Longitudinal Data Analysis:

Understanding Visit Irregularities

Kasra Vakiloroayaei

Department of Mathematics and Statistics

McGill University

Montréal (Québec) Canada

April 7, 2024

A project presented for the degree of Master of Science

Abstract

There are many existing modeling approaches for longitudinal data that have been established based on a balanced data structure satisfying various assumptions. Some of these approaches are described in order to show how they perform when the ideal of a perfectly repeated structure is compromised by irregular data. A better understanding of the extent of this problem can be helpful before carrying out any longitudinal data analysis. A study on the pharmacokinetics of remifentanil is used as an illustration.

Résumé

Plusieurs approches ont été proposées pour la modélisation de données longitudinales issues d'un plan équilibré et satisfaisant divers postulats. Certaines d'entre elles sont présentées afin d'étudier leur comportement lorsque l'idéal d'une structure parfaitement répétée est compromis en raison d'irrégularités dans les données. Une meilleure compréhension du problème peut s'avérer utile avant d'effectuer une analyse de données longitudinales. Une étude sur la pharmacocinétique du rémifentanyl illustre le propos.

Acknowledgements

First, I would like to give special thanks to my supervisor, Dr. Christian Genest, for his continuous belief in me throughout my graduate studies, for being my mentor, and for the endless motivation he gave me. He helped me adapt to the changes in the world as we went through the COVID-19 Pandemic and was very understanding. Furthermore, he provided critical support to me during some of the toughest times in my life. He has forever left a positive impact on my career in statistics, and I will be eternally grateful.

Second, I would also like to thank two amazing professors, Dr. Jerry Brunner and Dr. Luai Al-Labadi, whose support extended past my studies. I am thankful to the former for allowing me to ask questions without being afraid and for his open-door policy. The latter provided the opportunity to be part of writing my first publications; I thank him for always investing his time in creating a foundation for me to learn.

To my family, who has given me love and support from the very beginning, I thank you from the bottom of my heart. I am grateful for the most patient, loving, and caring mother; I would not be the person I am today without her. Thank you also to my father for being

an amazing role model during my academic endeavors, providing direction and support to ensure that my full potential is reached. And thank you to my sister Ana, for always being someone I can turn to for advice. I am truly blessed to have my family there, celebrating every milestone.

Last but not least, I would like to thank the love of my life, Jaynevie. She deserves the highest amount of recognition, due to her strength through a very tough journey. Midway through my thesis, she was diagnosed with Stage 4 ovarian cancer, more specifically a rare germ cell tumour. Her passion and enthusiasm to fight through illness to continue to have a future with me was truly inspirational. This thesis is dedicated to her belief in our relationship to keep growing, no matter what challenges we face. Thank you so much, and I cannot wait to take on more challenges and continuing to grow with you.

This thesis benefited from critical comments by Professors Christian Genest and Russell Steele, who served as examiners. However, any remaining error rests with the author.

Funding in support of this work was provided through grants to Professor Genest by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program. This funding is gratefully acknowledged.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Longitudinal data consist of measurements recorded on the same set of individuals over time. This is in contrast to cross-sectional data, which only capture variables for a number of subjects at a given time. Longitudinal data host the benefit of assessing changes within subjects over time, while a cross-sectional study does not.

A famous example of longitudinal study is the Framingham Heart Study (FHS). It extends over several decades and utilizes longitudinal methods in order to identify and understand which factors or attributes of interest contribute to cardiovascular disease (CVD). Moreover, the FHS study is known for being a rich, longitudinal trans-generational study, that has provided many insights in developing strategies for prevention and early detection [1].

The first objective of this project is to give an overview of statistical methods for perfectly repeated measurements in a longitudinal study. The intent is to circumscribe and highlight

the usefulness of considering longitudinal data. In turn, this will then help to inform the issue that may occur when deviating from a perfect repeated measurement study. These deviations are found in many applied health science studies, where it is common to observe missingness or irregularity within the data.

The second objective of this project is to bring awareness of the importance of checking for irregularity or missingness before selecting the appropriate outcome approach. As will be seen, testing for irregularity in the data is critical before blindly using modeling methods designed for perfectly repeated measurement. A data illustration concerning the pharmacokinetics of Remifentanil will be used to this end.

As mentioned above, the defining feature of a longitudinal study design is that the measurements of the response are taken on the same subject over several visit times, often called occasions. It may be, however, that the number of repeated observations, and their timing, vary widely across subjects within a study. Pullenayegum et al. [13] point out that while there are techniques that can help overcome the resulting bias, the assumptions about the nature of the dependence between visit times and outcome processes generally differ across methods, as do model assumptions. As a result, no single method can handle all plausible visit scenarios. Instead, careful modeling of the visit process can better inform the choice of analytical method for the outcomes.

1.1 Longitudinal Data vs Cross-Sectional Study

The underlying feature that can define a longitudinal study is that measurements of the same individuals are taken repeatedly through time. After collection of data on these measurement occasions, the objective of a longitudinal study is to characterize the change in response over time, and also to examine the covariates that influence change.

Considering this unique feature of having repeated measures on individuals, it is possible to obtain the within-individual change. That is, a longitudinal study design allows the assessment of within-subject changes in the response over time. In contrast, in a cross-sectional study, where the response is measured on a single occasion, researchers can only study the between-subject differences in the response. This implies that a cross-sectional study is only capable of comparisons among sub-populations that can happen to be different in age, say, but will not necessarily provide any information about how individuals change during the duration of the study.

Consider the following example that will illustrate the key distinguishing differences between a longitudinal study and a cross-sectional study. It is important to highlight the distinction between the two methods in order to grasp the possibilities offered by a longitudinal design. Body fat in girls is understood to increase just before or around menarche, leveling off four years after menarche. Suppose that the investigators behind this study are interested in determining the increase in body fat in girls after menarche. In a cross-sectional design, researchers could separate the subjects into two distinct groups: a

group of 10-year-old girls (a pre-menarcheal cohort) and a group of 15-year-old girls (a post-menarcheal cohort). Once they have been separated, the investigators might obtain measurements of body fat on the girls in the two separate groups.

In a cross-sectional design, the statistician may want to use a two-sample (unpaired) t-test, in order to compare average body fat in the two groups. This will not provide an estimate of the change in body fat as girls age from pre-menarche to post-menarche. It is important to note that the effect of growing or aging is inherently a within-subject effect and cannot be studied unless repeated measurements are taken on individuals. Thus, this illustrates one of the limiting aspects of a cross-sectional design: it does not obtain measures of how individuals change with time. This example also illustrates that there are many characteristics that may differentiate the girls in the two different groups, that can possibly alter the relationship between age in girls and body fat in the study [7].

A longitudinal study design measures a single cohort of girls on two separate occasions, say at the age of 10 and at the age of 15. By obtaining repeated measurements on the same individuals, one can provide an estimate of the change in body fat as girls age through menarche. Our response in this illustration is the difference in body fat percentage within each girl; it can be studied using, say, a paired t-test. A longitudinal study will allow each girl in the study to behave as their own control, so that changes in precent body fat throughout the study are estimated free of any between-individual variation in body fat. Overall, this example depicts the value of collection of longitudinal data if possible, in order to study

more accurately possible change in time for subjects.

Another distinctive feature of longitudinal data is that they are clustered [7]. The clusters are composed of the repeated measurements obtained on the same individual at different occasions. It is common to witness positive correlation for observations within a cluster. In the analysis of longitudinal studies, therefore, it is important to account for this correlation. Longitudinal data also exhibit some temporal order, meaning that the ordering of the repeated measures has some important implications on the analysis. It can easily be surmised that as the distance of time intervals between the occasions increases, we may experience a decrease in the correlation.

1.2 Regression Models for Correlated Responses

With great advancements in technology, it has become possible for statisticians to further their methods in analyzing longitudinal and clustered data. In fact, based on a regression paradigm, a broad class of models has now been developed that are designed to deal with correlated data. It must be noted that the use of the term "regression model" is not limited to the uses of standard linear regression for a continuous random variable. Instead, the use of this term more broadly refers to the development of any model that is designed to describe the dependence of the mean of a response variable on its set of covariates. This is obviously accomplished in the form of a regression equation. More specifically, the regression parameters are used to express how the mean of the response variable depends on

its covariates, given in some form of regression equation.

For example, consider the case of a continuous response modeled by a linear regression, where the regression coefficients express the dependence of the mean of the outcome in terms of a linear combination of the covariates. Another example to consider is in the linear logistic model for a binary response, where the regression coefficients express the dependence of the log odds of a positive response in terms of a linear combination of the covariates. Another feature of the regression modeling approach is its ability to incorporate a mixture of both continuous and discrete explanatory variables in an effortless manner.

Regression models are formulated in a way that allows the researchers to have interpretations that bear directly on the scientific question of main interest. The regression paradigm allows for a very flexible approach for analyzing data on repeated measurements, more specifically longitudinal data. These models can also provide a parsimonious description of how the mean response in a longitudinal study changes with time, and also how these changes are related to the features of interest. Therefore, the use of regression models is primarily geared towards describing the discernible patterns of change in the response over time and, via the regression coefficients, their relation to the features.

Chapter 2

Basic Concepts

2.1 Longitudinal Data: Basic Concepts

Longitudinal analysis is concerned with estimating how individuals change throughout the duration of a study and observing how different factors may influence the heterogeneity among the subjects and how they change over time.

2.1.1 Objectives of Longitudinal Analysis

Longitudinal studies allow researchers to enhance their understanding of the development and persistence of disease. They do a great job of acknowledging the natural heterogeneity among individuals in terms of how diseases may develop or progress. There is general belief in the science community that this natural sense of heterogeneity sources from genetic,

environmental, social, and behavioral factors. A longitudinal study allows the discovery of individual characteristics that can illustrate these inter-individual differences in changes in health over time.

In a study done on identical twins, Wong et al. [18] measured DNA methylation across the promoter regions of the dopamine receptor 4 gene (DRD4), the serotonin transporter gene (SLC6A4/SERT), and the X-linked monoamine oxidase A gene (MAOA) using DNA sampled at both ages 5 and 10 years in 46 MZ twin-pairs and 45 DZ twin-pairs; the total sample size was $n = 182$. Their data suggest that the differences are apparent already in early childhood, given that they are genetically identical individuals. This is a key example that demonstrates that however close two individuals can be genetically, there may always exist this natural sense of heterogeneity between individuals. Thus, each individual should be followed separately throughout the study.

In a longitudinal study, we are given the ability to directly assess changes in the response variable over time, as the study participants are measured repeatedly throughout the duration of the study. By obtaining measurements for the same subject repeatedly through time, a longitudinal study can answer the fundamental questions concerning the assessment of within-individual changes in the response. The idea behind within-individual change must be explained and conceptualized to understand the main objective behind a longitudinal study. It can be simply thought of as "change scores" or "difference scores," e.g., the difference between post-treatment and pre-treatment of the response.

The main objective of a longitudinal analysis is to describe trends in within-individual changes in the mean response, and to relate these changes to selected features (e.g., treatment group). This simple notion of expressing within-individual change extends naturally from "difference scores" to more general "response trajectories" over time. Therefore, we want to assess and describe the within-individual changes in the response over time, based on the comparison of the measurements taken on the same individual earlier in the study and at the end.

A longitudinal analysis of within-individual changes takes place in two conceptually distinct stages [7], viz.

a) the within-individual change in the response is characterized in terms of some appropriate summary of the changes in the repeated measurements on each individual over the course of the study (e.g., using "difference scores" or some form of "response trajectories");

b) these estimates of within-individual changes are then related to inter-individual differences in selected covariates.

These two stages of analysis can then be combined in a statistical model for longitudinal data. That is, a single statistical model for longitudinal data can be used to accomplish both, capturing the within-individual change over time and relating within-individual changes in the response to the selected covariates. It is prudent to acknowledge that the assessment of within-individual changes in the response over time can only be achieved with the use of a

longitudinal design. A cross-sectional study simply cannot estimate how individuals change over time, because the response is measured only at a single occasion.

Something to not take for granted here, in a longitudinal design, is that each subject inherently behaves as their own control. This is achieved by comparing each individual's responses at two or more occasions, thereby allowing longitudinal analysis to remove the unavoidable, extraneous sources of variability that naturally exist among individuals. The key point is to realize that there is natural heterogeneity among individuals, that appear in many extraneous variables. However, these extraneous variables are not of any scientific interest *per se*. Nevertheless, they can potentially have an impact on the response variable. It is valuable to note here that these extraneous factors may not even have been measured in the study.

The magic behind a longitudinal design is that any extraneous factors that influence the response (regardless of whether they have been measured) are eliminated out when an individual's responses are compared at two or more occasions. In summary, a longitudinal analysis is the assessment of within-individual changes in the response and also the explanation of systematic differences among individuals in their changes.

2.2 Defining Features of Longitudinal Data

In a study that utilizes a longitudinal design, the participants, or more generally the units being studied, are referred to as individuals or subjects. In many of the longitudinal

designs in practice, the individuals are mainly human subjects; however, in some others, the individuals may be animals. As mentioned earlier, a longitudinal design is composed of repeated measures on the individuals of interest at different time occasions. We refer to a study as "balanced" over time when all individuals have the same number of repeated measurements that were obtained at a common set of occasions.

In health sciences, an almost inescapable feature of longitudinal study is that some individuals will miss their scheduled visit or date of observation. This concern is highlighted, e.g., by Pullenayegum et al. [13], who emphasize that when data are collected longitudinally, measurement times often vary among patients. Note that this is especially the case when the repeated measurements extend over a relatively long study period. In some studies, this would imply that observations be made some time before or after the set occasion time. This will then result in a sequence of observation times that are no longer common to all individuals in the study, due to mistimed measurements. When this happens, the data are said to be "unbalanced" over time. That is, the term "unbalanced data" refers to repeated measurements that are not obtained at a common set of occasions. This is a frequent occurrence when the longitudinal study involves retrospectively collected data (e.g., longitudinal data collected from medical record databases).

Missing data is another common challenge that is often tackled by those leading a longitudinal study. Indeed, missing data are the rule, not the exception, in longitudinal studies in the health sciences. This takes into consideration that not always will subjects

complete the study, or show up for all scheduled visits. When some observations are missing, the data are inherently unbalanced over time, given that not all individuals have the same number of repeated measurements, obtained at a common set of occasions. However, in order to distinguish missing data from any other kind of unbalanced data, we refer to it as being "incomplete." This distinction can be of value, as it emphasizes the fact that an intended measurement on a subject could not be obtained.

A consequence of missing and/or unbalanced data is that additional attention is required in order to recover within-individual change. These ramifications for the analysis of longitudinal data that are incomplete go beyond whether a statistical method is sufficient to handle unbalanced longitudinal data. Anytime there is a case with incomplete data, some loss of information entails. Thus, there is a price to be paid in terms of efficiency or the precision with which changes over time can be estimated.

Besides causing inefficiency, it can also be seen that in some circumstances, missing data can introduce bias in the estimates of change. As longitudinal data in health sciences are rarely balanced and complete, they require examining them with closer detail. There currently exists methods used to quantify and assess the extent of irregularity that are demonstrated by Lokku et al. [11].

Another fundamental feature that is apparent in the statistical analysis on repeated measures on the same individual in a longitudinal design is that they usually are positively correlated. Correlated observations are a positive feature which can be advantageously

used, because they provide more precise estimates of the rate of change than one that is obtained from an equal number of independent observations of different individuals. However, it is important to note that a longitudinal design violates a fundamental assumption of independence between observations, the cornerstone to many standard regression techniques. More will be discussed on how a longitudinal design can combat this.

2.3 Notation

Let Y_{ij} denote the response variable for individual $i \in \{1, \ldots, N\}$ at occasion $j \in \{1, \ldots, n\}$. If the repeated measures on the subjects are assumed to be equally separated in time, this notation will hold. A more precise notation will be used in order to illustrate the different number of repeated occasions by each individual i, namely n_i. Just for simplicity, assume that the notation presented in this section will be expressed using an equal number of repeated measurements, with the same time occasions. Given that we have n repeated measures of the response variable on the same subject, we can now express it as an $n \times 1$ (column) response vector, denoted by $Y_i = (Y_{i1}, \ldots, Y_{in})^\top$, where $^\top$ denotes transposition.

The mean response is the primary focus in the analysis of longitudinal data [7], and in particular the changes that occur in the mean response over n measurement occasions. For each $i \in \{1, \ldots, N\}$, we will denote the mean or expectation of each response by

$$\mu_i = \mathbb{E}(Y_{i1} + \cdots + Y_{in})/n.$$

Additionally, in many longitudinal studies the main goal is to not only study the mean response change in time, but also its relation to the covariates over time. In order to allow the mean response and, in particular, the changes in the mean response, to vary from subject to subject as a function of individual-level covariates, we must make use of the double letter subscripts, viz. $\mu_{ij} = \mathbb{E}(Y_{ij})$.

Here, the expectation can be understood as a long-run average over a large sub-population of subjects who share similar values of covariates at the jth occasion. The quantity μ_{ij} is often referred to as the *conditional* mean response at the jth occasion. This notation allows change over time in the mean response, denoted by the dependence of μ_{ij} on the subscript j. Additionally changes in the mean response can also be related to individual-level covariates, denoted by the dependence of μ_{ij} on the subscript i.

Next, we can consider the correlation or dependence among the n repeated measures on the same subject. It is important to highlight that in a longitudinal analysis setting, a fundamental assumption from standard regression techniques is violated. That is, in a longitudinal analysis we experience a positive feature of correlation among the n repeated measures. In the setting where more than a single observation is obtained on the same subject, the assumption of independent observations is simply not attainable.

This feature can be more easily described in the following way: the response of an individual on one occasion is very likely to be predictive of the response on the same individual taken at a future occasion. A simple example of this occurs when a subject who

experiences a high LDL cholesterol level on one occasion will very likely again experience a high LDL cholesterol level taken on the following occasion. That is, repeated measures on the response for the same subject will result in past responses naturally being predictive of future responses [7]. The dependence among the repeated measures on the same subject can be characterized by their correlation, with a quantitative response variable of interest. Correlated observations help provide a more precise estimate of the rate of change or the effects of the covariates on the rate of change, than of that obtained from an equal number of independent observations of different individuals. As mentioned, this is a positive feature that can be used to our advantage when dealing with longitudinal data.

2.4 Dependence and Correlation

In order to simplify the discussion of dependence and correlation with a longitudinal setting, let us consider a simple longitudinal design, one that is balanced and complete. That is, one with n repeated measurements of the response variable, also made at a common set of occasions on N individuals. If we denote the conditional mean of Y_{ij} by $\mu_{ij} = \mathbb{E}(Y_{ij})$, then the conditional variance of Y_{ij} is defined as

$$\sigma_j^2 = \mathbb{E}\{Y_{ij} - \mathbb{E}(Y_{ij})\}^2 = \mathbb{E}(Y_{ij} - \mu_{ij})^2.$$

While μ_{ij} provides a measure of the location of the centre of the distribution of Y_{ij}, the conditional variance is used to provide a measure of the spread of the values of Y_{ij} around their conditional mean. Note that we have implicitly assumed that the variance can vary from occasion to occasion; this is illustrated by the use of a single-letter subscript, j. In principle, the variance can also be allowed to depend on individual-level covariates, which would then require the use of double subscripts [7].

Next we define a measure of the dependence among responses in a longitudinal study. The *conditional covariance* between the responses taken at two different occasions, say Y_{ij} and Y_{ik}, can be denoted by the following:

$$\sigma_{jk} = \mathbb{E}\{(Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})\}.$$

The equation above is utilized to provide a measure of the linear dependence between the two responses taken at both occasions j and k. The magnitude of the covariance depends not only on the degree of dependence, but also on their units of measurement. If the researchers were to make any changes in the measurement scales, then that would result in possibly significant changes in the value of the covariance. In order to provide a measure of linear dependence between Y_{ij} and Y_{ik} which is in some sense free of the variability resulting from the measurement units, Pearson's correlation is widely used.

The *conditional correlation* between Y_{ij} and Y_{ik} is denoted by

$$\rho_{jk} = \mathbb{E}\left\{\left(\frac{Y_{ij} - \mu_{ij}}{\sigma_j}\right)\left(\frac{Y_{ik} - \mu_{ik}}{\sigma_k}\right)\right\},$$

where σ_j and σ_k are the conditional standard deviations of Y_{ij} and Y_{ik}, respectively. Unlike the covariance, the correlation is a measure of linear dependence that is free of any units or scales of measurement [7]. As can easily be seen, this can be achieved by dividing each variable by its own respective standard deviation.

With longitudinal data, it can be expected that repeated measures on the same individuals are positively correlated. The n repeated measures that are collected on subject i are collected into a (column) vector $Y_i = (Y_{i1}, \ldots, Y_{in})^\top$. We can then define the variance-covariance matrix to be the following two-dimensional array of conditional variances and covariances, viz.

$$\operatorname{cov}\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \operatorname{var}(Y_{i1}) & \operatorname{cov}(Y_{i1}, Y_{i2}) & \cdots & \operatorname{cov}(Y_{i1}, Y_{in}) \\ \operatorname{cov}(Y_{i2}, Y_{i1}) & \operatorname{var}(Y_{i2}) & \cdots & \operatorname{cov}(Y_{i2}, Y_{in}) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(Y_{in}, Y_{i1}) & \operatorname{cov}(Y_{in}, Y_{i2}) & \cdots & \operatorname{var}(Y_{in}) \end{pmatrix},$$

where $\operatorname{cov}(Y_{ij}, Y_{ik}) = \sigma_{jk}$. Note that here we have implicitly assumed that the variances and covariances are constant across individuals.

Note that we often refer to the variance-covariance matrix of Y_i as the covariance matrix of Y_i or simply $\text{cov}(Y_i)$. Thus it can be shown that

$$\text{cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

It is now convenient to define the correlation matrix, $\text{corr}(Y_i)$, in terms of a similar two-dimensional array, viz.

$$\text{corr}(Y_i) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}.$$

We can see a defining feature in both the correlation and covariance matrix mentioned above: it is that both of them display matrices that are symmetric. For example, examining the correlation matrix, it can be seen that

$$\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = \text{corr}(Y_{ik}, Y_{ij}).$$

Let it be noted also that the diagonal elements in the correlation matrix are all equal to 1, because they denote the correlation of a variable with itself.

With longitudinal data, the usual assumptions for standard regression do not hold. More specifically, repeated observations on the same subject are not independent and, more importantly, the variance is not usually constant over the study period. Longitudinal data analysis will consider the heterogeneity of variance over time by allowing the elements on the main diagonal of the covariance matrix to differ. The lack of independence among the repeated measurements is accounted for by allowing the off-diagonal elements of the covariance and correlation matrices to be non-zero.

2.5 Sources of Correlation in Longitudinal Data

It is worth it to pause and wonder why longitudinal data are usually positively correlated. Many of the practical experiences gained from longitudinal data, arising from the biological and health sciences, have led to the following empirical observations about the nature of correlation among repeated measures on the same individual:

1. the correlations are positive;

2. the correlations often decrease with increasing time separation;

3. correlations between repeated measures rarely approach 0 (even in cases where they are taken many years apart);

4. the correlation between a pair of repeated measures taken very closely together in time, rarely approaches 1.

The aforementioned empirical observations have led researchers [7] to identify potentially three sources of variability that can influence the correlation among repeated measures on the same individual, namely

a) between-individual heterogeneity;

b) within-individual biological variation;

c) measurement error.

In fact, citing from Garcia and Marder [8]:

> "ignoring the different sources of correlation in longitudinal studies has severe consequences: higher false positive rates and invalid confidence intervals from underestimated standard errors."

Thus it is critical to consider the different sources of correlation that exists in each unique study, respectively.

2.5.1 Between-Individual Heterogeneity

It can be seen in any longitudinal study that some individuals have the propensity to consistently respond higher than the average, while others may consistently respond below

average. Therefore, one of the sources of a positive correlation in longitudinal studies is the heterogeneity or variability in the response variable that is seen between individuals in the population. Simply put, each individual's underlying propensity to respond — whether it be "high," "medium," or "low" and whether it be due to genetic, environmental, social, or behavioral factors (or some combination of those factors) — is still shared by all of the repeated measurements taken from the same individual.

In other words, it can be expected that a pair of repeated measures from one individual will be more similar than single observations obtained from two randomly selected individuals. This is one reason supporting the intuition of a positive correlation among repeated measures.

There can also be heterogeneity among individuals in their response trajectories over time. For example, when considering a longitudinal design, given a treatment or intervention that should lead to an "improvement" or "increase" in the response variable, different individuals show different levels of responsiveness, resulting in invariably different gains over time.

Changes in the response over time as a result of a treatment or intervention is not expected to be uniform across all individuals. Therefore, an important source of variability in longitudinal data, which has direct impact on the correlation among the repeated measures, is the between-subject variation in the response.

2.5.2 Within-Individual Biological Variation

Another source of variability that has an impact on correlation among repeated longitudinal responses is the within-individual biological differences. More specifically, the inherent biological variability of many health outcomes is a very important source of the variability on the correlation measures. These fluctuations may be due to many possible health related changes in a subject's life. For example, an individual's diet or perhaps circadian rhythm can cause fluctuations. This variability is sometimes referred to as the *inherent within-individual biological variability.*

The underlying idea is to notice that there may exist some underlying biological processes (or some combination of them) that causes changes through time in a relatively smooth and continuous fashion. This will result in random deviations from an individual's underlying response trajectory, which are likely to be more similar, especially when the measurements are obtained in close proximity with respect to time [7]. Moreover, consistent random deviations cannot be assumed to be independent.

To better grasp this notion of within-individual biological variation, consider the following example. If a subject were to have their blood pressure measured repeatedly at 30-minute intervals, the corresponding adjacent measurements will be more highly correlated than if the measurements were to be taken weeks apart. Similarly, if an individual were to record their weight on a daily versus weekly basis, the resulting outcomes that are recorded daily will also be more highly correlated. This will result in the

correlation matrix having a distinctive structure, such that the correlation is decreasing as the time separation between the repeated measurement occasions increases. Therefore, inherent within-individual biological variability in the response variable over time may result in the aforementioned unique correlation structure.

2.5.3 Measurement Error

The last source of variability that is found among longitudinal data is random measurement error. For some common health outcomes, such as height and weight, the variation due to measurement error can almost be ignored from the analysis. However, for many other outcomes, the variability as a result of measurement error can result in substantial effects.

Measurement error is an unavoidable component that exists across all studies, even if they are not longitudinal. Thus a coefficient of reliability has been utilized as a way to express the precision of the measurement procedure. A study was completed by van Smeden et al. [17], in which the authors debunk common myths in epidemiology studies on the prevalence of measurement error on the study. They demonstrated that it is invariant to a large sample size and that measurement error can still occur, regardless of sample size. Furthermore, they also demonstrated that measurement error can affect all types of epidemiological research. Thus measurement error is inevitable and cannot simply be dismissed.

Our understanding of reliability, in the statistics community, refers to the extent to which replicate measurements, taken under similar conditions, are similar. Given that most of the

responses in a resulting longitudinal study will contain measurement error, a question that arises is the potential impact of the measurement error on the analysis. In general, the effect of unreliability is to "attenuate" or shrink the correlation among the repeated measures towards zero.

As an example, if a measure on the response variable is known to contain measurement error and has a reliability of 0.8 in the population of interest, then we must attenuate the correlation among any pair of repeated measures by a factor of 0.8. Therefore, the larger the variance of the measurement errors, the greater the attenuation of the correlation among repeated measures. Clearly, measurement error has an impact on our overall analysis of longitudinal data.

2.5.4 A Final Note

Earlier in Section 2.5, four empirical observations were listed about the correlation among repeated measures in a longitudinal design. Now we can consider how the three aforementioned sources of variability in a longitudinal study can account for the four empirical observations made within many health studies.

First, it was noted that in longitudinal studies, we may experience positive correlations among the repeated measures. Both between- and within-individual biological variation in the response over time result in the positive correlations among repeated measures. The two aforementioned sources of variability behave in conjunction, in order to induce positive

correlation among the repeated measures. It was also noted that as time separation between measurements increases, the correlation tends to decrease. This also is a direct consequence of the inherent within-individual biological variation and/or between individual heterogeneity of response trajectories over time.

The third empirical observation was that the correlations between repeated measures rarely approach zero, even if taken many years apart. This is a direct consequence of between-subject heterogeneity in their underlying propensity to respond.

Our last observation is to the effect that the correlation between a pair of repeated measurements rarely approaches 1, even if taken very closely in time [7]. This final observation results directly from the presence of measurement error. In other words, regardless of how close the measurement occasions are, the correlation between any pair of repeated measurements is constrained by the reliability of the measurement procedure. Therefore, understanding the correlation between observations, and where their natural sources of variability stem from, is useful in considering a longitudinal design.

Chapter 3

Modeling the Mean:

Analyzing Response Profiles

3.1 Linear Models for Longitudinal Data

It is convenient to group the n_i repeated measures of the response variable for individual $i \in \{1, \ldots, N\}$ into an $n_i \times 1$ (column) vector, viz. $Y_i = (Y_{i1}, \ldots, Y_{in_i})^\top$. The vectors $Y_1, \ldots, Y_N$ of responses for the N subjects are assumed to be mutually independent. Associated with each response, Y_{ij}, there is a $p \times 1$ vector of covariates, more formally given, for each $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$, by $X_{ij} = (X_{ij_1}, \ldots, X_{ij_p})^\top$. That is, X_{i1} is a $p \times 1$ vector whose elements are the predictors value associated with the response variable for the ith subject at the first measurement occasion. Similarly, X_{i2} is a $p \times 1$ vector whose elements are the predictors

value associated with the response variable for the ith subject at the second measurement occasion, and so on [7].

Until now, it has been assumed that each individual in the study will have a vector, Y_i, of repeated responses and that associated with each repeated measure, there is a vector of p predictors. For convenience of notation, we can group together the vector of p covariates into a matrix X_i. Now we can consider a linear regression model in order to analyze the changes in the response over time and then further relate the changes to the predictors of interest in the study.

Consider the regression model defined, for all $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$, by

$$Y_{ij} = \beta_1 X_{ij_1} + \cdots + \beta_p X_{ij_p} + \epsilon_{ij},$$

in which $\beta_1, \ldots, \beta_p$ are considered to be unknown regression coefficients relating the mean of Y_{ij} to its corresponding predictors.

This can then be further broken down into n_i separate regression equations for the response, viz.

$$\begin{aligned} Y_{i1} &= \beta_1 X_{i11} + \cdots + \beta_p X_{i1p} + \epsilon_{i1} = X_{i1}^\top \beta + \epsilon_{i1}, \\ &\vdots \\ Y_{in_i} &= \beta_1 X_{in_i1} + \cdots + \beta_p X_{in_ip} + \epsilon_{in_i} = X_{in_i}^\top \beta + \epsilon_{in_i}. \end{aligned}$$

Here, we are considering ϵ_{ij} to be random errors with a zero mean, used to represent the deviations of the responses from their predicted means. Thus until now, there have been no further assumptions, other than observing the patterns of change in the mean response and finding ways to relate that to our predictors. Finally, using vector matrix notation, the regression model that has been developed until now is given by

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,$$

where for each $i \in \{1, \ldots, N\}$, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in})^\top$ is an $n_i \times 1$ vector of random errors that are associated with the corresponding elements of the responses with respect to the ith subject.

In turn, we can see that the response vector is comprised of two key components. First exists the "systematic component $\boldsymbol{X_i\beta}$," and second exists the "random component ϵ_i." Thus any assumptions that are made on the shape of the random errors will then translate into assumptions about the shape of the conditional distribution of Y_i given X_i. The vector of continuous responses is assumed to have a conditional multivariate Gaussian distribution with mean response vector

$$\mathbb{E}(Y_i \mid X_i) = \mu_i = \boldsymbol{X_i\beta}$$

and covariance given by $\Sigma_i = \text{cov}(Y_i \mid X_i)$.

This looks very familiar to the traditional regression settings that are conventionally used. However, recall that while observations from different individuals are assumed to be

independent of one another, we cannot make the same assumption on repeated measurements on the same individual. This lack of independence is captured by the off-diagonal elements in the covariance matrix. It is understood that any departure from independence of observations or assumption of constant variance of the errors can have a major impact on the analysis of a linear regression problem [7]. However, in longitudinal data analysis there are very similar results which suggest that it is the assumptions about dependence among the errors and, ultimately, assumptions about the variances and covariances that will have the greatest impact on statistical inference.

3.2 Modeling the Mean

When modeling the mean of a vector of longitudinal responses, there are two main approaches to be considered. They are described below.

1. Analysis of Response Profiles (ARP)

    a) This approach allows for arbitrary patterns in the mean response over time (more "within" flexibility).

    b) There is no specific time trend that is assumed.

    c) The times of measurement are regarded as levels of a discrete factor.

    d) This approach is most favorable when all individuals are measured at a common set of occasions, and usually the number of occasions is very small.

2. Parametric or Semi-Parametric Curves

    a) This approach assumes a parametric curve (e.g., linear or quadratic trend) for the mean response over time.

    b) It can reduce the number of parameters within the model.

    c) Parametric curves describe the mean responses as an explicit function of time. This notion contains the fact that time can be adjusted for as a covariate either as a discrete or continuous factor. Thus, in contrast to ARP, there is no necessity to require that all individuals in the study have a common set of measurement occasions, nor even the same number of repeated measurements.

Recall that one of the key features of longitudinal data is that repeated measures are obtained on the same subjects over time, and those resulting responses on that same individual are considered to be correlated. In order to yield valid inferences from longitudinal data, it is prudent to properly account for the covariance among repeated measures. Additionally, modeling the covariance carefully and correctly is often a requirement for obtaining valid estimates of the regression parameters, especially when there are missing data. In general, there are three broad approaches to modeling the covariance among repeated measures:

a) unstructured covariance;

b) covariance pattern models;

c) random-effect covariance structures.

3.2.1 Maximum Likelihood for Correlated Responses

We assume that the model being presented can be expressed in terms of a general linear regression model for the mean response vector, $\mathbb{E}(Y_i \mid X_i) = X_i\beta$, where its response vector, Y_i, is assumed to have a conditional distribution that is multivariate Gaussian, with covariance matrix $\text{cov}(Y_i \mid X_i) = \Sigma_i = \Sigma_i(\theta)$, where θ is a $q \times 1$ vector of covariance parameters.

More specifically, with balanced data (i.e., $n_1 = \cdots = n_N = n$), where an "unstructured" covariance matrix has been assumed, the elements of θ are the n variances, and $n(n-1)/2$ pairwise covariances. When there are n_i repeated measures on the same individual $i \in \{1, \ldots, N\}$, we cannot simply make an assumption that these repeated measures are independent.

Next, consider the joint probability density function for the vector of repeated measures. In order to obtain a maximum likelihood estimation of β, it is first assumed that Σ_i or θ is known. To obtain the Maximum Likelihood (ML) estimate of β, we must find the value that will maximize the log-likelihood function. Given that $\mathbf{Y}_i$ is assumed to have a conditional distribution that is multivariate Gaussian, we will maximize the following log-likelihood

$$\ell = -\frac{k}{2}\,\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\ln|\Sigma_i| - \frac{1}{2}\Big\{\sum_{i=1}^{N}(y_i - X_i\beta)^\top\Sigma_i^{-1}(y_i - X_i\beta)\Big\},$$

where $k = n_1 + \cdots + n_N$, which represents the total number of observations. Therefore, the estimator of β that will minimize the above expression, which is also known as the generalized least squares estimator (GLS), can be expressed as

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i^\top \Sigma_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i^\top \Sigma_i^{-1} y_i.$$

The initial assumptions that were imposed on Σ_i and θ are unrealistic and now we must consider how we can loosen those assumptions, within our setting of longitudinal data [7].

First, it is important to highlight a few key properties of the GLS estimate of β. Most notably, any choice of Σ_i will result in an unbiased estimate of β when taking into consideration a correct mean specification. Also, in large samples, the sampling distribution of $\hat{\beta}$ can be shown to be multivariate Gaussian, with mean β and covariance

$$\text{cov}(\hat{\beta}) = \sum_{i=1}^{N} X_i^\top \Sigma_i^{-1} X_i.$$

Here, it is understood that by "large sample" is meant that the sample size, N, grows larger while the number of repeated measures and model parameters stay fixed.

Moreover, although it can be shown that the GLS estimator of β is unbiased for any choice of Σ_i, it can also be shown that the most efficient GLS estimator of β is in fact the one that uses the true value of Σ_i. However, in practice $\Sigma_i(\theta)$ is typically estimated from the data at hand. It is also important to note that maximum likelihood estimates of θ proceed

similarly as β. Once the maximum likelihood estimate of θ has been obtained, then we can simply substitute the estimate $\Sigma_i(\theta)$, say $\hat{\Sigma}_i = \Sigma_i(\theta)$, into the generalized least squares estimator, resulting in

$$\hat{\beta} = \Big(\sum_{i=1}^{N} X_i^\top \hat{\Sigma}_i^{-1} X_i \Big)^{-1} \sum_{i=1}^{N} X_i^\top \hat{\Sigma}_i^{-1} y_i.$$

The latter estimate of β, using the ML estimate of Σ_i, surprisingly has the same properties as when Σ_i is actually known. This suggests that in large samples, $\hat{\beta}$ is a consistent estimator of β. Furthermore, if the distribution of the errors, e_i is assumed to be normal, or even based on further loosened assumptions, such as e_i follows a symmetric distribution, then $\hat{\beta}$ is also an unbiased estimator of β. Secondly, the sampling distribution of $\hat{\beta}$, when the ML estimate of Σ_i is used, will be approximately multivariate normal with a mean β and covariance of

$$\text{cov}(\hat{\beta}) = \Big(\sum_{i=1}^{N} X_i^\top \Sigma^{-1} X_i \Big)^{-1}.$$

This extends one step further, such that the properties of $\hat{\beta}$ hold in large samples, even when the assumption of Y_i following a multivariate normal is not satisfied, provided that the data are complete.

In conclusion, there is no apparent penalty with respect to the sampling distribution of $\hat{\beta}$ from needing to estimate Σ_i from the longitudinal data at hand. A very important note to bear in mind, is that this is a large-sample property of $\hat{\beta}$, such that as $N \to \infty$. Due to the limitations of the magnitude of data that can be collected during a typical longitudinal

study, we can expect the properties of the sampling distribution of $\hat{\beta}$ to be adversely affected.

An alternative approach to maximum likelihood estimation, known as Restricted Maximum Likelihood estimation (REML), is a valid approach. REML is an approach to be considered, since the maximum likelihood estimation of Σ_i has a well-known bias in finite samples, such as underestimating the diagonal values of Σ_i. The objective in REML estimation is to separate the data that are being used for estimation of Σ_i from those used for estimation of β. This will result in the REML estimator to be less seriously biased than the ML estimate of Σ_i. However, it should be understood that as the sample size, N, gets substantially larger than p (the dimension of β), the difference between the ML and REML estimation becomes less noticeable.

3.3 Modeling the Mean: Analysis of Response Profiles

Analysis of response profiles is a method used to model the patterns in the mean response over time. This method imposes a minimal structure or restriction on the mean responses over time on the covariance among the repeated measures. The analysis of response profiles is mainly used in applications where the longitudinal design is balanced, and also having a common set of measurements that were obtained among the subjects. The main objective of analysis of response profiles is to be able to characterize the mean change in the response over time in the groups. Furthermore, analysis of response profiles is also used to determine if the shapes of the mean response profiles differ among groups (i.e., treatment vs placebo).

Next it is important to understand the consequences of longitudinal data hypotheses concerning response profiles. Given a sequence of n repeated measurements on a distinct number of groups of subjects, there are three main questions regarding response profiles that can be further investigated [7]. They are as follows.

1. Are the mean responses between the groups similar, i.e., are the mean response profiles parallel also between the groups ?
    - This is a question that is concerned with group $\times$ time interaction effect.
2. Assuming that the population mean response profiles are parallel, are the means constant over time, in the sense that the mean response profiles are flat?
    - This is a question that is concerned with the time effect.
3. Assuming that the population mean response profiles are parallel, do the mean response profiles for the groups coincide?
    - This is a question that is concerned with the group effect.

3.3.1 Example of Analysis of Response Profiles

Next we can consider how to implement the analysis of response profiles in the general linear model specified, for each $i \in \{1, \ldots, N\}$, by

$$\mathbb{E}(Y_i \mid X_i) = \mu_i = X_i\beta$$

for appropriate choices of X_i. Consider the following example, where there exist two groups, a treatment and a placebo, and they are measured at three occasions. Let n be the number of repeated measures and N to be the total number of subjects. The model for this longitudinal design with $G = 2$ groups and $n = 3$ measurement occasions will require $G \times n$ parameters for the G mean response profiles.

Group	1	2	3
Treatment (t)	$\mu_1(t)$	$\mu_2(t)$	$\mu_3(t)$
Placebo (p)	$\mu_1(p)$	$\mu_2(p)$	$\mu_3(p)$

For the first group (Treatment group), let the design matrix be

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then for the second group (Placebo), let the design matrix be

$$X_i = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then in terms of the model $\mathbb{E}(Y_i \mid X_i) = \mu_i = X_i\beta$, where $\beta = (\beta_1, \ldots, \beta_6)^\top$, this can then be further categorized between the two groups, as follows:

Treatment:

$$\mu(T) = \begin{pmatrix} \mu_1(t) \\ \mu_2(t) \\ \mu_3(t) \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Placebo:

$$\mu(P) = \begin{pmatrix} \mu_1(p) \\ \mu_2(p) \\ \mu_3(p) \end{pmatrix} = \begin{pmatrix} \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix}.$$

As a result, the hypotheses about the change in mean response profiles in the two groups had been previously expressed in terms of μ, but can now be re-expressed in terms of hypotheses with respect to the components of β. More specifically as an illustration, if we are looking to test the hypothesis of no group $\times$ time interaction effect, it can be expressed as

$$\mathcal{H}_0 : \beta_1 - \beta_4 = \beta_2 - \beta_5 = \beta_3 - \beta_6.$$

When we fail to reject the hypothesis of parallel response profiles, it is common to further investigate the hypotheses concerning the main effects of time/group. This is dependent on the relevance of the design of the study when initially starting.

Suppose that the analysis of response profiles can be expressed in terms of the linear model, where β is a $p \times 1$ vector of regression coefficients (with $p = G \times n$). Furthermore, once the covariance of Y_i has been determined, estimates of the group $\times$ time interaction, and the main effects of time and group are then possible to find, using the maximum likelihood estimation of β.

In the analysis of response profiles, the covariance of Y_i is generally assumed to be unstructured with little to no constraints on the $n(n+1)/2$ covariance parameters. A condition to consider is that the covariance matrix must yield a symmetric matrix that is positive definite. While the repeated measures can be highly correlated, there must be no redundancy. The condition further ensures that no linear combination of the responses can have a negative variance. Given REML (Restricted Maximum Likelihood) or ML (Maximum Likelihood) estimates of β, and their standard errors, the test of group $\times$ time interaction and the main effects can be simply carried out using multivariate Wald tests.

In conclusion, it is fair to state that the analysis of response profiles is a conceptually straightforward way to analyze data. However, it does require for the longitudinal study to be balanced, and the timing of the repeated measurements to remain common among all subjects, for it to remain straightforward. This is certainly a limitation to many applied health and medical studies. The main feature of analysis of response profiles is that it allows for arbitrary patterns on the mean response over time, and arbitrary patterns in the covariance over time. This results in this method having a certain amount of robustness,

due to the fact that the potential risks of bias due to misspecification of the models of the mean and covariance are minimal.

There also exist important drawbacks to consider about the use of analysis of response profiles in a longitudinal design, especially when it is not balanced. In particular, the analysis of response profiles ignores the time ordering of the repeated measures taken on the subjects [7]. Moreover, in the analysis of response profiles, the number of estimated parameters ($G \times n$ mean parameters, and $n(n+1)/2$ covariance parameters) will grow rapidly with respect to the number, n, of measurement occasions. Consequently, it is not hard to see why this method may be preferred only in special settings, more specifically those where the total number of subjects, N, is relatively large in comparison to the number, n, of measurement occasions.

3.4 Modeling the Mean: Parametric Curves

In our previous approach of modeling the mean in a longitudinal design, analysis of response profiles was effectively imposing no structure on the underlying mean response trend over time. There are two major drawbacks that limit the usefulness of using analysis of response profiles within a longitudinal design. The first is that a statistical test of the null hypothesis of no Group $\times$ Time interaction is a global test and cannot really produce insightful results, but rather only a broad assessment of whether the mean response profiles are the same between the different groups.

Consider the following situation, where there are two groups (treatment vs placebo) and the subjects are all measured at a common set of measurements. For each $i \in \{1, \ldots, N\}$ and $j \in \{1, \ldots, n_i\}$, let

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Group}_i + \beta_4 \text{Group}_i \times \text{Time}_{ij},$$

where

$$\text{Group}_i = \begin{cases} 1 & \text{if } i \in \text{treatment}, \\ 0 & \text{if } i \in \text{placebo}. \end{cases}$$

Then, the null and alternative hypotheses can be expressed as $\mathcal{H}_0 : \beta_4 = 0$ vs $\mathcal{H}_1 : \beta_4 \neq 0$. If $\mathcal{H}_0$ is rejected, we are still left unable to indicate the specific ways in which the mean responses differ between the placebo and treatment. This will result in further analysis.

The second drawback that limits the usefulness of analysis of response profiles is that it completely ignores the time-ordering of the repeated measurements [7]. The use of analysis of response profiles is unable to detect that the repeated measurements can be considered as observations of some continuous, underlying response process over time.

Therefore, the analysis of response profiles uses a saturated model for the mean response over time, which results in a fit that is almost perfect, thereby not allowing the method to describe the most noteworthy aspect of the changes of the mean response over time. This would in turn make prediction not possible to be obtained. More specifically, in terms of

some pattern that can be described or explained in some substantive or theoretical manner.

In contrast, when introducing the use of fitting parametric or semi-parametric curves to longitudinal data, it can certainly be justified both on statistical and substantive grounds. Naturally in many application of longitudinal studies, one can observe that the true underlying mean response process is likely to change over time in a relatively smooth, monotonically increasing or decreasing pattern, at least during the duration of the study itself. We can see that from a statistical perspective, the fitting of a parsimonious model for the mean response can result in statistical tests of covariate effects (e.g., treatment $\times$ time interactions) that have proved to have a greater power than seen in analysis of response profiles.

3.4.1 Polynomial Trends in Time

This approach is useful to model the means as an explicit function of time. It is also useful to handle highly unbalanced designs in a relatively simply manner. In this model, the slope for time can have a direct interpretation, in terms of a constant change in the mean response over time for a single unit of change.

Once again, consider the hypothetical two-group study comparing a novel treatment and a control (i.e., placebo). We can adopt the following linear trend model if the mean response

changes are in an approximately linear fashion, viz.

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Group}_i + \beta_4 \text{Group}_i \times \text{Time}_{ij},$$

where

- ✓ $\text{Group}_i = 1$ if the ith individual was assigned to treatment;
- ✓ $\text{Group}_i = 0$, if the ith individual was assigned otherwise (in control/placebo group);
- ✓ Time_{ij} denotes the measurement time for the jth measurement, on the ith individual.

Group_i requires a single index since any given subject does not change treatment groups over the duration of the study. Also, it is useful to note the use of two indices for Time_{ij}, because we are implicitly allowing for the fact that there may exist mistimed measurements (i.e., $\text{Time}_{ij} \neq \text{Time}_{i'j}$, where i and i' denote two different subjects). Furthermore, the model for the mean for subjects assigned to the control group ($\text{Group}_i = 0$) can be expressed as

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij}.$$

Additionally, the model for the mean for subjects assigned to the novel treatment group can be expressed as ($\text{Group}_i = 1$), viz.

$$\mathbb{E}(Y_{ij}) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\, \text{Time}_{ij}.$$

Thus each group's mean response is assumed to change linearly over time. The parameter estimates have the following interpretations:

- ✓ β_1 is the intercept in the control group ("reference group"), while $\beta_1+\beta_3$ is the intercept within the treatment group. Both intercepts for each of the two respective groups have similar interpretation in terms of the mean response when $\text{Time}_{ij} = 0$. More generally, β_1 is interpreted as the mean response when all other covariates are set to zero.

- ✓ The slope of change in the mean response per unit change in time is β_2 within the control group, and $\beta_2 + \beta_4$ within the corresponding treatment group. The main objective of a longitudinal design is typically concerned with a comparison of the changes in the mean response over time; this can directly be translated into a comparison of slopes. Thus if $\beta_4 = 0$, then the two groups do not differ in terms of changes in the mean response over time.

When changes in the mean response do not appear to be linear over time, it is possible to take into consideration higher-order polynomial trends. In a quadratic trend model, changes in mean response are not constant, as they were in a linear trend model over the period of the study. Alternatively, the rate of change in the mean response for a quadratic trend depends on time. More specifically, the rate of change in the mean response depends on whether the focus is on changes that occur early or late in the study. This implies that the rate of change must be expressed in terms of two parameters.

Reconsider the hypothetical scenario with the two-group study comparing placebo and treatment. Assuming that the changes in the mean response can be approximated by some quadratic trends, one can formulate the model as

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Time}_{ij}^2 + \beta_4 \text{Group}_i + \beta_5 \text{Group}_i \times \text{Time}_{ij} + \beta_6 \text{Group}_i \times \text{Time}_{ij}^2.$$

The model for the mean for subjects assigned to the control group ($\text{Group}_i = 0$) can then be expressed as

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Time}_{ij}^2.$$

Additionally, the model for the mean for subjects assigned to the novel treatment group ($\text{Group}_i = 1$) can be expressed as

$$\mathbb{E}(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_{ij} + (\beta_3 + \beta_6)\text{Time}_{ij}^2.$$

Depending on Time_{ij}, the mean response changes at a different rate when considering quadratic trend models. It is also important to consider that a quadratic trend will experience a turning point where the trend changes, i.e., it can be from an increasing trend over time to a decreasing trend over time or vice versa. When taking into consideration higher-order polynomials trend models, there exists a natural hierarchy of effects that

imposes few implications for testing hypotheses with regards to linear, quadratic, and higher-ordered polynomial trends. More specifically, higher-ordered terms should be tested before lower-ordered terms [7]. If found to be appropriate, the higher-ordered terms should then be removed from the model.

3.4.2 Linear Splines

With the introduction of higher-order polynomials in time, there are many non-linearities apparent within longitudinal data that can simply be accommodated for. However, it is valuable to note that as the degree of the polynomial increases, we lose interpretability of our regression coefficients. In some applications of longitudinal data, we can see that it is difficult to model accurately the change in the mean response over time when only characterized by first- or second-degree polynomials. This is most often the case in circumstances where the mean response increases (or decreases) rapidly for some duration, and then more slowly thereafter or vice versa. When this pattern of change is present, we can consider using linear spline models instead.

Linear splines offer a very useful and flexible approach to modeling non-linear trends that cannot be approximated by simple polynomials in time. The foundational idea behind linear splines is quite simple. In fact, the time axis is divided into a series of segments; then for each respective segment, consider a model for the trend over time. This leads to piece-wise linear trends which may have different slopes but which are joined or tied together at fixed

times. The locations in which the lines meet are also known as knots.

The advantage of this model is that it allows the mean response to increase or decrease as time proceeds. Furthermore, the sign and magnitude of the regression slope can vary from one segment to the next.

Once again, we return to the two-group hypothetical study between a treatment and placebo group. If the mean response changes in a piece-wise linear way, we can fit the following linear spline model with knot at t^*:

$$\begin{aligned}\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3(\text{Time}_{ij} - t^*)_+ + \beta_4 \text{Group}_i \\ + \beta_5 \text{Group}_i \times \text{Time}_{ij} + \beta_6 \text{Group}_i \times (\text{Time}_{ij} - t^*)_+,\end{aligned}$$

where $x_+ = \max(x, 0)$, known as a truncated line function, is defined as a function that equals x when x is positive and is equal to zero otherwise. Thus,

$$(\text{Time}_{ij} - t^*)_+ = \begin{cases} (\text{Time}_{ij} - t^*) & \text{when Time}_{ij} > t^*, \\ 0 & \text{when Time}_{ij} \leq t^*. \end{cases}$$

So in this model, the means for the subjects in the placebo (control) group are

$$\mathbb{E}(Y_{ij}) = \beta_1 + \beta_2 \text{Time}_{ij} + \beta_3(\text{Time}_{ij} - t^*)_+,$$

which can be further expressed in terms of the mean response prior to and after t^*, viz.

$$\mathbb{E}(Y_{ij}) = \begin{cases} (\beta_1 - \beta_3 t^*) + (\beta_2 + \beta_3)\text{Time}_{ij} & \text{when Time}_{ij} > t^*, \\ \beta_1 + \beta_2 \text{Time}_{ij} & \text{when Time}_{ij} \leq t^*. \end{cases}$$

Thus, in the control group:

a) Slope prior to t^*: β_2;

b) Slope after t^*: $\beta_2 + \beta_3$.

Similarly the means for the subjects in the treatment group can be given as follows:

$$\mathbb{E}(Y_{ij}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_{ij} + (\beta_3 + \beta_6)(\text{Time}_{ij} - t^*)_+.$$

Once again, when it is further expressed in terms of the mean response prior to and after t^*, one gets [7]:

$$\mathbb{E}(Y_{ij}) = \begin{cases} \{(\beta_1 + \beta_4) - (\beta_3 + \beta_6)t^*\} + (\beta_2 + \beta_3 + \beta_5 + \beta_6)\text{Time}_{ij} & \text{when Time}_{ij} > t^*, \\ (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_{ij} & \text{when Time}_{ij} \leq t^*. \end{cases}$$

Then in terms of group comparisons, the null hypothesis of no group differences in patterns of change over time can be expressed as $\mathcal{H}_0 : \beta_5 = \beta_6 = 0$.

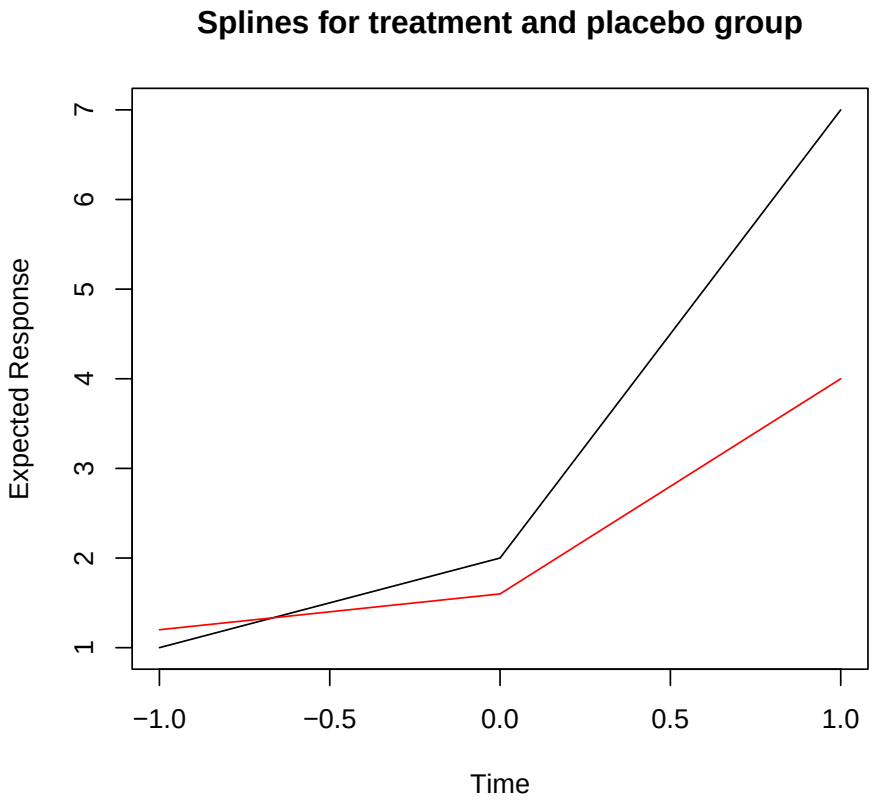


Figure 3.1: Spline regression by group for Treatment (black) and Placebo (red) with knot at Time = 0 for both groups

Comparisons of the groups before and after t^* are also possible. In the aforementioned example, we only considered the case of a single knot. However, further applications can be extended to include more knots, joining the line segments. More generally, it is convenient to consider that a spline model with k knots will produce $k+1$ line segments with $k+1$ corresponding slopes. Thus, it is possible to accommodate complex non-linear patterns for the changes in the mean response by including a sufficient and effective number of variables $(\text{Time}_{ij} - t^*)_+$, with knots located at t_k^*, for $k \in \{1, \ldots, K\}$. However, in many longitudinal

applications, the data can be sufficiently approximated by simple piece-wise linear models with at most one or two knots that are located at precisely selected locations. Choosing such locations for each respective knot is not always an easy task; it may require careful subject matter expert input.

Through the use of both parametric and semi-parametric modeling, both polynomial trends and spline models can be expressed in terms of the general linear regression model, $\mu_i = X_i\beta$. Furthermore, once the covariance of Y_i has been specified, the use of restricted maximum likelihood estimation of β and the construction of confidence intervals and tests of hypotheses can be achieved. Recall that in the use of analysis of response profiles, the covariance of Y_i is assumed to be unstructured with no constraints on the covariance parameters, other than yielding a symmetric matrix and one that is positive-definite. However, through the use of parametric or semi-parametric modeling, more parsimonious models for the covariance can be considered. The use of parametric curves is most appealing in circumstances where the longitudinal design is dealing with inherently unbalanced data over time. In conclusion, given models for both the mean and covariance, REML estimates of β, and their standard errors (based on the estimated covariance of $\hat{\beta}$), can be obtained.

Chapter 4

Modeling the Covariance

The most defining feature of longitudinal data is that they are correlated. This further demonstrates the level of importance behind appropriately modeling the covariance or the time dependence between repeated measure obtained on the same subjects. If an appropriate model is selected for the covariance, that may then lead to correct standard errors, valid inferences about the regression parameters can be made. By accounting for this important feature of longitudinal data, we can increase the precision in which the regression parameters can be estimated. This is due to the fact, that positive correlation among repeated measures reduces the variability of the estimate of change within the subjects. Furthermore, in settings where there are missing data, correct modeling of the covariance is required in order to obtain valid estimates and inferences.

In longitudinal data, our primary objective is to model the conditional mean response over time and the conditional covariance among repeated measures. Although we model these two aspects separately, it is important to note that they are in fact interrelated. This interdependence occurs because the vector of residuals (observed responses minus fitted responses) depends on the specification of the model for the conditional mean. Therefore, any misspecification of the model for the mean can potentially result in a different choice when selecting to model the covariance. Accordingly, it is important to consider this interdependence when developing a modeling strategy with longitudinal data.

Longitudinal data not only have the feature of being correlated, but in fact for the most part they are also positively correlated. This is something that can be taken advantage of. Consider a simple longitudinal design that is interested in the change in a particular health measure that was obtained before and after receiving some health intervention. Since we only have two repeated measurements, the analysis will be focused on the difference scores, say $Y_{i2} - Y_{i1}$, for each subject. Then the variability of the difference scores can be given by

$$\begin{aligned}\operatorname{var}(Y_{i2} - Y_{i1}) &= \operatorname{var}(Y_{i1}) + \operatorname{var}(Y_{i2}) - 2\operatorname{cov}(Y_{i2}, Y_{i1}) \\ &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} = \sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2,\end{aligned}$$

where ρ_{12} is the correlation among the pair of responses, Y_{i1} and Y_{i2}.

However, suppose that an alternate study was designed to assess the impact of this particular health intervention. This time a cross-sectional design is adopted, where the study participants are randomly assigned among two groups. One group will receive the intervention, while the other is the control group receiving a placebo. The variance of the difference between the responses of any two individuals, when one is randomly selected from treatment and the other from the control group, is given by

$$\text{var}(Y_{i2} - Y_{i1}) = \text{var}(Y_{i1}) + \text{var}(Y_{i2}) = \sigma_1^2 + \sigma_2^2.$$

Therefore, if the correlation is said to be positive among the repeated measures, then the variability of the within-individual differences will always be smaller than the variability of between-individual differences. Thus in a longitudinal design that consists of data that exhibit a relatively strong positively correlated data, then the variability of the within-individual differences can be substantially smaller than that for corresponding between-individual differences. The covariance among repeated measures is not usually the primary aspect behind modeling in a longitudinal design; however, it must not be ignored [7]. Furthermore, the positive correlation in longitudinal data allows us to estimate changes in the mean response, and their relations to the covariates, with much better precision than if the data happened to be uncorrelated.

4.1 Unstructured Covariance

In a setting where the study design only requires a relatively small number of measurement occasions and all subjects are measured at a common set of occasions, it may be reasonable to allow a covariance structure to be arbitrary, with all of its elements unconstrained. The only requirement of the matrix that exists is that it is symmetric and positive definite. When no explicit structure has been assumed for the covariance among repeated measures, the resulting covariance matrix is referred to as "unstructured."

With n measurement occasions, the unstructured covariance matrix has $n(n+1)/2$ parameters, which is comprised of n variances at each occasion and the $n(n-1)/2$ pairwise covariances, viz.

$$\operatorname{cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n1} & \dots & \sigma_n^2 \end{pmatrix}. \tag{4.1}$$

The advantage that lies in assuming an unstructured covariance model is that no assumptions are needed to be made about the variances and covariances [7]. However, it is easy to notice one of its potential drawbacks by observing equation (4.1). It is evident that the number of covariance parameters to be estimated will grow rapidly with respect to the number of measurement occasions. This then will further lead to unstable results in estimation, especially when the number of covariance parameters that need to be estimated

is large relative to the sample size.

Therefore, it can be seen that the setting best to use an unstructured covariance is when N is relatively large compared to the number of covariance parameters, $n(n+1)/2$. Furthermore, an unstructured covariance structure is not favorable within a setting that consists of mistimed measurements, or more generally measurements made at irregular intervals. In conclusion, when the sample size is not large enough or the longitudinal data is subject to irregular data, imposing some structure on the covariance may be favorable.

4.2 Covariance Pattern Models

There is a fine balance that is desired upon imposing structure on a covariance matrix. Imposing too little structure on the covariance can result in weaker inferences concerning the regression coefficients. When some structure is introduced and imposed on the covariance, it is then possible to improve estimation efforts of the regression coefficients, β. However, finding the balance is key, since imposing too much structure may lead to a potential risk of model misspecification. This could ultimately result in misleading inferences concerning β. This is the classic trade-off between bias and precision when modeling a covariance.

Structure can be developed into the covariance by adopting a covariance pattern model. Covariance pattern models were originally developed for time series data. Many of the models for time series data result in relatively parsimonious models for the covariance that can also be used in longitudinal studies. Below, we will describe the most widely used covariance

pattern models for longitudinal data.

4.2.1 Compound Symmetry

Historically this is one of the first covariance pattern models used for analysis of correlated measurements. A compound symmetry covariance structure assumes that the variance is constant across occasions, say σ^2 and $\text{corr}(Y_{ij}, Y_{ik}) = \rho$, for all j and k. That is,

$$\text{cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}$$

with the constraint that $\rho \geq 0$. A compound symmetry covariance is very parsimonious, with only two parameters regardless of the number, n, of measurement occasions. However, it is important to notice that it makes a strong assumption, that the correlation between any pair of measurements is the same regardless of the time interval between the measurement occasions [7]. This is an unappealing feature in longitudinal data, given that with increasing time separation between measurement occasions the correlations are expected to decay. Furthermore, it is unrealistic to assume constant variance across time in many longitudinal applications.

Interestingly, the compound symmetry covariance can also be considered as a random effects model. The theoretical justification is as a result of the mean response being thought to depend on a combination of both population parameters, β and also a single individual-specific random effect, b_i, viz.

$$Y_{ij} = X_{ij}^{\top}\beta + b_i + \epsilon_{ij},$$

where b_i is considered a random effect and ϵ_{ij} is a within-individual measurement error. Thus averaged over the random effect, this will create a compound symmetry structure on the covariance matrix such that $\rho \geq 0$; for details, refer to [7].

4.2.2 Toeplitz

The Toeplitz covariance pattern makes the assumption that any pair of responses that are equally separated in time have the same correlation. In a longitudinal design when the covariance matrix takes on the Toeplitz form, it is assumed that the variance is constant across occasions, say σ^2, and $\text{corr}(Y_{ij}, Y_{ik}) = \rho_k$, for all j and k. That is,

$$\mathrm{cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{pmatrix}.$$

The Toeplitz structure assumes that the correlation among responses taken on adjacent occasions is constant, ρ_1. This suggests that a Toeplitz form should only be considered within a setting that consists of measurement occasions being taken at equal intervals of time [7]. Finally, we note that the Toeplitz covariance has n parameters (one variance parameter, and $n-1$ correlation parameters). A special case of the Toeplitz structure is the (first-order) autoregressive covariance.

4.2.3 Autoregressive

The autoregressive model for the covariance assumes that the variance is constant across occasions, say σ^2 and $\mathrm{corr}(Y_{ij}, Y_{ik}) = \rho^k$, for all j and k, and for any $\rho \geq 0$. That is,

$$\operatorname{cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

The autoregressive covariance is also very parsimonious and only has two parameters, regardless of the number of measurement occasions, unlike the Toeplitz form. The autoregressive model for the covariance has a Toeplitz form. Once again, this form is only desirable in settings where the measurement occasions are at equal intervals (or approximately equal) of time. However, unlike the Toeplitz form, the autoregressive model experiences a decline in the correlations over time, as separation between the pairs of repeated measures increases.

In conclusion, the compound symmetry, Toeplitz, and autoregressive covariances all assume that the variance is constant across occasions. However, this assumption can be relaxed to further incorporate covariance pattern models with heterogeneous variances, i.e., $\operatorname{var}(Y_{ij}) = \sigma_j^2$; see [7].

As an example, consider the following autoregressive covariance model with heterogeneous variance, viz.

$$\text{cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \dots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \dots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \dots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_n\sigma_1 & \rho^{n-2}\sigma_n\sigma_2 & \rho^{n-3}\sigma_n\sigma_3 & \dots & \sigma_n^2 \end{pmatrix}.$$

Note that within the aforementioned example, there are $n + 1$ parameters (n variance parameters, and one correlation parameter).

4.2.4 Banded

The banded covariance patterns make the assumption that the correlation is zero beyond some specified interval. For example, a banded covariance pattern with a band size of 3 will imply that $\text{corr}(Y_{ij}, Y_{ik}) = 0$ for $k \geq 3$. In fact it is possible to apply a banded pattern to any of the covariance models mentioned thus far. In general, banding makes quite a strong assumption about how quickly the correlation decays to zero with some increasing time separation between the measurement occasions. Through empirical observations made from many longitudinal data applications, there have been very few instances in health sciences where the correlation decays to zero, even in studies with lengthy follow-up periods.

4.2.5 Exponential

In a longitudinal design where the measurement occasions are not equally spaced over time, the formulation of the autoregressive covariance model can be generalized with the following approach.

Let $t_{i1}, \ldots, t_{in}$ denote the observation times for the ith individual. Also assume that the variance is constant across measurement occasions, say σ^2 and $\text{corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij}-t_{ik}|}$, for $\rho \geq 0$. In other words, the correlation between any pair of repeated measurements will decrease exponentially with the time separations between them. This can further be re-expressed in the following way:

$$\text{cov}(Y_{ij}, Y_{ik}) = \sigma^2 \rho^{|t_{ij}-t_{ik}|} = \sigma^2 \exp(-\theta|t_{ij} - t_{ik}|),$$

where $\theta = -\ln(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$. A feature that must be considered about the exponential covariance pattern model is that it assumes a correlation of 1 if measurements are repeatedly made at the same occasion [7]. This is an unrealistic assumption, because it is considering measurements to be made with no measurement error. Additionally, this model considers the correlation to decay quite rapidly as the time separation between the measurements increases. This is also an unappealing feature to many longitudinal designs.

4.2.6 Hybrid Models

By combining the autoregressive and compound symmetry models, it is possible to overcome a lot of the unappealing features of the models that have been discussed above. Consider a model for the covariance in which

$$\operatorname{cov}(Y_i) = \Sigma_1 + \Sigma_2,$$

where

$$\Sigma_1 = \sigma_1^2 \begin{pmatrix} 1 & \rho_1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \dots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \dots & 1 \end{pmatrix}$$

and

$$\Sigma_2 = \sigma_2^2 \begin{pmatrix} 1 & \rho_2^{|t_{i1}-t_{i2}|} & \rho_2^{|t_{i1}-t_{i3}|} & \dots & \rho_2^{|t_{i1}-t_{in}|} \\ \rho_2^{|t_{i2}-t_{i1}|} & 1 & \rho_2^{|t_{i2}-t_{i3}|} & \dots & \rho_2^{|t_{i2}-t_{in}|} \\ \rho_2^{|t_{i3}-t_{i1}|} & \rho_2^{|t_{i3}-t_{i2}|} & 1 & \dots & \rho_2^{|t_{i3}-t_{in}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2^{|t_{in}-t_{i1}|} & \rho_2^{|t_{in}-t_{i2}|} & \rho_2^{|t_{in}-t_{i3}|} & \dots & 1 \end{pmatrix}.$$

In this model, one has

$$\operatorname{var}(Y_{ij}) = \sigma_1^2 + \sigma_2^2, \quad \operatorname{cov}(Y_{ij}, Y_{ik}) = \rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2,$$

and

$$\text{corr}(Y_{ij}, Y_{ik}) = \frac{\rho_1\sigma_1^2 + \rho_2^{|t_{ij}-t_{ik}|}\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Finally, this implies that the correlation between a measure that has been replicated (taken at same measurement occasion) on a subject is $(\rho_1\sigma_1^2+\sigma_2^2)/(\sigma_1^2+\sigma_2^2)$. This measure is clearly less than 1, when $\rho_1 < 1$. Lastly as the time separation increases, the correlation no longer will decay to zero, but will instead experience a minimum of $\rho_1\sigma_1^2/(\sigma_1^2 + \sigma_2^2)$; see [7].

As mentioned, compound symmetry is also considered as a random effects model, thus in our hybrid model the same argument can be made with respect to Σ_1. Re-writing the compound symmetry covariance to have a random effects model, one gets

$$\Sigma_1 = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \dots & \sigma_b^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix}.$$

In conclusion, this new consideration of our hybrid model in which we are combining an autoregressive and random effects compound symmetry models, can lead to a new interpretation of the total variance. It can be decomposed into the sum of three sources of variability captured in Y_{ij}, where σ_2^2 is the autoregressive variance, σ_b^2 is the subject-to-subject variability and lastly, σ_ϵ^2 the measurement error source of variability.

4.3 Choosing a Covariance Model

Recall the choice of models for the mean and covariance are interdependent. This illustrates the importance of following a sensible choice of models for both aspects of longitudinal data. Confidence intervals and tests of hypotheses concerning the regression coefficients, β, depend critically on the correct model being specified for the covariance. This will result in us naturally looking to choose a suitable model for the covariance as our first step.

The choice of model for the covariance requires a "maximal" model for the mean that will then minimize any potential misspecification of the model for the mean. The choice of a maximal model in a longitudinal study where there is a balanced design and a small number of discrete covariates (i.e., treatment assignment, exposure levels, or some characteristics of the subjects) is relatively simply. This is due to the fact that it is possible to choose a maximal model that includes the main effects of time (our within-subject factor), and all other main effects, in addition to their two-way and higher-order interactions. However, in a longitudinal study that has many covariates, in which some of them may be quantitative rather than discrete, the choice of maximal model is somewhat less obvious. In this case, it may not be realistic to consider a saturated model for the mean response. In summary, there is no real clear blueprint to follow when choosing the maximal model, especially when there are many covariates that can be included in the model. The maximal model for the mean ideally consists of one that excludes higher-order interactions.

Given a maximal model for the mean, a sequence of covariance pattern models can then be fitted to the data at hand. The choice between the models will be made by comparing the maximum likelihoods for each of the fitted covariance pattern models. When we obtain any pair of models that are nested, we can utilize a likelihood ratio test statistic to compare between the "Full" and "Reduced" model. Suppose two covariance models are said to be nested, when the "reduced" model is a special case of our "Full model," so that when we say the reduced model holds, then it is necessary that the full model also holds.

The likelihood ratio test for the two nested covariance models can be constructed by comparing the maximized REML log-likelihoods, say $\hat{\ell}_{\text{Full}}$ and $\hat{\ell}_{\text{Reduced}}$, for the "Full" and "Reduced" model, respectively. Note that the use of REML is preferred as an alternate to maximum likelihood because it reduces the well-known finite-sample bias in the estimation of the covariance. Then the likelihood test can be given as

$$G^2 = 2\Big(\hat{\ell}_{\text{Full}} - \hat{\ell}_{\text{Reduced}}\Big),$$

and comparing the statistic to percentiles of the χ^2 distribution with degrees of freedom $\Delta_{\text{Full}-\text{Reduced}}$ equal to the difference between the number of covariance parameters in the full and the reduced model.

In general, likelihood ratio tests provide a valid method for comparing nested models for the covariance [7]. However, in certain settings the likelihood ratio test may no longer be

valid; this depends on the nature of the null hypothesis. This implies that testing a null hypothesis that is "on the boundary of the parameter space," which is equivalent to the null hypothesis that the variance is zero, then the usual conditions for a likelihood ratio test no longer hold.

A likelihood ratio test such that the null hypothesis is that a variance is zero is considered to be on the boundary of the parameter space. Recall that this is due to the fact that variances are bounded from zero to infinity. A consequence of this result is that the usual null distribution for the likelihood ratio test is no longer valid. This is due to the fact that the resulting null distribution for the likelihood ratio test is no longer χ^2 with degrees of freedom $\Delta_{\text{Full}-\text{Reduced}}$, but instead the null is a mixture of chi-squared distributions [7].

Therefore, when testing a null hypothesis that is on the boundary of the parameter space, the usual null distribution will no longer hold, resulting in the selection of a model for the covariance that is too parsimonious, if not approached cautiously.

Usually in longitudinal designs we are often more interested in comparing non-nested models for the covariance. In order to achieve this, an alternate approach is the Akaike Information Criterion (AIC). According to AIC, given a set of competing models for the covariance, one should select the model that will minimize the following:

$$\text{AIC} = -2(\text{maximized log-likelihood}) + 2(\text{number of parameters}) = -2(\hat{\ell} - c),$$

where $\hat{\ell}$ is the maximized REML log-likelihood, and c is the number of covariance parameters.

Another criterion is the Bayesian Information Criterion (BIC). According to the BIC, given a competing set of competing models for the covariance, one should select the model that will minimize

$$\begin{aligned} \text{BIC} &= -2(\text{maximized log likelihood}) + \ln N^*(\text{number of parameters}) \\ &= -2\{\hat{\ell} - \ln(\sqrt{N^*}c)\}, \end{aligned}$$

where N^* is the number of subjects; see [7].

An illustration of using different methods for selecting the covariance structures for ecological studies is presented by Barnett et al. [2]. Ecological data sets often use repeated sampling in a longitudinal design. Thus, choosing the correct covariance structure is an important step in the analysis of such data, as the covariance describes the degree of similarity among the repeated observations. In particular, these authors utilized three methods for choosing the covariance which were: the Akaike information criterion (AIC), the quasi-information criterion (QIC), and the deviance information criterion (DIC).

The objective of this study was to determine the optimal criterion for model selection in ecological data. It was demonstrated that the three information criteria of interest have identical form and shared a common goal, to balance the model fit and complexity. It is important to note that the number of parameters is fixed for AIC, based on the actual

number of parameters. In contrast, for QIC and DIC an effective number of parameters was estimated. In conclusion it was recommended to use DIC because it adjusts for correlated parameters when using an unstructured covariance model.

In conclusion, we are able to see that the objective of correctly selecting our covariance pattern model is an attempt to account for all potential sources of variability that impact the covariance among obtained measurements on the same individual. Recall that some of the covariance pattern models discussed are only appropriate when the repeated measures are collected at equal intervals and cannot handle data subject to irregularity in the measurements obtained between each individual, such as the Toeplitz, autoregressive, and banded covariance pattern models.

Chapter 5

Linear Mixed Effects Model

The underlying idea behind a linear mixed effects model is that some subset of the regression parameters may vary from one individual to another. As a result, it is important to account for the sources of natural heterogeneity in the population. Individuals in the population are assumed to have their own subject-specific mean response trajectories over time. So now, in linear mixed effects models a subset of the regression parameters are now being considered as random. The mean response is then modeled as some combination of population parameters, β, and subject-specific effects that are unique to that particular subject. The population characteristics are regarded as being shared by all individuals within the study. In linear mixed effects models, population characteristics are referred to as fixed effects, while subject-specific effects are referred to as random effects. It is formally known as a linear mixed effects model due to the fact that it will contain a mix of both fixed and random effects.

The linear mixed effects model assumes that the responses will depend on a combination of fixed and random effects. This will further lead to a model for the marginal mean, expressed as $\mathbb{E}(Y_i \mid X_i) = X_i\beta$. In turn, the introduction of random effects induces covariance among the repeated measures and $\text{cov}(Y_i \mid X_i) = \Sigma_i$ will have a distinctive random effects structure. The covariance structures in the previous sections cannot explicitly distinguish between the sources of variability for between-subject and within-subject. Given that we can distinguish between random and fixed effects in linear mixed models, one can then allow the analysis of between-subject and within-subject sources of variation in the longitudinal responses; see [7].

It is also possible to predict how individual response trajectories change over time, in addition to modeling the mean response change. There are some very appealing features of linear mixed effects models, one of which is their incredible flexibility in accommodating any degree of unbalance within the dataset. In addition, they have the ability to account for the covariance among the repeated measures in a relatively parsimonious manner.

The underlying idea of a linear mixed effects models is to inherently allow some subset of the regression parameters to randomly vary from subject to subject. In the following model composition of linear mixed effects, suppose that there are N individuals on whom we have collected n_i repeated measures with the response variable Y_{ij} measured at time t_{ij}. Thus the longitudinal data can be unbalanced over time.

As a result, using vector and matrix notation, the linear mixed effects model can be expressed in the following way:

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{5.1}$$

where

- ✓ β is a $p \times 1$ vector of fixed effects,
- ✓ b_i is a $q \times 1$ vector of random effects,
- ✓ X_i is a $n_i \times p$ matrix of covariates,
- ✓ Z_i is a $n_i \times q$ matrix of covariates, in which $q \leq p$.

The matrix Z_i of covariates is a known design matrix linking the vector b_i of random effects to Y_i. Furthermore for many applications involving longitudinal data, it is possible to see that the columns of Z_i are a subset of the columns of X_i. In particular, the subset of regression parameters from β that vary randomly are determined by the columns of X_i that comprise Z_i. This simply means that any component of β can be allowed to vary, by inducing the columns of X_i comprised of the random effects, in Z_i, the design matrix for the random effects.

Furthermore, the vectors b_i of random effects are assumed to be independent of the covariates, X_i. The vector of random effects is taken to come from a multivariate normal

with mean 0 and covariance matrix G. In turn, the random effects given in Equation (5.1) will have an interpretation in terms of how the subset of regression parameters for the ith subject will deviate from those in the population.

In linear mixed effects models, it is important to note the distinction between the conditional and marginal means of Y_i. As a result, the conditional or subject-specific mean of Y_i given b_i is

$$\mathbb{E}(Y_i \mid b_i) = X_i\beta + Z_i b_i. \tag{5.2}$$

In contrast, the marginal mean of Y_i, when averaged over the distribution of the random effects, b_i, is

$$\begin{aligned}\mathbb{E}(Y_i) = \mu_i &= \mathbb{E}\{\mathbb{E}(Y_i \mid b_i)\} = \mathbb{E}(X_i\beta + Z_i b_i) \\ &= X_i\beta + Z_i\mathbb{E}(b_i) = X_i\beta. \end{aligned} \tag{5.3}$$

Also consider the $n_i \times 1$ vector ϵ_i of errors, which are assumed to be independent of b_i and to follow a multivariate normal distribution with mean 0 and covariance matrix, R_i. Traditionally, it is also assumed that R_i is the diagonal matrix $\sigma^2 I_{n_i}$, such that I_{n_i} denotes the $n_i \times n_i$ identity matrix. This is often referred to as the conditional independence assumption, where given the random effects, b_i, the measurement errors are independently distributed with a common variance of σ^2. This implies that ϵ_{ij} and ϵ_{ik} are uncorrelated, with equal variance [7]. These can be thought of as sampling or measurement errors.

In general it is possible to allow correlation among the ϵ_{ij}s, by assuming one of the covariance pattern models discussed in Chapter 4. However, the introduction of the aforementioned covariance pattern models can raise two potential complications:

(i) The ϵ_{ij}s will no longer have a simple interpretation of sampling error, in turn changing the interpretation of the b_is. This further implies that the ϵ_{ij}s include a component of model misspecification at the individual level.

(ii) In many longitudinal applications, there may be insufficient information to support the estimation of both G and R_i separately. This will result in subtle issues of model identification in the non-diagonal covariance matrix R_i.

Consider the following hypothetical example that is comparing a two-study group of treatment and control (placebo). If the mean response changes in an approximately linear fashion over time, but with the means of the intercepts and slopes depending on their group, then the following linear mixed effects model can be considered:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \,\text{Group}_i + \beta_4 t_{ij} \times \text{Group}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij},$$

where

$$\text{Group}_i = \begin{cases} 1 & \text{if assigned to treatment,} \\ 0 & \text{if assigned to placebo.} \end{cases}$$

In this model the design matrix, X_i, has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ 1 & t_{i2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 0 & 0 \end{pmatrix}.$$

Meanwhile the design matrix for the treatment group can be given as

$$X_i = \begin{pmatrix} 1 & t_{i1} & 1 & t_{i1} \\ 1 & t_{i2} & 1 & t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 1 & t_{in_i} \end{pmatrix}.$$

Additionally we can note that in this example, the design matrix Z_i has the same form for both the control and treatment group, namely

$$Z_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix}.$$

Now we can take into consideration the covariance among the components of Y_i with randomly varying intercepts and slopes. Let $\text{var}(b_{1i}) = g_{11}$, $\text{var}(b_{2i}) = g_{22}$, and $\text{cov}(b_{1i}, b_{2i}) = g_{12}$. If we also assume that $R_i = \text{cov}(\epsilon_i) = \sigma^2 I_{n_i}$, then

$$\begin{aligned}
\text{var}(Y_{ij}) &= \text{var}(X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}) \\
&= \text{var}(Z_{ij}^\top b_i + \epsilon_{ij}) \\
&= \text{var}(b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}) \\
&= \text{var}(b_{1i}) + t_{ij}^2 \text{var}(b_{2i}) + \text{var}(\epsilon_{ij}) + 2\text{cov}(b_{1i}, b_{2i}) \\
&= g_{11} + t_{ij}^2 g_{22} + \sigma^2 + 2 t_{ij} g_{12}.
\end{aligned} \tag{5.4}$$

Similarly, it can be shown that

$$\begin{aligned}
\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}(X_{ij}^\top \beta + Z_{ij}^\top b_i + \epsilon_{ij}, X_{ik}^\top \beta + Z_{ik}^\top b_i + \epsilon_{ik}) \\
&= \text{cov}(Z_{ij}^\top b_i + \epsilon_{ij}, Z_{ik}^\top b_i + \epsilon_{ik}) \\
&= \text{cov}(b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}, b_{1i} + b_{2i} t_{ik} + \epsilon_{ik}) \\
&= \text{var}(b_{1i}) + (t_{ij} + t_{ik}) \text{cov}(b_{1i}, b_{2i}) + t_{ij} t_{ik} \text{var}(b_{2i}) \\
&= g_{11} + (t_{ij} + t_{ik}) g_{12} + t_{ij} t_{ik} g_{22}.
\end{aligned} \tag{5.5}$$

Thus, we get this resulting model for the longitudinal data, in which the covariance matrix can be expressed as a function of time, t_{ij}; see [7].

5.1 Random Effects Covariance Structure

Consider the following linear mixed effects model, where

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

and $R_i = \text{cov}(\epsilon_i)$ describes the covariance among the repeated observations, in particular focusing on the conditional mean response profile of a specific subject. It is the covariance of the ith individual's deviation away from that same subject's mean response profile. In a linear mixed effects model, the conditional mean (5.2) of Y_i given b_i can be distinguished from the marginal mean (5.3) of Y_i. In a similar fashion, we can also distinguish between marginal and conditional covariances. Thus the conditional covariance of Y_i given b_i is

$$\text{cov}(Y_i \mid b_i) = \text{cov}(\epsilon_i) = R_i.$$

Meanwhile, the marginal covariance is given by

$$\begin{aligned}\text{cov}(Y_i) &= \text{cov}(X_i\beta + Z_i b_i + \epsilon_i) = \text{cov}(Z_i b_i) + \text{cov}(\epsilon_i) \\ &= Z_i \text{cov}(b_i) Z_i^\top + R_i = Z_i G Z_i^\top + R_i.\end{aligned}$$

An important consideration that should be made is in the case when $R_i = \text{cov}(\epsilon_i) = \sigma^2 I_{n_i}$. Given that our expression of covariance remains as follows, $\text{cov}(Y_i) = Z_i G Z_i^\top + \sigma^2 I_{n_i}$, it is

possible to note that it is in fact not a diagonal matrix; see [7]. As a result, $\text{cov}(Y_i)$ will in general have non-zero off-diagonal elements, in turn accounting for the correlation that exists among the repeated measures.

In conclusion, it has been observed that the introduction of random effects, b_i, induces correlation among the components of Y_i. Another property of linear mixed effects model is that $\text{cov}(Y_i)$ is described in two components. That is, in linear mixed effects models one can carry out an explicit analysis of between-subject (G) and within-subject (R_i) sources of variation in the response. Lastly, it has been observed that the marginal covariance of Y_i is a function of the times of measurement, as demonstrated in Equations (5.4) and (5.5). It is also important to distinguish that these induced random effects covariance structures do not require a balanced longitudinal design, whereas the other covariance structures seen in Chapter 4 require a balanced design. As a result, this makes the use of linear mixed effects models well suited for dealing with inherently unbalanced data.

Finally, unlike many of the covariance pattern models listed in Chapter 4 that make a strong assumption about the homogeneity of the variance over time, the random effects covariance structure will allow the variance or covariance to fluctuate as a function of the times of measurements.

5.2 Two-Stage Random Effects Formulation

The linear effects model that was presented in Equation (5.1) can be motivated by a two-stage random effects formulation of the model.

5.2.1 Stage 1

In Stage 1 of this two-step formulation, the subjects are assumed to have their own unique individual-specific mean response trajectory. Moreover, the repeated measures on each subject follow regressions model that have the same set of covariates but with distinct regression coefficients for each subject. This can be expressed in the form $Y_i = Z_i\beta_i + \epsilon_i$, where

a) ϵ_i can be thought of as measurement or sampling error, in which $\epsilon_i \sim \mathcal{N}(0, \sigma^2_{n_i})$;

b) the dimension of β_i is q, regardless of the number, n_i, of longitudinal responses;

c) these individual-specific regression coefficients represent the ith subject's "true" regression coefficients;

d) the matrix Z_i represents how a subject's mean response changes over time and/or how the mean response changes with other time-varying covariates.

Consider the following longitudinal study, which assumes that individual-specific trajectories are linear in time. Then the first-stage model can be expressed in the following

way:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}. \tag{5.6}$$

The goal in the first stage is to fit separate linear regression models to the data for each subject. However, it is assumed that these regressions will involve the same set, Z_i, of covariances. This further highlights that given a sufficient number of repeated measures obtained on each individual, it should be possible to estimate β_i and σ^2, while using only the data obtained on that ith individual.

Lastly, a feature of this first-stage formulation is that the matrix of covariates Z_i is said to contain only time-varying covariates (i.e., within-individual covariates), with the exception of a column of 1's for the intercept terms; see [7]. Therefore, any time-invariant (between-individual) covariates (e.g., gender, treatment group, etc.) cannot be included in Z_i because their effects will simply be absorbed by the intercept term. Instead, in the second stage of this formulation, we introduce time-invariant covariates.

5.2.2 Stage 2

In Stage 2, we now make the assumption that the subject-specific effects, β_i, are random. Given that our β_is are considered as random variables, they must then have some probability

distribution, with a mean and covariance. In turn, the population parameters in the second-stage of this formulation are the mean and the covariance of the β_is. More specifically, variation in β_i from one subject to another is modeled as a function of a set of time-invariant covariates; see [7]. Furthermore, the mean of β_i can be expressed as a function of a set of time-invariant covariates, namely A_i, such that

$$\mathbb{E}(\beta_i) = A_i\beta,$$

where A_i is a $q \times p$ matrix. Any remaining variation in β_i that cannot be explained by A_i is given by

$$\text{cov}(\beta_i) = G.$$

Finally, a specification of the model for the mean and covariance of β_i will in turn complete the second stage. It is important to note that β denotes a fixed parameter; meanwhile, β_i denotes a random variable.

Returning to the example used throughout this discussion, where we have a hypothetical two-group study (treatment vs placebo). Once again, if we assume that the subject-specific changes are linear in fashion over time, then the first stage of the model can be given by (5.6). In the second stage, we can now allow the mean of β_i to depend on group, viz.

$$\mathbb{E}(\beta_{1i}) = \beta_1 + \beta_3 \,\text{Group}_i,$$

$$\mathbb{E}(\beta_{2i}) = \beta_2 + \beta_4 \, \text{Group}_i.$$

In this model, β_1 can be interpreted as the mean intercept for the control group, while $\beta_1 + \beta_3$ is the mean intercept for the treatment group. Accordingly, β_3 represents the treatment group difference in the mean intercept. Similarly, β_2 can be interpreted as the mean slope or rate of change in the mean response over time within the placebo (control) group. Meanwhile, $\beta_2 + \beta_4$ represents the mean slope or rate of change within the treatment group. Furthermore, it can be understood that β_4 has interpretation as a treatment group difference rate of change in the mean response over time.

Continuing with the formation of the two-stage random effects structure, the design matrix A_i has the form

$$A_i = \begin{pmatrix} 1 & 0 & \text{Group}_i & 0 \\ 0 & 1 & 0 & \text{Group}_i \end{pmatrix},$$

in which the model for the control group can be expressed as

$$\mathbb{E} \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Similarly, for the treatment group, the model for the mean can be expressed as

$$\mathbb{E}\begin{pmatrix}\beta_{1i}\\ \\ \beta_{2i}\end{pmatrix} = \begin{pmatrix}1 & 0 & 1 & 0\\ \\ 0 & 1 & 0 & 1\end{pmatrix}\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \\ \beta_4\end{pmatrix} = \begin{pmatrix}\beta_1+\beta_3\\ \\ \beta_2+\beta_4\end{pmatrix}.$$

Furthermore, in the two-stage formulation, recall that it is also assumed that the remaining residual variation in β_i, which cannot be explained by the group effect, is given by

$$\text{cov}(\beta_i) = G = \begin{pmatrix}g_{11} & g_{12}\\ \\ g_{21} & g_{22}\end{pmatrix},$$

where $g_{11} = \text{var}(\beta_{1i})$, $g_{22} = \text{var}(\beta_{2i})$ and $g_{12} = g_{21} = \text{cov}(\beta_{1i}, \beta_{2i})$.

In order to yield our desired linear mixed effects model for Y_i, we now combine the two components of this two-stage formulation. However, this formation may encounter some restrictions. We can rewrite the subject-specific effects, β_i as, $\beta_i = A_i\beta + b_i$, where $b_i \sim \mathcal{N}[\mu = 0, \text{cov}(B_i) = G]$. In this case, b_i represents the ith subject's deviation from the population mean response. Combining the two components of the two-stage model formulation results in

$$Y_i = Z_i\beta_i + \epsilon_i = Z_i(A_i\beta + b_i) + \epsilon_i = (Z_iA_i)\beta + Z_ib_i + \epsilon_i = X_i\beta + Z_ib_i + \epsilon_i,$$

where $Z_i A_i = X_i$. Also when averaged over the random effects b_i, one has

$$\mathbb{E}(Y_i) = \mathbb{E}\{(Z_i A_i)\beta + Z_i b_i + \epsilon_i\} = (Z_i A_i)\beta = X_i \beta$$

and also $\text{cov}(Y_i) = Z_i G Z_i^\top + \sigma^2 I_{n_i}$.

In the two-stage formulation, the design matrix Z_i appears in both models for the marginal mean and covariance; see [7]. The model formed here as a result of this two-stage formulation is quite similar to the models introduced at the beginning of this chapter for linear mixed effects. However, there remains one critical difference between the two.

The two-stage model places a constraint on the choice of the design matrix for the fixed effects. More specifically, the two-stage formulation requires that the design matrix for the fixed effects has a particular structure, in which $X_i = Z_i \times A_i$, where A_i contains only the time-invariant (between-subject) covariates, and Z_i only contains the time-varying (within-subject) covariates.

The underlying objective of many longitudinal data applications is focused on the fixed effects, β. These regression parameters have interpretations with respect to changes in the mean response over time, and on their relations to the covariates at hand. In contrast, some longitudinal studies may want to predict subject-specific response profiles. Due to the advantageous feature of a linear mixed effects model, which distinguishes between fixed and random effects, we can also predict individual-specific response trajectories over time.

This implies that in a linear mixed effects model, it is possible to obtain predictions on subject-specific effects, b_i, or of the subject-specific response trajectory, $X_i\beta + Z_i b_i$.

5.3 Fixed Effects vs Random Effects Model

A substantial aspect of the use of regression models is the control for confounding variables. Recall that in longitudinal studies, randomization cannot be used; in turn greater caution must be placed on the measurement and control of important confounding variables. In many standard regression applications, the models allow the assessment of the effects of the covariates in which we have scientific interest in, while statistically adjusting (or controlling) for the confounding variables.

However, it is evident that there are some limitations for this type of adjustment for confounding variables. First, no matter how many confounding variables have been accounted for in the regression model, there will always exist open criticism that some critical confounding variables have been eliminated. Secondly, even if it were possible to include all of the confounding variables that may influence the study, it is inherently difficult or too expensive to measure. In order to overcome these limitations within a longitudinal design, fixed effects models were developed.

The underlying idea behind a fixed effects model is the control of all potential confounding variables that remain stable across the repeated measures obtained and whose effects on the response are assumed to be constant across time. Fixed effects models require two features

of the data, for their application to longitudinal data, namely

1. two or more repeated measures on the response;

2. values of the covariates of main scientific interest must vary over measurement occasions, for at least some subset of the sample.

The first requirement is quite trivial and is met, by definition, by all longitudinal studies. The second requirement suggests that fixed effects models will be best applied in settings where the main covariates of scientific interest are time-varying. In contrast, it can be seen that fixed effects models are not useful in settings where it is also of interest to estimate the effects of time-invariant covariates.

5.3.1 Statistical Formulation of Linear Fixed Effects Model

In order to accommodate for unbalanced data over time, we assume that there are n_i repeated measures of the response on the ith subject and that each Y_{ij} is observed at time t_{ij}. Associated with each response Y_{ij}, there is a $p \times 1$ vector of covariates. The vector of covariates can be further partitioned into two main types of covariates, time-varying covariates (within-subject effects) and time-invariant covariates (between-subject effects).

For the formulation of the linear fixed effects model, let X_{ij} denote the $q \times 1$ vector of time-varying covariates and W_{ij} to denote the $(p-q) \times 1$ vector of time-invariant covariates. For the time-invariant covariates, W_{ij}, we can drop the notation to W_i, given that for $j \in \{1, \ldots, n_i\}$

the same values of the covariates are replicated. This results in the linear fixed effects model to be given by

$$Y_{ij} = X_{ij}^{\top}\beta + W_i\gamma + \alpha_i + \epsilon_{ij},$$

where the α_is are the time-invariant effects (fixed effects) representing stable characteristics of the subjects, that were not otherwise accounted for by the inclusion of time-invariant effects, W_i. Additionally, the model assumes that ϵ_{ij}s are normally distributed with a mean of 0 and a covariance given by σ_{ϵ}^2.

This model formulation may strangely appear to resemble the model formulation for linear mixed effects. However, in the formulation of the fixed effects model, the α_is are considered to be fixed effects; meanwhile in the linear mixed effects models the α_i are considered to be random. It can be advantageous to compare the underlying model assumptions, when comparing between a fixed effects model and a mixed effects model. The fixed effects model makes the following assumptions about the relationships between X_{ij}, W_i, α_i, and ϵ_{ij}:

- X_{ij} is strictly exogenous, i.e., X_{ij} is assumed to be completely independent of the random errors, not only ϵ_{ij} but also for $\epsilon_{ij'}$, for $j \neq j'$. This assumption implies that the current value of Y_{ij} given X_{ij} does not predict the subsequent value of $X_{i,j+1}$.
- The fixed effects model allows the α_is to be correlated with X_{ij}.

It is the second assumption that really creates the distinction between a fixed effects model and a linear mixed effects model. This is due to the fact that in linear mixed effects, it is also

assumed that X_{ij} is strictly exogenous. However, an additional assumption is made, in which the α_is are considered as random rather than fixed effects, and are independent of X_{ij} (and independent of W_i and ϵ_i). In conclusion, the mixed effects model requires the additional assumption that the random subject effects are uncorrelated with X_{ij} for all measurement occasions, where $j \in \{1, \ldots, n_i\}$; see [7]. Additionally, the most notable feature of a fixed effects model is that it only provides estimates of regression parameters for time-varying covariates. Moreover, the effects of time-invariant covariates cannot be estimated in the fixed effects model formulation, due to its perfect collinearity with α_i.

5.3.2 Selecting a Modeling Outcome

In order to appropriately choose the method of modeling that will be applied to a longitudinal study, it is necessary to understand the way in which the data were collected. More specifically, it is prudent to capture the visit timing variability from subject to subject within a study. If there are perfectly repeated measurements, then one of the aforementioned methods of modeling the mean and covariance will be applicable and can provide statistically sound analysis.

However, there are many instances where a study statistician will have to manage missing data and data that are subject to irregularities. This is a very common feature of many longitudinal studies. For example a patient that is experiencing more severe symptoms of illness may require more visits or result in missed visits. This inherently creates an

unbalanced data structure. In the TMS data example provided by Garcia et al. [8], the authors studied approaches to longitudinal data analysis in neurodegenerative diseases: it can be seen that longitudinal studies experience varying participation frequency and the total number of scheduled visits will vary between individuals. This is in particular highlighted among those individuals with "high" or "medium" disease category, who may have limited mobility, resulting in missed or mistimed scheduled visits.

In longitudinal data the observation times of each subject are often assumed to be independent of the outcomes. According to Chen et al. [5], there are many studies in which this assumption is violated, and hence the standard inferential approaches may lead to biased inference. This highlights the importance of the step of examining the data structure before performing any analysis. If for a linear mixed effects model, model misspecification can lead to biased parameter estimates and incorrect inferences, then the addition of the data balance not taken into consideration, and investigated before selecting modeling approach, may also be introducing an even larger amount of bias. Thus, it is meaningful to investigate the balance of the data first, in order to reduce any bias before specifying the correct model for the mean and covariance structure.

Chapter 6

Study of the Influence of Visit Irregularity in a Longitudinal Design

Irregular visits are a common feature of observational longitudinal data, thus resulting in variability in the timings of visits across individuals. Visit irregularities require additional attention as they can be associated with the outcome trajectory and have the potential to lead to biased results, if not addressed appropriately. As mentioned in previous chapters, as important as it is to consider missing data as a part of the modeling strategy, it is also as equally important to assess visit irregularity within a study. Furthermore, investigating the magnitude of irregularity can help determine whether specialized models are needed to handle irregular visits.

The exploration of irregular visits is a new area of research to which Dr. Lokku contributed with a thesis in which he provided visual and numerical measures in order to help quantify the extent of irregularity [11]. The visual and numerical measures of irregularity he proposed are created based on bins constructed across the study duration. The bin widths can vary, and the mean proportions of individuals with 0, 1 and more than 1 visit per bin are plotted. This structured approach allows for a deeper visual understanding of visit irregularity across individuals across the duration of the study.

The ideal scenario is to achieve perfect repeated measures. Given the bins, this would imply that the mean proportion of individuals with 1 visit per bin is 1, while the mean proportions of individuals with 0 or > 1 visits are 0. There is a distinction to be made between experiencing missingness and irregularity within a longitudinal study, which can also be seen in relation to the construction of the bins. Missingness has to do with individuals with 0 visit per bin across the study, while irregularity is reflected in both individuals with 0 and more than 1 visits per bin. This highlights the impact of measuring and quantifying the extent of irregularity before performing any longitudinal analysis to the study.

To numerically assess the extent of visit irregularity, one can use the area under the curve (AUC) with respect to constructed bins of the mean proportions of individuals with 0 visit per bin plotted against the mean proportions of individuals with > 1 visit per bin. In order for AUC to effectively quantify the extent of irregularity, the curve should be increasing as irregularity increases [11]. Furthermore, the AUC must remain invariant to sample size and

follow up length (this refers to the time duration between visit times across individuals). It is important to note that the AUC is not invariant to missingness; in fact, the curve will increase with respect to missingness increasing as well.

In conclusion, these visual and numerical measures lead to a more informed modeling approach prior to considering the longitudinal landscape. Irregularity and missingness are an inherent part of natural variation that occurs in collecting data on repeated measurements, and thus they should be carefully considered in the modeling the outcome. With more and more health data becoming available, it is critical to assess the magnitude of irregularity, in order to perform specialized methods for irregular data, to ensure we have minimized the potential risk of any associated bias as a result of irregularity.

It is quite costly and time-consuming to perform randomized controlled trials, whence the readily increasing availability of observational longitudinal data. In addition, a common trend within health care is the aggregation of the Big Data that is being collected from multiple sources. Electronic health records across multiple places like hospitals, clinics, and laboratories are aggregated among individuals and used as Big Data. This is a very powerful feature for determining any population-level hypothesis of interest.

At the same time, however, there is a loss in overall data quality as a result of how much information is being collected at high frequency [15]. The underlying concern of data quality required for Bid Data modeling is not getting the attention it deserves. An example of this is in the Ontario Drug Benefit Program [9], which studied the trends for information

on opioid prescriptions for those prescribed. In this particular example, it is common to observe individuals that provide unrepresentative information about their usages to professional health care workers. This ultimately results in a set of challenges, including data missingness and irregular observations across each individual. These examples illustrate the importance of addressing the irregularity and missingness within observational longitudinal data; failing to do so could lead to significant bias. Therefore, specialized methods for dealing with irregular data should be considered within these particular settings.

6.1 Study Objective

The objective of this data demonstration is to illustrate measures for quantifying the extent of visit irregularity within an observational longitudinal study. These measures are intended to provide direction for whether specialized methods for dealing with irregular data are necessary or if the data should be treated as repeated measures.

It is necessary to provide first a visual measure of understanding the irregularity captured within the data. This will be completed by visual plotting the measures of irregularity based on bins across the duration of the study. In certain settings where the bin width is not obvious nor stated within the study protocol, we can vary the bin widths and plot the respective mean proportions of individuals with 0, 1 and > 1 visit per bin. The second section of the data investigation entails providing a single numerical measure for quantifying the extent

of irregularity that rises with increasing irregularity (visits with > 1 visit per bin). This is achieved by plotting the mean proportion of individuals with 0 visit per bin against the mean proportion of individuals with more than 1 visit per bin, while using the AUC as a numerical measure quantifying the extent of irregularity.

A common feature in observational longitudinal data is that often it features visit times that vary across individuals. Furthermore, it can potentially lead to timings of the visits and the frequency of visits to be associated with the outcome of the study. This in turn suggests that visit irregularity has the potential to lead to biased results and thus should be considered in the analyses of outcome trajectories [12]. It can be seen that when the visit process has not been accounted for, it could lead to misleading results.

An example of this was found in a study [4] in which the authors estimated the prevalence of pneumonia amongst Kenyan mothers with HIV-1 to be 2.89% when not taking into consideration the visit process. Later it was seen that the same estimate was decreased substantially to 1.48% after taking into consideration the visit process. This underscores the value in studying the visit process or intensity.

Observational data that are collected in aggregation across all electronic medical records of subjects are expected to be susceptible to irregular visit patterns among the individuals. The problem of visit irregularity is comparable to missing data, the key difference being that missing data occur when a scheduled visit is missed and no measurement is recorded, whereas with visit irregularity it is observed as an imbalance of visit patterns across individuals. Visit

irregularity is often most present in settings where there is an absence of a study-wide follow-up schedule. Representing missingness statistically often involves visit times that are fixed by the study design (protocol), and whether the visit occurs is a random variable. With visit irregularity, however, the timings of the visits itself are considered as random.

Studies that feature missingness have also be known to causes biases within the analysis. However, this has led to more developed methods for missing data patterns being recommended [3]. This is often achieved by examining the frequency of subjects with missing values for each variable of interest within the study, in which the severity of missingness can be assessed to determine if methods for missing data patterns are necessary. If some of the data are missing at random (or missing completely at random), then there are approaches available to model longitudinal data where missingness applies.

Techniques such as inverse-probability weighting (estimation of regression coefficients when some regressors are not always observed – James Robins) are often used [14]. As irregularity can lead to biased results, it should be also further explored. Within a lot of practical settings, scheduled visits are intended to be perfect repeated measures, but the timings of scheduled visits vary across all subjects, visits may be missed by the subject or there are unscheduled visits. Irregularity captures this variability observed across subjects and allows for judgment on whether to use appropriate analyzing techniques when dealing irregular data.

It must be noted that it is difficult to determine at which point the data are no longer treated as repeated measures and are examined using different techniques for irregular data. In fact, a study performed by Farzanfar et al. [6] demonstrated how infrequent irregularity is reported or analyzed in practice. Of the 44 qualifying studies, it was shown that 86% of the studies did not account for irregularity. To further emphasize this point, there was only one study in which specialized methods were used when interacting with irregular data. Thus the aim of is to present an intuitive visual measure for irregularity and a measure to quantify the extent of irregularity, which will better allow future researchers to consider checking irregularity in order to select an appropriate statistical approach for the outcome.

6.2 Data: Remifentanil

The pharmacokinetics of the Remifentanil dataset is used here to demonstrate the importance of assessing irregularity before selecting the appropriate statistical approach for the outcome [10]. Intravenous infusion of Remifentanil (a strong analgesic) was applied to a total of 65 subjects at different rates over varying time periods.

Concentration measurements of Remifentanil were taken as the parameter of interest, along with several covariates, thereby creating the Remifentanil data frame with 2107 rows and 12 columns. This data set is part of R package `MEMSS`.

It must be noted that within this data demonstration, there is no protocol available to inform us of the common set of scheduled measurement occasions for all subjects. A

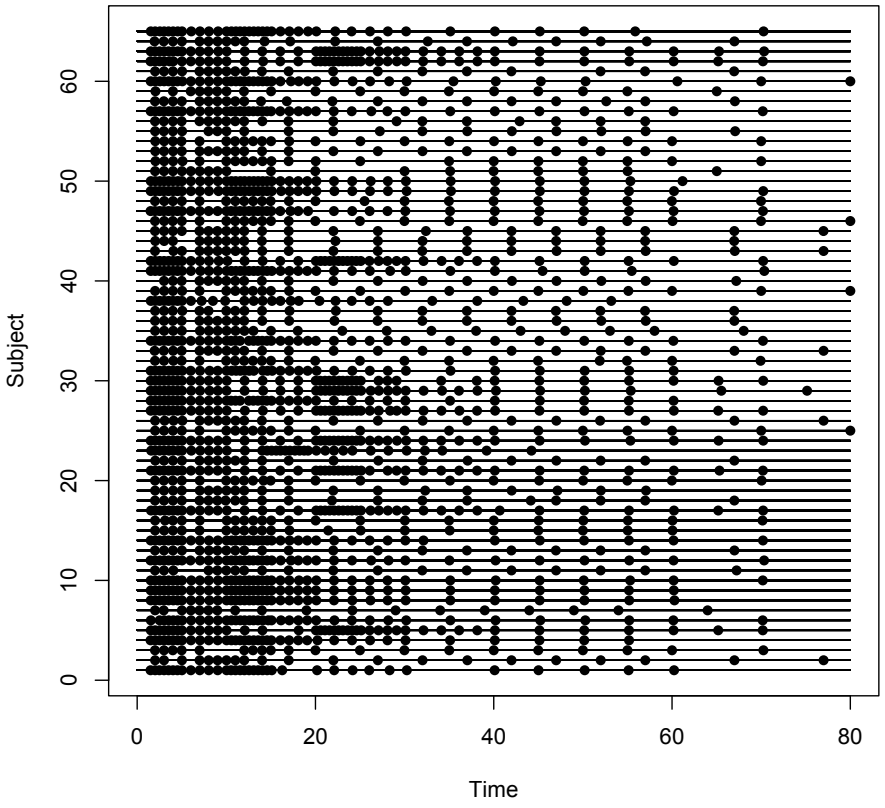


Figure 6.1: Visit timings for the 65 individuals from the Remifentanil Study

clinical research protocol will usually contain information such as data recording requirements, statistical considerations and more.

6.2.1 Illustrating the Measures of Irregularity

The visit timings for all 65 subjects from the beginning of the study are presented in Figure 6.1 above.

From Figure 6.1, we are able to see that we did not achieve a perfectly repeated measures study for Remifentanil. This is because all the observations across each subject are not lining up vertically, as they would if there were perfect repeated measures. Note the x-axis "Time" is referring to time from beginning of infusion in minutes. However, we are still able to capture the visit timings via this measure based on the frequency of visits per subject.

It is evident from the abacos plot that there was variability in visit timings across the 65 subjects. Furthermore, it can also be seen that there were subjects that had 0 visits at other people's scheduled times. This is a starting basis for a visual measure for illustrating the variability across visit times which may signify irregularity.

Recall that this pharmacokinetics of Remifentanil data have no preexisting protocol, which in turn does not allow for bin widths defined according to the prespecified visit times scheduled for each subject. Observing Figure 6.2 below demonstrates a method for visually assessing the number of bins that are constructed. In fact, if we wanted to treat the data like a repeated measures study, then we would be choosing the number of bins to be 25. This is where we can roughly see the curve of one observation per bin reach its maximum.

The second curve just below in Figure 6.2 represents the Area Under the Curve (AUC) plot, whose value is 0.122. Since this particular study has no pre-specified protocol, that results in a high AUC, which in turn demonstrates high irregularity across visits for subjects throughout the study.

By construction, the AUC is bounded between 0 to 0.5. When the AUC = 0 this implies perfectly repeated measures. As a result, we can see that the AUC value of 0.122 signifies that there is some sense of irregularity that exists within this observational dataset. This is aligned with the abacos plot, in which we were also able to visually identify that there is irregularity across visits for subjects within the study. In conclusion, it can be seen from the visual measures that the data could be viewed as repeated measures subject to a degree of irregularity and missingness.

6.2.2 Visit Process Model

In order to appropriately determine a valid modeling approach for the outcome, a more in-depth assessment is needed of the underlying assumptions with respect to the relationship between the outcome (lag concentration) and the visit process. It is essential to assess any potential predictors of visit intensity/frequency, because the standard approach of longitudinal data could be invalid if such predictors exist. A semi-parametric Cox proportional hazards regression model using the Anderson–Gill formulation was thus fitted in order to identify any predictors of visit intensity. The available baseline characteristics that are assessed at the individual-level, include concentration (lag concentration), age, and weight.

The concentration measure in the visit model were lagged by one visit for each individual (i.e., parameter name, lag concentration). This is due to the fact that we can only use the

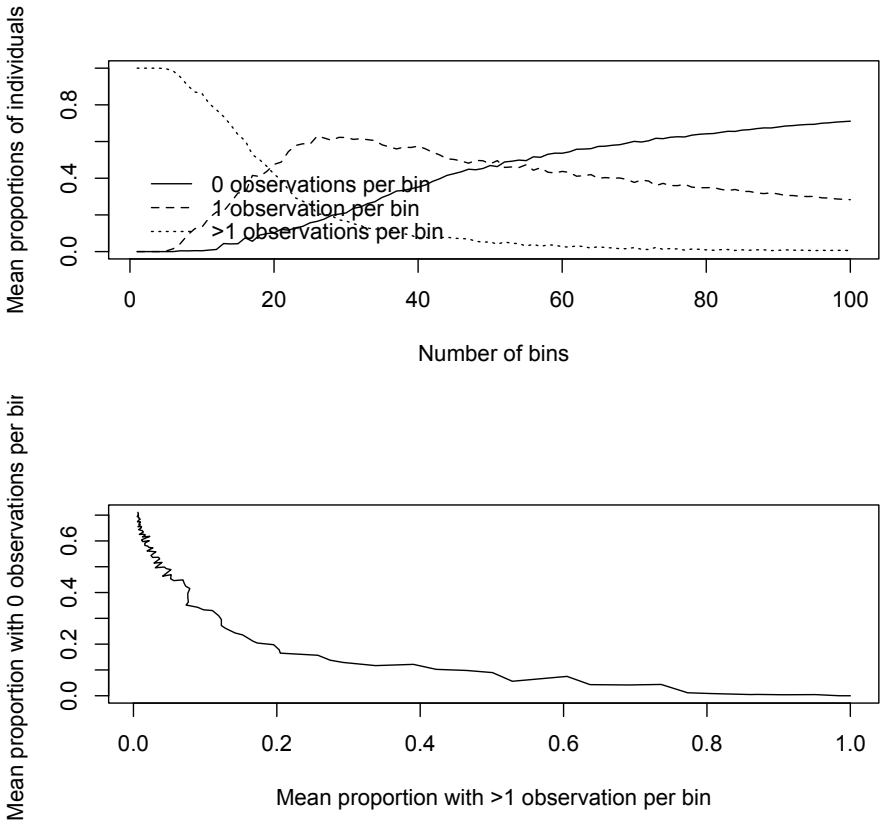


Figure 6.2: The mean proportions of individuals with 0, 1 and >1 visit per bin from the Remifentanil data

value at previous time points to predict current visit intensity. Model selection was done by fitting a regression model with the aforementioned predictors and subsequently determining the predictors that have a statistically significant influence on the model, namely those with p-value < 0.05 in the final model. This analysis was performed using the `coxph` function in R, version 3.1.0, with cluster robust standard errors [16]. Table 6.2 below presents the results from the model which included the aforementioned predictors.

Variable	Time-Varying	Hazard Ratio	Standard Error	p-value
Lag Concentration	Yes	1.0011914	0.0008113	0.0402
Age at Baseline	No	0.9845204	0.0012125	< 0.0010
Weight at Baseline	No	0.9986470	0.0016023	0.3443

Table 6.1: Visit process modeling results including all predictors of interest and the concentration measurement at the previous visit for the Remifentanil Study

Based on Table 6.2, it was suggested that an increase in lag concentration predictor was associated with more frequent visits. As a result, there is some correlation between the outcome and the visit process that is being considered. Additionally, increasing age was demonstrated to be associated with an decrease in the frequency of visits. The following section will describe how to model the study outcome, based on the visit process modeling results in conjunction with the measures of irregularity obtained earlier.

6.2.3 Modeling the Outcome

In order to determine which approach of modeling is most suitable, one must study the relationship between the outcome and the visit process modeled in Section 6.2.2. The analysis of visit intensity was aligned with the visual measures of irregularity and the AUC was also consistent in demonstrating that the data can indeed be viewed as repeated measures subject to both missingness and variable scheduled times among the individuals of interest.

The modeling of the visit process resulted in identifying an association between visit intensity and the concentration (lag) at the previous visit, as well as age, thus suggesting

that the visits between individuals are at best VAR (visiting at random). Likelihood-based methods (e.g., standard mixed effects model) are considered as invalid because the visit intensity model showed an association to its covariates that were not intended to be included in the outcome model.

As a result of identifying predictors of visit intensity, the visit process cannot be VCAR (Visiting completely at random). VCAR would further imply that the visit process and outcomes are completely independent. In the analysis of Remifentnil we have at best VAR (visiting at random), which can imply that they are conditionally independent. The AUC analysis indicated that the sources of irregularity were as a result of missingness and variable visit timings. Thus, we can consider methods such as inverse-intensity weighted generalized estimating equations, under the assumption that the data are VAR.

The data set for Remifentanil has demonstrated signals of it being subject to irregular data. This suggests that we can consider specialized methods for irregular longitudinal data such as the inverse-intensity weighted generalized estimating equations, that can account for predictors of visit intensity, in addition to analyzing the outcome with the presence of missingness.

When selecting a modeling approach to assess the trend in the mean outcome over time, the use of inverse-intensity weighted generalized estimating equations is valid, such that there are no latent predictors of visit intensity that would be correlated to the outcome. This analysis was performed using R, version 3.1.0. The inverse-intensity weighted generalized

Model Approach	Estimate (Time)	Standard Error	p-value
IIW GEE	-0.9204	0.04179	< 0.001
Naive	-0.9907	0.0693	< 0.001

Table 6.2: Modeling results comparing the IIW GEE model to a naive model

estimating equations (IIW GEE) approach was using the `iiwgee` function [12]. In addition, a "naive" model was considered, in which the visit process was not accounted for. Thus, this second model is blind to the irregularity that the data have been subject to. Table 6.2 above compares the results between modeling through both approaches.

In either approach, it can be seen that the mean concentration decreased over time (p-value < 0.001). The inverse-intensity weighted generalized estimating equations approach yielded in a slightly higher coefficient estimate.

In Figure 6.3, we can observe that the behavior, earlier in time, of the naive general estimating equation (green line) overestimates the mean concentration. As the study continues, we can then observe another change, in which the naive general estimating equation (green line) is now underestimating the mean concentration towards the end of the study. This implies that at different time points of the study, the naive fitted method is either overestimating or underestimating the mean concentration. This in turn shows that not taking into consideration the irregularity of the data can lead to biased results when modeling the outcome. In conclusion by taking an additional step, in assessing visually and numerically the extent of irregularity we are able to make a sound decision in modeling the

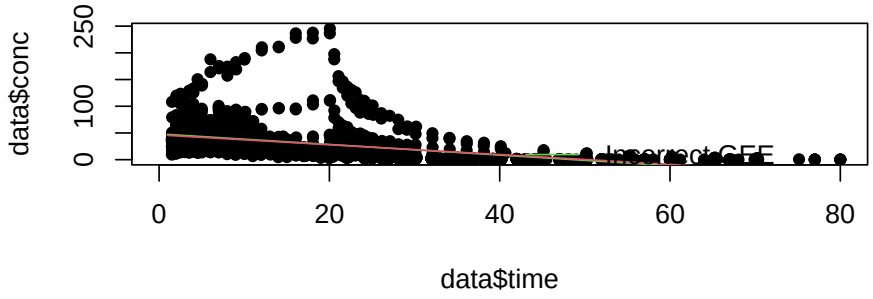


Figure 6.3: The mean proportions of individuals with 0, 1 and >1 visit per bin from the Remifentanil data

outcome approach, thereby allowing us to consider specialized methods that are developed to deal with data irregularity.

6.2.4 Conclusion

This data demonstration was used to illustrate the importance of using the measures of irregularity, exploring visit frequency/intensity in order to select the most valid modeling approach for a longitudinal outcome, which can in turn minimize the potential for biased

results. In this study the visual measures of irregularity in Figure 6.1 and the AUC both indicated that there was some variability in the visit timings between individuals, including missed visits. This in turn suggested to model the visit process of the concentration level at the previous visit, age, and weight at baseline, and indicated that age and lag concentration are associated with the visit intensity.

This finding was important, as it was able to rule out the use of more standard approaches such as standard mixed effects model. Ignoring the visit process and modeling the outcome using the unadjusted generalized estimating equations leads to biased results. Thus by taking these appropriate steps and measures, we can correctly use inverse-intensity weighting to correct for the potential bias caused as a result of the data being subject to irregularities.

Chapter 7

Conclusion

The ultimate objective of this thesis project was to provide an introduction to longitudinal data analysis while developing a sound modeling approach to a longitudinal study. This includes the introduction of complementary measures that can improve the overall modeling approach and help to recognize the choice of an appropriate model for the interdependent covariance and mean, in a longitudinal study.

Furthermore, prior to considering any modeling approach, it is necessary to examine the raw captured data themselves. It is crucial to confirm that the data are balanced, to check whether there are missing observations or if, as commonly observed, visit irregularities occur. The demonstration of these visual and numerical measures will lead to a significantly more informed modeling approach when considering the mean and covariance.

Bringing overall awareness of these highlighted features in this project will lead to a more concrete statistical longitudinal study, which will have less bias and result in more sound contributions to applied and health science studies. These approaches are quite simple to implement prior to beginning any longitudinal analysis and can achieve tremendous improvement in the modeling outcome.

Appendix - R Code for Chapter 6:

Remifentanil Data

RemiFentanil Data Demonstration

Kasra Vakiloroayaei

2023-11-13

```
#Packages utilized to complete demonstartion
#install.packages("lme4")
#install.packages("MEMSS")
#install.packages("IrregLong")
#install.packages("nlme")
#install.packages("survival")
#install.packages("data.table")
#install.packages("geepack")
#install.packages("gee")
```

#Header of the raw data from the MEMMSS Package, analyzing Remifentanil.

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
library(MEMSS)
```

```
## Warning: package 'MEMSS' was built under R version 4.2.3
```

```
##
## Attaching package: 'MEMSS'
```

```
## The following objects are masked from 'package:datasets':
##
##     CO2, Orange, Theoph
```

```
data(Remifentanil)
head(Remifentanil)
```

```
##   ID Subject Time  conc  Rate      Amt   Age  Sex  Ht Wt    BSA     LBM
## 1  1       1 0.00    NA 71.99 107.9850 30.58 Male 171 72 1.8393 56.5075
## 2  1       1 1.50  9.51 71.99  35.9950 30.58 Male 171 72 1.8393 56.5075
## 3  1       1 2.00 11.50 71.99  37.4348 30.58 Male 171 72 1.8393 56.5075
## 4  1       1 2.52 14.10 71.99  35.9950 30.58 Male 171 72 1.8393 56.5075
## 5  1       1 3.02 16.70 71.99  43.9139 30.58 Male 171 72 1.8393 56.5075
## 6  1       1 3.63 17.10 71.99  30.2358 30.58 Male 171 72 1.8393 56.5075
```

```
library(IrregLong)
```

```
## Warning: package 'IrregLong' was built under R version 4.2.3
```

```
library(nlme)
```

```
## Warning: package 'nlme' was built under R version 4.2.3
```

```
##
## Attaching package: 'nlme'

## The following object is masked _by_ '.GlobalEnv':
##
##     Remifentanil

## The following objects are masked from 'package:MEMSS':
##
##     Alfalfa, Assay, BodyWeight, Cefamandole, Dialyzer, Earthquake,
##     ergoStool, Fatigue, Gasoline, Glucose, Glucose2, Gun, IGF,
##     Machines, MathAchieve, Meat, Milk, Muscle, Nitrendipene, Oats,
##     Orthodont, Ovary, Oxide, PBG, Phenobarb, Pixel, Quinidine, Rail,
##     RatPupWeight, Relaxin, Remifentanil, Soybean, Spruce,
##     Tetracycline1, Tetracycline2, Wafer, Wheat, Wheat2

## The following object is masked from 'package:lme4':
##
##     lmList
```

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.2.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.2.3
```

```
library(geepack)
```

```
## Warning: package 'geepack' was built under R version 4.2.3
```

```
library(gee)
```

```
## Warning: package 'gee' was built under R version 4.2.3
```

```
###Format Data
Remifentanil$event <- 1-as.numeric(is.na(Remifentanil$conc))
tmax = 80
data <- Remifentanil
data <- data[data$event==1,]
data$time = data$Time
data$Time = NULL
data <- data[data$time<=tmax,]
data$id <- as.numeric(data$ID)
```

```
###Create lag conc variable
data$conc.lag = c(NA)
for(i in 2:length(data$conc))
{
  data$conc.lag[i] = data$conc[i-1]
}

for(i in 1:length(unique(data$id)))
{
  data$conc.lag[data$id == unique(data$id)[i]][1] = NA
}
```

```
###Create lag time variable
data$time.lag1 = c(NA)
```

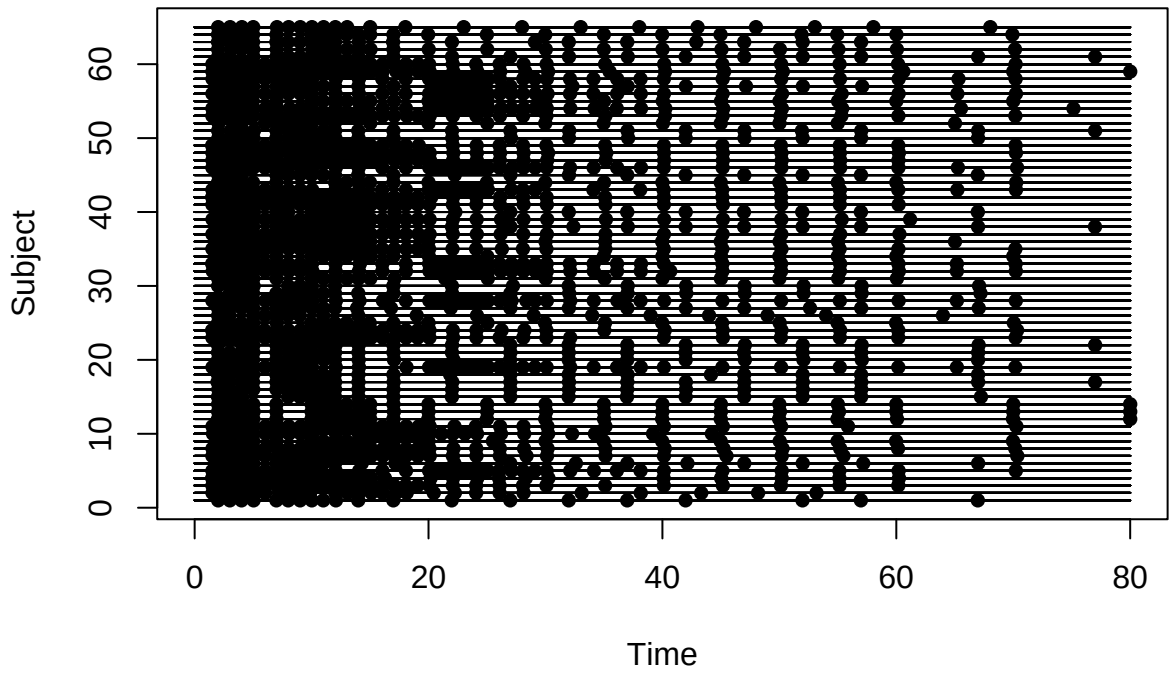
```r
for(i in 2:length(data$time))
{
  data$time.lag1[i] = data$time[i-1]
}

for(i in 1:length(unique(data$id)))
{
  data$time.lag1[data$id == unique(data$id)[i]][1] = NA
}
```

```r
head(data)
```

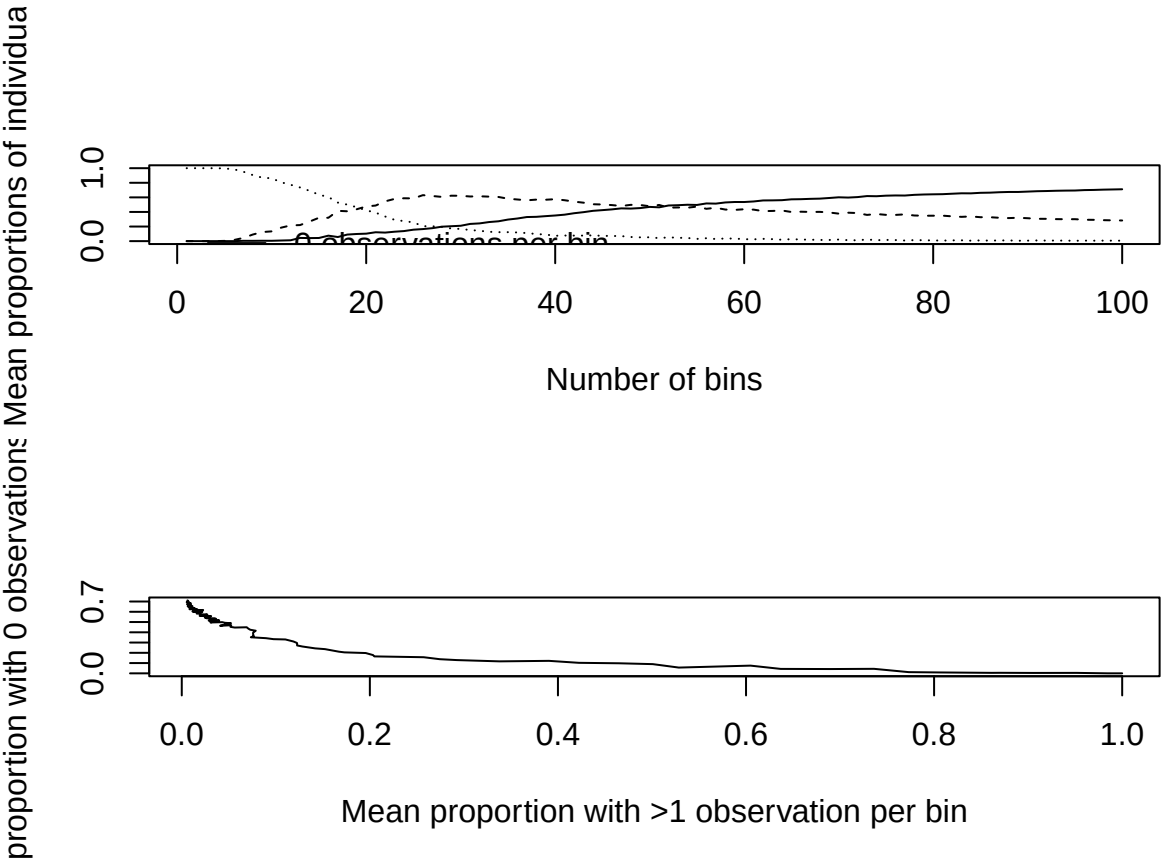
```
##   ID Subject  conc  Rate     Amt   Age  Sex  Ht Wt    BSA     LBM event time id
## 2  1       1  9.51 71.99 35.9950 30.58 Male 171 72 1.8393 56.5075     1 1.50  1
## 3  1       1 11.50 71.99 37.4348 30.58 Male 171 72 1.8393 56.5075     1 2.00  1
## 4  1       1 14.10 71.99 35.9950 30.58 Male 171 72 1.8393 56.5075     1 2.52  1
## 5  1       1 16.70 71.99 43.9139 30.58 Male 171 72 1.8393 56.5075     1 3.02  1
## 6  1       1 17.10 71.99 30.2358 30.58 Male 171 72 1.8393 56.5075     1 3.63  1
## 7  1       1 16.80 71.99 69.8303 30.58 Male 171 72 1.8393 56.5075     1 4.05  1
##   conc.lag time.lag1
## 2       NA        NA
## 3     9.51      1.50
## 4    11.50      2.00
## 5    14.10      2.52
## 6    16.70      3.02
## 7    17.10      3.63
```

#Visual Measures of Irregularity

```r
abacus.plot(n=length(unique(data$id)), time="time", id="id", data=data, tmin=0, tmax=tmax, xlab.abacus =
```




```
counts <-  extent.of.irregularity( data=data, time = "time", id = "id", scheduledtimes = NULL, cutpoints
```




```
counts$counts
```

```
##      [,1] [,2] [,3] [,4]        [,5]       [,6]        [,7]        [,8]
## [1,]    0    0    0    0 0.000000000 0.00000000 0.004395604 0.003846154
## [2,]    0    0    0    0 0.003076923 0.01538462 0.043956044 0.090384615
## [3,]    1    1    1    1 0.996923077 0.98461538 0.951648352 0.905769231
##             [,9]       [,10]       [,11]      [,12]      [,13]      [,14]
## [1,] 0.005128205 0.004615385 0.008391608 0.01153846 0.04378698 0.04175824
## [2,] 0.124786325 0.133846154 0.184615385 0.21538462 0.22011834 0.26813187
## [3,] 0.870085470 0.861538462 0.806993007 0.77307692 0.73609467 0.69010989
##           [,15]     [,16]     [,17]      [,18]      [,19]     [,20]     [,21]
## [1,] 0.04307692 0.0750000 0.0561086 0.08974359 0.09797571 0.1023077 0.1216117
## [2,] 0.32000000 0.3201923 0.4153846 0.40940171 0.43724696 0.4753846 0.4879121
## [3,] 0.63692308 0.6048077 0.5285068 0.50085470 0.46477733 0.4223077 0.3904762
##          [,22]     [,23]     [,24]     [,25]     [,26]     [,27]     [,28]
## [1,] 0.1167832 0.1290970 0.1378205 0.1569231 0.1650888 0.1777778 0.1978022
## [2,] 0.5454545 0.5785953 0.5878205 0.5858462 0.6301775 0.6188034 0.6065934
## [3,] 0.3377622 0.2923077 0.2743590 0.2572308 0.2047337 0.2034188 0.1956044
##          [,29]     [,30]     [,31]     [,32]     [,33]     [,34]     [,35]
## [1,] 0.2042440 0.2107692 0.2357320 0.2432692 0.2615385 0.2719457 0.2953846
## [2,] 0.6228117 0.6220513 0.6119107 0.6149038 0.6102564 0.6054299 0.5819780
## [3,] 0.1729443 0.1671795 0.1523573 0.1418269 0.1282051 0.1226244 0.1226374
##          [,36]     [,37]      [,38]      [,39]      [,40]      [,41]      [,42]
## [1,] 0.3102564 0.3301455 0.33319838 0.34280079 0.35153846 0.36285178 0.38241758
## [2,] 0.5709402 0.5596674 0.56923077 0.56765286 0.57500000 0.56060038 0.54175824
## [3,] 0.1188034 0.1101871 0.09757085 0.08954635 0.07346154 0.07654784 0.07582418
```

```
##           [,43]      [,44]      [,45]      [,46]     [,47]      [,48]
## [1,] 0.39713775 0.41538462 0.42495726 0.43511706 0.4490998 0.44615385
## [2,] 0.52701252 0.50594406 0.50222222 0.49397993 0.4818331 0.49743590
## [3,] 0.07584973 0.07867133 0.07282051 0.07090301 0.0690671 0.05641026
##           [,49]      [,50]      [,51]      [,52]      [,53]      [,54]
## [1,] 0.45306122 0.46923077 0.46334842 0.48875740 0.49259797 0.49886040
## [2,] 0.49513344 0.47815385 0.49592760 0.45887574 0.46153846 0.45811966
## [3,] 0.05180534 0.05261538 0.04072398 0.05236686 0.04586357 0.04301994
##           [,55]      [,56]      [,57]      [,58]      [,59]      [,60]
## [1,] 0.49594406 0.51675824 0.51417004 0.53050398 0.53663625 0.53615385
## [2,] 0.47356643 0.44340659 0.45721997 0.43289125 0.42842243 0.43717949
## [3,] 0.03048951 0.03983516 0.02860999 0.03660477 0.03494133 0.02666667
##           [,61]      [,62]      [,63]      [,64]      [,65]      [,66]
## [1,] 0.54375788 0.55781638 0.55897436 0.56129808 0.57325444 0.57482517
## [2,] 0.43177806 0.41091811 0.41391941 0.41971154 0.39952663 0.40349650
## [3,] 0.02446406 0.03126551 0.02710623 0.01899038 0.02721893 0.02167832
##           [,67]     [,68]      [,69]      [,70]     [,71]      [,72]     [,73]
## [1,] 0.57979334 0.5828054 0.59152731 0.60087912 0.5973998 0.60405983 0.6177028
## [2,] 0.40022962 0.4013575 0.39264214 0.37714286 0.3898158 0.38461538 0.3595364
## [3,] 0.01997704 0.0158371 0.01583055 0.02197802 0.0127844 0.01132479 0.0227608
##           [,74]      [,75]      [,76]       [,77]      [,78]      [,79]
## [1,] 0.61559252 0.62317949 0.62611336 0.624975025 0.63471400 0.63933788
## [2,] 0.37193347 0.36020513 0.35951417 0.366833167 0.35147929 0.34819864
## [3,] 0.01247401 0.01661538 0.01437247 0.008191808 0.01380671 0.01246349
##            [,80]       [,81]      [,82]      [,83]       [,84]      [,85]
## [1,] 0.642115385 0.643684710 0.64990619 0.65579240 0.655128205 0.66244344
## [2,] 0.348461538 0.349097816 0.33958724 0.33290083 0.338644689 0.32796380
## [3,] 0.009423077 0.007217474 0.01050657 0.01130677 0.006227106 0.00959276
##            [,86]       [,87]       [,88]       [,89]       [,90]       [,91]
## [1,] 0.664758497 0.670380195 0.674300699 0.673984443 0.679145299 0.684530854
## [2,] 0.327906977 0.320070734 0.316083916 0.320484010 0.314017094 0.306001691
## [3,] 0.007334526 0.009549072 0.009615385 0.005531547 0.006837607 0.009467456
##            [,92]       [,93]       [,94]       [,95]       [,96]       [,97]
## [1,] 0.686789298 0.690488007 0.692962357 0.693927126 0.699679487 0.701982554
## [2,] 0.305685619 0.301736973 0.300163666 0.300890688 0.293269231 0.291831879
## [3,] 0.007525084 0.007775021 0.006873977 0.005182186 0.007051282 0.006185567
##            [,98]       [,99]    [,100]
## [1,] 0.705808477 0.708469308 0.7109231
## [2,] 0.287284144 0.285003885 0.2830769
## [3,] 0.006907378 0.006526807 0.0060000
```

```
counts$auc
```

```
## [1] 0.1224512
```

#IIW GEE Analysis of Concentration

```
###Model visit process
coxph(Surv(time.lag1,time,event)~conc.lag + Age + Wt+ cluster(id), data = data)
```

```
## Call:
## coxph(formula = Surv(time.lag1, time, event) ~ conc.lag + Age +
##     Wt, data = data, cluster = id)
##
##                coef  exp(coef)   se(coef)  robust se       z      p
## conc.lag  0.0011907  1.0011914  0.0008113  0.0005805   2.051 0.0402
```

```
## Age      -0.0156006  0.9845204  0.0012125  0.0008732 -17.865 <2e-16
## Wt       -0.0013540  0.9986470  0.0016023  0.0014316  -0.946 0.3443
##
## Likelihood ratio test=183.9  on 3 df, p=< 2.2e-16
## n= 1861, number of events= 1861
##    (65 observations deleted due to missingness)
```

```
###Model outcome process
miiwgee <- iiwgee(conc ~ time, Surv(time.lag,time,event)~conc.lag + Age + Wt + cluster(id), id="id",tim
summary(miiwgee$geefit)
```

```
##
## Call:
## geeglm(formula = formulagee, family = family, data = data, weights = useweight,
##     id = iddup, corstr = "independence")
##
##  Coefficients:
##             Estimate  Std.err  Wald Pr(>|W|)
## (Intercept) 46.36395  2.92154 251.8   <2e-16 ***
## time        -0.92044  0.04179 485.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = independence
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)    600.6   202.8
## Number of clusters:   65  Maximum cluster size: 44
```

```
#summary(miiwgee$phfit)
```

```
### Compare to Results Without Weighting (i.e Naive model)
m <- geeglm(conc ~ time , id=id, data=data, corstr = 'exchangeable')
summary(m)
```

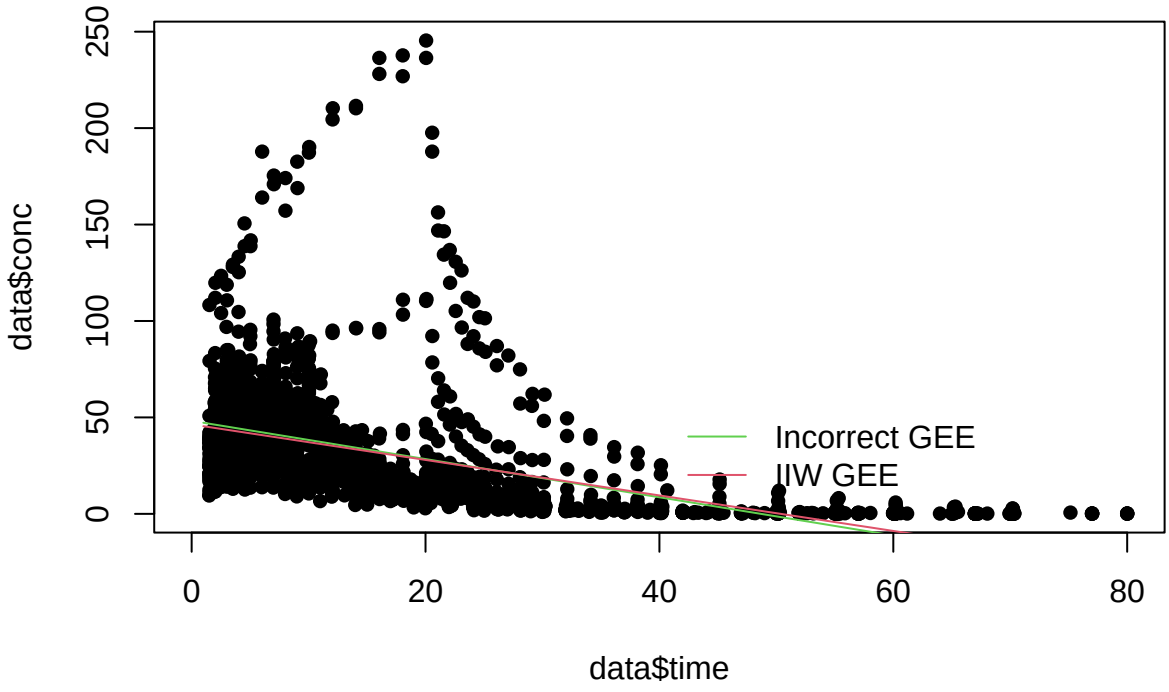
```
##
## Call:
## geeglm(formula = conc ~ time, data = data, id = id, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate Std.err Wald Pr(>|W|)
## (Intercept)  48.2255  3.5631  183   <2e-16 ***
## time         -0.9907  0.0693  205   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)      763     310
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
```

```
## alpha    0.651   0.102
## Number of clusters:   65  Maximum cluster size: 44
```

```
#m <- geeglm(conc ~ time , id=id, data=data)###Independence
```

```
time <- (1:tmax)
unweighted <- cbind(rep(1,tmax),time)%*%m$coefficients
weighted <- cbind(rep(1,tmax),time)%*%miiwgee$geefit$coefficients
```

```
plot(data$time,data$conc,xlim=c(0,tmax),pch=16)
lines(time,unweighted,type="l", col= 3)
lines(time,weighted,col=2)
legend (40,60,legend=c("Incorrect GEE","IIW GEE"),col=3:2,bty="n",lty=1)
```



Bibliography

[1] C. Andersson, A. D. Johnson, E. J. Benjamin, D. Levy, and R. S. Vasan. 70-year legacy of the Framingham Heart Study. *Nature Reviews Cardiology*, 16(11):687–698, 2019.

[2] A. G. Barnett, N. Koper, A. J. Dobson, F. Schmiegelow, and M. Manseau. Using information criteria to select the correct variance-covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, 1(1):15–24, 2010.

[3] D. Bolignano, F. Mattace-Raso, C. Torino, G. D'Arrigo, S. A. ElHafeez, F. Provenzano, C. Zoccali, and G. Tripepi. The quality of reporting in clinical research: The CONSORT and STROBE initiatives. *Aging Clinical and Experimental Research*, 25(1):9–15, 2013.

[4] P. Bůžková, E. R. Brown, and G. C. John-Stewart. Longitudinal data analysis for generalized linear models under participant-driven informative follow-up: An application in maternal health epidemiology. *American Journal of Epidemiology*, 171(2):189–197, 2010.

[5] Y. Chen, J. Ning, and C. Cai. Regression analysis of longitudinal data with irregular and informative observation times. *Biostatistics*, 16(4):727–739, 2015.

[6] D. Farzanfar, A. Abumuamar, J. Kim, E. Sirotich, Y. Wang, and E. Pullenayegum. Longitudinal studies that use data collected as part of usual care risk reporting biased results: A systematic review. *BMC Medical Research Methodology*, 17(1):1–12, 2017.

[7] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, 2012.

[8] T. P. Garcia and K. Marder. Statistical approaches to longitudinal data analysis in neurodegenerative diseases: Huntington's disease as a model. *Current Neurology and Neuroscience Reports*, 17(14), 2017.

[9] T. Gomes, D. Juurlink, I. Dhalla, A. Mailis-Gagnon, J. M. Paterson, and M. Mamdani. Trends in opioid use and dosing among socio-economically disadvantaged patients. *Open Medicine*, 5(1):e13, 2011.

[10] N. Hagenbuch and M. Maechler. Remifentanil: Pharmacokinetics of remifentanil in `nlme`: Linear and nonlinear mixed effects models. `https://rdrr.io/cran/nlme/man/Remifentanil.html`, Last visited on November 24, 2023.

[11] A. Lokku, L. S. Lim, C. S. Birken, and E. M. Pullenayegum. Summarizing the extent of visit irregularity in longitudinal data. *BMC Medical Research Methodology*, 20:1–9, 2020.

[12] E. M. Pullenayegum. Analysis of longitudinal data with irregular observation times. *R package version 0.1. 0*, 2019.

[13] E. M. Pullenayegum and L. S. Lim. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, 25(6):2992–3014, 2016.

[14] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[15] S. R. Sukumar, R. Natarajan, and R. K. Ferrell. Quality of Big Data in health care. *International Journal of Health Care Quality Assurance*, 28(6):621–634, 2015.

[16] T. Therneau. A package for survival analysis in R. `https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf`, Version dated August 13, 2023.

[17] M. van Smeden, T. L. Lash, and R. H. H. Groenwold. Reflection on modern methods: Five myths about measurement error in epidemiological research. *International Journal of Epidemiology*, 49(1):338–347, 2020.

[18] C. C. Y. Wong, A. Caspi, B. Williams, I. W. Craig, R. Houts, A. Ambler, T. E. Moffitt, and J. Mill. A longitudinal study of epigenetic variation in twins. *Epigenetics*, 5(6):516–526, 2010.