Spring 2016

Accounting for Correlation in the Analysis of Randomized Controlled Trials with Multiple Layers of Clustering

Adam Baumgardner

ACCOUNTING FOR CORRELATION IN THE ANALYSIS OF RANDOMIZED CONTROLLED TRIALS WITH MULTIPLE LAYERS OF CLUSTERING

A Thesis

Submitted to the McAnulty College & Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for
the degree of Master of Science

By

Adam Baumgardner

May 2016

ACCOUNTING FOR CORRELATION IN THE ANALYSIS OF RANDOMIZED CONTROLLED TRIALS WITH MULTIPLE LAYERS OF CLUSTERING

By

Adam Baumgardner

Thesis approved April 8, 2016.

Dr. Frank D'Amico Professor of Statistics (Committee Chair)	Dr. John Kern Professor of Statistics (Department Chair)
Dr. John Kern Professor of Statistics (Committee Member)	Dr. James Swindal Dean, McAnulty College and Graduate School of Liberal Arts Professor of Philosophy

ABSTRACT

ACCOUNTING FOR CORRELATION IN THE ANALYSIS OF RANDOMIZED CONTROLLED TRIALS WITH MULTIPLE LAYERS OF CLUSTERING

By

Adam Baumgardner

May 2016

Thesis supervised by Dr. Frank D'Amico.

A common goal in medical research is to determine the effect that a treatment has on subjects over time. Unfortunately, the analysis of data from such clinical trials often omits several aspects of the study design, leading to incorrect or misleading conclusions. In this paper, a major objective is to show via case studies that randomized controlled trials with longitudinal designs must account for correlation and clustering among observations in order to make proper statistical inference. Further, the effects of outliers in a multi-center, randomized controlled trial with multiple layers of clustering are examined and strategies for detecting and dealing with outlying observations and clusters are discussed.

ACKNOWLEDGEMENT

This thesis truly would not have been possible without the support of my advisor, Dr. Frank D'Amico. His passionate dedication to the study of statistics and intellectual curiosity have been an inspiration over the past two years, especially during our work on this thesis. He spent countless hours working with me to solidify my understanding of statistics and, for that, I am forever grateful. I would also like to thank Dr. John Kern who took time out of his busy schedule to read and critique this thesis. His patient and charismatic pedagogical approach initially sparked my interest in statistics during my first semester at Duquesne.

As a Teaching Assistant, I had the pleasure of sharing an office with Rich Boyles and Michael Guyer. Their positivity, humor, and passion for mathematics helped make my experience at Duquesne truly memorable.

Finally, I would like to thank my family. Since childhood, I have had the encouragement of my parents, David and Sherrie, throughout all of my academic and personal endeavours. Their guidance and support have been driving forces behind my academic ambitions and I could not have reached this point without them. And my soon-to-be wife, Jamie, whose unwavering love and commitment helped me stay positive and happy every day.

TABLE OF CONTENTS

List of Figures

List of Tables

1 Introduction and Background

A common goal in medical research is to determine the effect that a treatment has on subjects over time. Repeated measurements taken on the same subjects over time are often positively correlated, creating a cluster of observations for each subject. Correlation among observations brings about a few analytical challenges that must be taken into consideration. Unfortunately, statistical analysis of data from such trials often fails to take into account the clustering, which leads to incorrect inference. In this paper, a few techniques for properly accounting for this correlation in clusters will be presented and implemented. Additionally, a few approaches to detecting outlying observations in clustered data are introduced. Then, the information and techniques presented will culminate in a case study analysis of a multi-center, randomized controlled trial with a longitudinal design. First, a few basic concepts must be discussed to provide a foundation for understanding the statistical analysis that follows.

1.1 Clinical Trials and Experimental Design

When testing the efficacy of a new treatment or drug, drawing conclusions based on anecdotal evidence alone is ill-advised. For example, before the 19th century it was commonly believed that patients could be cured of illnesses by blood letting. This typically entailed applying leeches to the bodies of sick patients to suck out the bad blood. There was no scientific evidence of the efficacy or safety of this approach because methods for testing hypotheses through systematic data collection and statistical analysis had not been developed.[17] In this example, not only was there a lack of evidence that blood letting worked, it was dangerous and likely caused many people harm. This illustrates the importance of having appropriately designed research studies along with correct analysis when determining whether or not a treatment is safe and effective.

Research in medicine can be characterized as pre-clinical and clinical. Pre-clinical studies occur before the treatment of interest is given to human subjects. The main objective of pre-clinical research is typically to identify a safety profile for the treatment and to determine a safe dosage for initial human testing. A clinical trial, on the other hand, involves experimentation using human subjects and can be classified into four phases. In a Phase I clinical trial, investigators are primarily interested in exploring the potential side-effects of the treatment or drug being tested. Once a tolerable dosage is established and side-effects have been documented, investigators begin to examine the efficacy of the drug in a Phase II clinical trial. This phase typically consists of smaller-scale studies designed to determine whether the drug seems to be effective enough to warrant more costly, larger-scale Phase III clinical trials. During a Phase III clinical trial the new treatment or drug is compared either to the current standard of treatment or to a placebo. Finally, after the treatment has been deemed safe and effective, it either goes to market or becomes the new standard treatment. However, there is still a possibility of side-effects surfacing that had not been discovered during prior testing. Phase IV trials are observational studies that are implemented to monitor these potential issues.[17]

While each of the four phases of clinical trials plays a critical role in medical research, the focus of this paper will be primarily on Phase III trials. Approval of a new treatment by regulatory agencies typically depends on the results of Phase III testing. That being said, since Phase III clinical trials help dictate whether a drug should be made available to the general population, it is crucial that care is taken in both the design and analysis of such trials. As was stated previously, the objective of this type of clinical trial is to compare a new treatment to the current standard of treatment or to a placebo. To accomplish this, investigators must design the experiment according to statistical principles so results can be properly interpreted. Two key principles of experimental design that will be discussed are randomization and sample size determination.

Comparing groups of subjects receiving different treatments introduces a few technical issues. If the groups being compared are fundamentally different from one another, bias is introduced and the effect of the treatment may be confounded. In other words, a difference in the response variable between treatment groups may be due to variables unrelated to the treatment itself. This dilemma establishes the need for randomization. Randomly assigning subjects to different treatment groups is considered the gold standard for Phase III clinical trials. This type of study is generally referred to as a randomized controlled trial. The goal of randomization is to ensure all subjects within treatment groups in the trial are alike in all aspects except the treatment they receive. When this is the case, it can be assumed that any difference in the response variable between treatment groups can be credited directly to the difference in treatment. This causal relationship allows researchers to make proper statistical inference about the effect of the treatment(s) of interest.[1] Figure 1.1 gives a visual representation of a basic randomization scheme.

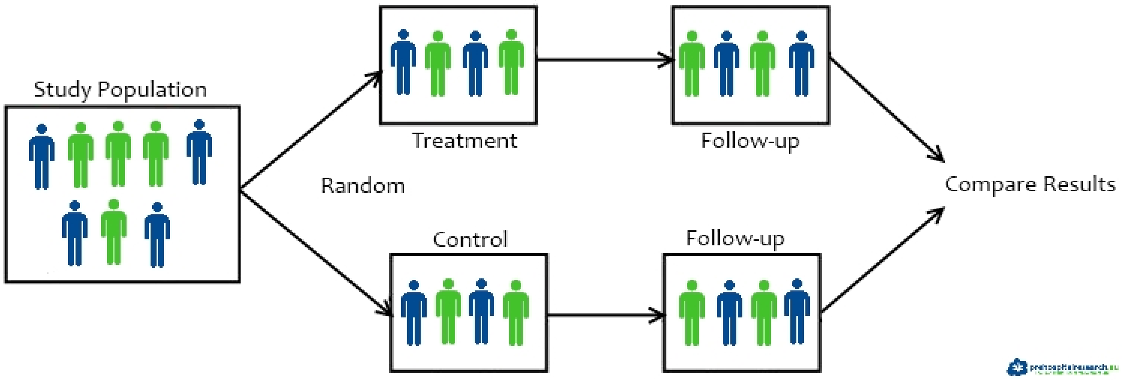


Figure 1.1: a simple randomization scheme

It is worth noting that both the treatment and control groups have two green subjects and two blue subjects. This is a trivial visualization of an important aspect of randomization: ensuring homogeneity. In the hypothetical example depicted in Figure 1.1, imagine that a blue subject represents a male and a green subject represents a female. If the randomization instead resulted in a treatment group of four males and a control group of four females, differences in response may have been due to sex rather than the treatment. As was stated earlier, this introduces bias and makes inference about the effect of the treatment misleading. Thus, if the randomization worked, there should be homogeneity among treatment groups. This can mean checking for relatively equal distributions of several variables such as sex, race, weight, etc. among treatment groups.

Another critical aspect of experimental design is determining an appropriate sample size to be used in randomized controlled trials. Recruiting too many subjects to participate in the experiment may be costly while recruiting too few subjects will result in the inability to draw any statistically significant conclusions. Researchers must find a proper balance between an experiment that is cost efficient and one that allows for meaningful analysis. The appropriate sample size to use for a particular experiment (usually denoted by n) is typically determined by the desired Type I and Type II error rates. The Type I error rate or significance level, α, is the probability of rejecting the null hypothesis, H_0, when it is true. On the other hand, the Type II error rate, β, is the probability of failing to reject H_0 when it is false. Another way of thinking about a Type II error is in terms of the *power* of an experiment. The power, $1-\beta$, is the probability of rejecting H_0 in favor of the alternative hypothesis, H_A, when H_A is true. It is rather intuitive that one would like for an experiment to have a low Type I error rate, α, along with high power, $1-\beta$. However, α and β are inversely related when n is fixed. That is, choosing a lower value for α directly leads to a higher value for β. If instead a value for α is chosen and fixed, increasing n generally leads to a decrease in β (and an increase in power). Thus, researchers can first choose a significance level and then determine the minimum sample size that is needed to achieve the desired power. As was mentioned earlier, one barrier that may arise when determining the sample size is cost. If the minimum sample size needed to maintain the desired values of α and β is too large to be financially feasible, concessions must be made. It is up to the researcher to determine whether it is more important to protect against Type I errors or against Type II errors, which is highly dependent on the context of the experiment. The idea of statistical power will be important in this paper and will be revisited in a future section.[6]

1.2 Longitudinal and Clustered Data

In a longitudinal design of a randomized controlled trial the main objective is to characterize the change in response to a treatment over time and to study the factors that influence the change. The identifying feature of a *longitudinal study* is that measurements of the response variable of interest are taken on the same subject repeatedly over time. With repeated measures, both within-subject and between-subject change can be captured using statistical models, allowing for a direct study of the change over time. Figure 1.2 shows an example of a longitudinal study that measures how the concentration of a substance changes over time in eight subjects.

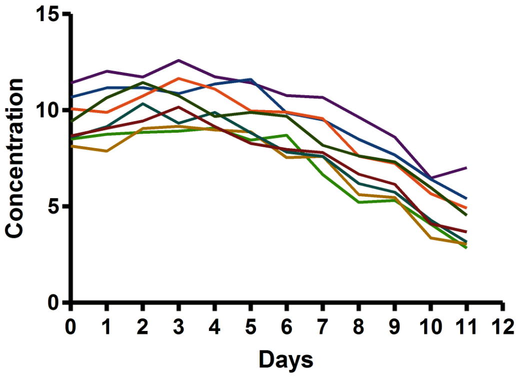


Figure 1.2: a basic longitudinal study involving eight subjects

Since the same individuals are being examined over time, the repeated measures for each individual form a *cluster*. In randomized controlled trials, the word "cluster" can have various meanings. In this paper, a cluster refers to a group of observations that are not independent from one another. In longitudinal studies, observations within these clusters have a natural ordering by time and will often be positively *correlated*. The correlation between two variables is the degree to which they are related in a linear fashion. The correlation is a standardized statistic usually denoted by ρ and is bounded between -1 and 1. If two variables are highly correlated ($\rho \approx 1$ or $\rho \approx -1$), then an almost direct linear relationship exists between them. Positive values for correlation imply that, as one variable increases (decreases), the other variable also increases (decreases). Conversely, a negative correlation implies that, as one variable increases (decreases), the second variable decreases (increases). *In the longitudinal case with repeated measures taken on the same subjects, the clustering arises because observations taken on one individual are more likely to be similar to each other than to the measurements taken on a different individual.* These effects of clustering will be important in the analysis and will be discussed at length in this paper. In non-longitudinal studies, clusters can still form either intentionally or naturally. For example, if an experiment is designed to determine if a new instructional technique is effective for improving math scores on a standardized test, a natural approach would be to compare the results from a class that did receive the new instructional technique to the results from a class that did not receive the new instructional technique. However, it is quite possible that the students within those classes should not be treated as statistically independent. For example, students enrolled in an honors math course are probably more similar to each other than they are to students in a standard math course. Thus, the honor students' test scores may change one way while the non-honors students' scores change in a different way. This lack of independence must be accounted for in the analysis before any meaningful conclusions can be made regarding the effect of the new instructional technique.[4]

Now that it has been mentioned that observations within clusters in longitudinal studies tend to be positively correlated, it is worth considering the potential sources of this correlation. According to Fitzmaurice[4], this correlation is generally impacted by three different sources of variation: between-subject heterogeneity, within-subject biological variation, and measurement error. The first source of variation, between-subject heterogeneity, arises due to the natural variation in humans' propensity to respond. In other words, in any longitudinal study, some subjects will be high respondents and others will be low respondents. High respondents will have consistently higher responses than average while low respondents will be consistently lower. Thus, a pair of repeated measures on one subject is likely to be more similar than a pair of measurements from two different subjects. The second source of variation that impacts correlation is within-subject biological variation. The idea behind this source of variation is that there are some underlying biological processes that cause a subject's response to deviate from their response trajectory. These deviations are likely to be more similar when the time between measurements is small. The third source of variation that impacts the correlation of clustered data that will be discussed is measurement error. Measurement error is a component of nearly all scientific studies and is often quantified as reliability. This reliability acts as a constraint for how closely correlated repeated measures can be. Now that we have seen the potential sources of correlation within clusters of repeated measures in longitudinal studies, we consider the consequences of ignoring this correlation in the analysis.

2 Correlation in Longitudinal Data: The Simple Case

In scientific research, investigators are often interested in whether or not a difference exists between two groups. This is usually examined by comparing the means of the two groups using a statistical test. Sometimes, the two groups under experimentation are simply pre- and post-intervention measurements on the same subjects. In this case, we can consider the study to be longitudinally designed with just two repeated measures (baseline and post-intervention).

As an example, consider the data found in Table 2.1 below. This data comes from a hypothetical study of $n = 5$ subjects where the response variable of interest Y is heart rate measured in beats per minute (BPM). Each subject's heart rate was measured at rest (time t_0) and then again after performing a physical activity (time t_1).

Subject	Resting HR (t_0)	Post-Activity HR (t_1)
1	68	74
2	80	96
3	82	92
4	76	80
5	74	82
mean: $\bar{X}_i$	76	84.8
variance: s_i^2	30.03	81.18

Table 2.1: heart rates measured on 5 subjects at 2 time points

The investigators are interested in determining if the mean resting heart rate is significantly different than the mean heart rate after activity. This is a simple example of a common situation that arises in scientific research. In this situation, the hypothesis of interest is given by

$$H_0 : \mu_0 = \mu_1$$

$$H_A : \mu_0 \neq \mu_1$$

where μ_0 and μ_1 represent the mean response at time t_0 and t_1, respectively. A typical approach to this problem is to perform a Student's t-test (One-way linear models approach). Unfortunately, investigators often fail to account for the correlation in the data by using this approach and often publish incorrect results. In this section, it will be shown how ignoring the correlation impacts the statistical inference being made.

2.1 Unpaired t-test: Ignoring the Correlation

An approach to testing the hypothesis of equal means between two time points that is commonly used by investigators is the unpaired Student's t-test. This is a very simple test that is taught in many introductory statistics courses. However, this test assumes that the two groups under experimentation are independent and, when dealing with repeated measures on the same subjects, this assumption is violated. This will be shown using both a traditional approach and a linear models approach. A linear model is a statistical model that describes a continuous random variable as a linear function of a set of predictor variables.

2.1.1 Traditional Approach to Unpaired Analysis

The traditional approach to performing the unpaired t-test is to calculate a test statistic, t, and then compare it to a Student's t-distribution with $n_1 + n_2 - 2$ degrees of freedom. First, the test statistic must be calculated:

$$t = \frac{\bar{X_1} - \bar{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.1}$$

where $\bar{X_1}$ and $\bar{X_2}$ are the means at time 1 and time 2, respectively, and n_1 and n_2 are the sample sizes at both times. In this equation, the denominator represents the *standard error* estimate for the difference in means. One component of the standard error is the pooled standard deviation of the two time points, s_p, which is calculated as follows:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2.2}$$

In effect, this equation yields a weighted average of the two sample variances which relies on the assumption that the population variances of the two time points are equal ($\sigma_1^2 = \sigma_2^2$). Using the data from the heart rate study, the following is obtained:

$$s_p = \sqrt{\frac{(5-1)(30.03) + (5-1)(81.18)}{5 + 5 - 2}} = \sqrt{55.6} = 7.46 \tag{2.3}$$

which yields a test statistic of

$$t = \frac{76 - 84.8}{7.46\sqrt{\frac{1}{5} + \frac{1}{5}}} = -1.87 \tag{2.4}$$

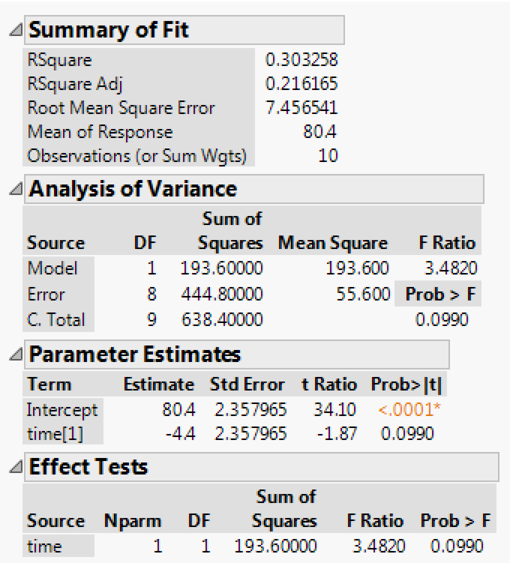
When compared to a t-distribution with 8 degrees of freedom, this test statistic proves to be significant ($p < 0.05$), allowing for the rejection of H_0 in favor of H_A. However, the pooled variance term used in the standard error does not account for correlation because it relies on the assumption of independence between the two time points. Since a longitudinal study almost always results in correlated data, this analysis is basically incorrect. Using this test would only be appropriate if the two time points were independent, which, in this case, they are not. This introduces the need for pairing the data and performing the analysis on the pairs.

2.1.2 Linear Models Approach: One-way ANOVA

An analagous approach to performing an unpaired t-test is the use of a One-way Analysis of Variance (ANOVA) model. In a One-way ANOVA, the goal is to model the difference between two or more independent groups. For the heart rate study, the model can be expressed mathematically as

$$Y_{ij} = \mu + \alpha_i + e_{ij} \text{ with } i = 1, 2 \text{ and } j = 1, \cdots, 5 \tag{2.5}$$

where μ is the overall mean of the observations, α_i is the effect of time group i, and $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ is the random error term. Here, α_i is considered the only effect in the model; μ is a constant and e_{ij} is the random error. Using JMP® statistical software[13], the following output is produced:



Summary of Fit

RSquare	0.303258
RSquare Adj	0.216165
Root Mean Square Error	7.456541
Mean of Response	80.4
Observations (or Sum Wgts)	10

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	1	193.60000	193.600	3.4820
Error	8	444.80000	55.600	Prob > F
C. Total	9	638.40000		0.0990

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	80.4	2.357965	34.10	<.0001*
time[1]	-4.4	2.357965	-1.87	0.0990

Effect Tests

			Sum of		
Source	Nparm	DF	Squares	F Ratio	Prob > F
time	1	1	193.60000	3.4820	0.0990

Figure 2.1: JMP: One-way ANOVA table, parameter estimates, and summary statistics

In the ANOVA table, the mean square column shows the amount of variation in the data that is due to the groups (model) and the amount that is due to the error (error). In fact, the mean square error of 55.6 is exactly the same value that was used for the pooled variance following the traditional approach, verifying that the correlation has still been ignored using this approach. The ratio of the mean squares gives the F-statistic, which is used to test the same hypothesis as in the traditional approach. When there is only 1 degree of freedom in the model, $F = t^2$, which is the case here ($3.482 = (-1.87)^2$). Thus, whether the traditional unpaired Student's t-test approach or the linear models approach using a One-way ANOVA is used to test the null hypothesis of equal group means, any correlation in the data will be ignored as both of these approaches rely on the assumption of independence.

2.2 Paired t-test: Accounting for the Correlation

To appropriately account for correlation, the analysis of this study must take into account the pairing of data (i.e., clustering). To do so, a few simple adjustments must be made when calculating the test statistic. As with the unpaired t-test, two analagous approaches will be shown: the traditional approach and the linear models approach.

2.2.1 Traditional Approach to Paired Analysis

Similar to the unpaired case, the traditional approach to performing a paired t-test is to calculate a test statistic, t, and compare it to a Student's t-distribution with $n - 1$ degrees of freedom where n is the number of pairs of observations. Here, the test statistic is given by

$$t = \frac{\bar{X}_d - \mu_0}{\frac{s_d}{\sqrt{n}}} \tag{2.6}$$

where $\bar{X}_d$ is the mean of the paired differences between time 1 and time 2, μ_0 is a hypothesized value for the mean difference (in this example, 0), and s_d is the standard deviation of the differences. The major difference between this method and the unpaired method is that calculations are based on the difference between the two time points as opposed to the pooling of the data. This allows the correlation to be accounted for in the analysis by calculating an appropriate standard error estimate. Specifically, the standard deviation of the differences between time 1 and time 2 is calculated as follows:

$$s_d = \sqrt{s_1^2 + s_2^2 - 2\sigma_{12}} \tag{2.7}$$

where σ_{12} is the covariance of the measurements at time 1 and time 2. It is through this term that the correlation enters the picture, i.e., $\sigma_{12} = \rho_{12} * s_1 s_2$ where $\rho_{12} = 0.9117$ is the correlation coefficient for the two time groups. Thus, the test statistic for this example is

$$t = \frac{-8.8 - 0}{\frac{4.6}{\sqrt{5}}} = -4.27 \tag{2.8}$$

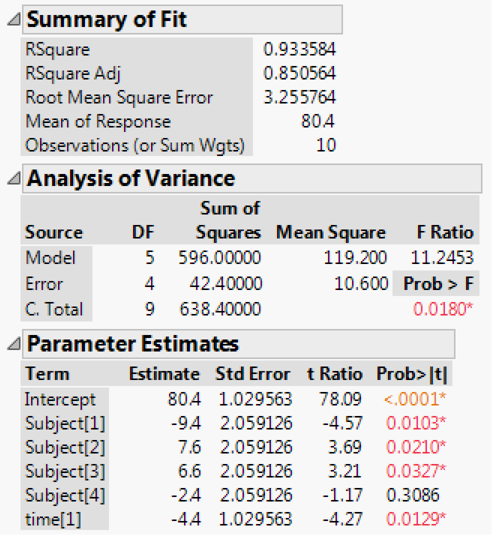
which is different than the test statistic that was found using an unpaired test. In this case, the test statistic found using paired samples is more extreme than that found with unpaired samples. Even though both tests lead to the rejection of the null hypothesis in this particular example, this is not always the case. Failing to account for correlation could be the difference between rejecting the null hypothesis or failing to do so. If the correlation is positive, the power of the study is lower than anticipated while if the correlation is negative, the study is not as significant as anticipated.

2.2.2 Linear Models Approach: Two-way ANOVA

An analagous approach to the paired t-test is the use of a Two-way Analysis of Variance model. In a Two-way ANOVA, the goal is to determine the effect of two factors on a continuous response variable. Similar to the One-way ANOVA, one of the factors in this model will be the time group. However, this model will also include a second factor: the subject. This helps account for the correlation in the data and allows for the direct study of the change in heart rate from time 1 to time 2. Written mathematically, the univariate form of the model is given by

$$Y_{ijk} = \mu + \alpha_i + s_j + e_{ijk} \text{ with } i = 1, 2 \text{ and } j = 1, \cdots, 5 \text{ and } k = 1 \tag{2.9}$$

where s_j represents the subject effect. Again using JMP to run the model, the following output is produced:



Summary of Fit

RSquare	0.933584
RSquare Adj	0.850564
Root Mean Square Error	3.255764
Mean of Response	80.4
Observations (or Sum Wgts)	10

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	5	596.00000	119.200	11.2453
Error	4	42.40000	10.600	Prob > F
C. Total	9	638.40000		0.0180*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	80.4	1.029563	78.09	<.0001*
Subject[1]	-9.4	2.059126	-4.57	0.0103*
Subject[2]	7.6	2.059126	3.69	0.0210*
Subject[3]	6.6	2.059126	3.21	0.0327*
Subject[4]	-2.4	2.059126	-1.17	0.3086
time[1]	-4.4	1.029563	-4.27	0.0129*

Figure 2.2: JMP: Two-way ANOVA table, parameter estimates, and summary statistics

In the parameter estimates, the t ratio for the variable time is -4.27, which is exactly the test statistic that was calculated using the traditional approach. This shows that both approaches will yield the same result. The advantage of using the Two-way ANOVA approach is being able to examine the source of variation in the ANOVA table. In the One-way ANOVA table produced earlier, the model only had one degree of freedom and the majority of the variation was due to random error. In the Two-way situation here, the model has five degrees of freedom and accounts for more of the variation. This shows how including the subject effect in the model allows for a more accurate examination of the variance and a direct study of the change in heart rate over time.

3 Correlation in Longitudinal Data with Multiple Repeated Measures

In the previous section, the effect of correlation was examined when the study was longitudinal with measurements taken at two points in time. In this section, a more complex study with four repeated measures will be examined. The data used in this section is taken from Fitzmaurice[4] and is a study of the level of lead in the blood of 50 exposed children at four different times. A snapshot of the data is shown in Table 3.1 below.

Subject	time0	time1	time2	time3
1	26.5	14.8	19.5	21
2	25.8	23	19.1	23.2
3	20.4	2.8	3.2	9.4
4	20.4	5.4	4.5	11.9
5	24.8	23.1	24.6	30.9
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
mean: $\bar{X}_i$	26.54	13.522	15.514	20.762
variance: s_i^2	25.21	58.867	61.657	85.495

Table 3.1: lead levels in blood of exposed children

With this data, the hypothesis of primary interest is

$$H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3$$

$$H_A : \text{At least one of the population means differs.}$$

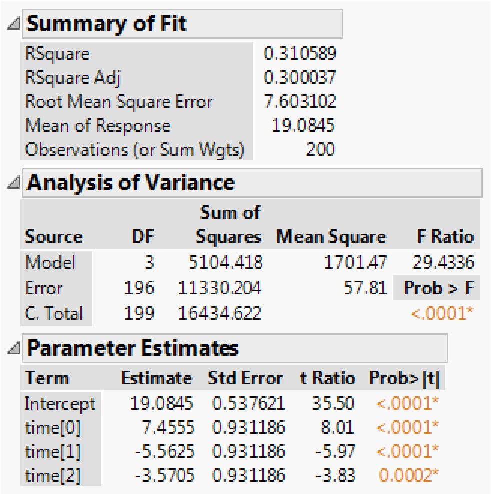
where μ_i is the population mean of the blood lead level at time i. As in the previous section, this data set will be analyzed using a few different techniques. First, the familiar One-way and Two-way ANOVA approaches will be implemented. Then, random effects will be introduced to develop a mixed effects model.

3.1 One-Way Analysis of Variance

The simplest and probably most common approach to analyzing the difference between groups is to use a One-way ANOVA model. In a One-way ANOVA, the effect of one nominal factor on a continuous dependent variable is examined. In the context of the lead-exposure study, the nominal factor is *time* and the dependent variable is *lead level*. Mathematically, the univariate form of the model for this data can be expressed as

$$Y_{ij} = \mu + \alpha_i + e_{j(i)} \text{ with } i = 1, \cdots, 4 \text{ and } j = 1, \cdots, 50 \tag{3.1}$$

where μ is the mean of all 200 observations, α_i is the effect at time i, and $e_{j(i)} \sim$ N$(0, \sigma^2)$ is the random error. Using JMP to run the model, the following output is obtained:



Summary of Fit

RSquare	0.310589
RSquare Adj	0.300037
Root Mean Square Error	7.603102
Mean of Response	19.0845
Observations (or Sum Wgts)	200

Analysis of Variance

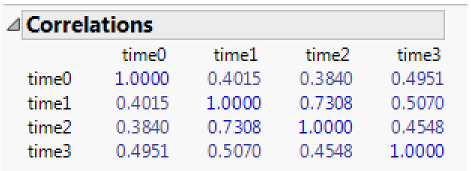
		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	5104.418	1701.47	29.4336
Error	196	11330.204	57.81	Prob > F
C. Total	199	16434.622		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	19.0845	0.537621	35.50	<.0001*
time[0]	7.4555	0.931186	8.01	<.0001*
time[1]	-5.5625	0.931186	-5.97	<.0001*
time[2]	-3.5705	0.931186	-3.83	0.0002*

Figure 3.1: JMP: One-way ANOVA table, parameter estimates, and summary statistics

A careless analyst may look at the F-statistic of 29.4336 in the ANOVA table and immediately conclude that a difference in means exists among the four time points (reject H_0) and begin performing multiple comparisons tests to pinpoint the source of the difference. However, it was shown in the previous section that the One-way ANOVA approach does not account for correlation in the data, so one must be careful before making hasty conclusions. Since there is only one nominal factor with four levels, the model only has three degrees of freedom while the random error term gets the remaining 196 degrees of freedom. This implies that the model only accounts for the variation between groups (i.e., time points) and any other deviation from the group mean is due to random error alone. When there is correlation among the groups, this is a misclassification because the model is ignoring the variation between subjects. Additionally, an R^2 value of 0.310589 implies only about 31% of the variation in the data is accounted for in the model. This is another indication that, even though the F-statistic proved to be significant ($p < 0.001$), there is still a large portion of the variation in the data that has not been modelled. To determine if there is correlation between the time points that needs to be modelled, the correlation matrix can be examined. Figure 3.2 below shows the correlation matrix produced by JMP.



Correlations

	time0	time1	time2	time3
time0	1.0000	0.4015	0.3840	0.4951
time1	0.4015	1.0000	0.7308	0.5070
time2	0.3840	0.7308	1.0000	0.4548
time3	0.4951	0.5070	0.4548	1.0000

Figure 3.2: JMP: correlation matrix for lead-exposure data

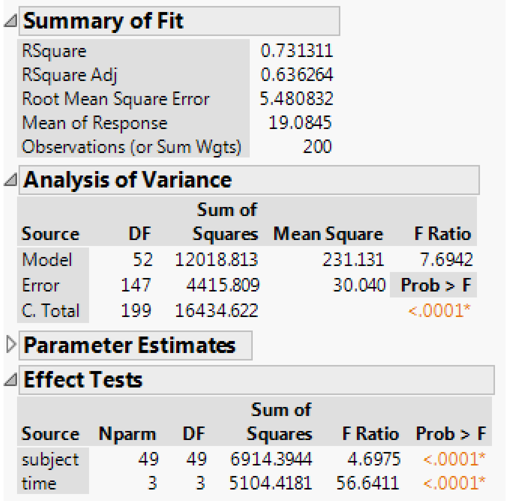
Each entry in Figure 3.2 is the correlation between the group in the row and the group in the column. In this matrix, it appears that a positive correlation exists between each pair of the four time points. This correlation must be accounted for in the analysis, meaning the One-way ANOVA approach is incorrect. To account for the correlation, a Two-way ANOVA model may be more appropriate.

3.2 Two-Way Analysis of Variance

In a Two-way ANOVA model, a second nominal factor is included in the model to help explain the variance in the continuous dependent variable. In the lead-exposure example, the second factor to be added is the subject effect. For this example, subject will be treated as a fixed effect. Entering this term in the model acknowledges the fact that a possible source of variation is between the subjects in the study. In its univariate form, the model is

$$Y_{ijk} = \mu + \alpha_i + s_j + e_{k(ij)} \text{ with } i = 1, \cdots, 4 \text{ and } j = 1, \cdots, 50 \text{ and } k = 1 \qquad (3.2)$$

where s_j represents the subject effect. The output produced by JMP is shown below in Figure 3.3.



Summary of Fit

RSquare	0.731311
RSquare Adj	0.636264
Root Mean Square Error	5.480832
Mean of Response	19.0845
Observations (or Sum Wgts)	200

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	52	12018.813	231.131	7.6942
Error	147	4415.809	30.040	Prob > F
C. Total	199	16434.622		<.0001*

Parameter Estimates

Effect Tests

			Sum of		
Source	Nparm	DF	Squares	F Ratio	Prob > F
subject	49	49	6914.3944	4.6975	<.0001*
time	3	3	5104.4181	56.6411	<.0001*

Figure 3.3: JMP: ANOVA table, effect tests, and summary statistics

When the One-way ANOVA model was used previously, it was stated that the only variation accounted for in the model was due to the difference in time point. The remaining variation in the data was attributed to random error. In the Two-way case here, the model accounts for the variance between time points and between subjects. This is most clearly seen in the ANOVA table in Figure 3.3. Now, the model has 52 degrees of freedom (3 due to the four levels of α and 49 due to the fifty subjects in the study) and accounts for about 73% of the variance in the data ($R^2 = 0.731311$). The F-statistic used for testing the null hypothesis $H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3$ is still significant ($p < 0.0001$), but the variance has been more accurately modelled than it was using a One-way ANOVA model.

3.3 Linear Mixed Effects Model

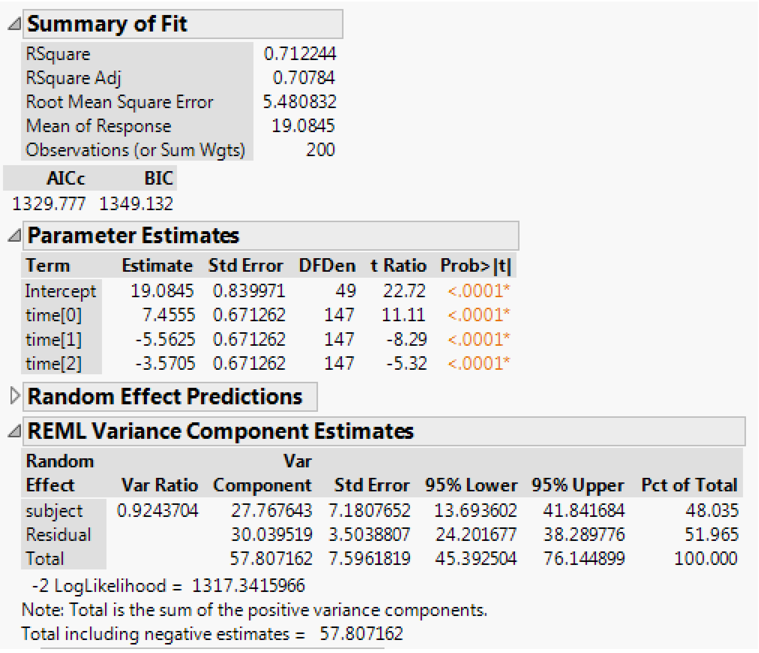
In both the One-way and Two-way ANOVA models, the factors were treated as fixed effects. Here, the lead-exposure study will be analyzed using a linear mixed effects model, treating subject as a *random effect*. In other words, subject represents a random sample from all subjects who satisfy the inclusion criteria for the trial. By considering subject a random effect, the lead level is assumed to vary randomly from one subject to another, meaning that each subject in the study has their own mean response trajectory over time.[4] In contrast, subject was treated as a *fixed effect* in the previous examples. When subject is a fixed effect, it is assumed that the 50 subjects in the study are the only subjects of interest.[6] The general linear mixed effects model is most easily expressed in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{ZU} + \boldsymbol{e} \tag{3.3}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects, **U** is a vector of random effects, **X** and **Z** are design matrices of covariates, and **e** is a vector of residuals. The random effects **U** are assumed to vary following a multivariate normal distribution with mean 0 and covariance matrix V(**U**)=**G** while the random error vector **e** is assumed to vary following a multivariate normal distribution with mean 0 and covariance matrix V(**e**)=**R**. Thus, the covariance structure of $\mathbf{Y}$ can be expressed explicitly as

$$\Sigma = \mathrm{V}(\boldsymbol{Y}) = \boldsymbol{ZGZ}' + \boldsymbol{R} \tag{3.4}$$

where $\mathbf{Z}'$ is the transpose of the matrix $\mathbf{Z}$. In longitudinal applications, $\mathbf{ZGZ}'$ is considered to represent the between-subject portion of the covariance matrix while $\mathbf{R}$ represents the within-subject portion. The proper specification of $\mathbf{G}$ and $\mathbf{R}$ is crucial to obtaining correct estimates of standard error and will be discussed at length in this section. Additionally, it is assumed that $\mathbf{U}$ and $\mathbf{e}$ are independent.[8] The matrix $\boldsymbol{\beta}$ of fixed regression parameters is the same for all subjects and can be interpreted as population-averaged estimates. On the other hand, the matrix of random regression coefficients $\mathbf{U}$ is subject-specific and creates separate mean trajectories for each subject.[4] Before further discussing the specification of the covariance matrix, JMP will be used to run the model for the lead-exposure data, with subject specified as a random effect. Figure 3.4 shows the output produced by JMP.



Summary of Fit

RSquare	0.712244
RSquare Adj	0.70784
Root Mean Square Error	5.480832
Mean of Response	19.0845
Observations (or Sum Wgts)	200

AICc	BIC
1329.777	1349.132

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	19.0845	0.839971	49	22.72	<.0001*
time[0]	7.4555	0.671262	147	11.11	<.0001*
time[1]	-5.5625	0.671262	147	-8.29	<.0001*
time[2]	-3.5705	0.671262	147	-5.32	<.0001*

Random Effect Predictions

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Pct of Total
subject	0.9243704	27.767643	7.1807652	13.693602	41.841684	48.035
Residual		30.039519	3.5038807	24.201677	38.289776	51.965
Total		57.807162	7.5961819	45.392504	76.144899	100.000

-2 LogLikelihood = 1317.3415966

Note: Total is the sum of the positive variance components.

Total including negative estimates = 57.807162

Figure 3.4: JMP: Summary Statistics and REML Variance Components Estimates

The major difference between the output produced for a mixed model and a fixed effects model is the inclusion of the Variance Components Estimates. Here, the Restricted Maximum Likelihood method (REML) was used to produce the estimates found in the table. This information shows how the variance is partitioned between the random effects in the model (in this case, subject) and the random error term (residual). In particular, 48.035% of the variation in the model is due to the clustering of data by subject. This value is known as the *intra-class correlation* or ICC and is calculated as follows:

$$\rho = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_e^2} \tag{3.5}$$

where σ_S^2 is the variance due to subject and σ_e^2 is the variance due to random error. If ρ is close to 1, then the majority of the variance in the model can be explained by the effect of clustering. In the lead-exposure example, clustering occurs within each subject, i.e., each subject's set of four observations forms a cluster. The ICC is also crucial in determining an appropriate sample size for a randomized controlled trial that involves clustering. For example, in a trial to determine if a difference in outcome exists between two treatment groups, the number of observations required in each treatment group is given by

$$N = \frac{2(z_{\alpha/2} + z_\beta)^2 \sigma^2 [1 + (m-1)\rho]}{(\mu_1 - \mu_2)^2} \tag{3.6}$$

where α is the desired level of significance, $1 - \beta$ is the desired power, σ^2 is the variance of the outcome, m is the number of observations per cluster, ρ again is the ICC, and $\mu_1 - \mu_2$ is the minimum difference considered to be clinically significant. Thus, when designing the trial, failing to account for correlation within clusters leads to setting $\rho = 0$ in this equation. This will result in recruiting fewer subjects than is necessary to maintain the desired statistical power. In effect, the term $[1 + (m-1)\rho]$ is the factor by which the sample size must be increased in order for the trial to have the same power as it would if no clustering were present. This term is often referred to as the Variance Inflation Factor (VIF) or the design effect.[15] Since the sample size calculation for a Phase III clinical trial must be done before the trial begins, the value of ρ is typically estimated. The estimate can be obtained from prior testing, like from a Phase II trial, or from empirical evidence. This again shows how failing to account for clustering of data can lead to incorrect inference.[5]

3.3.1 A Brief Overview of PROC MIXED using the SAS® System

Up to this point, JMP has been used to run all models and produce outputs. To provide a little more flexibility when specifying covariance structures in mixed effects models, the SAS System will be used.[12] In particular, programs will be written using the PROC MIXED procedure in SAS. Since JMP is simply a program that implements SAS, the SAS code used to run models in JMP can easily be produced. The SAS code for the mixed effects model created by default in JMP (Figure 3.4) is listed below. This will be used as a starting point from which other mixed effects models can be written.

```
PROC MIXED DATA=lead_exposure ALPHA=0.05;
CLASS subject time;
MODEL  lead_level =  time / SOLUTION;
RANDOM subject / SOLUTION;
RUN;
```

Explanations for the important statements and options in the code are listed below. In addition to the statements already included in the code, a few more statements and options that will be crucial when modelling covariance structures are discussed.

- CLASS - declares certain variables as nominal

- MODEL - specifies the response variable on the left-hand side of the equal sign and the fixed effects portion of the model, $\boldsymbol{X\beta}$, on the right-hand side

- RANDOM - specifies the random effects portion of the model, $\mathbf{ZU}$ and $\mathbf{G}$
- REPEATED - specifies the structure of $\mathbf{R}$
- TYPE - option for specifying the type of covariance structure to be used
- R, G, and V - options to print the $\mathbf{R}$, $\mathbf{G}$, and $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ matrices in covariance form
- RCORR, GCORR, and VCORR - options to print the $\mathbf{R}$, $\mathbf{G}$, and $\boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$ matrices in correlation form
- SUBJECT - specifies variables whose levels are used in defining $\mathbf{G}$ and $\mathbf{R}$

3.3.2 Modelling the Covariance

In the initial analysis of the lead-exposure data using a linear mixed effects model, JMP was used to run the model. As was shown in the SAS code produced by JMP, the structure of the covariance matrix $\mathbf{V}$ was not explicitly specified. In this section, a few options for modelling the covariance structure to fit the data will be presented. Then, the lead-exposure data will be analyzed by altering the linear mixed effects model to include the specification of the covariance structure. Before continuing, it is important to recall that observations on different patients are assumed to be independent. The correlation within each subject, which has been the theme of this paper, arises because the observations are taken at different times on the same subject. Thus, the structures that will be discussed refer to the covariance pattern of measurements on the same subject.[8] It is for this reason that the REPEATED statement will be used rather than the RANDOM statement. In the following examples, let σ_i^2 be the variance at the ith time point and let σ_{ij} be the covariance between the ith and jth time points. Also, recall that $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ where ρ_{ij} is the correlation between time i and j and

σ_i is the standard deviation of time i. Thus, the covariance structures listed below can be analagously expressed in terms of correlation.

3.3.2.1 Independent

When no covariance structure is specified in SAS for a linear mixed effects model, an *independent* structure is assumed. This is the simplest form of the covariance structure because it assumes that all observations are independent of each other. This creates a matrix consisting of the variance terms along the main diagonal and zeroes in all entries off the diagonal. When assuming homoscedasticity among time points, only one parameter for the structure must be estimated: the variance. Thus, the Independent Covariance Structure can be written as

$$\boldsymbol{V}^{(IND)} = \sigma^2 \boldsymbol{I} = \begin{array}{c c} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} \end{array}$$

where **I** is the identity matrix. The SAS code for running this model is shown below.[6]

```
PROC MIXED ASYCOV NOBOUND DATA=lead_exposure ALPHA=0.05;
CLASS subject time;
MODEL lead_level = time / SOLUTION DDFM=KENWARDROGER;
REPEATED / SUBJECT=subject R RCORR;
RUN;
```

3.3.2.2 Compound Symmetric

When the covariance is assumed to be the same for any pair of observations on the same subject, a *compound symmetric* covariance structure is appropriate. In other words, $\sigma_{ij} = \sigma_{kl}$ for all i, j, k and l. For simplicity, let v represent the covariance between any two time points. Then, in a compound symmetric covariance matrix that assumes homoscedasticity among time points, the off-diagonal elements are all v while the main diagonal elements are σ^2. Thus, there are 2 covariance parameters to be estimated when choosing the compound symmetric structure.

$$\boldsymbol{V}^{(CS)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} \sigma^2 & v & v & v \\ v & \sigma^2 & v & v \\ v & v & \sigma^2 & v \\ v & v & v & \sigma^2 \end{pmatrix} \end{array}$$

In a compound symmetric structure, the terms along the diagonal can be thought of as the *between-subject* variance and the terms off the main diagonal are the *within-subject* variance. Below is the SAS code for running a model of the lead-exposure data using a compound symmetric covariance structure.[6]

```
PROC MIXED ASYCOV NOBOUND DATA=lead_exposure ALPHA=0.05;
CLASS subject time;
MODEL lead_level = time / SOLUTION DDFM=KENWARDROGER;
REPEATED / TYPE=cs SUBJECT=subject R RCORR;
RUN;
```

3.3.2.3 Autoregressive (order 1)

The third structure that will be discussed in this paper is the *autoregressive* covariance matrix. When using this approach, it is assumed that observations on a subject that happen closer together are more highly correlated than observations that happen far apart. In terms of the correlation, $\rho_{ij} = \rho^{|t_j - t_i|}$ where t_i and t_j are the times at which observations i and j were taken. So, since $|\rho| \leq 1$, the correlation between observations at two time points decreases exponentially as the time between them increases. Here, only the variance σ^2 and the correlation ρ need to be estimated.

$$\boldsymbol{V}^{(AR(1))} = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ 1 & \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ 2 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 3 & \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{array}$$

The SAS code for running the lead-exposure data using an autoregressive (order 1) covariance structure is given below.

```
PROC MIXED ASYCOV NOBOUND DATA=lead_exposure ALPHA=0.05;
CLASS subject time;
MODEL lead_level = time / SOLUTION DDFM=KENWARDROGER;
REPEATED / TYPE=ar(1) SUBJECT=subject R RCORR;
RUN;
```

The SAS code for the autoregressive structure is very similar to the code for the compound symmetric structure. The only thing that has changed is the argument for the TYPE option.[6]

3.3.2.4 Unstructured

The final covariance structure that will be discussed is the *unstructured* matrix. This is probably the simplest covariance structure to understand because it places no structure on the covariance matrix at all. In other words, all values for pairwise covariances are allowed.

$$
\boldsymbol{V}^{(UN)} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array}
\begin{array}{c}
\begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\
\begin{pmatrix}
\sigma_0^2 & \sigma_0\sigma_1\rho_{01} & \sigma_0\sigma_2\rho_{02} & \sigma_0\sigma_3\rho_{03} \\
\sigma_0\sigma_1\rho_{01} & \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \sigma_1\sigma_3\rho_{13} \\
\sigma_0\sigma_2\rho_{02} & \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \sigma_2\sigma_3\rho_{23} \\
\sigma_0\sigma_3\rho_{03} & \sigma_1\sigma_3\rho_{13} & \sigma_2\sigma_3\rho_{23} & \sigma_3^2
\end{pmatrix}
\end{array}
$$

While the unstructured covariance matrix is easy to understand, it is difficult to model because there is a parameter to estimate for every pairwise relationship in the data. When the number of observations (and subjects if homoscedasticity is violated) gets large, using the unstructured covariance matrix is inefficient. Thus, it is often used as a starting point when determining if a simpler structure can be used.[6] The SAS code for running a mixed effects model with an unstructured covariance matrix is given below.

```
PROC MIXED ASYCOV NOBOUND DATA=lead_exposure ALPHA=0.05;
CLASS subject time;
MODEL lead_level = time / SOLUTION DDFM=KENWARDROGER;
REPEATED / TYPE=un SUBJECT=subject R RCORR;
RUN;
```

3.3.3 Comparison of Covariance Structures

Now that several covariance structures have been presented, a method for determining which structure best fits the data under examination. Littell asserts that either Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) can be used to compare models with the same fixed effects but different covariance structures.[8] These criteria provide an idea of relative goodness-of-fit for the models produced. The equations for calculating the AIC and BIC of a model are given by

$$\text{AIC} = 2k - 2ln(L) \tag{3.7}$$

$$\text{BIC} = -2ln(L) + kln(n) \tag{3.8}$$

where k is the number of parameters in the covariance matrix, $ln(L)$ is the maximized log-likelihood, and n is the number of subjects. Using these equations, a model with a smaller AIC and BIC indicates a better fit and is considered preferable. The two models will not always agree on the best model and, in fact, the BIC is often preferred because it carries a harsher penalty for increasing the number of parameters than does the AIC. By default, when using SAS to run the models with different covariance structures, k, $-2ln(L)$, AIC, and BIC are all included in the output. Running a linear mixed effects model for the lead-exposure data using each of the four covariance structures discussed above yields the following results in Table 3.2.

Structure	**Cov. Parameters**	$-2ln(L)$	**AIC**	**BIC**
Independent	1	1367.1	1369.1	1371.0
Compound Symmetry	2	1314.6	1318.6	1322.4
Autoregressive	2	1319.4	1323.4	1327.3
Unstructured	10	1280.3	1300.3	1319.5

Table 3.2: goodness-of-fit statistics for lead-exposure data

With the lead-exposure data, it appears that the unstructured covariance matrix is the best choice when comparing the values for either BIC or AIC. Since there are only $n = 50$ subjects in the study and $k = 10$ covariance parameters that must be modelled, the AIC and BIC for the unstructured covariance matrix did not carry enough of a penalty to make a simpler covariance structure preferable. Thus, it has been shown that, even with a simple example of fifty subjects being observed at four different points in time, the default independent covariance structure was not considered the best fit for the data.

Additionally, it was stated previously that modelling the covariance appropriately has a major impact on the standard errors used in parameter estimates of fixed effects. Using an incorrect covariance structure can lead to using incorrect standard errors. When this happens, conclusions made using statistical inference are often incorrect. To illustrate this effect, Table 3.3 contains the standard error term for the parameter estimates of the fixed effects in the lead-exposure data. Since time is a nominal fixed effect, dummy variables were used to distinguish the four different time points where the reference cell was time 3.

Time	**IND**	**CS**	**AR(1)**	**UN**
0	1.0962	1.0962	1.4131	1.1378
1	1.0962	1.0962	1.3022	1.2036
2	1.0962	1.0962	1.0549	1.2736
3 (INTERCEPT)	1.0962	1.0962	1.0717	1.3076

Table 3.3: standard errors of fixed effects parameter estimates

4 Detecting Outliers

One of the objectives of this paper is to examine various strategies for identifying outliers in clustered data. Before discussing different outlier detection techniques and applying them to a case study, it is important to have a strong understanding of what it means for an observation to be an outlier. In the simplest sense, an outlier is an extreme observation. When performing data analysis, data sets are assumed to have been generated by a particular probability model. This assumption is what allows statisticians to perform hypothesis tests and make statistical inferences about the data. When a subset of observations within a data set appear to be inconsistent with the assumed probability model, they are considered *suspect values*. Then, statistical tests can be carried out to determine if it is statistically unreasonable to assume that the suspect values belong to the assumed probability model. Observations that are deemed to have been generated from a different probability model are then referred to as *outliers*.[7] There are many different techniques for determining whether or not a particular observation is an outlier. Most of these techniques involve quantifying an observation's standardized distance from either a sample mean or predicted value or determining the influence the observation has on the overall fit of a statistical model. In this section, a few approaches to detecting outliers will be introduced and discussed. In the proceeding section, these techniques will be applied to real data from a multi-center, randomized controlled trial.

4.1 Outlier Detection in Models with a Univariate Response

When working with one variable within a data set, it is quite simple to determine which observations can be suspected of being outliers. A general rule of thumb for identifying suspect values that is introduced in many basic statistics courses is the $1.5 * IQR$ *rule* or the *interquartile range rule*. The interquartile range (IQR) for a data set is found by calculating the difference between the upper and lower quartile of the data. Any points that are farther than $1.5 * \text{IQR}$ from the median of the data set (in either direction) are considered suspect values. For example, consider the data set of commute times (in minutes) for 12 people shown in Table 4.1.

5	17	21	22
24	29	29	30
31	31	33	36

Table 4.1: commute times (in minutes)

For this data set, the median is 29, the upper quartile is 31, and the lower quartile is 20.5. Thus, the interquartile range is calculated as $\text{IQR} = Q_3 - Q_1 = 31 - 20.5 = 10.5$, which means any observation greater than $29 + 1.5 * 10.5 = 44.75$ or less than $29 - 1.5 * 10.5 = 13.25$ will be considered a suspect value. It is clear that the first observation of 5 is the only suspected value. This is most clearly seen visually with the boxplot shown in Figure 4.1. Here, the ”whiskers“ of the boxplot extend in either direction to the maximum or minimum data point that is within $1.5 * \text{IQR}$ of the median. The outlying observation with a value of 5 is the lone point that falls outside of the box and whiskers.

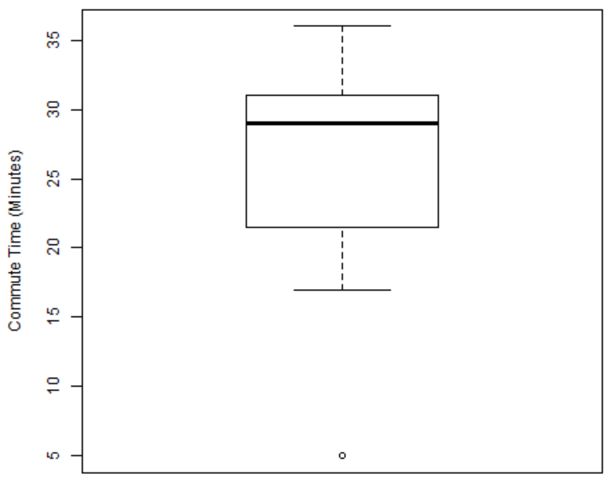


Figure 4.1: boxplot of commute times

While this approach for finding suspect values is very easy to implement, it is simply a rule of thumb and cannot be used as the only method for detecting outliers. Additionally, it is only applicable when analyzing the distribution of each variable separately. When modelling the relationship between a univariate response variable and a set of independent variables, this technique will not be of much use.

A more statistically sound, yet simple method for detecting outliers in the univariate case is through the examination of studentized residuals. When modelling a relationship between a univariate response variable and a set of explanatory variables, a *residual* is defined as the difference between the expected value and the observed value for a particular input of each explanatory variable in the model. This essentially gives an idea of how far off the model is at a given point. Since different data sets often have different variance, it is often useful to standardize residuals for easier interpretation. One method of standardization is to *studentize* the residuals; that is, convert each residual so that it follows a Student's t-distribution with $n - k - 1$

degrees of freedom (where k is the number of explanatory variables in the model). Thus, if the residual falls within the rejection range of the Student's t-distribution, the corresponding observation is considered an outlier. The studentized residual for the ith observation in a data set can be calculated using the following equation:

$$r_i = \frac{\hat{y}_i - y_i}{\text{RMSE}\sqrt{1 - h_i}} \tag{4.1}$$

where $\hat{y}_i$ is the predicted value of Y at observation i, y_i is the observed value of Y at observation i, RMSE is the root of the mean square error, and h_i is the leverage of the observation.[6] Rather than calculating the studentized residual for each observation by hand, most software packages will provide the values for all observations at once. This makes outlier detection in the univariate case fairly straightforward. However, the central focus of this paper has been on clustered data arising from longitudinal studies. Now, a few techniques for detecting outliers from such studies will be introduced.

4.2 Outlier Detection in Clustered Data

When dealing with clusters of data, an investigator is often interested in learning which clusters are outliers in addition to which observations are outliers. To find single outlying observations, analysis of studentized residuals or similar approaches may be used. However, identifying outlying clusters introduces a new challenge since multiple observations make up a cluster. A graphical representation of how observations form clusters is shown in Figure 4.2 to help make this idea a bit more clear. Here, three clusters of observations are shown in 3-dimensional space. This figure represents fictitious data, but is shown for purposes of illustration. The main point that Figure 4.2 aims to illustrate is that comparing clusters of observations is similar to comparing collections of points.

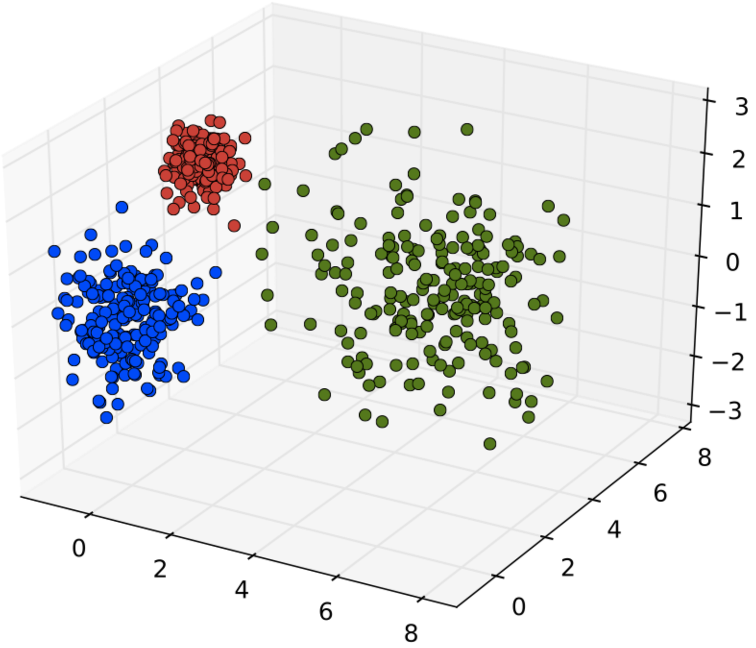


Figure 4.2: three clusters in 3-dimensional space

In a longitudinal study each subject is its own cluster of observations, so identifying outlying clusters equates to identifying outlying subjects. The first three approaches to detecting outlying clusters use the restricted likelihood distance, Cook's Distance, and the predicted residual sum of squares (PRESS). Each of these three approaches revolves around determining how the fit of a model is affected by removing a particular cluster from the data set and are available using the PROC MIXED procedure in the SAS System.

For each subject in the data set, the *restricted likelihood distance* (LD) is calculated by finding the difference between the maximized log likelihood functions of a model that includes the subject and one that does not. Mathematically, the restricted likelihood distance for the ith subject can be expressed as

$$\mathrm{LD}_i = 2[L(\hat{\theta}) - L(\hat{\theta}_{(i)})] \tag{4.2}$$

where $L(\hat{\theta})$ is the maximized log likelihood function of the full model and $L(\hat{\theta}_{(i)})$ is the maximized log likelihood function of the model with the ith cluster removed. In effect, this statistic yields the influence the ith subject has on the likelihood function. Subjects with large values for LD are considered to be suspect.[16]

A second measure of influence that can be utilized in the analysis of clustered data is *Cook's Distance* or, more commonly, Cook's D. As with the restricted likelihood distance, Cook's D measures the effect of removing a particular cluster of observations on the fit of the model. The calculation for Cook's D is given by

$$D_i = \frac{\boldsymbol{e_i^2 h_i}}{\mathrm{MSE} * p(1 - \boldsymbol{h_i})^2} \tag{4.3}$$

where $\boldsymbol{e_i}$ is the residual vector of the ith cluster, $\boldsymbol{h_i}$ is the leverage vector of the ith cluster, p is the number of observations within the cluster, and MSE is the mean squared error. Again, a large value for Cook's D typically indicates a suspect cluster. A general rule of thumb is that if $D_i > \frac{4}{n}$ where n is the number of clusters in the data set, then the subject can be considered an outlier.[3]

Another method that is useful in detecting outliers in clustered data is the *predicted residual sum of squares* (PRESS) statistic. The PRESS statistic is a measure of fit of a model based on how removing a particular subject from the data set impacts the residual sum of squares of the predicted values. To see the influence the ith subject has on the fit of the model, it is first removed from the data set. A new model is then fit using the remaining subjects. Finally, predicted values are found for the subject that was removed using on the newly fit model. This residual is referred to as the PRESS residual and is the figure of interest when determining the influence of each subject individually.

$$\text{PRESS}_i = y_{ij} - \hat{y}_{ij,(-i)} \tag{4.4}$$

where y_{ij} is the observed value of Y for the ith subject at the jth time point and $\hat{y}_{ij,(-i)}$ is the predicted value of Y for the ith subject at the jth time point obtained from the model that was fit excluding the ith subject. When interested in determining the overall fit of a model or the influence of multiple subjects, this process is performed iteratively and the PRESS statistic is the sum of the squared residuals of the excluded subjects' predicted values and their observed values.[14]

The final approach to detecting outlying clusters that will be discussed is using the *Mahalanobis distance*. Simply put, the Mahalanobis distance is a measure of standardized distance between two points in multiple dimensions. This can be viewed as a multivariate generalization of finding the standardized difference between two points that was discussed previously with studentized residuals. The equation for the Mahalanobis distance between two k-dimensional points $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ from a data set D is given by

$$M(\boldsymbol{x_1}, \boldsymbol{x_2}) = \sqrt{(\boldsymbol{x_1} - \boldsymbol{x_2})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x_1} - \boldsymbol{x_2})} \tag{4.5}$$

where $\boldsymbol{\Sigma}^{-1}$ is the $k \times k$ inverse of the covariance matrix of D.[10] When the data set under examination is clustered, the Mahalanobis distance between two clusters can be found with this equation. For example, in the lead-exposure data, the two points would be 4 dimensional vectors, representing the four measurements from two different subjects. To determine if a particular subject is an outlier, the Mahalanobis distance between that subject's vector of observations and the vector of means must first be found. The square of the Mahalanobis distance follows a χ^2 distribution with k degrees of freedom, so a critical value can be obtained to test whether or not the subject is an outlier.[9] A computational implementation of the Mahalanobis distance will be shown in the next section when analyzing real data.

It is worth reiterating the difference between the Mahalanobis distance approach and the first three approaches discussed. As was mentioned earlier, the first three approaches aim to quantify how the fit of a model changes when the subject or cluster under examination is removed. These calculations rely on the residuals between predicted values and observed values. To get predicted values, a model must be fit. On the other hand, the Mahalanobis distance in the context presented here considers only the observed values for a particular subject. These values are then compared to the vector of mean values. Thus, using the Mahalanobis distance in this way can help identify suspect subjects before a model is even fit.

5 Linear Mixed Effects Models: A Case Study

In this section, the topics introduced in the previous sections will be implemented to perform an in-depth analysis of the data from a multi-center, randomized controlled trial with a longitudinal design. The aim of this particular study was to determine if introducing a probiotic to babies with colic decreases the amount of time they spend crying. There had been prior Phase II randomized controlled trials that favored probiotics as a potential aid in reducing crying time. According to Mayo Clinic, colic is a condition identified by predictable periods of distress in an otherwise healthy baby. Babies with colic often cry for over three hours a day, three days per week for several weeks but tends to end after a few weeks or months. During the time when a baby has colic, however, it is very difficult to bring him or her any relief.[2]

This study was conducted at four different centers on a total of 292 babies with colic. At each center, babies were randomized into either the probiotic treatment group or to the placebo group. The amount of time each baby spent crying was reported by the baby's mother in a daily journal. Investigators then averaged the crying times for each baby by week at baseline, after one week, after two weeks, and after three weeks. The primary objective of the study was to determine if the crying time of the probiotic group decreased over time significantly more than the placebo group. Table 5.1 contains the raw mean daily crying time (in minutes) for the two treatment groups at each time point.

Treatment Group	Time (Days)			
	0	7	14	21
Placebo	201.95	155.80	127.89	110.71
Probiotic	215.20	127.39	101.32	77.51

Table 5.1: mean crying times at each time point

5.1 Developing the Model

Before proceeding with model development and analysis, it should be noted that this data set poses a couple of major analytical challenges. First, the data are longitudinal; each subject has four repeated measures for the response. This creates a cluster for each subject and introduces the need for modelling the correlation between observations. Overcoming this particular hurdle has been the primary focus of this paper up to this point. However, the subjects in this study were randomized into one of two treatment groups (probiotic or placebo) at one of four centers, which creates a nesting effect for the subjects. While the primary goal of this analysis is to determine whether or not there is a difference in response between the probiotic group and the placebo group over time, the challenges mentioned here must be taken into account in order to make proper statistical inference and draw meaningful conclusions.

In this analysis, the data will be examined using a linear mixed effects model with a random subject effect. The univariate form of the model can be expressed mathematically as

$$Y_{ijklm} = \mu + \alpha_i + \gamma_j + (\alpha * \gamma)_{ij} + \tau_k + (\alpha * \tau)_{ik} + (\gamma * \tau)_{jk} + s_{l(ijk)} + e_{m(ijkl)} \quad (5.1)$$

Each of these terms is outlined in Table 5.2 for clarity. The primary goal will be to make inference about $\gamma * \tau$ and γ in order to determine if there is a difference between treatment groups over time.

Symbol	**Variable**
Y	crying duration (response)
α	center
γ	trt_gp
$\alpha * \gamma$	center*trt_gp interaction
τ	time
$\alpha * \tau$	center*time interaction
$\gamma * \tau$	trt_gp*time interaction
s	random subject effect
e	random error term

Table 5.2: linear mixed effects model variables

As was shown in section 4, modelling the covariance structure is a crucial aspect of making proper statistical inference when random effects are involved. In this analysis, subject is a random effect and hence the covariance must be modelled. First, the SAS code for the basic structure of the mixed effects model is shown below. Here, a random intercept is estimated for each subject to give each subject its own mean trajectory.

```
PROC MIXED DATA=multi_center ALPHA=0.05;
CLASS center ID trt_gp time;
MODEL crying_duration = center trt_gp time trt_gp*time
         center*trt_gp center*time;
RANDOM intercept / SUBJECT=ID(center trt_gp );
RUN;
```

Here, no covariance strucure has been declared, so the independent structure is chosen by default. This structure assumes observations at different times for the same subject are pairwise independent. Since the observations are taken over time on the same subject, this is not likely to be the proper structure but is the most conservative approach. Nonetheless, it provides a starting point from which other models can be built. In Table 5.3, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are shown for the model using a few different covariance structures.

Structure	**Cov. Parameters**	$\boldsymbol{-2ln(L)}$	**AIC**	**BIC**
Independent	2	12639.8	12643.8	12651.2
Compound Symmetry	3	12639.8	12645.8	12656.9
Autoregressive	3	12606.4	12612.4	12623.4
Unstructured	11	12556.8	12578.8	12619.2

Table 5.3: goodness-of-fit statistics (lower values indicate better fit)

With this model, the values for the BIC and AIC agree that the unstructured covariance matrix best suits the data. It is important to keep in mind that this is not always the case. As the number of repeated measures grows, so too does the number of covariance parameters in the unstructured covariance matrix, driving up the values for AIC and BIC. This is because an unstructured covariance matrix requires a covariance parameter to be estimated for each entry in the matrix. Simpler structures, such as the autoregressive or compound symmetric structures, require fewer parameters to be estimated. However, in this case with four repeated measures, there are not enough covariance parameters to make a simpler structure preferable. Thus, the analysis will proceed assuming an unstructured covariance matrix for the subjects. The SAS code for the model is given below followed by the covariance matrix $\boldsymbol{V}$.

```
PROC MIXED DATA=multi_center ALPHA=0.05;
CLASS center ID trt_gp time;
MODEL crying_duration = center trt_gp time trt_gp*time
          center*trt_gp center*time / S;
RANDOM intercept / SUBJECT=ID(center trt_gp ) G GCORR
```

```
        V VCORR;
REPEATED / SUBJECT=ID(center trt_gp ) TYPE=un R RCORR;
RUN;
```

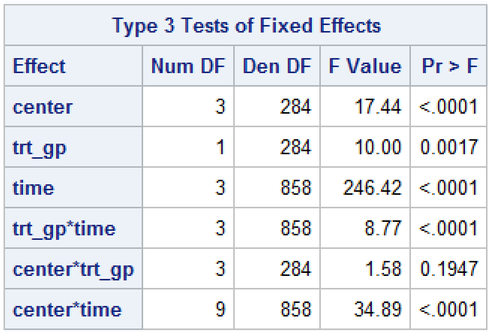
$$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} = \begin{array}{c c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 6594.52 & 3310.38 & 2686.41 & 2500.29 \\ 3310.38 & 4946.56 & 3166.80 & 2648.51 \\ 2686.41 & 3166.80 & 4762.72 & 3087.14 \\ 2500.29 & 2648.51 & 3087.14 & 4025.80 \end{pmatrix} \end{array}$$

Appropriately modelling the covariance leads to obtaining correct estimates of the standard error terms which ultimately results in making correct inference. Had this step been skipped, standard errors for testing various hypotheses would have been incorrect. In the context of medical research, this can have serious consequences. If a standard of treatment is changed (or fails to change) due to research that was analyzed incorrectly, patients may end up receiving suboptimal care. To see this, consider the table shown below. Table 5.4 shows a subset of the solutions for the fixed effect parameters obtained when assuming the independent covariance matrix and the unstructured covariance matrix. Each row in the table is a test of whether or not the coefficient of the corresponding parameter is statistically different from zero. It is quite clear from Table 5.4 that the standard error estimates do change. Since the statistic used in the hypothesis test of each row is simply the estimate divided by the standard error estimate, changes in the standard error can lead to changes in conclusions of statistical significance.

	Unstructured			**Independent**		
Effect	Estimate	St. Error	p-value	Estimate	St. Error	p-value
Intercept	68.8263	12.4486	$< .0001$	68.0902	13.4738	$< .0001$
center[1]	10.4374	15.6274	0.5047	12.2577	16.7818	0.4657
center[2]	-4.9703	17.3396	0.7746	-2.2496	18.6694	0.9042
center[3]	16.8045	14.6569	0.2525	16.2641	15.7439	0.3025
trt_gp[placebo]	36.6082	16.3544	0.026	37.9753	17.0015	0.0263
time[0]	53.5867	11.4216	$< .0001$	53.5867	10.0692	$< .0001$
time[7]	19.2648	9.2367	0.0373	19.2648	10.0692	0.056
time[14]	8.665	7.7953	0.2666	8.496	10.0747	0.3993

Table 5.4: fixed effect solutions using different covariance structures

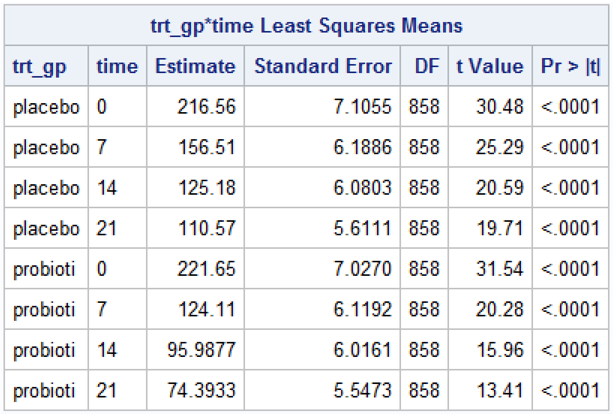
Now that the covariance structure has been modelled, proper inference can be made. Since the primary objective of the study is to determine the effect of the probiotic over time, the variables of primary interest are *trt_gp* and *trt_gp*time*. To determine whether or not trt_gp has a significant effect on crying time, the Type 3 test of the fixed effect can be used. Figure 5.1 the SAS output for performing the Type 3 Tests of Fixed Effects. With a p-value of .0017, it is clear that the treatment group does seem to have an effect on crying time. Also, since the estimate for the effect the placebo group has on crying time is 36.61 (shown in Table 5.4 above), it appears the probiotic did lead to a reduction in crying time.



Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
center	3	284	17.44	<.0001
trt_gp	1	284	10.00	0.0017
time	3	858	246.42	<.0001
trt_gp*time	3	858	8.77	<.0001
center*trt_gp	3	284	1.58	0.1947
center*time	9	858	34.89	<.0001

Figure 5.1: type 3 tests of fixed effects

An effective way to study the effect of the treatment group over time is to consider the least squares means at each level of trt_gp*time and how they differ from each other. If the probiotic is effective in reducing crying time, one would expect to see the difference in least squares means between the treatment groups be small at time 0 and then grow over time. The manner in which this difference changes over time will give a reasonable picture of what effect the probiotic has on crying time. In the three figures that follow, the least squares mean estimates will be explored. In Figure 5.2, the least squares mean estimates for each level of trt_gp*time are shown. Figure 5.3 graphs the least squares means by treatment group. This gives an intuitive visual for how crying duration changes over time in each treatment group. Finally, Figure 5.4 shows the estimated difference in least squares means between levels of trt_gp*time along with corresponding 95% confidence intervals.



trt_gp*time Least Squares Means						
trt_gp	time	Estimate	Standard Error	DF	t Value	Pr > |t|
placebo	0	216.56	7.1055	858	30.48	<.0001
placebo	7	156.51	6.1886	858	25.29	<.0001
placebo	14	125.18	6.0803	858	20.59	<.0001
placebo	21	110.57	5.6111	858	19.71	<.0001
probioti	0	221.65	7.0270	858	31.54	<.0001
probioti	7	124.11	6.1192	858	20.28	<.0001
probioti	14	95.9877	6.0161	858	15.96	<.0001
probioti	21	74.3933	5.5473	858	13.41	<.0001

Figure 5.2: least squares means estimates for each level of trt_gp*time

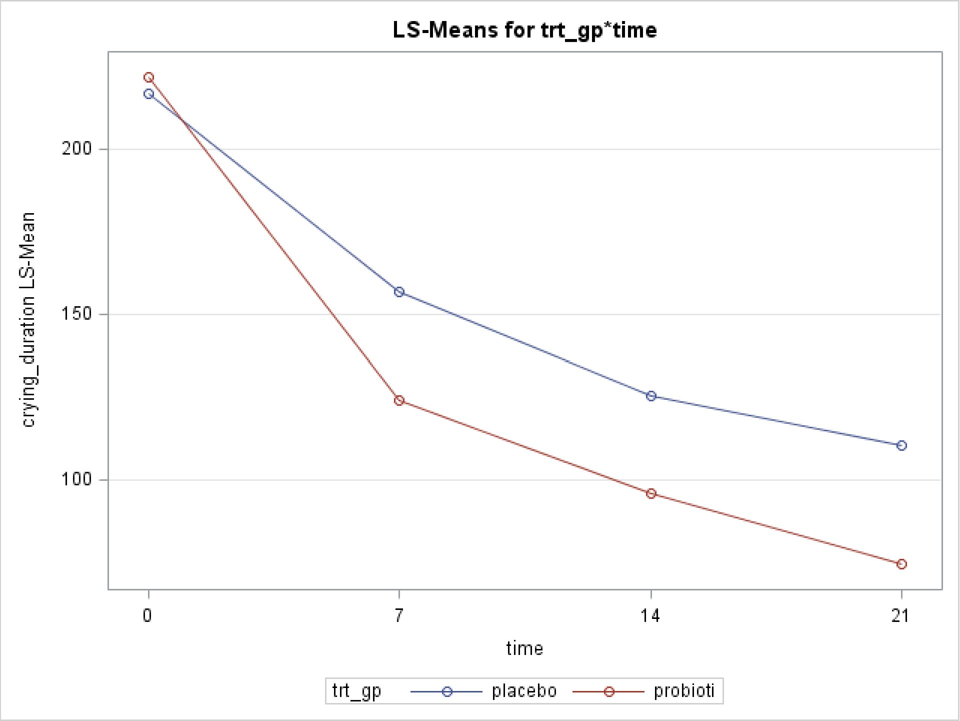
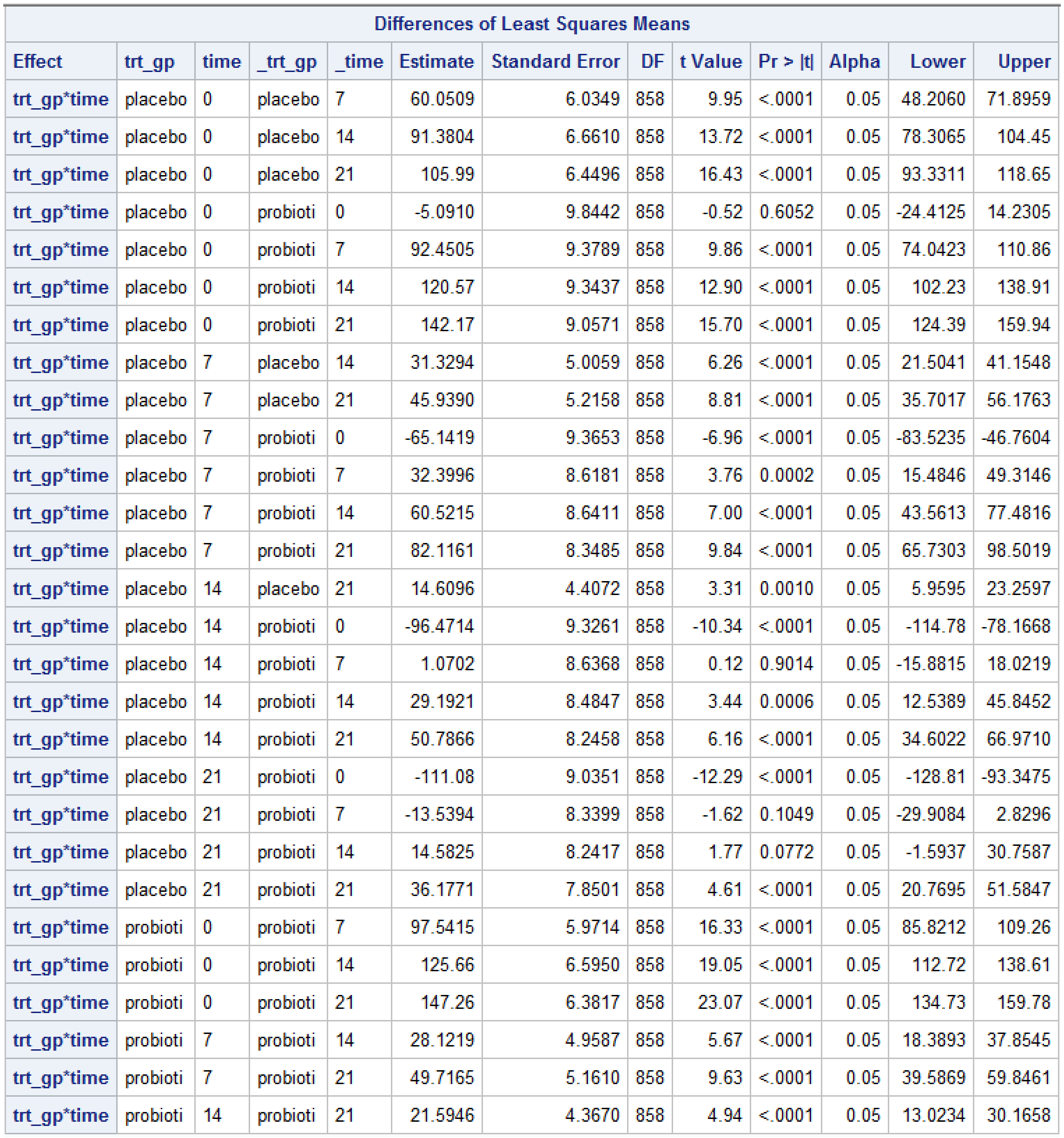


Figure 5.3: least squares means plot for trt_gp*time



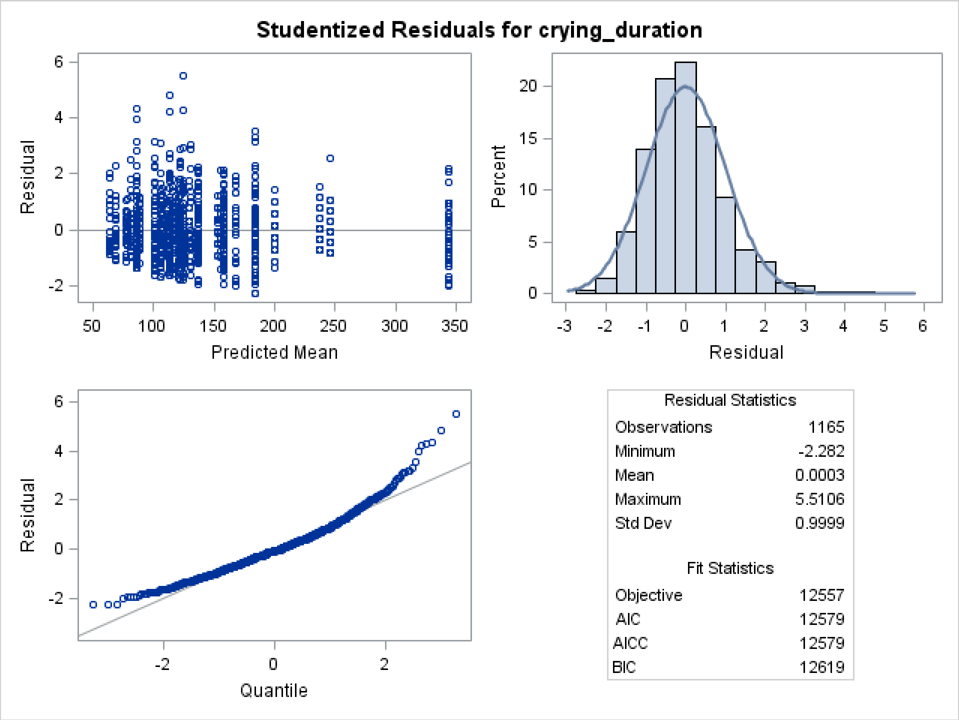
Differences of Least Squares Means												
Effect	trt_gp	time	_trt_gp	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
trt_gp*time	placebo	0	placebo	7	60.0509	6.0349	858	9.95	<.0001	0.05	48.2060	71.8959
trt_gp*time	placebo	0	placebo	14	91.3804	6.6610	858	13.72	<.0001	0.05	78.3065	104.45
trt_gp*time	placebo	0	placebo	21	105.99	6.4496	858	16.43	<.0001	0.05	93.3311	118.65
trt_gp*time	placebo	0	probioti	0	-5.0910	9.8442	858	-0.52	0.6052	0.05	-24.4125	14.2305
trt_gp*time	placebo	0	probioti	7	92.4505	9.3789	858	9.86	<.0001	0.05	74.0423	110.86
trt_gp*time	placebo	0	probioti	14	120.57	9.3437	858	12.90	<.0001	0.05	102.23	138.91
trt_gp*time	placebo	0	probioti	21	142.17	9.0571	858	15.70	<.0001	0.05	124.39	159.94
trt_gp*time	placebo	7	placebo	14	31.3294	5.0059	858	6.26	<.0001	0.05	21.5041	41.1548
trt_gp*time	placebo	7	placebo	21	45.9390	5.2158	858	8.81	<.0001	0.05	35.7017	56.1763
trt_gp*time	placebo	7	probioti	0	-65.1419	9.3653	858	-6.96	<.0001	0.05	-83.5235	-46.7604
trt_gp*time	placebo	7	probioti	7	32.3996	8.6181	858	3.76	0.0002	0.05	15.4846	49.3146
trt_gp*time	placebo	7	probioti	14	60.5215	8.6411	858	7.00	<.0001	0.05	43.5613	77.4816
trt_gp*time	placebo	7	probioti	21	82.1161	8.3485	858	9.84	<.0001	0.05	65.7303	98.5019
trt_gp*time	placebo	14	placebo	21	14.6096	4.4072	858	3.31	0.0010	0.05	5.9595	23.2597
trt_gp*time	placebo	14	probioti	0	-96.4714	9.3261	858	-10.34	<.0001	0.05	-114.78	-78.1668
trt_gp*time	placebo	14	probioti	7	1.0702	8.6368	858	0.12	0.9014	0.05	-15.8815	18.0219
trt_gp*time	placebo	14	probioti	14	29.1921	8.4847	858	3.44	0.0006	0.05	12.5389	45.8452
trt_gp*time	placebo	14	probioti	21	50.7866	8.2458	858	6.16	<.0001	0.05	34.6022	66.9710
trt_gp*time	placebo	21	probioti	0	-111.08	9.0351	858	-12.29	<.0001	0.05	-128.81	-93.3475
trt_gp*time	placebo	21	probioti	7	-13.5394	8.3399	858	-1.62	0.1049	0.05	-29.9084	2.8296
trt_gp*time	placebo	21	probioti	14	14.5825	8.2417	858	1.77	0.0772	0.05	-1.5937	30.7587
trt_gp*time	placebo	21	probioti	21	36.1771	7.8501	858	4.61	<.0001	0.05	20.7695	51.5847
trt_gp*time	probioti	0	probioti	7	97.5415	5.9714	858	16.33	<.0001	0.05	85.8212	109.26
trt_gp*time	probioti	0	probioti	14	125.66	6.5950	858	19.05	<.0001	0.05	112.72	138.61
trt_gp*time	probioti	0	probioti	21	147.26	6.3817	858	23.07	<.0001	0.05	134.73	159.78
trt_gp*time	probioti	7	probioti	14	28.1219	4.9587	858	5.67	<.0001	0.05	18.3893	37.8545
trt_gp*time	probioti	7	probioti	21	49.7165	5.1610	858	9.63	<.0001	0.05	39.5869	59.8461
trt_gp*time	probioti	14	probioti	21	21.5946	4.3670	858	4.94	<.0001	0.05	13.0234	30.1658

Figure 5.4: differences in least squares estimates for levels of trt_gp*time

Using these three figures, a few conclusions can be made. First, there is no significant difference in least squares means between the placebo group and probiotic group at time 0. This provides some evidence that the randomization of subjects worked as intended because there was no statistical difference in crying times between the two groups when the trial started. Then, at each of the next three time points, there are significant differences in crying time between the two groups, hinting that the probiotic did in fact reduce crying time. From the graph in Figure 5.3, it appears the major change in crying time between the two groups occurred from time 0 to time 7. Both groups saw a decrease in crying time, but the decrease was more drastic in the probiotic group. After time 7, the lines for the two groups are almost parallel, meaning both decreased at a fairly similar rate. This gives some evidence that the probiotic was most effective in reducing crying when first introduced. Then, both groups saw similar decreases in crying which can be attributed to time. This is consistent with the information provided by Mayo Clinic[2] on babies with colic that was mentioned earlier. After a few weeks or months, colic tends to go away. This gradual decrease over time seen in both groups may have been due to certain babies growing out of the colic stage.

5.2 Checking the Assumptions

Now that a model has been chosen and fit for the data, it is important to check that the assumptions for linear mixed effects models hold. The three major assumptions that are relevant to this model are normality of residuals, homoscedasticity, and independence. If the assumption of normality holds, the studentized residuals from the model will follow a (roughly) normal distribution. In Figure 5.5 below, the studentized residuals from the model are shown in a Residual vs. Predicted plot (top left), a histogram (top right), and a Q-Q plot (bottom left). All three plots show sufficient evidence of normality among the studentized residuals, which allow the assumption to be confirmed.



Residual Statistics	
Observations	1165
Minimum	-2.282
Mean	0.0003
Maximum	5.5106
Std Dev	0.9999

Fit Statistics	
Objective	12557
AIC	12579
AICC	12579
BIC	12619

Figure 5.5: studentized residual statistics for colic study

To check for homoscedasticity, plots of the studentized residuals can be used again. However, for this assumption to hold, the residuals will have roughly the same spread for each center. This will show that the variance is approximately equal across centers. Figure 5.6 shows the histograms for the studentized residuals for each center. It appears that the four centers have roughly the same spread of residuals, which allows the assumption of homoscedasticity to be confirmed.

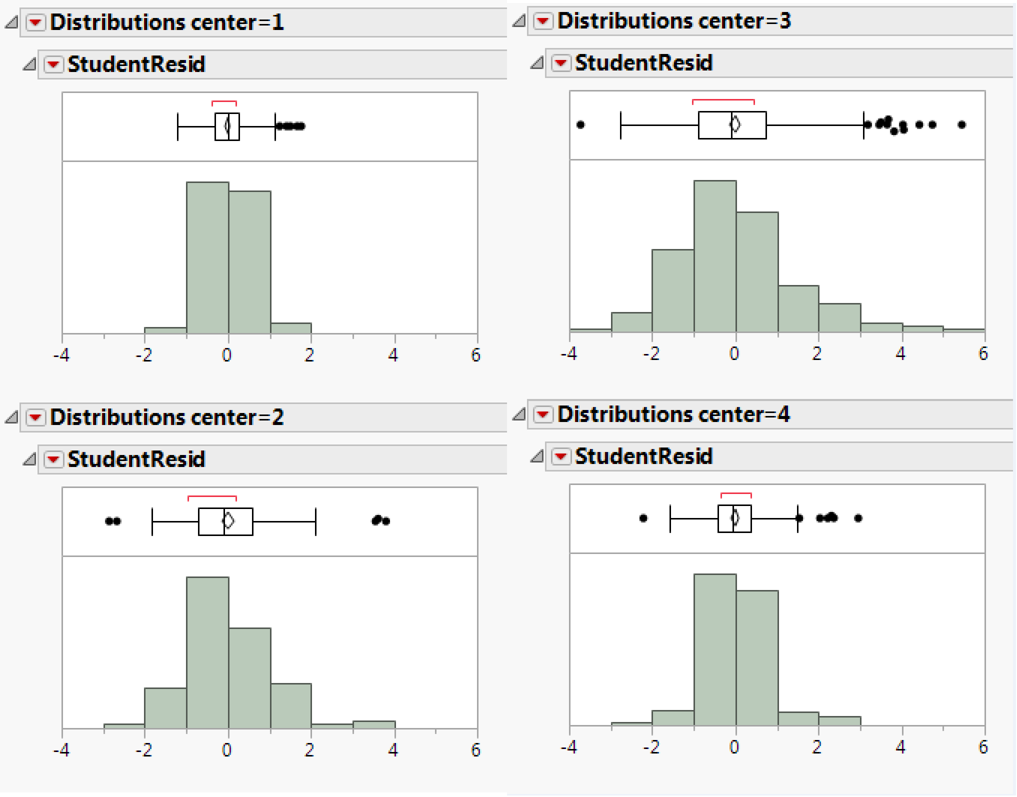


Figure 5.6: histograms for studentized residuals for each center

The assumption of independence is more difficult to confirm for this study. The fact that the observations for one subject are not independent of one another has already been discussed and accounted for through the modelling of the covariance structure. However, the linear models assumption of independence refers to the assumption that the order in which subjects were randomized and received treatment has no impact on the distribution of residuals. In other words, the residual error from one subject should not affect the residual error for the next subject.[1] Checking this assumption boils down to ensuring the randomization scheme was effective and the order of entry into the study had no impact on crying duration. To allow for this assumption to be checked, the order in which subjects are randomized should be recorded and included as a variable in the data set to be analyzed. Then a simple plot of studentized residuals vs. order could be produced to determine if any patterns or trends appear. Unfortunately, this information was not included in this study which means the assumption of independence can not be unequivocally confirmed at this time. For the sake of this analysis, it will be assumed without further justification that the randomization scheme used in the study was effective in preventing experimental dependence.

5.3 Detecting Outliers

At this point, the model has been fit, the covariance structure has been modelled, and inference about the effect of the probiotic over time has been made. To complete the analysis, any outliers will be determined. In this case study, outliers can be found at the observation level or at the subject level. At the observation level, the distribution of studentized residuals can be used to find any observations that fall far enough from the mean to be considered outliers. This method was discussed in Section 4 and will be skipped here because the focus of this paper has been on the clusters of observations that form for each subject. To find subjects that are outlying, the other influence diagnostics discussed in Section 4 can be implemented. Using SAS PROC MIXED, the restricted likelihood distance, Cook's Distance, and PRESS statistic can be found for each subject. The code is shown below. The major statement to make note of in the code is INFLUENCE. Including this option provides the relevant influence statistics as well as intuitive visuals. Here, the effect of interest is subject or ID(center trt_gp). The ITER statement accounts for how removing a subject impacts the covariance structure used in the model by performing a specified number of additional iterations to recompute the structure.[12]

```
PROC MIXED DATA=multi_center ALPHA=0.05;
CLASS center ID trt_gp time;
MODEL crying_duration = center trt_gp time trt_gp*time
          center*trt_gp center*time /
```

```
            INFLUENCE(EFFECT=ID(center trt_gp) ITER=5);
RANDOM intercept / SUBJECT=ID(center trt_gp );
REPEATED / SUBJECT=ID(center trt_gp ) TYPE=un;
RUN;
```

It is important to remember that these methods are used for determining the influence certain subjects have on the fit of the model. Their calculations are based on residuals that can only be obtained after a model is fit. The figures below show graphs for the restricted likelihood distance and Cook's D. Then, the tables that follow show the subjects with the largest values for the three measures of influence.

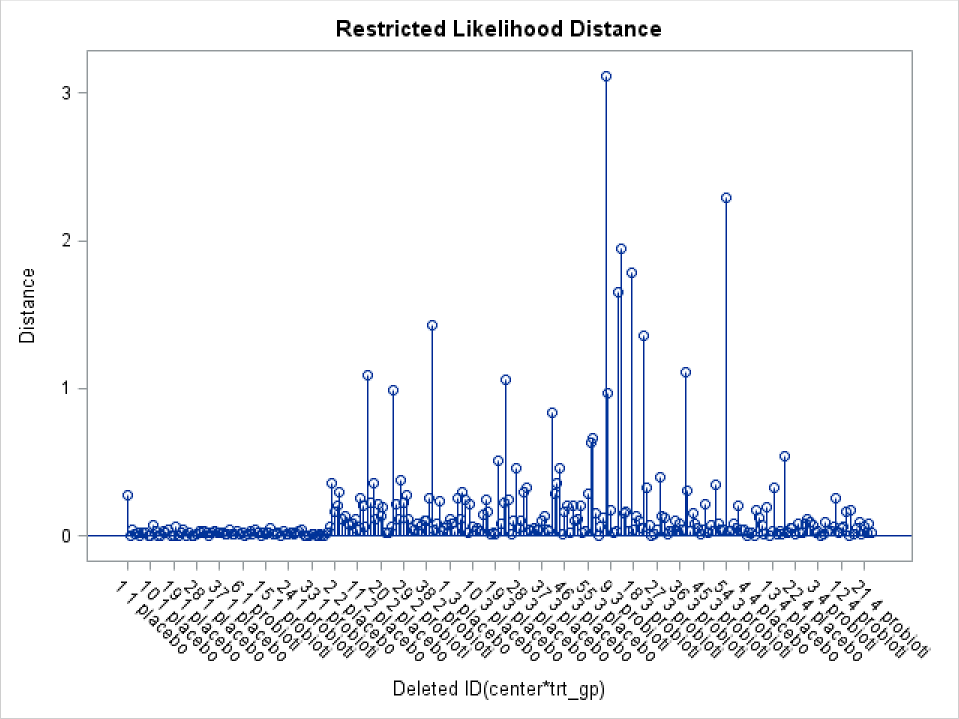


Figure 5.7: restricted likelihood distance for each subject

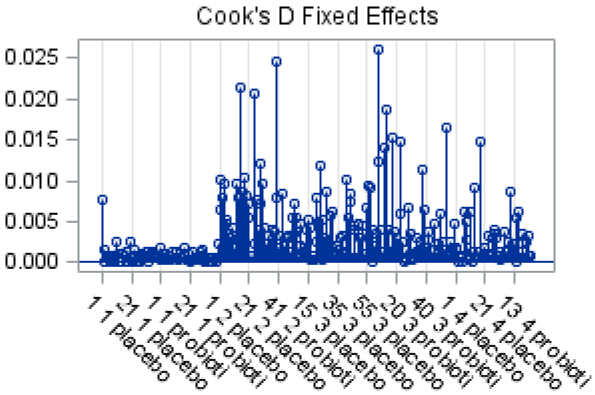


Figure 5.8: Cook's D for each subject

center	id	trt_gp	restricted likelihood distance
3	7	probiotic	3.1152
3	54	probiotic	2.2934
3	13	probiotic	1.9421
3	17	probiotic	1.7819
3	12	probiotic	1.6468
2	40	probiotic	1.4282
3	22	probiotic	1.3568
3	38	probiotic	1.1075
2	15	placebo	1.0882
3	23	placebo	1.063

Table 5.5: subjects with greatest restricted likelihood distance

center	id	trt_gp	Cook's D
3	7	probiotic	0.02604
2	40	probiotic	0.02456
2	15	placebo	0.0215
2	25	probiotic	0.02059
3	13	probiotic	0.01879
3	54	probiotic	0.01651
3	17	probiotic	0.01536
4	18	placebo	0.01477
3	22	probiotic	0.01467
3	12	probiotic	0.01409

Table 5.6: subjects with greatest Cook's D; $D > \frac{4}{n} = .014$ considered outlier

center	id	trt_gp	PRESS Statistic
3	7	probiotic	262878
3	13	probiotic	261239
3	17	probiotic	160234
3	22	probiotic	142293
3	8	probiotic	134210
3	43	placebo	116723
3	1	probiotic	102010
1	1	placebo	88543
2	28	probiotic	84836
3	12	probiotic	81078

Table 5.7: subjects with greatest PRESS

If more information had been recorded for each subject (e.g., height, weight, age, etc.), the distribution of outliers could have been examined in more detail to pinpoint the underlying cause for their status as outliers. Here, however, the only factor that can be considered is the center to which the subject belonged. As a demonstration of how one could analyze these outliers further, the distributions of the Cook's Distances are shown by center in Figure 5.9. It appears that center 3 has more subjects with extreme values for Cook's D than the other three centers. This may indicate that the model does not predict crying times for subjects from center 3 particularly well. This could be due to some natural heterogeneity between the four centers, which could be due to the clustering of subjects within each center. More information on the subjects within the centers would be needed before considering this a red flag.

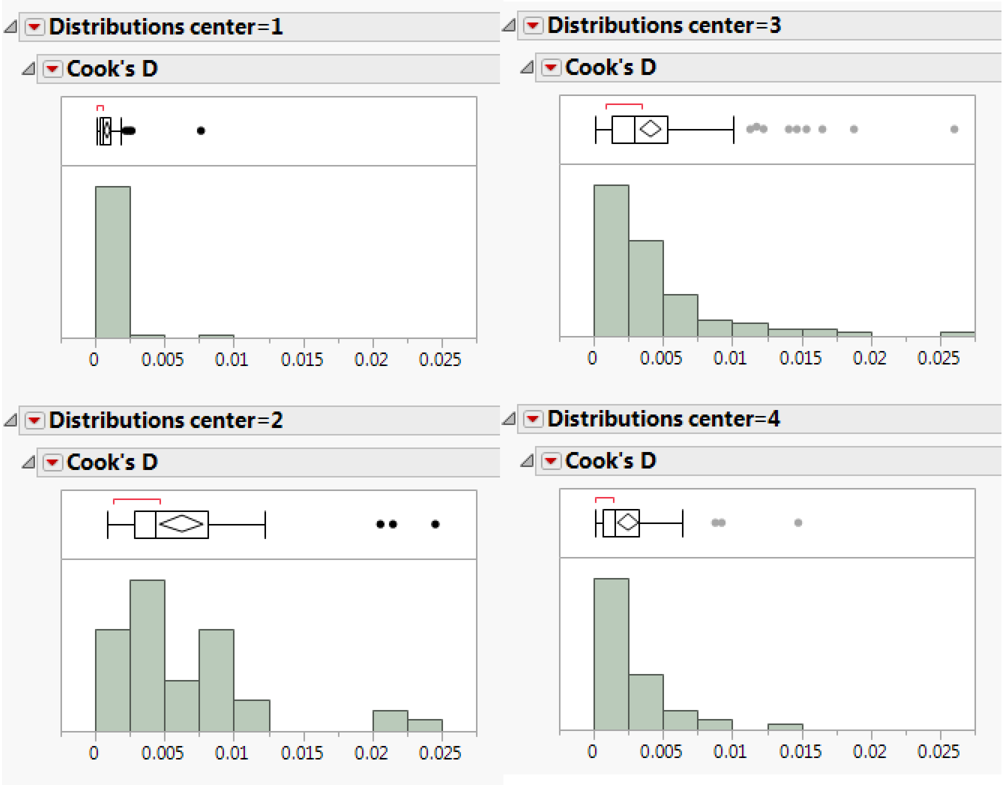


Figure 5.9: distribution of Cook's D for each center

In Section 4, it was also shown how the Mahalanobis distance can be used to identify suspicious subjects before even running a model. This will be implemented for the colic study using R software.[11] The objective is to find the Mahalanobis distance between each subject's vector of observations (time 0, time 7, time 14, time 21) and the vector of mean crying times at the four time points for each treatment group. Then, any subjects whose squared Mahalanobis distance falls in the rejection region of a χ^2 distribution with 4 degrees of freedom at the $\alpha = .025$ level of significance will be considered a suspect value. Then, these subjects will be examined to determine if there are any underlying causes for their large deviations. The relevant parts of the R code is shown below.

```
M_pro <-matrix(c(time0,time7,time14,time21),nrow=148,ncol=4,
          byrow=TRUE)
center_pro <- c(mean(time0),mean(time7),mean(time14),
          mean(time21))
##return squared mahalanobis distance##
mahal_pro <- mahalanobis(M_pro,center_pro,cov(M_pro))
id_pro <- ID
center_pro <- center

M_plac <- matrix(c(time0,time7,time14,time21),nrow=144,ncol=4,
          byrow=TRUE)
center_plac <- c(mean(time0),mean(time7),mean(time14),
          mean(time21))
##return squared mahalanobis distance##
mahal_plac <- mahalanobis(M_plac,center_plac,cov(M_plac))
id_plac <- ID
study_plac <- center


chi_stat <- qchisq(.975,df=4) #finds critical value of chi-sq
```

```
##save suspect observations in table##

suspect.id <- NULL
center.s <- NULL
trt.s <- NULL
mahal.s <- NULL

for(i in 1:length(mahal_pro)){
        if(mahal_pro[i]>chi_stat){
                suspect.id <- c(suspect.id,id_pro[i])
                center.s <- c(center.s,center_pro[i])
                trt.s <- c(trt.s,”pro”)
                mahal.s <- c(mahal.s,mahal_pro[i])
        }
}
for(i in 1:length(mahal_plac)){
        if(mahal_plac[i]>chi_stat){
                suspect.id <- c(suspect.id,id_plac[i])
                center.s <- c(center.s,center_plac[i])
                trt.s <- c(trt.s,”plac”)
                mahal.s <- c(mahal.s,mahal_plac[i])
        }
}
```

Running the program produces the following table which contains the subjects identified as suspects. Of the 36 subjects identified as suspects based on their Mahalanobis distance from the vector of means, 16 come from center 1, 6 come from center 2, 9 come from center 3, and 5 come from center 4. The number of outlying subjects in center 1 is of particular interest because there are only 80 subjects in center 1, meaning about 20% of the subjects are considered suspects. This is a rather large portion of the subjects within the center, indicating that some heterogeneity may exist between the centers. Further examination of each of these points would be required in order to fully understand their impact.

center	trt_gp	ID	Mahal Dist	center	trt_gp	ID	Mahal Dist
1	probiotic	11	13.64980207	3	probiotic	30	20.0792489
1	probiotic	12	22.85530468	4	probiotic	8	15.66687714
1	probiotic	13	11.29542303	4	probiotic	9	21.76780725
1	probiotic	14	14.34232228	4	probiotic	13	13.0016703
1	probiotic	17	17.90265271	1	placebo	11	25.02539749
1	probiotic	18	20.91496731	1	placebo	13	20.17655287
1	probiotic	20	15.85138331	1	placebo	15	17.70247995
1	probiotic	22	20.87020304	1	placebo	16	12.62205258
1	probiotic	23	12.03937911	1	placebo	17	20.4659725
1	probiotic	24	13.18695421	1	placebo	19	12.08646477
2	probiotic	8	14.83340416	2	placebo	10	13.60564029
2	probiotic	12	12.20388722	3	placebo	1	12.81123194
2	probiotic	14	14.79821831	3	placebo	3	11.80528215
2	probiotic	15	33.59425923	3	placebo	28	14.22762295
2	probiotic	17	15.3933404	3	placebo	37	14.61781584
3	probiotic	26	16.4418664	3	placebo	40	14.97477905
3	probiotic	27	12.40911079	4	placebo	9	12.94907788
3	probiotic	29	14.24526802	4	placebo	18	15.09555628

Table 5.8: suspect subjects with Mahalanobis distances

6 Summary

Clustered data arise naturally in several situations. One of the most common sources of clustering is longitudinal data. When repeated measures are taken on the same subject over time, a positive correlation often exists among the measurements and must be considered when performing data analysis. This paper has shown that clustered data needs to be treated with care when making statistical inferences, as failing to account for correlation leads to misleading conclusions. In the simple case where observations were paired (Section 2), adding a factor for subject to the Analysis of Variance model allowed the correlation to be included in the calculations of standard errors. Failing to account for the correlation led to obtaining incorrect standard errors, which ultimately led to making incorrect statistical inferences. In Section 3, a more complex example was shown where repeated measurements were taken on subjects at four different points in time. Here, subject was included in the model as a random effect which introduced the need for properly modelling the covariance structure. It is important to note that introducing random effects into the model changes the scope of inference. When using only fixed effects, regression models are used to make predictions of a response variable based on the values of explanatory variables. These predictions are simply expected values of $\mathbf{Y}$ conditional on the input values for $\mathbf{X}$. Any variance in the prediction is attributed to random error alone. Introducing random effects, however, allows the variance to be modelled by adding to the prediction a random component that can be thought of as a draw from a population that follows a multivariate normal distribution. Correctly modelling the covariance structure provided correct standard error estimates. Although it was not discussed at length previously, correctly specifying the covariance matrix also can have a major impact on the fixed parameter estimates themselves. For example, in a model with response vector $\boldsymbol{Y}$ and design matrix $\boldsymbol{X}$ of fixed effects, the vector of parameter estimates $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y} \tag{6.1}$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix. When the covariance structure is left unaccounted for, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{I}$.[6] This again shows how crucial modelling the covariance structure can be in statistical inference. Obtaining incorrect estimates for the parameters and standard errors leads to incorrect inference. Conclusions based on incorrect inference are misleading and can have serious consequences, especially in medical research. After stressing the importance of accounting for and modelling the correlation, a few techniques for detecting outliers in clustered data were discussed. In particular, it was shown how the Mahalanobis distance can be calculated to identify outlying clusters before even testing a statistical model. Finally, the concepts discussed in the paper were implemented in a case study analysis of a multi-center, randomized controlled trial for examining the effect of introducing a probiotic treatment to babies with colic. The objective of this paper was to show how proper statistical methods for accounting for correlation among observations within a cluster must be implemented in order to make correct statistical inference and draw conclusions.

References

[1] Ann R. Cannon, George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. *STAT2 Building Models for a World of Data*. W.H. Freeman and Company, New York, NY, 2013.

[2] Mayo Clinic. Diseases and conditions: Colic. `http://www.mayoclinic.org/diseases-conditions/colic/basics/definition/con-20019091`, Mar 2016.

[3] Dennis R. Cook. Detection of influential observations in linear regression. *Technometrics(American Statistical Association)*, 19(1):15–18, 1977.

[4] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. John Wiley and Sons Inc., Hoboken, NJ, 2004.

[5] F. Y. Hsieh, Philip W. Lavori, Harvey J. Cohen, and John R. Feussner. An overview of variance inflation factors for sample-size calculation. *Evaluation and the Health Professions*, 26(3):239–257, 2003.

[6] David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, and Keith E. Muller. *Applied Regression Analysis and Other Multivariable Methods*. PWS Publishing Co., Boston, MA, 1988.

[7] Ian H. Langford and Toby Lewis. Outliers in multilevel data. *Journal of the Royal Statistical Society*, 161(2):121–160, 1998.

[8] Ramon C. Littell, Jane Pendergast, and Ranjini Natarajan. Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19:1793–1819, 2000.

[9] Eberly College of Science Faculty. Multivariate normality and outliers: Lecture notes. Technical report, The Pennsylvania State University, 2016.

[10] Kay I. Penny. Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Journal of the Royal Statistical Society*, 45(1):73–81, 1977.

[11] R Foundation for Statistical Computing, Vienna, Austria. *R 3.2.5*, 2013.

[12] SAS Institute Inc., Cary, NC. *SAS 9.4*, 2013.

[13] SAS Institute Inc., Cary, NC. *JMP 12*, 2015.

[14] Oliver Schabenberger. Mixed model influence diagnostics. Technical report, SAS Institute Inc., 2004.

[15] Robert L. Wears. Advanced statistics: Statistical methods for analyzing cluster and cluster-randomized data. *Academic Emergency Medicine*, 9(4):330–341, 2002.

[16] Temesgen Zewotir and Jacky S. Galpin. Influence diagnostics for linear mixed models. *Journal of Data Science*, 3:153–177, 2005.

[17] Daowen Zhang. Statistical principles of clinical trials: Lecture notes. Technical report, North Carolina State University, 2009.