4-2017

Subgroup Analysis and Growth Curve Models for Longitudinal Data

Nichole Andrews
Western Michigan University, nicholeandrews@gmail.com

SUBGROUP ANALYSIS AND GROWTH CURVE MODELS FOR LONGITUDINAL DATA

by

Nichole Andrews

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Statistics
Western Michigan University
April 2017

Doctoral Committee:

Dr. Hyun Keun Cho, Ph.D., Chair
Dr. Joseph McKean, Ph.D.
Dr. Magdalena Niewiadomska-Bugaj, Ph.D.
Dr. Jun-Seok Oh, Ph.D.

SUBGROUP ANALYSIS AND GROWTH CURVE MODELS FOR LONGITUDINAL DATA

Nichole Andrews, Ph.D.

Western Michigan University, 2017

In clinical trials and biomedical studies, treatments are compared to determine which one is effective against illness. Growth curve analysis can be beneficial in longitudinal biomedical studies, as we can evaluate the treatment effect on the response over time. The generalized growth curve model using polynomial regression is proposed for longitudinal data. An optimal degree for the polynomial is obtained using the BIQIF, an adaptation of the Bayesian information criterion. Quadratic inference functions are used to estimate the parameters of the model, which takes into account the fact that repeated measurements from the same subject are more likely to be correlated. The equality of the growth curves is assessed using an asymptotically chi-square test statistic. Through this test, it could be shown that multiple treatments perform similarly, leading to the recommendation of either, however individuals can react to the same treatment very differently. A complete process for longitudinal data is also proposed that identifies subgroups of the population that would benefit from a specific treatment. A random effects linear model is used to evaluate individual treatment effects longitudinally where the random effects identify a positive or negative reaction to the treatment over. With the individual treatment effects and

characteristics of the patients, various classification algorithms are applied to build prediction models for subgrouping. While many subgrouping approaches have been developed recently, most of them do not check its validity. As such, a simple validation approach is proposed which not only determines if the subgroups used are appropriate and beneficial, but also compares methods to predict individual treatment effects. All proposed methods are confirmed with simulation studies and analysis of data from the Women Entering Care study on depression.

ACKNOWLEDGMENTS

While completing my undergrad degree, I would have never imagined that this is where I would end up. Fresh out of college and getting a teaching degree, things don't always work out the way we think they will. Starting grad school only a year after graduating was not the plan, but I can say that things work out for a reason and I'm happy with the way everything turned out. So many people have been with me throughout this journey that I'd like to thank, first being my family and friends. Their encouragement has gotten me through these last six years.

I'd also like to thank my committee members and all the professors I've had within the Statistics Department. All have taught me many things. While some classes were more challenging than others, they've taught me to persevere and always give it my all. While I never wish to study 20 hours for a final exam again, doing this has taught me a lot not only about the subject of study, but also about myself. I've learned my limits and my abilities. Thank you for your guidance. Within our department, I am thankful to Dr. Bugaj for the opportunities she's given me. The experience I've gained from working full time while finishing up my doctoral degree has been so worthwhile. While it hasn't always been easy to balance everything, I know that this has prepared me for other teaching opportunities down the road. Thank you for trusting me with these responsibilities. I'd also like to thank Michelle Hastings, the administrative assistant for our department. Not only has she greatly helped me with my job, but she's also been a big encouragement to me throughout this process and become a good friend.

A special thank you to my advisor, Dr. Cho, who has guided me along the way over the last two years. While my initial interest in the topic may have been minimal, it has grown significantly over my time of working with him. This has

been such an interesting and applicable topic for me to study. Thank you for not only advising me, but also keeping my interest in this relevant subject.

Lastly and most importantly, I'd like to thank my husband of nearly four years, Paul Andrews. Life has taken us in so many different directions. From graduating and moving many times to buying a home and starting a family, he has been by my side through it all and my biggest encourager. I can honestly say I don't know where I'd be without him. Paul, thank you for your endless love and support. Life has been an adventure and I wouldn't want to do it with anyone else. And thank you for giving me the greatest gift, our son, Mason. Mason, you bring your mommy so much joy. On my toughest days, your smiles and laughter get me through. Everything I do is for you. Mommy loves you so much.

Nichole Andrews

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Medical conditions can greatly impact one's way of life. While some conditions are not as severe, others need treatment to relieve symptoms and allow an individual to lead as normal of a life as possible. Biomedical studies then are used to test treatment effects. In such studies, where two or more treatments are compared, the goal is to identify which treatment will achieve a desired result and should be recommended for use to the population. It is common to use longitudinal data in biomedical studies. Longitudinal data consists of observations that are measured through time. This allows for analysis in the change of the response over time. While it is ideal that patients attend all follow up appointments for responses to be measured, it is common to have missing data points in longitudinal data.

Longitudinal data has been used to compare treatments and assess the response of a treatment on a patient over time. As such, two methods have been developed to analyze longitudinal data from medical studies and select a beneficial treatment for use to the population. We first introduce the idea of using growth curve models to choose an ideal treatment. If this method does not lead to the recommendation of a treatment that is deemed most beneficial, subgroup analysis, also known as personalized treatment, can be performed to determine a treatment beneficial for

subgroups of the population rather than the population as a whole.

We first outline the proposed procedures that will be presented in Chapters 2 and 3, followed by information on the Women Entering Care study, which will be analyzed throughout this paper.

1.1 Proposed Work

The ultimate goal is to determine the most beneficial treatment for a given condition that would be recommended for use to the population. We will first assess this with growth curves. Growth curves are used to investigate the trajectory for each treatment over time. This allows us to identify if a treatment is outperforming the others. Polynomial regression will be used to determine the coefficients of the growth curves and an asymptotically chi-square test statistic has been developed to assess the equality of the growth curves. This work is presented in Chapter 2.

With our proposed analysis of growth curves, it could be determined that two or more treatments outperform the others, however no significant difference can be found between the responses of these treatments. Which treatment should be recommended for use to the population then? Rather than selecting a single treatment, we will determine a personalized treatment for each patient that will produce optimal results. This work is presented in Chapter 3. A random effects model will be used to determine the treatment effect for each individual over time. Subgrouping is then performed to determine which treatment the patient should take. A validation approach will determine if the subgrouping was appropriate and beneficial.

We will assess both methods with simulation studies and analysis with data from the Women Entering Care study. The simulation studies confirm the effectiveness

of the proposed procedures while the data analysis allows us to see how this work can be applied to a real life situation.

1.2 Explanation of Dataset

In the Women Entering Care study, depression was studied among low-income minority women in the Washington D.C. area. 16,286 were screened major depressive disorder. Most were excluded for various reasons, including having an ethnicity that was not of interest for the study and having alcohol problems. In the end, 267 women were randomly assigned to treatments. The three treatments were antidepressant medication, cognitive behavioral therapy (CBT, psychotherapy), and referral to community mental health services (referral to community care, control).

Individuals in the medication group received the medication for 6 months. Paroxetine was initially given to the women, however if negative side effects or no improvement was seen by the ninth week, the women were then given buproprion. In total, 88 women were treated with medication. Of these, 67 received at least 9 weeks of guideline concordant medication therapy. Women assigned to the cognitive behavioral therapy group were treated by a psychologist with 8 weekly sessions, which could be either in a group or individually. While sessions were usually at a clinic, at-home sessions were an option, as some women were not able to travel to a clinic. In total, 90 women received cognitive behavioral therapy, of which 32 received at least six cognitive behavioral therapy sessions. Women in the referral to community care group were informed about depression and mental health treatments available to them in their communities. Of the 89 women in this group, only 15 attended at least one mental health visit.

The response for this dataset was the Hamilton Depression Score, which comes

as a result of an interview (Hamilton, 1960). A higher score indicates a more severe case of depression. Depression scores were obtained every month for the first six months, then every other month for the remainder of the year.

The duration of the treatments was no longer than six months, therefore data from only the first six months will be considered. Since the goal is to analyze the treatment effect over time, we only considered patients who had a baseline depression score and at least one follow-up. While all patients had an initial score, thirteen did not have any other depression scores and were therefore excluded from our analysis (medication $n = 86$, cognitive behavioral therapy $n = 83$, referral to community care $n = 85$).

Besides the treatment received and the depression scores, seven other variables were observed from each patient. These are:

- Age
- Marital Status
- School Status
- Housing Status
- Ethnicity
- Where You Were Born
- Whether Or Not You Work

Summary results of the categorical variables can be found in Table B.1 in Appendix B. More information on this dataset can be found in Miranda et al. (2003), where this dataset was first analyzed.

Tables 1.1 and 1.2 give us an initial look at our dataset with the means and confidence intervals for depression scores among all patients and broken down by

treatments, respectively. Once the assigned treatment has been used, we notice better (lower) scores among those in the medication and psychotherapy group. Without doing further analysis, this gives us an initial idea that one of these two treatments may be best for treating depression. In fact, in a previous study with this data, Miranda et al. (2003) found medication and psychotherapy to be significantly more effective at treating depression than receiving no treatment at all.

Table 1.1: Means and Confidence Intervals for Depression Scores

Time	Mean (95% CI)
Baseline	17.00 (16.30, 17.60)
Month 1	13.30 (12.40, 14.20)
Month 2	11.10 (10.20, 12.10)
Month 3	11.00 (10.00, 12.00)
Month 4	10.10 (9.06, 11.23)
Month 5	10.30 (9.24, 11.41)
Month 6	10.60 (9.69, 11.63)

Table 1.2: Means and Confidence Intervals for Depression Scores by Treatment

	Medication	CBT	Control
Time	Mean (95% CI)	Mean (95% CI)	Mean (95% CI)
Baseline	18.10 (17.00, 19.20)	16.30 (15.10, 17.50)	16.50 (15.40, 17.70)
Month 1	14.00 (12.50, 15.50)	13.10 (11.60, 14.60)	12.80 (11.00, 14.60)
Month 2	10.70 (9.15, 12.34)	11.40 (9.72, 13.13)	11.30 (9.43, 13.17)
Month 3	9.60 (8.02, 11.19)	10.20 (8.56, 11.92)	13.00 (11.20, 14.90)
Month 4	9.54 (7.68, 11.40)	9.07 (7.22, 10.92)	11.80 (9.86, 13.76)
Month 5	8.62 (6.83, 10.41)	10.5 (8.73, 12.22)	11.80 (9.72, 13.98)
Month 6	9.17 (7.41, 10.94)	10.7 (8.95, 12.52)	11.90 (10.10, 13.70)

The overall goal, then, is compare the depression scores over time for each treatment. While we wish to find a treatment that is most beneficial with growth curve models, this may not be the case. In such a case, subgroup analysis can be performed to determine which treatment is beneficial for an individual rather than the population as a whole.

Chapter 2

Growth Curve Models

2.1 Introduction

Growth curve analysis allows us to investigate the trajectory of the response variable for a given independent variable. Specifically, the growth curve models the change of the conditional mean of the response conditioned on an independent variable. Growth curve analysis can be advantageous in biomedical studies, where multiple treatment effects are compared, since it allows us to monitor the change in the response over time for each treatment.

In the Women Entering Care study, the goal was to investigate the treatment effect on depression longitudinally. As such, depression scores were collected monthly. Miranda et al. (2003) found that medication and cognitive behavioral therapy were better at treating depression than being referred to community care in the linear regression framework by incorporating the interaction term between treatment and time. On the other hand, Siddique et al. (2012) did not use all the depression scores, but rather at certain time points for those who took medication and cognitive behavioral therapy to investigate the response based on the severity of the depression. In this chapter, we propose the growth curve nonlinear mean regression

model that explores the change in the depression scores over a given time period for different treatment groups.

Growth curve analysis was first introduced by Potthoff and Roy in 1964. They provided a multivariate analysis of variance model for growth curves and conducted the hypothesis test concerning the coefficients in the model. Chou, Bentler, and Pentz (1998) introduced latent growth curve analysis that treats the initial status and the growth of the curves as latent variables to model the response variable. Curran, Obediat, and Losardo (2010) pointed out advantages to using growth curve models over traditional longitudinal models, which include, but are not limited to, the ability for the data to still be analyzed with missing data and nonlinear trajectories. These authors also note that a data requirement for analysis with growth curves is often that the response be continuous and normally distributed, while Barbosa and Goldstein (2000) proposed a multilevel model that is applicable for a discrete response variable.

In this chapter, we build the generalized growth curve model for longitudinal data under the polynomial regression framework. It is known that repeated measurements from the same subject are more likely to be correlated. Therefore, we employ the quadratic inference functions (Qu, Lindsay, and Li, 2000) to fit the growth curve model to longitudinal data. This yields more efficient estimators by accommodating the within-subject correlation without estimating the parameters associated with the correlation structure. When polynomial regression is modeled, the choice of a polynomial degree plays an important role in providing the most suitable growth curve model. We adopt the Bayesian information criterion (Schwarz, 1978) by treating the quadratic inference function as an objective function (Wang and Qu, 2009). This is able to choose the most parsimonious model consistently by imposing a penalty for overfitting the model. After the model is fitted to the data,

we further construct a hypothesis test to assess the equality of the growth curves. The proposed test statistic does not require the estimation of the covariance matrix of the regression parameter. In addition, it avoids the need to specify the likelihood function; specifying the likelihood function for correlated discrete response variables can be challenging. The entire proposed procedure is applicable for both continuous and discrete response variables. We apply our approach to the above-mentioned depression data and illustrate the change of treatment effects through the fitted growth curves of the depression scores.

2.2 Methodology

Suppose that $Y_i = (Y_{i1}, \ldots, Y_{im_i})'$ is a vector of m_i responses repeatedly measured at times $t_{i1}, \ldots, t_{im_i}$ for subject i where these measurements are more likely to be correlated. A generalized growth curve model is formulated as

$$E(Y_i|G_i, Z_i) = h(G_i' B Z_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where n is the number of subjects, $h(\cdot)$ is a known link function, $G_i = (G_{i1}, \ldots, G_{im_i})$ is a $(p+1) \times m_i$-dimensional matrix with $G_{ij} = (1, t_{ij}, t_{ij}^2, \ldots, t_{ij}^p)'$, Z_i is a q-dimensional vector representing the treatment received, and $B = (B_1, \ldots, B_q)$ is a $(p+1) \times q$-dimensional matrix of parameters. Note that Z_i models differences between q treatment groups and G_i specifies polynomial growth curves with an order of p; polynomial functions are dynamically consistent and a polynomial of sufficiently high order is guaranteed to provide an arbitrarily good fit to the observed longitudinal data, known as Taylor's theorem.

2.2.1 Estimation and Inference on the Regression Parameter

In this section, we first consider estimation of the parameter B in model (2.1). By letting $X_i = Z_i \otimes G_i$, where $\otimes$ is the kronecker product, and $\beta = (B_1', \ldots, B_q')'$, model (2.1) can be simplified as

$$E(Y_i|X_i) = h\left(X_i'\beta\right) = \mu_i. \tag{2.2}$$

Liang and Zeger (1986) extend a quasi-likelihood function (Wedderburn, 1974) and obtain an estimator of β by solving the generalized estimating equations

$$\sum_{i=1}^{n} \dot{\mu}_i' A_i^{-1/2} R_i(\alpha)^{-1} A_i^{-1/2} (Y_i - \mu_i) = 0, \tag{2.3}$$

where $\dot{\mu}_i = (\partial \mu_i / \partial \beta)$, A_i is the diagonal variance matrix of Y_i, and $R_i(\alpha)$ is a working correlation matrix of the responses with nuisance parameter α. Even though $R_i(\alpha)$ allows us to account for the within-subject correlation, the estimator of β can be inefficient if this working correlation structure is misspecified due to the need to estimate α. As such, Qu, Lindsay, and Li (2000) provide an alternative; it approximates an inverse of $R_i(\alpha)$ in (2.3) as

$$R_i(\alpha)^{-1} = \sum_{k=1}^{d} l_k D_{ik}, \tag{2.4}$$

where l_k is an unknown coefficient and D_{ik} is a known basis matrix for $k = 1, \ldots, d$. As an example, consider when $R_i(\alpha)$ is a compound symmetric matrix, where the diagonal elements are 1 and the remaining elements are α. Then $R_i(\alpha)^{-1}$ is a linear combination of two basis matrices, D_{i1} and D_{i2}, where D_{i1} is the identity matrix and D_{i2} contains 0's on the diagonal and all remaining elements are 1.

$$R_i(\alpha)^{-1} = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}^{-1}$$

$$R_i(\alpha)^{-1} = I_1 D_{i1} + I_2 D_{i2}$$

$$D_{i1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad D_{i2} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Likewise, if $R_i(\alpha)$ is a first-order autoregressive correlation matrix, where the element in the j^{th} row and m^{th} column is $\alpha^{|j-m|}$, $R_i(\alpha)^{-1}$ is a linear combination of three basis matrices: the identity matrix D_{i1}, D_{i2} with 1 on the two main off-diagonals and 0 elsewhere, and D_{i3} with a 1 on the upper left and lower right corners and 0 elsewhere.

$$R_i(\alpha)^{-1} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}^{-1}$$

$$R_i(\alpha)^{-1} = I_1 D_{i1} + I_2 D_{i2} + I_3 D_{i3}$$

$$D_{i1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} D_{i2} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} D_{i3} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

By replacing $R_i(\alpha)^{-1}$ in (2.3) with the basis matrices in (2.4), the estimator $\hat{\beta}$ is obtained by minimizing the quadratic inference functions

$$Q(\beta) = n\bar{g}(\beta)'C^{-1}\bar{g}(\beta), \tag{2.5}$$

where $\bar{g}(\beta) = n^{-1}\sum_{i=1}^{n} g_i(\beta)$ and $C = n^{-1}\sum_{i=1}^{n} g_i(\beta)g_i(\beta)'$ with

$$g_i(\beta) = \begin{pmatrix} \dot{\mu}_i' A_i^{-1/2} D_{i1} A_i^{-1/2}(Y_i - \mu_i) \\ \vdots \\ \dot{\mu}_i' A_i^{-1/2} D_{id} A_i^{-1/2}(Y_i - \mu_i) \end{pmatrix}. \tag{2.6}$$

This approach allows us to incorporate the correlation information without estimating the nuisance parameter α in $R_i(\alpha)$. Moreover, this approach optimally combines the estimating equations in (2.6) by assigning a lesser weight to the equation having a larger variation. This leads to the most efficient estimator among estimators solved by the same linear class of the equations in (2.6).

Theorem 1. *Under the regularity conditions in Appendix A, the estimator of β satisfies*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

where β_0 is the true parameter of β and $V = (\Phi'\Sigma^{-1}\Phi)^{-1}$ with $\Phi = E\{\partial g_i(\beta)/\partial\beta\}$ and $\Sigma = E\{g_i(\beta)g_i(\beta)'\}$. Moreover, $V_I - V$ is positive semidefinite, where V_I is

the asymptotic covariance matrix of $\hat{\beta}$ under the independent working correlation structure.

The proof of this theorem is presented in Appendix A.

Theorem 1 confirms that the estimator is asymptotically normal with true mean β and covariance matrix V. In addition, a positive semidefinite $V_I - V$ ensures that the asymptotic variance in V is no greater than the one obtained under the independent working structure. This accounts for the efficiency gain from incorporating the within-subject correlation commonly existing in longitudinal data.

For statistical inference on the regression parameter such as construction of a confidence interval for β, we can obtain a plug-in estimator of the asymptotic covariance matrix V in Theorem 1 as

$$\hat{V} = \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_i(\hat{\beta})}{\partial \beta} \right\}' \left\{ \frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\beta}) g_i(\hat{\beta})' \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_i(\hat{\beta})}{\partial \beta} \right\} \right]^{-1}.$$

2.2.2 Hypothesis Test for Identical Growth Curves

A main concern here that has not been addressed yet is which treatment produces better results over time. Which treatment should we recommend for use to the population? Do two or more treatments perform similarly? To investigate these inquiries, we have to determine whether or not all of the growth curves are the same. To assess the equality of the growth curves, we state the null hypothesis as

$$H_0 : B_1 = B_2 = \ldots = B_q.$$

Since the quadratic inference functions play an inferential role similar to minus twice the loglikelihood function, we construct the test statistic based on $Q(\beta)$ in (2.5)

for testing the null hypothesis as

$$T = Q(\check{\beta}) - Q(\hat{\beta}),$$

where $\check{\beta}$ is the minimizer of $Q(\beta)$ under the null hypothesis.

Theorem 2. *If the regularity conditions in Appendix A hold,* $T \xrightarrow{d} \chi^2_{(p+1)(q-1)}$ *under* H_0.

The proof of this theorem is presented in Appendix A.

Theorem 2 confirms that under the null hypothesis, this test statistic is asymptotically chi-square distributed with $(p+1)(q-1)$ degrees of freedom, thus we can use critical values to draw a conclusion on the equality of the growth curves. We remark that estimation of the asymptotic covariance matrix of the regression parameter is not required. Moreover, the proposed test is conducted without specifying the loglikelihood function, which can be especially challenging with discrete correlated responses.

If it is shown that not all growth curves are equal, one may wish to further investigate these curves and test the equality of some of them. We can readily extend the above test by setting the parameters of the growth curves of particular interest as equal for the null hypothesis.

2.2.3 Bayesian Information Criterion for Choice of a Polynomial Order

Up until this point, we have assumed that the polynomial with order p fits the data sufficiently well, however the process for selecting this order has yet to be discussed. One challenge that arises is that while order p is sufficient, so are orders higher than p. Therefore we seek to find the true model that neither overfits nor

underfits the data. As such, we provide a model selection procedure that selects the polynomial order for the parsimonious correct model. Although cross validation and generalized cross validation are commonly used to choose the polynomial order, these approaches tend to overfit the model (Wang, Li, and Tsai, 2007). An alternative procedure is the Bayesian information criterion (BIC), which enables us to identify the true model consistently.

We select a value of k that is large enough to capture p ($p \leq k$) and consider all models with polynomial order up to k. We index these candidate models by m, where $m = 0, \ldots, k$. For each candidate model with polynomial order m, we estimate the $(m+1) \times q$ matrix B in model (2.1). This is adapted to formulate the $(k+1) \times q$ matrix $\hat{B}(m) = (\hat{B}_1(m), \ldots, \hat{B}_q(m))$, where the first $(m+1)$ rows are $\hat{B}$ obtained from estimation with order m and all remaining elements are 0. As an example, consider when $k = 3$ and $q = 3$. Then

$$\hat{B}(0) = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 & \hat{\beta}_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \hat{B}(1) = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_3 & \hat{\beta}_5 \\ \hat{\beta}_2 & \hat{\beta}_4 & \hat{\beta}_6 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\hat{B}(2) = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_4 & \hat{\beta}_7 \\ \hat{\beta}_2 & \hat{\beta}_5 & \hat{\beta}_8 \\ \hat{\beta}_3 & \hat{\beta}_6 & \hat{\beta}_9 \\ 0 & 0 & 0 \end{bmatrix} \hat{B}(3) = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_5 & \hat{\beta}_9 \\ \hat{\beta}_2 & \hat{\beta}_6 & \hat{\beta}_{10} \\ \hat{\beta}_3 & \hat{\beta}_7 & \hat{\beta}_{11} \\ \hat{\beta}_4 & \hat{\beta}_8 & \hat{\beta}_{12} \end{bmatrix}$$

Let $Q_k(\beta)$ be the quadratic inference functions based on model (2.1) with polynomial order k. We adopt the BIC based on $Q_k(\beta)$ (Wang and Qu, 2009):

$$\text{BIQIF}_m = Q_k(\hat{\beta}(m)) + df_m \log(n), m = 1, \ldots, k, \tag{2.7}$$

where $\hat{\beta}(m) = (\hat{B}_1(m)', \ldots, \hat{B}_q(m)')'$ and df_m is the number of non-zero coefficients in $\hat{\beta}(m)$. The degree of the polynomial is determined by minimizing (2.7), denoted by $\hat{m} = \arg\min_m \text{BIQIF}_m$.

Theorem 3. *Under the regularity conditions in the appendix, as* $n \to \infty$

$$P(\hat{m} = p) \to 1.$$

See Appendix A for the proof of this theorem.

Theorem 3 ensures that the proposed criterion identifies the true order of the growth curve model consistently.

Note that if the selected polynomial order is 0, then all growth curves in model (2.1) are constant over time. In such a case, a growth curve might not be needed, however the benefit to our procedure is that it can identify this pattern.

2.3 Data Analysis

The proposed procedure was then applied to data from the Women Entering Care study outlined in Section 1.2. To assess the treatment effect over time, model (2.1) was applied, resulting in:

$$Y = \begin{pmatrix} 1 & 0 & \ldots & 0^p \\ 1 & 1 & \ldots & 1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 6 & \ldots & 6^p \end{pmatrix} \begin{pmatrix} \beta_1 & \beta_{p+2} & \beta_{2(p+1)+1} \\ \beta_2 & \beta_{p+3} & \beta_{2(p+1)+2} \\ \vdots & \vdots & \vdots \\ \beta_{p+1} & \beta_{2(p+1)} & \beta_{3(p+1)} \end{pmatrix} \begin{pmatrix} 1_{88} & 0 & 0 \\ 0 & 1_{90} & 0 \\ 0 & 0 & 1_{89} \end{pmatrix}' + \epsilon = G'BZ + \epsilon, \tag{2.8}$$

Table 2.1: BIQIF Values for Depression Dataset

m	$BIQIF_m$
0	58.528
1	50.350
2	14.514
3	21.743
4	29.254

where 1_j is a j-dimensional one vector. Note that the ith column vector of Z is $(1,0,0)'$ if the ith patient received medication, $(0,1,0)'$ if the patient received CBT, and $(0,0,1)'$ otherwise.

The true value of p was selected using BIQIF. Table 2.1 displays the values of the BIQIF for $k = 4$. This confirms that the quadratic form ($\hat{m} = 2$) fits the data sufficiently well.

We then obtained an estimator of B in (2.8) and its standard errors using the quadratic inference functions and plug-in approach in Section 2.2.1, respectively, and computed the Wald test statistics, as shown in Table 2.2. These estimates all resulted in high test statistics with p-values that were very close to zero. We also noticed that parameter estimates for medication and cognitive behavioral therapy are relatively close while those for control differ a bit more.

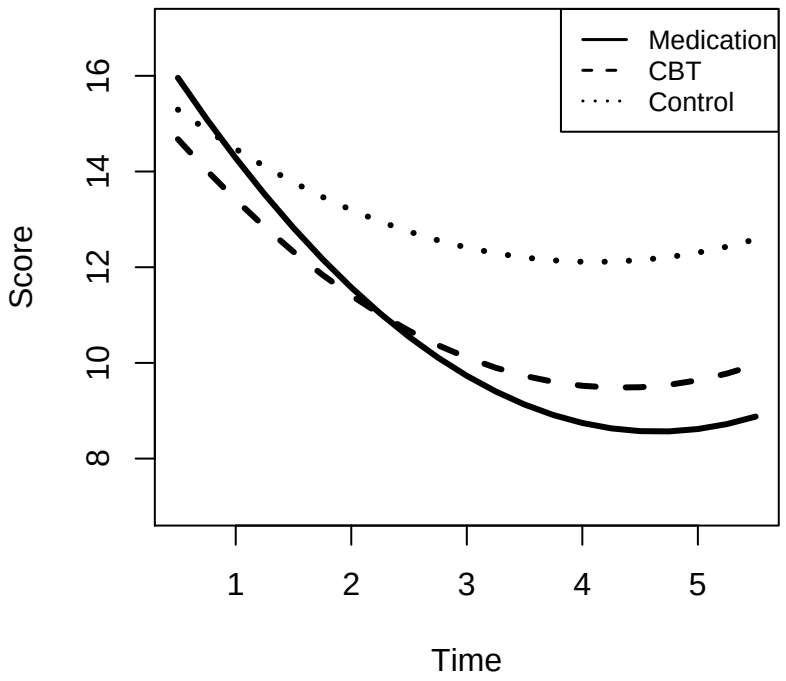
We also displayed the fitted curves in Figure 2.1. This figure indicates that all forms of treatment resulted in the depression score decreasing over time with medication and cognitive behavioral therapy producing lower depression scores.

In this manner, we tested whether all growth curves were the same using the proposed T test statistic from Section 2.2.2. The resulting test statistic was $T = 21.80$. Here, the degrees of freedom are calculated as $(p+1)(q-1) = (2+1)(3-1) = 6$. The p-value for $T = 21.80$ and 6 degrees of freedom is 0.0013. This suggests that differences between some of growth curves are statistically significant

Table 2.2: Estimated Coefficients, Standard Errors (SE), and Wald Test Statistics

	Estimate	SE	Wald
	Medication		
Intercept	17.849	0.504	1255.0
Linear	-3.997	0.397	101.2
Quadratic	0.430	0.062	47.5
	CBT		
Intercept	16.117	0.513	987.6
Linear	-3.060	0.366	69.7
Quadratic	0.354	0.061	33.1
	Control		
Intercept	16.235	0.533	927.4
Linear	-2.008	0.426	22.2
Quadratic	0.244	0.068	13.1

Figure 2.1: Fitted Growth Curves of Three Treatments for Depression Data



at a nominal level of 0.05.

From Figure 2.1, it is apparent that the growth curve for the control group is different from the other two. These curves indicate that medication and cognitive behavioral therapy perform better at treating depression than being referred to community care. The growth curves for medication and cognitive behavioral therapy, however, look similar, therefore they were compared to see if the two growth curves differ. Here, $T = 7.54$ and the degrees of freedom are calculated as $(2+1)(2-1) = 3$, resulting in a p-value of 0.0565. This indicates that the proposed test fails to reject the equality between the two growth curves at the significance level of 0.05, although it should be noted that the p-value is close to 0.05.

2.4 Simulation Studies

The simulation studies here reflect the results from the data analysis on the Women Entering Care study in Section 2.3. We generated correlated continuous responses based on model (2.8) with $p = 2$ by treating the estimated coefficients in Table 2.2 as B. In addition, the random errors matrix ϵ were generated from a multivariate normal distribution, $\epsilon \sim N(0, I_{267} \otimes R)$, where I_{267} is a 267×267-dimensional identity matrix and R is a 7×7-dimensional matrix of the AR(1) with a correlation coefficient of either 0.4 or 0.8, i.e.,

$$R = \begin{bmatrix} 1 & 0.4 & 0.4^2 & 0.4^3 & 0.4^4 & 0.4^5 & 0.4^6 \\ 0.4 & 1 & 0.4 & 0.4^2 & 0.4^3 & 0.4^4 & 0.4^5 \\ 0.4^2 & 0.4 & 1 & 0.4 & 0.4^2 & 0.4^3 & 0.4^4 \\ 0.4^3 & 0.4^2 & 0.4 & 1 & 0.4 & 0.4^2 & 0.4^3 \\ 0.4^4 & 0.4^3 & 0.4^2 & 0.4 & 1 & 0.4 & 0.4^2 \\ 0.4^5 & 0.4^4 & 0.4^3 & 0.4^2 & 0.4 & 1 & 0.4 \\ 0.4^6 & 0.4^5 & 0.4^4 & 0.4^3 & 0.4^2 & 0.4 & 1 \end{bmatrix}$$

and

$$R = \begin{bmatrix} 1 & 0.8 & 0.8^2 & 0.8^3 & 0.8^4 & 0.8^5 & 0.8^6 \\ 0.8 & 1 & 0.8 & 0.8^2 & 0.8^3 & 0.8^4 & 0.8^5 \\ 0.8^2 & 0.8 & 1 & 0.8 & 0.8^2 & 0.8^3 & 0.8^4 \\ 0.8^3 & 0.8^2 & 0.8 & 1 & 0.8 & 0.8^2 & 0.8^3 \\ 0.8^4 & 0.8^3 & 0.8^2 & 0.8 & 1 & 0.8 & 0.8^2 \\ 0.8^5 & 0.8^4 & 0.8^3 & 0.8^2 & 0.8 & 1 & 0.8 \\ 0.8^6 & 0.8^5 & 0.8^4 & 0.8^3 & 0.8^2 & 0.8 & 1 \end{bmatrix}$$

We generated 1000 simulated data sets and estimated the coefficients of B for all simulations under three working correlation structures: AR(1), compound symmetry, and independence.

We also computed 95% confidence intervals and determined whether or not the interval captured the true parameter. Moreover, the average length of the confidence intervals for each parameter estimate is reported in Table 2.3. This table also reports the proportion of times the true parameter was captured in the interval. We notice that the coverage probabilities were nearly the same, and at or close to 95%, for all coefficients under each working correlation structure. This could lead us to conclude that the working correlation structure used is irrelevant, however it is important to note the difference among the MSE's and average lengths of these

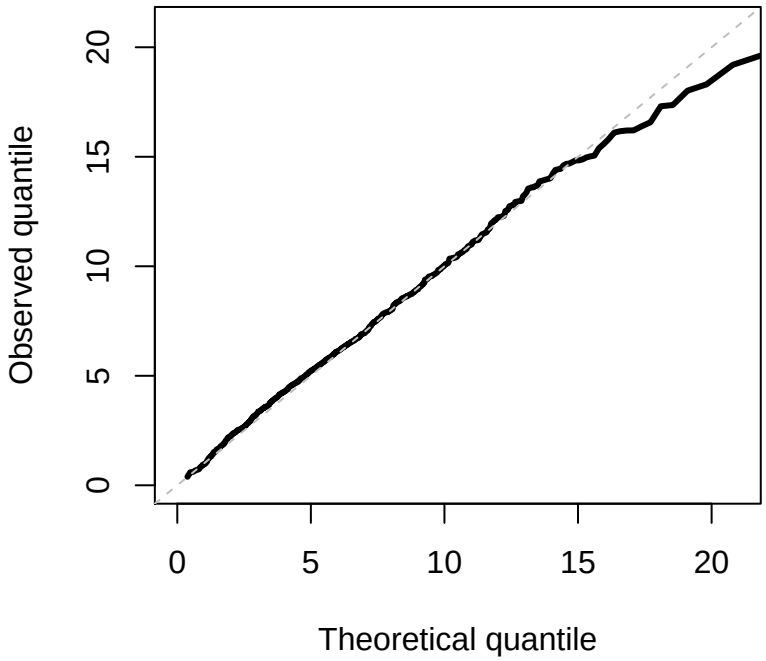
Table 2.3: Mean Squared Errors (MSE $\times$ 100), Coverage Probabilities (CP), and Means of Confidence Interval Lengths (Length) Under the AR(1), Compound Symmetry and Independent Working Correlation Structure

	AR(1)			Compound symmetry			Independent		
β	MSE	Length	CP(%)	MSE	Length	CP(%)	MSE	Length	CP(%)
$\rho = 0.8$									
β_1	1.429	0.438	93%	1.492	0.450	92%	1.550	0.480	94%
β_2	0.421	0.243	94%	0.445	0.250	94%	0.514	0.279	94%
β_3	0.010	0.038	94%	0.011	0.039	93%	0.013	0.044	95%
β_4	1.313	0.435	95%	1.374	0.445	94%	1.482	0.476	95%
β_5	0.400	0.241	95%	0.437	0.248	93%	0.513	0.277	94%
β_6	0.010	0.037	94%	0.011	0.038	94%	0.013	0.043	94%
β_7	1.343	0.435	93%	1.452	0.446	93%	1.458	0.476	95%
β_8	0.406	0.243	94%	0.428	0.251	95%	0.512	0.286	96%
β_9	0.010	0.038	94%	0.011	0.039	94%	0.013	0.045	95%
$\rho = 0.4$									
β_1	1.404	0.446	94%	1.415	0.453	94%	1.407	0.464	95%
β_2	0.750	0.329	95%	0.745	0.334	95%	0.744	0.344	95%
β_3	0.019	0.053	95%	0.019	0.053	94%	0.019	0.055	96%
β_4	1.446	0.449	94%	1.494	0.454	93%	1.453	0.466	94%
β_5	0.820	0.331	92%	0.810	0.335	93%	0.805	0.345	95%
β_6	0.020	0.052	92%	0.020	0.053	93%	0.020	0.054	94%
β_7	1.582	0.472	94%	1.666	0.478	94%	1.590	0.491	95%
β_8	0.822	0.339	93%	0.865	0.343	93%	0.835	0.353	94%
β_9	0.020	0.053	94%	0.021	0.054	93%	0.021	0.055	94%

confidence intervals. The confidence intervals for the correct working correlation structure, AR(1), are narrower than those of the other structures and are widest for independence, where we ignored correlation between observations.

For each simulation study, we also determined whether or not BIQIF identified the best polynomial order to be $p = 2$. In fact, BIQIF identified $\hat{m} = 2$ every time, regardless of the correlation coefficient. In addition, we also tested whether or not all growth curves were the same with our test statistic T. For $\rho = 0.8$, our test always identified that the growth curves were not all the same. For $\rho = 0.4$, the test had this same result 999 times of the 1000 simulations when the correct

Figure 2.2: Quantile-Quantile Plots for the Chi-Square Distribution with Six Degrees of Freedom Versus the Test Statistic when the Null Hypothesis is True



working correlation structure, AR(1), was used.

We also ran additional simulation studies with another 1000 datasets to further investigate Theorem 2. Here, we considered all growth curves to be the same as that of the CBT curve and tested how many times among the 1000 simulations we rejected the null hypothesis of equal growth curves at a significant level of 0.05. Here, we rejected the null hypothesis 53 times, which is close to the nominal level. We further drew the qq-plot in Figure 2.2 for the chi-square distribution versus the test statistic when the null hypothesis is true. All of this confirms Theorem 2.

2.5 Discussion

We have developed an entire process using a generalized growth curve model that fits the data to a polynomial and tests the equality of the curves. This is especially useful when comparing treatments to see which one results in a better response over time. In general, the proposed procedure is easy to implement and interpret. This procedure also overcomes the limitation of a linear relationship between the response and time. For our data analysis with the depression dataset, we found that there was a quadratic relationship between the response and time for each treatment. In addition, our method was able to show that medication and cognitive behavioral therapy outperformed referral to community care in treating depression.

Here, we assume that a polynomial curve fits the data sufficiently well. Since our growth curves are not necessarily linear, nonparametric regression is another approach for the growth curves. Our method could be extended to use of the spline curve. The spline curve is a global estimation, meaning it uses all the information to create the curve, as our approach did also. To extend our work, you would simply replace the polynomial regression matrix with the spline. This will create a curve between points called knots, which could also be identified with the BIC. This is beyond the scope of the current research, though.

Chapter 3

Subgroup Analysis

3.1 Introduction

For longitudinal data, we suppose that Y_{ij} is the response for the i^{th} subject at time T_{ij} where $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, and n_i is the number of times measurements are taken on the i^{th} patient. In addition, the n subjects are independent. To evaluate the treatment effect over time, the following marginal regression model could be considered:

$$Y_{ij} = \delta_0 + \delta_1 Z_i T_{ij} + \delta_2 T_{ij} + \xi_{ij}, \tag{3.1}$$

where $Z_i = 1$ or -1 represents the treatment assignment for patient i, $\delta = (\delta_0, \delta_1, \delta_2)'$ is the parameter vector, and ξ_{ij} are random errors. As outlined in Section 2.2.1, generalized estimating equations (Liang and Zeger, 1986) and quadratic inference functions (Qu, Lindsay, and Li, 2000) can be used to estimate δ. Both approaches can yield consistent and efficient estimators by accommodating the within-subject correlation commonly existing in longitudinal data.

Analysis with model (3.1), however, could indicate no difference in the outcome of two treatments, resulting in the recommendation of either treatment for use in

the population. Such was the case in Section 2.3, where no difference was found in the depression scores between the medication and cognitive behavioral therapy groups over time. At the same time, individuals can react very differently to the same treatment. Outside factors, such as biological or environmental influences, can have a significant impact on the outcome of a given treatment. As such, a method for identifying an ideal treatment based on patient characteristics is desired rather than identifying a single beneficial treatment for the entire population.

Song and Pepe (2004) proposed a method for subgrouping patients into a particular treatment according to a covariate determined by how this value compared to a pre-specified threshold. The use of a single covariate was also used by Bonetti and Gelber (2004), in which patients were grouped by the value of this covariate and analyzed with a moving average procedure. Moskowitz and Pepe (2004) used the concept of positive predictive values with a single covariate. The problem with these methods, though, is that more than one variable may be related to the outcome of the treatment. Cai et al. (2011) were able to utilize multiple baseline measurements with a two-stage method, where a parametric index score was calculated based on the estimated subject-specific mean response for the treatments with either a parametric or semiparametric model, followed by inference of the average treatment difference. Zhao et al. (2011) also used a parametric scoring system with multiple baseline covariates. Foster, Taylor, and Ruberg (2011) proposed the virtual twins method to identify a subgroup for which the treatment effect was better than the average treatment effect.

Recently, random effects linear models have been studied for personalized treatments, as the model allows each patient to be considered an individual rather than only a member of the population (Diaz, Yeh, and Leon, 2012, Diaz and de Leon, 2013). Diaz et al. (2007) used a random intercept model to model the log of

plasma concentrations given certain covariates. Diaz (2016) proposed benefit functions for treatment comparison and provided a graphical method for investigating the severity of a disease. Here, the random effects incorporate variability of the response differences in personal characteristics of the patient. Cho et al. (2016) used a random forest approach in an unspecified random effects model. Zhu and Qu (2016) personalized drug dosage over time with a log-linear mixed effect model. Diaz, Yeh, and Leon (2012) also noted that an empirical Bayesian approach under the mixed model framework may have better results for individualizing drug doses.

While the above mentioned procedures can subgroup the data, the effectiveness of their classification has not been fully discussed. Shen and He (2015) developed a procedure using a structured logistic-normal mixture model that not only classified the data, but also tested for the existence of subgroups. This work was extended by Wu, Zheng, and Yu (2016) for time-to-event data with the semiparametric logistic-cox mixture model. While these methods have advanced work in subgroup analysis, specifications for the data may not always be met.

In this chapter, we offer a complete process from subgrouping to validation for personalized treatments in longitudinal studies. Our procedure starts by providing a random effects linear model. The random effects in the model evaluate individual treatment effects over time, yet the fixed effects still allow us to look at the population as a whole. Since the variation in the random effects acts as the variation between characteristics of the patients (Diaz, Yeh, and Leon, 2012, Diaz, 2016, Senn, 2001), we employ various classification approaches to build prediction models based on the individual effects and characteristics of the patients; both linear and nonlinear classification approaches are considered, since the association between the characteristics of the patient and the outcome is unknown in practice. While subgrouping can be performed based on the prediction models, the question of its

appropriateness and which model is best remains unanswered. Therefore, a validation procedure has been developed to choose the best prediction model under the marginal regression framework.

While many methods have been developed for classifying data, the advantage to the proposed method is that it utilizes supervised learning algorithms already developed, making them easier to implement and interpret. In addition, the proposed procedure can be readily applied to a longitudinal medical study where all follow up appointments may not be attended, therefore resulting in missing measurements. Moreover, the validation approach allows us to not only analyze the treatment effect over time for those that received the treatment deemed beneficial with the prediction model, but also takes into account a time effect. This is an important aspect; while we may desire that the value of the outcome decreases over time, this may not happen. Including a time effect allows us to analyze whether or not the treatment slows the progression of the illness. Since we are able to assess the validity of our classification and determine the best prediction model, our steps outline the entire procedure for determining an appropriate subgroup.

3.2 Methodology

3.2.1 Evaluating an Individual Treatment Effect

Since a mixed model has been shown to be effective in the analysis of longitudinal data, we consider the model that evaluates the treatment effect, specifically its effect over time, on a response. Accordingly, the random slope intercept model is formulated as:

$$Y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i})Z_i T_{ij} + \beta_2 T_{ij} + e_{ij}, \quad i = 1, \ldots, n \quad j = 1, \ldots, n_i, \tag{3.2}$$

where α_{0i} and α_{1i} are the random intercept and slope for subject i, respectively. While β_1 represents the overall average of the treatment effect over time, α_{1i} enables us to take into account individual differences. By considering the interaction effect between the treatment and time, model (3.2) allows us to evaluate the treatment effect on the response over time.

We estimate the parameters in model (3.2) using maximum likelihood estimation. Without loss of generality, we suppose that the number of measurements taken on each subject are the same (i.e., $n_i = k$ for all i) and rewrite model (3.2) as:

$$Y = G\beta + D\alpha + e,$$

where Y is an nk-dimensional vector of the response and G and D are $nk \times 3$ and $nk \times 2n$ matrices of covariates corresponding to the fixed effect $\beta = (\beta_0, \beta_1, \beta_2)'$ and random effect $\alpha = (\alpha_{01}, \ldots, \alpha_{0n}, \alpha_{11}, \ldots, \alpha_{1n})'$, respectively. Assuming a multivariate normal distribution

$$\begin{pmatrix} \alpha \\ e \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega & 0 \\ 0 & \Sigma \end{pmatrix} \right),$$

Y can also be expressed as $Y = G\beta + e^*$, where $e^* = D\alpha + e$, resulting in $e^* \sim N(0, M)$ where $M = D\Omega D' + \Sigma$. If M were the identity matrix, $\hat{\beta} = (G'G)^{-1}G'Y$, however this is doubtful. Therefore, we will change our equation slightly to help us.

$$\begin{aligned} Y &= G\beta + \epsilon^* \\ M^{-\frac{1}{2}}Y &= M^{-\frac{1}{2}}G\beta + \underbrace{M^{-\frac{1}{2}}\epsilon^*}_{\sim N(0,I)} \end{aligned}$$

We are now able to estimate β.

$$\begin{aligned}\hat{\beta} &= (G'M^{-\frac{1}{2}}M^{-\frac{1}{2}}G)^{-1}G'M^{-\frac{1}{2}}M^{-\frac{1}{2}}Y \\ &= (G'M^{-1}G)^{-1}G'M^{-1}Y\end{aligned}$$

To estimate α, we will use the fact that $Y \sim N(X\beta, M)$ and $\alpha \sim N(0, \Omega)$. We will first find the joint distribution of Y and α. Since we know the expected value and variance of each, we need the covariance between the two vectors.

$$\begin{aligned}\text{COV}(Y, \alpha) &= \text{COV}(G\beta + D\alpha + \epsilon, \alpha) \\ &= \text{COV}(G\beta, \alpha) + \text{COV}(D\alpha, \alpha) + \text{COV}(\epsilon, \alpha) \\ &= 0 + D\Omega + 0 \\ &= D\Omega\end{aligned}$$

Therefore, $\begin{pmatrix} Y \\ \alpha \end{pmatrix} \sim N\left(\begin{pmatrix} G\beta \\ 0 \end{pmatrix}\begin{pmatrix} M & D\Omega \\ \Omega D' & \Omega \end{pmatrix}\right)$.

Recall that if $\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sum_A & \sum_{AB} \\ \sum_{BA} & \sum_B \end{pmatrix}\right)$, then

$$B|A \sim N(\mu_B + \textstyle\sum_{BA}\sum_A^{-1}(A - \mu_A), \sum_B - \sum_{BA}\sum_A^{-1}\sum_{AB})$$

Knowing this,

$$\begin{aligned} E(\alpha|Y) &= 0 + \Omega D' M^{-1}(Y - G\beta) \\ &= \Omega D' M^{-1}(Y - G\beta) \end{aligned}$$

Therefore, we have the following parameter estimates for our fixed and random effects:

$$\begin{aligned} \hat{\beta} &= (G'\hat{M}^{-1}G)^{-1}G'\hat{M}^{-1}Y \\ \hat{\alpha} &= \hat{\Omega} D' \hat{M}^{-1}(Y - G\hat{\beta}) \end{aligned}$$

Here $\hat{\Omega}$ and $\hat{\Sigma}$, and ultimately $\hat{M}$, are obtained by maximizing the following likelihood function:

$$\begin{aligned} I(\Omega, \Sigma) &= -\frac{1}{2}(Y - G(G'M^{-1}G)^{-1}G'M^{-1}Y)'M^{-1}(Y - G(G'M^{-1}G)^{-1}G'M^{-1}Y) \\ &\quad -\frac{1}{2}log|M| - \frac{nk}{2}log(2\pi), \end{aligned}$$

where $|M|$ is the determinant of the covariance matrix M. While this is computationally intensive, advances with technology and software make this a non-issue. Moreover, Hartley and Rao (1967) showed that these estimates are asymptotically consistent and efficient. When the estimate of M is biased, the restricted maximum likelihood is a viable alternative approach (Latra et al., 2010).

Since model (3.2) provides the individual treatment effect on the response over time, we can split all subjects into two groups according to whether or not they had a positive effect. For this, a value, C_i, is assigned to each subject based on

the sum of the fixed slope estimate and random slope estimate for the interaction between treatment and time ($\hat{\beta}_1 + \hat{\alpha}_{1i}$). The assignment is

$$C_i = \begin{cases} 1 & \hat{\beta}_1 + \hat{\alpha}_{1i} > 0 \\ -1 & \hat{\beta}_1 + \hat{\alpha}_{1i} \leq 0 \end{cases}$$

C_i indicates a positive or negative treatment effect for subject i over time.

3.2.2 Building the Prediction Model

Once individual effects have been identified, we classify our data accordingly. The response variable for our subgrouping approaches is the binary outcome of C_i. As such, we build a prediction model based on the independent variables X_i, which contains characteristics of patient i that are deemed influential to the assignment of the treatment. This could include variables such as, but not limited to, age, gender, and race. The use of C_i as the response is key, as it is determined by the parameter estimate for the interaction between treatment, Z_i, and time, T_{ij}, for each patient and is not an observed value form the dataset. Since we will be classifying observations into one of two groups, the desired prediction model is then specified as

$$f(X_i) = P(C_i = 1|X_i),$$

where $f(\cdot)$ is a function representing the relationship between C_i and X_i. In reality, this relationship is unknown. It could be either linear or nonlinear, however this lack of information makes the function $f(\cdot)$ unidentifiable. As such, various prediction models are constructed through both types of supervised learning algorithms:

- Linear
  - Logistic regression

  - Linear discriminant analysis (LDA)
  - Support vector machine (SVM) with linear kernel

- Nonlinear
  - Quadratic discriminant analysis (QDA)
  - Decision tree
  - Random forest
  - SVM with radial kernel

We denote the estimated model by $\hat{f}(X_i)$ and classify patient i as $\hat{C}_i = 1$ if $\hat{f}(X_i) > 0.5$ and $\hat{C}_i = -1$ otherwise.

Using logistic regression, we will calculate the probability $\mathrm{P}(C_i = 1|X_i)$. To do this, we will fit a model for $\mathrm{P}(C_i = 1|X_i)$ with the variables of X_i using the logistic function, that is

$$P(C_i = 1|X_i) = \frac{e^{\beta' X_i}}{1 + e^{\beta' X_i}}.$$

The regression coefficients of β are estimated using maximum likelihood estimation. If this probability is at least 0.5, we will classify patient i as $\hat{C}_i = 1$, otherwise we will let $\hat{C}_i = -1$. That is,

$$\hat{C} = \begin{cases} 1 & \text{if } P(C_i = 1|X_i) \geq 0.5 \\ -1 & \text{if } P(C_i = 1|X_i) < 0.5 \end{cases}$$

LDA is similar to logistic regression, however instead of only predicting $\mathrm{P}(C_i = 1|X_i)$, it will predict both $\mathrm{P}(C_i = 1|X_i)$ and $\mathrm{P}(C_i = -1|X_i)$. Our classification will then be based on which probability is higher. These probabilities are calculated using the prior probabilities, π_0 and π_1, that a randomly chosen observation is from a certain class and the density function of X_i. If the prior probabilities are

unknown, we will default to using the probability that an observation in the training dataset falls in the given class. It is assumed that X_i is from a multivariate normal distribution. Therefore, it is assumed that $X_i \sim N(\mu, \Sigma)$, then the density function of X_i is

$$f(X_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)}$$

where p is the number of variables in X_i. Classification is based on whether $\mathrm{P}(C_i = 1|X_i)$ or $\mathrm{P}(C_i = -1|X_i)$ is larger, where

$$P(C_i = k|X_i) = \frac{\pi_k f_k(X_i)}{\pi_0 f_0(X_i) + \pi_1 f_1(X_i)}.$$

Like LDA, QDA it will calculate a probability for each class and classify the observation according to the largest probability. The difference, however, is that QDA assumes that each class has its only covariance matrix, meaning $X_i \sim N(\mu_k, \Sigma_k)$ for the k^{th} class.

We will build a decision tree that splits our data into regions based on values of the personal characteristics. The classification groups are determined by the most occurring C_i value in that region from the training dataset.

The random forest algorithm works similar to that of a decision tree, however when building our tree, this takes bootstrap samples of the data and then averages the results. In addition, at each split, instead of considering all p variables, we will only consider $\sqrt{p}$ of them. This prevents strong variables from always being in the top split and makes the results more reliable (James et al., 2013). An added bonus of this algorithm is that it can identify important variables related to the classification.

We will also be using SVM, which extends beyond linear variables to quadratic and even cubic to find the relationship between C_i and X_i. SVM uses a function of

the inner products that we call a kernel. The default kernel in R is a radial kernel, but we will also consider a linear kernel. A linear kernel uses the inner product of each observation ($\sum_{l=1}^{p} x_{il}x_{i'l}$), ultimately finding a linear separation between the observations. A radial kernel uses the function $e^{-\gamma\sum_{l=1}^{p}(x_{il}-x_{i'l})^2}$, where γ is a positive constant.

Among these supervised learning algorithms, we expect logistic regression, LDA, and SVM with a linear kernel to provide an accurate prediction model if the predictors are linearly associated with the response. Likewise, we expect these methods to perform poorly and QDA, decision tree, random forest, and SVM with a radial kernel to perform well if the relationship between C_i and X_i is not linear.

Once classification is performed with the supervised learning algorithms, $\hat{C}_i$ indicates the recommended treatment for subject i that is believed to be beneficial. After we employ the supervised learning algorithms to build the prediction models, we assess these results with the validation approach outlined below.

3.2.3 Validating the Prediction Model

While classification can be performed on our dataset, the question still remains of whether the subgrouping was effective or not. Therefore, a validation approach has been developed to assess this problem. Suppose that a higher response is desired over time. Then treatment $Z = 1$ is deemed beneficial for patient i if $\hat{C}_i = 1$, as $\hat{\beta}_1 + \hat{\alpha}_{1i}$ is the parameter estimate for $Z_i T_{ij}$ and $C_i = 1$ means this estimate is positive. Likewise, treatment $Z = -1$ is deemed beneficial for patient i if $\hat{C}_i = -1$.

In this section, we assume that the desired outcome is for the response to decrease over time, which corresponds to our application of the depression study (i.e., treatments $Z = 1$ and -1 are deemed beneficial for patients whose $\hat{C}_i$ are

-1 and 1, respectively). For each subgrouping method described in Section 3.2.2, let U_i be the indicator that the patient received the treatment determined to be beneficial through the prediction model. We then formulate the following marginal regression model:

$$Y_{ij} = \gamma_0 + \gamma_1 U_i T_{ij} + \gamma_2 T_{ij} + \epsilon_{ij} \tag{3.3}$$

and estimate parameters γ_k, $k = 0, 1, 2$, using the generalized estimating equation approach (Liang and Zeger, 1986) that can yield unbiased and more efficient estimators than the one ignoring the within-subject correlation. For patients who receive the beneficial treatment, we should notice a decrease in their response over time, thus $\hat{\gamma}_1$ should be significantly negative. If this is the case, then the proposed subgrouping analysis is appropriate and beneficial. We remark that while this may appear to be similar to model (3.1), the key difference is the use of U_i rather than Z_i. We are no longer concerned with which treatment the patient receives, as we were in model (3.1), but rather with whether or not the subject received the treatment that was deemed beneficial in the building of the prediction model.

It may be the case that multiple subgrouping approaches prove to be beneficial but one must be chosen. The best classification approach is the one that distinguishes the two groups (did and did not receive the treatment predicted to be beneficial) the most. This is determined by the one with the largest Wald test statistic among effective prediction models.

3.3 Simulation Studies

In this section, we assess the proposed method through three simulation studies. First, we assume that subgrouping is appropriate and use both a linear and nonlinear form of the random slopes. Finally, we assume that subgrouping is not appropriate

and generate random slopes that are not dependent on the data. For these, a sample size of 200 for the training dataset and 100 for the testing dataset were modeled as:

$$Y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i}) Z_i T_{ij} + \beta_2 T_{ij} + e_{ij}, \quad j = 1, \ldots, 6, \tag{3.4}$$

where $(\beta_0, \beta_1, \beta_2)' = (0, 0, -0.2)'$, Z_i was randomly chosen as either -1 or 1 for the treatment assignment with a probability of 0.5, T_{ij} was the index of time j, α_{0i} was randomly generated from a uniform distribution between -1 and 1, and $e_{ij} = (e_{i1}, \ldots, e_{i6})'$ was randomly selected from a multivariate normal distribution with mean 0 and covariance matrix R, where R has a compound symmetry structure with a correlation coefficient of 0.7, i.e.,

$$\mathbf{R} = \begin{bmatrix} 1 & 0.7 & 0.7 & 0.7 & 0.7 & 0.7 \\ 0.7 & 1 & 0.7 & 0.7 & 0.7 & 0.7 \\ 0.7 & 0.7 & 1 & 0.7 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 1 & 0.7 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.7 & 1 & 0.7 \\ 0.7 & 0.7 & 0.7 & 0.7 & 0.7 & 1 \end{bmatrix}.$$

Six independent variables independent variables were used, which act as characteristics of the patient; X_{1i}, X_{2i}, and X_{3i} were generated randomly from a standard normal distribution, while X_{4i}, X_{5i}, and X_{6i} were binary variables assigned a value randomly chosen as either -1 or 1 for subject i with a probability of 0.5.

The training dataset was fit to model (3.4), and the seven supervised learning algorithms described in Section 3.2.2 were utilized based on the predictor vector $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})'$. Once prediction models were developed on the training dataset, subjects in the testing dataset were classified based on these models; misclassification error rates were computed, defined as the proportion of

times $C_i \neq \hat{C}_i$ for $i = 1, \ldots, 100$. Finally, the proposed validation approach was used to determine the appropriateness of our subgrouping and the best subgrouping method. A total of 1000 simulations were run for each type of random slope.

3.3.1 Linear Association

We will let our random slopes be:

$$\alpha_{1i} = -0.5X_{1i} + X_{4i} + \zeta_i,$$

where ζ_i is the error term randomly generated from a standard normal distribution. Note that we intentionally set X_4 to be a strong variable and X_1 to be weaker in order to see if logistic regression and the random forest algorithm will detect these variables as significant. From there, we fit our mixed model to our data using the equation for Y_{ij} specified above. Figure B.1 in Appendix B contains the histograms for both the random slopes and random slope estimates of our training data set for one of our simulations. We see that both of these plots are centered around 0. Since we are determining a value of C_i based on the sign of our random slope estimate, this symmetry is ideal. In addition, the shape of each histogram is similar. Again, this is ideal. In the end, 96.25% of our 200 data points had the same sign for their random slope and random slope estimate, on average. From here, our supervised learning algorithms were utilized and prediction models were developed.

The same initial steps were then performed with the testing dataset. Again, a random slope was calculated for each observation, then estimated by fitting our mixed model for the variable Y_{ij}. Here, the random slopes and random slope estimates had the same sign for 95.32% of the 100 data points, on average. Figure B.2 in Appendix B contains the plots of the random slopes and random slope

estimates for the testing dataset for one of our simulations. We notice a similar overall look to these histograms as we did above.

The classification methods developed for the training dataset were then utilized on the testing dataset. As stated section 3.2.2, the random forest algorithm, by default, considers $\sqrt{p}$ variables at each split, where p is the number of variables. In our case, this would be $\sqrt{6}$, which rounds to 2. We investigated using one through six variables at each split to see if we could obtain better results, however the results were not significantly different so we chose to simply use the default of randomly selecting two variables at each split.

The average of the misclassification error rates reported in Table 3.1 show that all methods except the decision tree perform similarly with the linear approaches producing slightly lower error rates.

Table 3.1: Average Misclassification Error Rates for Testing Data with Linear Random Slopes

Type	Method	Error Rate
Linear	Logistic	19.17%
	LDA	18.87%
	SVM (linear)	19.05%
Nonlinear	QDA	19.45%
	Decision tree	24.45%
	Random forest	20.57%
	SVM (radial)	19.12%

In addition, the validity of our classification was assessed on the testing dataset. Table 3.2 displays the proportion of times each method produced a significantly negative $\hat{\gamma}_1$, as well as the average and standard deviation of $\hat{\gamma}_1$ among the 1000 simulations. Since we were assuming that a lower response is desired over time, $\hat{\gamma}_1$ is significantly negative if the proposed prediction model is effective. This was achieved, as shown in Table 3.2; each subgrouping approach produced a significant parameter estimate all 1000 times, indicating that the proposed method performs

well in the case of linear random slopes. Moreover, the average and standard deviation of $\hat{\gamma}_1$ were approximately the same for all methods except the decision tree, suggesting that both linear and nonlinear classification approaches performed relatively equally. Due to ease of interpretation and simplicity, however, we would recommend the use of a linear classification approach here.

Table 3.2: Validation Results for Testing Data with Linear Random Slopes

Type	Method	Proportion	Mean($\hat{\gamma}_1$)	SD($\hat{\gamma}_1$)
Linear	Logistic	1.000	-1.976	0.233
	LDA	1.000	-1.994	0.227
	SVM (linear)	1.000	-1.988	0.228
Nonlinear	QDA	1.000	-1.970	0.234
	Decision tree	1.000	-1.670	0.301
	Random forest	1.000	-1.899	0.240
	SVM (radial)	1.000	-1.982	0.229

We note that we intentionally set X_4 to be a strong variable and X_1 to be weaker in order to find if logistic regression and the random forest algorithm would detect these variables as significant. For logistic regression, variables were considered significant if their corresponding p-value was less than 0.05. The number of times each variable was considered significant is displayed in Table B.2 in Appendix B. X_4 was always shown to be a significant variable in the model while X_1 was significant 99.3% of the time. Moreover, the remaining four variables were significant 5 to 6% of the time.

In addition, we were able to identify important factors when the random forest algorithm was implemented. Table B.3 in Appendix B displays the number of times each variable was ranked first, second, and third most important using the random forest algorithm. X_4 was always considered the most important factor and X_1 was the second most important variable over 90% of the time.

3.3.2 Nonlinear Association

We used the same information as above, however we added two nonlinear components to the random slopes in Section 3.3.1. The nonlinear random slopes were then generated as:

$$\alpha_{1i} = -0.5X_{1i} + X_{4i} + X_{1i}X_{2i} - 0.7X_{3i}^2X_{4i} + \zeta_i.$$

We would like the sign of each random slope and its corresponding random slope estimate to be the same, as we will classify our data based on the sign of the random slope estimates. For the training dataset, the random slopes and random slope estimates had the same sign for 95.83% of the 200 observations, on average. Histograms of the random slopes and random slope estimates for the training dataset for one of the simulations are in Figure B.3 in Appendix B.

The seven subgrouping methods were then used to create our prediction models on our training dataset. From there, our attention shifted towards the testing dataset. For this, 94.52% of the random slopes and random slope estimates had the same sign, on average. Figure B.4 in Appendix B contains the histograms of the random slopes and random slope estimates for the testing dataset for one of the simulations.

The prediction models developed for the training dataset were then utilized on the testing dataset. Table 3.3 summarizes the error rates for each method.

The results in Table 3.3 confirm that all nonlinear approaches outperformed the linear ones in terms of a lower misclassification error rate; SVM with a radial kernel had the lowest error rate with the random forest algorithm less than one percent behind.

We also remark that X_5 and X_6 were never considered among the top three most

Table 3.3: Average Misclassification Error Rates for Testing Data with Nonlinear Random Slopes

Type	Method	Error Rate
Linear	Logistic	39.06%
	LDA	39.02%
	SVM (linear)	39.27%
Nonlinear	QDA	33.24%
	Decision tree	34.73%
	Random forest	30.39%
	SVM (radial)	29.63%

influential variables among all 1000 simulations with the random forest approach. Considering these two variables were the only ones not used in the calculation of the above random slope, this result was not surprising. Table B.4 in Appendix B displays the number of times each variable was ranked first, second, and third most important using the random forest algorithm.

For logistic regression, variables were considered significant if their corresponding p-value was less than 0.05. The number of times each variable was considered significant is displayed in Table B.5 in Appendix B. Here, X_1 and X_4 were significant the most often. Since these variables are stronger than the others in the calculation of the random slopes, this is not surprising.

Table 3.4 shows that the prediction models based on the nonlinear classification approaches were better than those of the linear type in terms of a higher proportion of times that $\hat{\gamma}_1$ was significant; this estimate also has a smaller value, indicating that the beneficial treatment will lower the response more over time. When comparing the nonlinear algorithms, SVM with a radial kernel and random forest were the best classification methods studied. For these methods, the average $\hat{\gamma}_1$ was -1.615 and -1.587, respectively. On the other hand, the linear approaches all had the highest error rates and accordingly had the fewest significant parameter estimates with the validation approach. In fact, the proportion of significant pa-

rameter estimates decreased by about 30% with the linear approaches from when we had linear random slopes in Section 3.3.1. In addition, the average $\hat{\gamma}_1$ for these methods ranged from -0.737 to -0.815, indicating the subgrouping was beneficial but not as beneficial as that of the nonlinear methods.

Table 3.4: Validation Results for Testing Data with Nonlinear Random Slopes

Type	Method	Proportion	Mean($\hat{\gamma}_1$)	SD($\hat{\gamma}_1$)
Linear	Logistic	0.718	-0.815	0.383
	LDA	0.718	-0.814	0.385
	SVM (linear)	0.666	-0.737	0.400
Nonlinear	QDA	0.913	-1.088	0.369
	Decision tree	0.952	-1.280	0.418
	Random forest	0.997	-1.587	0.360
	SVM (radial)	0.999	-1.615	0.350

3.3.3 No Association

In the two previous simulation studies, the random slopes were calculated based on data values. We now investigate when the random slopes are not dependent on the data at all. This represents the null hypothesis, that subgrouping is not appropriate. Here, we will let our random slopes be randomly generated from a standard normal distribution.

$$\alpha_{1i} = \zeta_i.$$

For the training dataset, the random slopes and random slope estimates had the same sign for 95% of the 200 observations, on average. Histograms of the random slopes and random slope estimates for the training dataset for one of the simulations are in Figure B.5 in Appendix B. Once the prediction models were built the independent variables, our attention shifted to the testing dataset. Here, 92.4% of the random slopes and random slope estimates had the same sign, on

average. Figure B.6 in Appendix B contains the histograms of the random slopes and random slope estimates for the testing dataset for one of the simulations.

The averages of the misclassification error rates are also reported in Table 3.5. These are all near 50%, indicating that we are just as likely to correctly classify an individual as we are to misclassify them. This is because the random slopes are not associated with the data at all, yet the prediction models are built with the independent variables of X_i. As such, the classification approaches cannot perform well.

Table 3.5: Average Misclassification Error Rates for Testing Data for Randomly Generated Random Slopes

Type	Method	Error Rate
Linear	Logistic	50.1%
	LDA	50.1%
	SVM (linear)	50.1%
Nonlinear	QDA	50.1%
	Decision tree	49.9%
	Random forest	50.1%
	SVM (radial)	50.2%

Table 3.6 displays the results from the validation approach. Regardless of the classification approach, the proportion of times that the prediction model is deemed significant through validation is close to a nominal level of 0.05. This proportion also represents the type I error, where we recommend subgrouping when it is not appropriate. These results indicate that when subgrouping is not appropriate, all the classification methods do not recommend any subgrouping.

Logistic regression also assessed whether or not a variable was considered significant. Here, all variables were significant close to a nominal level of 5% of the time. In addition, variable importance was ranked with the random forest algorithm. Here, all variables were considered important a relatively equal amount of the time.

Table 3.6: Validation Results for Testing Data with Randomly Generated Random Slopes

Type	Method	Proportion	Mean($\hat{\gamma}_1$)	SD($\hat{\gamma}_1$)
Linear	Logistic	0.057	-0.0019	0.200
	LDA	0.057	-0.0019	0.200
	SVM (linear)	0.053	0.0016	0.209
Nonlinear	QDA	0.054	-0.0003	0.200
	Decision tree	0.064	-0.0025	0.202
	Random forest	0.048	0.0082	0.200
	SVM (radial)	0.060	0.0043	0.208

3.4 Data Analysis

In this section, the proposed subgrouping method was applied to the Women Entering Care study on depression that involved low-income and minority women. As stated in Section 1.2, seven independent variables were assessed for each patient: age, marital status, schooling, housing, ethnicity, where the patient was born, and whether or not the patient works. Only the first variable listed is numeric and the remaining categorical variables were converted to a binary variable. Grouping was based on logical grouping and an approximately equal number of observations in each group. Table B.8 and B.9 in Appendix B displays our assignment of binary variables and summary statistics of these variables, respectively.

To assess our procedure, we split the data into two smaller datasets with two thirds of the data in the training dataset and a third in the testing. Our proposed method had similar findings as that of Miranda et al. (2003). When using the training dataset to compare medication and cognitive behavioral therapy to the referral group, our mixed model estimated that the parameters for the interactions between treatment and time were $\hat{\beta}_1 = -0.445$ and -0.249 (t-value $= -4.02$ and -2.22), respectively. Therefore, our method was also able to show that medication and cognitive behavioral therapy are better at treating depression than being referred to community care.

Our attention shifted to comparing the medication and cognitive behavioral therapy groups. We start by fitting marginal model (3.1) to the training dataset and using the generalized estimating equations with an AR(1) correlation structure, as this is an established method. With this, the parameter estimate for the interaction between treatment and time for the training dataset was $\hat{\delta}_1 = 0.0117$ (Wald = 0.01, p-value = 0.92), meaning we could not determine a difference in outcomes of the treatment and would recommend either to a patient. Using proposed model (3.2) resulted in an estimate of $\hat{\beta}_1 = -0.0402$ (t-value = -0.38), again leading to the inability to conclude a significant difference in average outcomes between treatments. Therefore, analysis with the random effects of model (3.2) was performed, taking into account individual treatment effects over time. Figures B.7 and B.8 in Appendix B contain histograms of the random slope estimates for both the training and testing datasets. We see that these are symmetric around zero. In addition, the decision tree is in figure B.9 in Appendix B.

When building the prediction models, the seven independent variables were used for each patient. Table 3.7 displays the results when using model (3.3) to check the validity of our approach on the testing dataset with prediction models built on the training dataset. Not all of the parameter estimates for the interaction between treatment and time ($\hat{\gamma}_1$) were negative. These positive estimates result from linear classification methods, which indicate these are not appropriate subgrouping approaches. All parameter estimates for the interaction between treatment and time were negative with the nonlinear supervised learning algorithms, indicating that the depression score decreased over time for those individuals that received the treatment deemed to be beneficial. Not all these methods, however, produced significant results; the only method that found subgrouping to be appropriate and beneficial was the random forest algorithm with QDA being close to significant. The random

forest detects that housing and ethnicity were the two most important variables in classifying the data.

Table 3.7: Validation Results on Testing Dataset for Depression Data

Type	Method	$\hat{\gamma}_1$	SE	Wald	p-value
Linear	Logistic	0.163	0.385	0.18	0.6650
	LDA	0.163	0.385	0.18	0.6650
	SVM (linear)	−0.129	0.406	0.10	0.2750
Nonlinear	QDA	−0.549	0.383	2.05	0.0761
	Decision tree	−0.379	0.387	0.96	0.1635
	Random forest	−0.867	0.365	5.64	0.0090
	SVM (radial)	−0.251	0.387	0.42	0.2587

Subgrouping was not considered appropriate with any of the linear classification approaches. In fact, logistic regression did not detect any of the predictors as significant. Siddique, et al. (2012) found similar results in their study with growth mixture modeling. Their results with logistic regression were able to identify which trajectory class a patient should be in, however it did not determine a decision rule for beneficial treatment.

The validation results show the importance of performing multiple subgrouping approaches and comparing the results. Here, we do not know the relationship between the treatment effect and the independent variables, therefore we do not know which subgrouping approach will be best. Our analysis shows that the nonlinear approaches are best but also shows us that not all these nonlinear approaches perform well. Simply picking one method for subgrouping is not enough; we must perform multiple to find the most advantageous method.

3.5 Discussion

Unlike most of the existing methods referred to in Section refsection:SubgroupIntroduction, our proposed procedure offers a complete process for subgrouping and validation; it

utilizes a random effects linear model to assess the treatment effects over time for each subject, builds prediction models based on classification algorithms, and determines whether or not the subgroups are appropriate and beneficial. This whole process can be easily implemented using existing packages in statistical software such as R and SAS. To secure good performance for subgroup identification, repeated measures within the subject are required to separate variance components and identify individual treatment effects successfully (Senn, 2016).

With the numerical studies, all classification methods performed about the same with the linear random slopes, however for the nonlinear random slopes, the linear classification approaches performed poorly. This could lead to the recommendation of always using nonlinear classification approaches as the preferred method for subgrouping. While the results with such a nonlinear approach would still be good, the interpretation would not be as easy. As such, it is recommended that various classification approaches are considered to find the best subgrouping strategy. Real data analysis also confirms the importance of performing multiple subgrouping approaches and comparing the results. Our analysis showed that the nonlinear approaches were best but also showed that not all these nonlinear approaches perform well. In fact, the decision tree performed the worst among all classification approaches. Moreover, the simulation results indicate that our validation approach is not only simple, but also powerful. The validation approach produced significant results more often when the proper type of classification approach was used, while also having the probability of a type I error close to a nominal level under the null hypothesis.

Chapter 4

Conclusion

Two methods have now been proposed for comparing treatments over time. The first, in Chapter 2, used growth curves with polynomial regression. While other methods limit themselves to a linear trajectory, the BIQIF is able to identify the optimal degree of the polynomial, which may or may not be linear. Quadratic inference functions then calculate the coefficients of the each curve. From here, we are able to assess the equality of the growth curves with the asymptotically chi-square test statistic T. If the growth curves are not all equal, this indicates a difference in the performance of the treatments over time. Further analysis could be done to assess the equality of certain treatments. A limitation here, however, is that a difference in responses may not be found among several treatments. In such a case, it cannot be determined which treatment will perform better over time.

To overcome this limitation, an additional method was proposed in Chapter 3. While the mixed model is also able to identify a beneficial treatment, it can also have the same result as growth curve analysis in that it can show that the responses from two treatments are not significantly different over time. As such, the random effect for each individual is used to subgroup individuals according to a beneficial treatment for the patient. A validation approach was also developed to confirm

that the subgrouping is appropriate and beneficial.

There are several advantages to the two methods proposed. While effective, they are also easy to implement. Both these procedures can be employed with already existing packages in SAS and R. In addition, no assumptions are made about the data. An assumption made on most data is that the response is normally distributed. This is not a requirement for either of the proposed procedures. Finally, both methods can handle missing data. This is especially advantageous when working with longitudinal data, when not all follow up appointments are always attended.

Both these methods were applied to the Women Entering Care study. While these methods were shown to be effective here, they could be beneficial to other medical studies, as well. Both methods allow for the analysis with missing data, which can be prevalent in longitudinal studies.

Appendix A

Proofs From Chapter 2

The following conditions are required to establish the asymptotic properties of the estimator $\hat{\beta}$:

1) The parameter space $\mathbb{B}$ is compact and β_0 is in its interior.

2) There exists a β_0 such that $E\{g_i(\beta)\} = 0$, $i = 1, \ldots, n$, if and only if $\beta = \beta_0$.

3) $g_i(\beta)$ is almost surely continuously differentiable in β and $n^{-1}\sum_{i=1}^{n} \partial g_i(\beta)/\partial\beta$ converges in probability to a full rank matrix of $\Phi = E\{\partial g_i(\beta)/\partial\beta\}$.

4) $C = n^{-1}\sum_{i=1}^{n} g_i(\beta)g_i(\beta)'$ converges to Σ in probability, where $\Sigma = E\{g_i(\beta)g_i(\beta)'\}$ is positive definite.

They are the standard conditions commonly assumed in marginal regression procedures for longitudinal data such as the generalized estimating equations (Liang and Zeger, 1986) and quadratic inference functions (Qu, Lindsay and Li, 2000).

Proof of Theorem 1. Recall that $\hat{\beta}$ minimizes the quadratic inference functions, meaning $\hat{\beta} = \arg\min_{\beta} Q(\beta) = \arg\min_{\beta} n\bar{g}(\beta)'C^{-1}\bar{g}(\beta)$ with $\bar{g}(\beta) = n^{-1}\sum_{i=1}^{n} g_i(\beta)$. Taylor expansion leads to

$$\bar{g}(\hat{\beta}) = \bar{g}(\beta_0) + \dot{\bar{g}}(\check{\beta})(\hat{\beta} - \beta_0), \tag{A.1}$$

where $\dot{\bar{g}} = \partial\bar{g}(\beta)/\partial\beta$ and $\check{\beta}$ lies between $\hat{\beta}$ and β_0. By multiplying both sides in (A.1) by $\dot{\bar{g}}(\hat{\beta})'C^{-1}$, we have

$$\dot{\bar{g}}(\hat{\beta})'C^{-1}\bar{g}(\hat{\beta}) = \dot{\bar{g}}(\hat{\beta})'C^{-1}\bar{g}(\beta_0) + \dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\check{\beta})(\hat{\beta} - \beta_0). \tag{A.2}$$

The left hand side in (A.2) is zero since minimizing $Q(\beta)$ to obtain $\hat{\beta}$ is equivalent to solving $\dot{\bar{g}}(\beta)'C^{-1}\bar{g}(\beta)$. As such, (A.2) becomes

$$0 = \dot{\bar{g}}(\hat{\beta})'C^{-1}\bar{g}(\beta_0) + \dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\check{\beta})(\hat{\beta} - \beta_0). \tag{A.3}$$

Accordingly, (A.3) can be rearranged as

$$\begin{aligned} \dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\check{\beta})(\hat{\beta} - \beta_0) &= -\dot{\bar{g}}(\hat{\beta})'C^{-1}\bar{g}(\beta_0) \\ \sqrt{n}(\hat{\beta} - \beta_0) &= -\left\{\dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\check{\beta})\right\}^{-1}\dot{\bar{g}}(\hat{\beta})'C^{-1}\cdot\sqrt{n}\bar{g}(\beta_0) \end{aligned} \tag{A.4}$$

By the central limit theorem and condition 4, we have

$$\sqrt{n}\bar{g}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0) \xrightarrow{d} N(0, \Sigma). \tag{A.5}$$

Note that

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left\{\dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\check{\beta})\right\}^{-1}\dot{\bar{g}}(\hat{\beta})'C^{-1}\cdot\underbrace{\sqrt{n}\bar{g}(\beta_0)}_{N(0,\,\Sigma)}$$

It follows from (A.4), (A.5), and conditions 3 and 4 that

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, \left\{\dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\breve{\beta})\right\}^{-1}\dot{\bar{g}}(\hat{\beta})'C^{-1}\Sigma(\left\{\dot{\bar{g}}(\hat{\beta})'C^{-1}\dot{\bar{g}}(\breve{\beta})\right\}^{-1}\dot{\bar{g}}(\hat{\beta})'C^{-1})')$$

Let $\Phi = E\{\partial g_i(\beta)/\partial\beta\}$ and $\Sigma = E\{g_i(\beta)g_i(\beta)'\}$. Then

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, (\Phi'\Sigma^{-1}\Phi)^{-1}\Phi'\Sigma^{-1}\Sigma\Sigma^{-1}\Phi(\Phi'\Sigma^{-1}\Phi)^{-1'})$$

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, (\Phi'\Sigma^{-1}\Phi)^{-1}\Phi'\Sigma^{-1}\Phi(\Phi'\Sigma^{-1}\Phi)^{-1'})$$

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, (\Phi'\Sigma^{-1}\Phi)^{-1})$$

$$\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0, V)$$

where $V = (\Phi'\Sigma^{-1}\Phi)^{-1}$.

Now we prove that $V_I - V$ is positive semidefinite. Without loss of generality, $R_i(\alpha)^{-1}$ can be approximated by two basis matrices as $R_i(\alpha)^{-1} = d_1 I_{m_i} + d_2 D_i$, where I_{m_i} is an $m_i \times m_i$-dimensional identity matrix, D_i is a known basis matrix, and d_1 and d_2 are constants. With I_{m_i} and D_i, $g_i(\beta)$ is formulated as

$$g_i(\beta) = \begin{pmatrix} g_{i1}(\beta) \\ g_{i2}(\beta) \end{pmatrix} = \begin{pmatrix} \dot{\mu}_i' A_i^{-1/2} I_{m_i} A_i^{-1/2}(Y_i - \mu_i) \\ \dot{\mu}_i' A_i^{-1/2} D_i A_i^{-1/2}(Y_i - \mu_i) \end{pmatrix}. \tag{A.6}$$

By letting $\Phi_1 = E\left(\partial g_{i1}/\partial\beta\right)$ and $\Sigma_{11} = E\{g_{i1}(\beta)g_{i1}(\beta)'\}$, the asymptotic covariance matrix of $\hat{\beta}$ assuming an independent correlation structure is specified as $V_I = \left(\Phi_1'\Sigma_{11}^{-1}\Phi_1\right)^{-1}$.

We further define $\Phi_2 = E\left(\partial g_{i2}/\partial\beta\right)$ and $\Sigma_{21} = E\{g_{i2}(\beta)g_{i1}(\beta)'\}$, and orthogonalize $g_{i2}(\beta)$ against $g_{i1}(\beta)$, $g_{io}(\beta) = g_{i2}(\beta) - \Sigma_{21}\Sigma_{11}^{-1}g_{i1}(\beta)$, such that $E\{g_{i1}(\beta)g_{io}(\beta)'\} = 0$. By replacing $g_{i2}(\beta)$ in (A.6) with $g_{io}(\beta)$, we obtain $\Phi = E\{\partial g_i(\beta)/\partial\beta\}$ and $\Sigma = E\{g_i(\beta)g_i(\beta)'\}$

$$
\begin{aligned}
V^{-1} &= \Phi'\Sigma^{-1}\Phi \\
&= \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \\ E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \end{pmatrix}' \begin{pmatrix} E\{g_{i1}(\beta)g_{i1}(\beta)'\} & E\{g_{i1}(\beta)g_{i0}(\beta)'\} \\ E\{g_{i0}(\beta)g_{i1}(\beta)'\} & E\{g_{i0}(\beta)g_{i0}(\beta)'\} \end{pmatrix}^{-1} \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \\ E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \end{pmatrix} \\
&= \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \\ E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \end{pmatrix} \begin{pmatrix} E\{g_{i1}(\beta)g_{i1}(\beta)'\} & 0 \\ 0 & E\{g_{i0}(\beta)g_{i0}(\beta)'\} \end{pmatrix}^{-1} \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \\ E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \end{pmatrix} \\
&= \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\}'E\{g_{i1}(\beta)g_{i1}(\beta)'\}^{-1} & E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\}'E\{g_{i0}(\beta)g_{i0}(\beta)'\}^{-1} \end{pmatrix} \begin{pmatrix} E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \\ E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \end{pmatrix} \\
&= E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\}'E\{g_{i1}(\beta)g_{i1}(\beta)'\}^{-1}E\{\frac{\partial g_{i1}(\beta)}{\partial \beta}\} \qquad \text{(A.7)} \\
&\quad + E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\}'E\{g_{i0}(\beta)g_{i0}(\beta)'\}^{-1}E\{\frac{\partial g_{i0}(\beta)}{\partial \beta}\} \\
&= \Phi_1'\Sigma_{11}^{-1}\Phi_1 + (\Phi_2 - \Sigma_{21}\Sigma_{11}^{-1}\Phi_1)'\Sigma_0^{-1}(\Phi_2 - \Sigma_{21}\Sigma_{11}^{-1}\Phi_1) \\
&= V_I^{-1} + \Phi_o'\Sigma_o^{-1}\Phi_o, \qquad \text{(A.8)}
\end{aligned}
$$

where $\Phi_o = \Phi_2 - \Sigma_{21}\Sigma_{11}^{-1}\Phi_1$ and $\Sigma_o = E\{g_{io}(\beta)g_{io}(\beta)'\}$. The proof is completed since Σ_o in (A.8) is positive semidefinite. □

Proof of Theorem 2. The null hypothesis can be specified as

$$H_0 : b_{i1} = \ldots = b_{iq} \quad \text{for all } i = 1, \ldots, p+1, \tag{A.9}$$

where $B_j = (b_{1j}, \ldots, b_{(p+1)j})'$. With $(p+1)(q-1)$ contrasts $b_{ij} - b_{iq} = b^*_{ij}$ for $i = 1, \ldots, (p+1)$ and $j = 1, \ldots, q-1$, we can rewrite (A.9) as

$$H_0 : b^*_{ij} = 0 \quad \text{for all } i = 1, \ldots, p+1 \;\; j = 1, \ldots, q-1. \tag{A.10}$$

We let $\beta^* = (\beta_1^{*\prime}, \beta_2^{*\prime})'$, where β_1^* and β_2^* are vectors of $(p+1)(q-1)$ b^*_{ij}'s and $(p+1)$ b_{iq}'s, respectively. Then, the parameter β can be denoted by $\beta = (I_q \otimes H)\beta^*$ having a $q \times q$ identity matrix I_q and a square matrix H of order p with 1 on the diagonal and the last column, and 0 elsewhere. An estimator of β^* is accordingly obtained by minimizing $Q^*(\beta^*) = n\bar{g}^*(\beta^*)'C^{*-1}\bar{g}^*(\beta^*)$, where $\bar{g}^*(\beta^*) = n^{-1}\sum_{i=1}^n g_i^*(\beta^*)$, $C^* = n^{-1}\sum_{i=1}^n g_i^*(\beta^*)g_i^*(\beta^*)'$, and

$$g_i^*(\beta^*) = \begin{pmatrix} \dot{\mu}_i^{*\prime} A_i^{-1/2} D_{i1} A_i^{-1/2}(Y_i - \mu_i^*) \\ \vdots \\ \dot{\mu}_i^{*\prime} A_i^{-1/2} D_{id} A_i^{-1/2}(Y_i - \mu_i^*) \end{pmatrix}$$

with $\mu_i^* = h\{X_i'(I_q \otimes H)\beta^*\}$. Similar to Theorem 1, the estimator $\hat{\beta}^*$ satisfies

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_1^* - \beta_{10}^* \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_{\beta_1^*\beta_1^*} & \Omega_{\beta_1^*\beta_2^*} \\ \Omega_{\beta_2^*\beta_1^*} & \Omega_{\beta_2^*\beta_2^*} \end{pmatrix}^{-1}\right) = N\left(0, \Omega^{*-1}\right),$$

$$\sqrt{n}\left(\hat{\beta}_1^* - \beta_{10}^*\right) \xrightarrow{d} N\left(0, \left(\Omega_{\beta_1^*\beta_1^*} - \Omega_{\beta_1^*\beta_2^*}\Omega^{-1}_{\beta_2^*\beta_2^*}\Omega_{\beta_2^*\beta_1^*}\right)^{-1}\right), \tag{A.11}$$

where $(\beta_{10}^*, \beta_{20}^*)$ is a true value of (β_1^*, β_2^*) and $\Omega^* = \Phi^{*\prime}\Sigma^{*-1}\Phi^*$ with $\Phi^* =$

$E\{\partial g_i^*(\beta^*)/\partial\beta^*\}$ and $\Sigma^* = E\{g_i^*(\beta^*)g_i^*(\beta^*)'\}$.

Under H_0 in (A.10), the true parameter of β^* and its estimator are specified as $\beta_0^* = (0, \beta_{20}^*)$ and $\check{\beta}^* = (0, \check{\beta}_2^*)$, respectively. We note that $T = Q^*(0, \check{\beta}_2^*) - Q^*(\hat{\beta}_1^*, \hat{\beta}_2^*)$, and simplify the notations for ease of presentation as

$$\begin{aligned}
\frac{1}{n}\frac{\partial Q^*(\beta^*)}{\partial \beta^*} &= \begin{pmatrix} \frac{\partial Q^*(\beta^*)}{\partial \beta_1^*} \\ \frac{\partial Q^*(\beta^*)}{\partial \beta_2^*} \end{pmatrix} = \begin{pmatrix} \dot{Q}_{\beta_1^*} \\ \dot{Q}_{\beta_2^*} \end{pmatrix} = \dot{Q}_{\beta^*}, \\
\frac{1}{n}\frac{\partial^2 Q^*(\beta^*)}{\partial \beta^{*2}} &= \begin{pmatrix} \frac{\partial^2 Q^*(\beta^*)}{\partial \beta_1^{*2}} & \frac{\partial^2 Q^*(\beta^*)}{\partial \beta_1^* \partial \beta_2^*} \\ \frac{\partial^2 Q^*(\beta^*)}{\partial \beta_2^* \partial \beta_1^*} & \frac{\partial^2 Q^*(\beta^*)}{\partial \beta_2^{*2}} \end{pmatrix} = \begin{pmatrix} \ddot{Q}_{\beta_1^*\beta_1^*} & \ddot{Q}_{\beta_1^*\beta_2^*} \\ \ddot{Q}_{\beta_2^*\beta_1^*} & \ddot{Q}_{\beta_2^*\beta_2^*} \end{pmatrix} = \ddot{Q}_{\beta^*}.
\end{aligned}$$

By Taylor expansion, we have

$$Q^*(0, \check{\beta}_2^*)/n = Q^*(0, \beta_{20}^*)/n + \left(\check{\beta}_2^* - \beta_{20}^*\right)' \dot{Q}_{\beta_2^*}(\check{\beta}_2^*) + \frac{1}{2}\left(\check{\beta}_2^* - \beta_{20}^*\right)' \ddot{Q}_{\beta_2^*\beta_2^*}(\tilde{\beta}_2^*)\left(\check{\beta}_2^* - \beta_{20}^*\right),$$

where $\tilde{\beta}_2^*$ lies between $\check{\beta}_2^*$ and β_{20}^*. Similarly, $Q^*(\hat{\beta}_1^*, \hat{\beta}_2^*)/n$ can be extended as

$$Q^*(0, \beta_{20}^*)/n + \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \dot{Q}_{\beta^*}(\hat{\beta}^*) + \frac{1}{2}\begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \ddot{Q}_{\beta^*}(\check{\beta}^*) \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix},$$

where $\check{\beta}^*$ lies between $\hat{\beta}^*$ and β_0^*. It follows from $\dot{Q}_{\beta_2^*}(\check{\beta}_2^*) = 0$ and $\dot{Q}_{\beta *}(\hat{\beta}^*) = 0$

that

$$
\begin{aligned}
T/n &= Q^*(0,\check{\beta}_2^*)/n - Q^*(\hat{\beta}_1^*,\hat{\beta}_2^*)/n \\
&= \frac{1}{2}\begin{pmatrix} 0 \\ \check{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \ddot{Q}_{\beta_2^*\beta_2^*}(\tilde{\beta}_2^*) \begin{pmatrix} 0 \\ \check{\beta}_2^* - \beta_{20}^* \end{pmatrix} \\
&\quad -\frac{1}{2}\begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \ddot{Q}_{\beta^{*2}}(\check{\beta}^*) \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix} \\
&\quad + \underbrace{Q^*(0,\beta_{20}^*)/n - Q^*(\hat{\beta}_1^*,\hat{\beta}_2^*)/n}_{\to 0 \text{ as } n\to\infty} \\
&= \frac{1}{2}\begin{pmatrix} 0 \\ \check{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \ddot{Q}_{\beta_2^*\beta_2^*}(\tilde{\beta}_2^*) \begin{pmatrix} 0 \\ \check{\beta}_2^* - \beta_{20}^* \end{pmatrix} \\
&\quad -\frac{1}{2}\begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \ddot{Q}_{\beta^{*2}}(\check{\beta}^*) \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix} \\
&\quad + o_p(n^{-1})
\end{aligned} \tag{A.12}
$$

Taylor expansion also leads $\dot{Q}_{\beta_2^*}(\check{\beta}_2^*)$ and $\dot{Q}_{\beta *}(\hat{\beta}^*)$ to

$$
\dot{Q}_{\beta_2^*}(\check{\beta}_2^*) = \dot{Q}_{\beta_2}(\beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_2^*}(\check{\beta}_2^* - \beta_{20}^*) + o_p(n^{-1}) = 0, \tag{A.13}
$$

$$
\dot{Q}_{\beta *}(\hat{\beta}^*) = \dot{Q}_{\beta^*}(\beta_0^*) + \ddot{Q}_{\beta_2^*\beta_2^*}(\hat{\beta}_2^* - \beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_1^*}(\hat{\beta}_1^* - 0) + o_p(n^{-1}) = 0. \tag{A.14}
$$

By setting $\dot{Q}_{\beta_2^*}(\check{\beta}_2^*)$ and $\dot{Q}_{\beta *}(\hat{\beta}^*)$ from (A.13) and (A.14) to each other and solving for $(\check{\beta}_2^* - \beta_{20}^*)$, we have

$$
\dot{Q}_{\beta_2}(\beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_2^*}(\check{\beta}_2^* - \beta_{20}^*) + o_p(n^{-1}) =
$$

$$
\dot{Q}_{\beta^*}(\beta_0^*) + \ddot{Q}_{\beta_2^*\beta_2^*}(\hat{\beta}_2^* - \beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_1^*}(\hat{\beta}_1^* - 0) + o_p(n^{-1})
$$

$$
\begin{aligned}
\breve{\beta}_2^* - \beta_{20}^* &= \ddot{Q}^{-1}_{\beta_2^*\beta_2^*}(\underbrace{\dot{Q}_{\beta^*}(\beta_0^*) - \dot{Q}_{\beta_2}(\beta_{20}^*)}_{\to 0 \text{ as } n\to\infty} + \ddot{Q}_{\beta_2^*\beta_2^*}(\hat{\beta}_2^* - \beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_1^*}(\hat{\beta}_1^* - 0) + o_p(n^{-1})) \\
&= \ddot{Q}^{-1}_{\beta_2^*\beta_2^*}(\ddot{Q}_{\beta_2^*\beta_2^*}(\hat{\beta}_2^* - \beta_{20}^*) + \ddot{Q}_{\beta_2^*\beta_1^*}(\hat{\beta}_1^* - 0) + o_p(n^{-1})) \\
&= (\hat{\beta}_2^* - \beta_{20}^*) + \ddot{Q}^{-1}_{\beta_2^*\beta_2^*}\ddot{Q}_{\beta_2^*\beta_1^*}(\hat{\beta}_1^* - 0) + o_p(n^{-1}))
\end{aligned}
$$

$$
\begin{pmatrix} 0 \\ \breve{\beta}_2^* - \beta_{20}^* \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \ddot{Q}^{-1}_{\beta_2^*\beta_2^*}\ddot{Q}_{\beta_2^*\beta_1^*} & I \end{pmatrix} \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix} + o_p(n^{-1}). \tag{A.15}
$$

By substituting (A.15) into (A.12), we obtain

$$
\begin{aligned}
T &= \frac{n}{2}\begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix}' \begin{pmatrix} \ddot{Q}_{\beta_1^*\beta_2^*}\ddot{Q}^{-1}_{\beta_2^*\beta_2^*}\ddot{Q}_{\beta_2^*\beta_1^*} - \ddot{Q}_{\beta_1^*\beta_1^*} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1^* - 0 \\ \hat{\beta}_2^* - \beta_{20}^* \end{pmatrix} + o_p(1) \\
&= \frac{n}{2}(\hat{\beta}_1^* - 0)' \left(\ddot{Q}_{\beta_1^*\beta_2^*}\ddot{Q}^{-1}_{\beta_2^*\beta_2^*}\ddot{Q}_{\beta_2^*\beta_1^*} - \ddot{Q}_{\beta_1^*\beta_1^*}\right)(\hat{\beta}_1^* - 0) + o_p(1). \qquad \text{(A.16)}
\end{aligned}
$$

It consequently follows from $n^{-1}\partial^2 Q^*(\beta^*)/\partial\beta^{*2} \xrightarrow{p} 2\Phi^{*T}\Sigma^{*-1}\Phi^*$, (A.11), (A.16), and Theorem 10.2d in Arnold (1981) that T follows a chi-squared distribution asymptotically under the null hypothesis, and its degrees of freedom is the same as the dimension of β_1^*, $(p+1)(q-1)$. □

Proof of Theorem 3. This proof is similar to the proof of Theorem 3.1 in Wang and Qu (2009) and the proof of Lemma 2 in Cho and Qu (2013). □

Appendix B

Additional Tables and Figures

Table B.1: Summary of Depression Dataset

Variable	All	Medication	Psychotherapy	Control
Marital Status				
Married	76	31	28	17
Partner/Boyfriend	44	11	9	24
Widowed	4	1	3	0
Separated	33	12	13	8
Divorced	13	4	4	5
Never Married	84	27	26	31
School				
8th Grade Or Less	43	19	9	15
Some High School	50	17	15	18
High School Graduate/GED	80	30	25	25
Trade School	8	2	3	3
Some College	55	13	23	19
Completed College	18	5	8	5
Housing				
In Own House/Apartment	157	52	52	53
Projects	14	5	5	4
Parents	23	7	8	8
With Family Or Friends	59	21	18	10
Shelter/Hotel	1	1	0	0
Ethnicity				
Black	112	33	39	40
White	14	6	4	4
Latina	128	47	40	41
Born				
USA	130	41	44	45
Caribbean	6	2	3	1
Central America	89	37	24	28
Other North American Country	5	1	3	1
South America	24	5	9	10
Working				
Not Working	115	41	40	34
Working	141	45	45	51

Figure B.1: Histograms of the Linear Random Slopes and Linear Random Slope Estimates for the Training Dataset

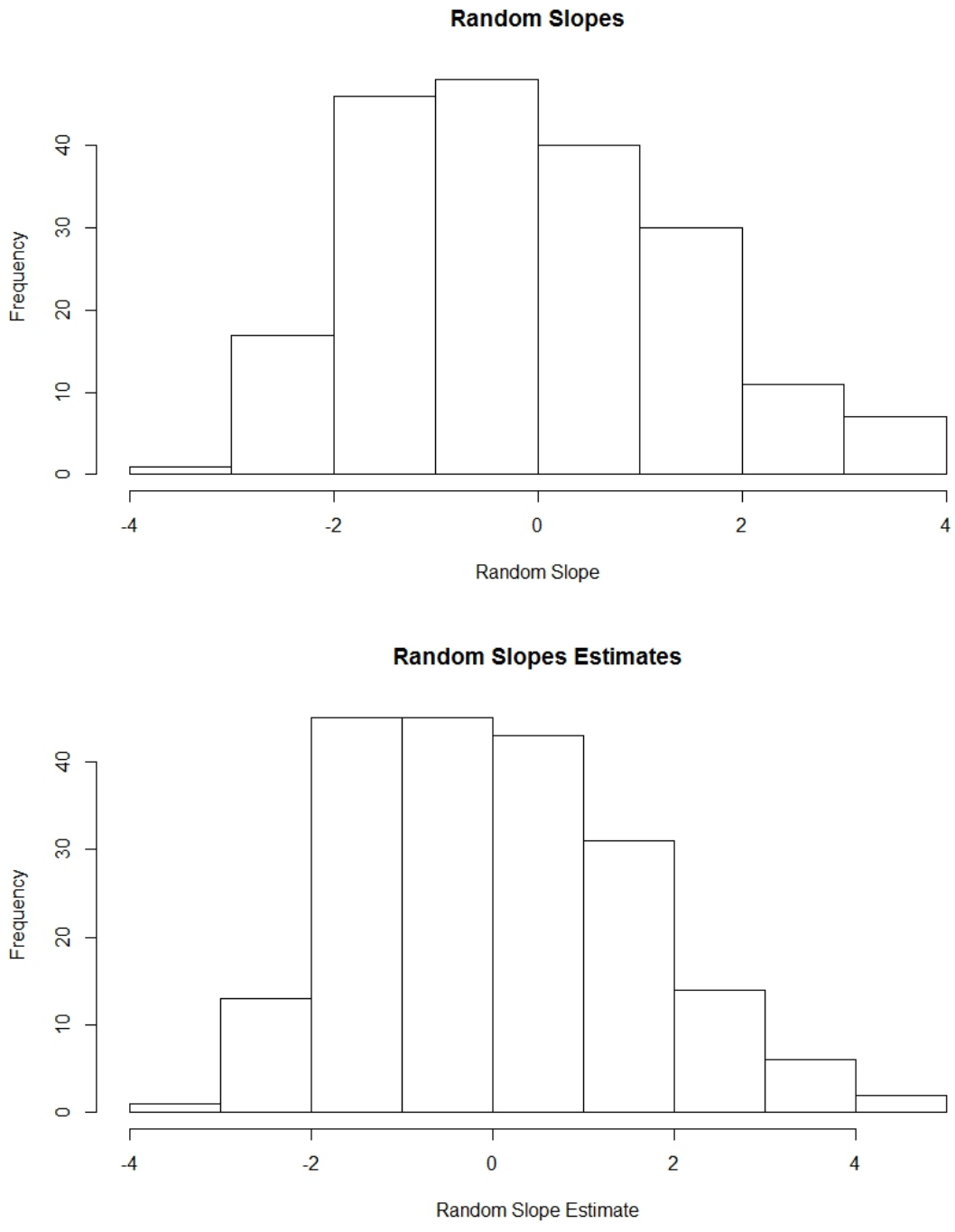
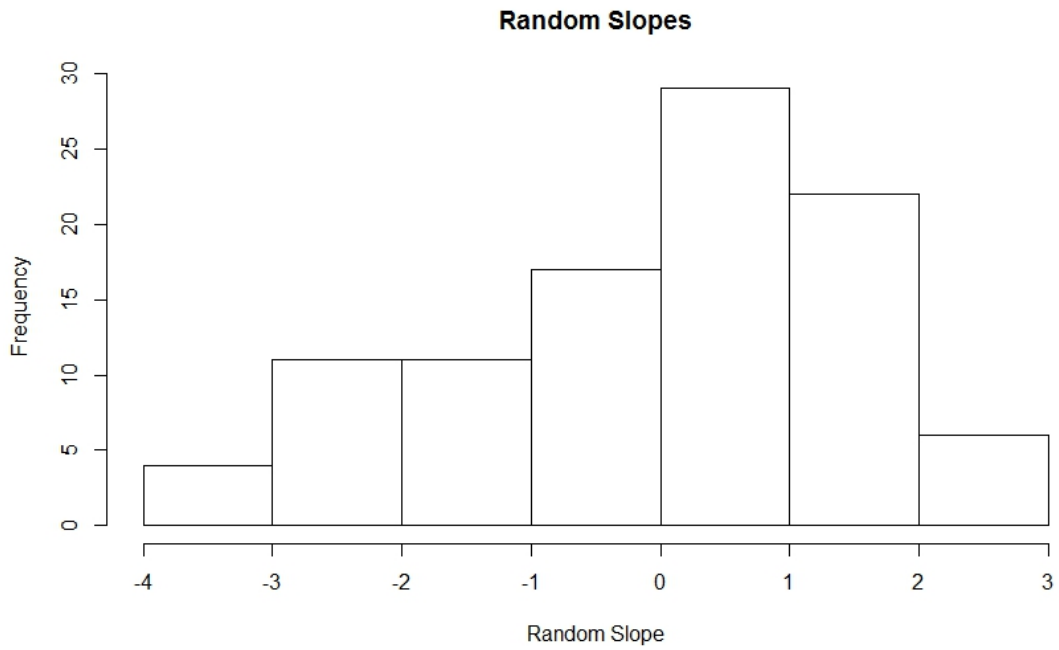


Figure B.2: Histograms of the Linear Random Slopes and Linear Random Slope Estimates for the Testing Dataset



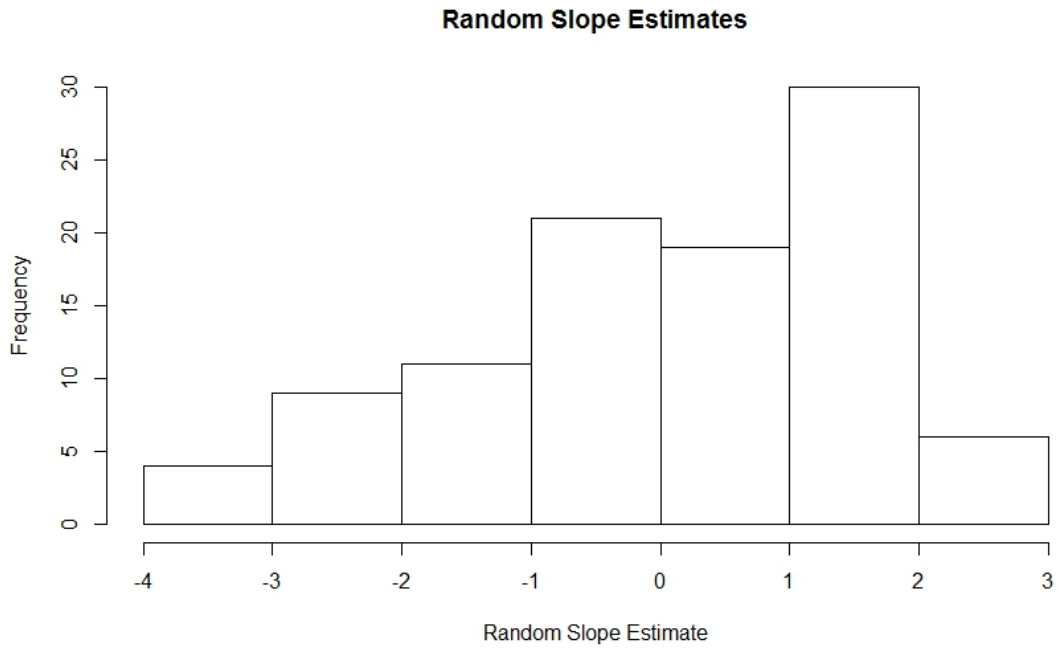


Table B.2: Significant Variables for Logistic Regression with Linear Random Slopes

Variable	Number of Times Variable Is Significant
X_1	993
X_2	61
X_3	51
X_4	1000
X_5	56
X_6	60

Table B.3: Variable Importance with the Random Forest Algorithm with a Nonlinear Random Slope

Variable	First	Second	Third
X_1	0	906	58
X_2	0	35	243
X_3	0	33	247
X_4	1000	0	0
X_5	0	13	234
X_6	0	13	218

Table B.4: Variable Importance with the Random Forest Algorithm with a Linear Random Slope

Variable	First	Second	Third
X_1	402	320	211
X_2	43	220	318
X_3	18	219	294
X_4	537	241	177
X_5	0	0	0
X_6	0	0	0

Figure B.3: Histograms of the Nonlinear Random Slopes and Nonlinear Random Slope Estimates for the Training Dataset

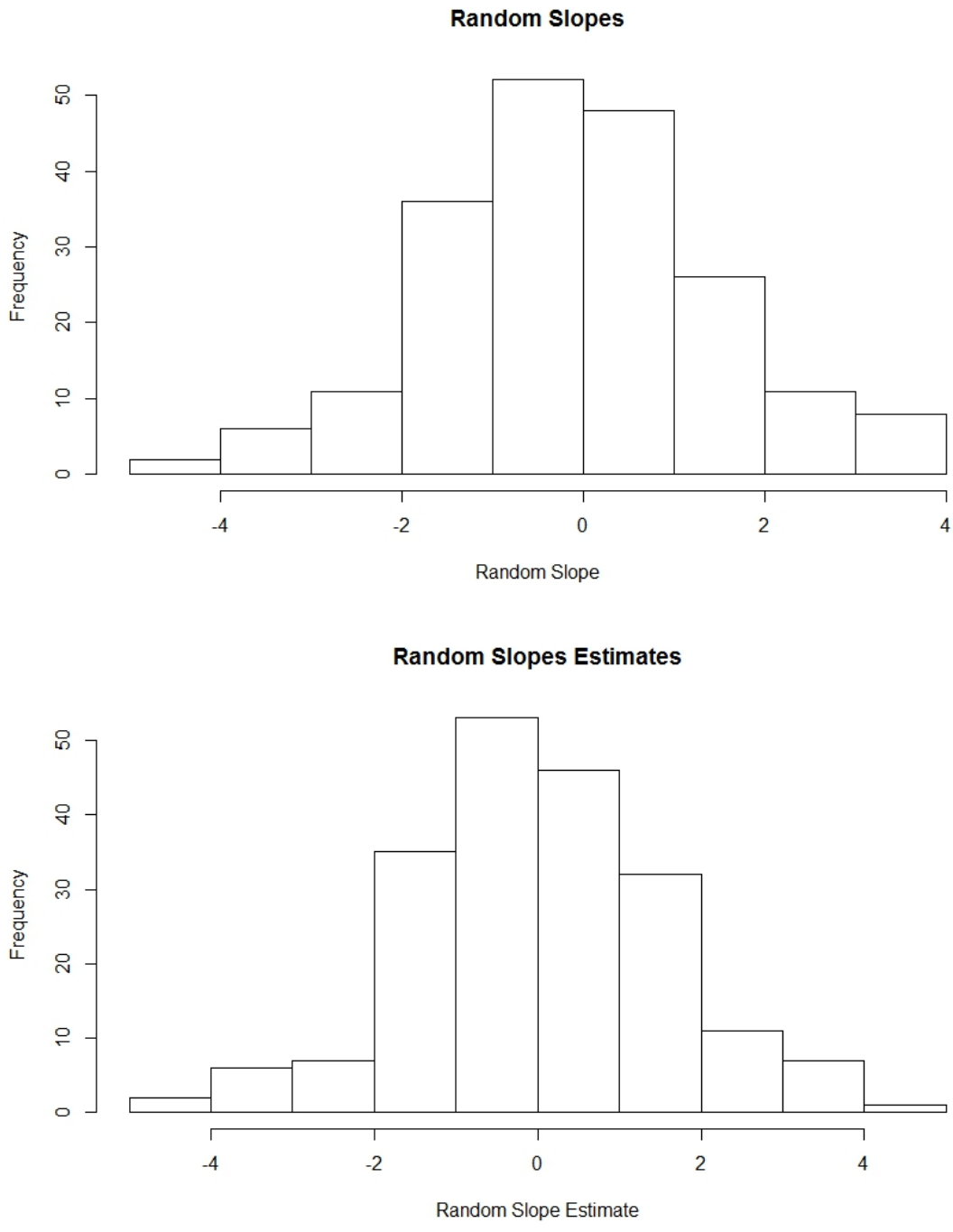


Figure B.4: Histograms of the Randomly Generated Random Slopes and Random Slope Estimates for the Testing Dataset

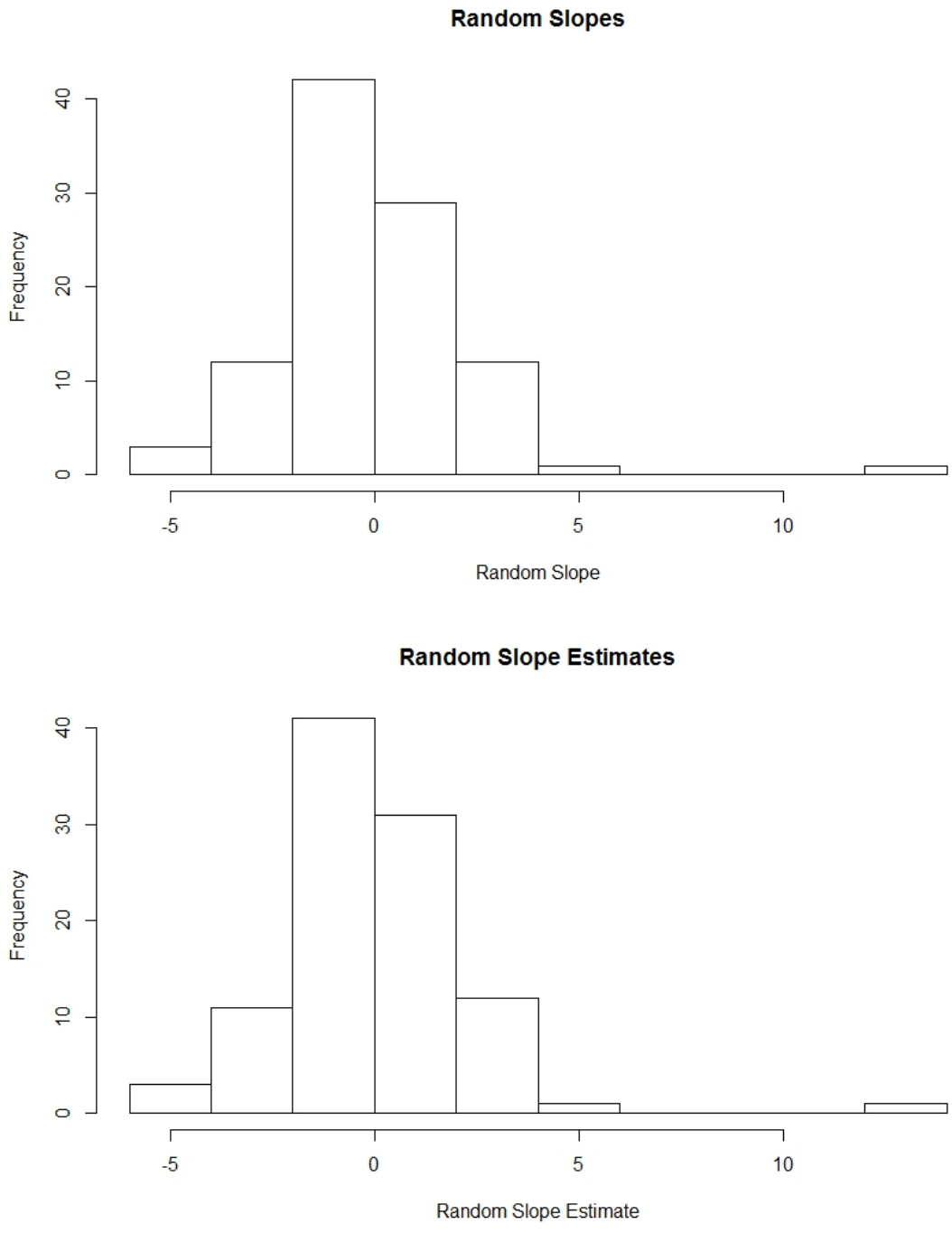


Table B.5: Significant Variables for Logistic Regression with Nonlinear Random Slopes

Variable	Number of Times Variable Is Significant
X_1	780
X_2	52
X_3	96
X_4	841
X_5	51
X_6	55

Table B.6: Significant Variables for Logistic Regression with Randomly Generated Random Slopes

Variable	Number of Times Variable Is Significant
X_1	55
X_2	52
X_3	47
X_4	56
X_5	39
X_6	49

Table B.7: Variable Importance with the Random Forest Algorithm with a Randomly Generated Random Slope

Variable	First	Second	Third
X_1	203	159	135
X_2	195	157	157
X_3	192	150	130
X_4	139	197	194
X_5	149	163	187
X_6	122	174	197

Figure B.5: Histograms of the Randomly Generated Random Slopes and Random Slope Estimates for the Training Dataset

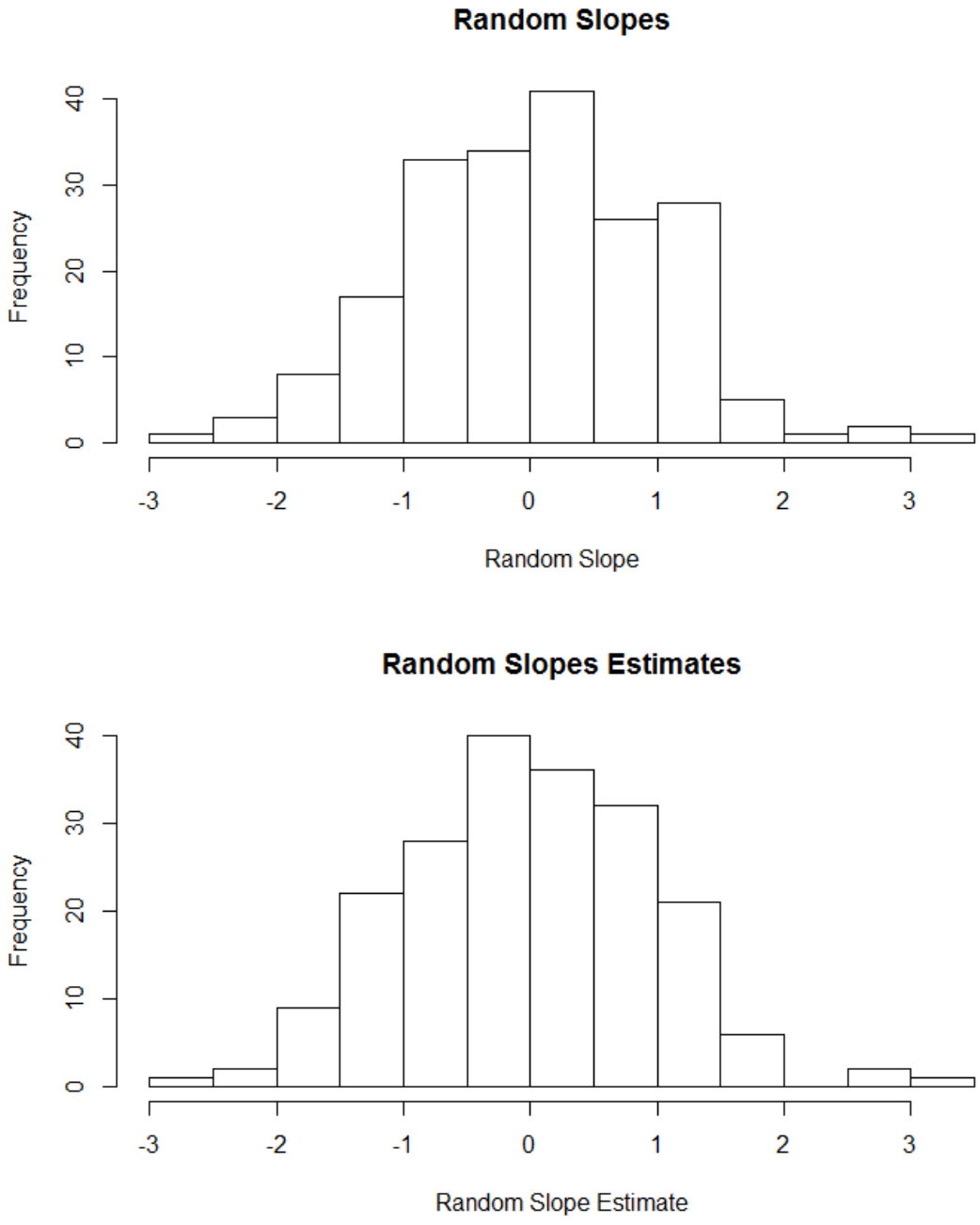


Figure B.6: Histograms of the Nonlinear Random Slopes and Nonlinear Random Slope Estimates for the Testing Dataset

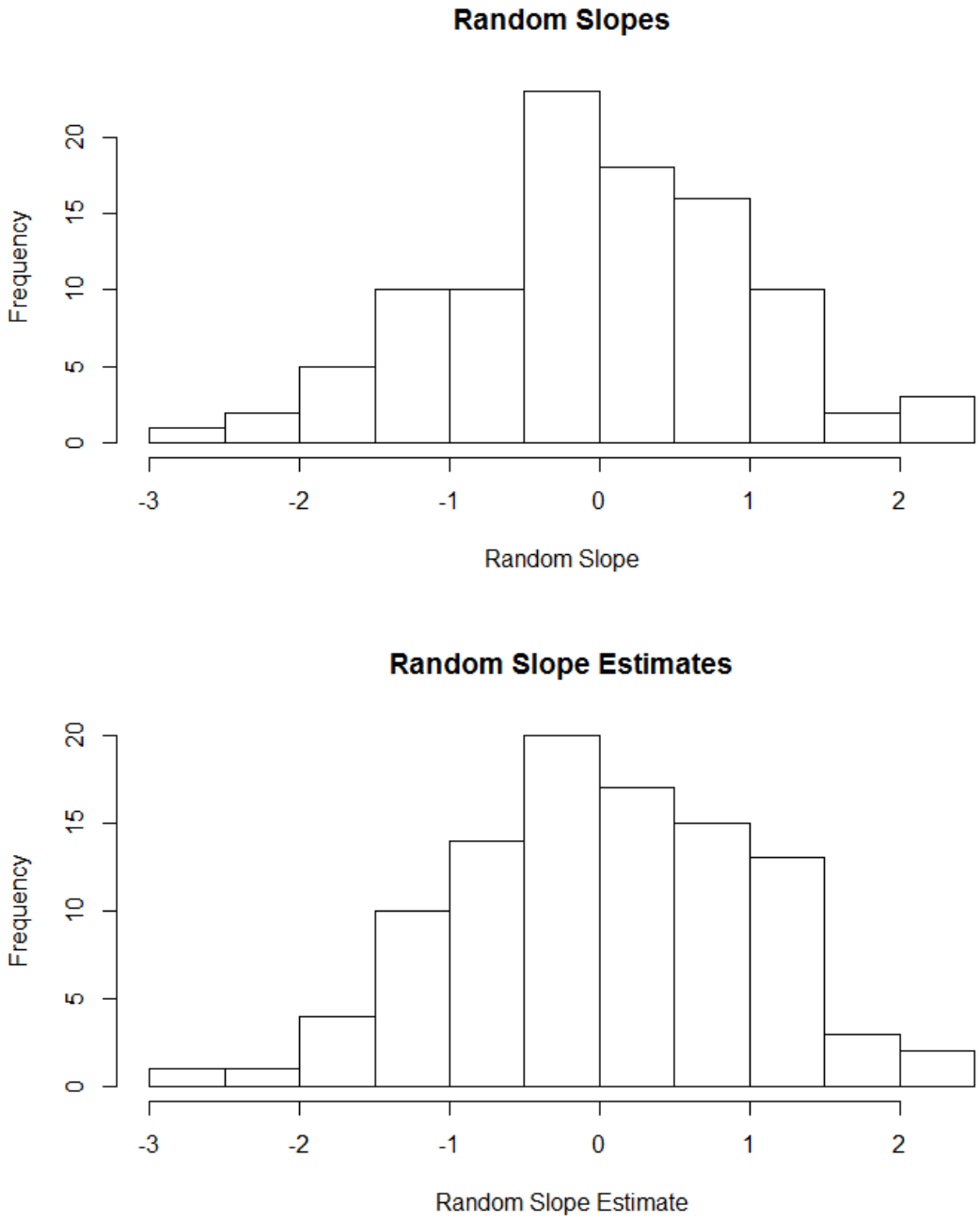


Table B.8: Assignment of Binary Variables for Depression Dataset

Variable	1	0
Marital Status	Married Partner/boyfriend	Widowed Separated Divorced Never married
School	8th grade or less Some HS	HS grad/GED Trade school Some college Completed college
Housing	In own house/apt	Projects Parents With family/friends Jail Shelter/hotel Car/street
Ethnicity	Black	African Asian White Indian Islander Latina More than one
Born	USA	Africa Asia Caribbean Central America Europe Middle East Oceania Other North America South America

Table B.9: Summary Statistics for Covariates in Depression Data

Variable		All	Medication	Psychotherapy	Control
Age	mean (sd)	29.4 (7.95)	28.8 (6.59)	29.7 (7.98)	29.7 (9.15)
Marital	1	120	42	37	41
	0	134	44	46	44
School	1	93	36	24	33
	0	161	50	59	52
Housing	1	157	52	52	53
	0	97	34	31	32
Ethnicity	1	112	33	39	40
	0	142	53	44	45
Born	1	130	41	44	45
	0	124	45	39	40
Working	1	115	41	40	34
	0	141	45	45	51

Figure B.7: Histogram of the Random Slope Estimates for the Training Dataset from Depression Data

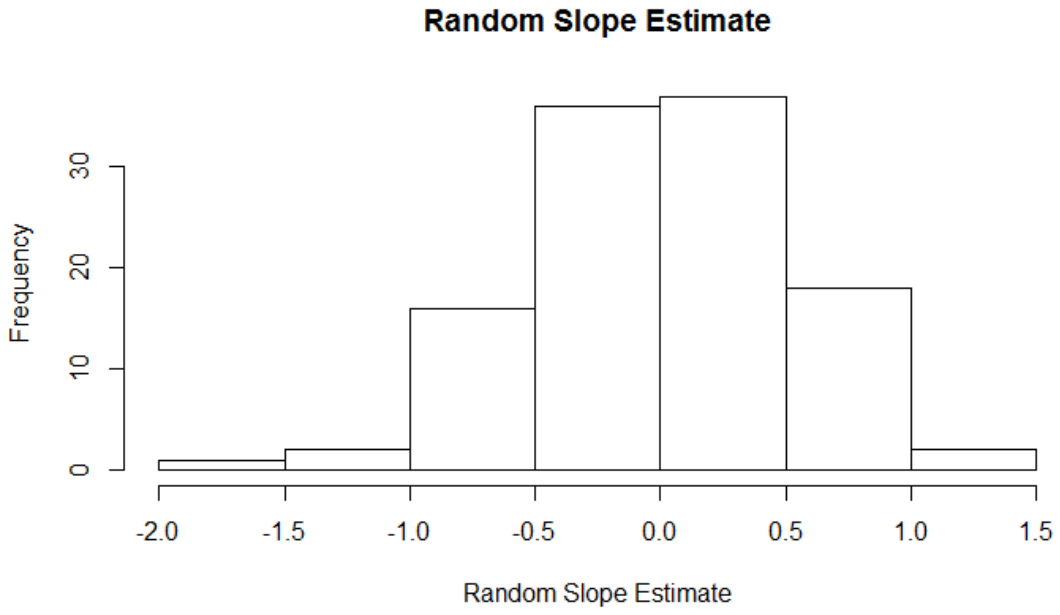


Figure B.8: Histogram of the Random Slope Estimates for the Testing Dataset from Depression Data

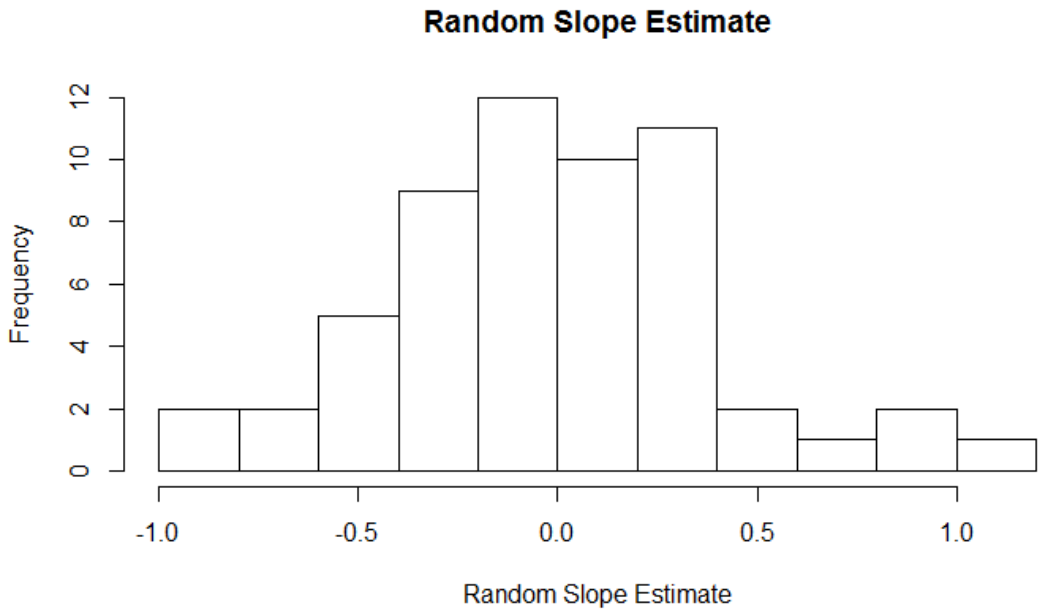
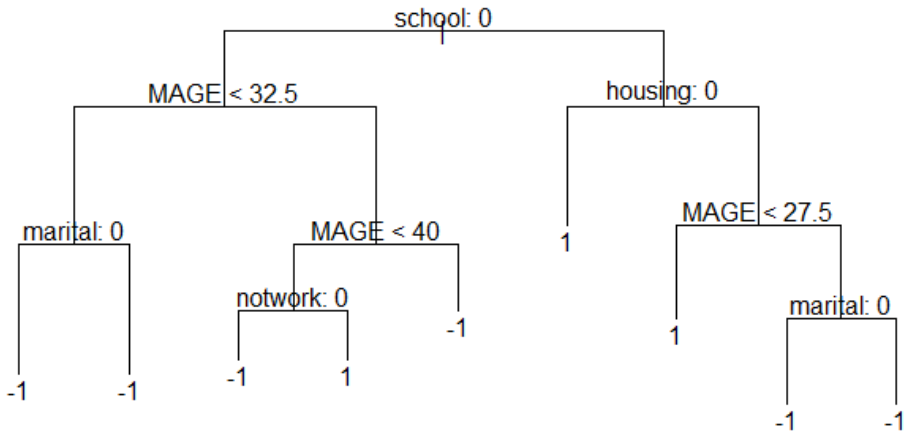


Figure B.9: Decision Tree for Subgrouping of Depression Dataset



Appendix C

HSIRB Approval Letter

WESTERN MICHIGAN UNIVERSITY

Human Subjects Institutional Review Board

Date: October 25, 2016

To: Hyunkeun Cho, Principal Investigator
Nichole Andrews, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 16-10-54

This letter will serve as confirmation that your research project titled "Subgroup Identification and Growth Curve Models for Longitudinal Data" has been **approved** under the **exempt** category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may **only** be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., ***you must request a post approval change to enroll subjects beyond the number stated in your application under "Number of subjects you want to complete the study***)." Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: October 24, 2017

1903 W. Michigan Ave., Kalamazoo, MI 49008-5456
PHONE: (269) 387-8293 FAX: (269) 387-8276
CAMPUS SITE: 251 W. Walwood Hall

References

Arnold, S. F. (1981). The theory of linear models and multivariate analysis. New York: John Wiley and Sons.

Barbosa, M. F. and Goldstein, H. (2000). Discrete response multilevel models for repeated measures: an application to voting intentions data. *Quality and Quantity* **34**, 323–330.

Bonetti, M. and Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5**, 465-481.

Cai, T., Tian, L., Wong. P. H., and Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12**, 270-282.

Cho, H. and Qu, A. (2013). Model selection for correlated data with diverging number of parameters. *Statistica Sinica* **23**, 901–927.

Cho, H., Wang, P., and Qu, A. (2016). Personalized treatment for longitudinal data using unspecified random-effects model. *Statistica Sinica*, in press.

Chou, C-P, Bentler, P. M., and Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent

curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal* **5**, 247-266.

Curran, P. J., Obeidat, K., and Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development* **11**, 121-136.

Diaz F.J. (2016) Measuring the individual benefit of a medical or behavioral treatment using generalized linear mixed-effects models. *Statistics in Medicine* **35**, 4077-4092.

Diaz F.J. and de Leon J. (2013) The mathematics of drug dose individualization should be built with random effects linear models. *Therapeutic Drug Monitoring* **35**, 276-277.

Diaz, F. J., Rivera, T. E., Josiassen, R. C., and Leon, J. (2007). Individualizing drug dosage by using a random intercept model. *Statistics in Medicine* **26**, 2052-2073.

Diaz, F. J., Yeh, H. W., and Leon, J. (2012). Role of statistical random-effects linear models in personalized medicine. *Current Pharmacogenomics and Personalized Medicine* **10**, 22-32.

Foster, J. C., Taylor, J. M. C., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867-2880.

Hamilton, M. (1960). A Rating Scale For Depression. *J. Neurol. Neurosurg. Psychiarty* **23**, 56-62.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93-108.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning. Vol. 6. New York: Springer.

Latra, N., Linuwih, S., Purhadi, and Suhartono (2010). Estimation for multivariate linear mixed models. *International Journal of Basic and Applied Sciences* **10**, 48-53.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 12–22.

Miranda, J., Chung, J. Y., Green, B. L., Krupnick, J., Siddique, J., Revicki, D. A., and Belin, T. (2003). Treating depression in predominantly low-income young minority women. *Journal of the American Medical Association* **290**, 57–65.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113-127.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Senn S. (2011) Individual therapy: new dawn or false dawn. *Drug Information Journal* **35**, 1479-1494.

Senn S. (2016) Mastering variation: variance components and personalised medicine. *Statistics in Medicine* **35**, 966-977.

Shen, J. and He, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *Journal of the American Statistical Association* **110**, 303-312.

Siddique, J., Chung, J. Y., Brown, C. H., and Miranda, J. (2012). Comparative effectiveness of medication versus cognitive behavioral therapy in a randomized controlled trial of low income young minority women with depression. *Journal of Consulting and Clinical Psychology* **80**, 995–1006.

Song, X. and Pepe, M. S. (2004). Evaluating markers for selecting a patientâĂŹs treatment. *Biometrics* **60**, 874-883.

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven tests for longitudinal data in the estimating equation approach. *Journal of the Royal Statistical Society, series B* **71**, 177–190.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Wu, R. F., Zheng, M., and Yu, W. (2016) Subgroup analysis with time-to-event data under a logistic-cox mixture model. *Scandinavian Journal of Statistics* **43**, 863–878.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J.. (2011). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108**, 527-539.

Zhu X, and Qu A. (2016) Individualizing drug dosage with longitudinal data. *Statistics in Medicine* **35**, 4474-4488.