Applications of Longitudinal Data Analysis Techniques: Metabolic Syndrome Trial

by

Anita M. Lloyd

A Practicum Submitted to

the Faculty of Graduate Studies

In Partial Fulfilment of the Requirements for the Degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION

Applications of Longitudinal Data Analysis Techniques: Metabolic Syndrome Trial

BY

Anita M. Lloyd

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirement of the degree

MASTER OF SCIENCE

Abstract

Longitudinal data arise when a response is measured at several measurement occasions on the same experimental unit. Special methods of statistical analyses are required to analyze longitudinal data due to the unique properties these data exhibit. In particular, the responses collected for each individual tend to be correlated. Longitudinal data analysis techniques are examined and applied to an experimental study of humans with metabolic syndrome. Initially, exploratory analyses are conducted by creating several graphs to visualize the data and make preliminary inferences. Next, simple methods of analysis for longitudinal data are applied to the data. Finally, more advanced techniques including mixed model methodology are examined.

Acknowledgements

Firstly, I would like to thank my advisor, Dr. Ken Mount, for his patience, support, guidance and advice over the past few years. I would also like to thank the members of my advisory committee, Dr. Brian Macpherson and Dr. James Friel for their invaluable comments.

I wish to thank Dr. Lisa Lix from the Department of Community Health Sciences at the University of Manitoba for providing me with the opportunity to work for her as a research assistant. She has been a positive role model for me and has made a significant contribution to my knowledge.

I would like to thank Tejal Patel for providing me with the dataset used in this practicum. I would also like to express my gratitude to the Faculty of Science for presenting me with a studentship award throughout the first two years of my Master's program.

Finally, I would like to thank my family and friends who have provided me with encouragement and support.

Dedicated to my mother,

Dr. Elli G. Roehm

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Longitudinal data arise when the *same* response variable is measured repeatedly on a subject across multiple occasions. Longitudinal data analysis techniques have applications in a variety of different research fields including agriculture and engineering as well as medical, physical and social sciences.

When analyzing longitudinal data, the main objectives are to (1) examine how the response changes over time, and (2) determine which factors influence the change in response (Singer & Willett, 2003). Statistical models appropriate for data that have a longitudinal structure must be used in order to analyze the data properly.

The goal of this practicum is three-fold: (a) Outline longitudinal data analysis techniques that can be applied in a biological setting, (b) apply techniques to a sample dataset, and (c) obtain results and make conclusions.

1.2 Comparing longitudinal and cross-sectional designs

Sometimes researchers will perform a cross-sectional rather than a longitudinal study to examine the change in response. In a cross-sectional study for example, a response is measured once for two or more cohort groups and the 'change' in response is subsequently analyzed. In contrast, a longitudinal study analyzes change by using the same set of subjects and measures the response repeatedly. While a cross-sectional study is both efficient and economical, only between individual differences in the response can be examined. Cohort effects can also be present. In a longitudinal study, each individual in the study acts as their own control and change over time for each individual can be distinguished from the effects of cohorts (Fitzmaurice, Laird & Ware, 2004; Hedeker & Gibbons, 2006).

1.3 Longitudinal Data Analysis Approaches

There are a variety of different methods to handle the analysis of longitudinal data, such as conducting separate statistical tests at each measurement occasion, univariate analysis of variance (ANOVA), multivariate analysis of variance (MANOVA) and mixed model methods. Computer software programs such as SAS, SPSS, S-Plus, R, and Stata have procedures of varying degrees of sophistication and capabilities that assist with analyses (Littell, Henry & Ammerman, 1998; Fitzmaurice et al., 2004). For this practicum, SAS (Version 9.1) was used for all statistical analyses. SAS and R (Version 2.1.1) were used for the graphs.

1.4 Terminology and Further Details

Longitudinal studies may be experimental or observational in design (Singer & Willett, 2003). As an example of the former, a treatment is randomly assigned to 100 individuals and blood pressure is assessed at 3 measurement occasions. An example of the latter design is measuring weight of infants from birth to 12 years annually.

The subjects (units, participants, or individuals) in a longitudinal study from which a particular outcome of interest can be measured may include anything from humans and animals to equipment components and geographical sites (Crowder & Hand, 1990).

The occasions (or times) of measurement for recording a response should occur at times suitable for collecting precise and sufficient data so that the change in response can be accurately studied (Singer & Willett, 2003). Measurement occasions can occur on a fixed schedule or when a particular event occurs (Ware, 1985). In the former case for example, weight of babies might be measured biennially from birth to age 10 (equal intervals) or at birth, 2 months, 6 months and 1 year (unequal intervals). In the latter case, a response might be recorded when a female subject reaches menarche. Alternatively, a measurement occasion may correspond to a certain experimental or observable variable attaining a specific level (Ware, 1985). For instance, a child's weight is measured when a height of 2 feet is realized. An additional detail to note about time is that the set of occasions at which subjects are measured may be the same for each individual or may vary across individuals (Fitzmaurice et al., 2004).

The response variable measured can be observed in quantitative, categorical or even count form (Fitzmaurice et al., 2004). Examples include measuring cholesterol levels in

humans, classifying the status of a person as 'obese' or 'not obese', and counting the number of weeds in a plot, respectively.

In longitudinal studies, covariates are often recorded at each measurement occasion in addition to the response variable. Covariates can be quantitative and/or qualitative and can be classified as either within-subject or between-subject covariates. A within-subject covariate is a subject attribute that may differ over measurement occasions. Examples of within-subject covariates include a subject's smoking status or age. On the other hand, a between-subject covariate is a variable for a particular subject that remains the same for the duration of the study. Variables such as sex, race and treatment group are all examples of between-subject covariates. Determining how a covariate is related to the pattern of change in the response over time is an objective in longitudinal studies (Ware, 1985).

In longitudinal studies, repeated measurements obtained from each subject form clusters. Clusters exhibit several important properties (Fitzmaurice et al., 2004):

i) Within a cluster, responses tend to be positively correlated (i.e., dependent).

ii) Within a cluster, responses tend to be more similar as compared to responses within a different cluster.

iii) Within a cluster, correlation among responses tends to decrease as time intervals increase.

The correlation among the repeated measurements must be accounted for so that correct estimates can be obtained followed by correct statistical interpretations and conclusions (Fitzmaurice et al., 2004). This topic will be dealt with in subsequent chapters of this practicum.

Fitzmaurice, Laird and Ware (2004) outline three sources of variability that arise in longitudinal data. They are (a) between-subject heterogeneity, (b) within-subject biological variation, and (c) measurement error.

Responses for each individual are influenced by their own specific genetic, environmental, social and behavioral factors which correspondingly lead to natural variation in response. This is known as between-subject heterogeneity and can explain why some individuals have a natural tendency to have response values that are above average while other individuals naturally respond below average. Next, consider that within each individual, biological processes exist (e.g., circadian rhythms) and are changing continuously through time. These biological processes subsequently influence an individual's response. This is called within-subject variation and can explain why responses that are measured closely in time tend to be more similar than responses further apart. Finally, unavoidable measurement error can result in variation. As an example, an investigator can expect that a poor measurement instrument may lead to smaller correlations among repeated measurements versus higher correlations if a precise instrument is used.

In longitudinal studies, especially when the subjects are humans, missing data often arise and can complicate statistical analyses. Missing data may occur when an individual in a study misses a scheduled time for a response to be measured or drops out of the study. Dealing with missing data will be investigated further in Chapter 8.

1.5 General Notation

In this practicum, the following preliminary notation will be adopted following Fitzmaurice, Laird and Ware (2004). Let Y_{ij} be the response for the i^{th} individual at the

j^{th} occasion where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n$. The repeated measurements for the i^{th} subject can subsequently be grouped into a vector

$$\boldsymbol{Y}_i = \begin{pmatrix} \mathrm{Y}_{i1} \\ \mathrm{Y}_{i2} \\ \vdots \\ \mathrm{Y}_{in} \end{pmatrix}.$$

The variance-covariance matrix for the i^{th} subject is defined as

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \mathrm{Var}(\mathrm{Y}_{i1}) & \mathrm{Cov}(\mathrm{Y}_{i1}, \mathrm{Y}_{i2}) & \ldots & \mathrm{Cov}(\mathrm{Y}_{i1}, \mathrm{Y}_{in}) \\ \mathrm{Cov}(\mathrm{Y}_{i2}) & \mathrm{Var}(\mathrm{Y}_{i2}) & \ldots & \mathrm{Cov}(\mathrm{Y}_{i2}, \mathrm{Y}_{in}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathrm{Y}_{in}, \mathrm{Y}_{i1}) & \mathrm{Cov}(\mathrm{Y}_{in}, \mathrm{Y}_{i2}) & \ldots & \mathrm{Var}(\mathrm{Y}_{in}) \end{pmatrix} = \begin{pmatrix} \sigma_1^{\,2} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^{\,2} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^{\,2} \end{pmatrix}$$

while the correlation matrix for the i^{th} subject is defined as

$$\mathrm{Corr}(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \ldots & 1 \end{pmatrix}.$$

Chapter 2

The Longitudinal Data Study

2.1 Background

During the winter semester of my Master of Science program in 2005 at the University of Manitoba, I elected to take the Statistical Consulting (5.729) course. As part of the course curriculum, each student was required to act as a statistical consultant for a Master's student in a different department. In February 2005, I was assigned my client, Tejal Patel, from the Faculty of Human Ecology. Ms. Patel needed assistance with analyzing and interpreting data from her experiment. The longitudinal study and corresponding data that will be used for the application of longitudinal data techniques is described in detail in the following sections.

2.2 Study Design and Details

Thirty-five humans that had been diagnosed as having Type II Diabetes were randomly assigned to one of three groups: control, milled flax, and flaxseed oil groups. The control (CO) group consisted of 10 individuals, the milled flax (MF) group consisted of 13 individuals and the remaining 12 individuals comprised the flaxseed oil (FO) group. Initial background analyses of the subjects' blood revealed that the participants had Metabolic Syndrome and were in a "pre-diabetes" stage. Type II Diabetes was being controlled by exercise and diet (Patel, 2005). Each individual ate a muffin 6 days per week for twelve weeks. The milled flax group ate muffins containing milled flax and the flaxseed oil group consumed muffins containing flaxseed oil. The muffins eaten by the control group contained canola oil. The dose of milled flax and flaxseed oil was equal in the treatment groups. Each muffin consumed contained 7.6 gm of α-linolenic acid (Patel, 2005). The muffins were normally consumed at breakfast or lunch.

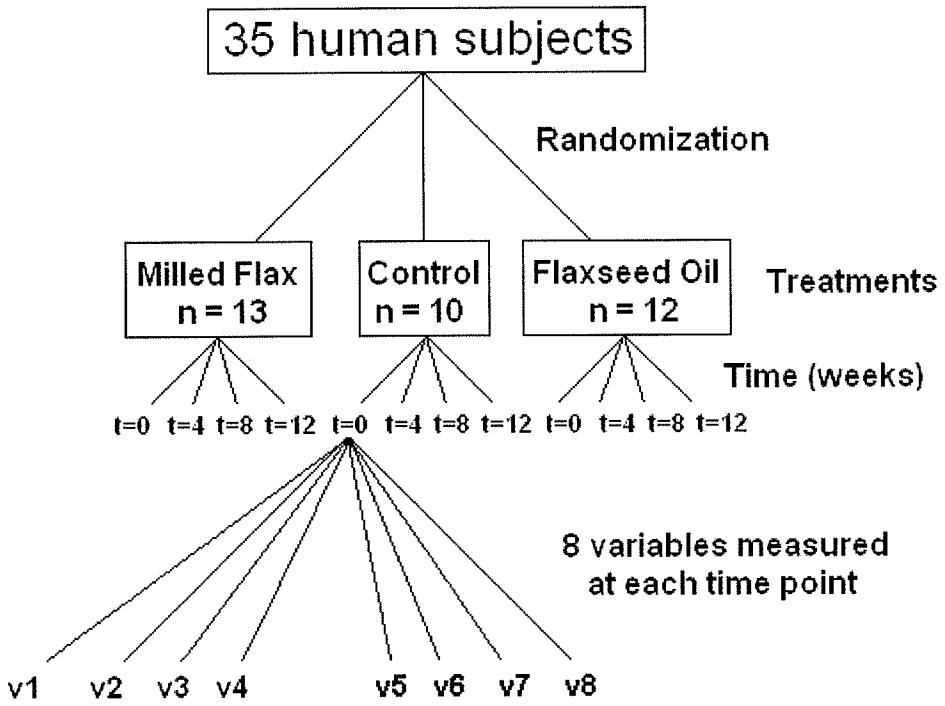

Figure 2.1: Study Randomization

Samples of blood and 24-hour urine were collected at four time levels (i.e., before any treatment was administered (baseline), and at 4, 8 and 12 weeks respectively). The samples were collected in the mornings at the Health Sciences Centre in Winnipeg, Manitoba. Eight different variables were measured for each subject at each measurement occasion. A schematic diagram of the randomization procedure is shown in Figure 2.1.

2.3 Selection of Experimental Units

Flyers were posted at hospitals and universities in Winnipeg, Manitoba and advertisements were placed in the newspaper to recruit volunteers for this experiment. It was determined that Type II Diabetes had previously been diagnosed in the selected subjects between 1 month to 11 years earlier.

2.4 Response Variables

The response variables were:

- Superoxide dismutase (U/mg protein)
- Catalase (U/mg protein)
- Vitamin A (μg/ml)
- Vitamin C (mg/dl)
- Vitamin E (μg/ml)
- Total Antioxidant Capacity (μM/L)
- Hydroperoxide (μM/L)
- Isoprostane (pg/mg Cr)

All of the variables except Isoprostane were measured by analyzing the blood samples. Isoprostane was measured by analyzing the 24-hour urine samples.

2.5 Method of Data Collection

Each variable was carefully analyzed using detailed procedures as outlined in the Appendices by Patel (2005). The methods had been previously published by various authors and standard methods were followed to obtain the response values in this study. To reduce chances of missing data, participants in the study were given reminder calls so that they would not miss their appointments.

2.6 Summary of Data

Below is a brief summary outlining the features of this particular study in terms of longitudinal data analysis terminology.

Subjects: 35 humans with Metabolic Syndrome

Occasions of measurement: Baseline (Week 0), 4, 8 and 12 weeks (i.e., fixed, equal intervals)

Response Variable(s): 8 different variables measured in the blood and urine (See Section 2.4)

Covariate(s): Treatment group (i.e. Between-subject covariate - Control, Milled Flax or Flaxseed Oil) and Time (i.e. Within-subject covariate – Time since baseline)

For this particular study, the main questions of interest are as follows:

(1) How does each response variable change from Week 0 to Week 12?

(2) Can changes in each response variable be predicted based on a subject's membership in the control group, milled flax group or flaxseed oil group?

Chapter 3

Initial Exploration of the Data

3.1 The Data

For the purpose of applying longitudinal data analysis techniques to these data, two response variables were randomly selected to focus on, namely Vitamin C and Vitamin E. To examine all variables in this paper would be too lengthy. Most techniques applied to the select variables would be appropriate for examining the remaining variables. Throughout this practicum, a nominal level of significance of $\alpha = 0.05$ is used.

3.2 Preliminary Glance

Initially, exploratory analyses of the data need to be conducted to obtain a description of what the data look like in order for appropriate models to be fit. By creating simple graphs, an investigator is able to clearly see each individual's response profile. Diggle, Liang and Zeger (1994) outline a few ideas to consider when

constructing graphs. They include (a) graphing raw data instead of data summaries, (b) graphing the data in a manner that illustrates the patterns of change, and (c) presenting data in order to easily identify outliers. Various graphing techniques are explored in the following sections.

3.3 Empirical Growth Plots

Graphs called 'empirical growth plots' as coined by Singer and Willett (2003) are displayed in Figure 3.1, Figure 3.2, and Figure 3.3. Here, Vitamin C data is plotted against the measurement occasions for six randomly chosen individuals from each treatment group. Potential problems that may arise when graphing only a random selection of individuals are that the samples may not be accurate representations of the treatment groups and may not include outlying individuals (Diggle, Liang & Zeger, 1994).

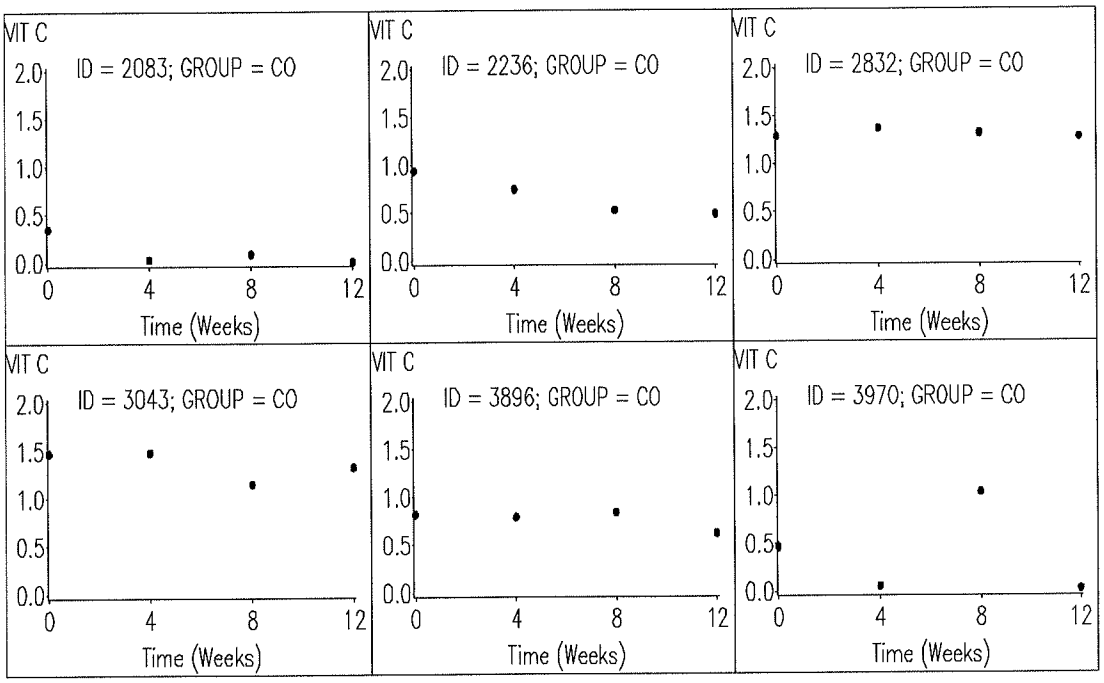

Figure 3.1: Empirical Growth Plots (Control Group)

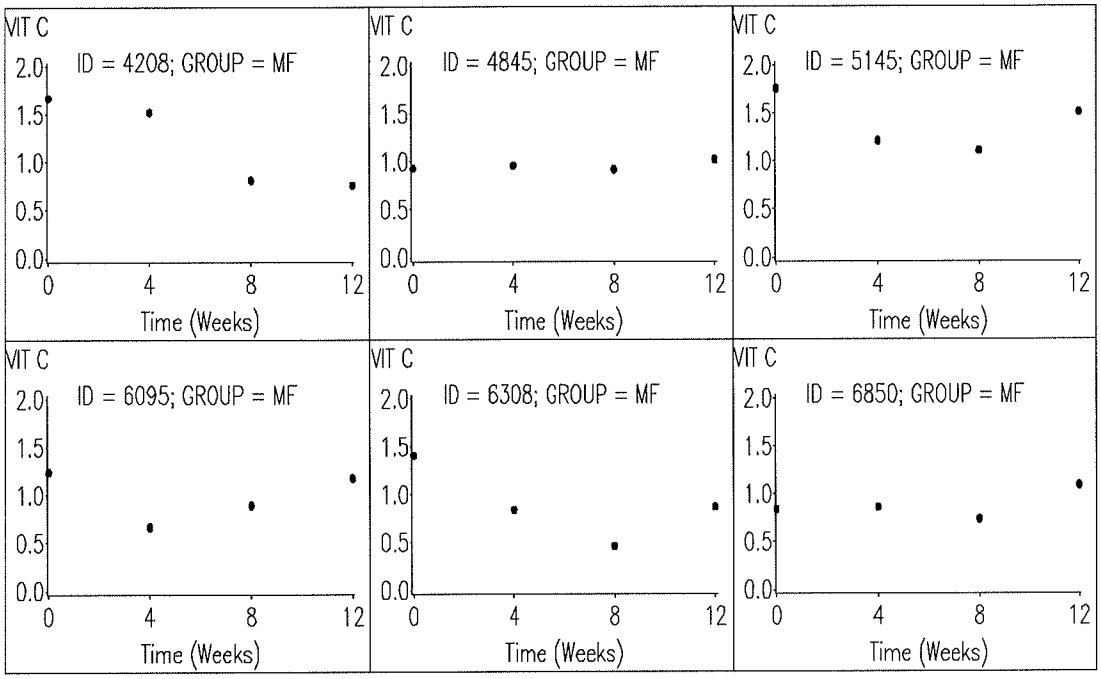

Figure 3.2: Empirical Growth Plots (Milled Flax Group)

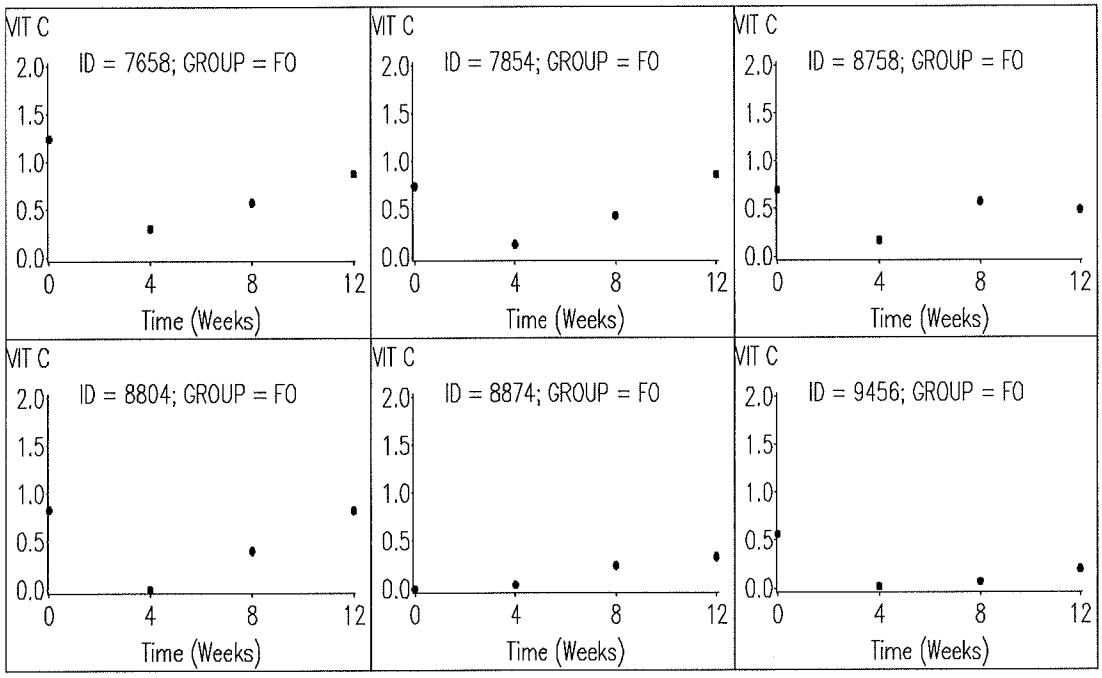

Figure 3.3: Empirical Growth Plots (Flaxseed Oil Group)

In Figure 3.1, Vitamin C levels for the control group seem quite stable across time for all of the individuals sampled with the exception of ID = 3970. This individual has what appears to be an outlying Vitamin C value at the third measurement occasion. For the milled flax group in Figure 3.2, subjects with ID values 5145, 6095, 6308, and 6850 all exhibit a decreasing then increasing pattern in Vitamin C levels over time. Subjects with ID numbers 4845 and 4208 do not follow this pattern and display constant and decreasing Vitamin C levels over time, respectively. Finally, empirical growth plots for six individuals sampled from the flaxseed oil group are shown in Figure 3.3. Here, a decreasing then increasing configuration over time, similar to that observed for select individuals in the milled flax group is exhibited by most individuals (i.e., ID = 7658, 7854, 8758, 8804, and 9456). In contrast, the subject with ID = 8874 demonstrates an increasing Vitamin C level pattern across the measurement occasions.

Once simple empirical growth plots of the data have been examined, lines can be fit to each individual's data using both non-parametric and parametric techniques.

3.4 Empirical Growth Plots - Nonparametric Approach

One method of smoothing is a nonparametric approach that simply connects the data points on empirical growth plots with smooth line segments. Adding the 'smoothed' lines to the plots allows the investigator to further examine individual-specific change in Vitamin C levels. Using the same six randomly chosen individuals for each treatment group from Figure 3.1, Figure 3.2, and Figure 3.3, nonparametric smoothing line segments are plotted on the empirical growth plots. The results are shown in Figure 3.4, Figure 3.5, and Figure 3.6 for the control, milled flax and flaxseed oil groups, respectively.

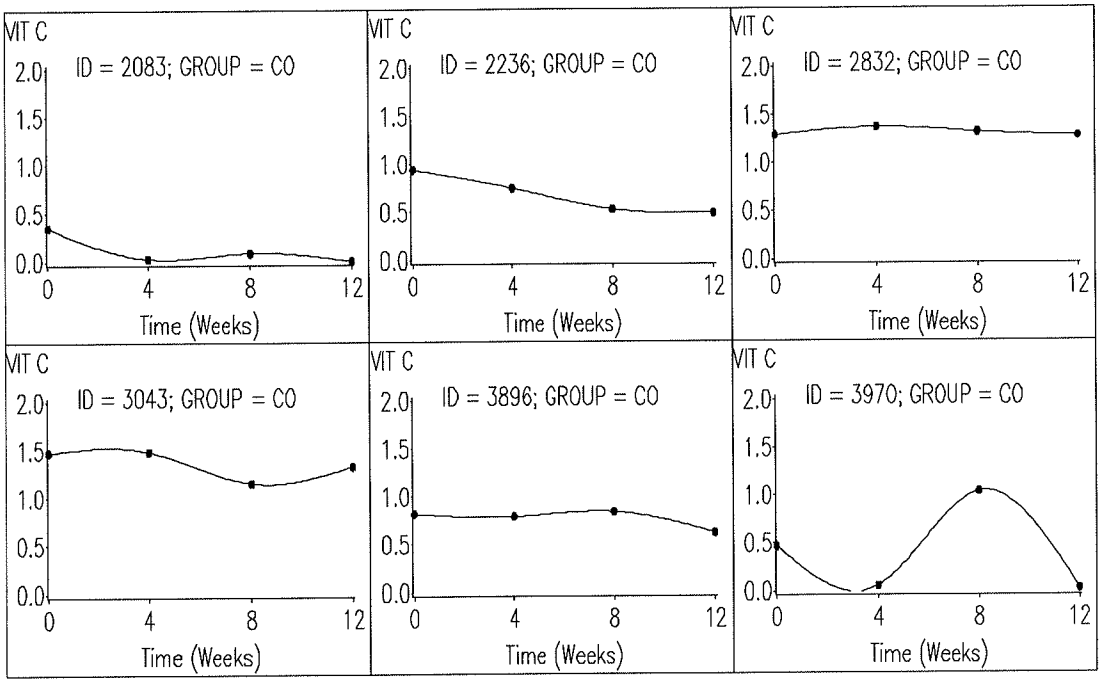

Figure 3.4: Empirical Growth Plots with Nonparametric Smoothing (Control)

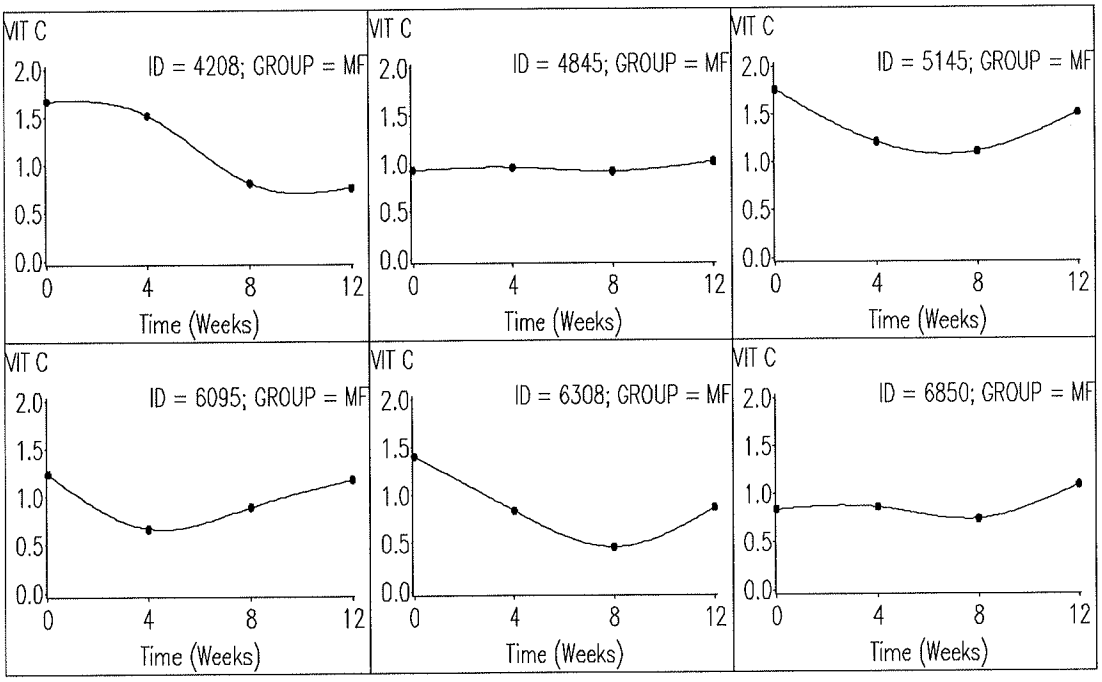

Figure 3.5: Empirical Growth Plots with Nonparametric Smoothing (Milled Flax)

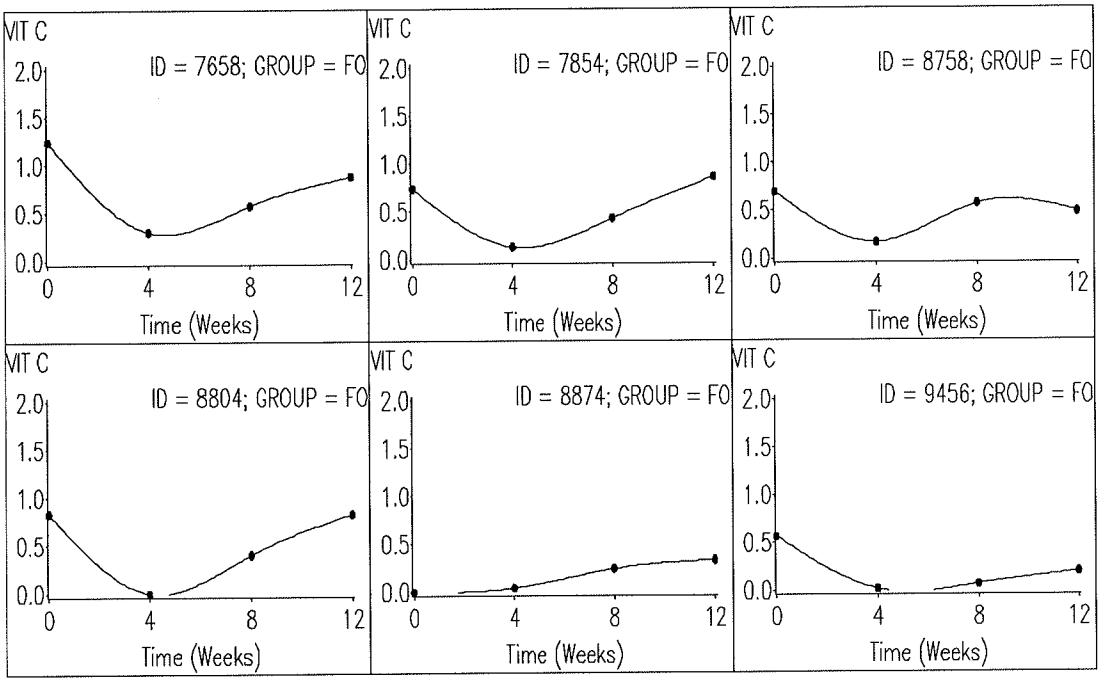

Figure 3.6: Empirical Growth Plots with Nonparametric Smoothing (Flaxseed Oil)

For the control group in Figure 3.4, most individuals have a level response profile with the exception of ID = 3970 which is strikingly different from the other subjects. For the milled flax and flaxseed oil groups in Figure 3.5 and Figure 3.6 respectively, there are many individuals that exhibit a curvilinear pattern of change resembling a 'U' shape. By fitting nonparametric smoothing curves to the sampled individuals, it is easy to see which subjects do not follow the common trends in each group.

3.5 Empirical Growth Plots – Parametric Approach

Another approach to facilitate data summary is to fit parametric curves to each subject's data using ordinary least squares (OLS) regression. According to Singer and Willett (2003), if the same OLS regression model is used for each individual, the researcher can discriminate between subjects with ease. To remain simplistic, a linear regression model is chosen to be fit to each subject's data in the control group while a

quadratic regression model is selected to be fit to each subject's data in the milled flax and flaxseed oil groups. Figure 3.7, Figure 3.8 and Figure 3.9 present the subject-specific OLS regression lines fit to the same randomly chosen individuals from Figure 3.1 to Figure 3.3 in the control, milled flax and flaxseed oil groups, respectively.

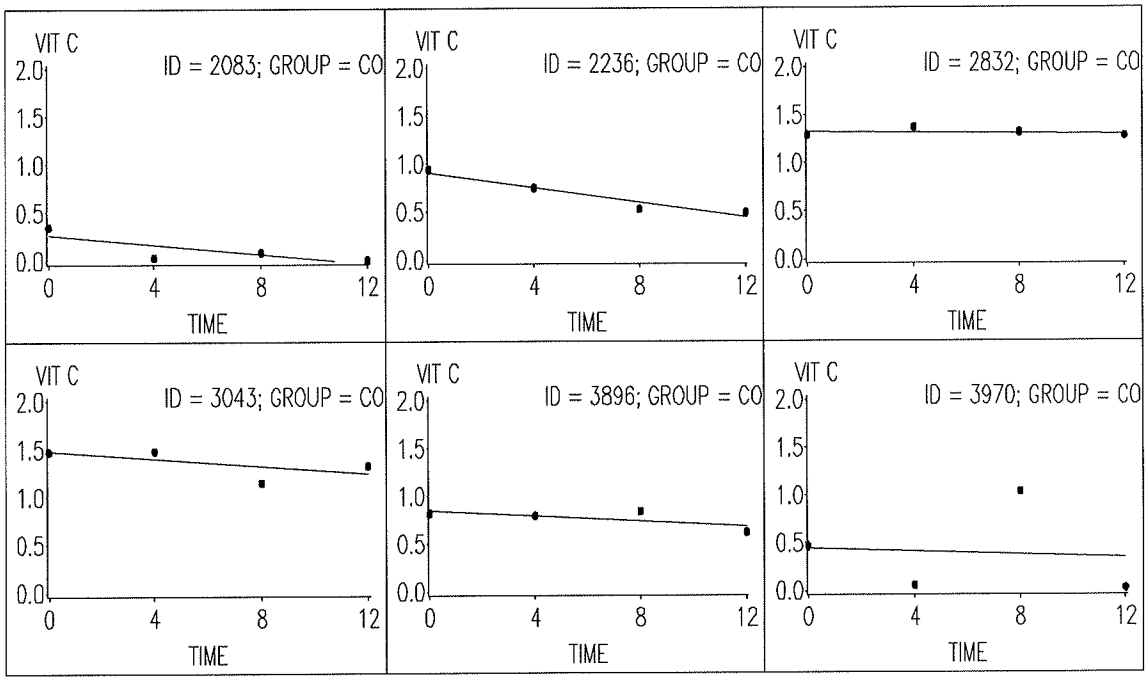

Figure 3.7: Empirical Growth Plots fit with OLS Regression Models (Control)

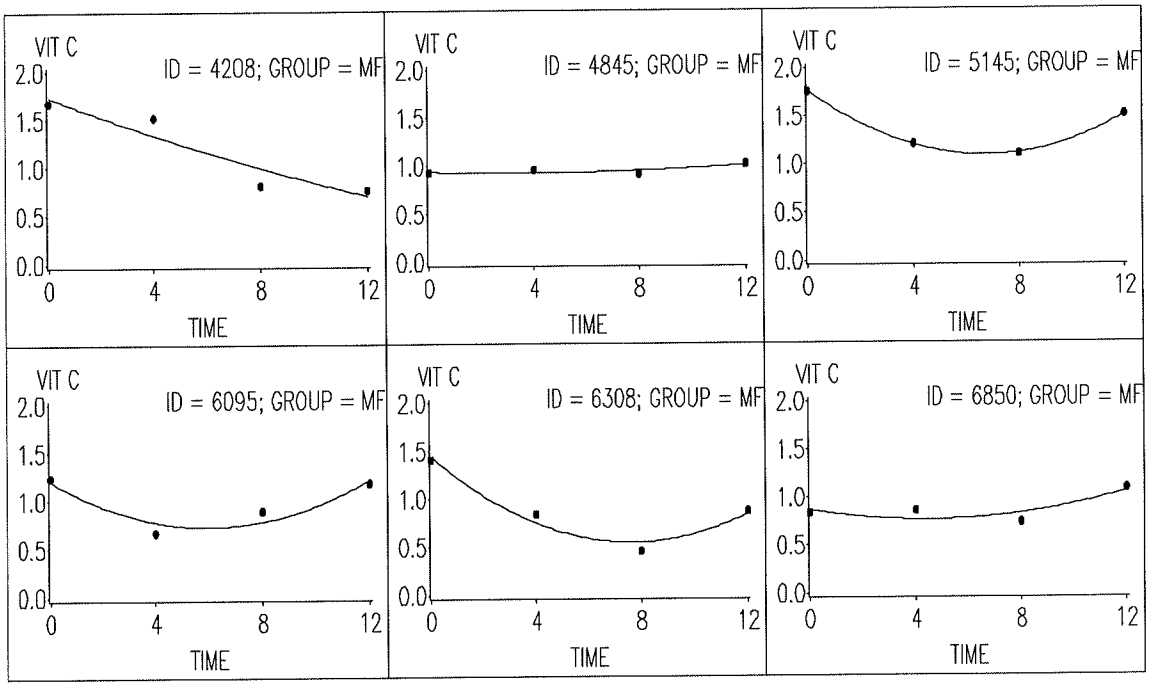

Figure 3.8: Empirical Growth Plots fit with OLS Quadratic Models (Milled Flax)

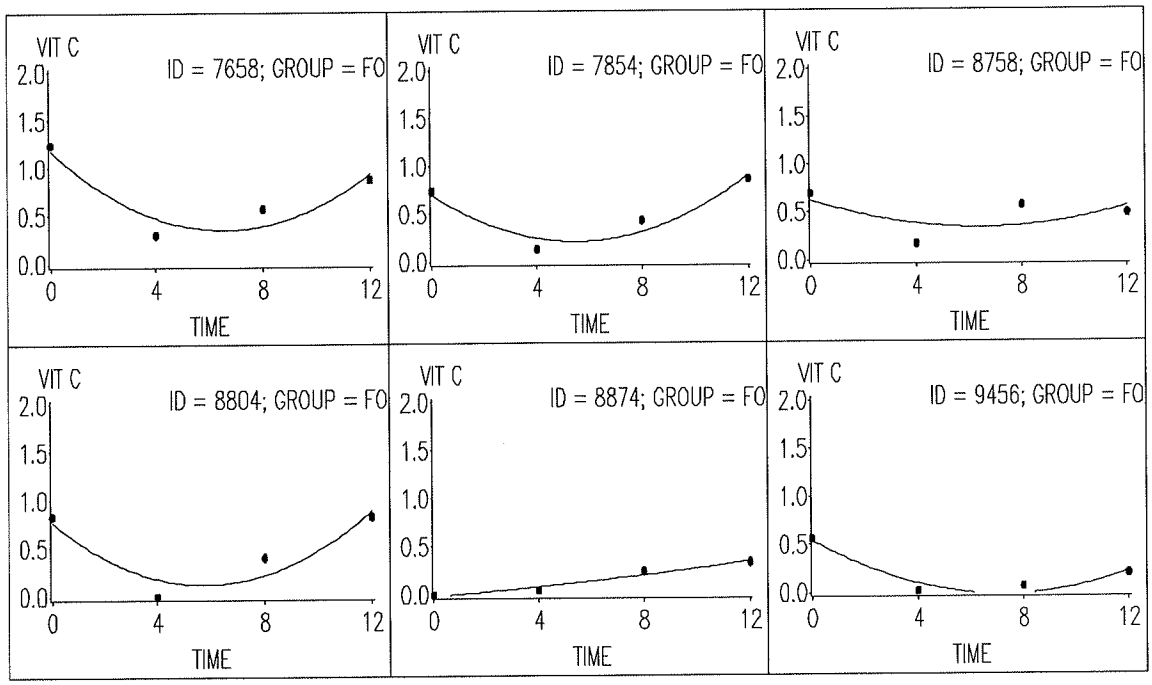

Figure 3.9: Empirical Growth Plots fit with OLS Quadratic Models (Flaxseed Oil)

For the control group in Figure 3.7, it appears that the OLS linear regression models fit the data well for each subject except for the individual with ID = 3970. This poor model fit was expected due to the outlying Vitamin C value at Week 8. The OLS quadratic regression models chosen for the milled flax and flaxseed oil groups also fit most subjects adequately as shown in Figure 3.8 and Figure 3.9, respectively. Summary statistics such as R^2 can aid in assessing goodness of fit for each model but will not be discussed in this practicum (Singer & Willett, 2003).

3.6 Time Plots

An alternative method of data exploration is to construct a graph (or time plot) that plots the response variable against the measurement occasions for each individual on the same graph. The responses for each individual are joined by line segments to distinguish individuals from one another. The time plot can be stratified by treatment group or another covariate to more readily identify response trends, outliers and changes in variability over time (Singer & Willett, 2003; Fitzmaurice et al., 2004). Figure 3.10(a), Figure 3.10(b) and Figure 3.10(c) present the raw Vitamin C data for the control, milled flax and flaxseed oil groups respectively. A researcher may choose yet another descriptive approach which is to plot the mean response at each measurement occasion stratified by treatment group (Fitzmaurice et al., 2004). Figure 3.10(d) plots the mean Vitamin C levels at each time point for each group.

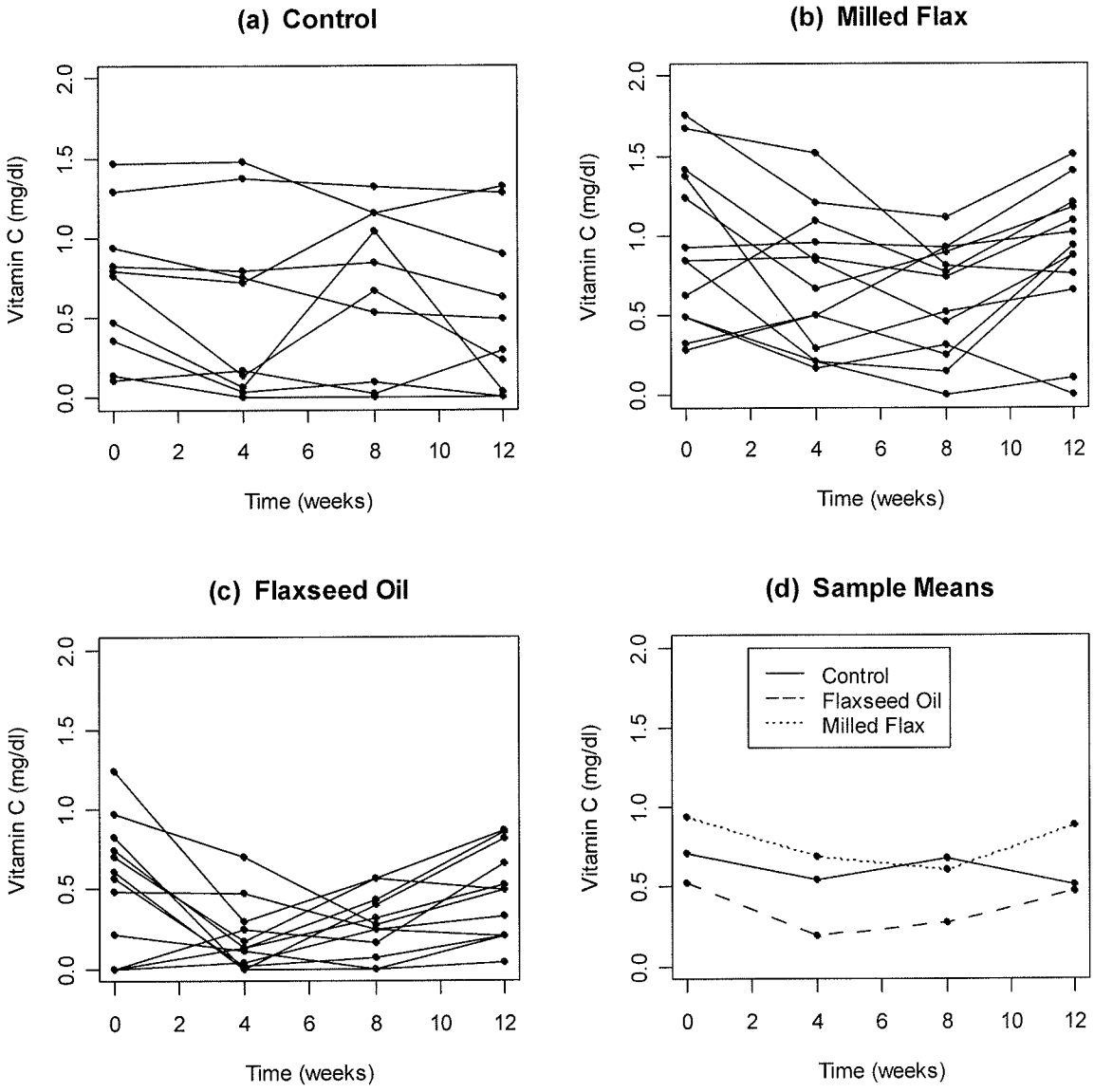

Figure 3.10: Time plots of Vitamin C Data

In Figure 3.10(a), there is no obvious trend in Vitamin C levels for the control group. There is no clear increase or decrease in the Vitamin C levels over time which is not surprising and variability across measurement occasions is approximately constant. Outliers do not appear to be present. Examining Figure 3.10(b) and Figure 3.10(c), it appears that the overall trend for both the milled flax and flaxseed oil groups is for Vitamin C to decrease over time. More specifically, the Vitamin C levels decrease until

Week 8 after which the overall trend changes as Vitamin C levels for both groups increase for the remainder of the study. Again, for both groups, there do not appear to be any obvious outliers. For the milled flax group, variability does not seem to increase or decrease over time while variability for the flaxseed oil group seems to be the greatest at baseline and decreases thereafter until Week 12. For parts (a), (b), and (c) of Figure 3.10, there is evidence of between-subject variability as shown by certain individuals having high Vitamin C levels and others having low Vitamin C levels throughout the study. Furthermore, within-subject variability is exhibited via the jagged lines that join the repeated measures for the individuals.

In Figure 3.10(d), plotting the means of each group at each measurement occasion reveals the overall response trends in a single, tidy diagram. At the baseline measurement, the groups appear to differ. This is an unexpected result as this measurement was taken before the commencement of treatment (i.e., muffin consumption). Mean Vitamin C values for the control group appear to be relatively constant from the beginning to end of the study. For the milled flax group, mean Vitamin C levels decrease until Week 8 and increase thereafter. At Week 12, the mean Vitamin C level for the milled flax group appears similar to what it was at baseline. This same pattern is evident for the flaxseed oil group with the mean Vitamin C level reaching a minimum at Week 4 and subsequently increasing close to the baseline level by the end of the study. By examining each time plot in Figure 3.10, it is apparent that there may be a difference between the control and treatment groups.

Group	Week 0	Week 4	Week 8	Week 12
Control	0.71	0.55	0.68	0.52
	(0.14)	(0.18)	(0.16)	(0.16)
Milled Flax	0.94	0.69	0.60	0.89
	(0.14)	(0.12)	(0.10)	(0.12)
Flaxseed Oil	0.53	0.19	0.27	0.47
	(0.12)	(0.06)	(0.06)	(0.08)

Table 3.1: Mean (Standard Error) of Vitamin C Data

Time plots can be augmented by including standard deviations (or standard errors) bars at each measurement occasion. Rather than doing this which would make the graph unnecessarily cluttered, the mean Vitamin C levels and their corresponding standard errors for each treatment group are presented in Table 3.1.

Chapter 4

Simple Analyses

4.1 Introduction

Researchers with a basic statistical repertoire may initially gravitate towards uncomplicated methods of analysis for longitudinal data. Since well-known procedures such as the simple *t*-test are relatively easy to conduct and interpret results, it is obvious why one might proceed in this manner. The simple methods of analysis to be discussed in this chapter are considered to be more *historical* in nature. The methods outlined include the *t*-test, univariate analysis of variance, and response feature analysis.

4.2 *t*-tests

Once data have been plotted and graphs examined, further analyses of the data can proceed. Everitt (1995) suggests that treatment groups be compared at each time point where a response is measured using either separate *t*-tests or a non-parametric equivalent.

Table 4.1 shows results of *t*-tests for the Vitamin C variable when comparing the control group to each of the milled flax and flaxseed oil groups at each measurement occasion.

Two implicit assumptions when conducting these *t*-tests are enumerated by Crowder and Hand (1990) as follows: (1) The sampled observed responses are independent and normally distributed, and (2) population variances are homogeneous. Table 4.1 shows that when comparing the control group to the milled flax group at each measurement occasion, no *p*-values are significant. This may be an indication of no treatment difference at each time point. When *t*-tests are conducted for comparing the control group to the flaxseed oil group, a significant *p*-value of 0.0338 is reported for Week 8 indicating there may be a difference between the treatments at this measurement occasion.

Time (weeks)		Control	Milled Flax	*t*	*p*-value
0	Mean	0.71	0.94		
	Std Err	0.14	0.14	-1.13	0.2727
4	Mean	0.55	0.69		
	Std Err	0.18	0.12	-0.7	0.4916
8	Mean	0.68	0.60		
	Std Err	0.16	0.10	0.44	0.6671
12	Mean	0.52	0.89		
	Std Err	0.16	0.12	-1.91	0.0696

Time (weeks)		Control	Flaxseed Oil	*t*	*p*-value
0	Mean	0.71	0.53		
	Std Err	0.14	0.12	1.01	0.3260
4	Mean	0.55	0.19		
	Std Err	0.18	0.06	1.92	0.0802
8	Mean	0.68	0.27		
	Std Err	0.16	0.06	2.41	**0.0338**
12	Mean	0.52	0.47		
	Std Err	0.16	0.08	0.24	0.8134

Table 4.1: Results of separate *t*-tests for Vitamin C

One must be cautious if choosing to use *t*-tests to analyze their data. Performing separate *t*-tests in this manner is not particularly useful. In fact, *p*-values produced in Table 4.1 can not even be compared because the *t*-tests are not independent due to the correlation between repeated measurements on the same subject (Crowder & Hand, 1990).

4.3 ANOVA

The one-way ANOVA procedure might be more appropriate for this particular dataset because it can compare three or more treatment groups at the same time rather than performing multiple *t*-tests. The sampled observations must be independent and normally distributed and variances must be homogeneous (Moore, 2003; Crowder & Hand, 1990). For observations that are non-normal, the Kruskal-Wallis test is an alternative procedure (Davis, 2002). Results from separate one-way ANOVA tests conducted at Weeks 0, 4, 8 and 12 for the Vitamin C variable are shown in Table 4.2.

Time (weeks)	*F*-statistic	*p*-value
0	2.56	0.0931
4	4.79	**0.0152**
8	4.17	**0.0245**
12	3.81	**0.0328**

Table 4.2: One-way ANOVA for Vitamin C

Based on the one-way ANOVA results, significant *p*-values of 0.0152, 0.0245, and 0.0328 at Weeks 4, 8, and 12 respectively, indicate that there may be a difference in mean Vitamin C levels between the control, milled flax, and flaxseed oil groups. Again, an investigator should be cautious as these results are based on repeated tests for the same subject.

4.4 Response Feature Analysis

Also known as *Summary Measure Analysis* and perhaps not familiar to every reader, this approach uses the observations from each individual to create a set of new numbers called *summary measures*. The summary measures represent a specific dimension of each individual's response trend (Fitzmaurice et al., 2004). There are a variety of different summary measures that can be adopted. These include the overall mean (i.e. the average of the repeated measurements for each individual), area under the curve (i.e. the area under each individual's response trajectory) and the maximum (or minimum) response (i.e. the maximum or minimum value of the repeated measurements for each individual) (Everitt, 1995; Fitzmaurice et al., 2004). Upon calculating the selected summary measures, simple univariate statistical methods (e.g., *t*-test, ANOVA, or non-parametric method) can be used to test the difference between groups. AUC, minimum response and overall mean are three different summary measures calculated for the Vitamin C data. The ANOVA procedure is used to produce *F*-statistics and *p*-values with results presented in Table 4.3.

Summary Measure	***F*-statistic**	***p*-value**
AUC	4.67	**0.0166**
Minimum Response	4.68	**0.0165**
Overall Mean	4.49	**0.0191**

Table 4.3: Response Feature Analysis Results

According to the response feature analysis approach, all three summary measures chosen yield significant *p*-values indicating that there may be a difference between the groups in terms of the AUC, minimum response and overall mean values.

There are several advantages and disadvantages to the response feature analysis approach. One advantage is that researchers with a basic knowledge of statistics are able to grasp this technique easily as introductory methods of analysis are applied to the summary measures. Secondly, the response feature analysis approach eliminates the correlation issues present in longitudinal data because the summary measures are independent. Finally, it works well with small sample sizes, especially in cases where correlation among repeated measurements is difficult to estimate. A disadvantage of response feature analysis is that the summary measure is a single feature of the individual's response profile; information is thus lost. Also, different individuals may have the same summary measure even though their individual response trends over time are not equivalent. Thirdly, only time-invariant covariates (e.g., treatment group) can be related to the summary measures. Lastly, response feature analysis becomes more complicated when there is missing data, or responses are measured at irregular time occasions because variances are not constant making methods of analysis described in this section not valid (Everitt, 1995; Crowder & Hand, 1994; Fitzmaurice et al., 2004).

It should be noted that simple methods of analysis such as the methods discussed in Chapter 4 are often omitted. These historical methods were presented so that the reader could see the natural progression of longitudinal data analysis over the years.

Chapter 5

General Linear Regression Model and Estimation

5.1 Introduction

In this chapter, a general linear model is introduced that will be used in modified forms in subsequent chapters. Two simple examples will be shown to illustrate the framework of the model. Subsequently, maximum likelihood and restricted maximum likelihood estimation will be examined to show how unknown parameters can be estimated in the various statistical models. Also, test statistics that will be used for statistical inference are introduced.

5.2 Notation

The next section examines the framework of linear models appropriate for longitudinal data. According to Fitzmaurice, Laird and Ware (2004), one model for

longitudinal data is an approximate multivariate normal distribution with mean $\boldsymbol{\mu}_i$ (i.e., $\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$) and covariance $\boldsymbol{\Sigma}_i$ (i.e., $\mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i$). More specifically, the repeated measurements for each individual are assumed to be multivariate normal, although this assumption is not absolutely required. The dependence among the repeated measures for each individual is accounted for by the off-diagonal values in $\boldsymbol{\Sigma}_i$. It will be shown in Chapter 6 how both ANOVA-type and regression-type models fit within the framework of linear models outlined in this section.

At this point, notation first introduced in Section 1.5 will be re-visited. Recall the vector

$$\boldsymbol{Y}_i = \begin{pmatrix} \mathrm{Y}_{i1} \\ \mathrm{Y}_{i2} \\ \vdots \\ \mathrm{Y}_{in_i} \end{pmatrix}$$

that defines the collection of n_i repeated measures for the i^{th} individual. Note that this vector has been slightly modified by using n_i to accommodate longitudinal studies that have missing data. The subscript i in n_i allows for each subject to have a different number of repeated measurements. For example, the Vitamin C dataset has no missing values therefore the vector $\boldsymbol{Y}_i$ defined in Section 1.5 is sufficient. On the other hand, variables such as Vitamin E and Vitamin A have incomplete datasets, making the vector $\boldsymbol{Y}_i$ defined in the current section appropriate. It is important to recall that the timing of measurements may not be the same for each individual. Let t_{ij} denote the j^{th} time at which a response is measured for individual i. This notation allows for the possibility of mistimed measurements. In this particular study, measurements were taken at the same occasions $\{t_{i1} = 0, t_{i2} = 4, t_{i3} = 8, t_{i4} = 12\}$ for all subjects. Thus, the subscript i in t_{ij} is not

strictly necessary. It should be noted that an important property of $\boldsymbol{Y}_i$ is that independence between the vectors is assumed (Fitzmaurice et al., 2004).

Next, the vector of covariates, X_{ij}, for the i^{th} subject at the j^{th} measurement occasion is denoted by

$$\boldsymbol{X}_{ij} = \begin{pmatrix} \mathrm{X}_{ij1} \\ \mathrm{X}_{ij2} \\ \vdots \\ \mathrm{X}_{ijp} \end{pmatrix},$$

where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, n_i$ and p is the total number of distinct covariates. The covariates can include both within-subject and between-subject covariates. The covariates for each individual i can also be set up in the matrix

$$\boldsymbol{X}_i = \begin{pmatrix} \mathrm{X}_{i11} & \mathrm{X}_{i12} & \cdots & \mathrm{X}_{i1p} \\ \mathrm{X}_{i21} & \mathrm{X}_{i22} & \cdots & \mathrm{X}_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{X}_{in_i1} & \mathrm{X}_{in_i2} & \cdots & \mathrm{X}_{in_ip} \end{pmatrix}.$$

Each row represents the p covariates at a measurement occasion while each column represents a unique covariate.

5.3 General Linear Model

Let us examine a linear regression model that regresses the repeated measurements on a set of covariates. The model equation is represented in matrix notation as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i \tag{5.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown coefficients, $\boldsymbol{e}_i$ is a vector of random errors, and $\boldsymbol{Y}_i$ and $\boldsymbol{X}_i$ are defined in Section 5.2. The expectation and covariance are $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i$, respectively.

Model (5.1) can be expanded as shown by model (5.2) below:

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + e_{ij} \tag{5.2}$$

where $j = 1, 2, \ldots, n_i$, and random errors are represented by e_{ij} with zero mean.

5.3.1 Example: Modeling Treatment Effects as Additive Constants

Next, an example of the general linear model framework is discussed in terms of the current study. Here, $n = 4$ for all i and j, and $t_1 = 0$, $t_2 = 4$, $t_3 = 8$ and $t_4 = 12$. The covariates in this study are: (1) the within-subject covariate of time since baseline, and (2) the between-subject covariate of treatment group. The matrix $\boldsymbol{X}_i$ for this study is denoted by

$$\boldsymbol{X}_i = \begin{pmatrix} X_{i11} & X_{i12} & X_{i13} \\ X_{i21} & X_{i22} & X_{i23} \\ X_{i31} & X_{i32} & X_{i33} \\ X_{i41} & X_{i42} & X_{i43} \end{pmatrix}.$$

The first column of $\boldsymbol{X}_i$ corresponds to the time elapsed since baseline (i.e., $X_{ij1} = t_j$) for the i^{th} individual at each occasion. Given that there are three treatment groups in this study, the second and third columns of $\boldsymbol{X}_i$ denote the between-subject covariates of treatment group and are necessary to distinguish between the groups. In other words,

$$X_{ij2} = \begin{cases} 1 & \textit{if the } i^{th} \textit{ individual belongs to the milled flax group} \\ 0 & \textit{otherwise} \end{cases}$$

and $$X_{ij3} = \begin{cases} 1 & \textit{if the } i^{th} \textit{ individual belongs to the flaxseed oil group} \\ 0 & \textit{otherwise} \end{cases}.$$

It follows that a series of regression equations for each measurement occasion can be written as follows:

$$\begin{aligned} Y_{i1} &= \beta_1 X_{i11} + \beta_2 X_{i12} + \beta_3 X_{i13} + e_{i1} \\ Y_{i2} &= \beta_1 X_{i21} + \beta_2 X_{i22} + \beta_3 X_{i23} + e_{i2} \\ Y_{i3} &= \beta_1 X_{i31} + \beta_2 X_{i32} + \beta_3 X_{i33} + e_{i3} \\ Y_{i4} &= \beta_1 X_{i41} + \beta_2 X_{i42} + \beta 3 X_{i43} + e_{i4} \end{aligned} \tag{5.3}$$

An intercept term β_1 can be added by setting $X_{ij1} = 1$ for all i and j. Model (5.3) can be alternatively expressed by:

$$\begin{aligned} Y_{i1} &= \beta_1 + \beta_2 X_{i12} + \beta_3 X_{i13} + \beta_4 X_{i14} + e_{i1} \\ Y_{i2} &= \beta_1 + \beta_2 X_{i22} + \beta_3 X_{i23} + \beta_4 X_{i24} + e_{i2} \\ Y_{i3} &= \beta_1 + \beta_2 X_{i32} + \beta_3 X_{i33} + \beta_4 X_{i34} + e_{i3} \\ Y_{i4} &= \beta_1 + \beta_2 X_{i42} + \beta_3 X_{i43} + \beta_4 X_{i44} + e_{i4} \end{aligned} \tag{5.4}$$

The $\boldsymbol{X}_i$ matrix is modified to reflect the addition of the intercept. X_{ij1} depicts the intercept term, X_{ij2} now represents the time elapsed since baseline, and X_{ij3} and X_{ij4} now represent the group effects.

The expectation of Y_{ij} given the set of covariates is

$$E(Y_{ij} \mid X_{ij}) = \beta_1 + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4}.$$

The conditional expectations for each group are as follows:

Control group $\quad E(Y_{ij} \mid X_{ij}) = \beta_1 + \beta_2 t_j$

Milled Flax group $\quad E(Y_{ij} \mid X_{ij}) = (\beta_1 + \beta_3) + \beta_2 t_j$

Flaxseed Oil group $\quad E(Y_{ij} \mid X_{ij}) = (\beta_1 + \beta_4) + \beta_2 t_j$

From the expectations above, the slopes are the same for each group. The treatment groups differ from the control group by the addition of β_3 and β_4 to the intercept term for the milled flax and flaxseed oil groups, respectively.

5.3.2 Example: Modeling Treatment Effects as Slopes

Another simple example of the general linear model framework in terms of the current study is presented next. Following Fitzmaurice, Laird and Ware (2004), a model is fit for the Vitamin C data in which the same intercept is set for all treatment group while slopes are allowed to differ. The model equation is

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij4} + e_{ij}$$

where $X_{ij1} = 1$ for all i and j, and the intercept is represented by β_1. X_{ij2} is a covariate set to represent the time elapsed since baseline (i.e., $X_{ij2} = t_j$). To allow differing slopes between treatment groups,

$$X_{ij3} = \begin{cases} t_j & \textit{if the } i^{th} \textit{ individual belongs to the milled flax group} \\ 0 & \textit{otherwise} \end{cases}$$

and $$X_{ij4} = \begin{cases} t_j & \textit{if the } i^{th} \textit{ individual belongs to the flaxseed oil group} \\ 0 & \textit{otherwise} \end{cases}.$$

The expected Vitamin C levels are $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. More specifically, the conditional expectations for each group are as follows:

Control group $$E(Y_{ij} \mid X_{ij}) = \beta_1 + \beta_2 t_j + \beta_3(0) + \beta_4(0) = \beta_1 + \beta_2 t_j$$

Milled Flax group $$E(Y_{ij} \mid X_{ij}) = \beta_1 + \beta_2 t_j + \beta_3 t_j + \beta_4(0) = \beta_1 + (\beta_2 + \beta_3) t_j$$

Flaxseed Oil group $$E(Y_{ij} \mid X_{ij}) = \beta_1 + \beta_2 t_j + \beta_3(0) + \beta_4 t_j = \beta_1 + (\beta_2 + \beta_4) t_j$$

In the above expectations, β_1 is the intercept or mean Vitamin C level at Week 0. The terms β_2, $\beta_2 + \beta_3$, and $\beta_2 + \beta_4$, all of which are associated with t_j, represent the slope or expected change in mean Vitamin C levels at each measurement occasion for the control, milled flax and flaxseed oil groups, respectively.

The corresponding design matrices for the control, milled flax and flaxseed oil groups respectively, are shown below. Notice how these design matrices are written differently from the design matrix in the example from Section 5.3.1. This reflects the varying slopes for each group.

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 1 & 8 & 0 & 0 \\ 1 & 12 & 0 & 0 \end{pmatrix}, \boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 4 & 0 \\ 1 & 8 & 8 & 0 \\ 1 & 12 & 12 & 0 \end{pmatrix}, \text{ and } \boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & 0 & 4 \\ 1 & 8 & 0 & 8 \\ 1 & 12 & 0 & 12 \end{pmatrix}$$

Various hypothesis tests can be conducted to help answer specific questions of interest. For example, the hypothesis H_0: $\beta_3 = \beta_4 = 0$ tests whether the groups have the same rate of change in Vitamin C levels over time. Hypothesis testing will be discussed in Section 5.4.2.

5.4 Estimation

Given that $E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, $\text{Cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_i$, and $\boldsymbol{Y}_i$ is multivariate normally distributed, we need to consider how the unknown parameters (i.e., $\boldsymbol{\beta}$ and the components that make up $\boldsymbol{\Sigma}_i$) are estimated. Let $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\phi)$, where ϕ represents the vector of covariance parameters of dimension $q \times 1$. Maximum likelihood estimation and restricted maximum likelihood estimation are two methods for parameter estimation that will be discussed next.

5.4.1 Maximum Likelihood Estimation

The maximum likelihood (ML) approach estimates $\boldsymbol{\beta}$ and ϕ by maximizing the log-likelihood function. The log-likelihood function is denoted as

$$\ell = -\frac{K}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\left\{\sum_{i=1}^{N}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right\}$$

where K is defined as $\sum_{i=1}^{N} n_i$, i.e., the total number of observations.

The likelihood function is the joint probability of the response variables calculated at their observed values. The resulting maximum likelihood estimates (MLE's) are denoted by $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_i(\hat{\phi})$, or alternatively $\hat{\boldsymbol{\Sigma}}_i$. The details for the derivation of the MLE's will be omitted in this practicum.

If $\boldsymbol{\Sigma}_i$ is *known*,

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right\}^{-1} \sum_{i=1}^{N} \left(\boldsymbol{X}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Y}_i \right). \tag{5.5}$$

Equation (5.5) is the general least squares (GLS) estimator of $\boldsymbol{\beta}$.

Some important properties of $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\Sigma}_i$ is *known* are:

(i) $\hat{\boldsymbol{\beta}}$ has a multivariate distribution with $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (i.e., $\hat{\boldsymbol{\beta}}$ is unbiased), and

$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \left\{ \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right\}^{-1}$ when $\boldsymbol{Y}_i$ is multivariate normally distributed.

(ii) $\hat{\boldsymbol{\beta}}$ has the smallest variance when $\boldsymbol{\Sigma}_i$ is correctly specified.

(iii) The estimate of $\boldsymbol{\beta}$ that is derived when multivariate normality is assumed is valid (but possibly not the best estimate) even if the multivariate normal assumption is violated.

If $\boldsymbol{\Sigma}_i$ is *unknown*, the MLE of $\boldsymbol{\Sigma}_i$ is obtained first and then substituted into equation (5.5) to yield

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{X}_i \right\}^{-1} \sum_{i=1}^{N} \left(\boldsymbol{X}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i \right) \tag{5.6}$$

The same properties hold approximately for the estimate in (5.6) as when Σ_i is known but with the additional requirement of large sample size. An important remark is that as the number of covariance parameters to estimate increases compared to sample size, estimation becomes problematic (Fitzmaurice et al., 2004).

5.4.2 Maximum Likelihood Inference

Hypothesis tests can be conducted using the MLE of $\boldsymbol{\beta}$. After estimating $\boldsymbol{\beta}$, we can estimate its corresponding covariance matrix by $\hat{\text{Cov}}(\hat{\boldsymbol{\beta}}) = \left\{ \sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{X}_i \right\}^{-1}$. Then,

H_0: $\beta_k = 0$ is tested using the Wald statistic

$$Z = \frac{\hat{\beta}_k}{\sqrt{\hat{\text{Var}}(\hat{\beta}_k)}}$$

where β_k is the k^{th} component of $\boldsymbol{\beta}$, and $\sqrt{\hat{\text{Var}}(\hat{\beta}_k)}$ is the square root of the variance component of $\hat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ for $\hat{\beta}_k$. The 95% confidence interval for $\hat{\beta}_k$ has endpoints

$$\hat{\beta}_k \pm 1.96\sqrt{\hat{\text{Var}}(\hat{\beta}_k)}.$$

If the investigator desires to test linear combinations of the elements of $\boldsymbol{\beta}$, H_0: $\boldsymbol{L\beta} = 0$ is an appropriate hypothesis where $\boldsymbol{L}$ is a contrast vector with only one row. In this situation, the Wald test statistic is denoted by

$$Z = \frac{\boldsymbol{L}\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{L}\hat{\text{Cov}}(\hat{\boldsymbol{\beta}})\boldsymbol{L}'}}.$$

The 95% confidence interval for $\boldsymbol{L}\hat{\boldsymbol{\beta}}$ has endpoints

$$\boldsymbol{L}\hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{L}\hat{\text{Cov}}(\hat{\boldsymbol{\beta}})\boldsymbol{L}'}.$$

Both of the Wald statistics above can be compared to the standard normal distribution.

Suppose we exploit the well-known property that the square of a standardized normal random variable (i.e., Z^2), is distributed according to a chi-square distribution with a single degree of freedom. Using this information, the hypothesis H_0: $\boldsymbol{L\beta} = 0$ just described above can also be tested with the test statistic

$$W^2 = (\boldsymbol{L}\hat{\boldsymbol{\beta}})\left\{\boldsymbol{L}\hat{\text{Cov}}(\hat{\boldsymbol{\beta}})\boldsymbol{L}'\right\}^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

The statistic W^2 can be compared to a chi-square distribution with one degree of freedom. This methodology can be extended to the case where the $\boldsymbol{L}$ matrix has more than one row, say r rows. The test statistic W^2 is then compared to a chi-square distribution with r degrees of freedom.

5.4.3 Restricted Maximum Likelihood Estimation

An alternative to maximum likelihood estimation is restricted maximum likelihood estimation (REML). This method of estimation is recommended by Fitzmaurice, Laird and Ware (2004) to estimate $\boldsymbol{\Sigma}_i$ due to the fact that diagonal elements of $\boldsymbol{\Sigma}_i$ are prone to underestimation when the ML method is used. In REML estimation, $\boldsymbol{\Sigma}_i$ is estimated by forming a residual likelihood. This likelihood involves $\boldsymbol{\Sigma}_i$ but not $\boldsymbol{\beta}$ from the likelihood function. See Fitzmaurice, Laird and Ware (2004) for further details on this topic.

5.4.4 The Likelihood Ratio Test

Next, the likelihood ratio test (LRT) will be examined as an alternative to the Wald test for testing the null hypothesis of H_0: $\boldsymbol{L\beta} = 0$ against the alternative hypothesis of H_1: $\boldsymbol{L\beta} \neq 0$. First, it is important to distinguish between the null and alternative hypotheses. In the model for the null hypothesis, a constraint is placed in the form of $\boldsymbol{L\beta} = 0$. For the alternative hypothesis the model is unconstrained (i.e., $\boldsymbol{L\beta} \neq 0$). The constrained model

can be viewed as a reduced model of the unconstrained, or full model. If the reduced model holds, the full model must hold as well. Thus, the models are nested. In order to compare these two models, their corresponding log-likelihoods (i.e., $\hat{l}_{\text{FULL}}$ and $\hat{l}_{\text{REDUCED}}$) are compared. The likelihood ratio test statistic is

$$G^2 = 2(\hat{l}_{\text{FULL}} - \hat{l}_{\text{REDUCED}})$$

and is compared to a chi-squared distribution. The degrees of freedom are found by subtracting the number of parameters in the reduced model from the number of parameters in the full model. Likelihood ratio tests can be used for inference for the fixed effects as well as the variance components. When examining the fixed effects using the likelihood ratio test, results are valid only when ML estimation is used. On the other hand, when testing hypotheses regarding variance components, both ML and REML estimation provide valid likelihood ratio test results (Verbeke & Molenberghs, 2000; Fitzmaurice et al., 2004).

Chapter 6

Modeling the Mean and Covariance Structures

6.1 Introduction

In this chapter, the process of finding models for the mean and covariance structures for longitudinal data will be examined. After this, various models for the mean and covariance structures will be fit to the Vitamin C and Vitamin E response variables from the longitudinal study.

6.2 Modeling the Mean Structure

There are two main approaches for modeling the mean structure of longitudinal data, namely the analysis of response profiles and the response curves method. Both of these approaches will be examined in Sections 6.2.1 and 6.2.2, respectively.

6.2.1 Response Profile Analysis

The purpose of response profile analysis is to examine the shape of the mean response across measurement occasions. No specific constraint is placed on the mean when using this method. Response profile analysis is ideally suited to the case where measurement occasions are the same for all subjects and there is only one discrete covariate. Concerning the present study, the primary goal is to determine whether the changes in the mean response for each of the eight outcome variables of interest are the same for the control, milled flax and flaxseed oil groups (Fitzmaurice et al., 2004).

The analysis of response profiles raises three questions:

(1) Is there a group × time interaction? (e.g., If there is no interaction effect, the mean response profiles for the three groups are parallel)

(2) If there is no group × time interaction, is there a time effect? (e.g., If there is no time effect, the mean response profiles are level)

(3) If there is no group × time interaction, is there a group effect? (e.g., If there is no group effect, the mean response profiles are the 'same')

The group and time effects are commonly referred to as *main* effects in the literature on longitudinal data. Typically, the main effects are not relevant in the presence of a group × time interaction. To illustrate, parts (a), (b), and (c) of Figure 6.1 depict the situations of no group × time interaction, no time effect (but a group effect), and no group effect (but a time effect), respectively.

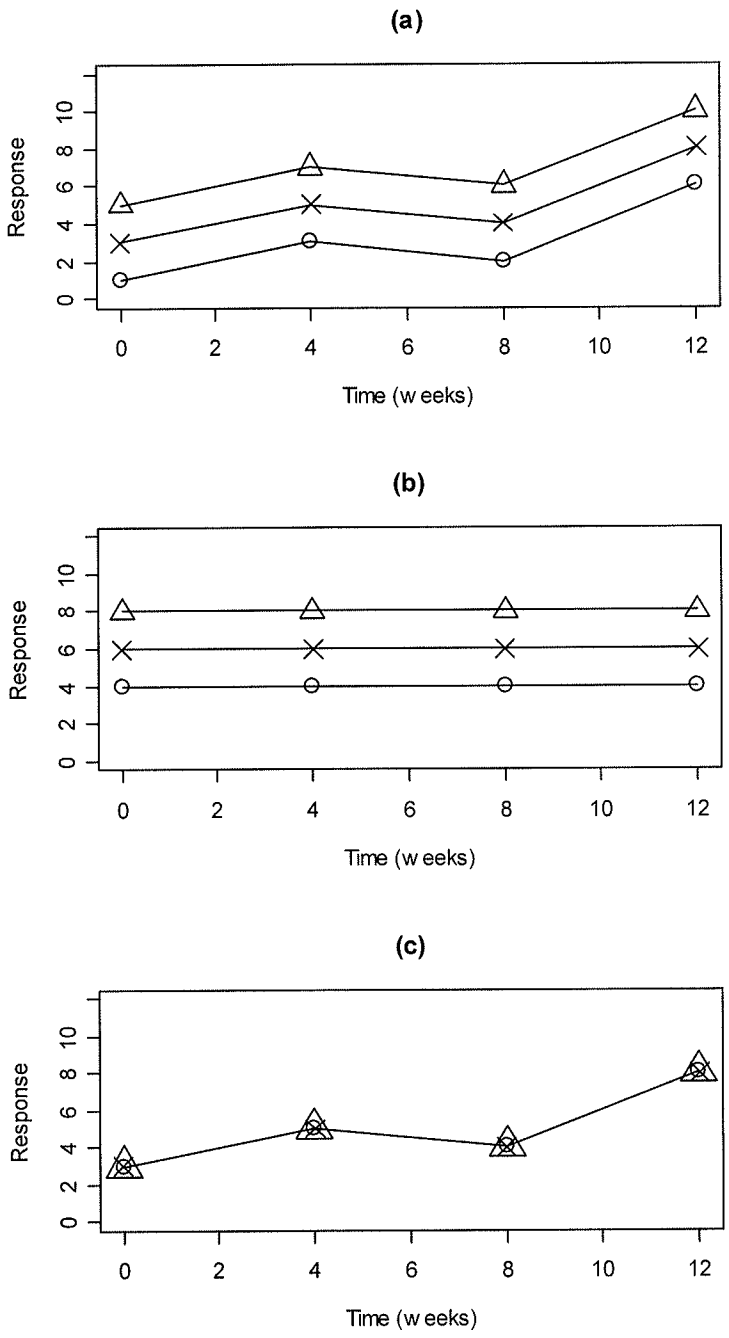

Figure 6.1: (a) No Group × Time Interaction Effect (b) No Time Effect (c) No Group Effect

Recall that the baseline measurements in this study were taken before any treatment was administered. In theory, the mean baseline responses for the groups should be equal as they do not depend on treatment group. Recollect the time plot in Figure 3.10(d) that

plotted the mean Vitamin C level at each measurement occasion stratified by treatment group. In this graph, the mean Vitamin C levels for the three groups at baseline did not appear to be equal, therefore adjustment for baseline differences may be necessary.

Let us examine the hypotheses outlined in the current section further. First, following Fitzmaurice, Laird and Ware (2004), define $\mu(g) = \{ \mu_1(g), \ldots, \mu_n(g)\}'$ as the mean response profile where $g = 1, \ldots, G$ denotes the group for $G \geq 2$ and n indicates the measurement occasion. To compare the G groups, define $\Delta_j(g) = \mu_j(g) - \mu_j(G)$ where j = 1, …, n and g = 1, …, $G - 1$. By defining notation in this manner, the null hypothesis test of no group × time interaction is

$$H_0: \ \Delta_1(g) = \Delta_2(g) = \ldots = \Delta_n(g)$$

where $g = 1, \ldots, G$–1 with $(G - 1) \times (n - 1)$ degrees of freedom.

Now that notation has been introduced, response profile analysis can be expressed in terms of the general linear model. Recall that

$$E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}$$

where $\boldsymbol{X}_i$ is the design matrix and $\boldsymbol{\beta}$ represents the vector of regression coefficients. In the current study, there are $G \times n$ (i.e., $3 \times 4 = 12$) parameters that comprise the response profiles for the three groups. There are a number of different ways that the design matrices can be written. We will examine two different parameterizations.

Parameterization 1:

The design matrices, $\boldsymbol{X}_i$, are written as follows:

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ for the control group,}$$

$$\mathbf{X}_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$ for the milled flax group, and

$$\boldsymbol{X}_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$ for the flaxseed oil group.

The vector of regression parameters is denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{12})'$.

Now, following the general linear model and multiplying each design matrix , $\boldsymbol{X}_i$ by $\boldsymbol{\beta}$,

$$\boldsymbol{\mu}_{\text{CO}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \ \boldsymbol{\mu}_{\text{MF}} = \begin{pmatrix} \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{pmatrix}, \text{ and } \boldsymbol{\mu}_{\text{FO}} = \begin{pmatrix} \beta_9 \\ \beta_{10} \\ \beta_{11} \\ \beta_{12} \end{pmatrix}.$$

As an example of interpretation, β_1, β_5, and β_9 are the mean values of the response variable at the first measurement occasion for the control, milled flax and flaxseed oil groups respectively. β_2 is the mean response at the second measurement occasion for the control group, and so on.

The null hypothesis test of no group × time interaction is now expressed in terms of $\boldsymbol{\beta}$ and is written as H_0: $(\beta_1 - \beta_5) = (\beta_2 - \beta_6) = (\beta_3 - \beta_7) = (\beta_4 - \beta_8)$ and

$$(\beta_1 - \beta_9) = (\beta_2 - \beta_{10}) = (\beta_3 - \beta_{11}) = (\beta_4 - \beta_{12})$$

and can be re-expressed as H_0: $\boldsymbol{L\beta} = \mathbf{0}$ where

$$\boldsymbol{L}=\begin{pmatrix}1&-1&0&0&-1&1&0&0&0&0&0&0\\1&0&-1&0&-1&0&1&0&0&0&0&0\\1&0&0&-1&-1&0&0&1&0&0&0&0\\1&-1&0&0&0&0&0&0&-1&1&0&0\\1&0&-1&0&0&0&0&0&-1&0&1&0\\1&0&0&-1&0&0&0&0&-1&0&0&1\end{pmatrix}.$$

Parameterization 2:

As a second example, the design matrices, $\boldsymbol{X}_i$, are written as follows in reference cell mode. An intercept is included for the control group which is set as the reference group. The milled flax and flaxseed oil groups are coded as comparisons to the control (i.e. reference) group.

$$\boldsymbol{X}_i=\begin{pmatrix}1&0&0&0&0&0&0&0&0&0&0&0\\1&1&0&0&0&0&0&0&0&0&0&0\\1&0&1&0&0&0&0&0&0&0&0&0\\1&0&0&1&0&0&0&0&0&0&0&0\end{pmatrix}$$ for the control group,

$$\boldsymbol{X}_i=\begin{pmatrix}1&0&0&0&1&0&0&0&0&0&0&0\\1&1&0&0&1&1&0&0&0&0&0&0\\1&0&1&0&1&0&1&0&0&0&0&0\\1&0&0&1&1&0&0&1&0&0&0&0\end{pmatrix}$$ for the milled flax group, and

$$\boldsymbol{X}_i=\begin{pmatrix}1&0&0&0&0&0&0&0&1&0&0&0\\1&1&0&0&0&0&0&0&1&1&0&0\\1&0&1&0&0&0&0&0&1&0&1&0\\1&0&0&1&0&0&0&0&1&0&0&1\end{pmatrix}$$ for the flaxseed oil group.

Now, following the general linear model,

$$\boldsymbol{\mu}_{\text{CO}}=\begin{pmatrix}\beta_1\\\beta_1+\beta_2\\\beta_1+\beta_3\\\beta_1+\beta_4\end{pmatrix},\ \boldsymbol{\mu}_{\text{MF}}=\begin{pmatrix}\beta_1+\beta_5\\(\beta_1+\beta_5)+(\beta_2+\beta_6)\\(\beta_1+\beta_5)+(\beta_3+\beta_7)\\(\beta_1+\beta_5)+(\beta_4+\beta_8)\end{pmatrix},\text{ and }\boldsymbol{\mu}_{\text{FO}}=\begin{pmatrix}\beta_1+\beta_9\\(\beta_1+\beta_9)+(\beta_2+\beta_{10})\\(\beta_1+\beta_9)+(\beta_3+\beta_{11})\\(\beta_1+\beta_9)+(\beta_4+\beta_{12})\end{pmatrix}.$$

For interpretation purposes, β_1, $\beta_1 + \beta_5$, and $\beta_1 + \beta_9$ are the mean values of the response variable at the first measurement occasion for the control, milled flax and flaxseed oil groups respectively. The rest of the parameters are interpreted accordingly.

The control group is a reference group and the null hypothesis test of no group × time interaction is modified from the previous example as

H_0: $\beta_6 = \beta_7 = \beta_8 = \beta_{10} = \beta_{11} = \beta_{12} = 0$

The null hypothesis does not need to be re-expressed in terms of $\boldsymbol{L\beta}$. Alternatively, hypothesis tests of the main effects may be of interest if the hypothesis of no group × time interaction holds true.

H_0: $\beta_2 = \beta_3 = \beta_4 = 0$ is the hypothesis of no time effect while

H_0: $\beta_5 = \beta_9 = 0$ is the hypothesis of no group effect.

6.2.1.1 Properties of Response Profile Analysis Method

The response profile analysis method is robust in nature due to the minimal constraints on the mean response profile and covariance structure. The mean can be different for each group at each measurement occasion, thus no structure is imposed. Consequently, risk of bias is nominal. It is unfortunately also a consequence of this feature and also in part due to the fact that response profile analysis does not take into account the order of repeated measurements, that only general conclusions can be made about between-subjects effects across time (Fitzmaurice et al., 2004).

6.2.2 Parametric Curves for Modeling the Mean Structure

A second approach for modeling the mean is to use parametric or semi-parametric curves. In contrast to the response profile analysis approach, specific constraints or structure are placed on the mean, and numerical values of the variable time are taken into

consideration. Also, individuals do not necessarily have to be measured at the same occasions. Furthermore, parametric curves for modeling the mean structure do so with less parameters than response profiles and have greater power for testing main and interaction effects (Fitzmaurice et al., 2004). The parametric curves to be discussed are linear and quadratic in nature. Similar to the response profile analysis approach, 'Group' membership is still treated as a categorical variable. In contrast to the response profile analysis approach, 'Time' is now a quantitative variable, thus making the parametric approach fit into the regression-type model category.

6.2.2.1 Linear Response Trends

Define

$$Y_{ij} = \beta_1 + \beta_2(\text{Time}_{ij}) + \beta_3(\text{Group}_i) + \beta_4(\text{Time}_{ij} \times \text{Group}_i) + e_{ij} \tag{6.1}$$

which models the mean response as a linear function of time for a two-group design. Model (6.1) can be augmented for the case of three groups to fit to the study in this practicum. This augmentation is shown in model (6.2). Define

$$\begin{aligned} Y_{ijkl} = \beta_1 + \beta_2(\text{Time}_j) + \beta_3(\text{Group}_k) + \beta_4(\text{Time}_j \times \text{Group}_k) + \\ \beta_5(\text{Group}_l) + \beta_6(\text{Time}_j \times \text{Group}_l) + e_{ij} \end{aligned} \tag{6.2}$$

where $\text{Group}_k = \begin{cases} 1 & \textit{if the } i^{th} \textit{ subject belongs to the milled flax group} \\ 0 & \textit{otherwise} \end{cases}$

and $\text{Group}_l = \begin{cases} 1 & \textit{if the } i^{th} \textit{ subject belongs to the flaxseed oil group} \\ 0 & \textit{otherwise} \end{cases}$.

The subscript i in Time_{ij} has been omitted in equation (6.2) because all subjects are measured at the same set of occasions.

Thus, in reference to the treatment groups, equation (6.2) is denoted as,

$Y_{ij00} = \beta_1 + \beta_2 \text{Time}_j + e_{ij}$ for the control group (i.e., $k = l = 0$),

$Y_{ij10} = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\text{Time}_j + e_{ij}$ for the milled flax group (i.e., $k = 1, l = 0$), and

$Y_{ij01} = (\beta_1 + \beta_5) + (\beta_2 + \beta_6)\text{Time}_j + e_{ij}$ for the flaxseed oil group (i.e., $k = 0, l = 1$).

One can see that each group has a unique intercept and slope.

In reference to the general linear model, $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i$, the design matrices for each group are written as follows:

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 & 0 & 0 \end{pmatrix}$$ for the control group,

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 & 0 & 0 \\ 1 & 8 & 1 & 8 & 0 & 0 \\ 1 & 12 & 1 & 12 & 0 & 0 \end{pmatrix}$$ for the milled flax group, and

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 & 4 \\ 1 & 8 & 0 & 0 & 1 & 8 \\ 1 & 12 & 0 & 0 & 1 & 12 \end{pmatrix}$$ for the flaxseed oil group.

The models defined above have the capability of handling both qualitative and quantitative covariates.

6.2.2.2 Quadratic Response Trends

Define

$$\begin{aligned} Y_{ij} = \beta_1 + \beta_2(\text{Time}_{ij}) + \beta_3(\text{Time}_{ij}^{\,2}) + \beta_4(\text{Group}_i) + \\ \beta_5(\text{Time}_{ij} \times \text{Group}_i) + \beta_6(\text{Time}_{ij}^{\,2} \times \text{Group}_i) + e_{ij} \end{aligned} \tag{6.3}$$

which models the mean response as a quadratic function of time for a two-group design. Again, the equation can be augmented so that it is appropriate for a three-group design that fits the current study. That is,

$$\begin{aligned} Y_{ij} = \beta_1 + \beta_2(\text{Time}_j) + \beta_3(\text{Time}_j^2) + \beta_4(\text{Group}_k) + \beta_5(\text{Time}_j \times \text{Group}_k) + \\ \beta_6(\text{Time}_j^2 \times \text{Group}_k) + \beta_7(\text{Group}_l) + \\ \beta_8(\text{Time}_j \times \text{Group}_l) + \beta_9(\text{Time}_j^2 \times \text{Group}_l) + e_{ij} \end{aligned} \tag{6.4}$$

where $\text{Group}_k = \begin{cases} 1 \text{ if subject belongs to the milled flax group} \\ 0 \text{ otherwise} \end{cases}$

and $\text{Group}_l = \begin{cases} 1 \text{ if subject belongs to the flaxseed oil group} \\ 0 \text{ otherwise} \end{cases}$.

Thus, the expectations based on equation (6.4) are

$E(Y_{ij00}) = \beta_1 + \beta_2(\text{Time}_j) + \beta_3(\text{Time}_j^2)$ for the control group (i.e., $k = l = 0$),

$E(Y_{ij10}) = (\beta_1 + \beta_4) + (\beta_2 + \beta_5)\text{Time}_j + (\beta_3 + \beta_6)\text{Time}_j^2$ for the milled flax group

(i.e., $k = 1, l = 0$), and

$E(Y_{ij01}) = (\beta_1 + \beta_7) + (\beta_2 + \beta_8)\text{Time}_j + (\beta_3 + \beta_9)\text{Time}_j^2$ for the flaxseed oil group

(i.e., $k = 0, l = 1$).

Notice that each group has a different intercept and the rate of change varies as a function of time.

6.3 Modeling the Covariance Structure

Modeling the covariance structure is an important topic in the analysis of longitudinal data. Recall that responses observed on the same individual over time are correlated. This correlation must be modeled properly in order to obtain valid statistical inferences. In fact, correctly modeling the covariance structure for longitudinal data

increases efficiency of the estimates and ensures correct standard errors (Fitzmaurice et al., 2004).

The covariance structures for longitudinal data that will be examined in this practicum are the (a) unstructured covariance, and (2) covariance pattern models.

6.3.1 Unstructured Covariance

A covariance structure in which no constraints are placed on the elements of the covariance matrix is the unstructured (UN) covariance. The covariance matrix must be symmetric as well as positive definite and is most suited to a balanced design with few measurement occasions. The unstructured covariance matrix has $n \times (n + 1)/2$ unique parameters and is written as

$$\text{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n2} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}.$$

When there are three measurement occasions, there are $3 \times (3 + 1)/2 = 6$ covariance parameters to estimate. With six measurement occasions, there are $6 \times (6 + 1)/2 = 15$ covariance parameters to estimate. As shown in these examples, the number of covariance parameters to estimate increases quickly as the number of repeated measurements increases. If sample size is small relative to the number of covariance parameters, it can result in unstable estimation (Fitzmaurice et al, 2004).

6.3.2 Covariance Pattern Models

There are numerous covariance structures which place constraints on the elements that comprise them. Collectively these structures are known as covariance pattern models.

6.3.2.1 *Simple Covariance Structure*

The 'simple' covariance structure is denoted by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

The variance σ^2 is constant over time.

6.3.2.2 *Compound Symmetric Structure*

The compound symmetric (CS) covariance structure is denoted by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}.$$

The variance σ^2, is constant over time and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho$, ($\rho \geq 0$) for all j and k, ($j \neq k$). For this structure, there are only two parameters to estimate, σ^2 and ρ, for any number of repeated measurements. A variation of the compound symmetric covariance structure is the heterogeneous compound symmetric structure which is denoted by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \dots & \rho\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_1\sigma_n & \rho\sigma_2\sigma_n & \dots & \sigma_n^2 \end{pmatrix}.$$

Notice how the variance is allowed to differ based on the specific measurement occasion. For the heterogeneous compound symmetric structure, there are $n + 1$ parameters to estimate.

6.3.2.3 *Autoregressive Structure*

This structure accounts for the decreasing magnitude in correlation between pairs of repeated measurements as time increases. This is a common feature of longitudinal studies. The autoregressive covariance structure of order 1 (AR-1) is written as

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

The variance, σ^2, is constant over time and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho^{|j-k|}$, ($\rho \geq 0$) for all j and k, ($j \neq k$). Similarly to the compound symmetric structure, the only two parameters to estimate are σ^2 and ρ. A variation of the AR-1 structure that allows for heterogeneous variances is aptly named the heterogeneous AR-1 (ARH-1) structure which is denoted by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \dots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \dots & \rho^{n-2}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \dots & \sigma_n^2 \end{pmatrix}.$$

This covariance structure has $n + 1$ parameters to estimate.

6.3.2.4 *Toeplitz Covariance Structure*

The Toeplitz (TOEP) structure is a variation of the AR-1 structure. It accounts for correlation by assigning the identical correlation between pairs of repeated measurements that are equally spaced. The variance, σ^2, is constant over time as in the AR-1 structure and $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho_k$ for all j and k. There are n covariance parameters to estimate. The Toeplitz structure is denoted by

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

A covariance structure that is most ideal is a structure that is both parsimonious in the number of parameters to estimate and also a good fit to the UN estimates for both the variance-covariance and correlation components. There are many more covariance structures that can be specified. PROC MIXED in SAS has the capability of implementing over 20 unique covariance structures. For further details on the various types of structures available in PROC MIXED, refer to the SAS Documentation.

6.3.3 Information Criteria

When fitting various covariance structures to longitudinal data with a specific mean model, it is useful to be able to compare the different models. This is especially true for the case where the models are not nested. In this case, the likelihood ratio test cannot be used to compare models. We can compare models here by this using various information criteria such as the Akaike Information Criterion (AIC), the Corrected Akaike Information Criterion (AICC) and the Bayesian Information Criterion (BIC).

The AIC is

$$\mathrm{AIC} = -2\log \mathrm{L} + 2p$$

where p is defined as the number of covariance parameters. The AIC has two main goals:

1. Determine whether the specified covariance structure fits the data well.
2. Determine whether the selected covariance structure not only fits the data well but is also parsimonious.

A covariance structure that fits the data well, but is not parsimonious in terms of the quantity of covariance parameters is effectively penalized.

The AICC is

$$\text{AICC} = -2\log \text{L} + p\log(N^* + 1)$$

where p is defined as the number of covariance parameters and N^* is defined as the number of 'subjects' (i.e., $N^* = N$ and $N^* = N + p$ for ML and REML estimation respectively).

The BIC is

$$\text{BIC} = -2\log \text{L} + p\log N^*$$

where N^* is defined as the number of subjects (i.e., $N^* = N$ and $N^* = N + p$ for ML and REML estimation respectively). The main goal of the BIC is to choose the model with the greatest Bayes factor, or posterior probability. The BIC tends to choose more simple models than the AIC. This is because the penalty incurred for increased covariance parameters is greater than for the AIC and is not recommended.

When choosing among the models with various covariance structures, the information criteria with the smallest value selects the best model. The reader should keep in mind that the same model will not always be selected by the various information criteria (Fitzmaurice et al., 2004; Hedeker & Gibbons, 2006; Lix & Lloyd, 2006).

6.4 Model Diagnostics

Upon selecting a model for the longitudinal data, residuals should be examined as a final evaluation of model fit. Residuals can assist a researcher in determining inadequacies for both the mean and covariance structures of the model. Furthermore, outliers and skewness can also be revealed through residual analyses. Define

$$\boldsymbol{r}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}$$

as the vector of residuals for the i^{th} individual.

An important property of $\boldsymbol{r}_i$ is that the observations that comprise $\boldsymbol{r}_i$ are correlated because of the within-subject heterogeneity that arises in longitudinal data. Due to this property, the residuals must be transformed so that standard residual diagnostics can be applied. Fitzmaurice, Laird and Ware (2004) recommend using the *Cholesky decomposition method* to transform the residuals. Although further details will not be discussed in this practicum, by transforming the residuals through the Cholesky decomposition, one can assume that the residuals have constant variance and are uncorrelated. Histograms, normal quantile plots and scatterplots of the transformed residuals can be examined to detect departures from normality. The resulting histograms, normal quantile plots and scatterplots are all interpreted in standard fashion.

6.5 Examples

Next, the Vitamin C and Vitamin E response variables in the metabolic syndrome study will be examined by applying some of the mean models and covariance structures discussed in the previous sections. Also, hypotheses tests will be conducted using the inference techniques discussed in Chapter 5.

6.5.1 Example 1: Modeling Vitamin C via Response Profile

Firstly, the Vitamin C data were fit using the response profile analysis method (Section 6.2.1) for the mean model with an unstructured covariance (Section 6.3.1) specified for the covariance structure. These particular models for the mean and covariance structures were indicated using the PROC MIXED procedure in SAS; also REML estimation was used (see Appendix B for SAS syntax). By default, SAS uses the

reference cell parameterization, or the second parameterization discussed in Section 6.2.1. The control group and the baseline measurement were set as the reference group and time, respectively. Multivariate Wald tests were output to test the interaction and main effects. For the response profile analysis method, both 'Group' membership as well as 'Time', were treated as categorical variables. Thus, the response profile model is an ANOVA-type model.

Table 6.1 displays the estimated covariance matrix for the Vitamin C data. Focusing on the diagonal elements of the matrix, we can see that the variability in Vitamin C levels decrease slightly from baseline to Week 4. From Week 4 to Week 8, the variability decreases even more. At Week 12, the variability rises and is almost as it was at baseline. The largest variance (0.2101) is not more than twice as large as the smallest variance (0.1293); therefore homogeneity of variances across the measurement occasions is a reasonable assumption.

0.2101	0.1224	0.0935	0.0835
0.1224	0.1715	0.0995	0.1172
0.0935	0.0995	0.1293	0.1113
0.0835	0.1172	0.1113	0.1730

Table 6.1: Estimated Covariance Matrix for Vitamin C Data

Table 6.2 displays the main and interaction effects test results. Recall Section 5.4 for Wald test statistic. For the group × time interaction test, the value of the Wald test statistic is 21.33 with a corresponding p-value of 0.0016. Formally, the null hypothesis of no group × time interaction is rejected and it can be concluded that mean response profiles for Vitamin C are not the same for the control, milled flax and flaxseed oil groups.

Effect	Df	*Wald*	p-value
Group	2	8.99	0.0112
Time	3	27.59	<.0001
Group × Time	6	21.33	0.0016

Table 6.2: Tests of Fixed Effects

Table 6.3 displays the estimates (and standard errors) of the β's. For ease of interpretation, the control group at Week 0 was used as the reference group in the SAS code. To relate the SAS output to the models just outlined, recall

$$\boldsymbol{\mu}_{CO} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + \beta_3 \\ \beta_1 + \beta_4 \end{pmatrix}, \quad \boldsymbol{\mu}_{MF} = \begin{pmatrix} \beta_1 + \beta_5 \\ (\beta_1 + \beta_5) + (\beta_2 + \beta_6) \\ (\beta_1 + \beta_5) + (\beta_3 + \beta_7) \\ (\beta_1 + \beta_5) + (\beta_4 + \beta_8) \end{pmatrix}, \text{ and } \boldsymbol{\mu}_{FO} = \begin{pmatrix} \beta_1 + \beta_9 \\ (\beta_1 + \beta_9) + (\beta_2 + \beta_{10}) \\ (\beta_1 + \beta_9) + (\beta_3 + \beta_{11}) \\ (\beta_1 + \beta_9) + (\beta_4 + \beta_{12}) \end{pmatrix}.$$

Effect	Group	Time	Estimate	Std Err	*Wald*	*p*-value
Intercept			0.7140	0.1450	4.93	<.0001
Group	MF		0.2291	0.1928	1.19	0.2436
Group	FO		–0.1848	0.1963	–0.94	0.3534
Time		Week 12	–0.1990	0.1470	–1.35	0.1854
Time		Week 8	–0.0330	0.1235	–0.27	0.7910
Time		Week 4	–0.1630	0.1170	–1.39	0.1731
Group × Time	MF	Week 12	0.1490	0.1956	0.76	0.4517
Group × Time	MF	Week 8	–0.3062	0.1643	–1.86	0.0715
Group × Time	MF	Week 4	–0.0855	0.1556	–0.55	0.5867
Group × Time	FO	Week 12	0.1440	0.1991	0.72	0.4747
Group × Time	FO	Week 8	–0.2228	0.1672	–1.33	0.1921
Group × Time	FO	Week 4	–0.1737	0.1584	–1.1	0.2811

Table 6.3: Solution for Fixed Effects

$\hat{\beta}_1 = 0.7140$, $\hat{\beta}_5 = 0.2291$, and $\hat{\beta}_9 = -0.1848$, are estimates for the first measurement occasion. Thus, for the control group, 0.7140 is the estimated Vitamin C level at the first measurement occasion. For the milled flax group, 0.7140 + 0.2291 = 0.9431 is the estimated Vitamin C level at baseline. Finally, for the flaxseed oil group, 0.7140 – 0.1848 = 0.5292 is the estimated Vitamin C level at Week 0. Table 6.4 displays the

observed and estimated mean response for Vitamin C at each measurement stratified by treatment group.

		Week 0	Week 4	Week 8	Week 12
Control	**Observed**	0.71	0.55	0.68	0.52
	Estimated	0.71	0.55	0.68	0.52
Milled Flax	**Observed**	0.94	0.69	0.60	0.89
	Estimated	0.94	0.69	0.60	0.89
Flaxseed Oil	**Observed**	0.53	0.19	0.27	0.47
	Estimated	0.53	0.19	0.27	0.47

Table 6.4: Observed and Estimated means at each measurement occasion

Notice in Table 6.4 that when the mean is modeled using the response profile analysis method, the observed and estimated means at each occasion are identical.

Let us go back to Table 6.3 and examine the tests of the group effect. Due to the reference cell parameterization for this model, the mean Vitamin C levels were compared between the groups at the *baseline* measurement. Non-significant p-values of 0.2436 and 0.3534 for the milled flax and flaxseed oil groups respectively indicate that the mean Vitamin C levels at baseline do not differ between the control, milled flax and flaxseed oil groups. Recall the initial observation that there may be a difference between the groups based on Figure 3.10(d). This result indicates that there may be no statistically significant difference between the groups. Based on the results due to this model, no adjustments are necessary. If adjustments were necessary, there are a few methods for handling differences at baseline. They include (a) subtracting the baseline response from the subsequent responses and analyzing the resulting differences, and (b) including the baseline response as a covariate in the model for the analysis of the post-baseline measurements (Fitzmaurice et al., 2004). When comparing the milled flax group to the control group, from baseline to Week 4, the milled flax group has a slightly greater

decrease (0.0855) in Vitamin C levels than the control group but the difference is not significant (p-value = 0.5867). From baseline to Week 8, the milled flax group again has a greater decrease (0.3062) in Vitamin C levels than the control group. Again, the difference is insignificant as suggested by a p-value of 0.0715. From baseline to Week 12, the milled flax group has a greater increase (0.1490) in Vitamin C levels than the control group but a p-value of 0.4517 indicates that the difference is not statistically significant. The comparison pattern of the flaxseed oil group to the control group is similar to that just described for the milled flax to control group. A greater decrease in Vitamin C levels from baseline to Week 4 (0.1737) and baseline to Week 8 (0.2228) is shown for the flaxseed oil group in comparison to the control group. From baseline to Week 12, the flaxseed oil group has a greater increase (0.1440) in Vitamin C levels than the control group. For the three comparisons just described, none are statistically significant as indicated by p-values of 0.2811, 0.1921, and 0.4747 respectively.

Next, residuals for this model will be examined commencing with a histogram plot.

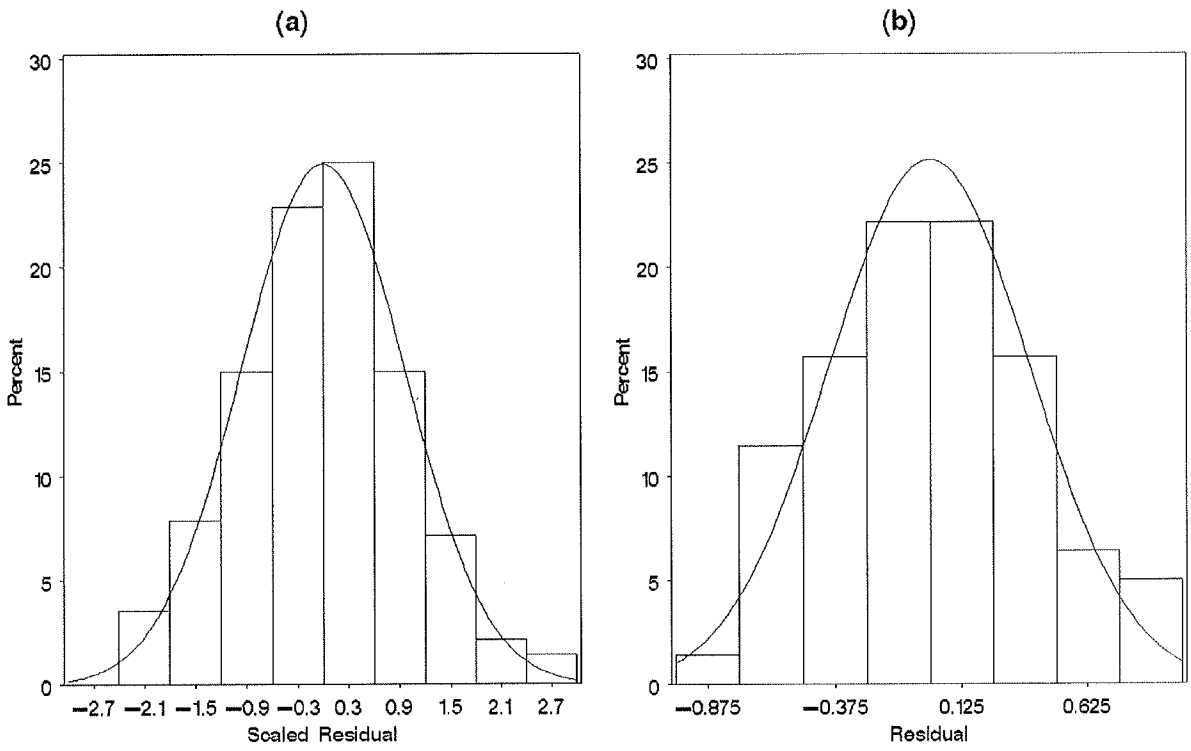

Figure 6.2: (a) Response Profile Model: Histogram plot of transformed residuals, (b) Histogram plot of untransformed residuals

Figure 6.2(a) displays the transformed residuals due to the Cholesky decomposition method while Figure 6.2(b) displays the untransformed residuals for comparison. Based on the histogram of the transformed residuals, the distribution appears normally distributed.

Next, normal quantile plots (or QQ plots) for the transformed and raw residuals are shown in Figure 6.7(a) and (b) respectively.

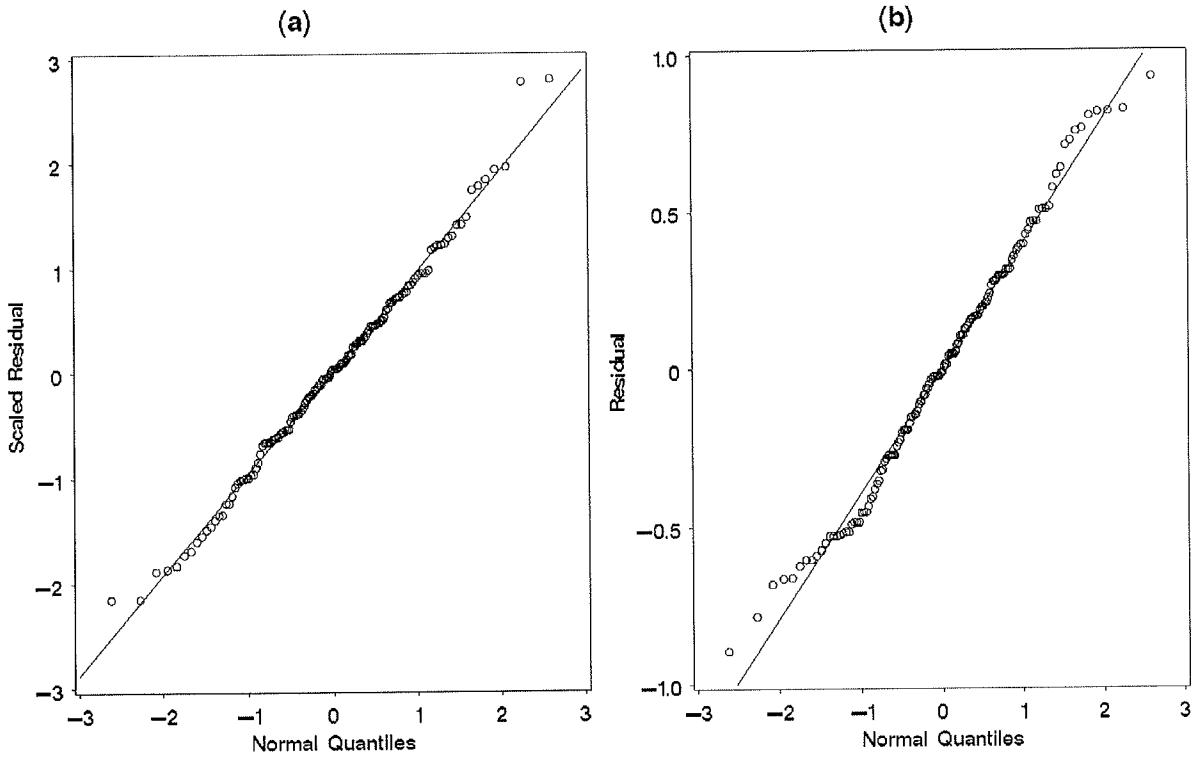

Figure 6.3: Response Profile Model: (a) QQ plot of transformed residuals, (b) QQ plot of untransformed residuals

Based on the QQ plot of the transformed residuals in Figure 6.3(a), there may be one or two outlying individuals. Otherwise, the residuals fit the line well indicating normality.

At this point the reader likely begins to wonder why the omnibus tests of the fixed effects in Table 6.2 were significant and none of the estimates of β in Table 6.3 were. While non-significant group effects can be explained by the reference cell parameterization, non-significant group × time interactions are more puzzling. Given these results, there is most likely a better model for these data. Therefore, we will re-model the Vitamin C data and commence with an entirely different model for the mean.

6.5.2 Example 1 (continued): Modeling Vitamin C via Parametric Curves

Recall that linear trends, quadratic trends and linear spline models (not discussed in this practicum) can all be examined in the form of the general linear model

$$\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta}.$$

We will examine both the linear and quadratic trends for the Vitamin C data in further detail. For the control, milled flax and flaxseed oil groups respectively, the design matrices, $\boldsymbol{X}_i$, for the quadratic trend are written as

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 16 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 8 & 64 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 12 & 144 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 16 & 1 & 4 & 16 & 0 & 0 & 0 \\ 1 & 8 & 64 & 1 & 8 & 64 & 0 & 0 & 0 \\ 1 & 12 & 144 & 1 & 12 & 144 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 4 & 16 & 0 & 0 & 0 & 1 & 4 & 16 \\ 1 & 8 & 64 & 0 & 0 & 0 & 1 & 8 & 64 \\ 1 & 12 & 144 & 0 & 0 & 0 & 1 & 12 & 144 \end{pmatrix}.$$

The regression coefficients are represented by the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_9)'$.

In the control, milled flax and flaxseed oil groups respectively, the mean responses are denoted by

$$\boldsymbol{\mu}_i = \begin{pmatrix} \beta_1 + 0\beta_2 + 0\beta_3 \\ \beta_1 + 4\beta_2 + 16\beta_3 \\ \beta_1 + 8\beta_2 + 64\beta_3 \\ \beta_1 + 12\beta_2 + 144\beta_3 \end{pmatrix}, \quad \boldsymbol{\mu}_i = \begin{pmatrix} (\beta_1 + \beta_4) + 0(\beta_2 + \beta_5) + 0(\beta_3 + \beta_6) \\ (\beta_1 + \beta_4) + 4(\beta_2 + \beta_5) + 16(\beta_3 + \beta_6) \\ (\beta_1 + \beta_4) + 8(\beta_2 + \beta_5) + 64(\beta_3 + \beta_6) \\ (\beta_1 + \beta_4) + 12(\beta_2 + \beta_5) + 144(\beta_3 + \beta_6) \end{pmatrix}, \text{ and}$$

$$\boldsymbol{\mu}_i = \begin{pmatrix} (\beta_1 + \beta_7) + 0(\beta_2 + \beta_8) + 0(\beta_3 + \beta_9) \\ (\beta_1 + \beta_7) + 4(\beta_2 + \beta_8) + 16(\beta_3 + \beta_9) \\ (\beta_1 + \beta_7) + 8(\beta_2 + \beta_8) + 64(\beta_3 + \beta_9) \\ (\beta_1 + \beta_7) + 12(\beta_2 + \beta_8) + 144(\beta_3 + \beta_9) \end{pmatrix}.$$

For the linear trend model, the design matrices above are modified by removing the columns that contain the quadratic term and adjusting the regression coefficient vector to include only six rather than nine β parameters. The mean responses can then be adjusted accordingly.

When using parametric curves for modeling the mean response, it should be noted that the covariance structure of $\boldsymbol{Y}_i$ need not be unstructured and thus a more parsimonious structure can be adopted (Fitzmaurice et al., 2004).

To apply the linear and quadratic response trend to a numerical example, PROC MIXED in SAS is fit to the Vitamin C data assuming an unstructured covariance matrix (See Appendix B for computer syntax). The control group was set as the reference group. First, a (a) linear trend model with an unstructured covariance was fit using ML estimation in order to be compared to the (b) response profile trend with an unstructured covariance structure discussed in Section 6.5.1. These models are nested, thus a likelihood ratio test can be conducted. The response profile trend model was re-fit using ML estimation in order to compare the models. As suspected, the response profile trend model was not adequate in comparison to the linear trend model. In order to compare the linear and quadratic models, maximum likelihood estimation was used to model these

response trends to the data using an unstructured covariance structure. The –2 log likelihood fit statistics are shown in Table 6.5.

	–2 Log Likelihood
Linear	96.1
Quadratic	64.9

Table 6.5: -2 Log Likelihood Fit statistics

The likelihood ratio test statistic can be calculated as $G^2 = 31.2$ (i.e., 96.1 – 64.9 = 31.2). The degrees of freedom are calculated by subtracting the number of parameters in the linear model from the number of parameters in the quadratic model (i.e., 12 – 8 = 4). The test statistic is then compared to a chi-squared distribution which in this case yields a *p*-value <0.005. Thus, the linear model is shown to provide an inferior fit to the data as compared to the quadratic model.

Recall the time plot of the Vitamin C means across measurement occasions shown in Figure 3.10(d) and reproduced here as Figure 6.4. Based on this time plot, the quadratic model seems appropriate for the milled flax and flaxseed oil groups. For the control group, a cubic trend model may be appropriate but the quadratic model may still provide an adequate fit for the data.

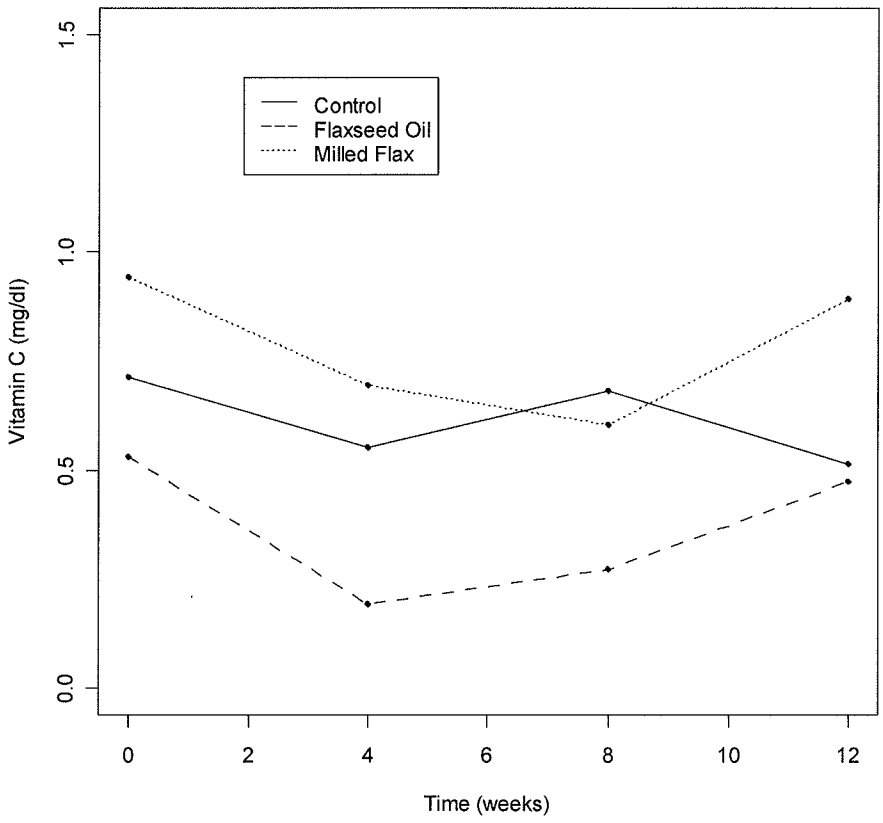

Figure 6.4: Time plot of means for Vitamin C

Next, upon re-fitting the quadratic model using REML estimation, the tests of the fixed effects and the estimated regression coefficients are displayed in Table 6.7 and Table 6.7 respectively.

Effect	**Df**	***Wald***	***p*-value**
Group	2	5.75	0.0564
Time	1	20.86	<.0001
Time × Time	1	25.83	<.0001
Group × Time	2	9.67	0.0079
Group × Time × Time	2	15.53	0.0004

Table 6.6: Tests of Fixed Effects

The time, time × time interaction, group × time interaction and group × time × time interaction effects are all significant. The most interesting result is the significance of the

group × time interaction effects. These significant interactions indicate that there may be a difference between the groups over time.

Effect		Estimate	Std Err	Test Statistic	*p*-value
Intercept		0.6904	0.1444	4.78	<.0001
Group	MF	0.2616	0.1921	1.36	0.1827
Group	FO	–0.1731	0.1955	–0.89	0.3825
Time		–0.0007	0.0284	–0.03	0.9794
Time×Time		–0.0007	0.0020	–0.37	0.7165
Time×Group	MF	–0.1100	0.0378	–2.91	0.0065
Time×Group	FO	–0.0969	0.0384	–2.52	0.0169
Time×Time×Group	MF	0.0094	0.0026	3.61	0.0010
Time×Time×Group	FO	0.0088	0.0026	3.32	0.0022

Table 6.7: Estimated Regression Coefficients for the Quadratic Trend Model

In Table 6.7, all of the group × time interaction and group × time × time interaction effects are significant. It is clear that the intercept and rate of change expressed as functions of 'Time' and 'Time × Time' respectively are different for the control, milled flax and flaxseed groups. The estimated mean responses for the treatment groups are as follows:

$$E(\hat{Y}_{ij}) = 0.6904 - 0.0007(\text{Time}) - 0.0007(\text{Time}^2),$$

$$E(\hat{Y}_{ij}) = (0.6904 + 0.2616) - (0.0007 + 0.1100)\text{Time} - (0.0007 - 0.0094)\text{Time}^2, \text{ and}$$

$$E(\hat{Y}_{ij}) = (0.6904 - 0.1731) - (0.0007 + 0.0969)\text{Time} - (0.0007 - 0.0088)\text{Time}^2$$

for the control, milled flax, and flaxseed oil groups respectively.

The observed and estimated means for each treatment group by measurement occasion are shown in Table 6.8 below.

		Week 0	Week 4	Week 8	Week 12
Control	**Observed**	0.71	0.55	0.68	0.52
	Estimated	0.69	0.68	0.64	0.58
Milled Flax	**Observed**	0.94	0.69	0.60	0.89
	Estimated	0.95	0.65	0.62	0.88
Flaxseed Oil	**Observed**	0.53	0.19	0.27	0.47
	Estimated	0.52	0.26	0.25	0.51

Table 6.8: Observed and Estimated Means

As evident from Table 6.8, the quadratic trend model seems to fit the data quite well. The only means that do not fit the data well are at Week 4 for both the control and flaxseed oil groups.

In order to see if any higher order polynomials would provide an even better fit to the data, a cubic trend model was fit. The cubic trend model was compared to the quadratic trend model but a likelihood ratio test revealed that the quadratic trend model was adequate.

Next, a comparison of the appropriateness of different covariance structures' fit to the data will be made using graphs, results of statistical tests and information criteria. As an unstructured (UN) covariance structure has already been fit to the data, it will be compared to various *structured* covariance matrices. Table 6.9 and Table 6.10 present the covariance and correlation matrices when an unstructured covariance structure is fit for the quadratic mean model.

0.2102	0.1218	0.0936	0.0832
0.1218	0.1743	0.0986	0.1186
0.0936	0.0986	0.1296	0.1109
0.0832	0.1186	0.1109	0.1738

Table 6.9: Unstructured Covariance Matrix for Quadratic Trend Model

1	0.6365	0.5673	0.4355
0.6365	1	0.6559	0.6817
0.5673	0.6559	1	0.7387
0.4355	0.6817	0.7387	1

Table 6.10: Unstructured Correlation Matrix for Quadratic Trend Model

From an examination of the diagonal elements of the matrix in Table 6.9 it appears the assumption of homogeneous variances is reasonable. As a result, equal variances at the four measurement occasions will be assumed. Table 6.11 displays the relevant elements for the covariance and correlation structures for simple, compound symmetric, autoregressive of order 1, and Toeplitz structures. In Table 6.11, within each structure, the top line shows the variance-covariance components while the bottom line shows the correlation components, both as a function of time lag (i.e., the entry for Week 0 is the variance while successive entries are the covariances with Week 0). The complete covariance and correlation matrices for the *covariance pattern models* can easily be computed using the structure formulas from Section 6.3.2.

Covariance/Correlation Structure	**Week 0**	**Week 4**	**Week 8**	**Week 12**
UN	0.2102	0.1218	0.0936	0.0832
	1	0.6365	0.5673	0.4355
Simple	0.1690	0	0	0
	1	0	0	0
CS	0.1714	0.1044	0.1044	0.1044
	1	0.6091	0.6091	0.6091
AR(1)	0.1781	0.1218	0.0833	0.0570
	1	0.6839	0.4677	0.3199
Toeplitz	0.1778	0.1222	0.1136	0.0810
	1	0.6874	0.6387	0.4557

Table 6.11: Variance-covariance and correlation components

Now, comparisons can be made between the structured covariance and correlation structures in Table 6.11 and the unstructured (sample) covariance and correlation

structures in Table 6.9 and Table 6.10. It is clear that the simple structure does not accurately reflect the trends in covariance and correlation when compared to the unstructured structure. The compound symmetric structure is quite similar at most measurement occasions in terms of the variance-covariance components. The correlation is a good reflection of the UN structure with the exception of Week 12. Similarly, the AR-1 structure also looks like a suitable fit for both variance-covariance and correlation components, again with the exception of Week 12. The structure that appears to fit best is the Toeplitz structure.

Alternatively, information criteria such as Akaike's Information Criterion can be compared for each structure. Table 6.12 compares the AIC, AICC, and BIC criteria for the structured and unstructured covariance models.

Covariance Structure	AIC	AICC	BIC
Simple	201.0	201.1	202.6
CS	145.2	145.2	148.3
AR(1)	144.8	144.9	148.0
Toeplitz	141.5	141.8	147.7
Unstructured	147.8	149.6	163.3

Table 6.12: Comparisons of Information Criterion

For each information criterion, the covariance structure corresponding to the smallest value is the best fit for the data. The various information criteria will not always select the same model. In this particular example, AIC, AICC and BIC all select the Toeplitz structure as the best fit.

Littell, Pendergast, and Natarajan (2000) describe a graph called a 'correlogram' that can also be useful when choosing an appropriate covariance structure. A correlogram is a plot of the correlation between the first measurement occasion and each successive measurement occasion. In other words, the correlation components for the covariance

structures from Table 6.11 are plotted against the measurement occasions. Figure 6.5 displays correlograms that compare the unstructured correlation structure to each of compound symmetric, first-order auto-regressive and Toeplitz correlation structures.

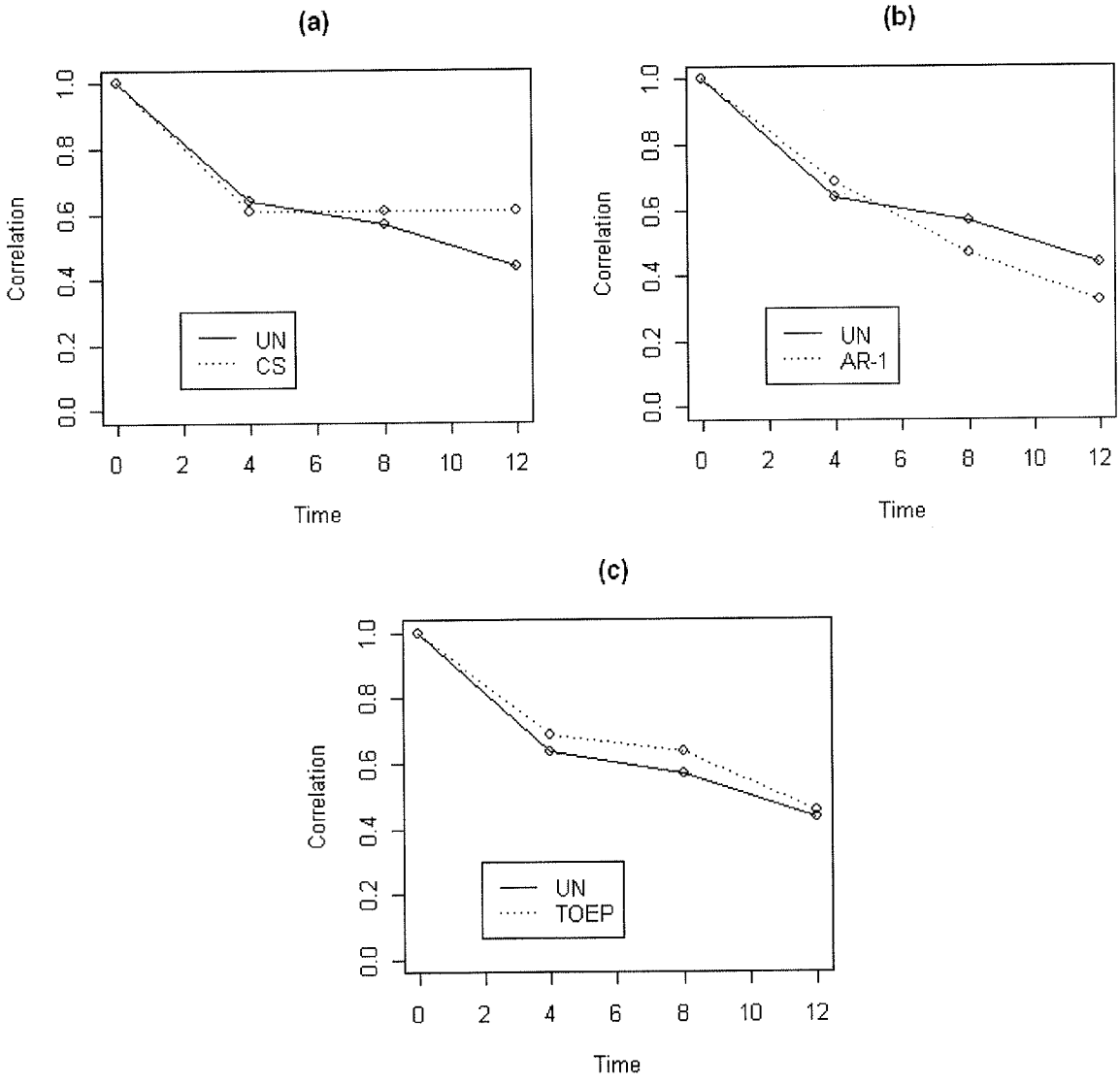

Figure 6.5: (a) UN vs. CS (b) UN vs. AR-1 (c) UN vs Toeplitz

Referring to Figure 6.5, it appears the Toeplitz correlation structure matches the Unstructured correlation structure best. It is especially evident by the correlograms that the compound symmetric structure provides a poor fit at Week 12. As shown in Figure

6.5, the correlogram can be a useful tool when choosing amongst covariance structures for the best fit.

Next, the results of the tests for the fixed effects (i.e., group, time, time× time, group × time interaction and group × time × time interaction) will be examined to see how the estimates of the *F*-statistics are affected when different covariance structures are fit to the data as compared to the unstructured covariance structure. Table 6.13 displays the fixed effects results.

Covariance Structure	Group	Time	Time×Time	Group×Time	Group×Time×Time
Simple	3.84	7.24	6.53	1.22	1.54
CS	3.66	18.26	16.48	3.07	3.89
AR(1)	3.08	18.71	17.70	2.76	4.18
Toeplitz	3.70	25.09	25.29	4.41	5.98
Unstructured	2.88	20.86	25.83	4.83	7.77

Table 6.13: *F*-statistics for fixed effects

For the simple covariance structure, *F*-values differ substantially. The most extreme differences in *F*-values for both the compound symmetric and autoregressive structures when compared to the unstructured covariance lie in the interaction effects. The Toeplitz structure yields similar estimates for all of the effects when compared to the estimates due to the unstructured covariance. Based on the comparisons of (a) the variance-covariance and correlation components, (b) the correlograms, (c) the information criterion, and (d) the fixed effect estimates as compared to the unstructured covariance structure, it can be concluded that the Toeplitz structure is adequate for modeling the covariance structure.

Given the quadratic trend model for the mean with a Toeplitz covariance structure, the residuals can also be examined. First, a histogram of the transformed (or scaled)

residuals followed by a histogram of the raw residuals will be examined in Figure 6.6 (a) and (b) respectively. The raw residuals are included as a comparison to the residuals transformed by the Cholesky decomposition method.

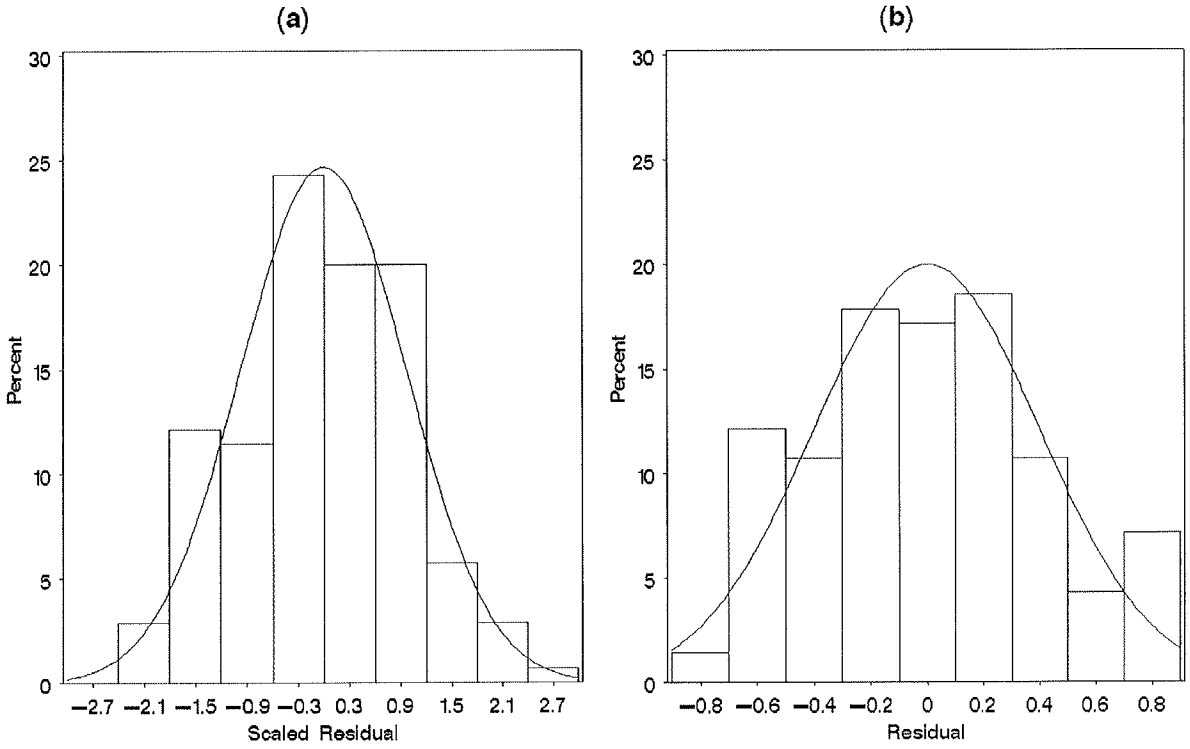

Figure 6.6: Quadratic Trend Model: (a) Histogram of Transformed Residuals (b) Histogram of Untransformed Residuals

As shown in Figure 6.6, the histogram of the transformed residuals appears to be normally distributed.

Next, normal quantile plots (or QQ plots) for the transformed and raw residuals are shown in Figure 6.7 (a) and (b) respectively.

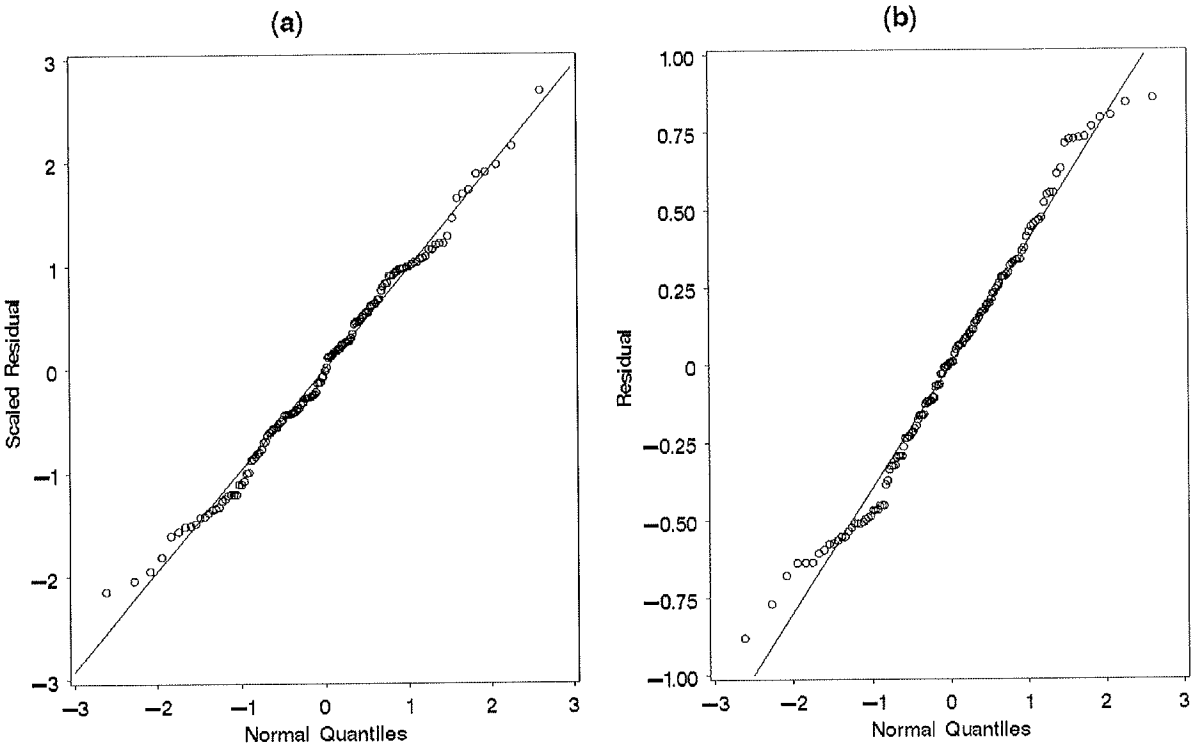

Figure 6.7: Quadratic Trend Model: (a) QQ Plot of Transformed Residuals (b) QQ Plot of Untransformed Residuals

By examining the transformed residuals from the QQ plot in Figure 6.7(a), they appear to be normally distributed. The quadratic trend model for the mean with a Toeplitz covariance structure appears to be adequate for modeling the Vitamin C data. As a final check for normality, the residuals plotted against the predicted values are shown in Figure 6.8.

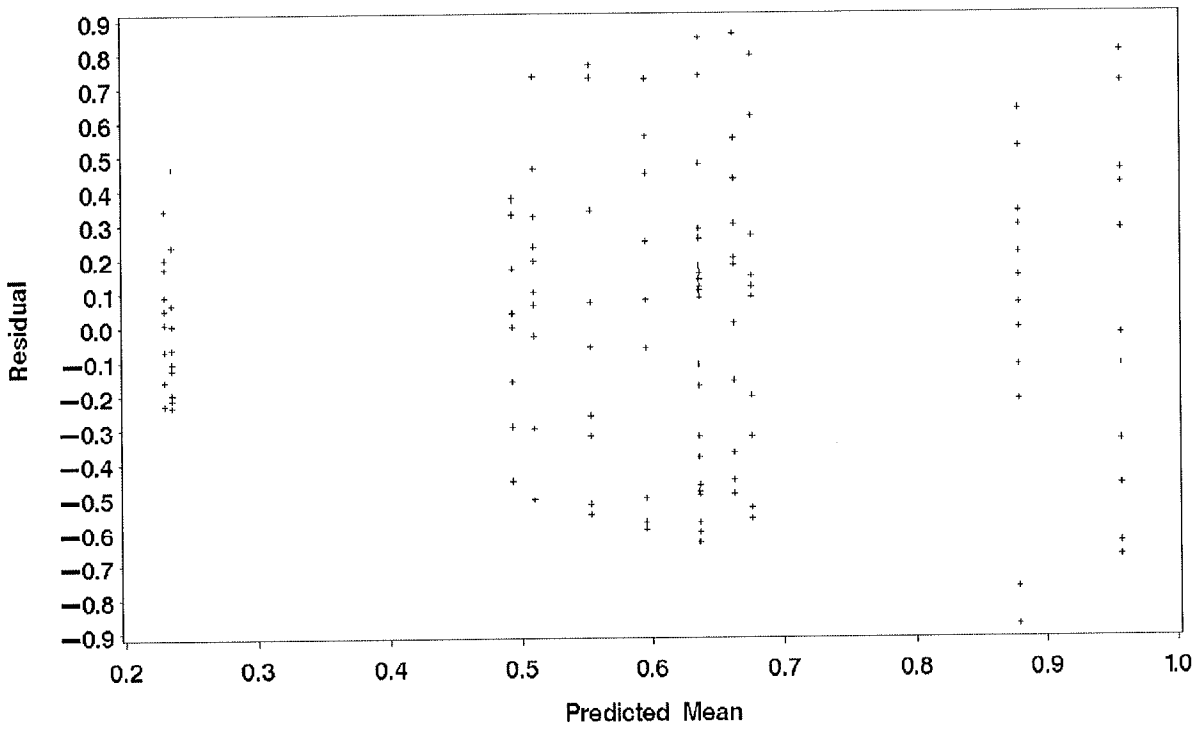

Figure 6.8: Residuals vs Predicted values for Vitamin C

As displayed by Figure 6.8, within each group, each measurement occasion has the same predicted mean thus leading to the striated appearance of the points. Initially the odd looking graph was thought to be a result of heterogeneous variances across the groups. Levene's test of homogeneity of variances was conducted which indicated that the hypothesis of heterogeneous covariances was correct. A modification in the SAS code was made to account for this variance heterogeneity (Littell, Milliken, Wolfinger & Schabenberger, 2006). The histograms, QQ plots and residuals versus the predicted plots were all adjusted accordingly but the new graphs appeared to look the same.

The results for this final model for the fixed effects are shown in Table 6.14 below.

Effect	Df	*Wald*	*p*-value
Group	2	7.40	0.0247
Time	1	25.09	<.0001
Time × Time	1	25.29	<.0001
Group × Time	2	8.83	0.0121
Group × Time × Time	2	11.95	0.0025

Table 6.14: Fixed Effects results for Vitamin C

Based on the fixed effects results in Table 6.14, we can conclude that there is a difference between the groups that changes depending on time.

6.5.3 Example 2: Modeling Vitamin E via Parametric Curve

Next, the Vitamin E data will be examined but in less detail than the Vitamin C example just discussed. The mean and corresponding standard errors for the Vitamin E data are shown in Table 6.15.

		Week 0	Week 4	Week 8	Week 12
Control	**Mean**	15.01	13.29	12.02	12.52
	Std Err	2.34	1.97	1.68	1.29
	n	10	10	10	10
Milled Flax	**Mean**	12.70	12.57	16.04	7.11
	Std Err	1.32	1.08	1.70	1.37
	n	13	13	13	13
Flaxseed Oil	**Mean**	11.81	10.14	12.07	10.40
	Std Err	0.84	0.76	0.62	0.95
	n	12	11	12	12

Table 6.15: Descriptive Statistics for Vitamin E data

It appears as though the mean Vitamin E levels are decreasing for the control group as time progresses. The milled flax group exhibits a substantial increase in Vitamin E from Week 4 to Week 8, and then drops approximately nine units from Week 8 to Week 12. Vitamin E levels for the flaxseed oil group are comparable for all of the measurement occasions with the exception of Week 8. It is not clear which model might provide an appropriate fit for these data.

Accompanying the raw descriptive statistics for the Vitamin E data is a time plot of the sample means from Table 6.15. The time plot in Figure 6.9 particularly emphasizes the drop in the Vitamin E level for the milled flax group after Week 8.

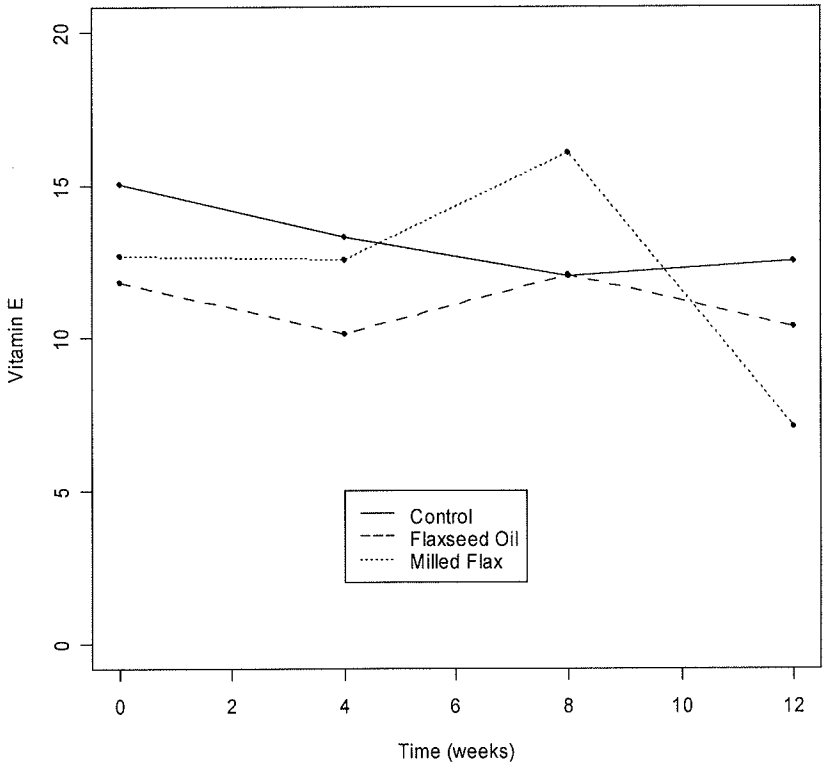

Figure 6.9: Sample means of Vitamin E data

Next, a linear trend model was fit to the data with an unstructured covariance. Following, both quadratic and cubic trend models with unstructured covariances were also fit to the data. The best model amongst the linear, quadratic and cubic trend models according to the likelihood ratio test was the cubic trend model.

Using the cubic trend model for the mean, various covariance structures were tested. The compound symmetric structure was the most appropriate as denoted by the lowest AIC value. Table 6.16 displays the estimates for the regression coefficients when a cubic trend mean model with a CS covariance structure was fit to the Vitamin E data.

Effect		Estimate	Std Err	Test Statistic	*p*-value
Intercept		15.01	1.47	10.18	<.0001
Group	MF	–2.31	1.96	–1.18	0.2477
Group	FO	–3.20	2.00	–1.60	0.1187
Time		–0.38	0.97	–0.39	0.6988
Time×Time		–0.03	0.21	–0.13	0.8988
Time×Time×Time		0.00	0.01	0.29	0.7706
Time×Group	MF	–1.44	1.29	–1.12	0.2665
Time×Group	FO	–1.17	1.33	–0.88	0.3810
Time×Time×Group	MF	0.64	0.28	2.25	0.0268
Time×Time×Group	FO	0.38	0.29	1.30	0.1970
Time×Time×Time×Group	MF	–0.05	0.02	–2.89	0.0048
Time×Time×Time×Group	FO	–0.02	0.02	–1.43	0.1558

Table 6.16: Estimates and Standard Errors for Vitamin E

Based on Table 6.16, the observed and estimated means for each treatment group were calculated and are shown below.

		Week 0	Week 4	Week 8	Week 12
Control	**Observed**	15.008	13.292	12.022	12.515
	Estimated	15.008	13.292	12.022	12.515
Milled Flax	**Observed**	12.699	12.573	16.035	7.113
	Estimated	12.699	12.572	16.032	7.104
Flaxseed Oil	**Observed**	11.808	10.138	12.069	10.399
	Estimated	11.808	10.040	12.073	10.409

Table 6.17: Observed and Estimated Means for Vitamin E

The estimated means derived from this model correspond well with the observed means. Although only three decimal places are shown, discrepancies between the observed and estimated means were often observed when the means were extended to the fifth or sixth decimal place. Using this final model, the cubic mean model with a CS covariance structure, the fixed effects results are shown in Table 6.18 below.

Effect	Test Statistic	*p*-value
Group	2.69	0.2606
Time	5.60	0.0179
Time×Time	7.27	0.0070
Time×Time×Time	9.15	0.0025
Time×Group	1.35	0.5085
Time×Time×Group	5.06	0.0795
Time×Time×Time×Group	8.36	0.0153

Table 6.18: Fixed effects results for Vitamin E

The Time, Time^2 and Time^3 effects are all significant as indicated by *p*-values of 0.0179, 0.0070 and 0.0025, respectively. More importantly, the Group × Time^3 interaction reveals a significant *p*-value of 0.0153. This indicates that there may be a difference between the groups in terms of Vitamin E levels that change at different rates depending on 'Time'. Referring back to the sample means plotted for each group across time in Figure 6.9, it appears that the cubic trend model best fits the milled flax and flaxseed oil groups while a quadratic trend model may have been an adequate model for the control group.

Chapter 7

Linear Mixed Effects Regression Models

7.1 Introduction

Linear mixed effects regression models are extremely useful in longitudinal studies as they enable researchers to examine response trends with more flexible models. In the current literature on longitudinal data, these models are also referred to as random effects models, multi-level models, hierarchical linear models, two-stage models and random coefficient models to name a few (Hedeker & Gibbons, 2006). By using a mixed model, the investigator can take into account the natural heterogeneity that exists among subjects. To study how the response profiles of *specific* individuals change over time, both fixed and random effects are included in the mean model. The fixed effects are the population parameters denoted by $\boldsymbol{\beta}$, which are common across individuals. The random effects are simply the individual-specific effects that, when combined with the

fixed effects, depict the mean response profile for each subject. Due to the incorporation of random effects in the model, covariance among the repeated measures arises (Fitzmaurice et al. 2004).

An advantage of the linear mixed effects regression model worth mentioning is that the data do not need to be balanced. One consequence of this feature is that individuals with missing responses are still included. This leads to improved statistical power and reduced bias. Furthermore, measurement occasions can be varied between individuals because the model incorporates time in a continuous manner. Both between-subject and within-subject covariates can be included in the model. As a result, determining how the covariates are related to the response can be studied. Also, estimates of change for each individual can easily be calculated (Fitzmaurice et al., 2004)

7.2 Linear Mixed Effects Model

The linear mixed model for the i^{th} subject ($i = 1, \ldots, n$) is written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \tag{7.1}$$

Here, $\boldsymbol{Y}_i$ denotes the response vector of dimension $n_i \times 1$. This model separates the fixed and random components. $\boldsymbol{\beta}$ is a $p \times 1$ vector that contains the fixed effects while $\boldsymbol{X}_i$ is a $n_i \times p$ design matrix associated with $\boldsymbol{\beta}$. For the random component of the model, $\boldsymbol{b}_i$ is a $q \times 1$ vector that contains the random effects while $\boldsymbol{Z}_i$ is a $n_i \times q$ design matrix associated with $\boldsymbol{b}_i$. The vector $\boldsymbol{e}_i$ denotes the random errors of dimension $n_i \times 1$.

Model Assumptions and Properties

1) Both vectors $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$ are assumed to exhibit a multivariate normal distribution and are independent of each other.

2) $\mathrm{E}(\boldsymbol{b}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{b}_i) = \mathbf{G}$

3) $\mathrm{E}(\boldsymbol{e}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{e}_i) = \mathbf{R}_i$

4) The columns that comprise $\boldsymbol{Z}_i$ are a subset of the columns that comprise $\boldsymbol{X}_i$. This property allows components of $\boldsymbol{\beta}$ (which are specified explicitly in $\boldsymbol{Z}_i$) to vary in a random fashion.

Usually it is assumed that the e_{ij}'s are uncorrelated, thus $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i}$. It is possible to assign a covariance structure such as those considered in Chapter 6 but this will not be explored in this practicum due to length.

Next, let us examine the mean and covariance of the $\boldsymbol{Y}_i$ vector. Firstly, the mean and covariance can be expressed *conditionally*, given the random effects $\boldsymbol{b}_i$. The mean is denoted by

$$\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i$$

while the covariance is

$$\mathrm{Cov}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i) = \mathrm{Cov}(\boldsymbol{e}_i) = \boldsymbol{R}_i .$$

Notice that the mean response for the i^{th} subject contains both fixed and random effects.

Alternatively, the mean and covariance can also be expressed *marginally*, averaged over the random effects $\boldsymbol{b}_i$. The mean is

$$\begin{aligned}
\mathrm{E}(\boldsymbol{Y}_i) &= \mathrm{E}\{\mathrm{E}(\boldsymbol{Y}_i \mid \boldsymbol{b}_i)\} \\
&= \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) \\
&= \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta}) + \mathrm{E}(\boldsymbol{Z}_i\boldsymbol{b}_i) \\
&= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathrm{E}(\boldsymbol{b}_i) \\
&= \boldsymbol{X}_i\boldsymbol{\beta}
\end{aligned}$$

while the covariance is

$$\begin{aligned}\text{Cov}(\boldsymbol{Y}_i) &= \text{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) + \text{Cov}(\boldsymbol{e}_i) \\ &= \boldsymbol{Z}_i\text{Cov}(\boldsymbol{b}_i)\boldsymbol{Z}_i' + \boldsymbol{R}_i \\ &= \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i\end{aligned}$$

Notice that the marginal mean response for the i^{th} subject does not contain the random effects $\boldsymbol{b}_i$ (Fitzmaurice et al., 2004; Laird & Ware, 1982; Schabenberger & Pierce, 2002).

7.3 Random Intercept Mixed Model

First let's introduce a mixed model that contains a random subject effect only. The model is adapted from Fitzmaurice, Laird and Ware (2004) and is

$$\text{Y}_{ij} = \beta_1 + \text{b}_{1i} + \beta_2 t_{ij} + \text{e}_{ij} \tag{7.2}$$

for the i^{th} subject at the j^{th} occasion. Here, $\beta_1 + \text{b}_{1i}$ represents the intercept for the i^{th} individual and is comprised of the fixed intercept β_1 common to all individuals plus the random subject effect b_{1i} which is unique to each individual. Furthermore,

$$\text{E}(\text{b}_{1i}) = 0\,; \;\; \text{Var}(\text{b}_{1i}) = \sigma_{\text{b}}^2, \text{ and}$$

$$\text{E}(e_{ij}) = 0\,; \;\; \text{Var}(e_{ij}) = \sigma^2$$

The e_{ij}'s are no longer independent of one another. Instead, the errors are conditionally independent given b_{1i}.

The conditional and marginal means of Y_{ij} are respectively

$$\text{E}(\text{Y}_{ij} \mid \text{b}_{1i}) = \beta_1 + \text{b}_{1i} + \beta_2 t_{ij}, \text{ and}$$

$$\text{E}(\text{Y}_{ij}) = \beta_1 + \beta_2 t_{ij}$$

while the marginal variance and covariance of Y_{ij} are respectively

$$\begin{aligned}\mathrm{Var}(\mathrm{Y}_{ij}) &= \mathrm{Var}(\beta_1 + \mathrm{b}_{1i} + \beta_2 t_{ij} + e_{ij}) \\ &= \mathrm{Var}(\mathrm{b}_{1i} + e_{ij}) \qquad \text{, and} \\ &= \sigma_{\mathrm{b}}^2 + \sigma^2\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}(\mathrm{Y}_{ij}, \mathrm{Y}_{ij'}) &= \mathrm{Cov}(\beta_1 + \mathrm{b}_{1i} + \beta_2 t_{ij} + e_{ij}, \beta_1 + \mathrm{b}_{1i} + \beta_2 t_{ij'} + e_{ij'}) \\ &= \mathrm{Cov}(\mathrm{b}_{1i} + e_{ij}, \mathrm{b}_{1i} + e_{ij'}) \\ &= \mathrm{Var}(\mathrm{b}_{1i}) \\ &= \sigma_{\mathrm{b}}^2.\end{aligned}$$

Thus the covariance structure for $\boldsymbol{Y}_i$ is compound symmetric and is

$$\mathrm{Cov}(\boldsymbol{Y}_i) = \begin{pmatrix} \sigma_{\mathrm{b}}^2 + \sigma^2 & \sigma_{\mathrm{b}}^2 & \cdots & \sigma_{\mathrm{b}}^2 \\ \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{b}}^2 + \sigma^2 & \cdots & \sigma_{\mathrm{b}}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\mathrm{b}}^2 & \sigma_{\mathrm{b}}^2 & \cdots & \sigma_{\mathrm{b}}^2 + \sigma^2 \end{pmatrix}.$$

Figure 7.1 shows the conditional means and measurement errors of two individuals, 'A' and 'B', as well as the marginal mean, or population trend for a random intercept model.

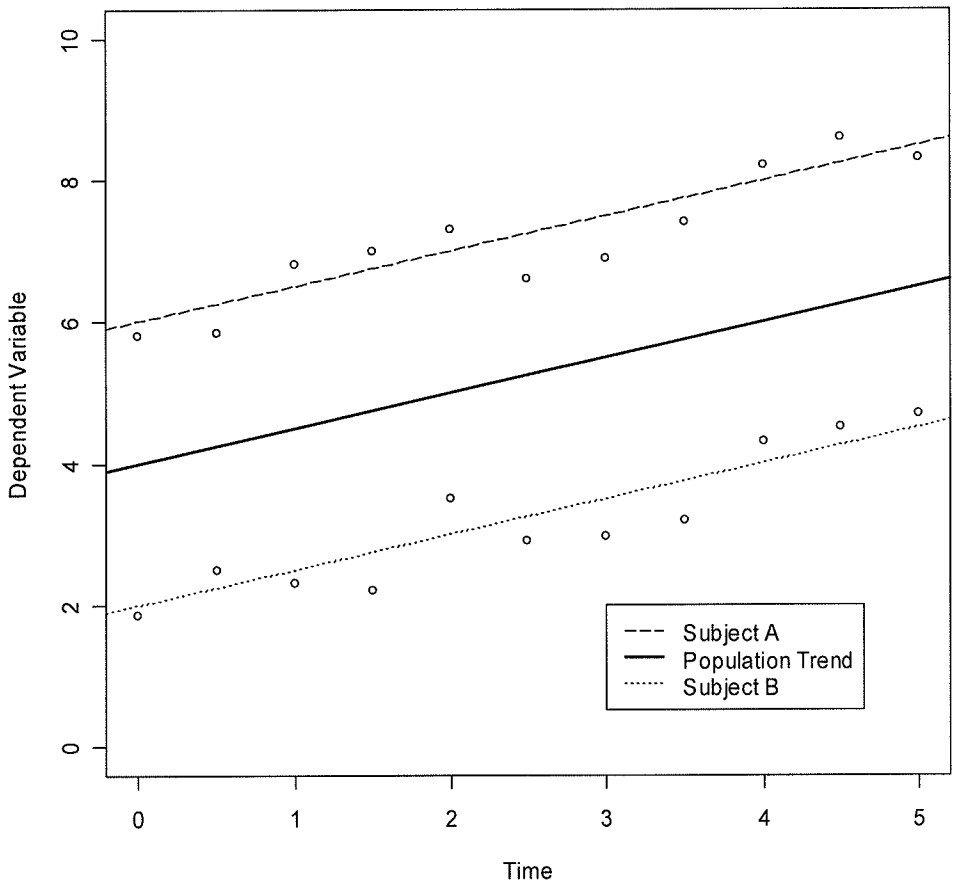

Figure 7.1: Random Intercept Mixed Model

Let us write model 7.2 in terms of the current study where there are three treatment groups. It must be modified as follows.

$$\begin{aligned} Y_{ijkl} = \beta_1 + b_{1ikl} + \beta_2 t_{ij} + \beta_3(\text{Group}_k) + \beta_4(t_{ij} \times \text{Group}_k) + \\ \beta_5(\text{Group}_l) + \beta_6(t_{ij} \times \text{Group}_l) + e_{ijkl} \end{aligned} \tag{7.3}$$

where $\text{Group}_k = \begin{cases} 1 & \textit{if subject belongs to the milled flax group} \\ 0 & \textit{otherwise} \end{cases}$

and $\text{Group}_l = \begin{cases} 1 & \textit{if subject belongs to the flaxseed oil group} \\ 0 & \textit{otherwise} \end{cases}$.

For the mixed model $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$, the design matrices for each group are specified as follows:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 & 0 & 0 \end{pmatrix}$$ for the control group (i.e., $k = l = 0$),

$$X_i = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 & 0 & 0 \\ 1 & 8 & 1 & 8 & 0 & 0 \\ 1 & 12 & 1 & 12 & 0 & 0 \end{pmatrix}$$ for the milled flax group (i.e., $k = 1$; $l = 0$), and

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 & 4 \\ 1 & 8 & 0 & 0 & 1 & 8 \\ 1 & 12 & 0 & 0 & 1 & 12 \end{pmatrix}$$ for the flaxseed oil group (i.e., $k = 0$; $l = 1$), while

$$Z_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$ for any group.

7.4 Random Intercept and Slope

Next, the framework for fitting a linear mixed model with a random intercept and slope for each individual is outlined. Referring to the current longitudinal study, the subjects belong to one of three treatment groups (control, milled flax, and flaxseed oil) and are measured at four occasions (0, 4, 8, and 12 weeks).

First, let us look at the general situation, specifically, a longitudinal study with a single group. The mixed model which allows for random intercepts and slopes for all subjects is written as

$$Y_{ij} = \beta_1 + b_{1i} + (\beta_2 + b_{2i}) t_{ij} + e_{ij} \qquad (7.4)$$

for the i^{th} subject at the j^{th} occasion. Here, $\beta_1 + \mathrm{b}_{1i}$ is the intercept for the i^{th} individual (i.e., β_1 is the population intercept and b_{1i} is the random subject effect for individual i). The slope for the i^{th} individual is $\beta_2 + \mathrm{b}_{2i}$ (i.e., β_2 is the population slope and b_{2i} is the random subject effect for individual i). The e_{ij}'s are normally distributed with mean 0 and variance σ^2 and are conditionally independent given b_{1i} and b_{2i}. The intercept and slope are bivariate normally distributed with mean $\mathbf{0}$ and covariance $\sum_b$ where

$$\sum_{\mathrm{b}} = \begin{pmatrix} \sigma^2_{\mathrm{b}_1} & \sigma_{\mathrm{b}_1\mathrm{b}_2} \\ \sigma_{\mathrm{b}_1\mathrm{b}_2} & \sigma^2_{\mathrm{b}_2} \end{pmatrix}.$$

The conditional and marginal means of Y_{ij} are respectively

$$\mathrm{E}(\mathrm{Y}_{ij} \mid \mathrm{b}_{1i}) = \beta_1 + \mathrm{b}_{1i} + (\beta_2 + \mathrm{b}_{2i})t_{ij}, \text{ and}$$

$$\mathrm{E}(\mathrm{Y}_{ij}) = \beta_1 + \beta_2 t_{ij}$$

while the marginal variance and covariance of Y_{ij} are respectively

$$\begin{aligned} \mathrm{Var}(\mathrm{Y}_{ij}) &= \mathrm{Var}(\beta_1 + \mathrm{b}_{1i} + (\beta_2 + \mathrm{b}_{2i})t_{ij} + e_{ij}) \\ &= \mathrm{Var}(\mathrm{b}_{1i} + \mathrm{b}_{2i}\, t_{ij} + e_{ij}) \qquad \text{, and} \\ &= \sigma^2_{\mathrm{b}_1} + 2t_{ij}\sigma_{\mathrm{b}_1\mathrm{b}_2} + t_{ij}{}^2\sigma^2_{\mathrm{b}_2} + \sigma^2 \end{aligned}$$

$$\begin{aligned} \mathrm{Cov}(\mathrm{Y}_{ij}, \mathrm{Y}_{ij'}) &= \mathrm{Cov}(\mathrm{b}_{1i} + \mathrm{b}_{2i}\, t_{ij} + e_{ij}, \mathrm{b}_{1i} + \mathrm{b}_{2i}\, t_{ij'} + e_{ij'}) \\ &= \sigma^2_{\mathrm{b}_1} + (t_{ij} + t_{ij'})\sigma_{\mathrm{b}_1\mathrm{b}_2} + t_{ij}\, t_{ij'}\, \sigma^2_{\mathrm{b}_2}. \end{aligned}$$

Figure 7.2 displays the conditional means and generated observations of two individuals, 'A' and 'B', as well as the population response trend. Notice the varying intercepts and slopes in the graph.

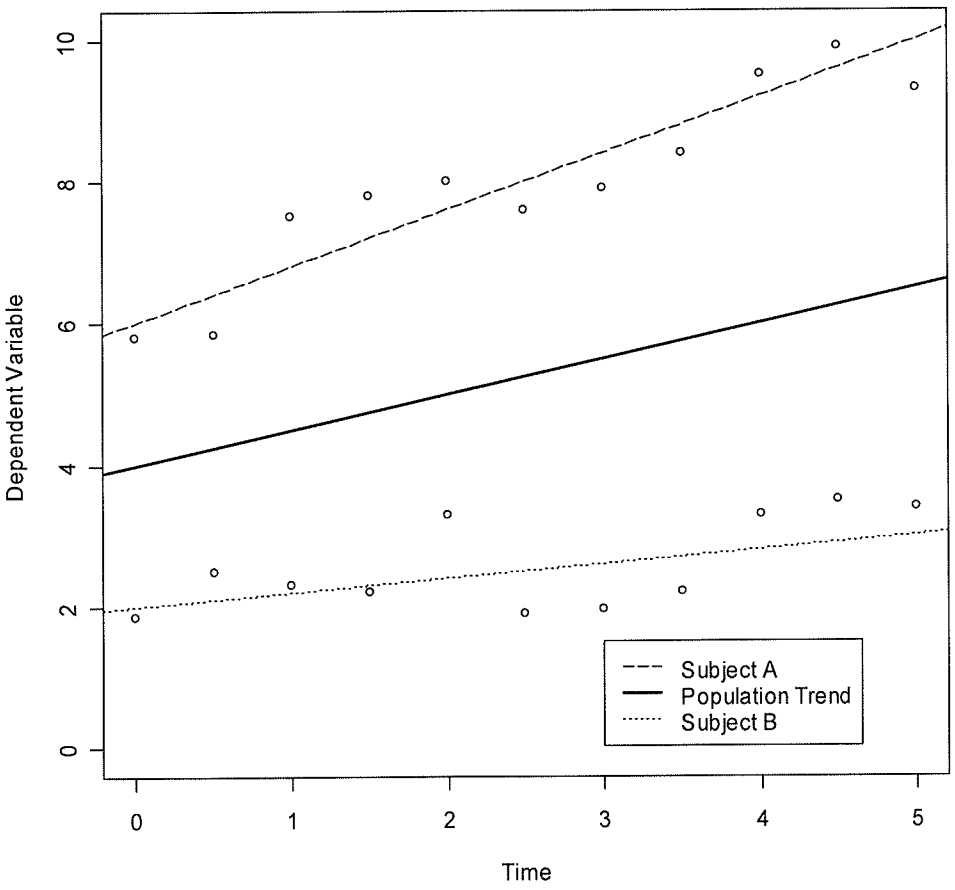

Figure 7.2: Random Intercept and Slope Model

If there were two groups, we could write this equation in terms of the mixed model $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$ in matrix form where

$$\boldsymbol{X}_i = \boldsymbol{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ 1 & t_{i3} \\ 1 & t_{i4} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{pmatrix}.$$

Model 7.4 does not take into account that there are three treatment groups in the study. It must be modified as follows.

$$\begin{aligned} \mathrm{Y}_{ijkl} = \beta_1 + \mathrm{b}_{1ikl} + (\beta_2 + \mathrm{b}_{2ikl})t_{ij} + \beta_3(\mathrm{Group}_k) + \beta_4(t_{ij} \times \mathrm{Group}_k) + \\ \beta_5(\mathrm{Group}_l) + \beta_6(t_{ij} \times \mathrm{Group}_l) + e_{ijkl} \end{aligned} \tag{7.5}$$

where $\text{Group}_k = \begin{cases} 1 & \textit{if subject belongs to the milled flax group} \\ 0 & \textit{otherwise} \end{cases}$

and $\text{Group}_l = \begin{cases} 1 & \textit{if subject belongs to the flaxseed oil group} \\ 0 & \textit{otherwise} \end{cases}$.

Notice how model (7.5) is similar to the linear trend model (6.1) in Section 6.2.2.1 but with the addition of the random effects b_{1ikl} and b_{2ikl}. In order to write this model in terms of the mixed model $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$, the design matrices for each group are specified as follows:

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 & 0 & 0 \end{pmatrix}$$ for the control group (i.e., $k = l = 0$),

$$X_i = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 & 0 & 0 \\ 1 & 8 & 1 & 8 & 0 & 0 \\ 1 & 12 & 1 & 12 & 0 & 0 \end{pmatrix}$$ for the milled flax group (i.e., $k = 1$; $l = 0$), and

$$X_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 & 4 \\ 1 & 8 & 0 & 0 & 1 & 8 \\ 1 & 12 & 0 & 0 & 1 & 12 \end{pmatrix}$$ for the flaxseed oil group (i.e., $k = 0$; $l = 1$), while

$$Z_i = \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{pmatrix}$$ for any group.

In matrix form, the mixed model for the control group is written as

$$\begin{pmatrix} Y_{i100} \\ Y_{i200} \\ Y_{i300} \\ Y_{i400} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{pmatrix} \begin{pmatrix} b_{1i00} \\ b_{2i00} \end{pmatrix} + \begin{pmatrix} e_{i100} \\ e_{i200} \\ e_{i300} \\ e_{i400} \end{pmatrix}$$

$$= \begin{pmatrix} (\beta_1 + b_{1i00}) + e_{i100} \\ (\beta_1 + b_{1i00}) + (\beta_2 + b_{2i00}) \times 4 + e_{i200} \\ (\beta_1 + b_{1i00}) + (\beta_2 + b_{2i00}) \times 8 + e_{i300} \\ (\beta_1 + b_{1i00}) + (\beta_2 + b_{2i00}) \times 12 + e_{i400} \end{pmatrix}.$$

For the milled flax group, the mixed model in matrix form is

$$\begin{pmatrix} Y_{i110} \\ Y_{i210} \\ Y_{i310} \\ Y_{i410} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 4 & 0 & 0 \\ 1 & 8 & 1 & 8 & 0 & 0 \\ 1 & 12 & 1 & 12 & 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{pmatrix} \begin{pmatrix} b_{1i10} \\ b_{2i10} \end{pmatrix} + \begin{pmatrix} e_{i110} \\ e_{i210} \\ e_{i310} \\ e_{i410} \end{pmatrix}$$

$$= \begin{pmatrix} (\beta_1 + \beta_3 + b_{1i10}) + e_{i110} \\ (\beta_1 + \beta_3 + b_{1i10}) + (\beta_2 + \beta_4 + b_{2i10}) \times 4 + e_{i210} \\ (\beta_1 + \beta_3 + b_{1i10}) + (\beta_2 + \beta_4 + b_{2i10}) \times 8 + e_{i310} \\ (\beta_1 + \beta_3 + b_{1i10}) + (\beta_2 + \beta_4 + b_{2i10}) \times 12 + e_{i410} \end{pmatrix}.$$

Finally, for the flaxseed oil group, the mixed model is written as

$$\begin{pmatrix} Y_{i101} \\ Y_{i201} \\ Y_{i301} \\ Y_{i401} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 1 & 4 \\ 1 & 8 & 0 & 0 & 1 & 8 \\ 1 & 12 & 0 & 0 & 1 & 12 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \end{pmatrix} \begin{pmatrix} b_{1i01} \\ b_{2i01} \end{pmatrix} + \begin{pmatrix} e_{i101} \\ e_{i201} \\ e_{i301} \\ e_{i401} \end{pmatrix}$$

$$= \begin{pmatrix} (\beta_1 + \beta_5 + b_{1i01}) + e_{i101} \\ (\beta_1 + \beta_5 + b_{1i01}) + (\beta_2 + \beta_6 + b_{2i01}) \times 4 + e_{i201} \\ (\beta_1 + \beta_5 + b_{1i01}) + (\beta_2 + \beta_6 + b_{2i01}) \times 8 + e_{i301} \\ (\beta_1 + \beta_5 + b_{1i01}) + (\beta_2 + \beta_6 + b_{2i01}) \times 12 + e_{i401} \end{pmatrix}.$$

7.5 Prediction of Random Effects

In the next section, the prediction of the random or individual-specific effects, $\boldsymbol{b}_i$ will be discussed. This material is adapted from Fitzmaurice, Laird and Ware (2004) and Littell, Milliken, Stroup, Wolfinger and Schabenberger (2006). There are two main reasons why this may be done: (a) It may be useful to predict the response trajectories for specific individuals, and (b) it may be useful to examine which individuals have extreme deviations in the response over time.

If $\boldsymbol{G}$ and $\boldsymbol{\Sigma}_i$ are known, the best linear unbiased predictor, otherwise known as the BLUP, of $\boldsymbol{b}_i$ is

$$\mathrm{E}(\boldsymbol{b}_i \mid \boldsymbol{Y}_i) = \boldsymbol{G}\boldsymbol{Z}_i\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$. The empirical BLUP of $\boldsymbol{b}_i$ is conditional on $\boldsymbol{Y}_i$ and depends on the unknown covariance matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{G}$. For this reason, $\boldsymbol{\Sigma}_i$ and G are replaced by estimates based on REML. Now, the predictor of $\boldsymbol{b}_i$ is

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{G}}\boldsymbol{Z}_i'\hat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}),$$

and is known as the empirical BLUP.

For the i^{th} subject, the predicted response trajectory is therefore

$$\hat{\boldsymbol{Y}}_i = \boldsymbol{X}_i\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i. \tag{7.6}$$

The topic of the prediction of random effects would not be complete without discussing 'shrinkage'. Using the empirical BLUP, the predicted response trajectory for the i^{th} individual can be 'shrunk' towards the mean response profile that is averaged over the population. This is done by modifying the predicted response trajectory for the i^{th} individual to be weighted by $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$. That is,

$$\hat{\boldsymbol{Y}}_i = (\hat{\boldsymbol{R}}_i \hat{\boldsymbol{\Sigma}}_i^{-1}) \boldsymbol{X}_i \hat{\boldsymbol{\beta}} + (\mathrm{I}_{n_i} - \hat{\boldsymbol{R}}_i \hat{\boldsymbol{\Sigma}}_i^{-1}) \boldsymbol{Y}_i . \tag{7.7}$$

The factors that affect the 'shrinkage' include $\boldsymbol{R}_i$ (within-subject variability), $\boldsymbol{\Sigma}_i$ (both within-subject and between-subject variability) and n_i (the number of repeated measurements for subject i). For example, when the within-subject variability is small compared to the between-subject variability, more weight is given to the observed response for the i^{th} individual. On the other hand, less weight is given to the observed response for the i^{th} individual when the relationship between the within and between-subject variability is opposite. Finally, when n_i is small, there is increased shrinkage towards the mean response trajectory for the population.

There are options using PROC MIXED in SAS to request the output of the BLUP's as well as the predicted values.

7.6 Examples

7.6.1 Random Intercept Model: Vitamin C data

Let us see how a random intercept model can be fit to the Vitamin C data from the study. First, recall the time plot of the Vitamin C values at each measurement occasion stratified by group in Figure 3.10. Upon examination of the baseline measurement for each group, there is evidence of between subject heterogeneity. Thus, fitting a random intercept model for the data makes sense.

Recall that the best fitting model for the Vitamin C data was a quadratic trend model fit for the mean with a Toeplitz covariance structure for the covariance. First, this model was run with the addition of a random intercept. An error message in SAS resulted. Perhaps the addition of the random intercept to this model was too complicated to be fit for the Vitamin C data. Referring back to Table 6.12 of information criteria values, the

next best choice of covariance structure for the Vitamin C data fit with a quadratic trend model for the mean was the AR(1) structure. This model was re-run in SAS with no errors. Table 7.1 displays the SAS output when the random intercept model was fit to the data. A brief explanation of the β parameters is included in the table to assist the reader with the interpretation.

Parameter		Estimate	Std Err	Z	p-value
β_1	Intercept	0.6985	0.1317	5.30	<.0001
β_2	Time	-0.0134	0.0317	-0.42	0.6721
β_3	Time^2	-0.0001	0.0025	-0.02	0.9850
β_4	Intercept Difference for MF	0.2504	0.1752	1.43	0.1561
β_5	Time Difference for MF	-0.0925	0.0421	-2.20	0.0304
β_6	Time^2 Difference for MF	0.0084	0.0033	2.56	0.0121
β_7	Intercept Difference for FO	-0.1772	0.1783	-0.99	0.3229
β_8	Time Difference for FO	-0.0906	0.0429	-2.11	0.0370
β_9	Time^2 Difference for FO	0.0084	0.0034	2.51	0.0137
$\sigma^2_{b_1}$		0.0886	0.0346	2.56	0.0052
σ^2		0.0855	0.0230	3.71	0.0001

Table 7.1: Random Intercept model for Vitamin C data

The observed versus the estimated means obtained from fitting the random intercept model to the Vitamin C data are displayed in Table 7.2. The model fits the data reasonably well. For comparison purposes, the observed versus the estimated means for the quadratic trend mean model with a Toeplitz covariance structure for the Vitamin C data are displayed in Table 7.3. Recall that the model fit to obtain the estimated values in Table 7.3 does *not* have a random intercept.

		Week 0	Week 4	Week 8	Week 12
Control	**Observed**	0.71	0.55	0.68	0.52
	Estimated	0.69	0.64	0.58	0.52
Milled Flax	**Observed**	0.94	0.69	0.60	0.89
	Estimated	0.95	0.66	0.64	0.89
Flaxseed Oil	**Observed**	0.53	0.19	0.27	0.47
	Estimated	0.52	0.29	0.33	0.63

Table 7.2: Observed and Estimated Means: Random Intercept for Vitamin C

		Week 0	Week 4	Week 8	Week 12
Control	**Observed**	0.71	0.55	0.68	0.52
	Estimated	0.68	0.64	0.60	0.55
Milled Flax	**Observed**	0.94	0.69	0.60	0.89
	Estimated	0.96	0.66	0.64	0.88
Flaxseed Oil	**Observed**	0.53	0.19	0.27	0.47
	Estimated	0.51	0.24	0.23	0.49

Table 7.3: Observed and Estimated Means: Quadratic Mean with Toeplitz Covariance for Vitamin C

Given that there are random effects in this model, the predicted values are unique for each individual. Thus, the plot of the residuals versus the predicted values will not have a striated appearance as in Chapter 6. Figure 7.3 shows this plot.

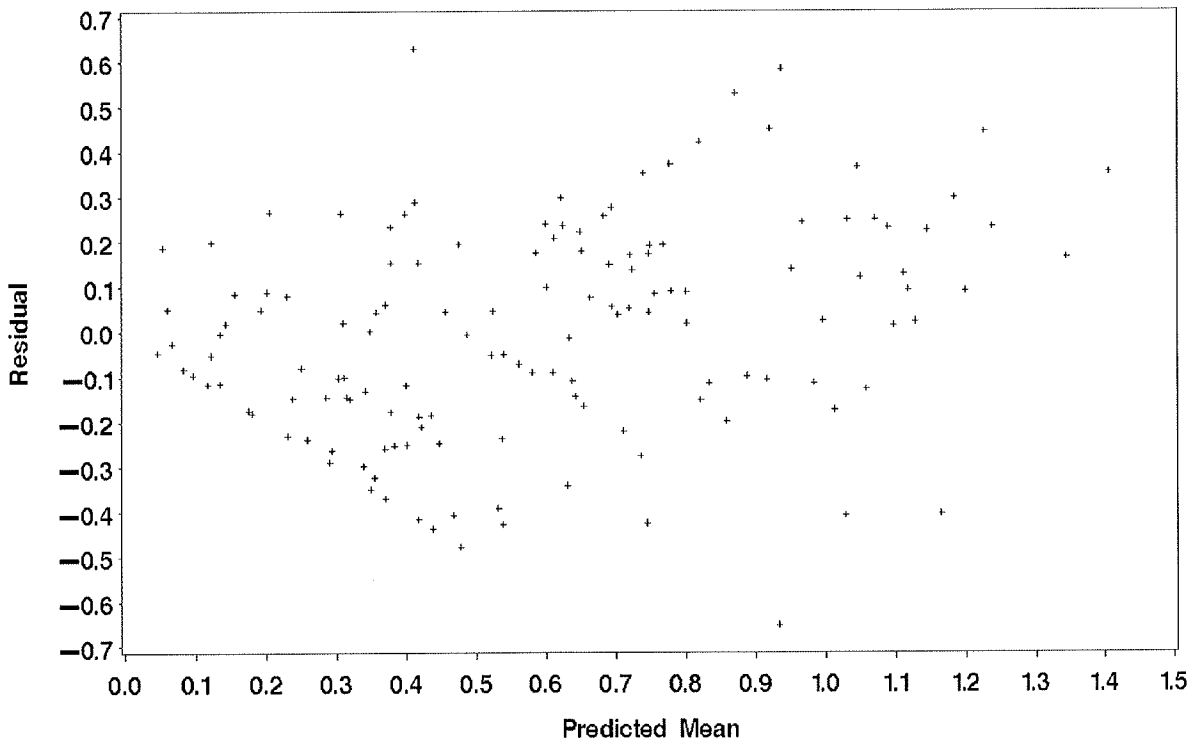

Figure 7.3: Residuals vs Predicted Values

Notice that the points in Figure 7.3 are scattered around zero indicating that the model provides a good fit for the data. The fixed effects results are shown in Table 7.4

Effect	Test Statistic	*p*-value
Group	6.66	0.0358
Time	19.14	<.0001
Time×Time	17.45	<.0001
Group×Time	5.98	0.0502
Group×Time×Time	8.25	0.0162

Table 7.4: Fixed Effects for Vitamin C data

The Group×Time×Time effect is significant in Table 7.4 indicating that there is a difference between the groups as a quadratic function of time.

7.6.2 Example – Vitamin C data with varying intercepts and slopes

Next we will fit a model, that has both intercepts and slopes that vary, to the quadratic trend model for the Vitamin C data from Chapter 6. Recall our attempt to

include a random intercept in the previous example which forced us to choose a different covariance structure for the errors. As suspected, when fitting the quadratic mean model with a Toeplitz structure, and a random intercept and slope, errors in SAS arose most likely due to the same reason previously. An AR-1 structure was switched with the Toeplitz structure and the model was re-run with no errors. The results are shown below.

Parameter		Estimate	Std Err	Z	p-value
β_1	Intercept	0.6914	0.1274	5.43	<.0001
β_2	Time	-0.0123	0.0302	-0.41	0.6881
β_3	Time^2	-0.0001	0.0024	-0.02	0.9843
β_4	Intercept Difference for MF	0.2602	0.1694	1.54	0.1293
β_5	Time Difference for MF	-0.0942	0.0402	-2.34	0.0222
β_6	Time^2 Difference for MF	0.0084	0.0031	2.69	0.0089
β_7	Intercept Difference for FO	-0.1736	0.1725	-1.01	0.3177
β_8	Time Difference for FO	-0.0912	0.0410	-2.23	0.0293
β_9	Time^2 Difference for FO	0.0084	0.0032	2.65	0.0102
$\sigma^2_{b_1}$		0.1003	0.0380	2.64	0.0041
$\sigma_{b_1 b_2}$		0.1577	0.2912	0.54	0.5880
$\sigma^2_{b_2}$		0.0003	0.0004	0.79	0.2156
σ^2		0.0634	0.0262	2.42	0.0077

Table 7.5: Random Intercept and Slope Model for Vitamin C

Let us look at the interpretation of the σ parameters in Table 7.5. The square root of $\sigma^2_{b_1}$ denotes the standard deviation for the population intercept (i.e., the estimate is $\sqrt{0.1003} = 0.32$). The square root of $\sigma^2_{b_2}$ denotes the standard deviation for the population slope (i.e., the estimate is $\sqrt{0.0003} = 0.02$). The resulting 95% confidence intervals are:

$0.6914 \pm (1.96 \times 0.32) = 0$ to 1.3186 for the intercept, and

$-0.0123 \pm (1.96 \times 0.02) = -0.0515$ to 0.0269 for the slope.

It is interesting to point out that the 95% confidence interval for the slope includes both positive and negative values. This is evidence of between-subject heterogeneity as some subjects have increasing Vitamin C levels over time while others have decreasing levels.

We can also request that SAS display the BLUP's and subsequently the predicted responses that are calculated using the BLUP's. More specifically, SAS displays the BLUP's of the random effects b_{1ikl} and b_{2ikl} in a table for each subject. A portion of the table is presented below.

Subject ID	Group	Effect	Estimate	Std Err	*t*	*p*-value
4208	MF	Intercept	0.4081	0.1670	2.44	0.0172
4208	MF	Time	-0.0154	0.0147	-1.04	0.3002
4563	MF	Intercept	0.0096	0.1670	0.06	0.9545
4563	MF	Time	-0.0118	0.0147	-0.80	0.4268
4756	MF	Intercept	-0.3954	0.1670	-2.37	0.0208
4756	MF	Time	-0.0131	0.0147	-0.89	0.3760
4845	MF	Intercept	0.1135	0.1670	0.68	0.4991
4845	MF	Time	0.0046	0.0147	0.32	0.7533
5041	MF	Intercept	-0.2810	0.1670	-1.68	0.0971
5041	MF	Time	0.0092	0.0147	0.62	0.5350
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Table 7.6: Empirical BLUP's for Vitamin C data

We can use the BLUP's to calculate the predicted Vitamin C values for each subject. The table below displays the actual observed value as well as the predicted value obtained using the BLUP's for a few subjects.

Subject ID	Group	Time	Actual	Predicted
4208	MF	0	1.67	1.36
4208	MF	4	1.52	1.01
4208	MF	8	0.81	1.92
4208	MF	12	0.76	1.11
4563	MF	0	1.37	0.96
4563	MF	4	0.29	0.62
4563	MF	8	0.52	0.55
4563	MF	12	0.66	0.75
4756	MF	0	0.49	0.56
4756	MF	4	0.17	0.21
4756	MF	8	0.31	0.14
4756	MF	12	0.00	0.33
.	.	.	.	.

Table 7.7: Actual vs. Predicted Vitamin C Response Values

Let's examine how a few of the predicted values are calculated referring to Table 7.5 for the fixed effects results and Table 7.6 for the random effects results. For the subject with ID = 4208 from the milled flax group, the predicted value at Week 0 is: (0.6914 + 0.2602 + 0.4081) = 1.36 which corresponds to the result in Table 7.7. As a second example, the individual with ID = 4563 from the milled flax group has a predicted value of 0.55 at Week 8. The calculation is (0.6914 + 0.2602 + 0.0096) + (–0.0123 – 0.0942 – 0.0118) ×8 + (–0.0001 + 0.0084) × 64 = 0.55. The rest of the predicted values can be calculated similarly.

The plot of the residuals versus the predicted values looks similar to the plot in Figure 7.3 so it will not be shown for this example.

7.6.3 Example – Vitamin E with random effects

Using the final model for Vitamin E from Chapter 6, which was a cubic trend model for the mean with a compound symmetric structure for the covariance, random effects were attempted to be added. Numerous errors resulted in SAS, namely convergence errors. Despite changing the covariance structure, no model that included a random

effect would converge. Again, this was most likely due to the complex nature of the mean model which had a cubic term combined with the small number of subjects.

Chapter 8

Missing Data

8.1 Introduction

Missing data are common in longitudinal studies especially in health research where subjects are humans. In this chapter, implications that missing data have on the methods for analyzing longitudinal data will be discussed.

8.2 Missing Data Patterns

Missing or incomplete data can follow either an intermittent or monotonic pattern. For instance, a subject that has missing observations at sporadic measurement occasions has an intermittent pattern of missingness. A subject exhibits a monotonic pattern of missingness if, when there is a missing observation at a particular measurement occasion, all observations at future occasions are missing. This latter pattern of missingness is also known as dropout.

8.3 Complications Due to Missing Data

The lack of observations at certain time points creates an unbalanced dataset. This can be problematic because some methods of analysis are based on the assumption that the data are balanced. When the data are incomplete, precision decreases, and potential bias can occur due to loss of information. Both the reduction of precision and bias can have an effect on parameter estimation. For these reasons, the investigator needs to determine if possible why the data are missing so appropriate methods of analysis can be used (Fitzmaurice et al, 2004; Hedeker & Gibbons, 2006).

8.4 Missing Data Mechanisms

Missing data mechanisms represent the collection of reasons why data can be missing in a study. Reasons for missingness may be either related or unrelated to the topic of study. The missing data mechanisms that will be examined in this practicum are MCAR (missing completely at random), MAR (missing at random), and MNAR (missing not at random).

8.4.1 Missing Completely at Random

The first missing data mechanism to be considered is missing completely at random, or MCAR. Under the MCAR classification, the probability that an observation is missing does not depend on any previously observed responses or the response that should have been collected. Due to the completely random reason for missingness, the data at hand are simply a random sample of the complete dataset and thus exhibit the same distributional properties. Consequently, if valid inferences result from a method of analysis appropriate for complete data, that method is also appropriate and will lead to valid inferences for data that have an MCAR pattern of missingness. An example of

MCAR is a lab technician dropping a test tube of blood containing the sample of a participant.

8.4.2 Missing at Random

The second missing data mechanism that will be examine is missing at random, or MAR. The probability that an observation is missing (e.g. at time t), depends on the previously observed responses (e.g. before time t) but does not depend on the response that *should have been* obtained at time t. Consequently, the observed data are **not** a random sample of the complete dataset and thus do **not** have the same distributional properties. Thus, sample means (covariances) calculated when data are MAR yield biased estimates of the means (covariances) in the population. It should be noted that if the mean response and covariance structure is correctly specified for data that are MAR, valid inferences for the mean response can be made using likelihood based techniques. An example of MAR is the removal of a subject if their response value does not attain a specific value.

8.4.3 Missing Not at Random

Data are classified as missing not at random, or MNAR, if the probability that an observation is missing depends on the response(s) that should have been collected. An investigator cannot ignore this type of missingness, and it is accordingly referred to as *nonignorable* missingness. Most methods of analysis will lead to biased estimates of mean response. Modeling the mean response and missing data mechanism is a requirement for obtaining valid estimates (Fitzmaurice et al., 2004). An example of MNAR is a participant dropping out of the study when their quality of life is compromised.

8.5 Simulation Study

To demonstrate the effects that missing data mechanisms have on longitudinal data analysis, repeated measurements were generated according to a multivariate normal distribution. Observations were subsequently deleted according to MCAR, MAR, and MNAR patterns in a monotonic fashion. To determine which observations to delete, two models were used. The models were based on (i) Algina, Keselman and Othman (2003), and (ii) Fitzmaurice, Laird and Ware (2004). Each simulation study is outlined below followed by an explanation of the simulation study executed in this practicum.

Simulation Study Using Algina, Keselman and Othman (2003) Model

1. Data were generated for a two-group repeated measures design based on the model equation

$$Y_{ijk} = \beta_{0i} + \beta_{1i} t_j + \varepsilon_{ijk}$$

for the i^{th} subject ($i = 1,\ldots,n_k$), j^{th} measurement occasion ($j = 1,\ldots,J$) and k^{th} treatment group ($k = 1,2$). Notice that the model is the same for each treatment group and only differs in the error term. See the Algina, Keselman and Othman (2003) paper for details regarding coding of t_j, the intercept and slope.

2. Once the data were generated in their complete form, observations were deleted according to the model

$$Z_{ijk} = \theta_{1j} + \theta_2 \beta_{0i} + \theta_3 \beta_{1i} + \theta_4 Y_{i(j-1)k} + \theta_5 Y_{ijk}.$$

For each observation Y_{ijk}, a uniformly distributed random variable U_{ijk} was compared to $\Phi(Z_{ijk})$ where Φ represented the standard normal distribution function. Each U_{ijk} was generated using a random number function in SAS. If $U_{ijk} < \Phi(Z_{ijk})$,

then the observation was deleted based on MCAR, MAR, or MNAR missing data mechanisms. In order to specify the missing data mechanism by which the observations were deleted, the *theta* parameters were manipulated.

3. The *theta* parameters were set as follows:

MCAR: $\theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$,

MAR: $\theta_2 = \theta_3 = \theta_5 = 0$,

MNAR-Y: $\theta_5 \neq 0$, and

MNAR-SI: $\theta_2 \neq 0$; $\theta_3 \neq 0$.

A cumulative rate of missingness at the last measurement occasion was set to vary between 30 - 40% by manipulating the θ_{1j}'s for each missing data mechanism. The MCAR and MAR mechanisms were defined as in Sections 8.4.1 and 8.4.2 in this practicum, respectively. A distinction was made between two types of MNAR mechanisms, namely MNAR-Y and MNAR-SI. Missingness in the MNAR-Y mechanism was based on the observation that should have been collected (as defined in Section 8.4.3 of this practicum) while missingness in the MNAR-SI mechanism was based on the subject's intercept and slope.

Simulation Study Using Fitzmaurice, Laird and Ware (2004) Model

1. Repeated measures data were generated for the i^{th} subject ($i = 1,\ldots,N$) at the j^{th} measurement occasion ($j = 1,\ldots,5$) following the mean model

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

with covariance structure of the observations being autoregressive of order one,

$$Corr(Y_{is}, Y_{it}) = \rho^{|s-t|}$$

where $\rho \geq 0$.

2. Values of β_1, β_2, and ρ were set at 5, 0.5, and 0.7 respectively.

3. The model to determine missingness for either MCAR, MAR, or MNAR missing data mechanisms was

$$\log\left\{\frac{\mathrm{P}(D_i = k \mid D_i \geq k, \mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{ik})}{\mathrm{P}(D_i > k \mid D_i \geq k, \mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{ik})}\right\} = \theta_1 + \theta_2 \mathrm{Y}_{i(k-1)} + \theta_3 \mathrm{Y}_{ik}.$$

D_i was defined as the 'dropout indicator variable' where $D_i = k$ if missingness occurred between the $(k-1)^{\text{th}}$ and k^{th} time points.

4. The missing data mechanisms were all defined as in Sections 8.4.1, 8.4.2, and 8.4.3 of this practicum. To set the type of missingness, the *theta* parameters were set as follows:

MCAR: $\theta_2 = \theta_3 = 0$,

MAR: $\theta_3 = 0$, and

MNAR: $\theta_3 \neq 0$.

8.5.1 Simulation Study: Details

To create the simulation study in this practicum, different aspects of the models just presented by Algina, Keselman and Othman (2003) and Fitzmaurice, Laird and Ware (2004) were taken to form a final model which will now be described. The SAS code was based on a poster presentation by Lloyd & Lix (2007).

Data were generated for a single-group design with four repeated measurements. The linear model used to model the mean structure was

$$\mathrm{E}(\mathrm{Y}_{it}) = \beta_1 + \beta_2 t$$

where i = 1, 2, …, N, and t = {1, 2, 3, 4}. Values of β_1 and β_2 were set at 0.735 and 0.323 respectively. An autoregressive structure with order one with $\rho = 0.7$ was used to model the covariance. To specify missingness according to either MCAR, MAR, or MNAR missing data mechanisms, the following model was adopted:

$$Z_{ij} = \theta_1 + \theta_2 Y_{i(j-1)} + \theta_3 Y_{ij}$$

where i indicates the subject and j specifies the measurement occasion. If $U_{ij} < \Phi(Z_{ij})$ then the observation is deleted. The model $U_{ij} < \Phi(Z_{ij})$ is a simplified version of the model used in the Algina, Keselman and Othman (2003) simulation study. U_{ij} represents a random variable that is uniformly distributed and Φ represents the standard normal distribution function. Again, each U_{ij} was generated using a random number function in SAS. To create a setting where missing data is MCAR, both θ_2 and θ_3 were set to zero. For MAR, $\theta_3 = 0$, and for MNAR, only $\theta_3 \neq 0$.

The parameters manipulated were:

(a) type of missing data mechanism (i.e., MCAR, MAR, and MNAR),

(b) percentage of missingness (i.e., 20%, 40%, 60%, and 80%), and

(c) sample covariance structure (i.e., AR-1, CS, and UN).

Each combination of conditions was performed when N = 20000. The values of θ_1, θ_2, and θ_3 were determined by trial and error to obtain various percentages of missingness. U_{ij} was randomly selected for each iteration using the random uniform function in SAS. All subjects had complete data for the first measurement occasion with potential for missingness at the second and subsequent occasions. The percentage of missingness was calculated based on the number of observations missing at the last measurement occasion.

Once data were deleted, PROC MIXED was used to fit a simple mean model to the simulated data with covariance structures of AR-1, CS, and UN separately. The mean models were fit with fixed effects of group, time, and group × time interaction. There were no random effects. The main objective was to show how the different missingness patterns have an effect on the regression parameter estimates under a range of conditions.

Figure 8.1 displays the sample means at each measurement occasion plotted against the population regression line when the generated data were complete and an AR-1 covariance structure was used. Notice how the sample means represented by the triangles virtually coincide with the population regression line.

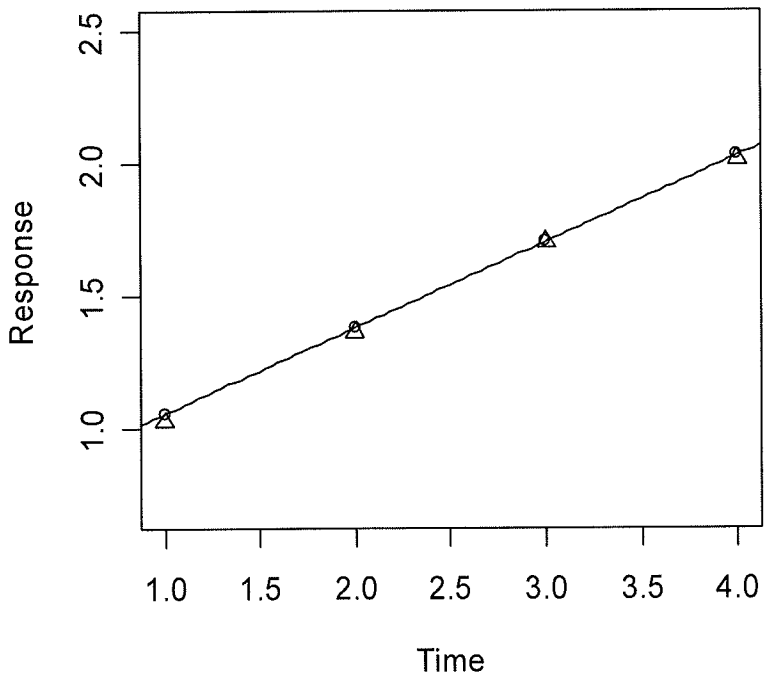

Figure 8.1: Sample means when data were complete

As a comparison, Figure 8.2 (a), (b), and (c) show the sample means at each time point plotted against the population regression line when the missing data were MCAR,

MAR, and MNAR, respectively. Again, an AR-1 covariance structure was specified in the model.

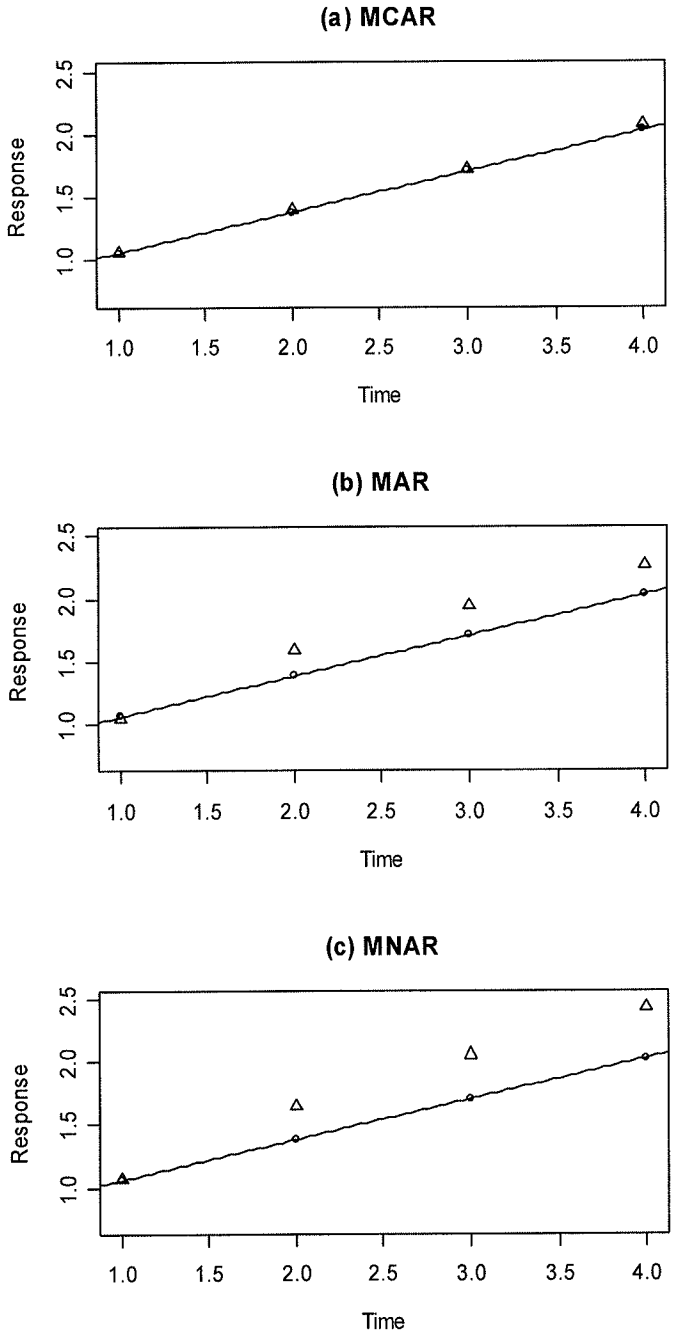

Figure 8.2: Sample means when missingness is 40% for data that is (a) MCAR, (b) MAR, and (c) MNAR

The graphs in Figure 8.2 reflect missingness that is approximately 40% at the last measurement occasion. When the missing data are MCAR, the sample means are very closely matched to the population regression line. This is no surprise as MCAR data are a simply a random subset of the complete dataset. When missing data are MAR, the sample means begin to deviate from the population regression line, and deviate even further when the missing data are MNAR. Based on Figure 8.2, it appears as though subjects with *lower* response values are more likely to drop out.

Next, we examine the estimated regression parameters for the simulated conditions. The results are displayed in Table 8.1. The heading of the first column is an abbreviation for missing data mechanism (i.e. MDM). Recall that the population covariance structure is AR-1. When there are no missing data, all three covariance structures fit to the generated data yield similar results. When the percentage of missing observations is approximate 20% at the last measurement occasion, parameter estimates are similar across the sample covariance structures when data are MCAR. These estimates were also comparable to results when data were complete. When the data were MAR and MNAR, we can see differences in the parameter estimates not only when comparing them to the estimates resulting when data were complete but also when the covariance structure was misspecified (e.g., CS). When missingness was 20% and the data were MAR, the intercept estimates (standard errors) were 0.7367 (0.0089), 0.7357 (0.0089), and 0.7752 (0.0079) when sample covariance structures were UN, AR-1, and CS, respectively. In this situation, the slope estimates (standard errors) were 0.3197 (0.0029), 0.3206 (0.0029), and 0.2998 (0.0022) when covariance structures were UN, AR-1, and CS, respectively. As the percentage of missingness increased, parameter estimates were

further over-estimated or under-estimated especially when the data were MNAR and the covariance structure was CS. The results of this illustration in terms of parameter estimates coincide with what was expected when data were either MCAR, MAR, or MNAR (i.e. biased estimates of the mean response trend). Furthermore, correctly specifying the covariance structure is also important and consequences of misspecification are shown in Table 8.1.

		UN		AR-1		CS	
MDM	**Parameter**	**Estimate**	**Std Err**	**Estimate**	**Std Err**	**Estimate**	**Std Err**
				No Missing Data			
None	Intercept	0.7424	0.0088	0.7423	0.0088	0.7433	0.0078
	Slope	0.3203	0.0027	0.3204	0.0027	0.3198	0.0021
				20% Missingness			
MCAR	Intercept	0.7218	0.0089	0.7218	0.0090	0.7213	0.0079
	Slope	0.3219	0.0029	0.3219	0.0029	0.3218	0.0023
MAR	Intercept	0.7367	0.0089	0.7357	0.0089	0.7752	0.0079
	Slope	0.3197	0.0029	0.3206	0.0029	0.2998	0.0022
MNAR	Intercept	0.7277	0.0089	0.7213	0.0085	0.7613	0.0075
	Slope	0.3661	0.0029	0.3641	0.0028	0.3499	0.0022
				40% Missingness			
MCAR	Intercept	0.7526	0.0090	0.7526	0.0090	0.7542	0.0080
	Slope	0.3209	0.0030	0.3209	0.0030	0.3202	0.0023
MAR	Intercept	0.7192	0.0089	0.7180	0.0090	0.7604	0.0080
	Slope	0.3253	0.0030	0.3264	0.0030	0.3030	0.0023
MNAR	Intercept	0.7209	0.0090	0.7085	0.0085	0.7511	0.0075
	Slope	0.3568	0.0028	0.3561	0.0027	0.3420	0.0021
				60% Missingness			
MCAR	Intercept	0.7243	0.0093	0.7244	0.0093	0.7236	0.0083
	Slope	0.3254	0.0035	0.3253	0.0035	0.3262	0.0028
MAR	Intercept	0.7477	0.0092	0.7465	0.0092	0.7870	0.0083
	Slope	0.3138	0.0036	0.3144	0.0036	0.2900	0.0029
MNAR	Intercept	0.6462	0.0093	0.6511	0.0090	0.6879	0.0081
	Slope	0.4148	0.0035	0.4096	0.0035	0.3922	0.0028
				80% Missingness			
MCAR	Intercept	0.7552	0.0095	0.7550	0.0096	0.7551	0.0087
	Slope	0.3195	0.0043	0.3196	0.0043	0.3194	0.0035
MAR	Intercept	0.7458	0.0095	0.7438	0.0096	0.7734	0.0087
	Slope	0.3211	0.0044	0.3229	0.0044	0.3025	0.0036
MNAR	Intercept	0.6301	0.0096	0.6352	0.0095	0.6702	0.0086
	Slope	0.4261	0.0042	0.4211	0.0043	0.4010	0.0036

Table 8.1: Parameter estimates under various missing data mechanisms (i.e. MDM's) and missingness rates

8.6 Dealing with Missing Data

There are a variety of different ways of handling missing data in longitudinal analysis. A few of these methods will be discussed in the following sections.

8.6.1 Complete Case Analysis

This method is the easiest to implement. If a subject has a missing observation at any measurement occasion, that subject is deleted and not included in the analysis. This method will yield unbiased estimates only if the data are assumed to be MCAR. This method is not recommended in general due to the loss of information.

8.6.2 Available Data Analysis

As the name suggests, this method uses all of the data available for each subject, even if a subject has missing observations. Again, like the complete cases analysis method, the available data method yields unbiased estimates only if the missing data are assumed to be MCAR.

8.6.3 Imputation

For the imputation method, new observations are created where there are missing data, thus creating a complete dataset. According to Fitzmaurice, Laird and Ware (2004), rather than performing a single imputation, it is recommended that multiple imputations are performed. This is done in order to reflect the *uncertainty* of the imputation values. There are several types of imputation. Some of these are: (a) Stratification, (b) Regression Imputation and (c) Last Observation Carried Forward (LOCF) (d) Deductive Imputation and (e) Nearest-Neighbor Imputation (Fitzmaurice et al, 2004; Lohr, 1999). Drawbacks of imputation methods include the underestimation of the variances and

covariances as well as smaller standard errors. Also, imputation methods can be unreliable and assume that the data are MCAR.

8.7 Missing Data in Current Study

In the current study, there are only a few missing observations. For the response variables Vitamin A, Vitamin E, and Total Antioxidant Capacity, subject 9508 from the flaxseed oil group had missing data at Week 4. For the response variables Isoprostanes, subject 5145 from the milled flax group had a missing observation at Week 0. Finally, for the response variable Hydroperoxide, subjects 6115 and 7658 had missing observations at Week 0 and Week 4 respectively. None of the subjects 'dropped out' of the study. For the subjects who did have missing observations, missingness often only occurred for a few of the response variables at the given measurement occasion. Perhaps the blood sample obtained for an individual could not be analyzed properly and could only yield results for certain response variables. Or, perhaps the technician made an error which resulted in missingness. Given these possible scenarios, it is reasonable to assume that the missingness in the current study is either missing completely at random (MCAR) or missing at random (MAR).

According to Cnaan, Laird and Slasor (1997), if missing data are MCAR, maximum likelihood estimates are valid and efficient. If missing data are MAR, estimates are valid and efficient given the model fit to the data for analysis is correct. If the missing data are assumed to be MAR, it is particularly important to execute the model diagnostics discussed in Section 6.4 to ensure that an appropriate model is fit to the data.

8.7.1 Example: Last Observation Carried Forward Imputation

The variable Hydroperoxide was used to demonstrate the Last Observation Carried Forward (LOCF) imputation method. The LOCF imputation technique imputes values using the last observed response for an individual. For example, the vector of response values for the subject with ID = 7658 was {0.48, missing, 0.07, 0.17}. The corresponding vector with the imputed value was {0.48. 0.48, 0.07, 0.17}. A linear model was fit for the mean response with a Toeplitz covariance structure. This model was fit to both the original Hydroperoxide data as well as the Hydroperoxide data that contained the imputed values. Table 8.2 below contains results for the original dataset that had two missing values.

Parameter	Effect	Estimate	Std Err	*t*	*p*-value
Intercept		0.6852	0.0599	11.44	<.0001
Group	MF	-0.0811	0.0806	-1.01	0.3215
Group	FO	-0.1931	0.0815	-2.37	0.0241
Time		-0.0169	0.0070	-2.41	0.0177
Group×Time	MF	0.0026	0.0094	0.27	0.7849
Group×Time	FO	0.0117	0.0095	1.24	0.2187

Table 8.2: Fixed effects for original Hydroperoxide data

Table 8.3 below contains the results for the Hydroperoxide dataset that was completed using the LOCF imputation technique.

Parameter	Effect	Estimate	Std Err	*t*	*p*-value
Intercept		0.6851	0.0594	11.54	<.0001
Group	MF	-0.0841	0.0790	-1.06	0.2949
Group	FO	-0.1922	0.0804	-2.39	0.0228
Time		-0.0169	0.0069	-2.43	0.0168
Group×Time	MF	0.0029	0.0092	0.32	0.7528
Group×Time	FO	0.0117	0.0094	1.25	0.2158

Table 8.3: Fixed effects for the imputed Hydroperoxide data

The results in the tables above are very similar. This suggests that any conclusions that would have made would not differ between the two datasets. For this reason, imputation for the rest of the response variables that have missing responses will not be explored as the results are not likely to change.

Chapter 9

Summary and Conclusions

9.1 Concluding Remarks

Longitudinal data analysis is important in several fields of research, particularly the health sciences. As discussed in this practicum, there are several approaches that can be used for the analysis of longitudinal data. These analyses would not be possible without the specific procedures for longitudinal data in various statistical software programs. These programs are continually being improved and developed.

There are several steps to follow when analyzing longitudinal data. First, it is important to represent the data in graphical form. A researcher may choose to select a sample of subjects and examine the response patterns across time. Or, it may be convenient to plot the mean response at each measurement occasion especially if there is more than one treatment group. Upon plotting the data, the investigator can gain a sense

of the overall response trend over time and think about specific models that might be appropriate for the data. Graphs of the data are also useful tools for presenting the information to an audience. After plotting the data, a mean model should be fit to the data. It might be appropriate to choose a very simple linear model, however, a more complicated model might be necessary. In order to find the best mean model, one can try different mean models for the data and examine the model fit using tools such as the likelihood ratio test. Once a satisfactory mean model is chosen, an appropriate covariance structure for the data should be specified. Various covariance structures may be fit separately and subsequently compared using tools such as the information criteria fit statistics discussed in this practicum. It is also important to consider whether or not to include random effects for the data. In this practicum, fitting a random intercept alone and then with a slope were explored. As a final check for appropriate model fit, residuals should be examined. Graphing tools such as a histogram of the residuals or quantile plots can be very useful in determining if the model is suitable for the data.

While this practicum did not go into depth regarding background theory in estimation, differences between ML and REML estimation techniques were discussed. Also, various tests of hypotheses were introduced.

The response variables that were examined in detail were Vitamin C and Vitamin E. These response variables were used to demonstrate how an investigator might examine longitudinal data. Based on the time plot of the mean Vitamin C levels over time (Figure 3.10), it appeared as though there may be a difference between the milled flax and control groups. Simple *t*-tests comparing the control to both the milled flax and flaxseed oil groups separately revealed that there may be a difference in Vitamin C between the

flaxseed oil and control group at Week 8. ANOVA tests comparing all of the groups revealed that there may be a difference in Vitamin C levels at Weeks 4, 8 and 12. Simple methods of analyses were not performed on the Vitamin E data. Next, more contemporary methods of fitting models to longitudinal data were fit to the data. First the response profile method was used to fit the mean model. An unstructured covariance structure was used for the covariance structure. Given this combination of mean and covariance structure, the group × time interaction effect was significant indicating that there may be a difference in Vitamin C levels among the groups across time. However, upon testing further mean models, the response profile model was not as adequate for the data as a quadratic mean model with a Toeplitz covariance structure. This model yielded a significant p-value for the interaction term (i.e., group × time^2). This indicated that there was a difference in Vitamin C levels among the groups as a quadratic function of time.

A similar procedure was followed for finding a suitable model for the Vitamin E data. The best model was a cubic trend model for the mean with a compound symmetric structure for the covariance. Based on this model, both the group × time interaction terms that included the quadratic and cubic time variables were significant. More specifically, the interaction terms were significant for the milled flax group indicating that there may be a difference in Vitamin E levels for the milled flax group as compared to the control and flaxseed groups across time.

Next, linear mixed models were fit to the data. These models included a model with a random intercept only followed by a model that included both a random intercept and slope. First, a random intercept was fit for the Vitamin C data using a quadratic mean

model with a Toeplitz covariance structure. Due to errors in the SAS program, the Toeplitz covariance structure was set to AR-1 instead. Based on this modified model, the group × time^2 interaction effect was significant. Next, a random slope was added to this model. The same conclusion of a difference between the treatment groups as a quadratic function of time was realized. The BLUP's were displayed for the random slope and intercept model to illustrate how they were calculated. Next, a random intercept and slope were added to the cubic trend mean model with a compound symmetric covariance structure for the Vitamin E data. This model would not converge despite trying various covariance structures. This was most likely due to the complicated model and the small sample size.

To explore the consequences of missing data in a longitudinal study, a simulation was conducted in SAS. A longitudinal dataset with four measurement occasions was generated following a linear mean model with an AR-1 covariance structure. Observations were deleted according to three missing data mechanisms: (1) Missing completely at random (MCAR), (2) Missing at random (MAR), and (3) Missing not at random (MNAR). The main results of the study revealed that missing data classified as MAR or MNAR showed increasing bias in the parameter estimates as the percentage of missing data increased. The bias was even more severe when the covariance structure was misspecified.

The three main methods discussed for handling missing data in this practicum included complete case analysis, available data analysis and imputation. For the current study, the 'last observation carried forward' method of imputation was used to complete the Hydroperoxide dataset which had two missing observations. The same model was fit

for both the Hydroperoxide data with missing values as well as the Hydroperoxide data that included the imputed values. The results (i.e., parameter estimates, standard errors, main conclusions) from the two Hydroperoxide datasets were very similar. Given that the rest of the response variables in the study had few missing values, imputation was not investigated further.

An approach for the analysis of longitudinal data that was not examined in this practicum was a 'Multivariate Repeated Measures' approach. Multivariate repeated measures data arise when multiple response variables are measured at each occasion. In the current study, eight dependent variables were each measured at Weeks 0, 4, 8, and 12 which corresponds to a multivariate repeated measures design. A decision was made to examine each variable separately. The variables Vitamin C and Vitamin E were randomly chosen to be examined. A disadvantage of analyzing the data in this manner is that the dependent variables may be correlated. The approaches examined in this practicum do not consider this potential correlation as only a single outcome variable is considered at a time. In multivariate repeated measures data, the correlation among the multiple dependent variables plus the correlation across time should be accounted for in the covariance structure. A more complicated covariance structure called a Kronecker product structure is often implemented to model this unique covariance structure that is typical of multivariate repeated measures data. For example, following notation in Njue (2001), let Δ be the covariance matrix among P dependent variables with dimension $P \times P$. Let Ω be the covariance matrix among T repeated measurements with dimension $T \times T$. By obtaining the Kronecker product (i.e., $\otimes$) of these matrices (i.e., $\Delta \otimes \Omega$), the resulting $TP \times TP$ matrix is known as a Kronecker product structure. By using a

Kronecker product structure to model the covariance, the number of covariance parameters to estimate is substantially reduced compared to ½ TC (TC + 1) parameter estimates if no covariance model is specified (Njue, 2001). Given the nature of this study, a multivariate approach is not suitable after all. There are eight dependent variables measured at each of four occasions, which yields a total of ½ 4*8(4*8 + 1) = 528 parameter estimates! This is a significant problem as the study only had a total of 35 subjects. Even if a multivariate analysis using only two of the eight dependent variable was attempted, there would still be ½ 4*2(4*2 + 1) = 36 parameter estimates. Thus, a multivariate analysis approach is unreasonable given the study size.

Although not all of the response variables in this study were examined, the steps outlined in this practicum could be applied to the remaining variables. The main steps for the analysis of longitudinal data are as follows: (1) Plot the data and make preliminary inferences, (2) Find a reasonable model for the mean, (3) Find a reasonable model for the covariance structure. If one wishes to separate the effect of experimental units from the experimental error, random effects for the experimental units can be introduced. The final step for the analysis of longitudinal data is to (4) Make final conclusions.

References

Algina, J., Keselman, H. J., and Othman, A. R. (2003). Analyzing group by time effects in longitudinal two-group randomized trial designs with missing data. *Journal of Modern Applied Statistical Methods, 2*, 2-280.

Cnaan, A., Laird, N. M., and Slasor, P. (1997). Tutorial in Biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16*, 2349-2380.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of Repeated Measures*. Chapman & Hall, London.

Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag, New York.

Diggle, P. J., Liang, K-Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Oxford University Press Inc., New York.

Everitt, B. S. (1995). The analysis of repeated measures: a practical review with examples. *The Statistician, 44*, 113-135.

Fitzmaurice, G., Laird, N., Ware, J. (2004). *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, New Jersey.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. John Wiley & Sons, Hoboken, New Jersey.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963-974.

Littell, R. C., Henry, P. R. and Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science, 76*, 1216-1231.

Littell, R. C., Milliken, G. A., Wolfinger, R. D. and Schabenberger, O. (2006). *SAS for Mixed Models, 2nd Edition*. Cary, NC, SAS Institute Inc.

Littell, R. C., Pendergast, J. and Natarajan, R. (2000). Tutorial in biostatistics: modelling covariance strucuture in the analysis of repeated measures data. *Statistics in Medicine, 19*, 1793-1819.

Lix, L. M. and Lloyd, A. M. (2006). A comparison of procedures for the analysis of multivariate longitudinal data. In Press.

Lloyd, A. M. and Lix, L. M. (2007, May). *Analysis of multivariate repeated measurements: missing observations and covariance misspecification.* Poster presented at the Canadian Society for Epidemiology and Biostatistics Conference. Calgary, AB.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Brooks/Cole Publishing Company, Pacific Grove.

Moore, D. S. (2003). *The basic practice of statistics 3rd Edition.* W. H. Freeman and Company, New York.

Njue, C. (2001). *On the efficiency of testing procedures in the linear model for multivariate longitudinal data.* Doctor of Philosophy thesis, University of Manitoba, Winnipeg, Manitoba.

Patel, T. (2005). *Assessing the oxidative status of individuals with the metabolic syndrome – before, during and after the consumption of milled flaxseed or flaxseed oil.* Master of Science thesis, University of Manitoba, Winnipeg, Manitoba.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

SAS Software, Version 9.1 of the SAS System for Windows. Copyright © 2002-2003, SAS Institute Inc., Cary, NC, USA

Schabenberger, O. and Pierce, F. (2002). *Contemporary Statistical Methods for the Plant and Soil Sciences.* CRC Press LLC, Boca Raton, Florida.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis.* Oxford University Press, Inc. New York, NY.

StataCorp. (2007). *Stata Statistical Software: Release 10.* College Station, TX: StataCorp LP.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* Springer-Verlag New York, Inc. New York, NY.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39,* 95-101.

Appendix A

The Longitudinal Data

All of the raw data in the metabolic syndrome study were presented in a multivariate format. As an example, the Vitamin C dataset is shown in the original multivariate structure followed by a transformed univariate structure. See Tables A.1 and A.2 respectively.

Subject	Group	ID	Week 0	Week 4	Week 8	Week 12
1	Control	2083	0.35	0.03	0.09	0.00
2	Control	2236	0.94	0.75	0.53	0.49
3	Control	2347	0.11	0.17	0.02	0.29
4	Control	2832	1.29	1.37	1.32	1.28
5	Control	2937	0.76	0.14	0.67	0.23
6	Control	3043	1.47	1.48	1.15	1.32
7	Control	3307	0.14	0.00	0.00	0.00
8	Control	3674	0.79	0.72	1.15	0.89
9	Control	3896	0.82	0.79	0.84	0.62
10	Control	3970	0.47	0.06	1.04	0.03
11	MF	4208	1.67	1.52	0.81	0.76
12	MF	4563	1.37	0.29	0.52	0.66
13	MF	4756	0.49	0.17	0.31	0.00
14	MF	4845	0.93	0.96	0.92	1.02
15	MF	5041	0.32	0.50	0.25	0.94
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Table A.1: Vitamin C Data in Multivariate Format

Subject	Group	ID	Time	Vitamin C Value
1	Control	2083	0	0.35
1	Control	2083	4	0.03
1	Control	2083	8	0.09
1	Control	2083	12	0.00
2	Control	2236	0	0.94
2	Control	2236	4	0.75
2	Control	2236	8	0.53
2	Control	2236	12	0.49
.	.	.	.	.
.	.	.	.	.
15	MF	5041	0	0.32
15	MF	5041	4	0.5
15	MF	5041	8	0.25
15	MF	5041	12	0.94
.	.	.	.	.
.	.	.	.	.

Table A.2: Vitamin C Data in Univariate Format

SAS: A Few Details

Different procedures used to analyze longitudinal data in SAS require the data to be input in different ways. Some procedures require that the data be entered in a multivariate format where each subject has only one row of data that contains all of the information collected for that particular subject. The procedure predominantly used in this practicum, PROC MIXED, uses data in a univariate format where each subject has multiple records of data. This means that a value collected at each time point has its own record. Data can be converted from its multivariate format to a univariate format and vice versa easily in SAS.

Appendix B

SAS: Sample Syntax

In order to assist the reader with the interpretation of the SAS code below, a brief explanation of the variables is included below.

`ID` = Unique identification number assigned to each individual

`GROUP` = Group membership, (i.e., Control, Milled Flax, or Flaxseed Oil)

`TIME` = Measurement occasion, (i.e., 0, 4, 8, 12)

`T` = A copy of the `'TIME'` variable, (i.e. 0, 4, 8, 12)

T^2 = Squared `'TIME'` values, (i.e., 0, 16, 64, 144)

Chapter 6
Response Profile Analysis

```
PROC MIXED DATA = VITC;
CLASS ID GROUP TIME;
MODEL Y = GROUP TIME GROUP*TIME / S CHISQ;
REPEATED TIME / TYPE = UN SUB = ID R;
RUN;
```

*To modify the covariance structure,
`TYPE = UN` for Unstructured
`TYPE = Simple` for Simple
`TYPE = AR(1)` for Autoregressive of order 1
`TYPE = CS` for Compound Symmetric
`TYPE = TOEPLITZ` for Toeplitz

SEE SAS DOCUMENTATION FOR COMPLETE LIST OF COVARIANCE STRUCTURES

Linear Response Trend

```
PROC MIXED DATA = VITC;
CLASS ID GROUP T;
MODEL Y = GROUP TIME GROUP*TIME / S CHISQ;
REPEATED T / TYPE = UN SUB = ID R;
RUN;
```

Quadratic Response Trend

```
PROC MIXED DATA = VITC;
CLASS ID GROUP T;
MODEL Y = GROUP TIME TIME**2 GROUP*TIME GROUP*TIME**2/ S
CHISQ;
REPEATED T / TYPE = UN SUB = ID R;
RUN;
```

Cubic Response Trend

```
PROC MIXED DATA = VITE;
CLASS ID GROUP T;
MODEL Y = GROUP TIME TIME*TIME TIME*TIME*TIME GROUP*TIME
GROUP*TIME*TIME GROUP*TIME*TIME*TIME/ S CHISQ;
REPEATED T / TYPE = UN SUB = ID R;
RUN;
```

Chapter 7

Random Intercept Model with group effect

```
PROC MIXED DATA = VITC;
CLASS ID GROUP T;
MODEL Y = GROUP TIME GROUP*TIME/ S CHISQ;
REPEATED T / TYPE = UN SUB = ID R;
RANDOM INT / SUB = ID;
RUN;
```

Random Intercept and Slopes Model with group effect

```
PROC MIXED DATA = VITC;
CLASS ID GROUP T;
MODEL Y = GROUP TIME GROUP*TIME/ S CHISQ;
REPEATED T / TYPE = UN SUB = ID R;
RANDOM INT TIME / SUB = ID;
RUN;
```

Simulation Study, Chapter 8

```
**********************************************************
**
*THE FOLLOWING PROGRAM DELETES OBSERVATIONS ACCORDING TO
VARIOUS MISSING DATA MECHANISMS;

**PART 1:  Multivariate data are generated from a normal
distribution, this is repeated 2500 times;
**Data are generated according to a linear mean model, i.e.
E(Yik) = Beta1k + Beta2k(time);
**The purpose is to examine the regression coefficients
Beta1k and Beta2k when the data are complete as compared to
when data are missing;
**Furthermore, we also examine the effects of missingness
when the covariance structure is correctly and incorrectly
specified;
**********************************************************
**
LIBNAME PART1 'E:\Writing\MDM\Part 1\a';

PROC IML;
RESET;
**----DEFINE CONDITIONS----**;
ALPHA=.05;
T=4; *number of repeated measurements;
GSIZE={2500}; *number of iterations;
%LET GSIZE=2500;
G=1; *number of groups;
%LET G=1;
NG=NCOL(GSIZE);
TYPECOV=1; *type of covariance structure;
%LET TYPECOV=1;

**-----------MISSING DATA MECHANISMS (MDM)-----------**;

*FOR MONOTONIC MISSINGNESS SET SWITCH AS FOLLOWS:
1=MCAR, 2=MAR, 3=MNAR;
SWITCH=2;
%LET SWITCH=2;

*FOR MISSINGNESS RATES SET MISS AS FOLLOWS:
1=APPROX 20%, 2=APPROX 30%, 3=APPROX 40%,
4=APPROX 60%, 5=APPROX 80%;
MISS=4;
%LET MISS=4;
```

```
***NOTE:  MISSINGNESS RATES ARE SET TO REPRESENT THE
PERCENTAGE OF VALUES MISSING AT THE LAST MEASUREMENT
OCCASION;
***THE FIRST MEASUREMENT OCCASION IS NEVER MISSING;

***--------DEFINE MISSINGNESS RATES WHEN SWITCH=1----------
***;
IF &MISS=1 THEN perc={0.08 0.08 0.08 0.08 0.08 0.08
                      0.08 0.08 0.08 0.08 0.08 0.08};
IF &MISS=2 THEN perc={0.115 0.115 0.115 0.115 0.115 0.115
                      0.115 0.115 0.115 0.115 0.115 0.115};
IF &MISS=3 THEN perc={0.16 0.16 0.16 0.16 0.16 0.16
                      0.16 0.16 0.16 0.16 0.16 0.16};
IF &MISS=4 THEN perc={0.26 0.26 0.26 0.26 0.26 0.26
                      0.26 0.26 0.26 0.26 0.26 0.26};
IF &MISS=5 THEN perc={0.42 0.42 0.42 0.42 0.42 0.42
                      0.42 0.42 0.42 0.42 0.42 0.42};

**------DEFINE MISSINGNESS RATES WHEN SWITCH = 2----------
**;
IF &MISS=1 THEN theta2 = -5.0;
IF &MISS=2 THEN theta2 = -1.9;
IF &MISS=3 THEN theta2 = -1.1;
IF &MISS=4 THEN theta2 = -0.65;
IF &MISS=5 THEN theta2 = -0.42;

**-----DEFINE MISSINGNESS RATES WHEN SWITCH = 3----------
**;
IF &MISS=1 THEN theta3 = -1.6;
IF &MISS=2 THEN theta3 = -3.8;
IF &MISS=3 THEN theta3 = -0.75;
IF &MISS=4 THEN theta3 = -0.55;
IF &MISS=5 THEN theta3 = -0.35;

*---DEFINE POPLN COV. STRUCTURE FOR REPEATED MEASUREMENTS---
*;
*Autoregressive Covariance for Repeated Measurements;
EPSCORR = {1 .7 .49 .343,
           .7 1 .7 .49,
           .49 .7 1 .7,
           .343 .49 .7 1};

**---DEFINE MEAN STRUCTURE---**;
XMAT={1 1,
      1 2,
      1 3,
```

```
      1 4};
BVECCO={0.735, 0.323};
XBCO=XMAT*BVECCO;
TXBCO=T(XBCO);
PMEAN=TXBCO;
*PMEAN IS THE MEAN STRUCTURE THAT RESULTS FROM THE ABOVE
EQUATION;

**-----------------GENERATE DATA--------------------------
**;
NTOT=SUM(GSIZE);
DO K=1 TO NG;
  X1=J(GSIZE[K],1,K);
  IF K=1 THEN X=X1;
  ELSE X=X//X1;
END;

Z=RANNOR(J(NTOT,T,0));
CT1=1;
CT2=0;
DO K=1 TO NG;
L=ROOT(EPSCORR);
GRPSZ=GSIZE[K];
CT2=CT2+GRPSZ;
GRPMN=PMEAN[K,];
TEMP=Z[CT1:CT2,]*L;
  IF K=1 THEN DO;
  Y=TEMP+REPEAT(GRPMN,GRPSZ,1);
  END;
  ELSE DO;
  Y=Y//(TEMP+REPEAT(GRPMN,GRPSZ,1));
  END;
  CT1=CT1+GRPSZ;
END;

**CREATE Y MATRIX WITH MISSING OBSERVATIONS;
*SET UP A REPLICATION OF THE Y MATRIX TO BE USED IN THE CODE
BELOW;
YFIX=Y;
CT1=1;
CT2=0;
DO K=1 TO NG;
GRPSZ=GSIZE[K];
CT2=CT2+GRPSZ;
  do jj=CT1 to CT2;
    do k2k2=2 to T;
```

```
      *MDM - McAR, WHERE & switch=1;
      if switch=1 then do;
      uni=uniform(repeat(0,1,1));
      if uni< perc[1,k2k2] then Y[jj,k2k2]=.;
      end;

      *MDM - MAR, WHERE & switch=2;
     if switch =2 then do;
     COMPMAR=probnorm(perc[1,k2k2]+(theta2)*(yfix[jj,k2k2-
     1]));
     uni=uniform(repeat(0,1,1));
     if uni<COMPMAR then y[jj,k2k2]=.;
     end;

     *MDM - MNAR-Y, WHERE switch=3;
     if switch =3 then do;
     compmnar=probnorm(perc[1,k2k2]+(theta3)*yfix[jj,k2k2]);
     uni=uniform(repeat(0,1,1));
     if uni<compmnar then y[jj,k2k2]=.;
     end;

     DO LK=(K2K2+1) TO T;
     IF Y[JJ,K2K2]=. THEN Y[JJ,LK]=.;
     END;
    END;*k2k2;
  END; *jj;

CT1=CT1+GRPSZ;
END; *k;

**-----CREATE DATASET USING A UNIVARIATE STRUCTURE FOR INPUT
TO PROC MIXED-------**;
DATAALL=X||Y;

IF T=4 THEN create datatest from dataall[colname={GROUP
DV1RM1 DV13LAST DV12LAST DV1LAST}];
append from dataall;

DATAALL_1=J(T*NTOT,5,0);
ID_VAL=(1:NTOT)`;
F=1;
L=NTOT;
CNPT=0;
DO J=1 TO T;
  CNPT=CNPT+1;
  DATAALL_1[F:L,1]=DATAALL[,1];
  DATAALL_1[F:L,2]=ID_VAL;
```

```
  DATAALL_1[F:L,3]=J(NTOT,1,1);
  DATAALL_1[F:L,4]=J(NTOT,1,J);
  DATAALL_1[F:L,5]=Y[1:NTOT,CNPT];
  F=F+NTOT;
  L=L+NTOT;
END;
CREATE DATAFIN FROM DATAALL_1[COLNAME={GROUP ID DV RM VAL}];
APPEND FROM DATAALL_1;

FILENAME NEWOUT 'OUTFILE';

PROC PRINTTO PRINT=NEWOUT NEW;
RUN;

ODS LISTING CLOSE;
ODS OUTPUT SOLUTIONF=F;

PROC SORT DATA=DATAFIN;
BY ID;
RUN;

*SET UP PROC MIXED MODEL WITH AN AUTOREGRESSIVE COVARIANCE
STRUCTURE;
PROC MIXED DATA=DATAFIN;
CLASS ID;
MODEL VAL=RM /Solution;
REPEATED / TYPE=AR(1) SUBJECT=ID;
RUN;

DATA READIN2; SET F;
COV=1;
IF effect='Intercept';
BETA=1; Est=Estimate; Stde=stderr; tstat=tvalue; PVAL=ProbT;
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est Stde tstat PVAL MDM
PERCMISS;
PROC APPEND BASE=PART1.estimatescomp;
RUN;

DATA READIN3; SET F;
COV=1;
IF effect='RM';
BETA=2; Est=Estimate; Stde=stderr; tstat=tvalue; PVAL=ProbT;
```

```
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est Stde tstat PVAL MDM
PERCMISS;
PROC APPEND BASE=PART1.estimatescomp;
RUN;

**---OUTPUT INFORMATION CRITERIA FOR AN UNSTRUCTURED
COVARIANCE STRUCTURE---**;

ODS LISTING CLOSE;
ODS OUTPUT SOLUTIONF=F;

PROC MIXED DATA=DATAFIN ;
CLASS ID;
MODEL VAL= RM /s;
REPEATED / TYPE=UN SUBJECT=ID;
RUN;

DATA READIN2; SET F;
COV=2;
IF effect='Intercept';
BETA=1; Est=Estimate; StdE=Stderr; tstat=tvalue; PVAL=ProbT;
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est StdE tstat PVAL MDM
PERCMISS;
PROC APPEND BASE=PART1.estimatescomp;
RUN;

DATA READIN3; SET F;
COV=2;
IF effect='RM';
BETA=2; Est=Estimate; StdE=Stderr; tstat=tvalue; PVAL=ProbT;
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est StdE tstat PVAL MDM
PERCMISS;
```

```
PROC APPEND BASE=PART1.estimatescomp;
RUN;



**---COMPOUND SYMMETRIC COVARIANCE STRUCTURE--**;

ODS LISTING CLOSE;
ODS OUTPUT SOLUTIONF=F;

PROC MIXED DATA=DATAFIN ;
CLASS ID ;
MODEL VAL= RM /s;
REPEATED / TYPE=CS SUBJECT=ID;
RUN;

DATA READIN2; SET F;
COV=3;
IF effect='Intercept';
BETA=1; Est=Estimate; StdE=Stderr; tstat=tvalue; PVAL=ProbT;
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est StdE tstat PVAL MDM
PERCMISS;
PROC APPEND BASE=PART1.estimatescomp;
RUN;

DATA READIN3; SET F;
COV=3;
IF effect='RM';
BETA=2; Est=Estimate; StdE=Stderr; tstat=tvalue; PVAL=ProbT;
NVAL=&GSIZE;
GRPTYPE=&G;
ep=&TYPECOV;
MDM=&SWITCH;
PERCMISS=&MISS;
KEEP GRPTYPE COV BETA ep NVAL Est StdE tstat PVAL MDM
PERCMISS;
PROC APPEND BASE=PART1.estimatescomp;
RUN;

**----CODE TO DETERMINE PERCENTAGE OF MISSINGNESS AT LAST
MEASUREMENT OCCASION FOR EACH SIMULATION---***;
```

```
*CALCULATE NUMBER OF NON-MISSING VALUES AT THE THIRD-LAST,
SECOND-LAST AND LAST OCCASIONS;
PROC MEANS n data=datatest;
VAR dv13LAST;
VAR dv12LAST;
VAR dv1LAST;
BY group;
OUTPUT OUT=outtest n=nonmiss3LAST nonmiss2LAST nonmissLAST;
run;

*CALCULATE NUMBER OF MISSING VALUES AT THE THIRD-LAST,
SECOND-LAST AND LAST OCCASIONS;
*CALCULATE PERCENT MISSINGNESS AT THE THIRD-LAST, SECOND-
LAST AND LAST OCCASIONS;
*CALCULATE TOTAL PERCENTAGE OF MISSING VALUES;
data percmissing;
set outtest;
no_missing2=(_freq_ - nonmiss3LAST);
no_missing3=(_freq_ - nonmiss2LAST);
no_missing4=(_freq_ - nonmissLAST);
no_miss_tot=no_missing2 + no_missing3 + no_missing4;
percmissG13LAST=(no_missing2/_freq_)*100;
percmissG12LAST=(no_missing3/_freq_)*100;
percmissG1LAST=(no_missing4/_freq_)*100;
percmisstotG1=((no_miss_tot)/(_freq_*4))*100;
run;

*SET UP OUTPUT FOR GROUP 1;

*MERGE DATASETS ABOVE TO CREATE OUTPUT DATASET WITH ALL
RELEVANT PERCENTAGE INFORMATION;
*NOTE - THE OUTPUT DATASET 'PERCENTAGEFINAL1' CAN BE CHANGED
TO REFLECT THE MISSING
DATA MECHANISM USED;
DATA PERCFINAL;
MERGE PERCMISSing readin2;
KEEP PERCMISSG13LAST PERCMISSG12LAST PERCMISSG1LAST
PERCMISSTOTG1 grptype MDM PERCMISS EP NVAL;
PROC APPEND BASE=part1.PERCDELETE;
RUN;

**-----END OF CODE TO DETERMINE PERCENTAGE MISSINGNESS----
***;

**---MEANS AT EACH REPEATED MEASUREMENT-----**;
PROC MEANS DATA=DATATEST;
VAR DV1RM1 DV13LAST DV12LAST DV1LAST;
```

```
OUTPUT OUT=MEANS MEAN=MEANRM1 MEANRM2 MEANRM3 MEANRM4
STD=STDRM1 STDRM2 STDRM3 STDRM4 STDERR=SERM1 SERM2 SERM3
SERM4 N=N;
RUN;

DATA MEANS2; merge readin2 MEANS;
MDM=&SWITCH;
PERCMISS=&MISS;
EP=&TYPECOV;
KEEP MDM PERCMISS EP GRPtype N
MEANRM1 MEANRM2 MEANRM3 MEANRM4 STDRM1 STDRM2 STDRM3 STDRM4
SERM1 SERM2 SERM3 SERM4;
PROC APPEND BASE=part1.MEANScomp;
RUN;

quit;
```