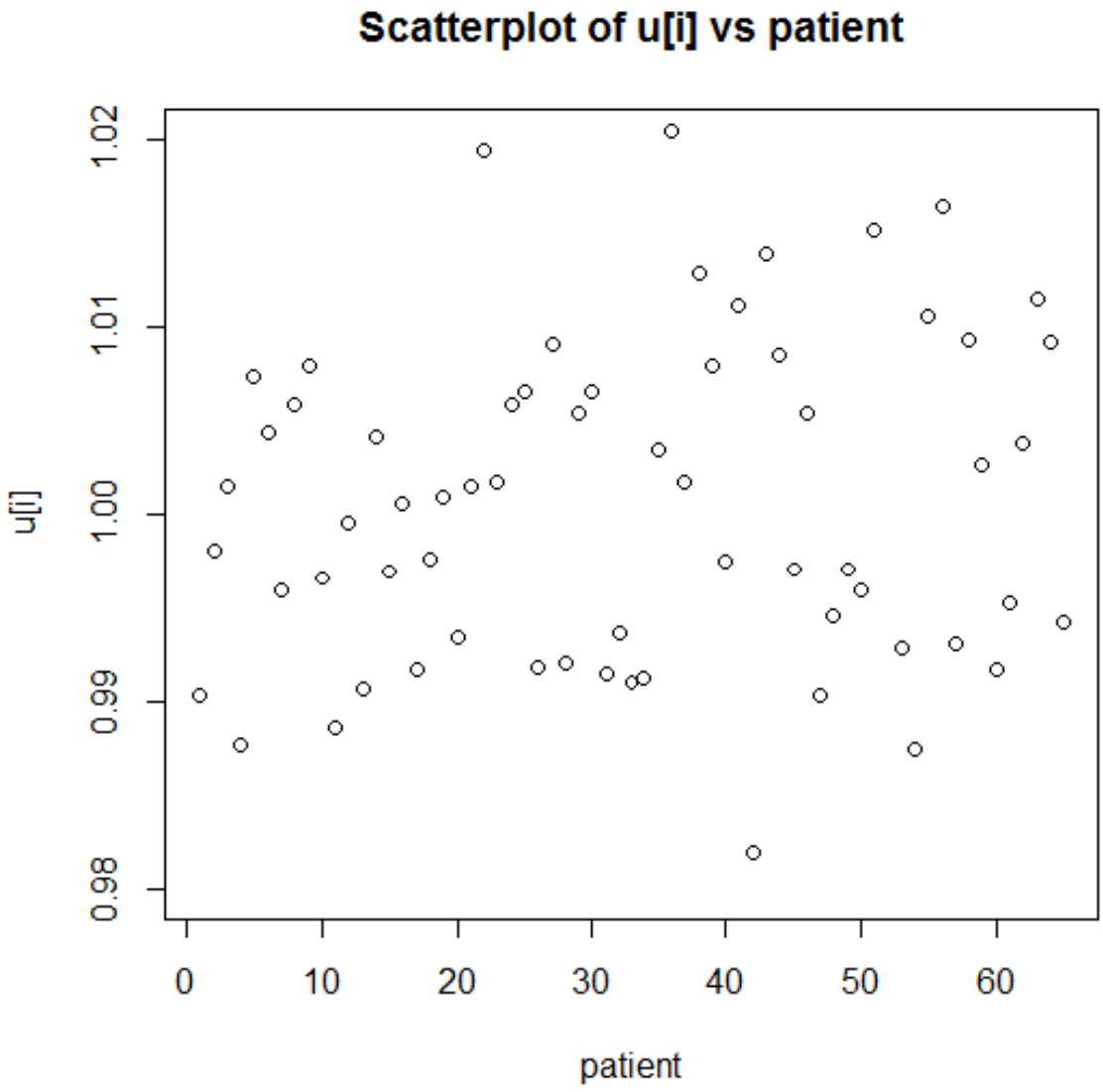


Figure 4.5: Predicted subject-specific random effects of each patients

rameter estimates are similar. The correlation parameter estimate is slightly higher in the unequally spaced method versus the equally spaced model, thus indicating a stronger relationship.

However, the two methods give the same conclusion as there is in fact a differ-

ence in the expected count response in individual patients between the two different pain management regimes.

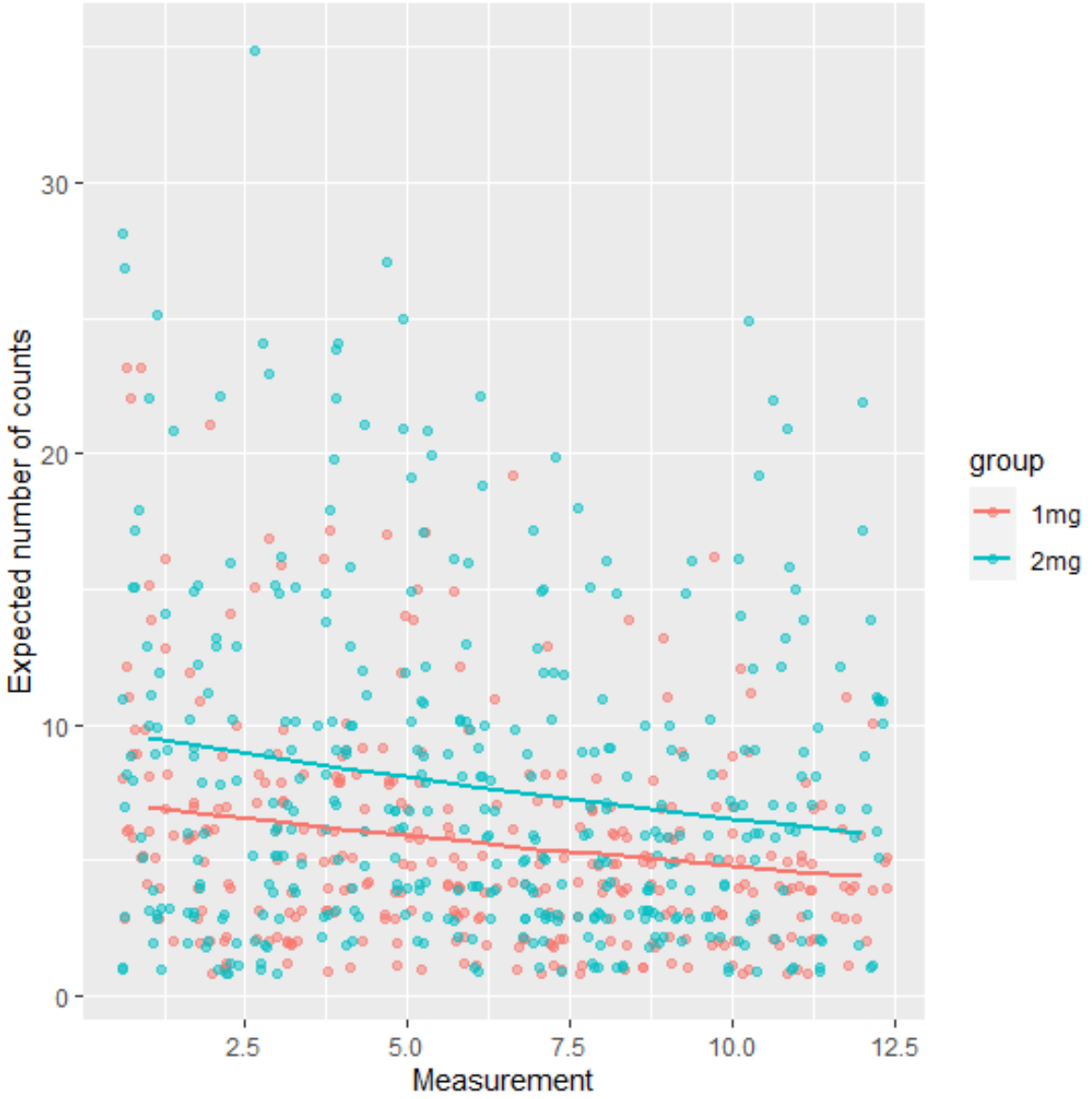


Figure 4.6: Expected number of counts for patients

We can also graph the predicted number of counts. The graph indicates higher response counts are predicted for patients in the group 2 (2mg), especially for the

earlier measurements following surgery.

## 4.4 Simulation

A simulation study was done in this section in order to evaluate the performance of the proposed model, try to assess the bias of the model parameters, as well as their standard errors. To do that, we carried out simulation runs for the unequally spaced model with random effects.

### 4.4.1 Simulation results

Each simulated dataset was generated using the following steps.

• Step 1:

Generate 65 independent observations, $U_1, \cdots, U_{65}$, from a Gamma distribution with mean 1 and variance $\sigma^2$;

• Step 2:

Generate covariates $V_{i1}, \cdots, V_{ij}$ for each $V_i$, following a lognormal distribution with mean $U_i$ and variance $\tau^2\rho(j, j')$ for each $U_i$, where $i = 1, \cdots, 65$. The $j$ measurements used are the same measurement points from our data;

• Step 3:

Generate 719 observations, $Y_{11}, \cdots, Y_{65,j}$ following the Poisson distribution by using

$$Tw_{p(\mu_{ij}V_{ij}, \epsilon^2 {V_{ij}}^{1-p})} \tag{4.1}$$

The conditional mean and variance of the Tweedie distribution is given as

$$E(Y_{ij}|W) = \mu_{ij}V_{ij} \quad \text{and} \quad \text{Var}(Y_{ij}|W) = \epsilon^2\mu_{ij}^q {V_{ij}}^{1-p} \tag{4.2}$$

where $\mu_{ij} = exp(x'_{ij}\beta)$

The results of the simulation are summarized in Table 4.3. Over 500 simulations were conducted and the estimated parameters are the averaged value of those simulations.

Table 4.3: Summary statistics for the Simulation

| Parameter | True Value | Estimated Value | Bias |
|---|---|---|---|
| $\beta_0$ | 1.0736 | 1.0633 | -0.0103 |
| $\beta_1$ | -0.0069 | -0.0034 | 0.0035 |
| $\beta_2$ | 0.0989 | 0.099 | 0.001 |
| $\sigma^2$ | 0.0013 | 0.0053 | 0.004 |
| $\tau^2$ | 0.0045 | 0.0012 | -0.0032 |
| $\epsilon^2$ | 1.0000 | 1.0000 | 0.0000 |
| $\rho$ | 0.8900 | 0.8847 | -0.0053 |

The estimated regression parameters are very close to the true values and the random effects parameters are reasonably estimated through the model.

# Chapter 5

# Discussion

## 5.1 Conclusion

Typically, the main objective of a longitudinal study is to estimate how a response variable is affected by the covariates and how it changes over time. Researchers have always had tremendous interest in developing new approaches and had made determined efforts to deal with different types of longitudinal data. This is particularly important in dealing with clinical studies and medical assessments.

The analysis of longitudinal data can often present several difficulties for researchers. In order to simplify the analysis, most longitudinal studies are designed in a balanced structure, where every subject is assessed on the same exact schedule with the same number of measurements taken. However, despite the efforts of researchers to design a perfect time structured data set, in reality, difficulties may arise with executing a completely balanced study, due to a number of reasons.

In dealing with the variably spaced longitudinal data sets, we proposed a method to estimate parameters for unevenly spaced longitudinal data by working with the flexible class of Tweedie generalized linear mixed models, with both subject-

specific and time-specific random effects. This class of models are able to handle a variety of data types, including continuous, discrete and mixed data.

Our proposed method of estimating parameters is less restrictive since the measurement time points do not have to be evenly spaced. Here we used a generalized AR(1) structure for analyzing the correlation. The proposed method was demonstrated with the analysis of patient-controlled analgesia dataset from Henderson and Shimacura (2003).

The estimates of the regression and random effects parameters for the responses showed that the group effect (the two different drug dosage and administration regime) had an effect on the responses in terms of count. It seemed that patients in Group 1 reported fewer number of times they required a dosage of drug per period of time in comparison with group 2

## 5.2 Extension to the approach

There are a variety of different approaches in dealing with longitudinal data. In further work, one thing to consider would be that observations of each of the patients are serially correlated since the measurements are done over a period of time. In our model, the correlation parameter $\rho$ is assumed under the AR(1) structure. However, it is of interest to investigate other structures in the model.

The correlation $\rho$ between two measurement times for the same object can be

expressed as follows:

$$\Re = \left[\rho_{(j,j')}\right] = \begin{bmatrix} 1 & \rho_{(1,2)} & \rho_{(1,3)} & \cdots & \rho_{(1,J-1)} \\ \rho_{(2,1)} & 1 & \rho_{(2,3)} & \cdots & \rho_{(2,J-1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{(J-1,1)} & \rho_{(J-1,2)} & \rho_{(J-1,3)} & \cdots & 1 \end{bmatrix} \tag{5.1}$$

We might consider the following approach.

Let $J_i$ be a subset of $T_i$ which only includes the count responses greater than zero at the evenly spaced time intervals $T_i$. We define an auxiliary variable $\delta_{ij}$ such that

$$\delta_{ij} \quad = \quad \begin{cases} 1 & \text{indicates that a response for individual i has occurred at the time point } t_{ij} \\ 0 & \text{indicates that a response for individual i has not occurred at the time point } t_{ij} \end{cases}$$

Let's consider a study where measurement responses of a number of individuals had been taken each week, over a period of four weeks, in other words, $T_i = 4$. Each individual could have either had a response recorded at each week or not. For example, one individual could have had four measurements recorded (one for each week) whereas some other individual could have had three measurements recorded (for example, a response for week one, two and four). Distribution of $N = N_1 + N_2 + N_3$ individuals based on unevenly spaced response type when $T = 4$: Number of $(Responses, Individuals) = (4, N_1), (3, N_2)$ and $(2, N_3)$ . Where $N = N_1 + N_2 + N_3 =$ Total number of individuals. Considering all the possible ways individuals could have provided measurements, we get the following sets:

1) Where the individual reported for all four measurements:

$\delta_{i1} = 1, \delta_{i2} = 1, \delta_{i3} = 1, \delta_{i4} = 1$ for $i = 1 \ldots N_{11} = N_1$

2) Where the individual reported three out of four measurements:

$\delta_{i1} = 1, \delta_{i2} = 1, \delta_{i3} = 1, \delta_{i4} = 0$ for $i = 1 \ldots N_{21}$

$\delta_{i1} = 1, \delta_{i2} = 1, \delta_{i3} = 0, \delta_{i4} = 1$ for $i = 1 \ldots N_{22}$

$\delta_{i1} = 1, \delta_{i2} = 0, \delta_{i3} = 1, \delta_{i4} = 1$ for $i = 1 \ldots N_{23}$

$\delta_{i1} = 0, \delta_{i2} = 1, \delta_{i3} = 1, \delta_{i4} = 1$ for $i = 1 \ldots N_{24}$

$N_{21} + N_{22} + N_{23} + N_{24} = N_2$

3) Where the individual reported two out of four measurements:

$\delta_{i1} = 1, \delta_{i2} = 1, \delta_{i3} = 0, \delta_{i4} = 0$ for $i = 1 \ldots N_{31}$

$\delta_{i1} = 1, \delta_{i2} = 0, \delta_{i3} = 1, \delta_{i4} = 1$ for $i = 1 \ldots N_{32}$

$\delta_{i1} = 1, \delta_{i2} = 0, \delta_{i3} = 0, \delta_{i4} = 1$ for $i = 1 \ldots N_{33}$

$\delta_{i1} = 0, \delta_{i2} = 1, \delta_{i3} = 1, \delta_{i4} = 0$ for $i = 1 \ldots N_{34}$

$\delta_{i1} = 0, \delta_{i2} = 1, \delta_{i3} = 0, \delta_{i4} = 1$ for $i = 1 \ldots N_{35}$

$\delta_{i1} = 0, \delta_{i2} = 0, \delta_{i3} = 1, \delta_{i4} = 1$ for $i = 1 \ldots N_{36}$

$N_{31} + N_{32} + N_{33} + N_{34} + N_{35} + N_{36} = N_3$

Notice that in case.1, we may compute the means and variances for $N_1$ patient at all 4 time points. However, under case.2, the means and variances cannot be computed at all 4 time points. The pair-wise correlation index parameters are shown as below:

1) Where the individual reported for all four measurements:

$$\delta_{i1,1,2} = 1(\rho_{12}), \delta_{i1,1,3} = 1(\rho_{12}, \rho_{23}), \delta_{i1,1,4} = 1(\rho_{12}, \rho_{23}, \rho_{34}), \delta_{i1,2,3} = 1(\rho_{23}), \delta_{i1,2,4} = 1(\rho_{23}, \rho_{34}), \delta_{i1,3,4} = 1(\rho_{34})$$

2) Where the individual reported three out of four measurements:

$$\delta_{i2,1,2} = 1(\rho_{12}), \delta_{i2,1,3} = 1(\rho_{12}, \rho_{23}), \delta_{i2,1,4} = 0, \delta_{i2,2,3} = 1(\rho_{23}), \delta_{i2,2,4} = 0, \delta_{i2,3,4} = 0,$$

$$\delta_{i2,1,2} = 1(\rho_{12}), \delta_{i2,1,3} = 0, \delta_{i2,1,4} = 1(\rho_{12}, \rho_{24}), \delta_{i2,2,3} = 0, \delta_{i2,2,4} = 1(\rho_{24}), \delta_{i2,3,4} = 0$$

$$\delta_{i2,1,2} = 0 \delta_{i2,1,3} = 1(\rho_{13}), \delta_{i2,1,4} = 1(\rho_{13}, \rho_{34}), \delta_{i2,2,3} = 0, \delta_{i2,2,4} = 0, \delta_{i2,3,4} = 1(\rho_{34}),$$

$$\delta_{i2,1,2} = 0, \delta_{i2,1,3} = 0, \delta_{i2,1,4} = 0, \delta_{i2,2,3} = 1(\rho_{23}), \delta_{i2,2,4} = 1(\rho_{23}, \rho_{34}), \delta_{i2,3,4} = 1(\rho_{34}),$$

And so on. There are 6 pair-wise correlation index parameters to estimate: $(\rho_{1,2}), (\rho_{1,3}), (\rho_{1,4}), (\rho_{2,3}), (\rho_{2,4})$and $(\rho_{3,4})$. For example, to write an observed function for $(\rho_{1,2})$, we need to look at all the covariance terms that involve $(\rho_{1,2})$. Let $f_{s,1,2}$ denote the sum of all pair-wise product terms whose covariance would contain $(\rho_{1,2})$ under all cases mentioned above. With $\text{Var}(y_i t_j)^{1/2} = [\epsilon^2 \mu_i t_j + \mu_i {t_j}^2 \sigma^2]^{1/2}$, and the unconditional expectation of the response $y_i t_j$ can be expressed as: $E(y_i t_j) = \mu_i t_j$. By using the covariance formula $\mu_i t_j \mu_i t_{j'} (\sigma^2 + \tau^2 \rho_{(t_j, t_{j'})})$, we find the expectation of $f_{(s,1,2)}$, which can be denoted by $f_s(\rho_{12}) = E[f_{(s,1,2)}]$. The moment estimating equation for $\rho_{12}$ would be $f_{(s,1,2)} - f_s(\rho_{12}) = 0$.

$$\begin{aligned}
& f_{s,1,2} \\
= & \frac{\sum_{i_1=1}^{N_1} \sum_{j=0}^{2} \left\{ (y_{i1t_1} - \mu_{i1t_1})(y_{i1t_{2+j}} - \mu_{i1t_{2+j}}) \right\}}{\left[ \left\{ \sum_{i_1=1}^{N_1} \sum_{j=0}^{2} (\epsilon^2 \mu_{i1t_1} + \mu_{i1t_1}^2 \sigma^2)(\epsilon^2 \mu_{i1t_{2+j}} + \mu_{i1t_{2+j}}^2 \sigma^2) \right\} \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{21}=1} \sum_{j=0}^{1} \left\{ (y_{i2t_1} - \mu_{i2t_1})(y_{i2t_{2+j}} - \mu_{i2t_{2+j}}) \right\}}{\left[ \left\{ \sum_{i_2=1}^{N_{21}=1} \sum_{j=0}^{1} (\epsilon^2 \mu_{i2t_1} + \mu_{i2t_1}^2 \sigma^2)(\epsilon^2 \mu_{i2t_{2+j}} + \mu_{i2t_{2+j}}^2 \sigma^2) \right\} \right]^{1/2}} \\
+ & \frac{\sum_{i_2=N_{21}+1}^{N_{21}+N_{22}} \left\{ (y_{i2t_1} - \mu_{i2t_1})(y_{i2t_2} - \mu_{i2t_2}) \right\}}{\left[ \left\{ \sum_{i_2=N_{21}+1}^{N_{21}+N_{22}} (\epsilon^2 \mu_{i2t_1} + \mu_{i2t_1}^2 \sigma^2)(\epsilon^2 \mu_{i2t_2} + \mu_{i2t_2}^2 \sigma^2) \right\} \right]^{1/2}} \\
+ & \frac{\sum_{i_2=N_{21}+1}^{N_{21}+N_{22}} \left\{ (y_{i2t_1} - \mu_{i2t_1})(y_{i2t_4} - \mu_{i2t_4}) \right\}}{\left[ \left\{ \sum_{i_2=N_{21}+1}^{N_{21}+N_{22}} (\epsilon^2 \mu_{i2t_1} + \mu_{i2t_1}^2 \sigma^2)(\epsilon^2 \mu_{i2t_4} + \mu_{i2t_4}^2 \sigma^2) \right\} \right]^{1/2}} \\
+ & \frac{\sum_{i_3=1}^{N_{31}} \left\{ (y_{i3t_1} - \mu_{i3t_1})(y_{i3t_2} - \mu_{i3t_2}) \right\}}{\left[ \left\{ \sum_{i_3=1}^{N_{31}} (\epsilon^2 \mu_{i3t_1} + \mu_{i3t_1}^2 \sigma^2)(\epsilon^2 \mu_{i3t_2} + \mu_{i3t_2}^2 \sigma^2) \right\} \right]^{1/2}}
\end{aligned}$$

$$
\begin{aligned}
& f_s(\rho_{12}) \\
= & \frac{\sum_{i_1=1}^{N_1}\sum_{j=1}^{3}\left\{\mu_{i1t_j}\mu_{i1t_{(j+1)}}\sigma^2+(\mu_{i1t_j}\mu_{i1t_{(j+1)}}\tau^2\rho_{(1,2)}[1+(\rho_{(2,3)})+(\rho_{(2,3)})(\rho_{(3,4)})])\right\}}{\left[\left\{\sum_{i_1=1}^{N_1}\sum_{j=1}^{3}(\epsilon^2\mu_{i1t_j}+\mu_{i1t_j}^2\sigma^2)(\epsilon^2\mu_{i1t_{j+1}}+\mu_{i1t_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{21}}\sum_{j=1}^{2}\left\{\mu_{i2t_j}\mu_{i2t_{(j+1)}}\sigma^2+(\mu_{i2t_j}\mu_{i2t_{(j+1)}}\tau^2\rho_{(1,2)}[1+(\rho_{(2,3)})])\right\}}{\left[\left\{\sum_{i_2=1}^{N_{21}}\sum_{j=1}^{2}(\epsilon^2\mu_{i2t_j}+\mu_{i2t_j}^2\sigma^2)(\epsilon^2\mu_{i2t_{j+1}}+\mu_{i2t_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=N_{21}+1}^{N_{21}+N_{22}}\sum_{j=3}^{3}\left\{\mu_{i2t_2}\mu_{i2t_{(j+1)}}\sigma^2+(\mu_{i2t_2}\mu_{i2t_{(j+1)}}\tau^2\rho_{(1,2)}[1+(\rho_{(2,4)})])\right\}}{\left[\left\{\sum_{i_2=N_{21}+1}^{N_{21}+N_{22}}\sum_{j=3}^{3}(\epsilon^2\mu_{i2t_2}+\mu_{i2t_2}^2\sigma^2)(\epsilon^2\mu_{i2t_{j+1}}+\mu_{i2t_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_3=1}^{N_3}\sum_{j=1}^{1}\left\{\mu_{i2t_j}\mu_{i2t_{(j+1)}}\sigma^2+\mu_{i2t_j}\mu_{i2t_{(j+1)}}\tau^2\rho_{(1,2)}\right\}}{\left[\left\{\sum_{i_3=1}^{N_3}\sum_{j=1}^{1}(\epsilon^2\mu_{i2t_j}+\mu_{i2t_j}^2\sigma^2)(\epsilon^2\mu_{i2t_{j+1}}+\mu_{i2t_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}}
\end{aligned}
\tag{5.3}
$$

Note that the moment equation involves some other pair-wise unevenly spaced correlation index parameters, thus, the moment equations for correlation index parameters have to be solved iteratively.

An illustration of the covariance matrix for patient=26 with $T_i = 12$ and $J_i = 5$, where for patient number 26, only 5 responses have occurred at 5 time points indicated by

$\delta_{i1} = 1, \delta_{i2} = 1, \delta_{i3} = 0, \delta_{i4} = 0, \delta_{i5} = 0, \delta_{i6} = 0, \delta_{i7} = 1, \delta_{i8} = 0, \delta_{i9} = 1, \delta_{i10} = 0, \delta_{i11} = 1, \delta_{i12} = 0$

Since $J_i = 5$, we can construct an $5 \times 5$ covariance matrix as $Var(Y_{i1}) \Leftrightarrow$

$Var(Y_{i1}) = \epsilon^2\mu_{i1}^p + \mu_{i1}^2(\sigma^2+\tau^2), Var(Y_{i2}) \Leftrightarrow Var(Y_{i2}) = \epsilon^2\mu_{i2}^p + \mu_{i2}^2(\sigma^2+\tau^2), Var(Y_{i3}) \Leftrightarrow Var(Y_{i7}) = \epsilon^2\mu_{i7}^p + \mu_{i7}^2(\sigma^2+\tau^2), Var(Y_{i4}) \Leftrightarrow Var(Y_{i9}) = \epsilon^2\mu_{i9}^p + \mu_{i9}^2(\sigma^2+\tau^2), Var(Y_{i5}) \Leftrightarrow Var(Y_{i11}) = \epsilon^2\mu_{i11}^p + \mu_{i11}^2(\sigma^2+\tau^2);$

$$\text{Cov}(Y_{i1}, Y_{i2}) = \mu_{i1}\mu_{i2}(\sigma^2 + \tau^2\rho_{(1,2)}) \Leftrightarrow \text{Cov}(Y_{i1}, Y_{i2}) = \mu_{i1}\mu_{i2}(\sigma^2 + \tau^2\rho_{(1,2)}),$$
$$\text{Cov}(Y_{i1}, Y_{i3}) = \mu_{i1}\mu_{i3}(\sigma^2 + \tau^2\rho_{(1,3)}) \Leftrightarrow \text{Cov}(Y_{i1}, Y_{i7}) = \mu_{i1}\mu_{i7}(\sigma^2 + \tau^2\rho_{(1,7)}),$$
$$\text{Cov}(Y_{i1}, Y_{i4}) = \mu_{i1}\mu_{i4}(\sigma^2 + \tau^2\rho_{(1,4)}) \Leftrightarrow \text{Cov}(Y_{i1}, Y_{i9}) = \mu_{i1}\mu_{i9}(\sigma^2 + \tau^2\rho_{(1,9)}),$$
$$\text{Cov}(Y_{i1}, Y_{i5}) = \mu_{i1}\mu_{i5}(\sigma^2 + \tau^2\rho_{(1,5)}) \Leftrightarrow \text{Cov}(Y_{i1}, Y_{i11}) = \mu_{i1}\mu_{i11}(\sigma^2 + \tau^2\rho_{(1,11)}),$$
$$\text{Cov}(Y_{i2}, Y_{i3}) = \mu_{i2}\mu_{i3}(\sigma^2 + \tau^2\rho_{(2,3)}) \Leftrightarrow \text{Cov}(Y_{i2}, Y_{i7}) = \mu_{i2}\mu_{i7}(\sigma^2 + \tau^2\rho_{(2,7)}),$$
$$\text{Cov}(Y_{i2}, Y_{i4}) = \mu_{i2}\mu_{i4}(\sigma^2 + \tau^2\rho_{(2,4)}) \Leftrightarrow \text{Cov}(Y_{i2}, Y_{i9}) = \mu_{i2}\mu_{i9}(\sigma^2 + \tau^2\rho_{(2,9)}),$$
$$\text{Cov}(Y_{i2}, Y_{i5}) = \mu_{i2}\mu_{i5}(\sigma^2 + \tau^2\rho_{(2,5)}) \Leftrightarrow \text{Cov}(Y_{i2}, Y_{i11}) = \mu_{i2}\mu_{i11}(\sigma^2 + \tau^2\rho_{(2,11)}),$$
$$\text{Cov}(Y_{i3}, Y_{i4}) = \mu_{i3}\mu_{i4}(\sigma^2 + \tau^2\rho_{(3,4)}) \Leftrightarrow \text{Cov}(Y_{i7}, Y_{i9}) = \mu_{i7}\mu_{i9}(\sigma^2 + \tau^2\rho_{(7,9)}),$$
$$\text{Cov}(Y_{i3}, Y_{i5}) = \mu_{i3}\mu_{i5}(\sigma^2 + \tau^2\rho_{(3,5)}) \Leftrightarrow \text{Cov}(Y_{i7}, Y_{i11}) = \mu_{i7}\mu_{i11}(\sigma^2 + \tau^2\rho_{(7,11)},$$
$$\text{Cov}(Y_{i4}, Y_{i5}) = \mu_{i4}\mu_{i5}(\sigma^2 + \tau^2\rho_{(4,5)}) \Leftrightarrow \text{Cov}(Y_{i9}, Y_{i11}) = \mu_{i9}\mu_{i11}(\sigma^2 + \tau^2\rho_{(9,11)})$$

For example, the computation of the covariance between the last two observed responses for individual number 26, will require the estimate of $\rho_{(t_{i9},t_{i11})})$, which indicates that there is a time interval of 2 units between the fourth and fifth observed responses.

Distribution of $N = N_1 + N_2 + N_3 + N_4 + N_5 + N_6 + N_7$ individuals based on unevenly spaced response type when $T = 12$: Number of $(Responses, Individuals) = (12, N_1), (11, N_2), (10, N_3), (9, N_4), (8, N_5), (7, N_6), (5, N_7)$ for our PCA data.

Where $N = N_1 + N_2 + N_3 + N_4 + N_5 + N_6 + N_7 = 65$ patients in the PCA data. We keep in mind that these cases are specific to our data set. In reality there are more different ways that the responses can occur for each case. For example, in case.2, the 11 responses can occur in 12 different ways, and in case.3, the 10 responses can occur in 66 different ways. Notice that in case.1, we may compute the means and variances for $N_1$ patient at all 12 time points. However, under case.2, the means and variances cannot be computed at all 12 time points. For example, for case.3, the mean and variances can be computed at time points 1,2,7,9 and 11 only. Same

goes for the computation of pair-wise correlations, where it is necessary that the two responses selected have indeed occurred.

As an example, let's look at case.7

Case 7. Where $J_i = 5$ for $i = 1, \ldots, N_7$; $N_7 = 1$

$$\delta_{i7,1,2} = 1(\rho_{12}), \delta_{i7,1,3} = 0, \delta_{i7,1,4} = 0, \delta_{i7,1,5} = 0, \delta_{i7,1,6} = 0, \delta_{i7,1,7} = 1(\rho_{12}, \rho_{27}), \delta_{i7,1,8} = 0, \delta_{i7,1,9} = 1(\rho_{12}, \rho_{27}, \rho_{79}), \delta_{i7,1,10} = 0, \delta_{i7,1,11} = 1(\rho_{12}, \rho_{27}, \rho_{79}, \rho_{9,11}), \delta_{i7,1,12} = 0; \delta_{i7,2,3} = 0, \delta_{i7,2,4} = 0, \ldots, \delta_{i7,11,12} = 0$$

Moment equation for unevenly spaced pair-wise correlation index parameters for the PCA data: There are 66 pair-wise correlation index parameters to estimate: $(\rho_{1,2}), (\rho_{1,3}), (\rho_{1,4}), \ldots, (\rho_{1,12}), (\rho_{2,3}), (\rho_{2,4}), (\rho_{2,5}), \ldots, (\rho_{2,12}), (\rho_{3,4}), (\rho_{3,5}), \ldots, (\rho_{10,12}), (\rho_{11,12})$

For example, to write an observed function for $(\rho_{1,3})$, we need to look at all the covariance terms that involve $(\rho_{1,3})$. Let $f_{s,1,3}$ denote the sum of all pair-wise product terms whose covariance would contain $(\rho_{1,3})$ under all 7 cases mentioned above for the PCA data.

Where $N_{21} = 3$, patients number (48,54,62); $N_{31} = 1$, patient number (16); $N_{41} = 1$, patient number 10. Unequally spaced measurements for $N_{21}$ are as following: $1, 3, 4, \ldots, 11$. This means that patients number 48, 54 and 62 have not reported a response at $j = 2$ or $j = 12$. Similarly, unequally spaced measurements for $N_{31}$ and $N_{41}$ are as following respectively : $1, 3, 4, 5, 6, 8 \ldots, 11$, and $1, 3, 4, \ldots, 12$.

$$
\begin{aligned}
& f_{s,1,3} \\
= & \frac{\sum_{i_2=1}^{N_{21}=3}\sum_{j=2}^{11}\left\{(y_{it_1}-\mu_{it_1})(y_{it_{j+1}}-\mu_{it_{j+1}})\right\}}{\left[\left\{\sum_{i_2=1}^{3}\sum_{j=2}^{11}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{31}=1}\sum_{j=2}^{10}\left\{(y_{it_1}-\mu_{it_1})(y_{it_{j+1}}-\mu_{it_{j+1}})\right\}}{\left[\left\{\sum_{i_2=1}^{1}\sum_{j=2}^{10}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{41}=1}\sum_{j=2}^{5}\left\{(y_{it_1}-\mu_{it_1})(y_{it_{j+1}}-\mu_{it_{j+1}})\right\}}{\left[\left\{\sum_{i_2=1}^{1}\sum_{j=2}^{5}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{41}=1}\sum_{j=7}^{10}\left\{(y_{it_1}-\mu_{it_1})(y_{it_{j+1}}-\mu_{it_{j+1}})\right\}}{\left[\left\{\sum_{i_2=1}^{1}\sum_{j=7}^{10}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}}
\end{aligned}
\tag{5.4}
$$

$$
\begin{aligned}
& f_s(\rho_{13}) \\
= & \frac{\sum_{i_2=1}^{N_{21}=3}\left\{\mu_i t_1\mu_i t_3(\sigma^2+\tau^2\rho_{(1,3)}[1+\rho_{(3,4)}+\rho_{(3,4)}\rho_{(4,5)}+\ldots+\rho_{(3,4)}\rho_{(4,5)}\ldots\rho_{(11,12)}])\right\}}{\left[\left\{\sum_{i_2=1}^{3}\sum_{j=2}^{11}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{31}=1}\left\{\mu_i t_1\mu_i t_3(\sigma^2+\tau^2\rho_{(1,3)}[1+\rho_{(3,4)}+\rho_{(3,4)}\rho_{(4,5)}+\ldots+\rho_{(3,4)}\rho_{(4,5)}\ldots\rho_{(10,11)}])\right\}}{\left[\left\{\sum_{i_2=1}^{3}\sum_{j=2}^{10}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{41}=1}\left\{\mu_i t_1\mu_i t_3(\sigma^2+\tau^2\rho_{(1,3)}[1+\rho_{(3,4)}+\rho_{(3,4)}\rho_{(4,5)}+\rho_{(3,4)}\rho_{(4,5)}\rho_{(5,6)}])\right\}}{\left[\left\{\sum_{i_2=1}^{1}\sum_{j=2}^{5}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}} \\
+ & \frac{\sum_{i_2=1}^{N_{41}=1}\left\{\mu_i t_1\mu_i t_3(\sigma^2+\tau^2\rho_{(1,3)}[\rho_{(3,4)}\ldots\rho_{(5,6)}\rho_{(6,8)}+\ldots+\rho_{(3,4)}\ldots\rho_{(5,6)}\rho_{(6,8)}\ldots\rho_{(11,12)}])\right\}}{\left[\left\{\sum_{i_2=1}^{1}\sum_{j=7}^{11}(\epsilon^2\mu_{it_1}+\mu_{it_1}^2\sigma^2)(\epsilon^2\mu_{it_{j+1}}+\mu_{it_{j+1}}^2\sigma^2)\right\}\ \right]^{1/2}}
\end{aligned}
\tag{5.5}
$$

With $\mathrm{Var}(y_i t_j)^{1/2} = [\epsilon^2\mu_i t_j + \mu_i {t_j}^2\sigma^2]^{1/2}$, and the unconditional expectation of

the response $y_i t_j$ can be expressed as: $E(y_i t_j) = \mu_i t_j$. By using the covariance formula $\mu_i t_j \mu_i t_{j'}(\sigma^2 + \tau^2 \rho_{(t_j, t_{j'})})$, we find the expectation of $f_{(s,1,3)}$, which can be denoted by $f_s(\rho_{13}) = E[f_{(s,1,3)}]$. The moment estimating equation for $\rho_{13}$ would be $f_{(s,1,3)} - f_s(\rho_{13}) = 0$.

## 5.3 Further study

The work on longitudinal studies is ongoing. There are so many aspects of these types of data frame that researchers are interested in developing and expanding upon. For example, the standard errors for random effects parameters used in this model were not estimated. It might be of interest to come up with tests to show the significance of these parameters.

In order to extensively analyze an unequally spaced longitudinal data set, we need to come up with a different correlation structure. Potentially a new method to estimate a generalized AR(1) correlation structure could solve the problem.

We could also look into extending the work done by Li (2018), and incorporate the multivariate model with unequally spaced time points. It would be interesting to observe the difference between when the covariates for each subject have the same measurement occasions, as opposed to different measurement occasions between covariates, and whether different approaches will be needed in examining each.

Also, the problem of missing data could come up at some point. The issue of missing data can perhaps be addressed by using a best linear unbiased predictor (BLUP) to predict and impute any missing values.

# Bibliography


[1] Henderson R, Shimakura S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika.* 90(2), pp355-366.

[2] Sutradhar B., Jowaheer V. (2003). On familial longitudinal Poisson mixed models with gamma random effects *Journal of Multivariate Analysis* 87, pp398-412.

[3] Leppik IE, Dreifuss FE., Porter R.,Bowman T., Santilli N., Jacobs M., Crosby C., Cloyd J., Stackman J., Graves N. (1987). A controlled study of progabide in partial seizures: methodology and results. Neurology. *National Library of Medicine* 37(6), pp963-8.

[4] Oyet A., Sutradhar B.(2019). Analyzing Unevenly Spaced Longitudinal Count Data *Indian Statistical Institute*

[5] Thall P., Vail S. (1990). Some Covariance Models for Longitudinal Count Data with Overdispersion *International Biometric Society* 46, pp657-671.

[6] Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. Roy. Statist. Soc. Ser. B* 49, pp127-162.

[7] Ma, R. (1999). *An orthodox blup approach to generalized linear mixed models*, University of British Columbia, PhD thesis.

[8] Garret M. Fitzmaurice, Nan M. Laird and James H. Ware. (2011). *Applied Longitudinal Analysis ($2^{nd}$ ed.)*. John Wiley & Sons: New Jersey.

[9] Diggle, P.J., Heagerty, P.J., Liang, K.Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data ($2^{nd}$ ed.)*. Oxford Science Publications: Oxford.

[10] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC Taylor Francis Group: Boca Raton.

[11] Li, J. (2018). *Joint Tweedie Mixed Models for Longitudinal Data of Mixed* , University of New Brunswick, Master's thesis.

[12] Nedler, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135, pp370-384.

[13] Liang K.Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, pp13-22.

[14] Liang K.Y. and Zeger, S.L. (1994). Analysis of Longitudinal Data. *Oxford University Press Inc. New York* , pp85.

[15] Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38, pp963-974.

[16] Guo, X. (2011). *Longitudinal Data Analysis in Social Science Data*, University of Alberta, Master thesis.

[17] Lall, S. (2014). *A Study of Tweedie Family of Distribution for Rainfall Modelling*, Indian Agricultural Research Institute, Master thesis.

[18] Jørgensen, B. (1997). *The Theory of Dispersion Models.* Chapman and Hall: London.

[19] Jørgensen, B. and Souza, M.C.P. (1994). *Fitting Tweedie's compound Poisson model to insurance claims data* Jcand. Actuarial J. 1, B49, pp 69-93.

[20] Jørgensen, B. (1987). *Exponential dispersion models (with discussion)* J. Roy. Statist. Soc. Ser., B49, pp 127-162.

[21] Ma, R. and Jørgensen, B. (2007). *Nested generalized linear models: an orthodox best linear unbiased predictor approach,* Journal of Royal Statistical Society, B69, pp 625-641.

[22] Ma, R., Hasan, M.T., and Sneddon, G. (2009). Modelling heterogeneity in clustered count data with extra zeros using compound Poisson random effect. *Statistics in medicine*, 28, pp 2356-2369.

[23] Ha, I.D. and Lee, Y. (2005). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models. *Biometrika*, 92, pp717-723.

[24] Ha, I.D. and Lee, Y. (2005).Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B58, pp619-678.

[25] Nunez-Anton, V., Woodworth G.(1994). Analysis of Longitudinal Data with Unequally Spaced Observations and Time- Dependent Correlated Errors *Biometrics* . 50(2), pp445-456.

[26] Glasbey C., Nunez-Anton, V., Woodworth G.(1995). Unequally Spaced Longitudinal Data *Biometrics* . 51(1), pp375-377.

[27] Huang, W and Fitzmaurice, G.M., (2005). Analysis of longitudinal data unbalanced over time, *J. R. Stat. Soc.* B 67, pp. 135–155

[28] Molenberghs, G. and Verbeke, G. (2005). Models for Discrete Longitudinal Data, *Springer, NewYork*.

[29] Cnaan, A., Laird, N.M., and Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data, *Stat. Med.* 16, pp. 2349–2380.

[30] Heagerty, P.(2002). Marginalized transition models and likelihood inference for longitudinal categorical data, *Biometrics* 58, pp. 342–351.

[31] R Core Team (2020). R: A Language and Environment for Statistical Computing, *R Foundation for Statistical Computing*. Vienna, Austria. http://www.R-project.org/.

# Vita

**Candidate's full name**:

Nakisa Tamjidi

**Universities Attended:**

Master of Science, June 2022, University of New Brunswick,

Fredericton, NB, Canada

Bachelor of Science, May 2014, University of New Brunswick,

Fredericton, NB, Canada

**Publications**: None

**Conference Presentations**: None