STATISTICAL MODELS FOR IDENTIFICATION OF TREATMENT-SENSITIVE SUBGROUPS BASED ON LONGITUDINAL OUTCOMES IN CLINICAL TRIALS

by

XINYI GE

A thesis submitted to the Department of Mathematics and Statistics in conformity with the requirements for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
September 2021

Copyright © Xinyi Ge, 2021
Abstract

In randomized clinical trials, an average effect of a treatment is often evaluated over all patients enrolled. But in the era of personalized medicine, there is an increasing interest in identification of patients who may benefit from or be sensitive to a specific type of treatment. Various statistical methods have been proposed recently in the literature to help clinical researchers to identify treatment-sensitive subgroups of patients based on baseline covariates or biomarkers and assess the difference in the treatment effect between different subgroups of patients.

Time to an event, such as overall or disease-free survival, is the clinical outcome used to identify treatment-sensitive subgroups in majority of these methods proposed in the literature. Longitudinal outcomes, such as measurements on the quality of life of patients at different time-points, are also important outcomes collected in clinical trials. There are, however, very few statistical approaches available which can be used to identify treatment-sensitive subgroups based on longitudinal outcomes. All current methods proposed in the literature require a subjective definition of subgroups.

This thesis is devoted to the development of statistical methods for identification of treatment-sensitive subgroups based on longitudinal outcomes. Three new statistical models are introduced based on three different clinical scenarios. Specifically, a threshold linear mixed model is introduced when the longitudinal outcomes are
assumed to follow a normal distribution and there is only a single continuous covariate available to define the subgroups. When the longitudinal outcomes may subject to potential floor and ceiling effects but there is still a single continuous covariate available, a threshold mixed-effects Tobit model is introduced. Finally, when there are multiple covariates available, a generalized single-index linear threshold model is introduced by assuming the marginal distribution of the longitudinal outcomes is in an exponential family. Statistical procedures are proposed to make statistical inferences on the parameters in these models, which can be used to identify and assess treatment-sensitive subgroups. All of the proposed models and inference procedures are assessed through simulation studies, as well as applications to the analysis of data from randomized clinical trials.
Co-Authorship

Part of contents in this thesis have been published or submitted for publication:

1. **Chapter 2:**

   Xinyi Ge, Yingwei Peng, and Dongsheng Tu: A Selective Overview of Statistical Methods for Identification of Treatment-sensitive Subset, Submitted to *Proceedings of the Forth ICSA–Canada Chapter Symposium*.

2. **Chapter 3:**


3. **Chapter 5:**

   Xinyi Ge, Yingwei Peng, and Dongsheng Tu: A Generalized Single-Index Linear Threshold Model for Identifying Treatment-Sensitive Subsets Based on Multiple Covariates and Longitudinal Measurements, Submitted.

   In all of the above papers published or submitted, I derived statistical methodology, performed simulation studies and analyses of data, and wrote the first draft. Drs. Yingwei Peng and Dongsheng Tu contributed to the concept developments, discussions of the results, and reviews and revisions of the manuscripts.
Acknowledgments

I would like to extend my warmest appreciation and gratitude to all those who have guided and supported me throughout the last four years. Because of you, the last four years at Queen’s have been a wonderful journey: a journey full of joys, challenges, and opportunities.

First and foremost, I would like to send my deepest appreciation to my supervisors, Professors Yingwei Peng and Dongsheng Tu, who have given me invaluable opportunities, insightful guidance, brilliant advice, continual encouragement, and unwavering support, not only throughout my studies, but also in my daily life. I thank my supervisors for their flexibility, kindness, and especially for their patience in tolerating my shortcomings. I consider myself lucky to have benefited from Professor Peng’s experience in computation and Professor Tu’s experience in real-data applications. The simulating conversations I have had with them have given me a focus for my future research and my future career.

Secondly, I would like to acknowledge the Department of Mathematics and Statistics for its financial support. Sincere thanks to my PhD and examination committee members, Professors Wenyu Jiang, Yanglei Song, Wenqing He, Zihang Lu, who shared their expertise, advice, and time to help me improve this thesis. Special thanks to Professor Devon Lin for her unending help and advice. Many thanks to Professors
Wenyu Jiang, Bingshu Chen, and Glen Takahara for their interesting lectures, which have inspired me to pursue biostatistics research.

I also wish to thank my fellow students and friends (especially Heye, Lina, Jiani, Blair, Mengqian, and Fangge) for making my time at Queen’s enjoyable and unforgettable. Special thanks to lifelong friends, Spencer, Ian, and Theresa, for their generosity, encouragement, and support. They helped me regain my confidence during a difficult time. Also, thanks to Eric whose company made my days easier and happier during the COVID-19 pandemic.

Finally, I give profound thanks to my aunt Xiaohua and uncle Naijie without whose example I might never have started on, and more importantly might never have completed my PhD. Deep thanks also to my beloved parents, Yufu Ge and Yaping Gu, for their unconditional love, endless encouragement, and warm support during my time away from home. They showed patience and tolerance whenever I complained. They believed in me whenever I doubted myself, and they gave me support whenever I made a choice. I hope I have made them proud.
Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work and that any ideas or quotations from the work of other people, published or unpublished, are fully acknowledged in accordance with the standard referencing practices of the discipline.
# Contents

**Abstract**  
i  
**Co-Authorship**  
iii  
**Acknowledgments**  
iv  
**Statement of Originality**  
vi  
**Contents**  
vii  
**List of Tables**  
ix  
**List of Figures**  
xi  

## Chapter 1: Introduction  
1  
1.1 Quality of Life Measurements as an Example of Longitudinal Outcomes in Clinical Trials  
1  
1.2 Subgroup Analysis in Clinical Trials  
2  
1.3 Objective and Organization of This Thesis  
5  

## Chapter 2: Background and Literature Review  
8  
2.1 Statistical Methods for Subgroup Identification Based on Survival Data  
8  
2.1.1 Biomarker-Adaptive Threshold Design  
9  
2.1.2 A Hierarchical Bayesian Method  
11  
2.1.3 A Single-Index Threshold Cox Model  
13  
2.1.4 An Interaction Tree  
16  
2.2 Statistical Methods for Subgroup Identification Based on Longitudinal Outcomes  
19  
2.2.1 A Procedure Based on Multilevel Models  
20  
2.2.2 A Prediction Model Approach  
22  
2.2.3 A Procedure Based on Structured Mixed-Effects Models  
25
<table>
<thead>
<tr>
<th>Chapter 3: A Threshold Linear Mixed Model</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>28</td>
</tr>
<tr>
<td>3.2 Methodology</td>
<td>32</td>
</tr>
<tr>
<td>3.2.1 The Model and Notations</td>
<td>32</td>
</tr>
<tr>
<td>3.2.2 Inferences of Model Parameters</td>
<td>34</td>
</tr>
<tr>
<td>3.3 Simulation Studies</td>
<td>38</td>
</tr>
<tr>
<td>3.4 Applications</td>
<td>44</td>
</tr>
<tr>
<td>3.4.1 Analysis of Data from CO.17</td>
<td>44</td>
</tr>
<tr>
<td>3.4.2 Analysis of data from ACTG 193A</td>
<td>48</td>
</tr>
<tr>
<td>3.5 Conclusions and Discussions</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4: A Threshold Mixed-Effects Tobit Model</th>
<th>53</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Methodology</td>
<td>55</td>
</tr>
<tr>
<td>4.2.1 The Model and Notations</td>
<td>55</td>
</tr>
<tr>
<td>4.2.2 Estimation of Model Parameters</td>
<td>57</td>
</tr>
<tr>
<td>4.2.3 Random Weighting Estimates for the Variances of the Parameter Estimators</td>
<td>59</td>
</tr>
<tr>
<td>4.3 Simulation Studies</td>
<td>60</td>
</tr>
<tr>
<td>4.4 An Application</td>
<td>66</td>
</tr>
<tr>
<td>4.5 Conclusions and Discussions</td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 5: A Generalized Single-Index Linear Threshold Model</th>
<th>72</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Introduction</td>
<td>72</td>
</tr>
<tr>
<td>5.2 Methodology</td>
<td>76</td>
</tr>
<tr>
<td>5.2.1 The Model and Notations</td>
<td>76</td>
</tr>
<tr>
<td>5.2.2 Inference of Model Parameters</td>
<td>78</td>
</tr>
<tr>
<td>5.3 Simulation Studies</td>
<td>84</td>
</tr>
<tr>
<td>5.4 An Application</td>
<td>90</td>
</tr>
<tr>
<td>5.5 Conclusions and Discussions</td>
<td>94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 6: Summary and Future Extensions</th>
<th>97</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>102</td>
</tr>
<tr>
<td>Appendix A: Generating Missing at Random Longitudinal Outcomes</td>
<td>116</td>
</tr>
<tr>
<td>Appendix B: Generating Correlated Binary Data</td>
<td>118</td>
</tr>
<tr>
<td>Appendix C: Gauss-Hermite Quadrature Approximation for the Tobit Model</td>
<td>121</td>
</tr>
</tbody>
</table>
## List of Tables

3.1 Results for simulation studies over 500 replications with optimal $h$ for sample size $n = 400, 800, 1600$. The numbers inside the brackets next to the parameter estimate and the sample size are the true value of the parameter and the optimal value of $h$, respectively. ........................................... 40

3.2 Results for simulation studies over 500 replications with $h = 0.1, 0.07, 0.04, 0.01, 0.001, 0.0001$ for sample size $n = 800$. The number inside the brackets next to the parameter estimate is the true value of the parameter. ................................................................. 41

3.3 Comparison between empirical standard deviation and Hessian standard error from simulation studies over 500 replications with different $h$ values under subject sample size $n = 800$. ........................................... 42

3.4 Comparison between empirical standard deviation and Hessian standard error from simulation studies over 500 replications with optimal $h$ for sample size $n = 400, n = 800, n = 1600$. The numbers inside the brackets next to the parameters and the sample size are the true value of the parameters and the optimal value of $h$, respectively. ................. 43

3.5 Actual Coverage rate of confidence interval from simulation studies over 500 replications with different $h$’s values for sample size $n = 800$. 43
3.6 Results for simulation studies over 500 replications when the observations may be missing at random. .................................................. 44
3.7 Subgroup analysis of quality of life data in CO.17 ............................. 47
3.8 Subgroup analysis of CD4 count data in 193A .................................... 49

4.1 Simulation results on the relative bias and RMSE of the parameter estimators when sample size is 100 by using the proposed method (labeled as P) and the threshold linear mixed model (labeled as T). .......................... 63
4.2 Simulation results on the relative bias and RMSE of the proposed parameter estimators when sample size is respectively 200 and 300. .......................... 64
4.3 Comparison between the standard deviation of proposed random weighting estimators for the variance of parameter estimators (SDE) and empirical standard deviation (ESD) of the parameter estimators when the sample size N=100 .................................................. 65
4.4 Results for simulation studies over 1000 replications when the observations may be missing at random. .................................................. 66
4.5 Results of analysis for the physical subscores in CO.17 trial .................. 69

5.1 Simulation results for continuous longitudinal measurements under different values of $h$ ........................................................................... 87
5.2 Simulation results for binary longitudinal measurements under different values of $h$ ........................................................................... 88
5.3 Results for simulation studies over 500 replications when the observations may be missing at random .................................................. 89
5.4 Analysis of QoL data in PA.3 trial ......................................................... 92
List of Figures

3.1 Mean of QoL score (in log-transformation) over time in global health-status for older patients [cetuximab vs BSC] and younger patients [cetuximab vs BSC] ......................................................... 48

4.1 Plot of physical subscores over time with lowess curves in two treatment groups superimposed .......................................................... 68

4.2 Averages of observed physical subscores over time by treatment group for older patients [left panel] and younger patients [right panel] with dots representing the averages of observed physical subscores at each timepoint and lines the lowess curves ............................................ 70

5.1 QoL scores over time with Lowess curves by treatment arm (blue lines for ERLGEM and red lines for GEM) in treatment sensitive and insensitive subgroups, respectively ........................................ 93

5.2 Log-odds of QoL not worsened at each time-point with loess curves by treatment arm (blue lines for ERLGEM and red lines for GEM) in treatment sensitive and insensitive subgroups, respectively ........ 95
Chapter 1

Introduction

1.1 Quality of Life Measurements as an Example of Longitudinal Outcomes in Clinical Trials

Longitudinal outcomes, which are repeated observations measured on the same subjects at different points in time, are often collected in clinical trials and other medical studies. In the following, I use quality of life measurements in cancer clinical trials, which motivated my researches summarized in this thesis, as an example for the illustration on the practical importance of these outcomes.

In cancer clinical trials, although the treatment effect is traditionally evaluated by relatively objective endpoints such as tumor response, relapse-free survival, or overall survival, it is argued that these endpoints may not provide adequate information in understanding of the full effect of a treatment. Recently, the importance of evaluations based on more subjective endpoints, such as patient reported quality of life (QoL), has become increasingly recognized in cancer clinical trials. Studies of quality of life have several benefits (Coons, 2007): Firstly, improvement in quality of life measurements in cancer patients may be associated with improvement in more
objective endpoints and it has been demonstrated that QoL measurements are of prognostic value for overall survival in several disease settings (Montazeri et al., 2003; Montazeri, 2008; Trask et al., 2009; Quinten et al., 2009; Grande et al., 2009). Secondly, when the prolonging of survival time is limited by currently available therapy, patients’ QoL may be important to determine the values of a new therapy (Richards and Ramirez, 1997). For example, although there is no current therapy to improve survive for patients with advanced symptomatic pancreatic cancer, gemcitabine has been approved by regulatory agencies as a standard therapy based on data from a clinical trial which used a composite of measurements of pain and other disease-related symptoms reported by patients as the primary endpoint of the trial (Burris et al., 1997). Thirdly, these endpoints can help patients to make the treatment decisions by providing detailed information on the side-effects of the treatment (Blazeby et al., 2001). Also these endpoints can help future patients understand the consequences of their illness and treatment (Bezjak et al., 2006).

In practice, these patient reported outcomes are usually assessed at several time-points before, during and after patients have received the treatment by asking patients to answer questions from one or more questionnaires. Therefore, they can be considered as longitudinal outcomes.

1.2 Subgroup Analysis in Clinical Trials

For many diseases, such as cancer, it is often difficult to find a treatment that would benefit all patents. There may be a subgroup of patients, defined by individual characteristics, for example age, gender, blood test results, or gene expression levels, who may be more sensitive to a specific treatment and have a larger treatment effect.
Conversely, if a treatment is costly or has potential negative side effects, there is an interest to look for subgroups for which the treatment has less side effects. Therefore, it is important to identify subgroups of patients who are sensitive to a specific treatment, termed as the treatment-sensitive subgroups, based on data from clinical trials or other medical studies. For example, in a recent secondary analysis of data from CO.17 and CO.20 trials conducted by the Canadian Cancer Trials Group (CCTG), the investigators were interested to know whether older patients with advanced colorectal cancer treated by respectively cetuximab alone or cetuximab plus brivanib had a less benefit, in comparison with younger patients, in terms of various outcomes including quality of life (Wells et al., 2008).

Subgroup analysis, which includes identification of subgroups, estimation of treatment effects in different subgroups and tests for the significance of the differences in the treatment effects in these subgroups, is a main statistical tool to assess the heterogeneity in treatment effects in subgroups defined by certain characteristics of patients. For example, in CO.17 and CO.20 analyses mentioned above, patients were divided into two age subgroups based on whether their age was 70 years or older and differential treatment effects in these two age groups were assessed through a test of interaction between subgroup and treatment. However, it is unclear whether 70 years is an optimal cutpoint to define the two age subgroups when assessing the heterogeneity of treatment effects by age. This issue arises in many studies where the variable to define subgroups is continuous but a pre-specified cutpoint is not available from previous studies or clinical experience, and a statistical approach is often needed to determine the optimal cutpoints based on data.

When the outcomes used for the subgroup analyses are times to an event, such as
progression-free or overall survivals, several approaches have been proposed for the
determination of cutpoints in the definition of subgroups. For example, Jiang et al. (2007)
proposed a biomarker-adaptive threshold design, which combines a test for
overall treatment effect in all patients with the determination and validation of a cut
point for a biomarker which is used to define a treatment sensitive subgroup. Chen
et al. (2014) developed a hierarchical Bayesian procedure to estimate simultaneously
the interaction parameter and cut point in a threshold Cox model. He et al. (2018)
adopted a single-index threshold Cox proportional hazards model, which includes a
smoothly clipped absolute deviation (SCAD) penalty function, to select and linearly
combine multiple biomarkers in identification of treatment-sensitive subgroups. Su
et al. (2008) proposed an interaction tree, which recursively partitioned the data into
two subgroups based on the greatest interaction with treatment, to obtain treatment-
sensitive subgroups.

From our knowledge, however, no statistical procedure to determine the cut point
of a continuous covariate is available in the literature for the definition of treatment-
sensitive subgroups when the outcomes are longitudinal measurements. Moineddin
et al. (2008) used multi-level models including subject-specific random effects to iden-
tify subgroups with differential treatment effects of gabapentin versus placebo on
longitudinal measurements of hot flashes based on the baseline score in a double-
blind randomized controlled trial for treatment of hot flashes in women who enter
menopause naturally, but a median was used as the cutpoint in defining subgroups.
Andrews (2017) considered a random effects linear model for longitudinal outcomes
to determine whether a patient had a positive response to the treatment and su-
pervised learning algorithms were proposed to estimate a predictive function for the
positive response, but 0.5 was used as an ad-hoc cut point for the predictive function to assign patients into subgroups. Recently, Shen and Qu (2020) proposed a mixed-effects model where the treatment effect was modeled as a random effect from a two-component mixture model, and subgroup membership of the mixture model was incorporated using a logistic regression model with respect to some covariates. They adopted an EM-type algorithm to obtain parameter estimations and ranked scores from the defined score functions to determine the subgroups, but they could not give a specific cut point of subgrouping.

1.3 Objective and Organization of This Thesis

As discussed above, longitudinal outcomes are important endpoints to assess treatment effects in clinical trials and, as with other types of endpoints, there may be a subgroup of patients who would benefit from a specific treatment based on this type of endpoints. There is, however, no objective statistical method which can be used to identify a treatment-sensitive subgroup based on longitudinal outcomes. The objective of this thesis is to develop new statistical methods for the identification of subgroups of patients who might benefit more from or be sensitive to a specific treatment based on their baseline covariates, such as age, blood test results, or biomarkers and longitudinal outcomes collected in a clinical trial. Three different statistical models are introduced in this thesis based on three different clinical scenarios and statistical procedures are developed for the statistical inferences of the parameters in these models which can be used to define the statistical rules for the definition of treatment-sensitive subgroups.

Specifically, a threshold linear mixed model is introduced for the first scenario
where the longitudinal outcomes are assumed to be continuous and normally distributed and there is only one continuous baseline covariate available for the identification of treatment-sensitive subgroup. A smoothing maximum likelihood method is proposed to obtain the estimates of the parameters in the model and associated variance estimators.

In the second clinical scenario, there is still one continuous baseline covariate available but some of potential continuous longitudinal outcomes may not be observed because of so-called floor or ceiling effect when a patient reported a maximum or minimum score for some questions in a questionnaire. A threshold mixed effects Tobit model is proposed in the forth chapter of this thesis to take into account of potential truncation on the longitudinal outcomes. The Gauss–Hermite quadrature method is used to approximate the smoothed likelihood function to obtain the estimates of the parameters. A random weighting method is adopted to obtain the variance estimation of the parameter estimators.

In some clinical trials, multiple covariates may be available and combination of them may be more useful in the identification of treatment-sensitive subgroups. The longitudinal outcomes may also not always be continuous. A generalized single-index linear threshold model is proposed in the last part of thesis to identify the patients based on a linear combination of multiple covariates and exponential family distributed longitudinal measurements. A procedure based on smoothed generalized estimating equations is derived for the estimation of parameters in the model and associated variance estimators.

For all of these new models, the proposed inference approaches are evaluated by simulation studies, as well as applications to real data from clinical trials.
1.3. OBJECTIVE AND ORGANIZATION OF THIS THESIS

The remainder of this thesis is organized as follows. In Chapter 2, I review some statistical methods developed previously on subgroup analyses based on respectively survival data and longitudinal outcomes. Chapter 3 presents a threshold linear mixed model for continuous longitudinal outcomes and a single continuous covariate. A smoothed likelihood is derived to obtain estimations of model parameters and associated variance estimators. In Chapter 4, I propose a threshold mixed-effects Tobit model based on a single continuous covariate, as well as longitudinal measurements which may have floor and ceiling effects. The smoothed likelihood function is defined and approximated by the Gauss–Hermite quadrature to obtain estimates of parameters in the model. Chapter 5 presents a generalized single-index linear threshold model based on a linear combination of covariates and longitudinal measurements. Smoothed generalized estimating equations are proposed to estimate the parameters in the model. A summary of the research results in this thesis and discussion of future work is presented in Chapter 6. Some technical details are presented in Appendixes.
Chapter 2

Background and Literature Review

In this chapter, I review some statistical models developed in the literature for subgroup analyses based on respectively survival data and longitudinal outcomes.

2.1 Statistical Methods for Subgroup Identification Based on Survival Data

When the outcome in a clinical trial is the time to an event, such as overall survival or progression-free survival, a few approaches have been proposed to identify treatment-sensitive subgroups. Before giving detailed descriptions on these approaches, I introduce some notations and a commonly used statistical model for survival data.

Denote $F_i$ and $C_i$ respectively as the potential failure and censoring times of the subject $i$ and assume that $F_i$ and $C_i$ are independent. The observed survival time $T_i$ and survival status indicator $\delta_i$ are defined respectively as

$$
\begin{cases}
T_i = \min(F_i, C_i) \\
\delta_i = I(F_i < C_i)
\end{cases}
$$

(2.1)
Let $h(t)$ be the hazard function of failure time $F_i$. In the survival analysis, Cox proportional hazards model (Cox, 1972, 1975) is usually used to model the relationship between $h(t)$ and a vector of covariates $X$ as follows:

$$h(t) = h_0(t)g(X, \beta),$$

where $g(\cdot)$ is a given link function and $h_0(t)$ is an unknown baseline hazard function.

In the next three subsections, I introduce three methods which were developed for the identification of treatment-sensitive subgroups of patients.

### 2.1.1 Biomarker-Adaptive Threshold Design

Jiang et al. (2007) proposed a biomarker-adaptive threshold design, which combines a test for overall treatment effect of all patients with the determination and validation for a cutpoint of a biomarker, that is used to define a treatment-sensitive subgroup.

Consider the following threshold Cox proportional hazards model:

$$\log h(t) = \log h_0(t) + \beta_1 Z_{1i} + \beta_2 I(Z_{2i} > c) + \beta_3 Z_{1i} I(Z_{2i} > c),$$

(2.2)

where $\beta_1$ is the main treatment effect, $\beta_2$ is the main biomarker effect, $\beta_3$ is the treatment-by-biomarker interaction effect, $Z_{1i}$ is the treatment indicator with $Z_{1i} = 1$ if patient $i$ is assigned into the treatment group or $Z_{1i} = 0$ if patient $i$ is assigned into the control group, $Z_{2i}$ is the value of a continuous biomarker ($i = 1, 2, \ldots, n$), and $c$ is an unknown threshold parameter that defines the treatment-sensitive subgroup. Without loss of generality, $c$ and $Z_{2i}$ are assumed to take values in the open interval $(0, 1)$. 
The first step in the procedure is to assess the effect of treatment over all patients by taking $\beta_2 = \beta_3 = 0$ in model (2.2). And the likelihood ratio test is used to test the null hypothesis that $\beta_1 = 0$ in the following reduced model:

$$\log h(t) = \log h_0(t) + \beta_1 Z_{1i}. \quad (2.3)$$

If the test rejects the null hypothesis of the reduced model (2.3), then the procedure stops and one can conclude that the treatment will benefit all patients. Otherwise, the procedure will continue to detect whether there is a subgroup of treatment-sensitive patients by testing the null hypothesis of $\beta_3 = 0$ in the full model (2.2). If the test rejects the null hypothesis, then it demonstrates that a subgroup of treatment-sensitive patients does exist.

The second step of the procedure is to estimate the unknown threshold parameter $c$. First, if the test fails to reject the null hypothesis $\beta_1 = 0$ in the model (2.3), then a test statistic is defined to test the hypothesis $\beta_3 = 0$ in the model (2.2). Specifically, fixing $\beta_1 = 0$ in the model (2.2), a log likelihood ratio statistic $S(c)$ for testing hypothesis $\beta_3 = 0$ can be obtained for each candidate biomarker threshold $c$ in the range $(0, 1)$. Maximizing $S(c)$ over a range of possible cut point values will give a test statistic for testing null hypothesis $\beta_3 = 0$ with $c$ unspecified. In order to obtain a reasonable power, a test statistic $T$ is defined as $\max((S(0) + R), \max_{0 < c < 1} S(c))$, where $R$ is a positive constant which was suggested to be 2.2 by Jiang et al. (2007). The p-value of this test statistic can be calculated from a resampling-based approach by randomly permutating treatment labels. If the test rejects the null hypothesis
\( \beta_3 = 0 \), the optimal threshold \( c_0 \) can be estimated as

\[
\hat{c}_0 = \arg \max_{c_0} l(c_0),
\]

where \( l(c_0) \) is the partial log likelihood function based on model (2.2):

\[
l(c_0) = \max_{\beta_1, \beta_2, \beta_3} l(\beta_1, \beta_2, \beta_3, c_0).
\]

Therefore, the treatment-sensitive subgroup can be defined by \( I(Z_{i2} > \hat{c}_0) \), that is, the patients with the biomarker over \( \hat{c}_0 \) can be defined as the treatment-sensitive subgroup.

### 2.1.2 A Hierarchical Bayesian Method

Chen et al. (2014) proposed a hierarchical Bayesian method to estimate all unknown parameters, including the threshold \( c \), in model (2.2) simultaneously without assumption \( \beta_1 = 0 \).

For simplicity of presentation, denote \([ Z_{i1}, I(Z_{i2} > c), Z_{i1}I(Z_{2i} > c) ] \)' as \( Z_i(c) \) and \([ \beta_1, \beta_2, \beta_3 ] \)' as \( \beta \). With these notations, model (2.2) can be rewritten as

\[
h(t) = h_0(t) \exp\{ Z_i'(c)\beta \}. \tag{2.4}
\]

Chen et al. (2014) assumed that the threshold parameter \( c \) has a prior Beta distribution Beta(2, q) for a given hyper-parameter \( q > 1 \), which can be written as

\[
p_1(c|q) \propto q(q + 1)c(1 - c)^{q-1}.
\]
This prior is flexible enough to accommodate any prior distribution in the family with its mode taking any specific value in the interval $(0, 1)$. In order to give a specific prior distribution of $c$, instead of taking an arbitrary value for $q$, it is considered that $q$ has a hyper-prior distribution with the following density function form

$$p_2(q) \propto \frac{q-1}{q(q+1)}, \quad q > 1.$$ 

At the same time, $\beta$ is assumed to have a uniform improper prior distribution $p(\beta) \propto 1$. For every given $0 < c < 1$, the corresponding partial likelihood function of $\beta$ in model (2.4) is given by

$$p_3(\beta | c) = \prod_{i=1}^{n} \left[ \frac{\exp\{Z_i'(c)\beta\}}{\sum_{j \in R(T_i)} \exp\{Z_j'(c)\beta\}} \right]^{\delta_i},$$

where the risk set $R(T_i)$ is the index set of patients who are at risk of experiencing an event at time $T_i$. Consequently, given the observed data, the joint posterior distribution of $\beta, c, q$ can be written as

$$p(\beta, c, q | \text{data}) \propto p_1(c | q)p_2(q)p_3(\beta | c)$$

$$= \prod_{i=1}^{n} \left[ \frac{\exp\{Z_i'(c)\beta\}}{\sum_{j \in R(T_i)} \exp\{Z_j'(c)\beta\}} \right]^{\delta_i} c(1-c)^{q-1}(q-1).$$
Therefore, the marginal posterior distributions of $\beta$ and $c$ can be calculated respectively as

$$p(\beta) = \int_{c,q} p(\beta, c, q|\text{data}) dc dq,$$

$$p(c) = \int_{\beta,q} p(\beta, c, q|\text{data}) d\beta dq.$$

Statistical inferences, such as point estimation, confidence interval and hypothesis testing, of the threshold parameter $c$ and the regression coefficient $\beta$ can be obtained based on these marginal distributions. After obtaining the estimate of the threshold $c$, the treatment-sensitive subgroup consequently can be defined.

2.1.3 A Single-Index Threshold Cox Model

Sometimes, it can be difficult in clinical trials to identify a treatment-sensitive subgroup based on a single biomarker. For example, in a randomized control trial PA.3 conducted by NCIC Clinical Trials Group, 35 key proteins were selected from a global genetic analysis of pancreatic cancers with the purpose of identifying a subgroup of patients with locally advanced or metastatic pancreatic cancer who will be sensitive to the treatment of erlotinib in addition to gemcitabine (Shultz et al., 2016). However, no significant interaction was found between the treatment and any of these biomarkers, which implies that it is impossible to identify a treatment-sensitive subgroup based on a single biomarker. In fact, He et al. (2018) found that a combination of some of these biomarkers (CA 19-9 and Axl) had the potential to define a treatment-sensitive subgroup of patients with pancreatic cancer. It is more complicated to identify a treatment-sensitive subgroup based on multiple biomarkers, compared to the cases where there is only a single biomarker.

Several approaches have been proposed in subgroup analysis based on multiple
biomarkers. He et al. (2018) proposed a single-index threshold Cox proportional hazards model to identify a treatment-sensitive subgroup for each treatment using multiple biomarkers based on linear combinations of the multiple biomarkers. In this subsection, I will briefly review this approach.

Let \( X = (x_1, x_2, \cdots, x_d)' \) be a \( d \)-dimensional vector of exposure variables including treatment group indicator for subjects, and \( Z = (z_1, z_2, \cdots, z_p)' \) be a \( p \)-dimensional vector of observed biomarkers. Define the indicator function \( I(Z'\gamma_j > c_j) \) be the subgrouping indicator for the \( j \)-th treatment, where \( \gamma_j \) is a \( p \)-dimensional vector used to combine biomarkers linearly and \( c_j \) is the unknown threshold parameter. The proposed model is written as

\[
h(t) = h_0(t) \exp \left\{ \beta'X + \sum_{j=1}^{d} \eta_j I(Z'\gamma_j > c_j) + \sum_{j=1}^{d} \alpha_j x_j I(Z'\gamma_j > c_j) \right\},
\]

where \( h(t), h_0(t) \) and \( \beta \) are the same defined in last section. Parameters \( \eta = (\eta_1, \eta_2, \cdots, \eta_d)' \) and \( \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)' \) model main effects of biomarkers and the treatment-by-biomarker interactions, respectively. A significant treatment-by-biomarker interaction \( \alpha_j \) implies that the effect of the \( j \)-th treatment varies across subgroups, and the treatment-sensitive subgroup for this treatment can be determined.

To obtain parameter estimates in the model, a maximum penalized smoothing partial likelihood method has been proposed. First, assume that data are available from \( n \) independent subjects. Denote \( \Gamma = (\gamma_1, \gamma_2, \cdots, \gamma_d)' \), \( c = (c_1, c_2, \cdots, c_d)' \), and \( \theta = (\beta', \eta', \alpha', c', \Gamma')' \). Then the partial likelihood of the parameters in model (2.5)
can be written as

\[
L(\theta) = \prod_{i=1}^{n} \left[ \frac{\exp \left\{ \beta' X_i + \sum_{j=1}^{d} \eta_j I(Z'_i \gamma_j > c_j) + \sum_{j=1}^{d} \alpha_j x_{ij} I(Z'_i \gamma_j > c_j) \right\}^{d_i}}{\sum_{k \in R(T_i)} \exp \left\{ \beta' X_k + \sum_{j=1}^{d} \eta_j I(Z'_k \gamma_j > c_j) + \sum_{j=1}^{d} \alpha_j x_{kj} I(Z'_k \gamma_j > c_j) \right\}^{d_i}} \delta_i \right].
\]

Since the partial likelihood function is not continuous at some parameters, the estimator of \( \theta \) cannot be obtained by directly maximizing the partial likelihood function \( L(\theta) \). A local distribution function \( \Phi((Z'\gamma_j - c_j)/h) \) is adopted to obtain a smoothing approximation to the indicator function \( I(Z'\gamma_j > c_j) \), where \( \Phi \) is the distribution function of the standard normal variable, and the bandwidth \( h \) converges to zero as the sample size increases. With this approximation, the smoothed partial likelihood (SPL) function can be defined as

\[
S(\theta) = \prod_{i=1}^{n} \left[ \frac{\exp \left\{ \beta' X_i + \sum_{j=1}^{d} \eta_j \Phi((Z'_i \gamma_j - c_j)/h) + \sum_{j=1}^{d} \alpha_j x_{ij} \Phi((Z'_i \gamma_j - c_j)/h) \right\}^{d_i}}{\sum_{k \in R(T_i)} \exp \left\{ \beta' X_k + \sum_{j=1}^{d} \eta_j \Phi((Z'_k \gamma_j - c_j)/h) + \sum_{j=1}^{d} \alpha_j x_{kj} \Phi((Z'_k \gamma_j - c_j)/h) \right\}^{d_i}} \delta_i \right].
\]

Because a large number of covariates may be available in practice but usually only a few of them might be relevant in the definition of treatment-sensitive subgroups, a penalty function is added to the SPL function for efficiently selecting relevant
biomarkers from large amount of biomarkers. Specifically, the smoothly clipped absolute deviation (SCAD) penalty function is employed, and therefore the penalized smoothed partial likelihood (PSPL) function can be defined as

\[
L_n(\theta) = \log \{ S(\theta) \} - n \sum_{j=1}^{d} \sum_{k=1}^{p} P_{\lambda}(|\gamma_{jk}|),
\]

where \( \lambda_{jk} \) is the k-th component of \( \gamma_j \) and \( P_{\lambda}(\cdot) \) is SCAD penalty function with a regularization parameter \( \lambda \). By maximizing PSPL function \( (2.8) \), the estimates of \( \theta \) can be obtained. Therefore, the corresponding treatment-sensitive subgroups for the treatment \( j \) can be determined by estimated \( c_j \).

### 2.1.4 An Interaction Tree

Su et al. (2008), Foster et al. (2011), and Lipkovich et al. (2011) utilized tree-based approaches to categorize subgroups with an enhanced treatment effect. For example, Su et al. (2008) constructed an interaction tree (denoted by \( T \)) to assess the interaction between the treatment and baseline covariates for survival data. In this subsection, the procedure of this method is briefly reviewed.

The interaction tree is constructed by the following three steps: 1) growing a large initial tree; 2) pruning the initial tree; 3) selecting an optimal tree size. The first step is to grow a large initial tree. Let \( s \) be a single binary split of the data in the tree construction based on a covariate \( x \). If \( x \) is continuous, then the split \( s \) is induced by whether or not \( x \leq c \), where the threshold \( c \) can be any constant. However, in practice the threshold \( c \) is chosen as one of the observed values of \( x \). If \( x \) is ordinal, the split \( s \) can be induced by the similar procedure. If \( x \) is a categorical variable with categories \( C = \{ c_1, \cdots, c_r \} \), then the split can be induced by the form of \( x \in A \) with \( A \subset C \). In
order to reduce the computational burden, the treatment effect within each category is often estimated at first, and then the categories of \( x \) are reordered according to their treatment effects. As a result, splitting on \( x \) can then be induced by treating \( x \) as an ordinal variable. Next, the best split is selected among all possible splits by comparing their differences in the treatment effects. The best split should show the greatest difference in the treatment effect between its two child nodes. Specifically, the splitting selection approach proposed in the interaction tree is to choose the split to maximize a statistic for testing \( H_0: \beta_3 = 0 \) in the Cox model:

\[
h(t|\mathbf{X}_i) = h_0(t) \exp\{\beta_1 Z_i + \beta_2 I(s) + \beta_3 Z_i I(s)\}, \tag{2.9}\]

where \( I(s) = I(x \leq c) \) if \( x \) is continuous or ordinal, or \( I(s) = I(x \in A) \) if \( x \) is categorical. The following partial likelihood ratio test (PLRT) statistic is adopted to test the above hypothesis:

\[
G(s) = -2(l_2 - l_1), \tag{2.10}\]

where \( l_2 \) is the maximized partial likelihood \((\text{Cox}, 1975)\) of model \((2.9)\) and \( l_1 \) is the maximized partial likelihood of the reduced model under \( H_0: h(t|\mathbf{X}_i) = h_0(t) \exp\{\beta_1 Z_i + \beta_2 I(s)\}. \tag{2.11}\)

The best split \( s^* \) can be determined by \( G(s^*) = \max_s G(s) \). After choosing the best split, the data can be divided into two subgroups and therefore the tree grows two child nodes. The same procedure is then implemented to split both child nodes based
on different variables. With recursively doing so, a large initial tree is constructed, denoted by $T_0$.

After growing the large initial tree, the tree needs to be pruned until it has an appropriate size. Following the split-complexity pruning algorithm for survival tree (LeBlanc and Crowley, 1993), An interaction-complexity measure $G_\lambda(T)$ is utilized in the pruning algorithm:

$$G_\lambda(T) = G(T) - \lambda \cdot |T - \tilde{T}|,$$

(2.12)

where $T$ is any given tree, $\tilde{T}$ represents the set of all terminal nodes of $T$, $|T - \tilde{T}|$ denotes the number of all internal nodes of $T$, and $\lambda \geq 0$ is the complexity parameter. $G(T) = \sum_{h \in T - \tilde{T}} G(h)$, where $h$ is the node (including its split to its child nodes), and $G(h)$ is the splitting statistic defined in (2.10). As such, $G_\lambda(T)$ evaluates the overall goodness-of-interaction of $T$ and $\lambda$ is a penalty parameter of each added node. The pruning procedure is started from the initial tree $T_0$ and for each internal node $h$ of $T_0$, one can have

$$g(h) = \frac{G(T_h)}{|T_h - \tilde{T}_h|},$$

where $T_h$ is the branch of tree with $h$ as its root, $\tilde{T}_h$ represents the set of all terminal nodes of $T_h$, and $|T_h - \tilde{T}_h|$ denotes the number of all internal nodes of $T_h$. By minimizing $g(h)$ over all the internal nodes of $T_0$, the weakest link (or the most ineffective split) $h^*$ can be determined. Denote $T_1$ as the subtree after pruning off the branch $T_{h^*}$ from $T_0$ and apply the same pruning procedure to the subtree $T_1$. After the above process is repeated recursively, a nested sequence of subtrees can be defined as
2.2. STATISTICAL METHODS ON LONGITUDINAL OUTCOMES

\[ \mathcal{T}_M \prec \cdots \prec \mathcal{T}_m \prec \mathcal{T}_{m-1} \cdots \prec \mathcal{T}_1 \prec \mathcal{T}_0, \]
where \( \mathcal{T}_M \) is a tree only having the root node and \( \prec \) means “a subtree of”.

After the pruning procedure is finished, the last step of the method is to select the best size of the tree. Since the interaction-complexity measure \( G_\lambda(\mathcal{T}) \) defined in (2.12) can evaluate the overall goodness-of-interaction of the tree, the optimally sized tree \( \mathcal{T}^* \) can be determined by maximizing \( G_\lambda(\mathcal{T}) \) as following:

\[
G_\lambda(\mathcal{T}^*) = \max_{m=0, \ldots, M} \{ G(\mathcal{T}_m) - \lambda \cdot |\mathcal{T}_m - \tilde{\mathcal{T}}_m| \},
\]

where the complexity parameter \( \lambda \) can be pre-specified within the range \( 2 \leq \lambda \leq 4 \) (LeBlanc and Crowley, 1993). After the optimally sized tree is determined, the treatment-sensitive subgroups can be identified based on the terminal nodes of the tree \( \mathcal{T}^* \).

2.2 Statistical Methods for Subgroup Identification Based on Longitudinal Outcomes

Longitudinal outcomes are repeated observations measured on the same subject at different time points. Multilevel or hierarchical models are often used for longitudinal outcomes, as these models incorporate the variation at different levels of the hierarchy into analysis. This class of models includes multilevel model, linear mixed model, random-effects ANOVA model, generalized estimating equations (GEE), mixed-effect mixture model, etc. Moineddin et al. (2008) utilized a multilevel model to identify subgroups with an enhanced treatment, but the cut point of each continuous biomarker used to define the subgroups is given by its median. Andrews (2017) considered a random effects linear mixed model for longitudinal outcomes to determine
whether a patient had a positive response to a treatment and supervised learning algorithms were proposed to estimate a predictive function for the positive response but 0.5 was used as an ad-hoc cutpoint for the predictive function to assign patients into subgroups. A structured mixed-effects models is proposed by Shen and Qu (2020) to simultaneously model the subgroup distribution and subgroup membership. The subgroup can be determined by ranking the estimated score function of the subgroup membership, but they didn’t give a specific decision rule to define the subgroup. In this section, these three methods are reviewed in details.

2.2.1 A Procedure Based on Multilevel Models

To establish notations, let $y_{ij}$ be the longitudinal measurement of subject $i$ ($i = 1, 2, \cdots, N$) at $j$-th observation time $t_{ij}$ ($j = 1, 2, \cdots, n_i$). The observation times are usually called as level-1 units in a multilevel model, while subjects are called as the level-2 units. Also denote $Z_i$ as the treatment indicator with $Z_i = 1$ if the patient is assigned into the treatment group, and $Z_i = 0$ if the patient is assigned into the control group. Consider the following two-level linear regression model proposed in Moineddin et al. (2008) for the longitudinal outcomes: the first level of the model assumes that response variable $y_{ij}$ is linear function of observation time $t_{ij}$, which can be expressed as

$$ y_{ij} = \beta_0 i + \beta_1 i t_{ij} + \varepsilon_{ij}, $$

(2.13)

where $\varepsilon_{ij}$ is the random error term following a normal distribution with mean zero and a constant variance $\sigma^2$. $\beta_0 i$ and $\beta_1 i$ are respectively a random intercept and the random slope associated with subjects in longitudinal studies, which can be explained
by linear functions of a level-2 variable $Z_i$ in the following forms:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} Z_i + u_{0i},$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11} Z_i + u_{1i},$$

where $\gamma_{rs}, r = 0, 1$ and $s = 0, 1$ are population average fixed effect parameters, as well as $u_{0i}$ and $u_{1i}$ are random errors following a bivariate normal distribution with mean zero and variance-covariance $\text{Var}(u_{0i}) = \sigma_0^2$, $\text{Var}(u_{1i}) = \sigma_1^2$, and $\text{Cov}(u_{0i}, u_{1i}) = \sigma_{01}^2$. Since $Z_i$ is the treatment indicator, it can be seen that the fixed effects $\gamma_{00}$ and $\gamma_{10}$ are respectively the population average of response at baseline (intercept) and the population average of change over time (slope) for patients in the control group, while the parameters $\gamma_{01}$ and $\gamma_{11}$ can be interpreted as the differences in respectively the population averages of response at baseline (intercepts) and the population average of changes over time (slopes) between the treatment and the control groups. Parameter $\sigma_0^2$ is the residual variance of the response at baseline (intercept), $\sigma_1^2$ is the residual variance of the change rate (slope), and $\sigma_{01}^2$ is the residual covariance between the baseline response and rate of change.

It is known that $u_{1i}$ represents the residual of the regression slope across the subjects. When the variance of $u_{1i}$ is significant at a two-sided 0.05 level, Moineddin et al. (2008) suggested that treatment-sensitive subsets can be identified based on a baseline factor (age, gender, biomarker, etc.) of patients by correlating $u_{1i}$ with this factor using a t-test or analysis of variance if the factor is categorical and the Pearson or Spearman correlations if the factor is continuous. When the association is significant at two-sided 0.05 level, treatment-sensitive subsets can be defined by the natural grouping generated by the categories of the baseline factor when it is
categorical (for example, female and male subsets if the gender is the baseline factor). When the factor is continuous such as the age or value of a biomarker, however, a cutpoint is required. Only an ad-hoc approach using the median of the factor as a cutpoint was suggested and there was no formal procedure proposed to estimate the cutpoint.

2.2.2 A Prediction Model Approach

Andrews (2017) proposed a complete procedure which can be used for both identification of treatment-sensitive subsets and validation of the subsets identified based on longitudinal measurements. First step in the proposed procedure is to use a linear mixed model which includes a random effect term to evaluate the individual treatment effect and a fixed effect term to evaluate the population average treatment effect. Based on the estimates of individual treatment effect, various classifying methods can then be used to build prediction models which can be used to identify treatment-sensitive subsets based on characteristics of patients. A validation step is then followed to select the best prediction model under a marginal regression framework.

Specifically, consider the following random intercept-slope linear mixed model:

\[
y_{ij} = \beta_0 + \alpha_{0i} + (\beta_1 + \alpha_{1i})Z_{i}t_{ij} + \beta_2 t_{ij} + e_{ij},
\]

where \(t_{ij}, y_{ij}\) and random error term \(e_{ij}\) are the same as defined in the last subsection, \(\beta_0\) and \(\beta_1\) represent respectively the population average of the initial status and the treatment effect over time, \(\alpha_{0i}\) and \(\alpha_{1i}\) are respectively the random intercept and slope for subject \(i\), and \(\beta_2\) is the fixed effect of time. The interaction effect \(\beta_1 + \alpha_{1i}\) between
the treatment and time in this model describes the trend of individual treatment
effect over time.

To simplify the presentation of the procedure, model (2.14) can be rewritten in
matrix form as

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

where $Y$ is an $n$-dimensional vector of the responses with $n = \sum_{i=1}^{N} n_i$, $X$ and $D$ are
an $n \times 3$ and $n \times 2N$ matrices of covariates corresponding to the fixed effects
$\beta = (\beta_0, \beta_1, \beta_2)'$ and random effects $\alpha = (\alpha_{01}, \ldots, \alpha_{0N}, \alpha_{11}, \ldots, \alpha_{1N})$, respectively, and $e$ is a $m$-dimensional vector of the random errors. It is assumed that $E(\alpha) = 0$ and $E(e) = 0$. In addition, it is assumed that $\alpha$ and $e$ are independent and distributed as multivariate normal as

$$
\begin{bmatrix}
\alpha \\
e
\end{bmatrix}
\sim
N
\left(
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
G & 0 \\
0 & R
\end{bmatrix}
\right).
$$

By using the conventional maximum likelihood method for the linear mixed model,
the parameter estimates for the fixed and random effects can be obtained as following:

$$
\hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}Y, \\
\hat{\alpha} = \hat{G}D'\hat{\Sigma}^{-1}(Y - X\hat{\beta}),
$$

where $\Sigma = DGD' + R$ and $\hat{G}$ and $\hat{R}$ are obtained by maximizing the following
likelihood function:

\[
l(R, G|Y, X) = -\frac{1}{2}(Y - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y)'\Sigma^{-1}(Y - X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y) - \frac{1}{2}\log|\Sigma| - \frac{n}{2}\log(2\pi),
\]

where $|\Sigma|$ is the determinant of the variance-covariance matrix $\Sigma$. The asymptotic consistency and efficiency of these estimates was proved by [Hartley and Rao (1967)](https://www.jstor.org/stable/2333824). Furthermore, if the variance estimation is biased, the restricted maximum likelihood would be a viable alternative method ([Verbeke and Molenberghs, 2009](https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470713749.ch2)).

Since the random slope $\beta_1 + \alpha_{1i}$ describes the treatment effect over time, patients can be divided into two subsets based on whether its estimate $\hat{\beta}_1 + \hat{\alpha}_{1i}$ is positive. Define $C_i$ as the subset indicator based on this definition. That is,

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

Since some baseline characteristics or covariates of patients, such as age, gender, blood pressure, and gene expression, might influence the treatment effect, a prediction model

\[
f(X_i) = P(C_i = 1|X_i)
\]

based on the subset indicator $C_i$ and these baseline characteristics or covariates $X_i$ may be used to classify patients into two subsets which have differential treatment effects. In general, the relationship between $C_i$ and $X_i$ is unknown, which could be linear or nonlinear, so the predictive function $f(\cdot)$ in the prediction model needs to be estimated. [Andrews (2017)](https://www.jstor.org/stable/2027827) suggested various linear or nonlinear supervised learning.
algorithms, such as logistic regression, support vector machine (SVM) with linear kernel, linear discriminant analysis (LDA), decision tree, random forest, etc., may be used to estimate \( f(\cdot) \). Once the estimated prediction function \( \hat{f}(X_i) \) is obtained from the data, it was proposed that patient \( i \) can be classified in the subset of patients who may benefit from the treatment if \( \hat{f}(X_i) > 0.5 \).

Andrews (2017) also developed a validation procedure to assess the effectiveness of the method proposed above for treatment-sensitive subset identification but the choice of 0.5 as the cutpoint for estimated predictive function to define the subsets is ad-hoc, which may have large impact on the performance of the proposed method.

2.2.3 A Procedure Based on Structured Mixed-Effects Models

Shen and Qu (2020) utilized a structured mixed-effects model for identification of subgroups based on longitudinal outcomes. In this model, treatment effects are assumed to follow a normal mixture model. Specifically, denote \( y_{ij} \) as a continuous response for the \( i \)-th patient measured at \( j \)-th time point, \( i = 1, 2, \cdots, N \), \( j = 1, \cdots, n_i \), and define \( Y_i \) as a vector of \( \{y_{i1}, \cdots, y_{in_i}\} \). Let \( X_{ij} \) be a \( q_1 \)-dimensional vector of time-varying covariates, \( X_i \) a \( q_1 \times n_i \) matrix of \( \{X_{ij} : j = 1, \cdots, n_i\} \), and \( W_i \) a \( q_2 \)-dimensional vector of covariates associated with the subgroup membership through the following logistical regression model:

\[
\logit(P(\delta_i = 1|W_i, X_i)) = W_i' \gamma, \tag{2.16}
\]

where \( \delta_i \) is a latent subgroup indicator taking values as 0 and 1 and \( \gamma \) is a vector of unknown parameters.

Given the subgroup indicator \( \delta_i \) and the other covariates, the following linear
mixed model is first assumed for the longitudinal outcomes $Y_i$:

$$Y_i|\{\delta_i, X_i, W_i, b_i, Z_i\} = X_i'\beta + Z_i b_i + \varepsilon_i,$$

where $Z_i$ is the treatment indicator with $Z_i = 1$ if the patient is assigned into the treatment group or $Z_i = 0$ if the patient is assigned into the control group, $\beta$ is fixed effect, $\varepsilon_i$ is a $n_i \times 1$ random error vector distributed as $N(0, \sigma^2 R)$, and $b_i$ is an $n_i \times 1$ treatment effect vector which is assumed to follow a normal mixture distribution:

$$b_i \sim \delta_i N(\mu_1 1, \sigma^2 I_{n_1 \times n_1}) + (1 - \delta_i) N(\mu_0 1, \sigma^2 I_{n_1 \times n_1}),$$

where $1$ is a $n_i \times 1$ vector with each element as 1. Then, assuming that $\varepsilon_i$ is independent of the treatment effect $b_i$ and conditional on $X_i, W_i, Z_i, \delta_i$, the joint distribution of $Y_i$ and $b_i$ is given as

$$\begin{bmatrix} Y_i \\ b_i \end{bmatrix} \sim N \left( \begin{bmatrix} X_i'\beta + Z_i b_i \\ \mu_i 1 \end{bmatrix}, \begin{bmatrix} \Sigma_i & \sigma^2 Z_i I_{n_1 \times n_1} \\ \sigma^2 Z_i I_{n_1 \times n_1} & \sigma^2 I_{n_1 \times n_1} \end{bmatrix} \right), \quad (2.17)$$

where $\mu_i = \mu_1 \delta_i + \mu_0 (1 - \delta_i)$ and $\Sigma_i = \sigma^2 (R + Z_i^2 I_{n_1 \times n_1})$.

For simplicity, let $D_i = \sigma^2 I_{n_1 \times n_1}$ and $\theta = (\gamma, \beta, \mu_0, \mu_1, \sigma^2, R)$. Based on the joint distribution (2.17), the posterior expectation of $b_i$ conditional on $W_i, X_i, Z_i, \delta_i, Y_i$ can be written as

$$\hat{b}_i = E[b_i|W_i, X_i, Z_i, \delta_i, Y_i] = \mu_i 1 + Z_i D_i \Sigma_i^{-1}(Y_i - X_i'\beta - Z_i \mu_i 1).$$
2.2. STATISTICAL METHODS ON LONGITUDINAL OUTCOMES 27

The corresponding variance-covariance matrix of \( \hat{\mathbf{b}}_i \) is

\[
\hat{\mathbf{D}}_i = \mathbf{D}_i - Z_i^2 \mathbf{D}_i \mathbf{\Sigma}_i^{-1} \mathbf{D}_i,
\]

and the second moment of \( \mathbf{b}_i \) given \( \delta_i \) and \( \mathbf{Y}_i \) can be written as

\[
B_i = E(\mathbf{b}_i \mathbf{b}_i' | \delta_i, \mathbf{Y}_i) = \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i' + \hat{\mathbf{D}}_i.
\]

Based on equations (2.18) and (2.19), an EM-type algorithm can be used to obtain an estimate of \( \theta \). With an estimate of \( \gamma \), the probability with which patient would benefit more from the treatment can be calculated from the logistic model (2.16). Since a higher probability corresponds to having a better chance to benefit from the treatment, in practice, the subgroups can be formed by ranking patients’ predicted probabilities. However, no specific decision rule is given to determine a cut point of the rank which can be used objectively to define subgroups.
Chapter 3

A Threshold Linear Mixed Model

3.1 Introduction

Longitudinal data, which are repeated measurements made on the same subjects, are often collected in clinical trials or other medical studies. For example, although the evaluation of treatment effects in cancer clinical trials have been traditionally based on relatively objective endpoints such as response rate, disease or progression-free survival, or overall survival, it is argued that these endpoints may not be adequate in the understanding of treatment effects on patients. Recently, investigation on more subjective endpoints, such as patient reported quality of life, has been receiving increased attention since these endpoints can help future patients understand the consequences of their illness and treatment (Bezjak et al. 2006). These patient reported outcomes are usually longitudinal, collected from assessments made at several timepoints before, during and after patients have received the treatment. In clinical studies such as randomized control trials, average effects of a new treatment for all patients enrolled in the studies are usually assessed. However, because of heterogeneity in demographic
or disease characteristics, different patients may respond differently to the same treat-
ment. Therefore, it is important to identify subsets of patients who may be sensitive 
to a specific treatment in clinical trials. For example, in a recent secondary analysis of 
data from CO.17 and CO.20 trials conducted by the Canadian Cancer Trials Group (CCTG), the investigators were interested to know whether older patients with ad-
vanced colorectal cancer treated by respectively cetuximab alone or cetuximab plus 
brivanib had a less benefit, in comparison with younger patients, in terms of various 
outcomes including quality of life (QoL) (Wells et al., 2008).

Subgroup analysis, which includes estimation of treatment effects in different sub-
groups and tests for the significance of the differences in the treatment effects in these 
subgroups, is a main statistical tool to assess the heterogeneity in treatment effects in 
subgroups defined based on certain characteristics of patients. For example, in CO.17 
and CO.20 analyses mentioned above, patients were divided into two age subgroups 
based on whether their age was 70 years or older and differential treatment effects in 
these two age groups were assessed through a test of interaction between subgroup 
and treatment. However, it is unclear whether 70 years is an optimal cutpoint to 
define the two age subgroups when assessing the heterogeneity of treatment effects 
by age. This issue arises in many studies where the variable to define subgroups is 
continuous but a pre-specified cutpoint is not available from previous studies or clin-
ical experience, and a statistical approach is often needed to determine the optimal 
cutpoints based on data.

Various data-driven statistical methods have been proposed for subgroup analy-
ses in clinical trials, which include Bayesian penalized regression model with terms 
involving covariate-by-treatment interactions (Dixon and Simon, 1991), tree based
3.1. INTRODUCTION

partitioning with interactions (Loh, 2002), and “virtual twins” method (Foster et al., 2011). A Review on some of these methods may be found in Lipkovich et al. (2017).

Most of these methods deal with the situations where the outcomes are binary or continuous and there are multiple covariates. When the outcomes of clinical trials are times to an event, such as progression-free or overall survivals, and there is only single covariate, several simple approaches have been proposed for the determination of cutpoint in the definition of subgroups. For example, Jiang et al. (2007) proposed a biomarker-adaptive threshold design, which combines a test for overall treatment effect in all patients with the determination and validation of a cutpoint for a biomarker which is used to define a sensitive subgroup. Chen et al. (2014) developed a hierarchical Bayesian procedure to estimate simultaneously the interaction parameter and cutpoint in a threshold Cox model.

As reviewed in Chapter 2, no statistical procedure to determine the cutpoint of a continuous covariate is available in the literature, however, when the outcomes are longitudinal measurements. Part of the reasons may be due to existence of correlations between the longitudinal measurements, which requires different classes of models such as mixed effects models for the data. Estimations of parameters in these models may be more complicated and computation may be more challenging. Moineddin et al. (2008) used multi-level models including subject-specific random effects to identify subgroups with differential treatment effects of gabapentin versus placebo on longitudinal measurements of hot flashes based on the baseline score in a double-blind randomized controlled trial for treatment of hot flashes in women who enter menopause naturally but a median was used as the cutpoint in defining subgroups. Andrews (2017) considered a random effects linear model for longitudinal
3.1. INTRODUCTION

outcomes to determine whether a patient had a positive response to the treatment and supervised learning algorithms were proposed to estimate a predictive function for the positive response but 0.5 was used as an ad-hoc cutpoint for the predictive function to assign patients into subgroups.

In this chapter, a threshold linear mixed model is introduced which can be used simultaneously to determine the cutpoint of a continuous covariate, such as age or the expression level of a biomarker, in forming subgroups and to assess the interaction effect between the treatment and subgroup indicator on longitudinal outcomes. The standard likelihood method is difficult to apply to the inference of the parameters in the model because the likelihood function is not continuous for some parameters. A smoothing likelihood function is therefore proposed to approximate the original likelihood function and to make inferences on the model parameters based on the new likelihood function. The proposed procedure is evaluated through simulation studies and applications to a cancer trial and AIDS clinical trial.

The remainder of this chapter is organized as follows. Section 3.2 introduces a threshold linear mixed model and the proposed statistical procedure for the inference of the parameters in the proposed model. Section 3.3 presents results from simulation studies to assess the performance of the proposed procedure. An application of the proposed procedure to two randomized clinical trials is presented in Section 3.4. This chapter is concluded with some discussions in Section 3.5.
3.2 Methodology

3.2.1 The Model and Notations

Denote a column vector \( Y_i = (y_{i1}, y_{i2}, \cdots, y_{in_i}) \) for the longitudinal measurements of the \( i \)-th subject with \( n_i \) observations, and the \( j \)-th element \( y_{ij} \) is the \( j \)-th observation of subject \( i \) measured at the time \( t_{ij} \), where \( n_i \) is the number of observations from the \( i \)-th subject, \( i = 1, 2, \cdots, N \) and \( N \) is the number of subjects. For each subject, denote \( X_i = (x_{i1}, x_{i2}, \cdots, x_{in_i})' \) as a \((n_i \times p)\) designed matrix of covariates for fixed effect \( \beta \) and \( Z_i = (z_{i1}, z_{i2}, \cdots, z_{in_i})' \) as a \((n_i \times q)\) designed matrix of covariates for random effect \( \alpha_i \). Assume \( b_i \) is an indicator of the treatment received by subject \( i \) with either \( b_i = 1 \) if the patient receiving a new therapy or \( b_i = 0 \) if not. Denote \( w_i \) as a continuous covariate at baseline for subject \( i \) and assume two subgroups of subjects can be defined based on whether \( w_i \) exceeds an unknown cutpoint \( c \). I propose the following threshold linear mixed model to assess the potential differential treatment effects between these two subgroups:

\[
Y_i = X_i \beta + Z_i \alpha_i + \eta_1 I(w_i > c)1 + \eta_2 b_i I(w_i > c)1 + \varepsilon_i, \tag{3.1}
\]

where \( \varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{in_i})' \) is a vector of random errors and \( 1 \) is a \( n_i \)-dimensional vector with its all elements as 1. In model (3.1), the response \( y_{ij} \) of subject \( i \) measured at the time \( t_{ij} \) is modeled by three components: the fixed effects of all covariates \( x_{ij}' \beta + \eta_1 I(w_i > c) + \eta_2 b_i I(w_i > c) \), the subject effect \( z_{ij}' \alpha_i \), and the random error \( \varepsilon_{ij} \). The columns of \( X_i \) may include intercept, time or its function, treatment, and other confounding variables, and the columns of \( Z_i \) are assumed as a subset of the columns of \( X_i \). In order to simplify the presentation, I rewrite model (3.1) in the matrix form...
3.2. METHODOLOGY

defined as:

\[ Y = X\beta + W\eta + Z\alpha + \varepsilon, \tag{3.2} \]

where \( Y = [Y_1', Y_2', \cdots, Y_N']' \), \( X = [X_1', X_2', \cdots, X_N']' \), \( \alpha = (\alpha'_1, \alpha'_2, \cdots, \alpha'_N)' \), 
\( \varepsilon = (\varepsilon'_1, \varepsilon'_2, \cdots, \varepsilon'_N)' \) and \( W = [W_1', W_2', \cdots, W_N']' \), and

\[
Z = \begin{pmatrix}
Z_1 & 0 & 0 & \cdots & 0 \\
0 & Z_2 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & Z_N
\end{pmatrix}, \quad W_i = \begin{pmatrix}
I(w_i > c) & b_i \times I(w_i > c) \\
I(w_i > c) & b_i \times I(w_i > c) \\
\vdots & \vdots \\
I(w_i > c) & b_i \times I(w_i > c)
\end{pmatrix}_{n_i \times 2}.
\]

For the vector of random effects \( \alpha \) and vector of random errors \( \varepsilon \) in the model, it is assumed that \( E(\alpha) = 0 \) and \( E(\varepsilon) = 0 \). In addition, it is assumed that \( \alpha \) and \( \varepsilon \) are independent and distributed as multivariate normal, that is,

\[
\begin{bmatrix}
\alpha \\
\varepsilon
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\
0 \end{bmatrix}, \begin{bmatrix} G & 0 \\
0 & R \end{bmatrix} \right).
\]

It is also assumed that \( R = \sigma^2I \) (\( \sigma \) is an unknown parameter) and \( G = \sigma^2\rho^2I \) (\( \rho \) is also an unknown parameter). Following [Patterson and Thompson (1971)], the covariance-variance matrix of the observation \( Y \) can be written as

\[ Var(Y) = \Sigma = \sigma^2(\rho^2ZZ' + I) = \sigma^2H, \]

where \( H = \rho^2ZZ' + I \).
3.2. METHODOLOGY

3.2.2 Inferences of Model Parameters

Under assumptions and notations mentioned above, \( Y \) follows a multivariate normal distribution as \( N(X\beta + W\eta, \sigma^2H) \). Denote \( n = \sum_{i=1}^{N} n_i \) as the total number of observations, The log-likelihood for the unknown parameters \( \theta = (\beta, \eta, c, \rho^2, \sigma^2) \) in model (3.2) based on longitudinal outcomes \( Y \) can be written as

\[
l(\theta|Y, X, Z) = -\frac{1}{2} \left\{ \log(2\pi) + n \log \sigma^2 + \log |H| + \frac{(Y - X\beta - W\eta)'H^{-1}(Y - X\beta - W\eta)}{\sigma^2} \right\}. \tag{3.3}
\]

However due to the presence of the indicator functions \( I(w_i > c) \), the log-likelihood function is not continuous with respect to \( c \), which makes the conventional maximum likelihood theory and algorithm difficult to apply. Following a smoothing procedure used by Brown and Wang (2007), I propose to use a kernel smooth function

\[
\Phi \left( \frac{w_i - c}{h} \right) \tag{3.4}
\]

as a smooth approximation to the indicator function \( I(w_i > c) \), where \( \Phi \) is the distribution function of the standard normal distribution and \( h \) is a bandwidth which converges to zero as the sample size increases. Note that if \( w_i > c \), \( \Phi((w_i - c)/h) \to 1 \) as \( n \to \infty \), while \( \Phi((w_i - c)/h) \to 0 \) as \( n \to \infty \) when \( w_i < c \), that is, \( \Phi((w_i - c)/h) \to \).
3.2. METHODOLOGY

$I(w_i - c) > 0$ as $n \to \infty$. Using this approximation, I can define a smoothed log-likelihood function by replacing $W_i$ in the definition of $W$ in (3.3) with

$$\tilde{W}_i = \begin{bmatrix}
\Phi(\frac{w_i - c}{h}) b_i \times \Phi(\frac{w_i - c}{h}) \\
\Phi(\frac{w_i - c}{h}) b_i \times \Phi(\frac{w_i - c}{h}) \\
\vdots \\
\Phi(\frac{w_i - c}{h}) b_i \times \Phi(\frac{w_i - c}{h})
\end{bmatrix}_{n \times 2},$$

therefore the smoothed log-likelihood function of $\theta$ is given by

$$sl(\theta|Y, X, Z) = -\frac{1}{2} \left\{ \log(2\pi) + n \log \sigma^2 + \log |H| + \frac{(Y - X\beta - \tilde{W}\eta)'H^{-1}(Y - X\beta - \tilde{W}\eta)}{\sigma^2} \right\}, \quad (3.5)$$

where $\tilde{W} = [\tilde{W}_1, \tilde{W}_2, \ldots, \tilde{W}_n]'$. The maximum smoothed likelihood estimates (MSLE) of $\theta$ can be obtained by maximizing the smoothed log-likelihood function (3.5) or, equivalently, solving the following score equations:

$$\frac{\partial sl}{\partial \beta} = \frac{1}{\sigma^2} X' H^{-1} (Y - X\beta - \tilde{W}\eta) = 0,$$

$$\frac{\partial sl}{\partial \eta} = \frac{1}{\sigma^2} \tilde{W}' H^{-1} (Y - X\beta - \tilde{W}\eta) = 0,$$

$$\frac{\partial sl}{\partial c} = \tilde{W}' \eta H^{-1} (Y - X\beta - \tilde{W}\eta)/\sigma^2 = 0,$$

$$\frac{\partial sl}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + (Y - X\beta - \tilde{W}\eta)' H^{-1} (Y - X\beta - \tilde{W}\eta)/(2\sigma^4) = 0,$$

$$\frac{\partial sl}{\partial \rho^2} = -\frac{1}{2} tr(H^{-1} \tilde{H}_\rho) + \frac{1}{2\sigma^2} [(Y - X\beta - \tilde{W}\eta)' H^{-1} \tilde{H}_\rho H^{-1} (Y - X\beta - \tilde{W}\eta)] = 0,$$
3.2. METHODOLOGY

where \( \dot{W} = \frac{\partial W}{\partial c} \) and \( \dot{H}_{\rho^2} = \frac{\partial H}{\partial \rho^2} = \frac{\partial (Z^T G Z')}{\partial \rho^2} = ZZ'. \) The Hessian matrix of log-likelihood is given as

\[
J(\beta, \eta, c, \sigma^2, \rho^2) = \begin{pmatrix}
\frac{\partial^2 s_l}{\partial \beta^2} & \frac{\partial^2 s_l}{\partial \beta \eta} & \frac{\partial^2 s_l}{\partial \beta \sigma^2} & \frac{\partial^2 s_l}{\partial \beta \rho^2} \\
\frac{\partial^2 s_l}{\partial \beta \eta} & \frac{\partial^2 s_l}{\partial \eta^2} & \frac{\partial^2 s_l}{\partial \eta \sigma^2} & \frac{\partial^2 s_l}{\partial \eta \rho^2} \\
\frac{\partial^2 s_l}{\partial \beta \sigma^2} & \frac{\partial^2 s_l}{\partial \eta \sigma^2} & \frac{\partial^2 s_l}{\partial \sigma^2} & \frac{\partial^2 s_l}{\partial \sigma \rho^2} \\
\frac{\partial^2 s_l}{\partial \beta \rho^2} & \frac{\partial^2 s_l}{\partial \eta \rho^2} & \frac{\partial^2 s_l}{\partial \sigma \rho^2} & \frac{\partial^2 s_l}{\partial (\rho^2)^2}
\end{pmatrix},
\]

where each component in the Hessian matrix is defined as following:

\[
\frac{\partial^2 s_l}{\partial \beta^2} = -X' \Sigma^{-1} X,
\]
\[
\frac{\partial^2 s_l}{\partial \beta \eta} = -\frac{1}{\sigma^2} X' H^{-1} \dot{W},
\]
\[
\frac{\partial^2 s_l}{\partial \beta \sigma^2} = 0,
\]
\[
\frac{\partial^2 s_l}{\partial \beta \rho^2} = -\frac{1}{\sigma^2} X' \Sigma^{-1} (Y - X \beta - \dot{W} \eta),
\]
\[
\frac{\partial^2 s_l}{\partial \eta^2} = -X' \Sigma^{-1} \dot{H}_{\rho^2} \Sigma^{-1} (Y - X \beta - \dot{W} \eta),
\]
\[
\frac{\partial^2 s_l}{\partial \eta \sigma^2} = \dot{W}' \Sigma^{-1} \dot{W},
\]
\[
\frac{\partial^2 s_l}{\partial \eta \rho^2} = -\dot{W}' \Sigma^{-1} (Y - X \beta - \dot{W} \eta) + \dot{W}' \Sigma^{-1} \dot{W} \eta,
\]
\[
\frac{\partial^2 s_l}{\partial \sigma^2} = -\frac{1}{\sigma^2} \dot{W}' \Sigma^{-1} (Y - X \beta - \dot{W} \eta),
\]
\[
\frac{\partial^2 s_l}{\partial \sigma \rho^2} = -\frac{1}{\sigma^2} \dot{W}' \Sigma^{-1} \dot{H}_{\rho^2} \Sigma^{-1} (Y - X \beta - \dot{W} \eta),
\]
3.2. METHODOLOGY

\[
\frac{\partial^2 sl}{\partial c^2} = (\ddot{W}\eta)'\Sigma^{-1}(Y - X\beta - \tilde{W}\eta) - (\ddot{W}\eta)'\Sigma^{-1}(\ddot{W}\eta),
\]

\[
\frac{\partial^2 sl}{\partial c \partial \sigma^2} = -\frac{1}{\sigma^2}(\dot{W}\eta)'\Sigma^{-1}(Y - X\beta - \tilde{W}\eta),
\]

\[
\frac{\partial^2 sl}{\partial \sigma^2 \partial \rho^2} = (-\dot{W}\eta)'\Sigma^{-1}\dot{H}_{\rho^2}\Sigma^{-1}(Y - X\beta - \tilde{W}\eta)/\sigma^2,
\]

\[
\frac{\partial^2 sl}{\partial \sigma^2} = \frac{n}{2\sigma^4} - (Y - X\beta - \tilde{W}\eta)'\Sigma^{-1}(Y - X\beta - \tilde{W}\eta)/\sigma^4,
\]

\[
\frac{\partial^2 sl}{\partial \rho^2} = -(Y - X\beta - \tilde{W}\eta)'\Sigma^{-1}\dot{H}_{\rho^2}\Sigma^{-1}(Y - X\beta - \tilde{W}\eta),
\]

\[
\frac{\partial^2 sl}{\partial (\rho^2)^2} = \frac{1}{2}tr(H^{-1}\dot{H}_{\rho^2})^2 + \frac{1}{\sigma^2}(Y - X\beta - \tilde{W}\eta)'(H^{-1}\dot{H}_{\rho^2})^2
\]

\[
H^{-1}(Y - X\beta - \tilde{W}\eta),
\]

where \(\dddot{W} = \frac{\partial^2 W}{\partial c^2}\).

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm proposed in Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970), which belongs to quasi-Newton methods and is included in R package ‘maxLik’ (Henningsen and Toomet 2011), can be used to calculate the MSLE of \(\theta\). Specifically, in the case considered in this chapter, the algorithm is applied to minimize the minus smoothed log-likelihood function \(sl(\theta)\) from the function (3.5) and to find the corresponding parameters \(\theta\), which can be described in the following steps:

**Step 1**: Choose an initial value of parameters \(\theta_0\) and obtain the corresponding approximate Hessian matrix \(H_0\);

**Step 2**: In k-th iteration, Obtain a direction \(P_k\) by solving \(H_kP_k = -\nabla sl(\theta_k)\);

**Step 3**: Perform a one-dimensional optimization (linear search) to find an acceptable step-size \(\Delta \theta_k\) in the direction found in the step 2, that is, \(\Delta \theta_k = \arg \min sl(\theta + \Delta \theta P_k)\);

**Step 4**: Set \(s_k = \Delta \theta_k P_k\) and \(\theta_{k+1} = \theta_k + \Delta \theta_k P_k\);
Step 5: Obtain $y_k = \nabla s_l(\theta_{k+1}) - \nabla s_l(\theta_k)$;

Step 6: Obtain $H_{k+1} = H_k + \frac{y_k y_k'}{y_k' s_k} - \frac{H_k s_k' H_k'}{s_k' H_k s_k}$;

Step 7: Repeat Step 2 to Step 6 ($k = 0, 1, 2, \cdots$) until $\| \nabla s_l(\theta_k) \|$ converges to 0.

After the estimates of $\theta$ are obtained, their standard errors can be calculated based on the inverse matrix of the Hessian matrix $J(\beta, \eta, c, \sigma^2, \rho^2)$.

A comment on the selection of the bandwidth parameter $h$ before closing this section: Since our objective is to use a smooth function to approximate an indicator function, in principle, a smaller $h$ would lead to a better estimator but our simulation studies showed that an extremely small $h$ may cause instability in the proposed estimator. Based on the theoretical and empirical evidences presented by Lin et al. (2011) and He et al. (2018), I select $h$ as $h = \hat{d} n^{-1/3}$ for the proposed estimator, where $\hat{d}$ is the sample standard deviation of covariate $w_i$. The performance of this method of determining $h$ will be examined by simulation in the next section.

### 3.3 Simulation Studies

In the simulation studies, I have assessed the performance of the MSLEs for the parameters in the following threshold linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 b_i + \eta_1 I(w_i > c) + \eta_2 b_i I(w_i > c) + \alpha_i + \varepsilon_{ij},$$

where $\beta_0$ is an intercept, $\alpha_i$ is a random variable for the subject effect, $\beta_1, \beta_2, \eta_1,$ and $\eta_2$ are respectively the effects of measurement time, treatment indicator, subgroup indicator, and interaction between the treatment and subgroup indicators, and $\varepsilon_{ij}$ is a random error.

In the simulations, it is assumed that subjects are randomly assigned to either
treatment group $b_i = 1$ or control group $b_i = 0$, each with probability $1/2$ and each subject is measured at the same four equal spaced timepoints 1, 2, 3, and 4. For each subject $i$, the baseline covariate $w_i$ and random effect $\alpha_i$ are generated respectively from a uniform distribution over $[0, 1]$ and a normal distribution with mean zero and covariance $\tilde{\sigma}^2 = \rho^2 \sigma^2 = 0.64 \times 2.25 = 1.44$, while the random error $\varepsilon_{ij}$ is generated from a normal distribution with mean zero and covariance $\sigma^2 = 2.25$. The coefficients of fixed effects in the model are taken as $\beta_0 = 1$, $\beta_1 = 5$, $\beta_2 = 2$, $\eta_1 = 4$, $\eta_2 = 3$, while the true cutpoint of $w_i$ for the definition of subgroup is $c = 0.3$. The number of patients is set to $N = 100$, $N = 200$ and $N = 400$, which implies that the total observations from these patients are $n = 400$, $n = 800$ and $n = 1600$ since each has 4 longitudinal measurements. In each scenario, 500 replications are used to obtain the bias, standard deviation (SD), and the root of the mean square error (RMSE) of the parameter estimators.

The performance of procedure proposed in this chapter is first assessed when the bandwidth is calculated based on the optimal formula $h = \hat{d} n^{-1/3}$. The bias, SD and RMSE of the parameters estimates are summarized Table 3.1. From Table 3.1, one can see that the bias, SD and RMSE of the estimates for all parameters, except the bias of estimate for $\sigma^2$ and $\rho^2$, decrease as the total number of observations increases when $h$ is fixed at the optimal value and RMSE becomes very small for most of parameters when $n = 1600$, which suggests that the bandwidth $h = \hat{d} n^{-1/3}$ may truly minimize RMSE. To verify this and assess the impact of bandwidth on the estimation, I have fixed the number of observations as $n = 800$ and varied $h$ over a range of $h = 0.1, 0.07, 0.04, 0.01, 0.001, 0.0001$. Table 3.2 presents the bias, SD and RMSE of the estimates for each parameter from this investigation, which
Table 3.1: Results for simulation studies over 500 replications with optimal $h$ for sample size $n = 400, 800, 1600$. The numbers inside the brackets next to the parameter estimate and the sample size are the true value of the parameter and the optimal value of $h$, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\beta}_0(1)$</th>
<th>$\hat{\beta}_1(5)$</th>
<th>$\hat{\eta}_2(2)$</th>
<th>$\hat{\eta}_1(4)$</th>
<th>$\hat{\eta}_3(3)$</th>
<th>$\hat{c}(0.3)$</th>
<th>$\hat{\sigma}^2(2.25)$</th>
<th>$\hat{\rho}^2(0.64)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias</td>
<td>SD</td>
<td>RMSE</td>
<td>Bias</td>
<td>SD</td>
<td>RMSE</td>
<td>Bias</td>
<td>SD</td>
</tr>
<tr>
<td>$n = 400$ (.038)</td>
<td>0.111</td>
<td>0.432</td>
<td>0.446</td>
<td>0.089</td>
<td>0.594</td>
<td>0.601</td>
<td>0.002</td>
<td>0.064</td>
</tr>
<tr>
<td>$n = 800$ (.031)</td>
<td>0.070</td>
<td>0.302</td>
<td>0.310</td>
<td>0.078</td>
<td>0.390</td>
<td>0.398</td>
<td>0.001</td>
<td>0.049</td>
</tr>
<tr>
<td>$n = 1600$ (.024)</td>
<td>0.059</td>
<td>0.204</td>
<td>0.212</td>
<td>0.061</td>
<td>0.269</td>
<td>0.276</td>
<td>0.001</td>
<td>0.034</td>
</tr>
</tbody>
</table>

indicates that when the sample size is fixed as 800, the RMSE of the estimates for all parameters tends to be the smallest when $h$ is around 0.04 except for $\beta_2$, which is not very sensitive to the selection of $h$. This value of $h$ is very close to that calculated from the optimal bandwidth formula.

I also compared the empirical standard deviations of the parameter estimates with the averages of the estimated standard errors based on the inverse of the Hessian matrix. From results presented in Table 3.3, one may observe that the empirical
Table 3.2: Results for simulation studies over 500 replications with $h = 0.1$, 0.07, 0.04, 0.01, 0.001, 0.0001 for sample size $n = 800$. The number inside the brackets next to the parameter estimate is the true value of the parameter.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\hat{\beta}_0(1)$ Bias</th>
<th>$\hat{\beta}_0(1)$ SD</th>
<th>$\hat{\beta}_0(1)$ RMSE</th>
<th>$\hat{\beta}_1(5)$ Bias</th>
<th>$\hat{\beta}_1(5)$ SD</th>
<th>$\hat{\beta}_1(5)$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.395</td>
<td>0.381</td>
<td>0.550</td>
<td>0.298</td>
<td>0.483</td>
<td>0.568</td>
</tr>
<tr>
<td>0.07</td>
<td>0.211</td>
<td>0.336</td>
<td>0.397</td>
<td>0.187</td>
<td>0.454</td>
<td>0.491</td>
</tr>
<tr>
<td>0.04</td>
<td>0.107</td>
<td>0.305</td>
<td>0.323</td>
<td>0.074</td>
<td>0.406</td>
<td>0.412</td>
</tr>
<tr>
<td>0.01</td>
<td>0.019</td>
<td>0.429</td>
<td>0.430</td>
<td>0.001</td>
<td>0.447</td>
<td>0.447</td>
</tr>
<tr>
<td>0.001</td>
<td>0.104</td>
<td>0.478</td>
<td>0.489</td>
<td>0.077</td>
<td>0.494</td>
<td>0.500</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.216</td>
<td>0.536</td>
<td>0.580</td>
<td>0.187</td>
<td>0.568</td>
<td>0.598</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\hat{\beta}_2(2)$ Bias</th>
<th>$\hat{\beta}_2(2)$ SD</th>
<th>$\hat{\beta}_2(2)$ RMSE</th>
<th>$\hat{\eta}_1(4)$ Bias</th>
<th>$\hat{\eta}_1(4)$ SD</th>
<th>$\hat{\eta}_1(4)$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.002</td>
<td>0.045</td>
<td>0.045</td>
<td>0.508</td>
<td>0.410</td>
<td>0.654</td>
</tr>
<tr>
<td>0.07</td>
<td>0.001</td>
<td>0.047</td>
<td>0.047</td>
<td>0.283</td>
<td>0.364</td>
<td>0.461</td>
</tr>
<tr>
<td>0.04</td>
<td>0.002</td>
<td>0.047</td>
<td>0.047</td>
<td>0.150</td>
<td>0.337</td>
<td>0.369</td>
</tr>
<tr>
<td>0.01</td>
<td>0.002</td>
<td>0.048</td>
<td>0.048</td>
<td>0.023</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>0.045</td>
<td>0.045</td>
<td>0.129</td>
<td>0.509</td>
<td>0.525</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.005</td>
<td>0.047</td>
<td>0.047</td>
<td>0.321</td>
<td>0.535</td>
<td>0.624</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\hat{\eta}_2(3)$ Bias</th>
<th>$\hat{\eta}_2(3)$ SD</th>
<th>$\hat{\eta}_2(3)$ RMSE</th>
<th>$\hat{c}(0.3)$ Bias</th>
<th>$\hat{c}(0.3)$ SD</th>
<th>$\hat{c}(0.3)$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.386</td>
<td>0.556</td>
<td>0.677</td>
<td>0.008</td>
<td>0.015</td>
<td>0.017</td>
</tr>
<tr>
<td>0.07</td>
<td>0.233</td>
<td>0.532</td>
<td>0.581</td>
<td>0.003</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>0.04</td>
<td>0.116</td>
<td>0.469</td>
<td>0.483</td>
<td>0.008</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td>0.01</td>
<td>0.064</td>
<td>0.696</td>
<td>0.699</td>
<td>0.009</td>
<td>0.075</td>
<td>0.076</td>
</tr>
<tr>
<td>0.001</td>
<td>0.102</td>
<td>0.557</td>
<td>0.566</td>
<td>0.009</td>
<td>0.060</td>
<td>0.061</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.255</td>
<td>0.638</td>
<td>0.687</td>
<td>0.008</td>
<td>0.084</td>
<td>0.084</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\hat{\sigma}^2(2.25)$ Bias</th>
<th>$\hat{\sigma}^2(2.25)$ SD</th>
<th>$\hat{\sigma}^2(2.25)$ RMSE</th>
<th>$\hat{\rho}^2(0.64)$ Bias</th>
<th>$\hat{\rho}^2(0.64)$ SD</th>
<th>$\hat{\rho}^2(0.64)$ RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.003</td>
<td>0.128</td>
<td>0.128</td>
<td>0.278</td>
<td>0.143</td>
<td>0.313</td>
</tr>
<tr>
<td>0.07</td>
<td>0.005</td>
<td>0.129</td>
<td>0.129</td>
<td>0.196</td>
<td>0.127</td>
<td>0.234</td>
</tr>
<tr>
<td>0.04</td>
<td>0.005</td>
<td>0.120</td>
<td>0.120</td>
<td>0.100</td>
<td>0.126</td>
<td>0.161</td>
</tr>
<tr>
<td>0.01</td>
<td>0.010</td>
<td>0.117</td>
<td>0.118</td>
<td>0.051</td>
<td>0.358</td>
<td>0.362</td>
</tr>
<tr>
<td>0.001</td>
<td>0.012</td>
<td>0.133</td>
<td>0.134</td>
<td>0.173</td>
<td>0.477</td>
<td>0.507</td>
</tr>
<tr>
<td>0.0001</td>
<td>0.003</td>
<td>0.122</td>
<td>0.122</td>
<td>0.472</td>
<td>0.528</td>
<td>0.709</td>
</tr>
</tbody>
</table>
standard deviations are close to the averages of the estimated standard errors when $h$ lies in the range between 0.01 and 0.03. Table 3.4 presents the results from the same comparisons but with $h$ calculated from the optimal formula and three different sample sizes. It can be seen that under the optimal $h$ the empirical standard deviations and the average of estimated standard errors are close to each other and they both become smaller as the sample size increases.

Table 3.3: Comparison between empirical standard deviation and Hessian standard error from simulation studies over 500 replications with different $h$ values under subject sample size $n = 800$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
<th>$c$</th>
<th>$\sigma^2$</th>
<th>$\rho^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>Empirical</td>
<td>0.379</td>
<td>0.437</td>
<td>0.145</td>
<td>0.368</td>
<td>0.504</td>
<td>0.006</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Hessian</td>
<td>0.374</td>
<td>0.424</td>
<td>0.149</td>
<td>0.357</td>
<td>0.506</td>
<td>0.004</td>
<td>0.225</td>
</tr>
<tr>
<td>0.03</td>
<td>Empirical</td>
<td>0.384</td>
<td>0.453</td>
<td>0.145</td>
<td>0.380</td>
<td>0.525</td>
<td>0.008</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Hessian</td>
<td>0.382</td>
<td>0.437</td>
<td>0.149</td>
<td>0.369</td>
<td>0.521</td>
<td>0.005</td>
<td>0.225</td>
</tr>
<tr>
<td>0.05</td>
<td>Empirical</td>
<td>0.392</td>
<td>0.472</td>
<td>0.145</td>
<td>0.393</td>
<td>0.548</td>
<td>0.010</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Hessian</td>
<td>0.890</td>
<td>0.449</td>
<td>0.149</td>
<td>0.379</td>
<td>0.536</td>
<td>0.005</td>
<td>0.225</td>
</tr>
<tr>
<td>0.07</td>
<td>Empirical</td>
<td>0.405</td>
<td>0.495</td>
<td>0.145</td>
<td>0.409</td>
<td>0.576</td>
<td>0.012</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Hessian</td>
<td>0.396</td>
<td>0.460</td>
<td>0.149</td>
<td>0.389</td>
<td>0.550</td>
<td>0.005</td>
<td>0.225</td>
</tr>
<tr>
<td>0.09</td>
<td>Empirical</td>
<td>0.422</td>
<td>0.525</td>
<td>0.145</td>
<td>0.429</td>
<td>0.610</td>
<td>0.014</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>Hessian</td>
<td>0.403</td>
<td>0.472</td>
<td>0.149</td>
<td>0.398</td>
<td>0.562</td>
<td>0.006</td>
<td>0.225</td>
</tr>
</tbody>
</table>

I further examined the coverage probabilities of 95% confidence intervals constructed based on the estimated standard errors of the estimates as mentioned in the last section and the asymptotic normality of the estimates. Because of the results above, I only considered $h$ ranged from 0.01 to 0.09. The results presented in Table 3.5 show that the actual coverage rate of the confidence intervals for all parameters...
Table 3.4: Comparison between empirical standard deviation and Hessian standard error from simulation studies over 500 replications with optimal $h$ for sample size $n = 400$, $n = 800$, $n = 1600$. The numbers inside the brackets next to the parameters and the sample size are the true value of the parameters and the optimal value of $h$, respectively.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$n = 400$ ($0.041$)</th>
<th>$n = 800$ ($0.032$)</th>
<th>$n = 1600$ ($0.025$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical</td>
<td>Hessian</td>
<td>Empirical</td>
</tr>
<tr>
<td>$\beta_0(1)$</td>
<td>0.430</td>
<td>0.441</td>
<td>0.378</td>
</tr>
<tr>
<td>$\beta_1(5)$</td>
<td>0.620</td>
<td>0.577</td>
<td>0.454</td>
</tr>
<tr>
<td>$\beta_2(2)$</td>
<td>0.070</td>
<td>0.067</td>
<td>0.050</td>
</tr>
<tr>
<td>$\eta_1(4)$</td>
<td>0.491</td>
<td>0.483</td>
<td>0.402</td>
</tr>
<tr>
<td>$\eta_2(3)$</td>
<td>0.690</td>
<td>0.697</td>
<td>0.573</td>
</tr>
<tr>
<td>$c(0.3)$</td>
<td>0.013</td>
<td>0.009</td>
<td>0.008</td>
</tr>
<tr>
<td>$\sigma^2(2.25)$</td>
<td>0.176</td>
<td>0.183</td>
<td>0.128</td>
</tr>
<tr>
<td>$\rho^2(0.64)$</td>
<td>0.203</td>
<td>0.160</td>
<td>0.255</td>
</tr>
</tbody>
</table>

but $c$ are closer to 95% for the range of $h$. The coverage probabilities of the confidence intervals for $c$ are slightly under 95%.

In practice, some patients may drop out the study earlier, which may lead to some observations from these patients missing. To assess the robustness of our proposed procedures to the potential missing data, I performed additional simulation studies which assume the observations from some patients are missing at some timepoints due to drop out. Specifically, in the simulations, the first measurement $y_{i1}$ is assumed non-missing for every subject $i$. A logistic regression model proposed in Section 3.2 of
Table 3.6: Results for simulation studies over 500 replications when the observations may be missing at random.

<table>
<thead>
<tr>
<th></th>
<th>β₀</th>
<th>β₁</th>
<th>β₂</th>
<th>η₁</th>
<th>η₂</th>
<th>c</th>
<th>σ²</th>
<th>ρ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIAS</td>
<td>0.110</td>
<td>0.100</td>
<td>0.002</td>
<td>0.164</td>
<td>0.007</td>
<td>0.008</td>
<td>0.009</td>
<td>0.130</td>
</tr>
<tr>
<td>SD</td>
<td>0.707</td>
<td>0.716</td>
<td>0.088</td>
<td>0.947</td>
<td>1.272</td>
<td>0.121</td>
<td>0.228</td>
<td>0.512</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.715</td>
<td>0.722</td>
<td>0.088</td>
<td>0.961</td>
<td>1.271</td>
<td>0.121</td>
<td>0.228</td>
<td>0.528</td>
</tr>
</tbody>
</table>

Diggle and Kenward (1994), which links the probability that the $k$-th observation from the $i$-th subject with the previous observations from the same subject, was used to generate a data set with potentially missing at random observations for 100 subjects during 4 scheduled assessment timepoints with a rate of missing at around 23%. The detailed model is showed in Appendix A. The estimated bias, SD, RMSE from 500 replications are presented in Table 3.6. In comparison with the results in Table 3.1 with the full sample size 400, one can find that the biases of the estimates are similar for all parameters but, because of missing observations which reduced the effective size of the sample, SD and RMSE of the estimates are larger for most of parameters, which is a reasonable increase considering the amount of missing observations.

3.4 Applications to data from Cancer and AIDS Clinical Trials

I applied the proposed model and estimation procedures to the data from the randomized clinical trial CO.17 described in Section 3.1 and data from trial 193A conducted by the AIDS Clinical Trial Group (ACTG) (Henry et al., 1998).

3.4.1 Analysis of Data from CO.17

In the trial CO.17, 572 patients were randomized to receive either cetuximab with best supportive care (BSC) or BSC alone. The trial was primarily designed to evaluate the
3.4. APPLICATIONS

overall survival (OS) benefit from cetuximab in patients with colorectal cancer with comparison of quality of life between the treatment groups as one of the important secondary objectives. In the primary analysis including all randomized patients, cetuximab demonstrated superior overall survival (OS), progression-free survival (PFS) and quality-of-life (QoL) compared to BSC [Jonker et al., 2007]. Subsequent studies found that these benefits from cetuximab are limited to a subgroup of patients with wild-type K-ras mutation [Jonker et al., 2007; Karapetis et al., 2008; Au et al., 2009]. Whether similar subgroups can be found based on other baseline covariates or biomarkers, such as age, remains unknown. I applied the proposed model to identify a subgroup of patients based on age which may have different treatment effects on QoL from the rest of patients.

For each patient in the trial, the QoL was assessed using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire C30 (EORTC QLQ-C30) at baseline and 4, 8, 16 and 24 weeks after the start of treatment. EORTC QLQ-C30 includes nine scales, which are divided into five function scales, three symptom scales and one global health-status/quality of life dimension.

In this analysis, I only considered the scores of global health-status dimension assessed over time, which had a range between 0 and 100 with higher scores representing better QoL. A total of 443 patients are included in the analysis, of whom 240 were treated with cetuximab. Because the QoL scores may be skewed, 1 was first added to all scores and then a log-transformation was taken. The baseline covariate age was re-scaled to (0, 1]. Since the sample standard deviation of age is around 0.115 and the total number of observations is 1612, the bandwidth $h$ is chosen as 0.01 based on the approach mentioned in Section 3.3. After examining the lowess time plots of QoL
scores, a piece-wise linear spline with two knots at week 8 and week 16 was used to model the time effect. That is, the following threshold linear mixed model were used to fit the data:

\[ y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 8)_+ + \beta_3 (t_{ij} - 16)_+ + \beta_4 b_i + \eta_1 I(\text{age} > c) + \eta_2 b_i I(\text{age} > c) + \varepsilon_{ij}. \]

The estimates of parameters in the above model are presented in Table 3.7.

From Table 3.7, one can see that the estimated cutpoint of age is \( c = 0.765 \), which is 67.39 years old after scaling back to the original age scale. With this cutpoint, patients can be divided into two subgroups based on whether their age is older than 67.39 years or not. From Table 3.7, there is a statistically significant difference in treatment effect between these two age subgroups (\( p = 0.028 \) for the interaction between treatment and subgroup), which means age might be a predictive biomarker for the treatment benefit in global QoL for advanced CRC patients treated with cetuximab.

To understand the effects of treatment in these two subgroups, I plotted in Figure 3.1 the lowess curves for the scores of global health status over time respectively for patients in these two subgroups. As shown in Figure 3.1, in comparison with BSC alone, cetuximab plus BSC resulted in an apparently better QoL for younger patients but worse QoL for older patients. To investigate these apparently differential treatment effects further, I have fitted a linear mixed model with time and treatment as covariates separately for patients in these two subgroups. There were 291 (65.7\%) with age \( \leq 67.39 \) years, of whom 169 (58.1\%) of 291 patients received cetuximab treatment, while there were 152 patients with age \( > 67.39 \) years, of whom 71 were
### 3.4. APPLICATIONS

**Table 3.7: Subgroup analysis of quality of life data in CO.17**

<table>
<thead>
<tr>
<th></th>
<th>Estimation</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>all patients</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>3.933</td>
<td>0.052</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Time ($\beta_1$)</td>
<td>-0.013</td>
<td>0.005</td>
<td>.012</td>
</tr>
<tr>
<td>(Time-8)$_+$ ($\beta_2$)</td>
<td>0.002</td>
<td>0.009</td>
<td>.838</td>
</tr>
<tr>
<td>(Time-16)$_+$ ($\beta_3$)</td>
<td>0.012</td>
<td>0.012</td>
<td>.320</td>
</tr>
<tr>
<td>Treatment: cetuximab ($\beta_4$)</td>
<td>0.243</td>
<td>0.062</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Age $&gt;$ c ($\eta_1$)</td>
<td>0.089</td>
<td>0.085</td>
<td>.296</td>
</tr>
<tr>
<td>Treatment: Age $&gt;$ c ($\eta_2$)</td>
<td>-0.250</td>
<td>0.114</td>
<td>.028</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.765</td>
<td>0.023</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\rho^2$</td>
<td>0.293</td>
<td>0.013</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td></td>
<td>0.601</td>
<td>0.082</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

**age $\leq$ 67.39**

<table>
<thead>
<tr>
<th></th>
<th>Estimation</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>3.924</td>
<td>0.050</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Time ($\beta_1$)</td>
<td>-0.011</td>
<td>0.006</td>
<td>0.049</td>
</tr>
<tr>
<td>(Time-8)$_+$ ($\beta_2$)</td>
<td>-0.002</td>
<td>0.010</td>
<td>.872</td>
</tr>
<tr>
<td>(Time-16)$_+$ ($\beta_3$)</td>
<td>0.016</td>
<td>0.013</td>
<td>.215</td>
</tr>
<tr>
<td>Treatment: cetuximab ($\beta_4$)</td>
<td>0.249</td>
<td>0.060</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

**age $>$ 67.39**

<table>
<thead>
<tr>
<th></th>
<th>Estimation</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>4.047</td>
<td>0.073</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Time ($\beta_1$)</td>
<td>-0.016</td>
<td>0.010</td>
<td>.117</td>
</tr>
<tr>
<td>(Time-8)$_+$ ($\beta_2$)</td>
<td>0.010</td>
<td>0.018</td>
<td>.574</td>
</tr>
<tr>
<td>(Time-16)$_+$ ($\beta_3$)</td>
<td>0.004</td>
<td>0.013</td>
<td>.856</td>
</tr>
<tr>
<td>Treatment: cetuximab ($\beta_4$)</td>
<td>-0.027</td>
<td>0.089</td>
<td>.761</td>
</tr>
</tbody>
</table>

Note: The treatment variable is coded as 1 if a patient received cetuximab plus BSC and 0 if treated by BSC alone.

Treated with cetuximab. The results of the stratified analysis for these two groups of patients are also reported in Table 3.7. From these results, one can see that, in comparison with those treated with BSC alone, younger patients ($\leq 67.39$ years) received cetuximab with BSC had a statistically significant improvement in global health-status QoL ($p < .0001$), but no significant difference in global health-status QoL between the two treatment groups were observed for older patients ($p = .761$).
3.4. APPLICATIONS

![Graph](image1.png)

Figure 3.1: Mean of QoL score (in log-transformation) over time in global health-status for older patients [cetuximab vs BSC] and younger patients [cetuximab vs BSC]

Finally, I employed a residual bootstrap method (Flores-Agreda and Cantoni, 2019; Field and Welsh, 2007) to assess the stability of the classifying rule derived based on the cutpoint identified from the proposed model. For each bootstrap sample drawn randomly with replacement from a total of 443 patients, patients were divided into two age subgroups based on the cutpoint 67.23. A linear mixed model with time, treatment, subgroup indicator, and an interaction term between age-subgroup status and treatment were then fitted to the sample. Among 1000 bootstrap samples, a statistically significant interaction effect between age-subgroup status and treatment was found in 74.5% samples, which shows that the classifying rule is reasonably stable.

3.4.2 Analysis of data from ACTG 193A

In the AIDS Clinical Trial Group (ACTG) study 193A, 1309 patients with advanced immune suppression (CD4 counts of \( \leq 50 \) cells/mm\(^3\)) (Henry et al., 1998) were randomized to receive the triple therapy (zidovudine plus 400mg of zalcitabine plus 400mg of nevirapine) or non-triple therapy (the double therapy and the alternating therapy). CD4 count, the number of T-lymphocyte cells in the body, is directly
Table 3.8: Subgroup analysis of CD4 count data in 193A

<table>
<thead>
<tr>
<th></th>
<th>Estimation</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>all patients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>2.609</td>
<td>0.081</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Time ($\beta_1$)</td>
<td>-0.001</td>
<td>0.002</td>
<td>.719</td>
</tr>
<tr>
<td>(Time-16)$_+$ ($\beta_2$)</td>
<td>-0.018</td>
<td>0.003</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Treatment: triple therapy ($\beta_3$)</td>
<td>0.427</td>
<td>0.151</td>
<td>0.004</td>
</tr>
<tr>
<td>Age$&gt;$ c ($\eta_1$)</td>
<td>0.352</td>
<td>0.084</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Treatment: Age$&gt;$ c ($\eta_2$)</td>
<td>-0.235</td>
<td>0.161</td>
<td>0.143</td>
</tr>
<tr>
<td>Threshold (c)</td>
<td>0.391</td>
<td>0.005</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.451</td>
<td>0.011</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>$\rho^2$</td>
<td>1.000</td>
<td>0.047</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Note: The treatment variable is coded as 1 if a patient received the triple therapy and 0 if treated by the non-triple therapy.

affected by the HIV virus and is one of the primary outcomes in this study. A CD4 count is between 800 to 1000 for a normal adult, however, CD4 counts of the patients in this trial ranged from 0 to 50 (cells/mm$^3$) at baseline. Measurements of CD4 counts were collected from each patient at baseline and at 8-week intervals during follow-up in this study.

As in the last example, a log-transformation was applied to the outcomes and the baseline covariate age was re-scaled on $(0, 1]$ before the analysis. Since the sample standard deviation of age is around 0.11 and the total number of observations is 4914, the bandwidth $h$ is chosen as 0.006 based on the approach mentioned in Section 3.3. The time effect was also modeled by a piece-wise linear spline with a knot at week 16. That is, the following threshold linear mixed model was fitted:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 16)_+ + \beta_3 b_i + \eta_1 I(\text{age} > c) + \eta_2 b_i I(\text{age} > c) + \varepsilon_{ij}.$$  

The estimates of the parameters in the model are presented in Table 3.8.
The estimated cutpoint of age is \( c = 0.391 \), which is 29.0 years old after scaling back to the original age scale. With this cutpoint, patients can be divided into two different age subgroups based on whether the patient’s age is older than 29.0 or not. From Table 3.8 one can find, however, there was no statistically significant difference in treatment effects between the two age subgroups defined by this cutpoint \( (p = 0.143 \) for the interaction between the treatment and the subgroup), which implies that age may not be a predictive biomarker for the benefit in CD4 count for advanced AIDS patients treated with the triple therapy.

### 3.5 Conclusions and Discussions

In this chapter, a general threshold linear mixed model is proposed for identification of treatment-sensitive subsets based on longitudinal data. This model is very flexible and can include different effect of time and other covariates as both fixed and random effects. Statistical procedures are proposed based on smooth approximation of an indicator function to estimate the unknown parameters in the proposed model. Efficient computational algorithms are developed to compute the estimators and associated variances. Results from simulation studies showed the proposed procedure performed very well and the application to the analysis of data from two clinical trials demonstrates that the proposed procedure is practical and stable.

Our model assumes the outcomes from the clinical trials are normally distributed. Although transformations, such as the logarithm transformation used in our examples, may make the distribution of the outcomes close to a normal distribution, interpretations from the models with transformed outcomes may not be straightforward. Recently, longitudinal beta regression models [Hunger et al., 2012], which assume the
outcomes are restricted to an interval, were proposed to analyze quality of life scores. Threshold models considered in this paper may be defined based on these models to provide a better fit to the QoL data. However, computations required to fit these models may be more challenging since they are in the class of generalized linear mixed models, which require approximations to the smoothed likelihood functions.

As for the general linear mixed models, the inference procedures proposed for the model introduced in this chapter are valid under the assumption that data are missing at random (MAR). Simulation studies confirmed the robustness of the proposed procedures under the MAR assumption. In some applications, the observations may not be missing at random. For example, the quality of life score from a patient may be missing because of recurrence or death. Handling of data which may be missing not at random is a very difficult problem in the analysis of longitudinal data. Sensitivity analyses based on pattern mixture models or joint analysis of longitudinal outcomes and survival times may be used to assess the robustness of the results but these methods have not yet been extended to the threshold models considered in this paper, which would be interesting topics for future research.

I considered in this chapter the case where there is only a single covariate which is of interest or available for the definition of subgroups. In practise, there may be multiple covariates or biomarkers available. Combining them may improve subgroup identification. As reviewed in the introduction, various data-driven statistical methods have been proposed for subgroup identification based on multiple covariates when the outcomes from a clinical trial are continuous, binary or censored survival data. Recently, He et al. (2018) considered the use of a linear combination of multiple biomarkers for subgroup identification through a threshold Cox proportional hazards model.
model and used a penalized smooth likelihood procedure to select covaraites and estimate the parameters in the model. This model may be generalized to longitudinal outcomes and multiple covariates. A smoothed penalized profile likelihood which combines the approach proposed by Fan and Li (2012) with a smoothed indicator function may be used to select covariates and make inferences on the parameters in the model.
Chapter 4

A Threshold Mixed-Effects Tobit Model

4.1 Introduction

In previous chapter, I introduced a threshold linear mixed model which can be used to identify a subgroup of patients who may be sensitive to a specific treatment based on a continuous covariate and longitudinal continuous outcomes. This model assumes that, conditional on a random effect, the longitudinal measurements from the same patients are independent and follow a normal distribution.

When the longitudinal outcomes are obtained from the answers of patients to a question in a specific questionnaire, a problem with the use of the threshold linear mixed model is the potential ceiling and/or floor effects caused by a large portion of individuals who report either the maximum or minimum score in the assessments. For example, in the CO.17 trial mentioned in previous chapter, there are four possible answers to the first question ‘Do you have any trouble doing strenuous activities, carrying a heavy shopping bag or a suitcase?’ and the second question ‘Do you have any trouble taking a long walk?’ of EORTC QLQ-C30: not at all, a little, quite a little, and very much. It was found that about 19% of patients reported ‘not at all ’ to
both the first and second questions, while about 9% of patients answered ‘very much’ to these two questions. In this case, the estimates of parameters in the threshold linear mixed model may be biased because the normal distribution assumption for the longitudinal outcomes may be violated. To address this issue in the context of the conventional linear mixed models, [Twisk and Rijmen (2009)] proposed a longitudinal Tobit regression model as an alternative approach to analyze longitudinal outcomes with potential floor and ceiling effects. There are also some other papers in the literature which considered the use of Tobit models for the analysis of longitudinal outcomes with potential ceiling effects [Cowles et al., 1996, Dagne and Huang, 2012, Wang and Griswold, 2016, Ye et al., 2018, Sayers et al., 2020].

In this chapter, I propose a threshold mixed-effects Tobit model which can be used to identify a treatment-sensitive subgroup based on a continuous covariate and longitudinal continuous outcomes with floor and ceiling effects. Because of an unknown indicator function in the definition of the model, a smoothed likelihood function is also adopted for the estimation of parameters in the model. But, unlike the smoothed likelihood function defined in the previous chapter for the threshold linear mixed model, the smoothed likelihood function for the proposed threshold mix-effects Tobit model involves a complicated integration which dose not have a closed form and, therefore, it is very difficult to obtain the estimates of the parameters in the model by directly maximizing the smoothed likelihood function. A multidimensional Gauss-Hermite quadrature method is proposed to approximate the smoothed likelihood function and obtain the estimates of parameters. A randomly weighting method, which is more flexible than the bootstrap method, is proposed to estimate the variances for the parameter estimates. The proposed procedures is assessed through simulation studies.
and an application to the data from a randomized clinical trials on patients with advanced colorectal cancer.

The remainder of this chapter is organized as follows. Section 4.2 introduces the threshold mixed-effects Tobit model and statistical procedures for the inference of the parameters in the model. The results of simulation studies evaluating the performance of the proposed inference procedures are presented in Section 4.3. Section 4.4 presents the results from an application of the proposed method to the analysis of data in a randomized clinical trial on patients with advanced colorectal cancer. This chapter is concluded with some discussions in Section 4.5.

4.2 Methodology

4.2.1 The Model and Notations

Assume $y_{ij}^*$ is a latent response of the $i$-th subject measured at the $j$-th time-point, where $j = 1, \ldots, n_i$ and $i = 1, 2, \ldots, N$. Assume $b_i$ is an indicator of the treatment received by subject $i$ with either $b_i = 1$ if the patient is receiving a new therapy or $b_i = 0$ if not. Denote $w_i$ as a continuous covariate at baseline for subject $i$ and assume two subgroups of subjects can be defined based on whether $w_i$ exceeds an unknown cutpoint $c$. For each subject, denote $X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})'$ as a $(n_i \times p)$ designed matrix of other covariates (such as intercept, time) for fixed effect $\beta$ and $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{in_i})'$ as a $(n_i \times q)$ designed matrix of covariates for random effect $\alpha_i$. If all $y_{ij}^*$ are observed completely, then the following threshold linear mixed model as introduced in the last chapter can be used to identify and assess treatment-sensitive
subgroup:

\[ y_{ij}^* = X_{ij}' \beta + Z_{ij}' \alpha_i + \eta_1 I(w_i > c) + \eta_2 b_i I(w_i > c) + \epsilon_{ij}. \]

However, in practice, \( y_{ij}^* \) may only be observed in a pre-specified interval \([l, u]\) (\(l\) and \(u\) are some known constants) because of the potential floor and ceiling effects. Assume \( y_{ij} \) is corresponding observation of \( y_{ij}^* \) restricted in the interval \([l, u]\). Then following the approach proposed by Twisk and Rijmen (2009), the following threshold mixed-effect Tobit model may be used to assess the potential differential treatment effects between these two subgroups based on the observed responses:

\[
y_{ij}^* = X_{ij}' \beta + W_i \eta + Z_{ij}' \alpha_i + \epsilon_{ij},
\]

\[
y_{ij} = \begin{cases} 
  l & \text{if } y_{ij}^* \leq l \\
  y_{ij}^* & \text{if } l < y_{ij}^* < u \\
  u & \text{if } y_{ij}^* \geq u 
\end{cases}
\]  

(4.1)

where \( W_i = (I(w_i > c), I(w_i > c) b_i)' \) and \( \eta = (\eta_1, \eta_2)' \). If there exists only flooring or ceiling effect, one can set \( u = +\infty \) or \( l = -\infty \).

In model (4.1), the unobserved response \( y_{ij}^* \) of subject \( i \) measured at the time \( t_{ij} \) is modeled by three components: the fixed effects of all covariates \( X_{ij}' \beta + W_i \eta \), the random subject effect \( Z_{ij}' \alpha_i \), and the random error \( \epsilon_{ij} \). The elements of \( X_{ij} \) may include intercept, time or its function, treatment, and other confounding variables, and the elements of \( Z_{ij} \) are assumed as a subset of the elements of \( X_{ij} \). As in the conventional linear mixed model, it is assumed that random effects \( \alpha_i \) are independent
4.2. METHODOLOGY

of the random errors $\varepsilon_{ij}$ and

$$\alpha_1, \cdots, \alpha_i \overset{i.i.d.}{\sim} \text{MVN}_p(0, \Sigma),$$

$$\varepsilon_{i1}, \cdots, \varepsilon_{im_i} \overset{i.i.d.}{\sim} \text{N}(0, \sigma_\varepsilon^2).$$

4.2.2 Estimation of Model Parameters

In order to simplify the presentation, define $\theta = (\beta, \eta, c, \sigma_\varepsilon, \Sigma)$ and then under the model assumption, the likelihood function of $\theta$ is given by

$$L(\theta; y_{ij}) = \prod_{i=1}^{N} \int_{\alpha_i} \left\{ \frac{\phi \left( \frac{l - X_{ij}' \beta - W_i' \eta - Z_{ij}' \alpha_i}{\sigma_\varepsilon} \right)}{\phi \left( \frac{X_{ij}' \beta + W_i' \eta + Z_{ij}' \alpha_i}{\sigma_\varepsilon} \right)} \right\}^{I_{ij}^l} \left[ \frac{1}{\sigma_\varepsilon} \phi \left( \frac{y_{ij} - X_{ij}' \beta - W_i' \eta - Z_{ij}' \alpha_i}{\sigma_\varepsilon} \right) \right]^{(1 - I_{ij}^l - I_{ij}^u)} f(\alpha_i) d(\alpha_i),$$

where $f(\cdot)$ is the density function of $\alpha_i$, $\Phi(\cdot)$ and $\phi(\cdot)$ denote the commutative distribution function and the probability density function, respectively, of the standard normal distribution, and $I_{ij}^l$ and $I_{ij}^u$ are indicator functions with

$$I_{ij}^l = \begin{cases} 1 & \text{if } y_{ij} = l, \\ 0 & \text{if } y_{ij} > l, \end{cases}$$

$$I_{ij}^u = \begin{cases} 1 & \text{if } y_{ij} = u, \\ 0 & \text{if } y_{ij} < u. \end{cases}$$

Same as the threshold linear mixed model defined in the last chapter, due to the presence of the indicator functions $I(w_i > c)$, the likelihood function is not
4.2. METHODOLOGY

continuous with respect to \( c \), which makes the conventional maximum likelihood theory and algorithm difficult to apply. Similarly, following a smoothing procedure used by Brown and Wang (2007), a kernel smooth function

\[
\Phi \left( \frac{w_i - c}{h} \right)
\]

(4.3)
can be used as a smooth approximation to the indicator function \( I(w_i > c) \), where \( h \) is a bandwidth which converges to zero as the sample size increases. Note that if \( w_i > c \), \( \Phi((w_i - c)/h) \to 1 \) as \( n \to \infty \), while \( \Phi((w_i - c)/h) \to 0 \) as \( n \to \infty \) when \( w_i < c \), that is, \( \Phi((w_i - c)/h) \to I(w_i - c) > 0 \) as \( n \to \infty \). Using this approximation, a smoothed likelihood function can be defined by replacing \( W_i \) in (3.3) with \( \tilde{W}_i = (\Phi((w_i - c)/h), b_i\Phi((w_i - c)/h))' \). Specifically, the smoothed likelihood function is given by

\[
SL(\theta; y_{ij}) = \prod_{i=1}^{N} SL_i(\theta; y_{ij})
\]

\[
= \prod_{i=1}^{N} \int_{\alpha_i} \left\{ \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{l - X_{ij}'\beta - \tilde{W}_i'\eta - Z_{ij}'\alpha_i}{\sigma}\right) \right]^{I_{ij}} \right. \\
\left. \Phi \left( \frac{X_{ij}'\beta + \tilde{W}_i'\eta + Z_{ij}'\alpha_i - u}{\sigma} \right) \right]^{I_{ij}} \\
\left[ \frac{1}{\sigma} \phi \left( \frac{y_{ij} - X_{ij}'\beta - \tilde{W}_i'\eta - Z_{ij}'\alpha_i}{\sigma} \right) \right]^{(1-I_{ij}-I_{ij})} \} f(\alpha_i) d(\alpha_i).
\]

The smoothed likelihood function defined above includes a complicated integration which does not have a closed form and, therefore, it is very difficult to obtain
the estimation of $\theta$ by directly maximizing it. Various approaches, such as Gauss-Hermite quadrature method, EM algorithm, Markov Chain Monte Carlo (MCMC), Bayesian approaches can be used. Here I apply the multidimensional Gauss-Hermite quadrature [Jäckel, 2005] to numerically calculate the integrals in the likelihood function (4.4). Assuming $SL_i^*(\theta; y_{ij})$ is the Gauss-Hermite quadrature approximation of $SL_i(\theta; y_{ij})$, estimates of of $\theta$ can be obtained by

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^{N} \log(SL_i^*(\theta; y_{ij})).$$  (4.5)

In practice, Gauss-Hermite quadrature approximation can be obtained using function ‘ghq’ of the R package glmmML (Broström, 2009) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), which belongs to quasi-Newton methods, can be used to calculate $\hat{\theta}$ defined by (4.5) via BFGS option in R package maxLik (Henningsen and Toomet, 2011).

4.2.3 Random Weighting Estimates for the Variances of the Parameter Estimators

Since it is difficult to derive an analytic expression for the variances of the parameter estimators defined by (4.5), I use a random weighting method to estimate the variance of parameter estimators, which may be more stable than the bootstrap method. This is because the flooring and ceiling effects of longitudinal outcomes from the bootstrap samples may be more serious due to potential repeated sampling of the minimum or maximum scores (Cui et al., 2008). The basic idea of random weighting method can be found in Chapter 10 of Shao and Tu (1995). This method has also been called as perturbation method (Jin et al., 2001; Minnier et al., 2011). There are several papers
which have applied this method to outcomes with floor or ceiling effect to obtain variance estimates (Wang et al., 2009; Xiao et al., 2014).

To obtain random weighting estimates for the variances of parameter estimators, I first define, for the model proposed in the last section, the following random weighting estimate (RWE) of $\theta$:

$$
\tilde{\theta} = \arg \max_\theta \sum_{i=1}^{N} \omega_i \log(SL_i^*(\theta; y_{ij})),
$$

where $\{\omega_i : i = 1, \cdots, N\}$ is a sequence of independent and identically distributed random variables with mean 1 and variance 1 (Wang et al., 2009; Cai and Zheng, 2013). Similar algorithms introduced in the last section can be used to calculate $\tilde{\theta}$ defined in (4.6). A random weighting estimator for the variance of $\hat{\theta}$ can be obtained by generating $B$ independent sequences of random variables $\{\omega_i : i = 1, \cdots, N\}$ and calculating the corresponding random weighting estimate of $\theta$ for each sequence. Assume $\{\tilde{\theta}_1, \cdots, \tilde{\theta}_B\}$ are $B$ random weighting estimates. The variance of the parameter estimator $\hat{\theta}$ can be estimated by the sample variance of $\{\tilde{\theta}_1, \cdots, \tilde{\theta}_B\}$.

4.3 Simulation Studies

The performance of the procedures introduced in the last section is evaluated using simulated longitudinal data with potential floor and ceiling effects. All data in the simulation studies are generated from Model (4.1) with three covariates ($p = 3$): intercept, time $t_{ij}$, and treatment indicator $b_i$. In this case, Model (4.1) can be
rewritten as the following:

\[ y_{ij}^* = \beta_0 + \beta_1 t_{ij} + \beta_2 b_i + \eta_1 I(w_i > c) + \eta_2 b_i I(w_i > c) + \alpha_i + \varepsilon_{ij}, \]

\[ y_{ij} = \begin{cases} l & \text{if } y_{ij}^* \leq l \\ y_{ij}^* & \text{if } l < y_{ij}^* < u \\ u & \text{if } y_{ij}^* \geq u. \end{cases} \quad (4.7) \]

In all the simulations, the regression coefficients except the intercept $\beta_0$ in the model are set to respectively $\beta_1 = 2$, $\beta_2 = -2$, $\eta_1 = 1$, and $\eta_2 = 1$, and the variance of random error $\varepsilon_{ij}$ and random effect $\alpha_i$ are assumed as respectively $0.8^2$ and $0.4^2$. A total of $n_i = 4$ longitudinal measurements are assumed to be observed from each subject at equal-spaced timepoints from 0 to 2. The treatment indicator and the biomarker value of each subject are generated respectively from a Bernoulli distribution with probability 0.5 and a uniform distribution $U[0, 1]$.

To evaluate the impact of floor and ceiling effects on the proposed parameter estimators and associated variance estimators, four different cases with different proportions of patients who had a minimum or maximum scores are considered by varying the intercept $\beta_0$, and the lower and upper limits $l$ and $u$. Specifically, in the first case where $\beta_0 = 4$, $l = 0$ and $u = 6$, both the proportions of patients with minimum and maximum scores are around 7%. In the second case where the upper limit $u$ is changed to 5, the proportions of patients with minimum and maximum scores are respectively around 7% and 17%, while these proportions are reversed to respectively 17% and 7% in the third case where $\beta_0$ is changed to 3 but $l$ and $u$ are kept as respectively 0 and 6. Finally, in the last case where the lower limit $l$ is changed to 0.25 and the upper limit $u$ to 3.5 with $\beta_0 = 3$, both the proportions of patients with the
floor and ceiling effects are relative high and around 22%.

For each case defined above, simulation is repeated 1000 times, and the value of the bandwidth $h$ is calculated by $\hat{\rho}n^{-1/3}$ (He et al. 2018), where $\hat{\rho}$ is the sample standard deviation of the biomarker values and $n$ is the total number of longitudinal measurements. The Gauss-Hermite quadrature approximation to the smooth likelihood function and the corresponding gradients implemented in the algorithm for the model considered in the simulation studies are given in the Appendix C.

I first evaluated the performance of the proposed estimation procedure for the parameters in (4.7) when the number of subjects is $N = 100$ in terms of relative absolute bias (RBIAS), standard deviation (SD), and root mean square error (RMSE) of parameter estimators and misclassification rate (MR) of subgrouping. As a comparison, I also calculated these performance measures for the parameter estimators calculated from the threshold linear mixed model introduced in Chapter 3, which ignores the potential floor and ceiling effects in the longitudinal outcomes.

The simulation results are summarized in Table 4.1, where proposed and Method 1 refer to the methods based on respectively the models proposed in this and last chapter. From the results in this table, one can see that the relative absolute value of the bias for proposed estimators is much smaller than that of the estimators obtained under threshold linear mixed model for all parameters except the cutpoint $c$ but the RMSE of the proposed estimators is much smaller for all parameters. As expected, for all parameters, as floor or ceiling effect becomes large, the relative bias and RMSE of the estimators increase for both methods but the increase is smaller for the proposed method. Furthermore, the misclassification rate of the proposed method is smaller than that of the method obtained from the threshold linear mixed model. Overall, our
Table 4.1: Simulation results on the relative bias and RMSE of the parameter estimators when sample size is 100 by using the proposed method (labeled as P) and the threshold linear mixed model (labeled as T).

<table>
<thead>
<tr>
<th>Method</th>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \eta_1 )</th>
<th>( \eta_2 )</th>
<th>( c )</th>
<th>( \sigma_0 )</th>
<th>( \sigma_1 )</th>
<th>( \sigma_2 )</th>
<th>( \sigma_3 )</th>
<th>MR (10^{-2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBIAS</td>
<td>0.004</td>
<td>0.005</td>
<td>0.001</td>
<td>0.057</td>
<td>0.066</td>
<td>0.012</td>
<td>0.003</td>
<td>0.017</td>
<td>0.079</td>
<td>0.079</td>
<td>0.369</td>
</tr>
<tr>
<td>SD</td>
<td>0.135</td>
<td>0.151</td>
<td>0.098</td>
<td>0.269</td>
<td>0.365</td>
<td>0.013</td>
<td>0.046</td>
<td>0.079</td>
<td>0.079</td>
<td>0.079</td>
<td>3.500</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.136</td>
<td>0.151</td>
<td>0.098</td>
<td>0.275</td>
<td>0.371</td>
<td>0.015</td>
<td>0.046</td>
<td>0.079</td>
<td>0.079</td>
<td>0.079</td>
<td>0.105</td>
</tr>
<tr>
<td>RBIAS</td>
<td>0.154</td>
<td>0.323</td>
<td>0.752</td>
<td>0.312</td>
<td>0.650</td>
<td>0.011</td>
<td>0.489</td>
<td>0.392</td>
<td>0.308</td>
<td>0.308</td>
<td>3.000</td>
</tr>
<tr>
<td>SD</td>
<td>0.140</td>
<td>0.140</td>
<td>0.052</td>
<td>0.676</td>
<td>0.932</td>
<td>0.104</td>
<td>0.280</td>
<td>0.392</td>
<td>0.308</td>
<td>0.308</td>
<td>3.500</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.483</td>
<td>0.660</td>
<td>0.752</td>
<td>0.744</td>
<td>1.136</td>
<td>0.104</td>
<td>0.280</td>
<td>0.392</td>
<td>0.308</td>
<td>0.308</td>
<td>0.105</td>
</tr>
</tbody>
</table>

floor and ceiling effects (21%, 21%) floor and ceiling effects (17%, 7%) floor and ceiling effects (7%, 17%) floor and ceiling effects (7%, 7%) floor and ceiling effects (7%, 7%)

The simulation studies demonstrate that our proposed method, which taking into account of potential floor and ceiling effects, is better than the method which ignores the potential floor and ceiling effects in the estimation of model parameters. The good performance of the method proposed in this chapter is also confirmed by the results in (4.2) which evaluated the performance of the proposed estimation procedure when the sample size is increased to \( N = 200 \) and \( N = 300 \).

I then evaluated the performance of the proposed random weighting estimators for the variance of the parameter estimators by comparing the standard deviation of the
4.3. SIMULATION STUDIES

Table 4.2: Simulation results on the relative bias and RMSE of the proposed parameter estimators when sample size is respectively 200 and 300.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>floor and ceiling effects (21%, 21%)</th>
<th>floor and ceiling effects (17%, 7%)</th>
<th>floor and ceiling effects (7%, 17%)</th>
<th>floor and ceiling effects (7%, 7%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₀</td>
<td>RBIAS 0.001 0.0003 0.0008 0.0006&lt;br&gt;SD 0.104 0.083 0.099 0.078&lt;br&gt;RMSE 0.104 0.083 0.099 0.078</td>
<td>0.0008 0.0006 0.100 0.079&lt;br&gt;0.100 0.079</td>
<td>0.0004 0.0008 0.098 0.077&lt;br&gt;0.098 0.077</td>
<td>0.0005 0.0009 0.098 0.077&lt;br&gt;0.098 0.077</td>
</tr>
<tr>
<td>β₁</td>
<td>RBIAS 0.006 0.008 0.007 0.008&lt;br&gt;SD 0.109 0.089 0.104 0.083&lt;br&gt;RMSE 0.109 0.089 0.105 0.085</td>
<td>0.007 0.008 0.102 0.082&lt;br&gt;0.103 0.084</td>
<td>0.009 0.009 0.100 0.087&lt;br&gt;0.101 0.087</td>
<td>0.009 0.009 0.100 0.087&lt;br&gt;0.101 0.087</td>
</tr>
<tr>
<td>β₂</td>
<td>RBIAS 0.002 0.001 0.0009 0.0009&lt;br&gt;SD 0.071 0.058 0.058 0.049&lt;br&gt;RMSE 0.071 0.058 0.058 0.049</td>
<td>0.001 0.0004 0.057 0.048&lt;br&gt;0.057 0.048</td>
<td>0.004 0.0003 0.054 0.046&lt;br&gt;0.054 0.046</td>
<td>0.003 0.0003 0.054 0.046&lt;br&gt;0.054 0.046</td>
</tr>
<tr>
<td>η₁</td>
<td>RBIAS 0.050 0.004 0.048 0.041&lt;br&gt;SD 0.133 0.110 0.131 0.107&lt;br&gt;RMSE 0.142 0.118 0.139 0.115</td>
<td>0.049 0.041 0.129 0.105&lt;br&gt;0.138 0.112</td>
<td>0.048 0.040 0.128 0.104&lt;br&gt;0.137 0.111</td>
<td>0.048 0.040 0.128 0.104&lt;br&gt;0.137 0.111</td>
</tr>
<tr>
<td>η₂</td>
<td>RBIAS 0.015 0.003 0.018 0.002&lt;br&gt;SD 0.205 0.167 0.187 0.153&lt;br&gt;RMSE 0.205 0.167 0.188 0.153</td>
<td>0.016 0.004 0.192 0.160&lt;br&gt;0.192 0.160</td>
<td>0.019 0.003 0.185 0.151&lt;br&gt;0.186 0.151</td>
<td>0.019 0.003 0.185 0.151&lt;br&gt;0.186 0.151</td>
</tr>
<tr>
<td>c</td>
<td>RBIAS 0.004 0.014 0.002 0.012&lt;br&gt;SD 0.008 0.005 0.007 0.005&lt;br&gt;RMSE 0.009 0.011 0.008 0.010</td>
<td>0.004 0.014 0.008 0.005&lt;br&gt;0.007 0.005</td>
<td>0.007 0.005 0.007 0.005&lt;br&gt;0.007 0.005</td>
<td>0.007 0.005 0.007 0.005&lt;br&gt;0.007 0.005</td>
</tr>
<tr>
<td>σε</td>
<td>RBIAS 0.0002 0.0005 0.002 0.002&lt;br&gt;SD 0.032 0.027 0.027 0.022&lt;br&gt;RMSE 0.032 0.027 0.027 0.022</td>
<td>0.001 0.0009 0.027 0.022&lt;br&gt;0.027 0.022</td>
<td>0.002 0.002 0.025 0.020&lt;br&gt;0.025 0.020</td>
<td>0.002 0.002 0.025 0.020&lt;br&gt;0.025 0.020</td>
</tr>
<tr>
<td>σα</td>
<td>RBIAS 0.007 0.022 0.0012 0.027&lt;br&gt;SD 0.052 0.042 0.047 0.037&lt;br&gt;RMSE 0.052 0.043 0.047 0.039</td>
<td>0.100 0.022 0.047 0.038&lt;br&gt;0.047 0.038</td>
<td>0.012 0.025 0.044 0.036&lt;br&gt;0.045 0.036</td>
<td>0.012 0.025 0.044 0.036&lt;br&gt;0.045 0.036</td>
</tr>
<tr>
<td>MR (10⁻³)</td>
<td>2.988 1.871 2.082 1.259&lt;br&gt;2.616 1.503&lt;br&gt;2.037 1.205</td>
<td>1.871 1.259&lt;br&gt;1.503&lt;br&gt;1.205</td>
<td>2.082 1.259&lt;br&gt;1.503&lt;br&gt;1.205</td>
<td>2.037 1.205</td>
</tr>
</tbody>
</table>
4.3. SIMULATION STUDIES

Table 4.3: Comparison between the standard deviation of proposed random weighting estimators for the variance of parameter estimators (SDE) and empirical standard deviation (ESD) of the parameter estimators when the sample size N=100

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ESD</th>
<th>SDE</th>
<th>ESD</th>
<th>SDE</th>
<th>ESD</th>
<th>SDE</th>
<th>ESD</th>
<th>SDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.135</td>
<td>0.136</td>
<td>0.132</td>
<td>0.128</td>
<td>0.128</td>
<td>0.130</td>
<td>0.131</td>
<td>0.127</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.151</td>
<td>0.145</td>
<td>0.136</td>
<td>0.137</td>
<td>0.140</td>
<td>0.133</td>
<td>0.132</td>
<td>0.132</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.098</td>
<td>0.096</td>
<td>0.079</td>
<td>0.081</td>
<td>0.079</td>
<td>0.079</td>
<td>0.073</td>
<td>0.075</td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>0.269</td>
<td>0.236</td>
<td>0.204</td>
<td>0.229</td>
<td>0.254</td>
<td>0.226</td>
<td>0.200</td>
<td>0.226</td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>0.365</td>
<td>0.328</td>
<td>0.302</td>
<td>0.296</td>
<td>0.344</td>
<td>0.309</td>
<td>0.300</td>
<td>0.292</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>0.013</td>
<td>0.016</td>
<td>0.011</td>
<td>0.015</td>
<td>0.012</td>
<td>0.015</td>
<td>0.011</td>
<td>0.015</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>0.046</td>
<td>0.044</td>
<td>0.039</td>
<td>0.037</td>
<td>0.038</td>
<td>0.037</td>
<td>0.036</td>
<td>0.035</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>0.079</td>
<td>0.080</td>
<td>0.068</td>
<td>0.072</td>
<td>0.068</td>
<td>0.070</td>
<td>0.064</td>
<td>0.067</td>
</tr>
</tbody>
</table>

In practice, some patients may drop out the study earlier, which may lead to some observations from these patients missing. I performed additional simulation studies to assess the robustness of our proposed procedures to the potential missing data. Specifically, in these simulations, the first measurement $y_{i1}$ is assumed non-missing for every subject $i$. A logistic regression model proposed in [Diggle and Kenward (1994)], which links the probability that the $k$-th observation from the $i$-th subject with the previous observations from the same subject, was used to generate a data set with around 15% of longitudinal outcomes missing at random for 100 subjects. The estimated bias, SD, RMSE from 100 replications are presented in Table 4.4. In comparison with the results in Table 4.1, one can find that the biases of the estimates are similar for most parameters but, because the effective size of the sample is reduced by the presence of missing observations, SD and RMSE of the estimates and the misclassification rate become a little bit larger, which is reasonable considering the amount of missing observations.
4.4. AN APPLICATION

Table 4.4: Results for simulation studies over 1000 replications when the observations may be missing at random.

<table>
<thead>
<tr>
<th>β₀</th>
<th>μ</th>
<th>σ</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rbias</td>
<td>0.006</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>SD</td>
<td>0.135</td>
<td>0.129</td>
<td>0.135</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.135</td>
<td>0.135</td>
<td></td>
</tr>
<tr>
<td>β₁</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.007</td>
<td>0.176</td>
<td>0.176</td>
</tr>
<tr>
<td>SD</td>
<td>0.161</td>
<td>0.150</td>
<td>0.144</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.144</td>
<td>0.144</td>
<td></td>
</tr>
<tr>
<td>β₂</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.007</td>
<td>0.176</td>
<td>0.176</td>
</tr>
<tr>
<td>SD</td>
<td>0.161</td>
<td>0.151</td>
<td>0.144</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.144</td>
<td>0.144</td>
<td></td>
</tr>
<tr>
<td>η₁</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.058</td>
<td>0.307</td>
<td>0.307</td>
</tr>
<tr>
<td>SD</td>
<td>0.292</td>
<td>0.263</td>
<td>0.213</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.296</td>
<td>0.268</td>
<td>0.213</td>
</tr>
<tr>
<td>η₂</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.057</td>
<td>0.393</td>
<td>0.393</td>
</tr>
<tr>
<td>SD</td>
<td>0.361</td>
<td>0.351</td>
<td>0.307</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.364</td>
<td>0.356</td>
<td>0.331</td>
</tr>
<tr>
<td>c</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.014</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td>SD</td>
<td>0.012</td>
<td>0.012</td>
<td>0.011</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.017</td>
<td>0.015</td>
<td>0.018</td>
</tr>
<tr>
<td>σ₀</td>
<td>μ</td>
<td>σ</td>
<td>RMSE</td>
</tr>
<tr>
<td>Rbias</td>
<td>0.005</td>
<td>0.056</td>
<td>0.056</td>
</tr>
<tr>
<td>SD</td>
<td>0.045</td>
<td>0.042</td>
<td>0.039</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.045</td>
<td>0.042</td>
<td>0.039</td>
</tr>
<tr>
<td>MR(10⁻³)</td>
<td>4.380</td>
<td>3.760</td>
<td>3.480</td>
</tr>
</tbody>
</table>

4.4 Analysis of Data from a Cancer Clinical Trial

I applied the proposed model and inference procedures to the analysis of the data from a randomized cancer clinical trial CO.17 conducted by Canadian Cancer Trials Group, which randomized 572 patients to receive a new treatment, cetuximab, with best supportive care (BSC) or best supportive care alone. As introduced in Chapter 3, quality of life, assessed using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire C30 (EORTC QLQ-C30) at baseline and 4, 8, 16 and 24 weeks after the start of treatment, is an important patient-oriented
outcome to assess the benefit of cetuximab. EORTC QLQ-C30 consists of a total 30 questions and each question is answered with a scale which is converted to a numerical score from 1 to 4 for first 28 questions and from 1 to 7 for the last two questions with the higher score representing the better QoL. In Chapter 3 global health-status score of the QoL, which is defined as the sum of scores from the last two questions on EORTC-QLQC30, was considered as an important outcome and a threshold linear mixed model was used to identify the treatment-sensitive subgroup based on the age of the patients and longitudinal global-health status scores.

A physical QoL subscore, which is defined as the sum of scores from the first two questions of EORTC QLQ-C30 and re-scaled in the range of \([0, 10]\) for the data analysis, is important to evaluate patients’ physical abilities in doing strenuous activities, such as taking a long walk or carrying a heavy bag. It is also of interest to identify a subgroup of patients who are sensitive to cetuximab based on the age of patients and longitudinal physical subscores. But as mentioned in Section 4.1 and evidenced by the plots of longitudinal physical subscores over time in Figure 4.1 many patients reported floor and ceiling scores and, therefore, it may be difficult to use the threshold linear mixed model proposed in Section 4.1 to identify treatment-sensitive subgroup.

To apply instead the threshold mixed-effects Tobit model proposed in this chapter, I first also re-scaled the the baseline covariate age on \((0, 1]\). Since the sample standard deviation of age is around 0.115 and the total number of observations from 544 patients is 1651, the bandwidth \(h\) is chosen as 0.01 based on the approach mentioned in Section 4.2. After examining the lowess time plots of longitudinal physical subscores in Figure 4.1, a piece-wise linear spline with two knots at week 8 and week 16 is used to model the time effect. That is, the following threshold mixed-effects
4.4. AN APPLICATION

Tobit model was used to fit the data:

\[ y_{ij} = \beta_0 + \beta_1 b_i + \beta_2 t_{ij} + \beta_3 (t_{ij} - 8)_+ + \beta_4 (t_{ij} - 16)_+ + \eta_1 I(\text{age} > c) \]
\[ + \eta_2 b_i I(\text{age} > c) + \alpha_i + \varepsilon_{ij}. \]

The estimates of parameters in this model and associated standard errors of the estimates by using the procedures introduced in Section 4.2 are presented in Table 4.5. From this table, one can find that the estimated cutpoint of age for the definition of subgroup is \( c = 0.750 \), which is around 66 years after scaled back to the original age scale. With this cut point, patients can be divided into two different age subgroups based on whether their age is older than 66 years old or not. There are 206 patients who are older than 66 years old, while 320 patients are younger than 66 years old. From Table 4.5, one can also find that there is a trend that the difference in treatment effects between two age subgroups is statistically significant (\( p = 0.073 \) for the
4.5. CONCLUSIONS AND DISCUSSIONS

Table 4.5: Results of analysis for the physical subscores in CO.17 trial

<table>
<thead>
<tr>
<th></th>
<th>Estimation</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept ($\beta_0$)</td>
<td>4.020</td>
<td>0.180</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Treatment: cetuximab ($\beta_1$)</td>
<td>-0.345</td>
<td>0.180</td>
<td>0.003</td>
</tr>
<tr>
<td>Week ($\beta_2$)</td>
<td>-0.060</td>
<td>0.050</td>
<td>0.116</td>
</tr>
<tr>
<td>Week &gt; 8 ($\beta_3$)</td>
<td>0.338</td>
<td>0.105</td>
<td>0.0006</td>
</tr>
<tr>
<td>Week &gt; 16 ($\beta_4$)</td>
<td>-0.383</td>
<td>0.133</td>
<td>0.002</td>
</tr>
<tr>
<td>Age &gt; c ($\eta_1$)</td>
<td>0.135</td>
<td>0.193</td>
<td>0.241</td>
</tr>
<tr>
<td>Treatment: Age &gt; c ($\eta_2$)</td>
<td>0.264</td>
<td>0.181</td>
<td>0.073</td>
</tr>
<tr>
<td>Cut-point (c)</td>
<td>0.750</td>
<td>0.154</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>2.395</td>
<td>0.124</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>$\sigma_\alpha$</td>
<td>3.180</td>
<td>0.193</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Note: The treatment variable is coded as 1 if a patient received cetuximab plus BSC and 0 if treated by BSC alone.

interaction between treatment and subgroup under the significant level 0.1), which indicates that age might be the predictive biomarker for the benefit in the physical subscore for advanced CRC patients treated with cetuximab.

To further investigate the difference in the treatment effects of cetuximab between the two age groups, I plotted lowess curve of physical subscores over time for the different treatment in the two age subgroups. As showed in Figure 4.2 while the physical ability in doing strenuous activities was similar for patients older than 66 years old, the difference becomes larger after 8 weeks of treatment between two treatment groups for patients younger than 66 years old with the worst physical QoL for patients received cetuximab treatment.

4.5 Conclusions and Discussions

In this chapter, a threshold mixed-effects model is introduced to identify treatment-sensitive subgroups based on longitudinal outcomes with floor and ceiling effects and
a smoothing estimation procedure is derived for estimation of the parameters in the model through Gauss-Hermit Quadrature approximation to the likelihood function. A random weighting method is utilized to obtain the estimate for the variance of the proposed parameter estimators. Simulation studies show that, when the floor and ceiling effects are present, the proposed method leads to estimators which are less biased and have smaller RMSE, in comparison with the method which ignores the potential floor and ceiling effects of the longitudinal outcomes. The application to the analysis of quality of life data from a randomized clinical trial demonstrates that the proposed procedure is practical and stable.

There are a few potential extensions for the method proposed in this chapter. For example, only a single covariate is considered in the proposed model but, in practice, multiple biomarkers may be available and combination of them may be
necessary. There are a few papers which considered subgroup identification and assessment based on multiple covariates. For example, He et al. (2018) introduced a single-index Cox model to identify a treatment-sensitive subgroup for each treatment based on linear combinations of multiple biomarkers and adapted penalty functions to select associated biomarkers. Li et al. (2018) proposed a multi-threshold change plane regression model to identify subgroups with different treatment effects based on a linear function of subgroup covariates. Extension of these models to longitudinal outcomes with potential floor and ceiling effects would be an interesting and important future research topic.
Chapter 5

A Generalized Single-Index Linear Threshold Model

5.1 Introduction

A patient’s response to a treatment is in general heterogeneous, which may depend on characteristics of the patient. These characteristics can be related to the demographics of the patient (e.g., age, race, and gender), disease (e.g., size and grade of the tumor), and biology (e.g., the mutation status and expression level of a gene) and will be referred to as covariates throughout this chapter. Using information from these covariates measured at the time of onset or progression of a disease to predict potential benefits a patient would get from a new treatment has been long recognized as an important step in the process of medical decision making (Byar and Green, 1980). The subgroup of patients who benefit from a specific treatment is called as a treatment-sensitive subgroup. As presented in a recent literature review (Rekkas et al., 2020), various regression modeling approaches have been proposed to assess the heterogeneity of treatment effect within a randomized clinical trial and identify
5.1. INTRODUCTION

treatment-sensitive subgroups, which would provide decision makers with more reliable and individualized information on the benefits of treatments.

In Chapters 3 and 4, statistical procedures were developed to identify treatment-sensitive subgroups based on one single covariate and longitudinal continuous outcomes. It is, however, sometimes difficult to find a single covariate which could capture all the heterogeneity of treatment effect and provide accurate prediction for the treatment benefits. Therefore, it often needs to combine data from multiple covariates and integrate them into a single index. For example, the popular Nottingham prognostic index (NPI) used to determine prognosis following surgery for a women with breast cancer was derived by combining the following three pathological characteristics of the patient, the size of the tumour (S), the number of involved lymph nodes (N), and the grade of the tumour (G), as a single index \( NPI = 0.2 \times S + N + G \) (Phung et al., 2019). Recently, a composite measure of risk was created by combining the above characteristics with additional covariates including age to determine a subgroup of breast cancer patients who would benefit most from the treatment with adjuvant aromatase inhibitors such as letrozole. This measure was shown to be better than individual covariates to inform treatment selection (Viale et al., 2011). The definition of this measure was based on combination of observed values of all the covariates with the estimated parameter values from a Cox proportional hazards model. It was shown that patients with higher values of this measure had more benefits from the treatment with letrozole based on a subpopulation treatment effect pattern plot (Bonetti and Gelber, 2000) and a formal test of treatment-by-continuous composite measure of risk interaction. To define the subgroup of patients who would be sensitive to the letrozole treatment, they, however, used a convenient approach.
which took the upper tertile of the distribution as the cut-point for the continuous composite measure. Recently, He et al. (2018) proposed a single-index threshold Cox proportional hazards model which can be used simultaneously to identify a subgroup of patients who may be sensitive to a treatment based on multiple covariates and test the treatment-by-subgroup interaction for a formal assessment the heterogeneity of treatment effect between the patients in this subgroup and the rest of patients. This model assumes the outcome of clinical trials is the time to an event, such as progression-free or overall survival.

In some clinical trials, longitudinal measurements collected at different timepoints before, during, and after the treatment are important outcomes for measuring the treatment effect. For example, monthly measurements of CD4 counts are important outcomes in a clinical trial assessing a new treatment for HIV patients (Goldman et al., 1996) and patient reported Quality of Life (QoL) outcomes are routinely collected at different timepoints in a cancer clinical trial to assess the potential benefits and harms of a new treatment from the perspectives of patients (Au et al., 2009). There are some statistical procedures proposed recently for the identification of treatment-sensitive subgroups with multiple covariates and longitudinal measurements. For example, Andrews (2017) proposed a method based on a predictive model for the probability that an individual will have a positive treatment effect estimated from random effect terms in a linear mixed model for the longitudinal measurements. An individual is considered as sensitive to the treatment if the predicted probability estimated from the observed values of the covariates using a supervised classification method is higher than 0.5, an ad hoc value without a formal statistical justification. Recently, a generalized linear mixed-model tree (GLMM tree) (Fokkema et al., 2018)
and an exploratory tool called as "interaction tree for longitudinal trajectories" (Wei et al., 2020) were proposed to discover subgroups of patients who may potentially benefit from the treatment without a formal assessment of the interaction between the treatment and identified subgroups.

Similar to He et al. (2018), I propose in this chapter an approach to simultaneously identify a treatment-sensitive subgroup and provide a formal test of the interaction between the subgroup and the treatment when the clinical outcomes are longitudinal measurements. Specifically, an index which is a linear combination of all the covariates is introduced and a treatment-sensitive subgroup is defined based on whether the value of the index is larger than 1. The weight assigned to each covariate in the linear combination is estimated by using a generalized single-index linear threshold model which includes a function of the time, treatment, subgroup indicator, and interaction between the treatment and subgroup indicator as the covariates and allows the longitudinal measurements to be both continuous and categorical. Because the subgroup indicator is not continuous which causes a problem in the adoption of conventional approaches in the estimation of parameters in the model, smoothed generalized estimating equations are proposed to make inferences on the parameters. The proposed method is evaluated through simulation studies and applications to the data from clinical trials on advanced pancreatic cancer.

The reminder of this chapter is organized as follows. Section 5.2 introduces the proposed generalized single-index linear threshold model and the statistical procedure for the inference of unknown parameters in the model. Section 5.3 presents results of simulation studies evaluating the performance of the proposed statistical procedure. An application to data from clinical trials on pancreatic cancer is presented in Section
5.2. Methodology

5.2.1 The Model and Notations

Suppose there are $N$ subjects enrolled in a clinical trial which randomizes the subjects into a new treatment or control group. For subject $i$, let $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ be a $(n_i \times 1)$ vector of repeated measurements observed from this subject, where $y_{ij}$ is the measurement made at time $t_{ij}$, $1 \leq j \leq n_i$. Assume that the marginal distribution of $y_{ij}$ is in an exponential distribution family with the following density function:

$$f(y_{ij}) = \exp \left\{ \frac{y_{ij} \theta_{ij} - a(\theta_{ij})}{\phi} + b(y_{ij}) \right\},$$

where $\phi$ is a scale parameter and $\theta_{ij}$ is a location parameter. Also, $a(\theta_{ij})$ and $b(\theta_{ij})$ both are the functions of $\theta_{ij}$. It is known from this assumption that the first two moments of $y_{ij}$ are given by

$$\mu_{ij} = E(y_{ij}) = a'(\theta_{ij}), \quad var(y_{ij}) = a''(\theta_{ij})\phi = \phi V(\mu_{ij}),$$

where $V(\cdot)$ is a variance function.

Assume there are $m$ important continuous or categorical covariates, such as age, gender, biomarkers, etc., measured at baseline for each subject before randomized into the clinical trial and denote $W_i = (W_{i1}, W_{i2}, \ldots, W_{im})^T$ as an $(m \times 1)$ vector for the values of these covariates observed from the $i$th subject. As mentioned in the introduction, our objective is to identify a subgroup of subjects who may benefit more from the new treatment than the rest of the subjects by linearly combining these
covariates. Let \( I_i^{(c)} = I \left( \sum_{k=1}^{m} \gamma_k W_{ik} > c \right) \) be a subgroup indicator which equals 1 when the \( i \)th subject is in the treatment-sensitive subgroup. If the coefficients \( \gamma_1, \cdots, \gamma_m \) and cutpoint \( c \) are known and, therefore, the subgroups are pre-specified, differential treatment effects between subjects in the treatment-sensitive subgroup and the rest of the subjects can be assessed by testing the significance of interaction between this subgroup indicator and a treatment indicator \( b_i \), defined as \( b_i = 1 \) if the subject \( i \) is assigned to a new treatment or \( b_i = 0 \) if assigned to a control, i.e., testing \( H_0 : \eta_2 = 0 \), in the following generalized linear model:

\[
g(\mu_i) = X_i \beta + \eta_1 I_i^{(c)} 1 + \eta_2 b_i I_i^{(c)} 1, \tag{5.1}
\]

where \( g(\mu_i) = (g(\mu_{i1}), \cdots, g(\mu_{in_i}))' \), \( g(\cdot) \) is a known link function, \( X_i = (x_{i1}, \cdots, x_{in_i}) \) is an \( n_i \times p \) design matrix for \( p \) fixed effects, which include intercept, time \( t_{ij} \) or its functions, treatment indicator \( Z_i \), and other potential confounding variables but exclude any variable related to the subgroup indicator \( I_i^{(c)} \), \( 1 \) is an \( n_i \times 1 \) vector of 1’s, and \( \beta = (1, \beta_0, \cdots, \beta_{p-1})' \) and \( \eta = (\eta_1, \eta_2)' \) are unknown parameters. In practice, \( \gamma_1, \cdots, \gamma_m \) and \( c \) are unknown and required to be estimated from the data. In this case, these parameters are, however, not identifiable since \( B\gamma_1, \cdots, B\gamma_m \) and \( Bc \) produce the same value of \( I_i^{(c)} \) for any constant \( B \). To overcome this identifiability issue, \( c \) is fixed at 1 in the definition of subgroup indicator (which is similar as the methods used in [He et al., 2018] and [Li et al., 2018]) and denote this new subgroup indicator as the following:

\[
I_i \triangleq I_i^{(1)} = I \left( \sum_{k=1}^{m} \gamma_k W_{ik} > 1 \right).
\]
With this definition, model (5.1) can be rewritten as

\[ g(\mu_i) = X_i \beta + \eta_1 I_i 1 + \eta_2 b_i I_i 1. \]  \hspace{1cm} (5.2)

This model is called as a Generalized Single-Index Linear Threshold Model. The unknown parameters in this model are \( \theta = (\beta', \eta', \gamma')' \), where \( \gamma = (\gamma_1, \cdots, \gamma_m)' \). A procedure for the inference of these parameters is introduced in the next subsection.

### 5.2.2 Inference of Model Parameters

Because of the correlations among the repeated measurements in \( Y_i \) from the same subject, I adopt a Generalized Estimating Equation (GEE) approach for the inference of unknown parameters \( \theta \) in Model (5.2).

Let \( R_i(\alpha) \) be a \((n_i \times n_i)\) working correlation matrix of \( Y_i \) with an \((s \times 1)\) vector of unknown parameters \( \alpha \) and denote \( A_i \) as a \((n_i \times n_i)\) diagonal matrix with \( V(\mu_{ij}) \) as the \( j^{th} \) diagonal element. Then the working covariance matrix of \( Y_i \) can be written as

\[ V_i(\alpha) = \phi A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}. \]  \hspace{1cm} (5.3)

According to Zeger et al. (1988) and Zeger and Liang (1986), in principle, the estimate of \( \theta \) can be obtained by solving the following generalized estimating equations:

\[ U(\theta) = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \theta} V_i^{-1}(\alpha)(Y_i - \mu_i) = 0, \]  \hspace{1cm} (5.4)

and the estimate is consistent even when the working correlation matrix \( R_i(\alpha) \) is
misspecified.

Due to the presence of the indicator function $I_i$, the above equations cannot be defined for parameters $\gamma$ and, therefore, the conventional GEE approach is difficult to apply. Following a smoothing procedure suggested by Brown and Wang (2007), I propose to use a kernel smooth function

$$\Phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right)$$

as a smooth approximation to the indicator function $I_i$, where $\Phi$ is the distribution function of the standard normal distribution and $h$ is a bandwidth which converges to zero as the sample size increases. Note that if $\sum_{k=1}^{m} \gamma_k W_{ik} > 1$, $\Phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right) \to 1$ as $N \to \infty$, while $\Phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right) \to 0$ as $N \to \infty$ when $\sum_{k=1}^{m} \gamma_k W_{ik} \leq 1$, that is, $\Phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right) \to I(\sum_{k=1}^{m} \gamma_k W_{ik} > 1)$ as $N \to \infty$. Let $I_i^* = \Phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right)$ and $\mu_i^* = g^{-1}(\mathbf{X}_i \beta + \eta_1 I_i^* \mathbf{1} + \eta_2 b_i I_i^* \mathbf{1})$. Smoothed generalized estimating equations (SGEEs) can then be defined by replacing $\mu_i$ in equation (5.4) with $\mu_i^*$ as following:

$$U^*(\theta) = \sum_{i=1}^{N} \frac{\partial \mu_i^*}{\partial \theta} V_i^{-1}(\alpha)(\mathbf{Y}_i - \mu_i^*).$$  \hspace{1cm} (5.6)

Considering the normal distribution assumption and the identity link function, the detailed form of (5.6) can be expressed as

$$U^*(\theta) = \sum_{i=1}^{N} \begin{pmatrix} D'_{1i} \\ D'_{2i} \\ D'_{3i} \end{pmatrix} V_i^{-1}(\alpha)(\mathbf{Y}_i - \mu_i^*) = 0.$$
5.2. METHODOLOGY

where

\[ \begin{align*}
D_{1i} & \triangleq \frac{\partial \mu_i^T}{\partial \beta} = \frac{\partial \mu_i^T}{\partial \beta} = X_i', \\
D_{2i} & \triangleq \frac{\partial \mu_i^T}{\partial \eta} = \left( \begin{array}{cccc}
I_i^* & I_i^* & \cdots & I_i^*
\end{array} \right)_{2 \times n_i}, \\
D_{3i} & \triangleq \frac{\partial \mu_i^T}{\partial \gamma} = \left( \frac{\partial (D_{2i} \eta)'}{\partial \gamma} \right)_{m \times n_i},
\end{align*} \]

and

\[ \frac{\partial (D_{2i} \eta)'}{\partial \gamma_k} = \frac{n_i}{h} W_{ik} \phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right) + \frac{n_i b_i W_{ik}}{h} \phi \left( \frac{\sum_{k=1}^{m} \gamma_k W_{ik} - 1}{h} \right) \]

and \( \phi(\cdot) \) is the density function of the standard normal distribution.

When \( h \) is sufficiently small, \( U^*(\theta) \) will be sufficiently close to \( U(\theta) \) and the estimate of \( \theta \) from \( U^*(\theta) \) will be close to that from \( U(\theta) \). However, an extremely small \( h \) can cause numerical instability in computations. Following the ideas of Lin et al. (2011) and He et al. (2018), I use the following formula to calculate the bandwidth \( h \):

\[ h = \hat{\sigma} n^{-1/3}, \quad (5.7) \]

where \( \hat{\sigma} \) is the minimum of the sample standard deviations of all covariates and \( n = \sum_{i=1}^{N} n_i \). 
The computation for $\theta$ based on equation (5.6) can be accomplished using a wrapper of an efficient spectral algorithm (DF-SANE) (Cruz and Raydan, 2003; La Cruz et al., 2006) for solving nonlinear systems of equations as implemented by the function BBsolve in the R package BB (Varadhan and Gilbert, 2009). With this algorithm, $\pm U^*(\theta)$ in equation (5.6) is used as search directions in a systematic way with one of the spectral coefficients as the step-length and a non-monotone line-search technique for global convergence. Because the estimating equations (5.6) depend only on the marginal mean and variance of $Y_i$ and the distribution of $Y_i$ is not specified, the estimation approach based on the estimating equations is also called as the Quasi Generalized Pseudo Maximum Likelihood (QGPML) method (Wedderburn, 1974; Ziegler and Vens, 2010). Similar to other QGPML methods, the initial value of $\theta$ required by the spectral algorithm can be obtained by employing a standard generalized linear model and ignoring the correlations among observations within the same subject, i.e., by solving the estimating equations (5.6) with $R_i = I$, where $I$ is the $(n_1 \times n_i)$ identity matrix.

Estimating equations (5.6) involve also two nuisance parameters, the scale parameter $\phi$ in the definition of variance of $y_{ij}$ and $\alpha$ in the working correlation matrix, which need to be estimated. Followed by Zhao and Prentice (1990), $\phi$ can be estimated by

$$\hat{\phi} = \frac{1}{(\sum_{i=1}^{N} n_i) - k \sum_{i=1}^{N} \sum_{j=1}^{n_i} \tilde{r}_{ij}^2},$$

where $\tilde{r}_{ij} = (y_{ij} - \mu_{ij}^*)/\sqrt{V(\mu_{ij}^*)}$ is the $ij^{th}$ Pearson residual and $k = p + m + 2$ is the total number of the parameters in the model, while $\alpha$ can be estimated by the
5.2. METHODOLOGY

following estimating equations:

\[ U(\hat{\alpha}) = \sum_{i=1}^{N} \left( \frac{\partial \xi_i}{\partial \hat{\alpha}} \right)'H_i^{-1}(T_i - \xi_i) = 0, \]

where

\[ T_i = (\tilde{r}_{i1}\tilde{r}_{i2}, \tilde{r}_{i2}\tilde{r}_{i3}, \cdots, \tilde{r}_{i(n_i-1)}\tilde{r}_{im_i})', \]

\[ H_i = \text{diag}(\text{VAR}((T_i)_j))_{q_i \times q_i}, \]

\[ \xi_i = E(T_i)_{q_i \times 1} \]

with \( q_i = \binom{n_i}{2} \). If an exchangeable correlation matrix is employed as a working correlation matrix, \( \alpha \) reduces to a single parameter \( \alpha \) which can be estimated by

\[ \hat{\alpha} = \frac{1}{\hat{\phi}} \sum_{i=1}^{N} \left\{ \frac{\sum_{j=1}^{n_i} \sum_{j' \neq j} \tilde{r}_{ij}\tilde{r}_{ij'}}{n_i \times (n_i - 1)} \right\}. \]

With the above approaches to determining the initial value of \( \theta \) and estimate \( \hat{\alpha} \) and \( \hat{\phi} \), (5.6) can be solved using the Newton-Raphson method in which the new estimate \( \hat{\theta} \) is used to update \( \hat{\alpha} \) and \( \hat{\phi} \) and the process iterates until convergence.

For simplicity, let \( \hat{\theta} = (\hat{\beta}', \hat{\eta}', \hat{\gamma}')' \) be the solution of the proposed smoothed GEE (5.6) and \( \theta \) is the true value of parameters. The Taylor expansion of \( U^*(\hat{\theta}) \) at \( \theta \) is

\[ \hat{\theta} - \theta = -\left( \frac{\partial U^*(\theta)}{\partial \theta} \right)^{-1} \sum_{i=1}^{N} \frac{\partial \mu_i^*}{\partial \theta} V_i^{-1}(\alpha)(Y_i - \mu_i^*) + o_p(N^{-1}). \]
5.2. METHODOLOGY

From the definition of $U^*(\theta)$ in equation (5.6), I have

$$
\frac{\partial U^*(\theta)}{\partial \theta} = \sum_{i=1}^{N} \frac{\partial \mu_i^T}{\partial \theta} V_i^{-1}(\hat{\alpha}) \left( - \frac{\partial \mu_i^*}{\partial \theta} \right) + \sum_{i=1}^{N} \frac{\partial^2 \mu_i^T}{\partial \theta^2} V_i^{-1}(Y_i - \mu_i^*).
$$

Denote the first term of the above equation as $-\Omega$. Since the second term of the above equation is of order $o_p(N^{-1})$ because $Y_i - \mu_i^*$ has an approximate mean as zero, it is known that $\hat{\theta} - \theta$ can be approximated by

$$
\Omega^{-1} \sum_{i=1}^{N} \left( \frac{\partial \mu_i^T}{\partial \theta} \right) V_i^{-1}(\hat{\alpha})(Y_i - \mu_i^*)
$$

and, therefore, the covariance matrix of $\hat{\theta}$ can be consistently estimated by the sandwich estimator

$$
V_{s administrators}(\hat{\theta}) = \Omega^{-1} \left\{ \sum_{i=1}^{N} \left( \frac{\partial \mu_i^T}{\partial \theta} \right) V_i^{-1}(\hat{\alpha}) Cov(Y_i) V_i^{-1}(\hat{\alpha}) \left( \frac{\partial \mu_i^T}{\partial \theta} \right)^T \right\} \Omega^{-1} \tag{5.8}
$$

Zeger and Liang (1986) proposed to estimate $Cov(Y_i)$ in equation (5.8) by $S_i S_i^T$, where $S_i = Y_i - \mu_i^*$. However, Pan (2001) argued that the covariance calculated within the sandwich estimator is not an optimal estimator of $Cov(Y_i)$ because it is based on the data from the $i$th subject only and, therefore, is not efficient. He proposed to replace $Cov(Y_i)$ with

$$
P_i = \phi A_i^{1/2} = A_i^{1/2} \left( \sum_{i=1}^{N} A_i^{-1/2}(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)^T A^{-1/2} / N \right) A_i^{1/2},
$$

which is a pooled, or averaged, covariance based on all subjects, in the sandwich estimator (5.8). I adopt this improved sandwich estimator to estimate the variances
of the estimated parameters.

In the next section, I will report results of simulation studies which assessed the performance of the proposed procedures for the estimation of the parameters in model (5.1) and the variances of the parameter estimators.

5.3 Simulation Studies

The data in the simulation studies are generated from Model (5.1). Without loss of generality, I consider the case where there are two covariates ($m = 2$) and $X_i$ contains the intercept, time $t_{ij}$ and treatment indicator $b_i$. In this case, Model (5.1) can be written as

$$g(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 b_i + \eta_1 I_i + \eta_2 b_i I_i,$$

where $I_i = I(\gamma_1 W_{i1} + \gamma_2 W_{i2} > 1)$. Two different link functions are considered, $g(\mu) = \mu$ and $g(\mu) = \log \frac{\mu}{1-\mu}$, to generate the continuous and binary longitudinal measurements, respectively. The regression coefficients and the nuisance parameter are set to respectively $\beta_0 = 1$, $\beta_1 = -1$, $\beta_2 = 1.5$, $\eta_1 = -1$, $\eta_2 = 1.5$, and $\phi = 0.64$ for the continuous longitudinal measurements and $\beta_0 = 1$, $\beta_1 = -1.5$, $\beta_2 = 1.5$, $\eta_1 = 1$, $\eta_2 = -1.5$, and $\phi = 1$ for binary longitudinal measurements. Two sets of parameters $(\gamma_1, \gamma_2) = (1.5, 1)$ and $(-1, 1.5)$ are considered, which lead to respectively two balanced subgroups with roughly equal number of subjects in two subgroups and two unbalanced subgroups with roughly one quarter of subjects in the subgroup with $I_i = 1$ and three quarters in the subgroup with $I_i = 0$. Equal-spaced longitudinal measurements are assumed from each subject with the total number of measurements $n_i = 4$ when $(\gamma_1, \gamma_2) = (1.5, 1)$ and $n_i = 5$ when $(\gamma_1, \gamma_2) = (-1, 1.5)$ and an
exchangeable correlation $\alpha = 0.2$ is assumed among these measurements. In all the simulations, the number of subjects is assumed to be $N = 200$ and the treatment indicator $Z_i$ and the observations of two covariates $(W_{i1}, W_{i2})$ for each subject are generated respectively from a Bernoulli distribution with probability 0.5 and a bivariate normal distribution with mean $(0.5, 0.5)$, variance $(1, 1)$, and correlation 0.5. Normal-distributed longitudinal measurements were generated using the R Package \texttt{mvtnorm} \cite{Mi2009} and the details of the procedure for generating binary longitudinal measurements is summarized in Appendix B. Although a formula is given in (5.7) for the calculation of the bandwidth $h$ required for the proposed procedure, I also varied the value of $h$ in the simulation studies to assess the robustness of the proposed formula.

For each parameter in Model (5.9), the absolute relative bias (RBIAS) (defined as the absolute value of the ratio of the bias over the true value of the parameter), standard deviation (SD), and the root of the mean square error (RMSE) of its estimator, and the coverage probability (CP) of its 95% confidence intervals using the sandwich variance estimator are estimated from 500 replications. I also estimated the misclassification rate (MR), defined as the average of the proportions of subjects who is known from one subgroup based on the real value of $(\gamma_1, \gamma_2)$ but mis-classified into another subgroup based on the estimates of $(\gamma_1, \gamma_2)$, to evaluate the accuracy of the subgroup classification.

Tables 5.1 and 5.2 present the results of simulations when the longitudinal measurements are respectively continuous and binary. From these tables, one can see that the RBIASs and RMSEs of the estimators for $\beta$ are stable with the change of $h$, which indicates that the estimation of $\beta$ would not be sensitive to the choice of $h$ defined by
equation (5.7). For all the parameters, the relative absolute biases of their estimators are small in majority of cases considered in the simulations and the smallest RMSE is achieved when \( h \) is around 0.1, which is close to the optimal \( h \) value 0.114 calculated from (5.7) when the subgroups are balanced and 0.101 when the subgroups are unbalanced. Although the optimal choice of \( h \) based on minimization of the RMSE in the parameter estimation may not guarantee a good performance in terms of the coverage probability of the confidence interval [Imbens and Kalyanaraman, 2012; Cheng et al., 2006], the coverage probabilities of the confidence intervals based on the proposed sandwich variance estimator are close to the nominal level for all parameters when \( h \) takes the value from 0.1 to 0.3. The misclassification rate is quite low for all the \( h \) considered, although there is an increasing trend when \( h \) increases.

In practice, some longitudinal measurements may be missing from a patient due to the drop-out or other reasons. I have conducted additional simulations to assess the robustness of the proposed procedures to the potential missing data. Specifically, for each data set generated in previous simulations, the first longitudinal measurement for each subject is kept but the remaining observations may be missing and the probability of missing depends on the previous observed measurements via a logistic regression model [Diggle and Kenward, 1994]. The detailed procedure is presented in Appendix A. The overall missing rate is around 25% for simulated data sets when the subgroups are balanced and around 18% when the subgroups are unbalanced. The results of these additional simulation studies are presented in Table 5.3. In comparison with the results in Tables 5.1 and 5.2, one may find that there is no substantial change in the biases of the estimators for all parameters but SD and RMSE of the estimators are larger for most of parameters, which is expected because the presence of missing
Table 5.1: Simulation results for continuous longitudinal measurements under different values of $h$.

<table>
<thead>
<tr>
<th>$h$</th>
<th>Balanced subgroups</th>
<th>Unbalanced subgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unbalanced subgroups</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>RBIAS</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>93.8</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>RBIAS</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>96.0</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>RBIAS</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>95.0</td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>RBIAS</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.127</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>95.0</td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>RBIAS</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.182</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.184</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>94.8</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>RBIAS</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.053</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>2.049</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>84.4</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>RBIAS</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.335</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>2.353</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>89.2</td>
</tr>
<tr>
<td>MR(%)</td>
<td>0.926</td>
<td>1.347</td>
</tr>
</tbody>
</table>
Table 5.2: Simulation results for binary longitudinal measurements under different values of $h$.

<table>
<thead>
<tr>
<th></th>
<th>Balanced subgroups</th>
<th>Unbalanced subgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
<td>0.09</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>RBIAS</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.260</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.261</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>96.2</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>RBIAS</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>94.4</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>RBIAS</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.331</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>91.4</td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>RBIAS</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.298</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>91.4</td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>RBIAS</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.470</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.486</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>90.6</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>RBIAS</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.313</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>88.0</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>RBIAS</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.296</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>85.4</td>
</tr>
</tbody>
</table>
Table 5.3: Results for simulation studies over 500 replications when the observations may be missing at random

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Balanced</td>
<td>Unbalanced</td>
<td>Balanced</td>
<td>Unbalanced</td>
<td>Balanced</td>
<td>Unbalanced</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>RBIAS</td>
<td>0.114</td>
<td>0.003</td>
<td>0.018</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.100</td>
<td>0.073</td>
<td>0.282</td>
<td>0.243</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.101</td>
<td>0.073</td>
<td>0.282</td>
<td>0.242</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>RBIAS</td>
<td>0.0004</td>
<td>0.007</td>
<td>0.028</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.108</td>
<td>0.082</td>
<td>0.335</td>
<td>0.287</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.108</td>
<td>0.082</td>
<td>0.337</td>
<td>0.296</td>
<td></td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>RBIAS</td>
<td>0.007</td>
<td>0.008</td>
<td>0.077</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.112</td>
<td>0.083</td>
<td>0.408</td>
<td>0.332</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.112</td>
<td>0.084</td>
<td>0.424</td>
<td>0.340</td>
<td></td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>RBIAS</td>
<td>0.031</td>
<td>0.020</td>
<td>0.131</td>
<td>0.232</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.101</td>
<td>0.103</td>
<td>0.308</td>
<td>0.411</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.105</td>
<td>0.105</td>
<td>0.224</td>
<td>0.472</td>
<td></td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>RBIAS</td>
<td>0.024</td>
<td>0.019</td>
<td>0.153</td>
<td>0.188</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.170</td>
<td>0.154</td>
<td>0.547</td>
<td>0.623</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.174</td>
<td>0.156</td>
<td>0.592</td>
<td>0.683</td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>RBIAS</td>
<td>0.014</td>
<td>0.046</td>
<td>0.032</td>
<td>0.219</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.137</td>
<td>0.128</td>
<td>0.517</td>
<td>0.538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.138</td>
<td>0.136</td>
<td>0.519</td>
<td>0.580</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>RBIAS</td>
<td>0.082</td>
<td>0.034</td>
<td>0.031</td>
<td>0.080</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.135</td>
<td>0.133</td>
<td>0.398</td>
<td>0.463</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.157</td>
<td>0.143</td>
<td>0.399</td>
<td>0.478</td>
<td></td>
</tr>
<tr>
<td>MR(%)</td>
<td></td>
<td>0.889</td>
<td>1.332</td>
<td>3.983</td>
<td>4.203</td>
<td></td>
</tr>
</tbody>
</table>

Observations would reduce the overall sample sizes. The MRs are, however, similar to what are observed before when there is no missing data, indicating that the proposed procedure works well for longitudinal data even when some observations are missing at random.
5.4 Application to Data from a Clinical Trial on Pancreatic Cancer

I have applied the proposed model and procedure to the data from a PA.3 trial conducted by the Canadian Cancer Trials Group which randomized 569 patients with locally advanced or metastatic pancreatic cancer into two arms: erlotinib plus gemcitabine (ERLGEM) or gemcitabine alone (GEM) (Moore et al., 2007). A small benefit in overall survival was found for patients treated with ERLGEM but there was no significant difference between the arms in any of the Quality of Life (QoL) function domains defined by the European Organisation for Research and Treatment of Cancer (EORTC) Core Quality of Life Questionnaire C30 (QLQ-C30). It is of interest to identify a subgroup of patients who may benefit from ERLGEM in terms of QoL.

I considered scores of physical function domain from the EORTC QLQ-C30, which range from 0 to 100 with higher scores corresponding to better QoL and measured every 4 weeks until documentation of progressive disease, as the clinical outcome of the interest. A prior study has identified carbohydrate associated antigen 19-9 (CA19-9) and AXL receptor tyrosine kinase (Song et al., 2011; Bauer et al., 2013; Shultz et al., 2016) as two important biomarkers for identification of a subgroup of patients who may benefit from ERLGEM in overall survival (He et al., 2018). I explored in our application whether these two biomarkers can also define a subgroup of patients who may have better physical function when treated by ERLGEM.

A total of 387 patients who had at least one assessment of QoL physical function and biomarker data available were included in our analysis. Among them, 175 were treated by ERLGEM. The measurements on the two biomarkers were logarithmized and standardized so that each of them has mean 0 and variance 1 before the conduct
I first fitted model (5.2) with a link function \( g(\mu) = \mu \) and \( X \), including intercept, time and its square, and treatment indicator to the observed longitudinal scores of QoL physical function domain. The bandwidth calculated from formula (5.7) based on the sample size and the measurements of biomarkers was 0.85. The results are presented in the first three columns of the top panel in Table 5.4, which include, for each parameter in the model, the estimate and its standard error and associated \( p \)-value for the significance of the parameter. From these results, one can see that a treatment-sensitive subgroup may be defined by all patients with \(-1.196 \times CA19-9 + 1.441 \times AXL > 1\). The interaction between the subgroup status and the treatment was highly significant \((p < 0.0001)\), which implies that patients in this subgroup may have larger benefit in QoL physical function when treated by ERLGEM than the rest of patients. To verify this, I have first plotted in Figure 5.1 the observed scores of QoL physical function over time by treatment group with lowess curves superimposed for patients in each of these two subgroups. These plots suggest that, for 71 patients in the identified treatment-sensitive subgroup, there is a large difference between treatment groups in QoL physical function and patients on ERLGEM did better, especially at later timepoints, but almost no difference between treatment groups was observed for the rest of patients. This observation was also formally confirmed by fitting a regular GEE model separately to patients in each of these two subgroups with the R package gee (Carey 2002). Results presented in the first three columns of lower panel in Table 5.4 indicate that, while a significant treatment effect was observed for patients in the treatment-sensitive subgroup, there was no significance between treatment groups for the rest of patients.
I further used a residual bootstrap method (Flores-Agreda and Cantoni, 2019; Field and Welsh, 2007) to assess the stability of the subgroup classification based on the rule developed from our procedure. For each bootstrap sample drawn by the residual bootstrap method, patients are divided into two subgroups based on the above rule, and then a GEE model with time, quadratic term of time, treatment indicator, subgroup indicator, and an interaction term between the subgroup status and treatment indicator were fitted. Among 1000 bootstrap samples, a statistically significant interaction effect was found in 94.6% samples, which shows that the classifying rule is very stable.

For easier interpretation, a patient is often classified as with QoL “worsened” or
Figure 5.1: QoL scores over time with Lowess curves by treatment arm (blue lines for ERLGEM and red lines for GEM) in treatment sensitive and insensitive subgroups, respectively

“not worsened” at a specific time-point based on whether the observed QoL score is reduced by more than 10% compared to the baseline scores (Osoba et al., 1998; Cella et al., 2002). I fitted the same model (5.2) with a link function \( g(\mu) = \log \frac{\mu}{1-\mu} \) to the binary longitudinal QoL outcomes created by this classification. The results are presented in the last three columns of the top panel in Table 5.4, which suggest two subgroups defined by whether \(-1.116 \times \text{CA19-9} + 1.078 \times \text{AXL} > 1\) would have significantly different treatment effects as measured by the binary outcome of QoL physical function status \((p < 0.0001\) for the interaction between the subgroup status and treatment). The plots over time by treatment group with lowess curves superimposed for the log of the odds for “not worsened” in QoL physical function
presented in Figure 5.2 for patients in each of these two subgroups suggests that 65 patients in the treatment-sensitive subgroup defined by \(-1.116 \times \text{CA19-9} + 1.078 \times \text{AXL} > 1\) were significantly more likely to not have a sizable worsened physical function when treated by ERLGEM, while there was no significant difference between treatments for the rest of patients. This observation is confirmed by the results from a formal analysis which fitted a GEE model separately for patients in these two subgroups and presented in the last three columns of the lower panels in Table 5.4. The stability of subgroup classification in this case was also evaluated using a parametric bootstrap method [Adjei and Karim, 2016]. Among 1000 bootstrap samples generated, a statistically significant interaction effect was found in 89.5% of them, which indicates that the classifying rule generated by our proposed approach is reasonably stable.

5.5 Conclusions and Discussions

In this chapter, a procedure based on smoothed generalized estimating equations is proposed to make statistical inferences on the parameters in a generalized single-index linear threshold model, which can be used for identification of treatment-sensitive subgroups based on multiple biomarkers and longitudinal measurements in clinical trials and other clinical studies. Efficient computation algorithms are employed to compute the estimators and the corresponding sandwich variances estimators. Results from simulation studies showed that the proposed procedures work well for both continuous and binary outcomes and the application to the analysis of data from a randomized clinical trial suggests that the proposed procedures are feasible and stable.
Figure 5.2: Log-odds of QoL not worsened at each time-point with loess curves by treatment arm (blue lines for ERLGEM and red lines for GEM) in treatment sensitive and insensitive subgroups, respectively.

The proposed model is applicable to both continuous and binary longitudinal measurements and can include functions of time and other covariates as fixed effects. But the marginal distribution of the longitudinal measurements is assumed from a family of exponential distributions and the function of time needs to be known and pre-specified, which may not be suitable to all important applications. For example, since the QoL scores are restricted in an interval, a simplex distribution, which is not in the exponential distribution family, may be more appropriate to model the marginal distribution of this type of longitudinal proportional data (Qiu et al., 2008). Different procedures with spline functions are suggested to provide more flexible approaches to model potential non-linear relationships between the longitudinal measurements.
and time (Wu and Zhang, 2006). Extensions of the model proposed in this chapter to these more complicated cases of statistical applications would be of interest but would bring more technical challenges.

The number of covariates included in our model is fixed. In some applications, there may be a large number of potential covariates available and some selection procedures, such as that based on a smoothly clipped absolute deviation (SCAD) penalty function (Wang et al., 2012), may have to be used to determine a small subgroup of covariates to be included in the final model. When the number of covariates is ultra-high, interaction pursuit methods developed by Kong et al. (2017) may be generalized to longitudinal outcomes for the identification of treatment-sensitive subgroups. These will be important and interesting topics for future investigations.

For the construction of an index based on a linear combination of multiple covariates which can be used for the diagnosis of a disease or determine the prognosis of a patient after the diagnosis with a disease, various statistical methods have been developed to find for an optimal linear combination by maximizing some criteria such as partial area under the receiver operating characteristic curve (pAUC) (Pepe and Thompson, 2000; Yan et al., 2018). It is difficult to generalize these methods to find the best linear combination for the definition of a treatment-sensitive subgroup since, as noted by Janes et al. (2015), criteria to evaluate the accuracy of the biomarkers in the prediction of differential treatment effect are difficult to define, which will need medical researchers and statisticians to work together to find out a solution on the best approach to evaluate and compare approaches for the prediction of treatment benefits.
Chapter 6

Summary and Future Extensions

In this thesis, I reviewed some statistical methods proposed in the literature which can be used to identify treatment-sensitive subgroups based on the survival data and longitudinal outcomes, respectively. Since existing methods for longitudinal outcomes require subjective determination of subgroups, new statistical models are introduced in this thesis for the determination of subgroups based on a single continuous covariate or multiple covariates with longitudinal comes. Statistical procedures are proposed for the inference of parameters in these models. The proposed new models and statistical inference procedures are evaluated through simulation studies and applied to the analysis of the data from cancer clinical trials conducted by Canadian Cancer Trials Group.

In Chapter 3, a threshold linear mixed model is proposed for identification of treatment-sensitive subgroups based on a single covariate and longitudinal outcome. This model is very flexible and can include different effect of time and other covariates as both fixed and random effects. Statistical estimation methods are proposed based on smooth approximation of an indicator function in the link function to estimate the unknown parameters in the proposed model. Efficient computational algorithms
are developed to compute the estimators and associated variances. Results from simulation studies showed the proposed procedure performed very well and the application to the analysis of data from two clinical trials demonstrates that the proposed procedure is practical and stable.

In Chapter 4, a threshold mixed-effects Tobit model is introduced as a modification to the threshold linear mixed model proposed in Chapter 3 to accommodate the longitudinal outcomes with floor and ceiling effects caused by a large portion of patients reporting minimum and maximum values. The modification is necessary since some assumptions in the threshold linear mixed model may be violated in the longitudinal outcomes with floor and ceiling effects. Smoothed likelihood function is defined similarly based on smooth approximation of an indicator function but a Gauss-Hermite Quadrature approximation to the smoothed likelihood function is adopted to obtain the estimates of parameters in the model because smoothed likelihood function of this model is difficult to evaluate analytically. A random weighting method is utilized to estimate the variance of the proposed parameter estimators. Simulation studies show that the proposed method is less biased and has smaller RMSE for longitudinal outcomes with different degrees of floor and ceiling effects, in comparison with the method based on threshold linear mixed model which ignores the potential floor and ceiling effects of longitudinal outcomes. The application to the analysis of data from a randomized clinical trial demonstrates that the proposed procedure is also practical and stable.

In Chapter 5, a generalized single-index linear threshold model is introduced for identification of treatment-sensitive subsets based on multiple biomarkers and longitudinal outcomes which may be continuous or discrete. A procedure based on smoothed
generalized estimating equations is proposed to make statistical inferences on the parameters in the model. Efficient computation algorithms are employed to compute the estimators and the corresponding sandwich variances estimators. Results from simulation studies showed that the proposed procedures work well for both continuous and binary outcomes and the application to the analysis of data from a randomized clinical trial suggests that the proposed procedures can be applied in practice.

I have discussed at the end of Chapters 3-5 potential extensions of the models and methods introduced. In general, this thesis only considered identification of treatment-sensitive subgroups based on a single covariate or a fixed set of a small number of covariates. In practice, a very large number of covariates may be available but only a few may be important in defining treatment-sensitive subgroups and the number of covariates may increase as sample size increases. Extension of proposed models and procedures to this situation is important and requires further investigations.

My research work presented in this thesis considered only the scenario where the outcomes of clinical trials are longitudinal measurements, although I have reviewed several methods for identification of treatment-sensitive subgroups when the outcomes of the clinical trials are survival times. Joint models have proposed recently to simultaneously analyze the longitudinal and survival outcomes, which may lead to more efficient estimation of the treatment effects on both the longitudinal and survival outcomes and reduce the bias in the estimation of the overall treatment effect when longitudinal and survival outcomes are correlated and the missing of longitudinal outcomes may be caused by the death of patients (Rizopoulos, 2012). Development of
joint threshold models to identify subgroups of patients who are sensitive to a treatment with respect to both longitudinal and survival outcomes is an interesting topic for future research.

In all the models proposed in this thesis, the biomarkers that are used to identify subgroups are assumed fixed at baseline. In practice, it may also be useful to identify subgroups by using time-varying biomarkers such as levels of gene expression measured at timepoints post-baseline. Bertolet et al. (2016) proposed a tree-based method to identity subsets of time-varying biomarkers with respect to survival data, but their method cannot be adopted directly to identifying treatment-sensitive subsets. Generalization of the procedures proposed in their paper to identification of treatment-sensitive subsets with respect to the longitudinal data is also of interest.

In this thesis, I only considered the scenario where there are only two treatment-sensitive subgroups. For some treatments, the number of subgroups may be more than two. Investigating directly the interaction between the treatment and the biomarker on a continuous scale may be used to explore whether there exist more than two treatment-sensitive subgroups. Liu et al. (2015) developed a local partial-likelihood bootstrap method to assess whether treatment effect varies as a function of biomarkers when the outcomes of the clinical trials are survival times. Extending their procedures to the longitudinal data is another interesting topic for the future investigations.

In some clinical trials, there may be more than two treatment groups. He et al. (2018) considered multiple treatment groups in their proposed procedures to identify treatment-sensitive subsets for survival data. It will be interesting and important to extend also our models and methods to this situation.
The performance of the proposed models and procedures are evaluated when longitudinal data may be missing at random (MAR). The MAR assumption, however, may not always be clinically plausible (Sterne et al., 2009). Handling of data which may be missing not at random is a difficult problem in the longitudinal analysis. Extensions of the models developed in this thesis to take into account of other missing data mechanisms are also of interest and important in the applications.

Some other potential future research topics would be to develop statistical procedures for selection and diagnosis of the proposed models and extend our proposed methods to other models, such as mixture models which can incorporate non-Gaussian longitudinal outcomes (Shen and Qu, 2020), when it is found some assumptions of the proposed models are not satisfied.
Bibliography


BIBLIOGRAPHY


Appendix A

Generating Data Missing at Random

In simulation studies for the threshold linear mixed model presented in Chapter 3, it was assumed the observations from some patients are missing due to random drop-out, which depends on the observed measurements. I used a logistic regression model (Diggle and Kenward, 1994) to generate a dataset with potential missing at random observations. The logistic model can be written as

\[
\logit\{P_{ik}(y_{i1}, \cdots, y_{ik}; \gamma)\} = \gamma_0 + \gamma_1 y_{ik} + \sum_{j=2}^{k} \gamma_j y_{i,k+1-j} \\
i = 1, \cdots, N; \; k = 2, \cdots, n_i.
\]

In my implementation, I have assumed that \(\gamma_1 = \gamma_3 = \cdots = \gamma_k = 0\), which simplified the logistic model as

\[
\log \frac{P_{ik}}{1 - P_{ik}} = \gamma_0 + \gamma_2 y_{i,k-1} \text{ where } i = 1, \cdots, N; \; k = 2, \cdots, n_i.
\]

For each subject \(i\), if \(P_{ik} \geq 0.5\) and all \(P_{i2, \cdots, P_{i,k-1}}\) are less than 0.5, then the first \(k\) observations will be retained and all the rest observations are deleted. In the
simulation study, $\gamma_0$ was chosen as $-1.5$ and $\gamma_2$ as 0.1, which leads to a missing rate at around 23%. The full sample size was 200 subjects with 4 measurements from each subject.

In simulation studies for a generalized single-index linear threshold model presented in Chapter 5, the above logistic model cannot be directly used to generate missing binary longitudinal outcomes. Since response $\mu_{ij}$ follows Model 5.9, which leads to lacking of distinction for the responses. To avoid this problem, I added a random part to $\mu_{ij}$ to make a little bit larger distinction. Specifically, the logistic model can be revised as

$$\log \frac{P_{ik}}{1 - P_{ik}} = \gamma_0 + \gamma_2(\mu_{i,k-1} + \text{error})$$

where $\mu_{ik} = E(y_{ik})$ and error follows $N(0, V_i)$. For each subject $i$, if $P_{ik} \geq 0.5$ and all $P_{i2}, \cdots, P_{i,k-1}$ are less than 0.5, then the first $k$ observations are retained and all the rest observations deleted. In the simulation studies, I chose $\gamma_0$ as $-0.25$ and $\gamma_2$ as 0.12 and the original sample size as 200 subjects with 5 measurements of each subject, then the dataset generated by using the above logistic model had a rate of missing at around 24.3% (38/200 subjects have one observations, 19/200 subjects have two observations, 9/200 subjects have three observations, 16/200 subjects have four observations and 118/200 subjects have five observations).
Appendix B

Generating Correlated Binary Data

Many methods have been proposed to generate correlated binary data. Qaqish (2003) proposed a method based on a conditional linear family of multivariate Bernoulli distributions. Qaqish’s method is attractive in simulation studies because of the following reasons. First, it allows unequal means of observations within the same subject and both positive and negative correlations between observations within the same subject. Second, it avoids to compute the full joint distribution of the multivariate binary variables, which consumes lots of computation time and was impractical for the bigger cluster size. Third, this method has been employed in some simulation studies, such as Evans and Li (2005). I adopted this method in the simulation studies presented in Chapter 5. The method is described as below.

First I introduce some notations to present multivariate binary distributions and define a subfamily. For $i = 1, \ldots, n$ define $X_i \triangleq (Y_1, \ldots, Y_{i-1})'$, $\theta_i \triangleq E(X_i)$, $G_i \triangleq Cov(X_i)$ and $s_i \triangleq Cov(X_i, Y_i)$.

Generally, for $i = 2, \ldots, n$, the conditional linear family is given as

$$E(Y_i|X_i = x_i) = \mu_i + k_i(x_i - \theta_i),$$
where each $k_{i1}$ is an $(i-1) \times 1$ integer vector. Furthermore, for given marginal means and correlations, that is, for a given $(\mu, V)$, the parameters are obtained in the closed form as $k_{i1} = G_i^{-1}s_i \triangleq b_i$ for $i = 2, \cdots, n$ and the corresponding conditional Bernoulli distribution of $y_i$ given $X_i = x_i$ can be expressed as

$$\lambda_i := \lambda_i(x_i; \mu, V) = E(Y_i|X_i = x_i) = \mu_i + b_i^T(x_i - \theta_i) \quad (i = 2, \cdots, n). \tag{B.1}$$

If the correlation structure is exchangeable, namely $r_{ij} = \alpha$ for $i \neq j$ and $-1/(n-1) < \alpha < 1$. The $j$-th element of $b_i$ is obtained as

$$b_{ij} = \frac{\alpha}{1 + (i-2)\alpha} \left( \frac{v_{ii}}{v_{jj}} \right)^{\frac{1}{2}} (j = 1, \cdots, i-1).$$

Then take $b_{ij}$ plug into equation (B.1) and therefore $\lambda_i$ is given as

$$\lambda_i = \mu_i + \frac{\alpha}{1 + (i-2)\alpha} \sum_{j=1}^{i-1} \left( \frac{v_{ii}}{v_{jj}} \right)^{\frac{1}{2}} (y_j - \mu_j) \quad (i = 2, \cdots, n).$$

If the correlation structure is AR(1), namely $r_{ij} = \alpha^{|i-j|}$ for $i \neq j$ and $|\alpha| < 1$. The conditional mean $\lambda_i$ is given as

$$\lambda_i = \mu_i + \alpha(y_{i-1} - \mu_{i-1}) \left( \frac{v_{ii}}{v_{(i-1),(i-1)}} \right)^{\frac{1}{2}} \quad (i = 2, \cdots, n).$$

Based on the procedure introduced above, the algorithm to simulate the clustered binary data is described as follow. For each subject, the first step is to generate the first observation $y_1$ from the Bernoulli distribution with given mean $\mu_i$, and then, for $i = 2, \cdots, n$, generate the $i$-th observation $y_i$ for the subject from the conditional
Bernoulli distribution with conditional mean $\lambda_i$ obtained by equation (B.1).
Appendix C

Gauss-Hermite Quadrature Approximation for the Tobit Model

Applying the Gauss-Hermite quadrature to the equation, one can get

\[
L_i = \frac{1}{\sqrt{\pi}} \sum_{m=1}^{M} \omega_m^* \left\{ \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2}\sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right) \right]^{I_{ij}} \Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2}\sigma_\alpha \psi_m^* - u}{\sigma_\varepsilon} \right) \right\}^{I_{ij}} \left(1 - I_{ij} - I_{uj} \right),
\]

where \( H \) is the number of quadrature points, \( \psi_i, \cdots, \psi_H \) are the abscissae, and \( \omega_1, \cdots, \omega_H \) are the corresponding weights (Liu and Pierce, 1994). Therefore, the log-likelihood function is given by

\[
\log L = \sum_{i=1}^{N} \log L_i
\]
and its gradients given the vector of parameters \((\beta^T, \eta^T, c, \sigma, \sigma_\alpha)^T\) are the following:

\[
\frac{\partial L}{\partial \beta_p} = \sum_{i=1}^{N} \frac{1}{\sqrt{n}L_i} \sum_{m=1}^{M} \omega_m \left\{ \left( \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi^*_m}{\sigma_\epsilon} \right) \right] ^{I_{ij}} \right) \right\} ^{I_{ij}} \]

\[
= \sum_{i=1}^{N} \frac{1}{\sqrt{n}L_i} \sum_{m=1}^{M} \omega_m \left\{ \left( \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi^*_m}{\sigma_\epsilon} \right) \right] ^{I_{ij}} \right) \right\} ^{I_{ij}} \]
\[
\frac{\partial L}{\partial \eta_q} = \sum_{i=1}^{N} \frac{1}{\sqrt{\pi L_i}} \sum_{m=1}^{M} \omega_m^* \left\{ \left( \prod_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m^*}{\sigma \varepsilon} \right) \right)^{I_{ij}}\right.
\]
\[
\left[ \Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m^* - u}{\sigma \varepsilon} \right) \right]^{I_{ij}^\text{u}} \left[ \left( \sum_{j=1}^{n_i} \right) - \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m^*}{\sigma \varepsilon} \right) \right]^{I_{ij}^\text{v}} \left( \frac{[W_i]_q}{\sigma \varepsilon} \right)^{I_{ij}^\text{q}} \left[ \Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m^* - u}{\sigma \varepsilon} \right) \right]^{I_{ij}^\text{w}} \left[ \left( \sum_{j=1}^{n_i} \right) - \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m^*}{\sigma \varepsilon} \right) \right]^{I_{ij}^\text{x}} \left( \frac{[\tilde{W}_i]_q}{\sigma \varepsilon} \right)^{I_{ij}^\text{y}} \left\} \right.
\]
\[
\frac{\partial L}{\partial c} = \sum_{i=1}^{N} \frac{1}{\sqrt{\pi} L_i} \sum_{m=1}^{M} \omega_m^* \left\{ \left( \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma_{\alpha} \psi^*_m}{\sigma_{\varepsilon}} \right) \right] \right) \right\}.
\]
\[
\begin{align*}
\frac{\partial L}{\partial \sigma} &= \sum_{i=1}^{N} \frac{1}{\sqrt{\pi} \lambda_i} \sum_{m=1}^{M} \omega_m^* \left\{ \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{1 - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right]^I_{ij} \right. \\
&\quad \left. \Phi \left( \frac{X_{ij}^T \beta + \bar{W}_i^T \eta + \sqrt{2} \alpha \psi^*_m - u}{\sigma} \right) \right] \left[ \Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right] (1 - I_{ij}^w) \right) \\
&= \sum_{i=1}^{N} \frac{1}{\sqrt{\pi} h L_i} \sum_{m=1}^{M} \omega_m^* \left\{ \prod_{j=1}^{n_i} \left[ \Phi \left( \frac{1 - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right]^I_{ij} \right. \\
&\quad \left. \Phi \left( \frac{X_{ij}^T \beta + \bar{W}_i^T \eta + \sqrt{2} \alpha \psi^*_m - u}{\sigma} \right) \right] \left[ \Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right] (1 - I_{ij}^w) \right) \\
&\quad \left. \left[ \Phi \left( \frac{1 - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right]^I_{ij} \right. \\
&\quad \left. \Phi \left( \frac{X_{ij}^T \beta + \bar{W}_i^T \eta + \sqrt{2} \alpha \psi^*_m - u}{\sigma} \right) \right] \left[ \Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right] (1 - I_{ij}^w) \right) \\
&\quad \left. \left[ \Phi \left( \frac{1 - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right]^I_{ij} \right. \\
&\quad \left. \Phi \left( \frac{X_{ij}^T \beta + \bar{W}_i^T \eta + \sqrt{2} \alpha \psi^*_m - u}{\sigma} \right) \right] \left[ \Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \bar{W}_i^T \eta - \sqrt{2} \alpha \psi^*_m}{\sigma} \right) \right] (1 - I_{ij}^w) \right) \\
\end{align*}
\]
\[
\left( \sum_{j=1}^{n_i} \left[ \frac{\phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)}{\Phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)} \right] l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right) \] 

\[
\left[ \frac{\phi \left( X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u}{\sigma_\varepsilon} \right)}{\Phi \left( X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u \right)} \right] X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u \]

\[
\left[ \frac{1}{\sigma_\varepsilon} \left( \frac{\phi \left( y_{ij} - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right)}{\Phi \left( y_{ij} - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right)} \right) \right] \]

\[
\left( \sum_{j=1}^{n_i} \left[ \frac{\phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)}{\Phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)} \right] l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right) \] 

\[
\left[ \frac{\phi \left( X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u}{\sigma_\varepsilon} \right)}{\Phi \left( X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u \right)} \right] X_{ij}^T \beta + W_i^T \eta + \sqrt{2} \sigma_\alpha \psi_m^* - u \]

\[
\left[ \frac{1}{\sigma_\varepsilon} \left( \frac{\phi \left( y_{ij} - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right)}{\Phi \left( y_{ij} - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^* \right)} \right) \right] \]

\[
\frac{\partial L}{\partial \sigma_\alpha} = \sum_{i=1}^{N} \frac{\sqrt{2}}{\sqrt{\pi} L_i} \sum_{m=1}^{M} \omega_m^* \left( \left( \prod_{j=1}^{n_i} \left[ \frac{\phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)}{\Phi \left( \frac{l - X_{ij}^T \beta - \widetilde{W}_i^T \eta - \sqrt{2} \sigma_\alpha \psi_m^*}{\sigma_\varepsilon} \right)} \right] \right) \right) \]
\[
\begin{align*}
&\Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m - u}{\sigma_\varepsilon} \right)^1_{ij} \\
&\Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1}_{ij} \\
&\Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m - u}{\sigma_\varepsilon} \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1-1}_{ij} \\
&= \sum_{i=1}^{N} \frac{1}{\sqrt{2\pi L_i}} \sum_{m=1}^{M} \omega_m \left\{ \left( \prod_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1}_{ij} \\
&\Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m - u}{\sigma_\varepsilon} \right)^{1}_{ij} \\
&\Phi \left( \frac{y_{ij} - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1}_{ij} \\
&\Phi \left( \frac{X_{ij}^T \beta + \tilde{W}_i^T \eta + \sqrt{2} \sigma \psi_m - u}{\sigma_\varepsilon} \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1-1}_{ij} \\
&\left( \sum_{j=1}^{n_i} \Phi \left( \frac{l - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma \psi_m}{\sigma_\varepsilon} \right) \psi_m \right)^{1-1}_{ij} \right\}.
\end{align*}
\]
\[
\begin{pmatrix}
\left( y_{ij} - X_{ij}^T \beta - \tilde{W}_i^T \eta - \sqrt{2} \sigma_m \psi^*_m \right) \psi^*_m \\
\frac{\sigma^2}{\epsilon}
\end{pmatrix}^{(1-I_{ij}^1-I_{ij}^u)}
\]

where \([X_{ij}]_p (p = 0, 1, 2)\) is the \(p\)-th element of \(X_{ij}\), \([\tilde{W}_i]_q (q = 1, 2)\) is the \(q\)-th element of \(\tilde{W}_i\) and the following fact is used in simplifying the gradients:

\[
\frac{[\phi(g(x))]'}{\Phi(g(x))} = -g(x)g'(x)
\]