Abstract

In this thesis, we explore semiparametric marginal models for binary longitudinal data with dropouts. We are specifically interested in the joint estimation of the marginal mean parameters and association parameters by second order generalized estimating equations when the marginal mean response model is partially linear.

First, we propose and explore a set of weighted generalized estimating equations (GEEs) for fitting regression models to longitudinal binary responses when there are dropouts. Under a given missing data mechanism, the proposed method provides unbiased estimators of the regression parameters and association parameters. Simulations were carried out to study the robustness properties of the proposed method under both correctly specified and misspecified correlation structures. The method is also illustrated in an analysis of some actual incomplete longitudinal data on cigarette smoking trends, which were used to study coronary artery development in young adults.

We also developed a semiparametric approach to analyzing longitudinal binary data. We applied second order GEE approach to analyze longitudinal binary responses under partially linear single-index models. We use the local polynomial smoothing technique to estimate the single-index parameters. We study the empirical properties of the proposed method in simulations. Our simulation study demonstrates that if the true underlying model is partially linear, then our proposed
method generally provides unbiased and efficient estimators. The proposed method is also applied to some real data sets obtained from two longitudinal studies. Next, we propose a weighted local linear kernel method and a weighted second order GEE approach for simultaneous estimation of the single-index, regression and association parameters in partially linear single-index models with dropouts. Finally, the proposed methods are applied to some clinical data obtained from a Genetic and Inflammatory Marker of Sepsis (GenIMS) study. The estimates of the single-index showed strong evidence of nonlinear trend in the mean response function. It is evident from this application that our proposed methods can be used to improve the efficiency of the estimates obtained from an ordinary GEE model for binary longitudinal data with dropouts.
Acknowledgments

First and foremost I want to thank my supervisor Dr. Sanjoy K. Sinha. It has been an honor to be his Ph.D. student. He has taught me, both consciously and unconsciously, how good statistical theory and practice is done. I appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example he has provided as a successful Bangladeshi-Canadian statistician and professor.

I also thank the School of Mathematics and Statistics for providing me the opportunity to pursue graduate study in Statistics. I would like to thank Cate Palmer, the Graduate Administrator of the school, for her valuable help and suggestions. It is certainly a privilege to thank Dr. Patrick Farrell and Dr. Saleh, A.K.Md.E. for providing immense support and encouragement.

For this dissertation I would like to thank my committee members: Dr. Lang Wu, Dr. Mayer Alvo, Dr. Chul Gyu Park and Dr. Jose Galdo for their time, interest, insightful questions and helpful comments.

My time at Carleton was made enjoyable in large part due to the many friends and groups that became a part of my life. I am grateful for time spent with friends, for my travel buddies and our memorable trips into the mountains and seas. Special
thanks to Aadyean, Sanam, Mahbub, Abeer, Jakir, Tabrez, Zulhasnine, Rabbi and Tipu for being the sources of laughter, joy and support. My time at Carleton was also enriched by the Carleton University Bangladeshi Students Association (CUBSA) and my musical band “OPartheeb”.

Lastly, I would like to thank my family, my aunt Nina Ahmed, uncle Faruq Hos-sain and my lovely cousin Ishti Hossain for for all their precious support, love and encouragement. For my hard-working parents who have sacrificed their lives for my brother and myself and provided unconditional love and care. I love them so much, and I would not have made it this far without them. For the presence of my brother Mursalin and his family here at Ottawa. And most of all for my loving, supportive, encouraging, and patient wife Disha (Asfin Haider) whose faithful support during the all stages of this Ph.D. is so appreciated. These past few years have not been an easy ride, both academically and personally. I truly thank Disha for sticking by my side, even when I was irritable and depressed. I feel that what we both learned a lot about life and strengthened our commitment and determination to each other and to live life to the fullest.
I dedicate this thesis to
My Wife
Asfin Haider Disha
for her constant support and unconditional love.
## Table of Contents

Abstract i  

Acknowledgments iii  

Dedication v  

Abbreviations xiii  

1 Introduction 1  

1.1 Longitudinal Studies 1  

1.2 Marginal Models for Longitudinal Data 3  

1.3 Semiparametric Models for Longitudinal Data 4  

1.4 Dropouts in Longitudinal Data 5  

1.5 Statement of the Problems 6  

1.6 Organization of Thesis 8  

2 Weighted Generalized Estimating Equations for Incomplete Binary Data 10  

2.1 Introduction 10  

2.2 Model and Notation 11
2.2.1 Response Model .................................. 11
2.2.2 Dropout Model .................................. 13
2.3 Methods of Estimation .................................. 15
  2.3.1 Generalized Estimating Equations ....................... 15
  2.3.2 Weighted Generalized Estimating Equation ............... 18
  2.3.3 Approximate Variance of the WGEE Estimator .......... 21
2.4 Asymptotic Properties of WGEE Estimates ..................... 21
2.5 Simulation Study .................................. 26
  2.5.1 Response Models for Simulations ....................... 27
  2.5.2 Dropout Models for Simulations ....................... 28
  2.5.3 Estimating Dropout Probabilities ..................... 29
  2.5.4 Diagnostic methods ............................. 30
  2.5.5 Results.................................. 31
2.6 Application: Analysis of Smoking Data ....................... 37
2.7 Conclusions .................................. 42

3 Semiparametric Generalized Estimating Equations .......................... 44
  3.1 Introduction .................................. 44
  3.2 Model and Notation .................................. 47
    3.2.1 Response Models .................................. 47
      3.2.1.1 Binary Logistic Model ....................... 47
3.2.1.2 Partially Logistic Linear Single-index Model . . . . . . . . 49
3.3 Methods of Estimation ........................................ 50
  3.3.1 Generalized Estimating Equation .......................... 50
  3.3.2 GEE for Partially Linear Single-index Model ............... 51
  3.3.3 Estimating the Response Model: Computational Details ...... 55
3.4 Asymptotic Properties ........................................ 58
  3.4.1 Asymptotics for $\hat{\varphi}(\cdot)$ .......................... 58
  3.4.2 Asymptotics for $(\hat{\beta}, \hat{\xi}, \hat{\alpha})$ ................. 60
3.5 Simulation Study ............................................. 64
  3.5.1 Models for Simulation Study ............................... 64
  3.5.2 Diagnostic methods ........................................ 65
  3.5.3 Results .................................................. 66
3.6 Applications .................................................. 68
  3.6.1 Analysis of ICHS Data ..................................... 68
  3.6.2 Analysis of Smoking Data ................................. 72
3.7 Conclusions ................................................... 75

4 Weighted Semiparametric Generalized Estimating Equations .... 77
  4.1 Introduction .................................................. 77
  4.2 Model and Notation .......................................... 79
    4.2.1 Response Model ....................................... 79
4.2.2 Dropout Model ........................................ 80
4.3 Methods of Estimation ....................................... 82
  4.3.1 WGEE for Partially Linear Single-index Model ............ 82
  4.3.2 Estimating Dropout Probabilities ........................... 85
  4.3.3 Estimating the Response Model .............................. 87
4.4 Conclusions .................................................. 90

5 Application: GenIMS Study ...................................... 91
  5.1 Introduction .................................................. 91
  5.2 Model and Notation ........................................... 93
  5.3 Methods of Estimation ....................................... 94
  5.4 Results ....................................................... 96

6 Conclusions ..................................................... 102
  6.1 Future Research ................................................ 104

Bibliography .......................................................... 105

A R Codes for CARDIA Study Data Analysis ......................... 116

B R Codes for GenIMS Data Analysis ............................... 130
List of Figures

3.1 Estimated nonlinear curves for ICHS study. ............................. 70
3.2 Estimated nonlinear curves for the smoking data from the CARDIA Study .................................................. 74
5.1 Boxplots of Antithrombin Data ............................................. 92
5.2 Estimated nonlinear curves for GenIMS study. ....................... 99

List of Tables

2.1 Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout. [True parameter values: \( \beta = (-1,1,.2)' \), \( \alpha = 0.5 \) and \( \tau = (-2,2,0)' \).] .................................................. 33
2.2 Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout. [True parameter values: \( \beta = (-1,1,.2)' \), \( \alpha = 0.5 \) and \( \tau = (-2,3,0)' \).] .................................................. 34
2.3 Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout. [True parameter values: \( \beta = (-1, 1, .2)' \), \( \alpha = 0.5 \) and \( \tau = (-2, 2, 0.5)' \)]. ...................................................... 35

2.4 Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout. [True parameter values: \( \beta = (-1, 1, .2)' \), \( \alpha = 0.5 \) and \( \tau = (-2, 2, 1)' \)]. ...................................................... 36

2.5 ML Estimates and Standard Errors of Missing Data Model Parameters for the CARDIA Study. .......................................................... 41

2.6 Estimates and Standard Errors of Regression and Association Parameters for the CARDIA Study. .................................................. 41

3.1 Comparison of the usual Prentice’s second order GEE (Method 1) with the proposed method (Method 2) when \( u_i \)'s are continuous. [True parameter values: \( \beta_1 = 1, \xi = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})' \) and \( \alpha = 0.3. \)] ................................................. 66

3.2 Comparison of the usual Prentice’s second order GEE (Method 1) with the proposed method (Method 2) when \( u_i \)'s are discrete. [True parameter values: \( \beta_1 = 1, \xi = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})' \) and \( \alpha = 0.3. \)] 67
3.3 Estimates and Standard Errors of Regression and Association Parameters for the ICHS Study. ........................................ 71

3.4 Estimates and Standard Errors of Regression and Association Parameters for the CARDIA Study. ............................ 74

5.1 ML Estimates and Standard Errors of Missing Data Model Parameters for the GenIMS Study. ........................................ 97

5.2 Estimates and Standard Errors of Regression and Association Parameters for the GenIMS Study. ............................ 100
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARDIA</td>
<td>Coronary Artery Development in Young Adults</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>GEE</td>
<td>Generalized Estimating Equation</td>
</tr>
<tr>
<td>GenIMS</td>
<td>Genetic and Inflammatory Marker of Sepsis</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized Linear Model</td>
</tr>
<tr>
<td>GPLM</td>
<td>Generalized Partially Linear Model</td>
</tr>
<tr>
<td>GPLSIM</td>
<td>Generalized Partially Linear Single-index Model</td>
</tr>
<tr>
<td>ICHS</td>
<td>Indonesian Childrens Health Study</td>
</tr>
<tr>
<td>LM</td>
<td>Linear Model</td>
</tr>
<tr>
<td>MAR</td>
<td>Missing at Random</td>
</tr>
<tr>
<td>MAR</td>
<td>Missing Completely at Random</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NI</td>
<td>Non-Ignorable</td>
</tr>
<tr>
<td>NR</td>
<td>Newton-Raphson</td>
</tr>
<tr>
<td>SIM</td>
<td>Single-index Model</td>
</tr>
<tr>
<td>WGEE</td>
<td>Weighted Generalized Estimating Equation</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Longitudinal Studies

The defining characteristic of a longitudinal study is that individuals are measured repeatedly through time. Longitudinal studies are in contrast to cross-sectional studies, in which a single outcome is measured for each individual. While it is often possible to address the same scientific questions with a longitudinal or a cross-sectional study, the major advantage of the former is its capacity to separate what in the context of population studies is called cohort or age effect (Diggle et al. 2003).

Longitudinal data can be collected either prospectively, following subjects forward in time, or retrospectively, by extracting multiple measurements on each person from historical records. But they are more commonly collected prospectively since the quality of repeated measurements collected from past records or from a
person’s recollections may be inferior. We often collect data in biological, medical
and environmental studies where measurements from the same subjects are taken
repeatedly over a given period of time. A common goal of a longitudinal study is
to characterize the change in response over time and the factors that influence the
change.

Longitudinal data require special statistical methods because the set of observ-
vations on one subject tends to be intercorrelated. This correlation must be taken
into account to draw valid scientific inferences. Analysis of longitudinal data tends
to be simpler, as different subjects can be assumed independent. Valid inferences
can be made by borrowing strength across subjects. That is, the consistency of a
pattern across subjects is the basis for substantive conclusions. For this reason, in-
ferences from longitudinal studies can be made more robust to model assumptions
than those from time series data, particularly to assumptions about the nature of
the correlation (Diggle et al. 2003).

Sociologists and economists often refer to longitudinal studies as panel studies.
Although many of the statistical methods which we discuss in this thesis will be
applicable to the analysis of data from the social sciences, our emphasis is firmly
motivated by longitudinal data problems arising in the biological or health sciences.
1.2 Marginal Models for Longitudinal Data

A common approach to estimating the regression parameters of marginal models for longitudinal binary responses is the generalized estimating equations (GEEs). Liang and Zeger (1986) introduced a class of estimating equations that give consistent estimators of the regression parameters under mild assumptions about the time dependence. The estimating equations are derived without specifying the joint distribution of a subject’s observations, yet they reduce to score equations for multivariate Gaussian outcomes. The GEE approach is based on a “working” correlation structure, and provides consistent estimators even under a misspecified correlation structure. This first-order GEE approach is widely viewed as attractive because it does not require complete specification of the joint distribution of the longitudinal responses, but rather is based only on specification of the first two moments.

Although primary interest most frequently lies in making inferences about the parameters in regression models, there has been increasing interest in association parameters in recent years. When the association parameters are of central importance, second-order GEEs can be constructed to facilitate their efficient estimation (Grace and Cook, 2002). Prentice (1988) extended the GEE approach for estimating both regression and association parameters of marginal models. Lipsitz et al. (1991), Carey et al. (1993) and Fitzmaurice and Lipsitz (1995) considered analyzing the longitudinal data by modelling the association among repeated responses
in terms of marginal odds ratios.

1.3 Semiparametric Models for Longitudinal Data

Often the relationship between the response and covariates is very complex and linear terms may not be adequate to feature that relationship. It may be even worse than useless to fit a linear model to a nonlinear relationship (Grace et al., 2009). There has been substantial recent interest in developing nonparametric and semiparametric regression methods for longitudinal or clustered data. See, for example, Grace et al. (2009), Wang et al. (2005), Fan and Li (2004), Carroll et al. (1997), and Heckman (1986) among others.

In a partial linear model, the mean of the outcome variable is assumed to depend on some covariates \( x \) parametrically and some other covariate \( u \) nonparametrically. This model is particularly appealing when the effects of \( x \) (e.g., treatment) are of major interest, while the effects of \( u \) (e.g., confounders) are nuisance parameters. These models are a natural combination and generalization of simpler models already in the literature, namely single-index models and partial linear models. Carroll et al. (1997) called it generalized partially linear single-index models (GPLSIM). Furthermore, it is a common practice to include a nonparametric function into the model for covariates \( u \) that have a large dimension and are of little interest (e.g., confounders). This allows us to make inference on the effects of \( x \) while mak-
Marginal semiparametric models based on GEE methods and their various extensions have become increasingly popular. See, for example, Grace et al. (2009), Fan and Li (2004), Xia and Hardle (2006), Carroll et al. (1997), and Hardle et al. (1993) among others. However, they did not consider the joint estimation of marginal mean parameters and association parameters.

1.4 Dropouts in Longitudinal Data

Although many longitudinal studies are designed to collect data on every individual in the sample at each time of follow-up, commonly not all responses are observed at all occasions. In general, an individual’s response can be missing at one follow-up time and be measured at the next follow-up time, resulting in a large class of distinct missingness patterns.

We focus on marginal models for analyzing longitudinal binary outcomes with dropouts. When there are missing data, the classical GEE approaches of Liang and Zeger (1986) and Prentice (1988) are valid only when the data are missing completely at random (MCAR) (Rubin, 1976); that is, given the covariates, the missing data process is independent of both the observed and unobserved outcomes. Under a weaker assumption of missing at random (MAR), where missingness depends on the observed but not the unobserved outcomes, the classical GEE estimator may be biased (Fitzmaurice et al., 1995).
1.5 STATEMENT OF THE PROBLEMS

Robins et al. (1995) proposed an inverse probability-weighted first order GEE approach, in which a subject’s contribution to the ordinary GEE is reweighted by the estimated probability of dropout at the time of attrition. The probabilities forming these weights are obtained by modelling indicators of the missing data status as a function of the responses. Estimation of these weights is possible under a MAR scenario, but sensitivity analyses must be conducted when data are nonignorably missing. This method yields unbiased estimating equations and consistent estimators for the mean parameters when the missing data model is MAR and the probability of dropout is correctly specified.

1.5 Statement of the Problems

In this thesis, we propose and explore semiparametric marginal models for binary longitudinal data with dropouts. Our specific interest lies in the joint estimation of the marginal mean parameters and association parameters by second order generalized estimating equations when the marginal model is partially linear. We explore these problems in three parts as stated below.

First, we investigate a set of weighted generalized estimating equations (GEEs) for fitting regression models to longitudinal binary responses when there are dropouts. Under a given missing data mechanism, our proposed method provides unbiased estimators of the regression parameters and association parameters. Simulations
were carried out to study the robustness properties of the proposed method under both correctly specified and misspecified correlation structures. The method is also illustrated in an analysis of some actual incomplete longitudinal data on cigarette smoking trends, which were used to study coronary artery development in young adults.

We also propose and explore a semiparametric method based on second order GEE to analyze longitudinal binary responses under partially linear single-index models. We use a local polynomial smoothing technique to estimate the single-index. In our proposed method, we incorporate the true correlation structure into the estimation algorithm that jointly estimates the single-index, marginal mean parameters and association parameters. We study the empirical properties of the proposed method in simulations. We also applied this method to some actual data sets obtained from two different longitudinal studies.

Next, we propose a weighted local linear kernel method and a weighted second order GEE approach to conduct simultaneous estimation of the single-index, regression and association parameters for partially linear single-index models with dropouts. Finally, we present an application of the proposed methods using some clinical data from a Genetic and Inflammatory Marker of Sepsis (GenIMS) study. This study (Milbrandt et al., 2009) was a longitudinal cohort study of patients with community acquired pneumonia who were recruited from 28 US academic and community hospitals. Roughly, two-third of the patients have missing observations
due to drop-outs. In this study, one of the important coagulation biomarkers for understanding the mechanisms of pneumonia progression to sepsis is antithrombin. Here our goal is to draw inferences on the change in antithrombin level of patients in the presence of missing data and intraperson correlation based on our proposed methods.

1.6 Organization of Thesis

The thesis is organized as follows. In Chapter 2, we introduce the ordinary unweighted GEE and weighted GEE approaches for analyzing incomplete longitudinal data. Then we discuss our proposed methodology and estimation algorithm based on weighted generalized estimating equations for fitting regression models to incomplete longitudinal binary responses. We also present results from a simulation study as well as from a real data analysis in this chapter.

In chapter 3, we describe our proposed methodology based on second order GEE approach for analyzing longitudinal binary responses under partially linear single-index models. We also discuss results from a simulation study and present applications using two different data sets.

In chapter 4, we propose a weighted local linear kernel method and a weighted second order GEE approach to conduct simultaneous estimation of the single-index, regression and association parameters for partially linear single-index models with
1.6. ORGANIZATION OF THESIS

dropouts. Chapter 5 presents an application of the proposed methods using the GenIMS data analysis introduced earlier. Chapter 6 concludes the thesis with some discussion and direction for further research.
Chapter 2

Weighted Generalized Estimating Equations for Incomplete Binary Data

2.1 Introduction

Here our focus is on regression models for longitudinal binary responses, in which the mean binary response at a given time is related to a set of covariates and a time trend by a known link function. The analysis of longitudinal data is often complicated by the fact that not all outcomes are observed at all occasions. In general, the outcome of a subject can be missing at one follow-up time and be observed at the next follow-up time, resulting in a large class of missing data patterns. Here we restrict our attention to the monotone missingness patterns: once a subject drops out, no more measurements are taken on that subject.
2.2. MODEL AND NOTATION

We incorporate the inverse probability-weights of Robins et al. (1995) into the GEE approach of Prentice (1988) for analyzing longitudinal binary responses with dropouts. We study the empirical properties of the weighted GEE estimators in simulations. This chapter is organized as follows. Section 2.2 introduces the model and notation to define the response process and missing data mechanism for incomplete binary longitudinal data. Section 2.3 reviews the ordinary unweighted GEE and weighted GEE approaches for analyzing the incomplete data. Section 2.4 discusess the asymptotic properties of the proposed weighted GEE estimates. Section 2.5 presents results from a simulation study, which was carried out to investigate the empirical properties of the weighted GEE approach. Section 2.6 presents an application of the proposed method using actual longitudinal data from a health study. Section 2.7 gives the conclusions of the chapter.

2.2 Model and Notation

2.2.1 Response Model

Suppose $K$ subjects are observed at a fixed set of $T$ time points. Let $Y_{it}$ represent a binary response variable from subject $i$, $i = 1, \ldots, K$, at visit $t$, $t = 1, \ldots, T$. For the $i$th subject, we can form a $T \times 1$ vector, $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, of binary response variables. Let the lower case letters $y_{it}$ and $y_i$ denote the realizations of $Y_{it}$ and $Y_i$, respectively. Also, let $x_{it} = (1, x_{it,1}, \ldots, x_{it,p-1})'$ be a $p \times 1$ vector of covariates from
subject \( i \) at time \( t \). The covariates may be time-dependent or fixed across the entire observation times. Let \( x_i = (x'_{i1}, \ldots, x'_{iT})' \).

Assume that the marginal distribution of \( Y_{it} \) is Bernoulli:

\[
Y_{it} \sim \text{Bernoulli}(p_{it}), \ i = 1, \ldots, K; \ t = 1, \ldots, T, \tag{2.1}
\]

with the probability of success,

\[
p_{it} = E(Y_{it}|x_i) = P(Y_{it} = 1|x_i). \tag{2.2}
\]

Let \( p_i = (p_{i1}, \ldots, p_{iT})' \). Assuming that the mean of response variable \( Y_{it} \) depends only on the covariate vector for subject \( i \) at time \( t \), i.e., \( p_{it} = E(Y_{it}|x_i) = E(Y_{it}|x_{it}) \) (Pepe and Anderson, 1994), we consider modelling the mean response by the logistic regression:

\[
\text{logit}(p_{it}) = \log \left( \frac{p_{it}}{1 - p_{it}} \right) = x'_{it}\beta, \tag{2.3}
\]

where \( \beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})' \) is the vector of regression parameters. The marginal variance of the response variable \( Y_{it} \) is specified as a function of the marginal mean as

\[
v_{it} = \text{var}(Y_{it}|x_{it}) = p_{it}(1 - p_{it}). \tag{2.4}
\]

We assume that \( Y_{it} \) and \( Y_{i't} \) are uncorrelated when \( i \neq i' \). Let

\[
\text{corr}(Y_{it}, Y_{i't}) = \alpha_{it'}, \tag{2.5}
\]

represent the correlation between \( Y_{it} \) and \( Y_{i't} \) for given \( x_{it} \), where

\[
\alpha = (\alpha_{12}, \ldots, \alpha_{1T}, \alpha_{23}, \ldots, \alpha_{T-1,T})' \] is the vector of correlation parameters.
2.2. MODEL AND NOTATION

2.2.2 Dropout Model

Note that attrition due to dropout and staggered entry is common in many longitudinal studies, where subjects drop out prior to the end of the study and do not return. For example, in panel surveys members of the panel may drop out because they move to a location that is inaccessible to the researchers, or, in a clinical trial, some subjects drop out of the study for known or unknown reasons, possibly due to side effects of drugs, or curing of a disease. The pattern of attrition is an example of monotone missing data, where variables can be arranged so that all $Y_{i,t+1}, \ldots, Y_{iT}$ are missing for cases where $Y_{it}$ is missing and all previous outcomes $Y_{i,1}, \ldots, Y_{i,t-1}$ are observed. In practice, the pattern of missing data is rarely monotone, but it is close to monotone (Little and Rubin, 1987).

In many cases, the responses are missing due to some stochastic missing data mechanism. A different issue concerns the mechanism that lead to missing data, and in particular, the question of whether the variables that are missing are related to the underlying values of the variables in the data set. Missing data mechanisms are crucial since the properties of missing data methods depend very strongly on the nature of the dependencies in the missingness mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simple device of treating the missing data indicators as random variables and assigning
them a distribution.

To introduce a dropout model, let \( R_i = (R_{i1}, \ldots, R_{iT})' \) denote the response indicators for the vector \( Y_i = (Y_{i1}, \ldots, Y_{iT})' \), i.e., \( R_{it} \) be the indicator variable taking the value 1 if the response \( Y_{it} \) is observed and 0 otherwise. We assume a monotone missing data pattern, where \( R_{i1} \geq \ldots \geq R_{iT} \) and \( R_{i1} = 1 \) for all subjects.

In general, the dropout mechanism can depend on the full vector of responses \( Y_i \) (including the unobserved components of \( Y_i \)) and the vector of covariates \( x_i \). Let

\[
\lambda_{it} = P(R_{it} = 1 | R_{i1} = \ldots = R_{i,t-1} = 1, y_i, x_i, \tau) \tag{2.6}
\]

be the probability that the \( i \)th subject is observed at time \( t \), given that the subject is observed at previous \( t - 1 \) time points and given the response vector \( y_i \) and covariates \( x_i \). Here the components of \( \tau \) are referred to as the “nuisance parameters” of the missing data model. We denote the observed components of the response vector \( Y_i \) by the vector \( Y_i^o \) and unobserved components by the vector \( Y_i^u \). In (2.6), as missingness depends on the unobserved values of the response variables, it is referred to as nonignorable (NI) missingness. Data are called missing at random (MAR) if

\[
\lambda_{it} = P(R_{it} = 1 | R_{i1} = \ldots = R_{i,t-1} = 1, y_i^o, x_i, \tau). \tag{2.7}
\]

Data are called missing completely at random (MCAR) if \( \lambda_{it} = P(R_{it} = 1 | R_{i1} = \ldots = R_{i,t-1} = 1, x_i, \tau) \).

Note that in the case of dropouts, the random vector \( R_i = (R_{i1}, \ldots, R_{iT})' \) of
binary indicators can be characterized by a single random variable

\[ M_i = 1 + \sum_{t=1}^{T} R_{it}, \quad (2.8) \]

which indicates the time of dropout. In this case, the missing data or dropout process can be defined by

\[ \nu_{im_i} = f_{M_i}(m_i|y_i, x_i, \tau) = P(M_i = m_i|y_i, x_i, \tau). \quad (2.9) \]

If we assume that all subjects are observed on the first occasion, then \( M_i \) takes on values between 2 and \( T + 1 \), where the maximum value \( (T + 1) \) corresponds to a complete measurement sequence. It can be shown that

\[
P(M_i = m|y_i, x_i, \tau) = P(R_{i2} = \ldots = R_{i,m-1} = 1, R_{im} = 0|y_{i1}, \ldots, y_{im}, x_i, \tau)
= \left\{ \prod_{t=2}^{m-1} P(R_{it} = 1|R_{i1} = \ldots = R_{i,t-1} = 1, y_{i1}, \ldots, y_{it}, x_i, \tau) \right\}
\times \left\{ P(R_{im} = 0|R_{i1} = \ldots = R_{i,m-1} = 1, y_{i1}, \ldots, y_{im}, x_i, \tau) \right\} I\{m \leq T\}
\]

where \( I\{\} \) denotes an indicator variable.

### 2.3 Methods of Estimation

#### 2.3.1 Generalized Estimating Equations

Our primary interest lies in estimating the regression parameters \( \beta \) as well as the association parameters \( \alpha \), with \( \tau \) being viewed as nuisance parameters of the
missing data model. In our marginal modelling approach, we specify only the first two moments of the responses for each person. With Gaussian data, the first two moments fully determine the likelihood, but this is not the case with other members of the GLM family. To specify the entire likelihood, additional assumptions about higher-order moments are also necessary. Even if additional assumptions are made, the likelihood is often intractable and involves many nuisance parameters in addition to $\beta$ and $\alpha$ that must be estimated. For this reason, generalized estimating equations (GEEs) are widely used in the literature, which is a multivariate analogue of quasi-likelihood approach.

Recall that we partitioned $Y_i$ into the observed components $Y_i^o$ and the unobserved components $Y_i^u$. Similarly, we consider partitioning the mean vector $p_i$ into $p_i^o$ and $p_i^u$.

As a naive method, one can consider analyzing the longitudinal data by simply ignoring the missing data pattern and then estimating the model parameters based on the observed data only. In such a case, the generalized estimating equations (GEEs) of Liang and Zeger (1986) can be used for estimating the regression parameters $\beta$:

$$U_\beta(\beta, \alpha) = \sum_{i=1}^{K} D_i^t V_i^{-1} (Y_i^o - p_i^o) = 0,$$

(2.11)

where $D_i = \partial p_i^o / \partial \beta$, $V_i = (B_i^o)^{1/2} R^o(\alpha)(B_i^o)^{1/2}$, $B_i = \text{diag}\{p_{i1}(1 - p_{i1}), p_{i2}(1 - p_{i2}), \ldots, p_{iT}(1 - p_{iT})\}$, $B_i^o$ has the same form as $B_i$, but with $p_i$ replaced by $p_i^o$, and $R^o(\alpha)$ is a “working” correlation matrix for $Y_i^o$ depending on the vector $\alpha$ of
correlation parameters. The above equations can be solved numerically for \( \hat{\beta} \) using an iterative method. Liang and Zeger (1986) show that the solution to the above equation is asymptotically as efficient as if \( \alpha \) were known.

Liang and Zeger (1986) considered estimating the association parameters \( \alpha \) by the method of moments, which uses the Pearson residuals

\[
\hat{r}_{it} = \frac{(y_{it} - \hat{p}_{it})}{(1 - \hat{p}_{it})^{1/2}}. \tag{2.12}
\]

The moment estimators of \( \alpha_{tt'} \) may be obtained as

\[
\hat{\alpha}_{tt'} = \sum_{i=1}^{K} \hat{r}_{it}\hat{r}_{it'}/(K - p) \tag{2.13}
\]

where \( p \) is the dimension of \( \beta \).

Prentice (1988) considered an extension of the GEE approach to allow joint estimation of the regression parameters \( \beta \) and the association parameters \( \alpha \). Specifically, a GEE estimator of the correlation parameter \( \alpha \) may be obtained from a second set of estimating equations by noting that the “sample correlation”

\[
Z_{itu} = Z_{itu}(\beta) = \frac{(Y_{it} - p_{it})(Y_{iu} - p_{iu})}{(p_{it}q_{it}p_{iu}q_{iu})^{1/2}} \tag{2.14}
\]

has mean \( \rho_{itu} \) and variance

\[
w_{itu} = 1 + (1 - 2p_{it})(1 - 2p_{iu})(p_{it}q_{it}p_{iu}q_{iu})^{-1/2} \rho_{itu} - \rho_{itu}^2, \tag{2.15}
\]

for \( t < u < T, i = 1, \ldots, K, \) and \( t = 1, \ldots, T \). Let \( Z_i = (Z_{i12}, \ldots, Z_{iT1}, Z_{i23}, \ldots, Z_{iT-1,T})' \) and \( \rho_i = (\rho_{i12}, \ldots, \rho_{iT1}, \rho_{i23}, \ldots, \rho_{iT-1,T})' \). We denote the observed component of \( Z_i \)
by $Z_i^o$ and that of $\rho_i$ by $\rho_i^o$. Similarly, let $w_{itu}^o$ denote the function $w_{itu}$ obtained from the elements of $p^o$ and $\rho^o$.

Then following Prentice (1988), the GEE estimators of $(\beta, \alpha)$ may be obtained by solving the estimating equations

$$U_\beta(\beta, \alpha) = \sum_{i=1}^{K} D'_i V^{-1}_i (Y_i^o - p_i^o) = 0, \quad (2.16)$$

$$U_\alpha(\beta, \alpha) = \sum_{i=1}^{K} G'_i W^{-1}_i (Z_i^o - \rho_i^o) = 0, \quad (2.17)$$

where $G_i = \partial \rho_i^o / \partial \alpha$ and $W_i = \text{diag}\{w_{i12}^o, \ldots, w_{i1T}^o, w_{i23}^o, \ldots, w_{iT-1,T}^o\}$.

Note that under an MAR or NI process, $E(Y_i^o|x_i, \beta) \neq p_i^o$, in general, and consequently, the ordinary GEE approach may provide biased estimators of both the regression parameters $\beta$ and the association parameters $\alpha$ (Fitzmaurice et al., 1995).

In the next section, we discuss the use of weighted generalized estimating equations for estimating the regression and association parameters when there are dropouts.

### 2.3.2 Weighted Generalized Estimating Equation

Robins et al. (1995) proposed the weighted GEE (WGEE) approach for estimating the regression parameters. In the WGEE approach, a subject’s contribution to the ordinary GEEs is weighted by the inverse probability of dropout at the given time. The WGEE estimators of $\beta$ are obtained by solving:

$$\tilde{U}_\beta(\beta, \alpha) = \sum_{i=1}^{K} \frac{1}{\nu_{im}} D'_i V^{-1}_i (Y_i^o - p_i^o) = 0, \quad (2.18)$$
2.3. METHODS OF ESTIMATION

where \( \nu_{im} \) is introduced in (2.9). The estimating equations (2.18) are unbiased for 0 at the true \( \beta \) if \( \nu_{im} \) is correctly specified (Fitzmaurice et al., 1995).

Note that to estimate the association parameters \( \alpha \), here we consider extending the GEE approach of Prentice (1988). Specifically, we consider estimating \( \alpha \) by solving the weighted GEEs:

\[
\hat{U}_\alpha(\beta, \alpha) = \sum_{i=1}^{K} \frac{1}{\nu_{im}} G'_i W^{-1}_i (Z^o_i - \rho^o_i)
\]

\[
= \sum_{i=1}^{K} \sum_{m_i=2}^{T+1} \frac{I\{M_i = m_i\}}{\nu_{im}} G'_i(m_i) W^{-1}(m_i) \{Z_i(m_i) - \rho_i(m_i)\} = 0,
\]

(2.19)

where \( Z_i(m_i) \) and \( \delta_i(m_i) \) are the corresponding \( m_i - 1 \) elements of \( Z_i \) and \( \rho_i \). For instance, if \( m_i = m \) then \( Z_i(m) = (Z_{i12}, \ldots, Z_{i1,m-1}, Z_{i23}, \ldots, Z_{i,m-2,m-1})' \) and \( \rho_i(m) = (\rho_{i12}, \ldots, \rho_{i1,m-1}, \rho_{i23}, \ldots, \rho_{i,m-2,m-1})' \). We define \( G_i(m_i) \) and \( W_i(m_i) \) analogously. Note that the number of response patterns in \( Z_i \) is the same as the number of response patterns in \( Y_i \), and there are only \( T \) possible patterns. Since

\[
\nu_{im_i} = P(M_i = m_i|y_i, x_i, \tau) = E_{M_i|y_i} [I\{M_i = m_i\}],
\]

(2.20)

we can show that

\[
E \left[ \frac{I\{M_i = m_i\}}{\nu_{im_i}} G'_i(m_i) W^{-1}_i(m_i) \{Z_i(m_i) - \rho_i(m_i)\} \right]
\]

\[
= E_{Y_i} \left[ G'_i(m_i) W^{-1}_i(m_i) \{Z_i(m_i) - \rho_i(m_i)\} E_{M_i|y_i} \left[ \frac{I\{M_i = m_i\}}{\nu_{im_i}} \right] \right]
\]

\[
= E_{Y_i} \left[ G'_i(m_i) W^{-1}_i(m_i) \{Z_i(m_i) - \rho_i(m_i)\} \right] = 0.
\]

(2.21)
Thus the estimating equations (2.19) are unbiased for 0 at the true \( \alpha \) if \( \nu_{im} \) is correctly specified. Since the estimating equations for both \( \beta \) and \( \alpha \) are unbiased for 0, from the standard theory of method of moments, we can argue that the WGEE estimators \( (\hat{\beta}, \hat{\alpha}) \) obtained by solving (2.18) and (2.19) are consistent for \( (\beta, \alpha) \). If the drop out probabilities \( \nu_{im} \) are consistently estimated, then the WGEE estimators would still provide consistent estimators of \( (\beta, \alpha) \).

The iterative procedure for calculating the WGEE estimators \( (\hat{\beta}, \hat{\alpha}) \) begins with some starting values \( (\beta(0), \alpha(0)) \) and produces updated values \( (\beta_{s+1}, \alpha_{s+1}) \) from interim values \( (\beta_s, \alpha_s) \) by means of the iterative equations

\[
\beta_{s+1} = \beta_s + \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} D_i^t V_i^{-1} D_i \right)^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} D_i^t V_i^{-1} (Y_i^o - p_i^o), \quad (2.22)
\]

\[
\alpha_{s+1} = \alpha_s + \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i^t W_i^{-1} G_i \right)^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i^t W_i^{-1} (Z_i^o - \rho_i^o), \quad (2.23)
\]

for \( s = 0, 1, 2, \ldots \), where the second term on the right side of each estimating equation is evaluated at the current estimates \( (\beta_s, \alpha_s) \). Note that the association parameters \( \alpha \) are often estimated by the method of moments or by the GEE approach of Prentice (1988), without any use of inverse probability weights in the estimating equations for \( \alpha \). Here we consider finding the estimators by incorporating the weights into the GEEs for \( \alpha \). This weighted GEE is found to improve the efficiency of the estimators of \( \alpha \).
2.4. ASYMPTOTIC PROPERTIES OF WGEE ESTIMATES

2.3.3 Approximate Variance of the WGEE Estimator

Similarly to White (1982), we consider approximating the variance-covariance matrices of the WGEE estimators \( \hat{\beta} \) and \( \hat{\alpha} \) by using sandwich-type estimators. In particular, we approximate the variance-covariance matrix of \( \hat{\beta} \) from

\[
V(\hat{\beta}) = M_{\beta}^{-1} Q_{\beta} M_{\beta}^{-1},
\]

where the matrices \( M_{\beta} \) and \( Q_{\beta} \) are obtained as

\[
M_{\beta} = \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_{i} V_{i}^{-1} G_{i},
\]

\[
Q_{\beta} = \sum_{i} S_{\beta,i} S'_{\beta,i} \text{ with } S_{\beta,i} = \frac{1}{\nu_{im}} G_{i} V_{i}^{-1} (Y_{i} - p_{i}).
\]

Similarly, the variance-covariance matrix of the WGEE estimator \( \hat{\alpha} \) is obtained from

\[
V(\hat{\alpha}) = M_{\alpha}^{-1} Q_{\alpha} M_{\alpha}^{-1},
\]

where the matrices \( M_{\alpha} \) and \( Q_{\alpha} \) are given by

\[
M_{\alpha} = \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_{i} W_{i}^{-1} G_{i},
\]

\[
Q_{\alpha} = \sum_{i} S_{\alpha,i} S'_{\alpha,i} \text{ with } S_{\alpha,i} = \frac{1}{\nu_{im}} G_{i} W_{i}^{-1} (Z_{i} - \rho_{i}).
\]

The matrices \( M_{\beta}, M_{\alpha}, Q_{\beta}, \) and \( Q_{\alpha} \) are evaluated at the WGEE estimators \( \hat{\beta} \) and \( \hat{\alpha} \).

2.4 Asymptotic Properties of WGEE Estimates

Let us write Equation (2.18) and (2.19) for \( \beta \) and \( \alpha \) jointly as

\[
\sum_{i=1}^{K} \Phi_{K}(\beta, \alpha) = 0.
\]
Applying the first order Taylor series approximation about true values \((\beta, \alpha)\) we can write

\[
0 \equiv \sum_{i=1}^{K} \Phi_K(\hat{\beta}, \hat{\alpha}) \cong \sum_{i=1}^{K} \Phi_K(\beta, \alpha) + \left[ \sum_{i=1}^{K} \frac{\partial \Phi_k(\beta, \alpha)}{\partial (\beta, \alpha)} \right]' \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} - \begin{bmatrix} \beta \\ \alpha \end{bmatrix}
\]

\[
\Rightarrow \left[ \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} - \begin{bmatrix} \beta \\ \alpha \end{bmatrix} \right] \cong \left\{ - \left[ \sum_{i=1}^{K} \frac{\partial \Phi_k(\beta, \alpha)}{\partial (\beta, \alpha)} \right]' \right\}^{-1} \sum_{i=1}^{K} \Phi_K(\beta, \alpha)
\]

\[
\Rightarrow K^{-1/2} \left[ \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} - \begin{bmatrix} \beta \\ \alpha \end{bmatrix} \right] \cong \left\{ - K^{-1} \left[ \sum_{i=1}^{K} \frac{\partial \Phi_k(\beta, \alpha)}{\partial (\beta, \alpha)} \right]' \right\}^{-1} \left[ K^{-1/2} \sum_{i=1}^{K} \Phi_K(\beta, \alpha) \right]
\]

(2.27)

Thus under regularity conditions \(K^{-1/2} \left[ (\hat{\beta} - \beta)', (\hat{\alpha} - \alpha)' \right]'\) can be approximated by

\[
\left[ -K^{-1} \frac{\partial \hat{U}_\beta(\beta, \alpha)}{\partial \beta'} - K^{-1} \frac{\partial \hat{U}_\beta(\beta, \alpha)}{\partial \alpha'} \right]^{-1} \left[ \begin{bmatrix} K^{-1/2} \hat{U}_\beta(\beta, \alpha) \\ K^{-1/2} \hat{U}_\alpha(\beta, \alpha) \end{bmatrix} \right] \equiv J^{-1} L \quad (2.28)
\]

We can write

\[
Var(J^{-1} L) = J^{-1} Var(L)(J^{-1})', \quad (2.29)
\]
2.4. ASYMPTOTIC PROPERTIES OF WGEE ESTIMATES

where \( \text{Var}(L) \) may be obtained as

\[
\text{Var}(L) = K^{-1} \begin{pmatrix}
\text{Var}(\tilde{U}_\beta) & \text{Cov}(\tilde{U}_\beta, \tilde{U}_\alpha) \\
\text{Cov}(\tilde{U}_\beta, \tilde{U}_\alpha) & \text{Var}(\tilde{U}_\alpha)
\end{pmatrix}
\]

\[
= K^{-1} \begin{pmatrix}
\sum_{i=1}^{K} \frac{1}{\nu_i m_i} D_i' V_i^{-1} \text{Var}(Y_i) V_i^{-1} D_i & \sum_{i=1}^{K} \frac{1}{\nu_i m_i} D_i' V_i^{-1} \text{Cov}(Y_i, Z_i) W_i^{-1} G_i \\
\sum_{i=1}^{K} \frac{1}{\nu_i m_i} G_i' W_i^{-1} \text{Cov}(Y_i, Z_i) V_i^{-1} D_i & \sum_{i=1}^{K} \frac{1}{\nu_i m_i} G_i' W_i^{-1} \text{Var}(Z_i) W_i^{-1} G_i
\end{pmatrix}
\]

\[
\equiv K^{-1} \begin{pmatrix}
\Lambda_{11} & \Lambda_{12} \\
\Lambda_{21} & \Lambda_{22}
\end{pmatrix},
\]

with approximations

\[
\text{Var}(Y_i) \approx (Y_i - p_i) (Y_i - p_i)',
\]

\[
\text{Cov}(Y_i, Z_i) \approx (Y_i - p_i) (Z_i - \rho_i)',
\]

\[
\text{Var}(Z_i) \approx (Z_i - \rho_i) (Z_i - \rho_i)'.
\]

(2.31)

We have already seen that the estimating equations \( \tilde{U}_\beta \) in (2.18) and \( \tilde{U}_\alpha \) in (2.19) are unbiased. Hence, the linear functions \( L \) quite generally have an asymptotic normal distribution with mean zero and variance matrix

\[
\Sigma_L = \lim_{K \to \infty} K^{-1} \begin{pmatrix}
\Lambda_{11} & \Lambda_{12} \\
\Lambda_{21} & \Lambda_{22}
\end{pmatrix}.
\]

Thus the joint asymptotic distribution of \( K^{1/2} \left[ (\hat{\beta} - \beta)', (\hat{\alpha} - \alpha)' \right] = J^{-1} L \) is Gaussian with mean zero and variance matrix

\[
\Sigma = \lim_{K \to \infty} J^{-1} \Sigma_L \lim_{K \to \infty} (J^{-1})'.
\]

(2.33)
Also as $K \to \infty$,

$$-K^{-1} \frac{\partial \tilde{U}_\beta(\beta, \alpha)}{\partial \beta'} = -K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} D'_i V_i^{-1} \frac{\partial (Y^o_i - p^o_i)}{\partial \beta'}$$

$$- K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} \frac{\partial D'_i V_i^{-1}}{\partial \beta'} (Y^o_i - p^o_i)$$

$$= K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} D'_i V_i^{-1} D_i + o_p(1), \quad (2.34)$$

since $\frac{\partial (Y^o_i - p^o_i)}{\partial \beta'} = (-1)D_i$, and $\frac{\partial D'_i V_i^{-1}}{\partial \beta'}$ are fixed matrices that do not involve $Y_i$.

Similarly, as $\frac{\partial \nu^{-1}_i}{\partial \alpha'}$, $\frac{\partial E_i W^{-1}_i}{\partial \alpha'}$ and $\frac{\partial E_i W^{-1}_i}{\partial \beta'}$ are all free of $Y_i$, we can write as $K \to \infty$,

$$-K^{-1} \frac{\partial \tilde{U}_\beta(\beta, \alpha)}{\partial \alpha'} = -K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} D'_i V_i^{-1} \frac{\partial \nu^{-1}_i}{\partial \alpha'} (Y^o_i - p^o_i) = o_p(1), \quad (2.35)$$

$$-K^{-1} \frac{\partial \tilde{U}_\alpha(\beta, \alpha)}{\partial \alpha'} = -K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} G'_i W_i^{-1} \frac{\partial (Z^o_i - \rho^o_i)}{\partial \alpha'}$$

$$- K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} \frac{\partial G'_i W_i^{-1}}{\partial \alpha'} (Z^o_i - \rho^o_i)$$

$$= K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} G'_i W_i^{-1} G_i + o_p(1), \quad (2.36)$$

and

$$-K^{-1} \frac{\partial \tilde{U}_\alpha(\beta, \alpha)}{\partial \beta'} = -K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} G'_i W_i^{-1} \frac{\partial (Z^o_i - \rho^o_i)}{\partial \beta'}$$

$$- K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} \frac{\partial G'_i W_i^{-1}}{\partial \beta'} (Z^o_i - \rho^o_i)$$

$$= - K^{-1} \sum_{i=1}^{K} \frac{1}{\nu_{im}} G'_i W_i^{-1} \frac{\partial Z^o_i}{\partial \beta'} + o_p(1). \quad (2.37)$$
2.4. ASYMPTOTIC PROPERTIES OF WGEE ESTIMATES

Here,
\[
\frac{\partial Z_{itu}}{\partial \beta_j} = - \left\{ \frac{\partial p_{it}}{\partial \beta_j} (Y_{iu} - p_{iu}) + \frac{\partial p_{iu}}{\partial \beta_j} (Y_{it} - p_{it}) + \frac{1}{2} (Y_{it} - p_{it}) (Y_{iu} - p_{iu}) \right\} \times \left[ (1 - 2p_{it}) p_{it}^{-1} q_{it}^{-1} \frac{\partial p_{it}}{\partial \beta_j} + (1 - 2p_{iu}) p_{iu}^{-1} q_{iu}^{-1} \frac{\partial p_{iu}}{\partial \beta_j} \right] \left( p_{it} q_{it} p_{iu} q_{iu} \right)^{-1/2}. \tag{2.38}
\]

Hence as \( K \to \infty \), we can write
\[
J = K^{-1} \begin{bmatrix} \sum_{i=1}^{K} \frac{1}{\nu_{im}} D_i' V_i^{-1} D_i & 0 \\ - \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' W_i^{-1} \frac{\partial Z}_{itu} \frac{\partial Z}_{itu}' & \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' W_i^{-1} G_i \end{bmatrix} \tag{2.39}
\]
and
\[
J^{-1} = K \begin{bmatrix} \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} D_i' V_i^{-1} D_i \right)^{-1} & 0 \\ B^* & \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' W_i^{-1} G_i \right)^{-1} \end{bmatrix} = J^*, \tag{2.40}
\]
where
\[
B^* = \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' W_i^{-1} G_i \right)^{-1} \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' W_i^{-1} \frac{\partial Z}_{itu} \frac{\partial Z}_{itu}' \right) \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} D_i' V_i^{-1} D_i \right)^{-1}. \tag{2.41}
\]

Finally, the asymptotic variance matrix of \( K^{1/2} [ (\hat{\beta} - \beta)', (\hat{\alpha} - \alpha)']' \) can be expressed as
\[
\Sigma = J^* \Sigma_L J^*, \tag{2.42}
\]
where \( J^* \) and \( \Sigma_L \) are defined in (2.40) and (2.32) respectively.
To study the empirical properties of the GEE and WGEE estimators under incomplete longitudinal data, we ran two sets of simulations. In the first set, the estimators were studied under correctly specified MAR models. In the second set, the robustness properties of the estimators were studied under misspecified non-ignorable missing data models. In each set of simulations, data were generated under both exchangeable and serial correlation structures among the responses. The following three methods were compared in the simulations:

**GEE:** Estimators of both $\beta$ and $\alpha$ are obtained by solving the unweighted GEEs (2.16) and (2.17) following Prentice (1988).

**WGEE1:** Estimators of $\beta$ are obtained by solving the weighted GEEs (2.18), but estimators of $\alpha$ are obtained by solving the unweighted GEEs (2.17).

**WGEE2:** Estimators of both $\beta$ and $\alpha$ are obtained by solving the weighted GEEs (2.18) and (2.19).

Note that although the GEE approach of Prentice (1988) has been studied extensively for complete data, little is known about its properties under incomplete data with a stochastic missing data mechanism. We are not aware of any work that studies the performance of the WGEE1 and WGEE2 methods for incomplete longitudinal data.
2.5. SIMULATION STUDY

2.5.1 Response Models for Simulations

For the simulation study, we consider a two-group design configuration with a binary response measured on four occasions. The marginal model for the mean response \( E(Y_{it}|x_{it}, \beta, \alpha) \) is given by

\[
\text{logit} \{ E(Y_{it}|x_{it}, \beta, \alpha) \} = \beta_0 + \beta_1 x_i + \beta_2 t, \quad t = 1, \ldots, 4, \tag{2.43}
\]

where \( x_i \) is a dichotomous covariate indicating the group membership for the \( i \)th individual \((i = 1, \ldots, K)\) observed over a fixed set of \( T = 4 \) time-points, \( t = 1, 2, 3, 4 \).

Throughout the simulations, we consider \( P(X_i = 1) = 0.2 \). The marginal mean \( E(Y_{it}|x_{it}, \beta, \alpha) \) is defined as a function of both \( x_i \) and time-point \( t \).

The simulated data were generated using two types of correlation structures: exchangeable and serial, with the correlation parameter \( \alpha \). For the exchangeable correlation, we chose \( \text{corr}(Y_{it}, Y_{it'}) = \alpha \) and for the serial correlation, we chose \( \text{corr}(Y_{it}, Y_{it'}) = \alpha^{|t-t'|} \) for all \((t, t')\). We employ Bahadur (1961) multivariate binary model to generate the longitudinal data. We estimate the model parameters assuming both exchangeable and serial correlation structures. Under the “true” correlation structure, the “fitted” model assumes the same correlation as that of the true model, whereas under the “misspecified” correlation structure, these two correlations are different.

Throughout the simulations, the regression and association parameters were fixed at \( \beta = (\beta_0, \beta_1, \beta_2)' = (-1, 1, .2)' \) and \( \alpha = 0.5 \), respectively. Each simulation run
2.5. SIMULATION STUDY

was based on 1000 replications of data sets, with each data set containing $K = 500$ subjects and a maximum of $T = 4$ observations per subject.

2.5.2 Dropout Models for Simulations

The dropout model was assumed to be functionally independent of the group membership, but was assumed to be dependent on the current and previous values of the response variable. That is, we assumed that

$$P(M_i = m_i | y_i, x_i, \tau) = P(M_i = m_i | y_{i1}, \ldots, y_{im}, \tau).$$

(2.44)

We assumed that all subjects were measured at the first time-point. Since we observe the individuals at a fixed set of $T = 4$ time-points, the values of $M_i$ can vary between 2 and 5, $m_i = 2, \ldots, 5$.

To calculate $P(M_i = m_i | y_{i1}, \ldots, y_{im}, \tau)$ in (2.44), the individual conditional probabilities of the missing data indicators $R_{it}$ were obtained from

$$P(R_{it} = 0 | R_{i1} = \ldots = R_{i,t-1} = 1, y_{i1}, \ldots, y_{im}, \tau) = \frac{\exp(\tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it})}{1 + \exp(\tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it})},$$

(2.45)

for $t = 2, 3, 4$, i.e., the probability of being observed at a given time is entirely determined by the previous and the current, possibly unobserved, responses. Note that the choice $\tau_C = 0$ leads a MAR model, whereas the choice $\tau_P = \tau_C = 0$ leads to the assumption that the data are missing completely at random (MCAR). For $\tau_C \neq 0$, missingness depends on a current value of the response variable $Y$, and the
missing data become nonignorable (NI).

Using (2.10) and (2.45) we get

\[
\nu_{im_i} = P(M_i = m_i | y_{i1}, \ldots, y_{im_i}, \tau) = \left\{ \prod_{t=2}^{m_i-1} \frac{1}{1 + \exp(\tau_0 + \tau_y y_{i,t-1} + \tau_C y_{it})} \right\} \left\{ \frac{\exp(\tau_0 + \tau_P y_{i,m_i-1} + \tau_C y_{im_i})}{1 + \exp(\tau_0 + \tau_P y_{i,m_i-1} + \tau_C y_{im})} \right\}^{I\{m_i \leq T\}}
\]

(2.46)

2.5.3 Estimating Dropout Probabilities

For calculating the WGEE estimators, prior to solving the iterative equations (2.22) and (2.23), we estimate the response probability weights \( \nu_{im_i} \) using (2.10).

From (2.46) we find the pseudo-likelihood function for \( \tau \) as

\[
L(\tau) = \prod_{i=1}^{K} P(M_i = m_i | y_{i1}, \ldots, y_{im_i}, \tau) = \left\{ \prod_{i=1}^{K} \left\{ \prod_{t=2}^{m_i-1} \frac{1}{1 + \exp(\tau_0 + \tau_y y_{i,t-1} + \tau_C y_{it})} \right\} \right\} \left\{ \frac{\exp(\tau_0 + \tau_P y_{i,m_i-1} + \tau_C y_{im_i})}{1 + \exp(\tau_0 + \tau_P y_{i,m_i-1} + \tau_C y_{im})} \right\}^{I\{m_i \leq T\}}
\]

(2.47)

Let \( \logit\{p^*_\tau(t)\} = \tau_0 + \tau_P y_{i,t-1} + \tau_C y_{it} \). We find the pseudo-ML estimator of \( \tau \) by maximizing the above likelihood function. From (2.47), the pseudo-score equations for \( \tau \) takes the form

\[
S(\tau) = \sum_{i=1}^{K} \left\{ - \sum_{t=2}^{m_i-1} p^*_\tau(t) y^*_t + I(m_i \leq T)\{1 - p^*_\tau/im_i(\tau)\} y^*_im_i \right\} = 0,
\]

(2.48)

where \( y^*_t = (1, y_{i,t-1}, y_{it})' \). The approximate variance of the pseudo-ML estimator of \( \tau \) is obtained from the Information matrix

\[
I(\tau) = \sum_{i=1}^{K} \sum_{t=2}^{\min(m_i,T)} p^*_\tau(t)(1 - p^*_\tau(t)) y^*_t y^*_t.
\]

(2.49)
We apply the Newton-Rapson iterative algorithm to solve the above estimating equations for the pseudo-ML estimator $\hat{\tau} = (\hat{\tau}_0, \hat{\tau}_P, \hat{\tau}_C)'$. The predicted probability of a response for individual $i$ at time $t$ is obtained as

$$1 - \hat{p}_{it}^* = \frac{1}{1 + \exp(y_i^* \hat{\tau})}.$$  \hfill (2.50)

Then we estimate the probability of dropping out for individual $i$ at time $m_i$ by

$$\hat{\nu}_{im_i} = P(M_i = m_i | y_{i1}, \ldots, y_{im_i}, \hat{\tau}) = \left\{ \prod_{t=2}^{m_i-1} (1 - \hat{p}_{it}^*) \right\} \times \left\{ \hat{p}_{i,m_i}^* \right\} I_{\{m_i \leq T\}}.$$  \hfill (2.51)

We replace $\nu_{im_i}$'s in equations (2.22) and (2.23) by $\hat{\nu}_{im_i}$ and then solve these equations iteratively for the WGEE estimators $\tilde{\beta}$ and $\tilde{\alpha}$.

### 2.5.4 Diagnostic methods

We compare three methods based on empirical biases, mean squared errors and coverage probabilities of the estimators. Specifically, the bias of an estimator $\hat{\theta}$ of $\theta$ is estimated by

$$\text{bias}(\hat{\theta}) \approx \frac{1}{S} \sum_{s=1}^{S} \frac{\hat{\theta}_s - \theta}{S},$$  \hfill (2.52)

where $\hat{\theta}_s$ is the estimate of $\theta$ obtained from the $s$th simulated data set and $S$ is the simulation size. We calculate the percentage relative bias as

$$\frac{\text{bias}(\hat{\theta})}{\theta} \times 100.$$  \hfill (2.53)

The mean squared error (MSE) of $\hat{\theta}$ is estimated by

$$\text{MSE}(\hat{\theta}) \approx \frac{1}{S} \sum_{s=1}^{S} \frac{(\hat{\theta}_s - \theta)^2}{S}.$$  \hfill (2.54)
2.5. SIMULATION STUDY

We also find the coverage probabilities of an estimator $\hat{\theta}$ for 95% confidence intervals, $\hat{\theta} \pm 1.96 \times \text{SE}(\hat{\theta})$, where $\text{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$. The empirical coverage probability (CP) is obtained from

$$\text{CP}(\hat{\theta}) = \frac{1}{S} \sum_{s=1}^{S} I \left\{ |\hat{\theta}_s - \theta| \leq 1.96 \times \text{SE}(\hat{\theta}) \right\},$$

(2.55)

where $I \{}$ is an indicator variable.

2.5.5 Results

As mentioned earlier, in the first set of simulations, the data were generated using a MAR model. The parameter values of the dropout model (2.45) were chosen as $\tau = (\tau_0, \tau_P, \tau_C)' = (-2, 2, 0)'$ and $(-2, 3, 0)'$. For these two choices of $\tau$, the data contained roughly 30% and 40% missing values, respectively. Table 2.1 presents the empirical percentage relative biases, mean squared errors, and coverage probabilities of the GEE, WGEE1 and WGEE2 estimators of the regression parameters $\beta = (\beta_0, \beta_1, \beta_2)'$ and the correlation parameter $\alpha$ for $(\tau_0, \tau_P, \tau_C) = (-2, 2, 0)$. Table 2.2 repeats these results for $(\tau_0, \tau_P, \tau_C) = (-2, 3, 0)$.

It is clear from Table 2.1 that both WGEE1 and WGEE2 methods provide approximately unbiased estimates of the regression parameters under all simulation configurations considered. The biases of these estimators increase slightly when the proportion of missing observations increases, as shown in Table 2.2. On the other hand, the unweighted GEE approach, generally provides large biases in the esti-
mates of both the regression parameters and the correlation parameter, as expected. For example, as shown in Table 2.2, under the correctly specified serial correlation structure, the GEE estimator of $\beta_2$ gives $-46.39\%$ relative bias and that of $\alpha$ gives $-28.48\%$ relative bias. When comparing the WGEE1 and WGEE2 methods, our proposed WGEE2 method generally provides smaller bias for the correlation parameter $\alpha$. For example, from Table 2.2, under the correctly specified serial correlation, the WGEE1 estimator of $\alpha$ gives $-14.14\%$ relative bias, whereas the proposed WGEE2 estimator of $\alpha$ gives a much smaller relative bias of $-1.99\%$.

The WGEE2 estimators also give better coverage probabilities; in particular, for the correlation parameter $\alpha$. For example, as shown in Table 2.1, under the correctly specified exchangeable correlation structure, the WGEE2 estimator of $\alpha$ gives an empirical coverage probability of 97%, which is close to the nominal 95% confidence level. On the other hand, the GEE and WGEE1 estimators of $\alpha$ give empirical coverage probabilities of 35% and 49%, respectively.

The mean squared errors of the WGEE2 estimators are also smaller as compared to the WGEE1 and GEE estimators. Under misspecified correlation structures, although the WGEE2 method provides slightly larger biases, but these biases are still smaller than those obtained by the GEE and WGEE1 methods. In this sense, the WGEE2 method is considered to be more robust than the other two methods.
Table 2.1: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout. [True parameter values: $\beta = (-1, 1, .2)^T$, $\alpha = 0.5$ and $\tau = (-2, 2, 0)^T$.]

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
<th>Method</th>
<th>$%$ Relative Bias</th>
<th>MSE</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Exchangeable</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>3.02</td>
<td>0.35</td>
<td>-15.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-0.08</td>
<td>0.93</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-0.04</td>
<td>0.99</td>
<td>-0.09</td>
</tr>
<tr>
<td>Serial</td>
<td>GEE</td>
<td></td>
<td>9.67</td>
<td>-0.32</td>
<td>-37.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-0.07</td>
<td>0.96</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-0.02</td>
<td>1.01</td>
<td>-0.20</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>-5.27</td>
<td>1.21</td>
<td>6.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-0.68</td>
<td>0.87</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-0.66</td>
<td>0.87</td>
<td>0.62</td>
</tr>
<tr>
<td>Serial</td>
<td>GEE</td>
<td></td>
<td>2.20</td>
<td>0.77</td>
<td>-13.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-0.50</td>
<td>0.83</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-0.51</td>
<td>0.84</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Table 2.2: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under MAR dropout. [True parameter values: $\beta = (-1, 1, 2)'$, $\alpha = 0.5$ and $\tau = (-2, 3, 0)'$.]

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
<th>Method</th>
<th>% Relative Bias</th>
<th>MSE</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Exchangeable</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>9.46</td>
<td>-1.89</td>
<td>-52.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-1.63</td>
<td>-1.63</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-1.65</td>
<td>-1.34</td>
<td>1.07</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>23.69</td>
<td>-3.70</td>
<td>-105.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-1.14</td>
<td>-1.64</td>
<td>-0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-1.27</td>
<td>-1.25</td>
<td>-0.48</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>-4.12</td>
<td>-1.46</td>
<td>-6.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-2.06</td>
<td>-2.27</td>
<td>3.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-1.98</td>
<td>-1.97</td>
<td>2.97</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>8.55</td>
<td>-2.76</td>
<td>-46.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-1.78</td>
<td>-2.38</td>
<td>2.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-1.87</td>
<td>-2.00</td>
<td>2.82</td>
</tr>
</tbody>
</table>
Table 2.3: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout. [True parameter values: $\beta = (-1, 1, .2)'$, $\alpha = 0.5$ and $\tau = (-2, 2, 0.5)'$].

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
<th>Method</th>
<th>% Relative Bias</th>
<th>MSE</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\beta_0$</td>
<td>$\beta_1$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>Exchangeable</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>9.29</td>
<td>-1.54</td>
<td>-56.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>0.18</td>
<td>1.00</td>
<td>-0.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>0.23</td>
<td>1.06</td>
<td>-0.75</td>
</tr>
<tr>
<td>Serial</td>
<td></td>
<td>GEE</td>
<td>20.31</td>
<td>-2.75</td>
<td>-93.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>0.20</td>
<td>0.95</td>
<td>-0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>0.24</td>
<td>1.03</td>
<td>-0.85</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>3.60</td>
<td>-0.92</td>
<td>-37.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>9.12</td>
<td>-1.45</td>
<td>-44.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>9.28</td>
<td>-1.56</td>
<td>-45.23</td>
</tr>
<tr>
<td>Serial</td>
<td></td>
<td>GEE</td>
<td>12.73</td>
<td>-1.25</td>
<td>-61.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>9.48</td>
<td>-0.83</td>
<td>-45.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>9.69</td>
<td>-0.85</td>
<td>-45.51</td>
</tr>
</tbody>
</table>
Table 2.4: Empirical percentage relative biases, mean squared errors, and coverage probabilities of GEE, WGEE1 and WGEE2 estimators under NI dropout. [True parameter values: $\beta = (-1, 1, .2)'$, $\alpha = 0.5$ and $\tau = (-2, 2, 1)'$.]

<table>
<thead>
<tr>
<th>True Model</th>
<th>Fitted Model</th>
<th>Method</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\alpha$</th>
<th>MSE $\beta_0$</th>
<th>MSE $\beta_1$</th>
<th>MSE $\beta_2$</th>
<th>Coverage Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchangeable</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>18.82</td>
<td>-3.29</td>
<td>-114.36</td>
<td>-43.73</td>
<td>0.051</td>
<td>0.038</td>
<td>0.055</td>
<td>0.69 0.95 0.055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>-0.18</td>
<td>-0.59</td>
<td>-2.62</td>
<td>-15.78</td>
<td>0.045</td>
<td>0.127</td>
<td>0.004</td>
<td>0.012 0.94 0.92 0.96 0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>-0.21</td>
<td>-0.26</td>
<td>-2.64</td>
<td>-3.98</td>
<td>0.045</td>
<td>0.124</td>
<td>0.004</td>
<td>0.007 0.94 0.93 0.96 0.87</td>
</tr>
<tr>
<td>Serial</td>
<td>Serial</td>
<td>GEE</td>
<td>33.42</td>
<td>-5.03</td>
<td>-165.40</td>
<td>-36.47</td>
<td>0.130</td>
<td>0.039</td>
<td>0.113</td>
<td>0.036 0.31 0.95 0.00 0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>0.30</td>
<td>-0.49</td>
<td>-4.78</td>
<td>3.49</td>
<td>0.042</td>
<td>0.136</td>
<td>0.005</td>
<td>0.009 0.94 0.92 0.96 0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>0.20</td>
<td>-0.05</td>
<td>-4.49</td>
<td>14.74</td>
<td>0.043</td>
<td>0.132</td>
<td>0.005</td>
<td>0.013 0.95 0.92 0.96 0.81</td>
</tr>
<tr>
<td>Serial</td>
<td>Exchangeable</td>
<td>GEE</td>
<td>10.77</td>
<td>-3.49</td>
<td>-91.95</td>
<td>-57.84</td>
<td>0.032</td>
<td>0.038</td>
<td>0.038</td>
<td>0.085 0.87 0.94 0.12 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>19.07</td>
<td>-13.67</td>
<td>-98.61</td>
<td>-41.80</td>
<td>0.089</td>
<td>0.111</td>
<td>0.048</td>
<td>0.049 0.92 0.89 0.51 0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>18.95</td>
<td>-13.90</td>
<td>-98.16</td>
<td>-56.53</td>
<td>0.088</td>
<td>0.115</td>
<td>0.048</td>
<td>0.082 0.92 0.89 0.49 0.00</td>
</tr>
<tr>
<td>Serial</td>
<td>Serial</td>
<td>GEE</td>
<td>21.08</td>
<td>-4.30</td>
<td>-120.31</td>
<td>-42.37</td>
<td>0.066</td>
<td>0.038</td>
<td>0.062</td>
<td>0.047 0.68 0.94 0.03 0.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE1</td>
<td>20.21</td>
<td>-13.59</td>
<td>-99.95</td>
<td>-37.82</td>
<td>0.093</td>
<td>0.113</td>
<td>0.049</td>
<td>0.039 0.91 0.88 0.48 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WGEE2</td>
<td>20.61</td>
<td>-13.27</td>
<td>-100.63</td>
<td>-20.98</td>
<td>0.095</td>
<td>0.109</td>
<td>0.050</td>
<td>0.017 0.91 0.89 0.48 0.67</td>
</tr>
</tbody>
</table>
2.6. APPLICATION: ANALYSIS OF SMOKING DATA

In the second set of simulations, we study the effects of misspecified missing data models on the estimators of the regression and association parameters. As before, the data were generated from the binary longitudinal model (2.43), but with a non-ignorable missing data model (2.45) with non-zero \( \tau_C = 0.5, 1.0 \). Table 2.3 presents the empirical percentage relative biases, MSEs and coverage probabilities of the estimators for \( \tau = (\tau_0, \tau_P, \tau_C) = (-2, 2, 0.5) \) for which the data contained roughly 32% missing observations. Table 2.4 repeats the results for \( \tau = (\tau_0, \tau_P, \tau_C) = (-2, 1, 1) \) for which the data contained roughly 40% missing observations. In both cases, we estimated the model parameters under the misspecified MAR model.

It is clear from Tables 2.3 and 2.4 that all three methods provide biased estimators of the model parameters under the misspecified missing data model. However, the extent of the bias from the WGEE2 method is less severe as compared to the GEE and WGEE1 methods. The WGEE2 method also provides smaller MSEs and better coverage probabilities as compared to the other two methods.

2.6 Application: Analysis of Smoking Data

We present an analysis of data on cigarette smoking trends from the Coronary Artery Development in Young Adults (CARDIA) study, which is an epidemiological study that recorded cardiovascular risk factors on five occasions over a 10-year period in black and white males and females (Hughes et al., 1987). This study
was conducted in four urban centres (Birmingham, AL; Chicago, IL, Minneapolis, MN; and Oakland, CA) across the United States in which a total of 5,115 young adults aged 18-30 years were followed prospectively and examined up to five times from 1986 to 1996. Recruitment, restricted to blacks and whites, was carried out to achieve approximate balance in sample size with respect to age, race, gender, and education. Study participants were scheduled for visits at years 0, 2, 5, 7, and 10. We consider the first four visits and 5,078 (99.3%) young adults with self reported smoking status (yes/no) known at baseline (year 0). Data from person-exams occurring after a person’s first missed exam were omitted to create a data set with monotone missingness. Specifically, 578 person-exams of a total of 17,995 were omitted to create a monotone data set.

Here the goal is to draw inferences on the change in smoking prevalence of young adults in the presence of missing data and intraperson correlation based on our proposed method. The model parameters were estimated and compared using the unweighted GEE, WGEE1 and the proposed WGEE2 methods.

Let the binary response variable \( Y_{it} = 1 \) if the \( i \)th individual is smoker at the \( t \)th visit, and 0 if he/she is a nonsmoker. The marginal mean response \( p_{it} = E(Y_{it}) \) is defined as a function of the covariates in the form

\[
\text{logit}(p_{it}) = \beta_0 + \beta_1(\text{age}/10)_i + \beta_2x_t + \beta_3\text{edu}_i + \beta_4\text{edu}_i
\]

\[+\beta_5\text{racebf}_i + \beta_6\text{racewm}_i + \beta_7\text{racewf}_i, \tag{2.56}\]
for \( i = 1, \ldots, 5,078 \) and \( t = 1, \ldots, 4 \), where \( \text{age}_i = \text{age of individual } i \) in years at baseline time; \( x_t = \text{year since the baseline measurement} = 0, 2, 5, 7 \); the binary indicators \( \text{eduh}_i = 1 \) if \( i \)th individual's education level is high school or less, and 0 otherwise; \( \text{educ}_i = 1 \) if education level is up to some college, and 0 otherwise; \( \text{racebf}_i = 1 \) if the person is black female, and 0 otherwise; \( \text{racewm}_i = 1 \) if the person is white male, and 0 otherwise; and \( \text{racewf}_i = 1 \) if the person is white female, and 0 otherwise.

To estimate the model parameters in (2.56) by the WGEE methods, we first estimate the inverse probability weights based on the missing data model

\[
\logit(p^*_it) = \tau_0 + \tau_1 y_{i,t-1} + \tau_2 \text{racebf}_i + \tau_3 \text{racewm}_i + \tau_4 \text{racewf}_i,
\]

(2.57)

where \( p^*_it = P(R_{it} = 0| R_{i1} = \ldots, R_{i,t-1} = 1, y_i, x_i, \tau) \) is the conditional probability that the \( i \)th individual drops out at time \( t \).

Table 2.5 presents the pseudo-maximum likelihood estimates of the missing data model parameters \( \tau = (\tau_0, \tau_1, \tau_2, \tau_3, \tau_4) \), their standard errors, and the corresponding \( z \)-values. Results in this table suggest that the dropout probabilities vary across the race and gender as well as the smoking status of an individual at the previous visit. Young adults are likely to have \( \exp(0.389) = 1.47 \) times higher odds to miss a visit if they were a smoker (vs. nonsmoker) at the previous visit. Also the study suggests that the black males are more likely to miss a visit than any other race-gender combinations. The seven-year follow-up rates for the CARDIA study
were 62%, 68%, 81% and 79% for black males, black females, white males and white females, respectively. The fitted missing data model appears to reflect this scenario.

We finally find the GEE, WGEE1, and WGEE2 estimators of the regression parameters in the marginal model (2.56) assuming both exchangeable and serial correlation structures. The estimates of the model parameters, their standard error and the corresponding $z$-values are presented in Table 2.6. From these results, the ordinary GEE approach appears to give somewhat different results as compared to the WGEE1 and WGEE2 approaches. For example, under the exchangeable correlation structure, the covariate age has the coefficient 0.320 by the GEE method, whereas this coefficient is 0.481 by both WGEE1 and WGEE2 methods. The results are generally very close by the two weighted GEE methods, except for the correlation coefficient $\alpha$. For example, the estimate of $\alpha$ under the exchangeable correlation structure is 0.727 by the WGEE2 method, whereas the estimate is 0.706 by the WGEE1 method and 0.709 by the unweighted GEE method.
### Table 2.5: ML Estimates and Standard Errors of Missing Data Model Parameters for the CARDIA Study.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>S.E</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.923</td>
<td>0.038</td>
<td>-51.28</td>
</tr>
<tr>
<td>Previous smoking status</td>
<td>0.389</td>
<td>0.037</td>
<td>10.57</td>
</tr>
<tr>
<td>Race (Black Female)</td>
<td>-0.197</td>
<td>0.046</td>
<td>-4.25</td>
</tr>
<tr>
<td>Race (White Male)</td>
<td>-0.809</td>
<td>0.053</td>
<td>-15.18</td>
</tr>
<tr>
<td>Race (White Female)</td>
<td>-0.722</td>
<td>0.062</td>
<td>-11.68</td>
</tr>
</tbody>
</table>

### Table 2.6: Estimates and Standard Errors of Regression and Association Parameters for the CARDIA Study.

<table>
<thead>
<tr>
<th>Fitted Model</th>
<th>Covariates</th>
<th>GEE</th>
<th></th>
<th>WGE1</th>
<th></th>
<th>WGE2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate S.E</td>
<td>z-value</td>
<td>Estimate S.E</td>
<td>z-value</td>
<td>Estimate S.E</td>
<td>z-value</td>
<td></td>
</tr>
<tr>
<td>Exchangeable</td>
<td>Intercept</td>
<td>-2.544</td>
<td>0.221</td>
<td>-11.51</td>
<td>-2.781</td>
<td>0.326</td>
<td>-8.52</td>
</tr>
<tr>
<td></td>
<td>Age/10</td>
<td>0.320</td>
<td>0.082</td>
<td>3.91</td>
<td>0.481</td>
<td>0.119</td>
<td>4.04</td>
</tr>
<tr>
<td></td>
<td>Year Followup</td>
<td>-0.014</td>
<td>0.004</td>
<td>-3.35</td>
<td>-0.017</td>
<td>0.006</td>
<td>-2.74</td>
</tr>
<tr>
<td></td>
<td>Education (High School or less)</td>
<td>1.924</td>
<td>0.085</td>
<td>22.61</td>
<td>1.730</td>
<td>0.129</td>
<td>13.40</td>
</tr>
<tr>
<td></td>
<td>Education (Some college)</td>
<td>1.169</td>
<td>0.079</td>
<td>14.77</td>
<td>0.903</td>
<td>0.123</td>
<td>7.35</td>
</tr>
<tr>
<td></td>
<td>Race (Black Female)</td>
<td>-0.201</td>
<td>0.080</td>
<td>-2.52</td>
<td>-0.285</td>
<td>0.103</td>
<td>-2.77</td>
</tr>
<tr>
<td></td>
<td>Race (White Male)</td>
<td>-0.111</td>
<td>0.091</td>
<td>-1.22</td>
<td>-0.079</td>
<td>0.135</td>
<td>-0.58</td>
</tr>
<tr>
<td></td>
<td>Race (White Female)</td>
<td>-0.104</td>
<td>0.089</td>
<td>-1.18</td>
<td>-0.088</td>
<td>0.126</td>
<td>-0.70</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>0.709</td>
<td>0.017</td>
<td>40.92</td>
<td>0.706</td>
<td>0.016</td>
<td>43.56</td>
</tr>
<tr>
<td>Serial</td>
<td>Intercept</td>
<td>-2.621</td>
<td>0.223</td>
<td>-11.77</td>
<td>-2.773</td>
<td>0.323</td>
<td>-8.58</td>
</tr>
<tr>
<td></td>
<td>Age/10</td>
<td>0.343</td>
<td>0.082</td>
<td>4.17</td>
<td>0.472</td>
<td>0.117</td>
<td>4.02</td>
</tr>
<tr>
<td></td>
<td>Year Followup</td>
<td>-0.015</td>
<td>0.004</td>
<td>-3.58</td>
<td>-0.021</td>
<td>0.008</td>
<td>-2.69</td>
</tr>
<tr>
<td></td>
<td>Education (High School or less)</td>
<td>1.943</td>
<td>0.086</td>
<td>22.61</td>
<td>1.737</td>
<td>0.128</td>
<td>13.55</td>
</tr>
<tr>
<td></td>
<td>Education (Some college)</td>
<td>1.177</td>
<td>0.080</td>
<td>14.71</td>
<td>0.910</td>
<td>0.122</td>
<td>7.48</td>
</tr>
<tr>
<td></td>
<td>Race (Black Female)</td>
<td>-0.207</td>
<td>0.080</td>
<td>-2.57</td>
<td>-0.274</td>
<td>0.102</td>
<td>-2.70</td>
</tr>
<tr>
<td></td>
<td>Race (White Male)</td>
<td>-0.115</td>
<td>0.092</td>
<td>-1.25</td>
<td>-0.060</td>
<td>0.134</td>
<td>-0.45</td>
</tr>
<tr>
<td></td>
<td>Race (White Female)</td>
<td>-0.095</td>
<td>0.089</td>
<td>-1.06</td>
<td>-0.082</td>
<td>0.126</td>
<td>-0.65</td>
</tr>
<tr>
<td></td>
<td>α</td>
<td>0.809</td>
<td>0.013</td>
<td>62.00</td>
<td>0.805</td>
<td>0.012</td>
<td>66.09</td>
</tr>
</tbody>
</table>
2.7. CONCLUSIONS

It is clear from Table 2.6 that the covariates Age, Year Followup, and Level of Education have significant influence on the smoking trend in young adults. The older individuals are more likely to be a smoker, but there is an overall trend of quitting this habit over time. These results also suggest that the level of education has strong influence on the smoking status of the subjects. For example, the young adults are estimated to have $\exp(1.728) = 5.64$ times higher odds to be a smoker if their level of education is up to high school or less than those who have a college degree or more. The results are somewhat similar under the serial and exchangeable correlation structures. However, the serial correlation structure may be preferable here, as this appears to give slightly smaller standard errors of the estimates as compared to the exchangeable correlation structure.

2.7 Conclusions

The purpose of this research was to provide a better alternative to the unweighted GEE models of Prentice (1988) for analyzing incomplete longitudinal data. Our simulation study demonstrates that the proposed WGEE2 approach generally provides unbiased and efficient estimators when the missing data mechanism follows a MAR model. When the missing data are nonignorable, all three methods give biased estimators of the model parameters. However, the extent of the bias from the WGEE2 method is generally less severe as compared to the unweighted
GEE and WGEE1 methods.
Chapter 3

Semiparametric Generalized Estimating Equations

3.1 Introduction

The often used linear regression paradigm models a dependent variable $Y$ as a linear function of a vector-valued independent variable $x$. Namely, for some vector $\beta$, we assume that $E[Y|x] = x'\beta$. Generalized linear models (GLMs) provide a flexible extension of linear regression by assuming that the dependent variable $Y$ is of the form, $E[Y|x] = \varphi(x'\beta)$; $\varphi(.)$ is referred to as the inverse link function or transfer function. Generalized linear models include commonly used regression techniques such as the logistic regression, where $\varphi(z) = 1/(1 + e^z)$ is the logistic function.
In the case of linear regression, the least-squares method is a highly efficient procedure for parameter estimation. Unfortunately, in the case of GLMs, even in the setting when $\varphi(.)$ is known, the problem of fitting a model that minimizes squared error is typically not convex. The standard procedure is iteratively reweighted least squares based on the Newton-Raphson method.

The case when both $\varphi(.)$ and $\beta$ are unknown is referred to as Single Index Models (SIM). These models involve challenging problem of jointly estimating $\varphi(.)$ and $\beta$, where $\varphi(.)$ may come from a large non-parametric family such as the family of all monotonic functions.

In practice, the relationship between the response and covariates may be very complex and linear terms may not be adequate to feature that relationship. Under these circumstances, it might be preferable to include a linear term $x'\beta$ and a nonlinear term $\varphi(u'\xi)$ in a semiparametric regression, where $\varphi(.)$ is a smooth but unknown function. These models are a natural combination and generalization of simpler models already in the literature, namely single-index models and partially linear models. Carroll et al. (1997) called it generalized partially linear single-index models (GPLSIM) . Li (1991) noted that if $\varphi(.)$ is monotone, then $\xi$ takes on the same general meaning as “effect” parameters as would occur in ordinary linear models. Furthermore, it is a widely applied method to include a nonparametric function into the model for covariates $u$ that have large dimension and are of little interest (e.g., confounders). This allows us to make inference on the effects of $x$
while making minimal assumptions on $u$. Marginal semiparametric models based on using GEE methods and their various extensions have become increasingly popular. See, for example, Grace et al. (2009), Wang et al. (2005), Fan and Li (2004) and Carrol et al. (1997) among others. Carrol et al. (1997) propose estimates of unknown parameters $\beta, \xi$ and unknown function $\varphi(.)$ using quasi-likelihood and local polynomial regression and obtain their asymptotic distributions. However, they did not consider the joint estimation of marginal mean parameters and the association parameters.

In many applications, simply working on the marginal mean responses could be very restrictive. Estimation of the association parameters may be the central theme of the study. For example, in familial studies of inherited traits and developmental toxicology studies of laboratory animals, subjects in a family or cluster share common genetic traits or are subject to common environmental factors, and hence it is of prime scientific interest to study the association between responses.

In this chapter, we propose a method based on Prentice’s (1988) second order GEE approach to analyze longitudinal binary responses under partially linear single-index models. The computing algorithm for solving usual estimating equations based on the Newton-Raphson method cannot be directly employed due to the inclusion of a nonlinear function whose form is unknown. To circumvent this problem, we use the Carrol et al.’s (1997) local polynomial smoothing technique to perform estimation of the single index. We study the empirical properties of the
proposed method in simulations.

This chapter is organized as follows. Section 3.2 introduces the model and notation to define the generalized linear and partially linear single-index models for binary longitudinal data. Section 3.3 reviews the GEE approach and describes second order GEE estimation of partially linear single-index models. Section 3.4 discusses asymptotic properties of the proposed GEE estimator. Section 3.5 presents results from a simulation study, which was carried out to investigate the empirical properties of the weighted GEE approach. Section 3.6 presents applications of the proposed method using two different actual longitudinal data sets. Section 3.7 gives the conclusions of the chapter.

3.2 Model and Notation

3.2.1 Response Models

3.2.1.1 Binary Logistic Model

Suppose \( K \) subjects are observed at a fixed set of \( T \) time points. Let \( Y_{it} \) represent a binary response variable from subject \( i, i = 1, \ldots, K \), at visit \( t, t = 1, \ldots, T \). For the \( i \)th subject, we can form a \( T \times 1 \) vector, \( Y_i = (Y_{i1}, \ldots, Y_{iT})' \), of binary response variables. Let the lower case letters \( y_{it} \) and \( y_i \) denote the realizations of \( Y_{it} \) and \( Y_i \), respectively. Also, let \( x_{it} = (x_{it,1}, \ldots, x_{it,p})' \) be a \( p \times 1 \) vector of covariates from
subject \(i\) at time \(t\). The covariates may be time-dependent or fixed across the entire observation times. Let \(x_i = (x_{i1}, \ldots, x_{iT})'\).

Assume that the marginal distribution of \(Y_{it}\) is Bernoulli:

\[
Y_{it} \sim \text{Bernoulli}(p_{it}), \ i = 1, \ldots, K; \ t = 1, \ldots, T, \tag{3.1}
\]

with the probability of success,

\[
p_{it} = E(Y_{it}|x_i) = P(Y_{it} = 1|x_i). \tag{3.2}
\]

Let \(p_i = (p_{i1}, \ldots, p_{iT})'\). Assuming that the mean of response variable \(Y_{it}\) depends only on the covariate vector for subject \(i\) at time \(t\), i.e., \(p_{it} = E(Y_{it}|x_i) = E(Y_{it}|x_{it})\) (Pepe and Anderson, 1994), we consider modelling the mean response by the logistic regression:

\[
\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1 - p_{it}}\right) = x_{it}'\beta, \tag{3.3}
\]

where \(\beta\) is the \(p \times 1\) vector of regression parameters. The marginal variance of the response variable \(Y_{it}\) is specified as a function of the marginal mean as

\[
v_{it} = \text{var}(Y_{it}|x_{it}) = p_{it}(1 - p_{it}). \tag{3.4}
\]

We assume that \(Y_{it}\) and \(Y_{it'}\) are uncorrelated when \(i \neq i'\). Let

\[
\text{corr}(Y_{it}, Y_{it'}) = \alpha_{it'} \tag{3.5}
\]

represent the correlation between \(Y_{it}\) and \(Y_{it'}\) for given \(x_{it}\), where \(\alpha = (\alpha_{12}, \ldots, \alpha_{1T}, \alpha_{23}, \ldots, \alpha_{T-1,T})'\) is the vector of correlation parameters.
3.2. MODEL AND NOTATION

3.2.1.2 Partially Logistic Linear Single-index Model

Let \( u_{it} = (u_{it,1}, \ldots, u_{it,q})' \) be a \( q \times 1 \) vector of another set of covariates from subject \( i \) at time \( t \). The covariates may be time-dependent or fixed across the entire observation times. Let \( u_i = (u'_{i1}, \ldots, u'_{IT})' \). Define \( p_{it} = E(Y_{it}|x_i, u_i) = P(Y_{it} = 1|x_i, u_i) \). Provided the mean of \( Y_{it} \) depends only on the covariate vector for subject \( i \) at time \( t \), i.e., \( p_{it} = E(Y_{it}|x_i, u_i) = E(Y_{it}|x_{it}, u_{it}) \), we consider a partially linear single-index logistic model

\[
\logit(p_{it}) = \log \left( \frac{p_{it}}{1 - p_{it}} \right) = x'_{it}\beta + \varphi(u'_{it}\xi) \quad \text{with} \quad ||\xi|| = 1
\] (3.6)

where \( \beta \) and \( \xi \) are \( p \times 1 \) and \( q \times 1 \) unknown parameter vectors, respectively, and \( \varphi(.) \) is an unknown smooth function. The restriction \( ||\xi|| = 1 \) ensures identifiability of \( \xi \). We note that model (3.6) generalizes the usual logistic regression in a sense that a nonlinear term \( \varphi(u'_{it}\xi) \) is included in the model. This model has the same marginal variance and correlation structure as (3.4) and (3.5). If \( \varphi(.) \) is specified as the identity function, then model (3.6) becomes an ordinary logistic regression like (3.3).
3.3 Methods of Estimation

3.3.1 Generalized Estimating Equation

Our primary interest lies in estimating the regression parameters $\beta$, $\xi$ as well as the association parameters $\alpha$. As a naive method, one can consider analyzing the longitudinal data by simply assuming $\varphi(.)$ as the identity function and then estimating the model parameters based on the linear single index model. In such a case, the generalized estimating equations (GEEs) of Liang and Zeger (1986) can be used for estimating the regression parameters $(\beta, \xi)$:

$$U_{\beta, \xi}(\beta, \xi, \alpha) = \sum_{i=1}^{K} D_i' V_i^{-1} (Y_i - p_i) = 0,$$

(3.7)

where $D_i = (\partial p_i/\partial \beta, \partial p_i/\partial \xi)'$, $V_i = (B_i)^{1/2} R(\alpha)(B_i)^{1/2}$, $B_i = \text{diag}\{p_{i1}(1-p_{i1}), p_{i2}(1-p_{i2}), \ldots, p_{iT}(1-p_{iT})\}$ and $R(\alpha)$ is a “working” correlation matrix for $Y_i$ depending on the vector $\alpha$ of correlation parameters. The above equations can be solved numerically for $\hat{\beta}$ and $\hat{\xi}$ using an iterative method.

Liang and Zeger (1986) considered estimating the association parameters $\alpha$ by the method of moments, which uses the Pearson residuals

$$\hat{r}_{it} = \frac{y_{it} - \hat{p}_{it}}{(1 - \hat{p}_{it})^{1/2}}.$$  

(3.8)

The moment estimators of $\alpha_{it'}$ may be obtained as

$$\hat{\alpha}_{it'} = \sum_{i=1}^{K} \hat{r}_{it} \hat{r}_{it'}/(K - (p + q)).$$  

(3.9)
where $p$ is the dimension of $\beta$ and $q$ is the dimension of $\xi$.

Prentice (1988) considered an extension of the GEE approach to allow joint estimation of the regression parameters $\beta$ and the association parameters $\alpha$. Specifically, a GEE estimator of the correlation parameter $\alpha$ may be obtained from a second set of estimating equations by noting that the “sample correlation”

$$Z_{itu} = Z_{itu}(\beta) = \frac{(Y_{it} - p_{it})(Y_{iu} - p_{iu})}{(p_{it}q_{it}p_{iu}q_{iu})^{1/2}}$$

has mean $\rho_{itu}$ and variance

$$w_{itu} = 1 + (1 - 2p_{it})(1 - 2p_{iu})(p_{it}q_{it}p_{iu}q_{iu})^{-1/2}(\rho_{itu} - \rho_{itu}^2),$$

for $t < u < T$, $i = 1, \ldots, K$, and $t = 1, \ldots, T$. Let $Z_i = (Z_{i12}, \ldots, Z_{iT}, Z_{i23}, \ldots, Z_{i,T-1,T})'$ and $\rho_i = (\rho_{i12}, \ldots, \rho_{iT}, \rho_{i23}, \ldots, \rho_{i,T-1,T})'$.

Then following Prentice (1988), the GEE estimators of $(\beta, \xi, \alpha)$ may be obtained by solving the estimating equations

$$U_{\beta,\xi}(\beta, \xi, \alpha) = \sum_{i=1}^{K} D_i^{'} V_i^{-1} (Y_i - p_i) = 0,$$

$$U_{\alpha}(\beta, \xi, \alpha) = \sum_{i=1}^{K} G_i^{'} W_i^{-1} (Z_i - \rho_i) = 0,$$

where $G_i = \partial \rho_i / \partial \alpha$ and $W_i = \text{diag}\{w_{i12}, \ldots, w_{iT}, w_{i23}, \ldots, w_{i,T-1,T}\}$.

### 3.3.2 GEE for Partially Linear Single-index Model

For a partially linear single-index model, the estimating functions $U_{\beta,\xi}(\beta, \xi, \alpha, \varphi(.))$ and $U_{\alpha}(\beta, \xi, \alpha, \varphi(.))$ involve an unknown smooth function $\varphi(.)$. We use nonpara-
3.3. METHODS OF ESTIMATION

metric approaches to estimate this function locally in order to estimate $\beta$, $\xi$ and $\alpha$. Assuming $\varphi(c)$ has a second derivative, we may approximate $\varphi(c)$ by a locally linear function within the neighborhood of $c_0$ via the Taylor series expansion $\varphi(c) \approx \varphi(c_0) + \varphi^{(1)}(c_0)(c - c_0)$, for a given point $c_0$, where $\varphi^{(1)}(c_0)$ is the first order differentiation of $\varphi(c)$ with respect to $c$ at $c_0$. Denote $\eta_0(c_0) = \varphi(c_0)$, $\eta_1(c_0) = \varphi^{(1)}(c_0)$, and $\eta(c_0) = (\eta_0(c_0), \eta_1(c_0))'$.

We further introduce the following notation: $C_{it} = \mathbf{u}'_{it}\xi$, $\phi_{it}(c, \xi) = (1, C_{it} - c)'$, $\Phi_i(c)$ is a $t \times 2$ matrix with the $t$th row $\phi_{it}(c, \xi)$, and $\Delta_i = \text{diag}(p_{it}^{(1)}, t = 1, \ldots, T)$, where $p_{it}^{(1)}$ is the first order derivative of the mean function evaluated at $x_{it}'\beta + \varphi(u_{it}'\xi)$. Let $K(c)$ be a kernel function (or a symmetric density function) with a compact support and $h$ be a bandwidth. Denote $K_h(a) = K(a/h)/h$ and $\mathbf{K}_{ih}(c) = \text{diag}\{K_h(C_{it} - c), t = 1, \ldots, T\}$.

Using above notations, we describe below an algorithm for simultaneous estimation of the mean parameters $\beta$, $\xi$ and the correlation parameters $\alpha$. Here we adopt the local linear kernel method and Prentice's (1988) approach to conduct simultaneous estimation of $\varphi$, $\beta$, $\xi$ and $\alpha$.

**Step 0.** (Initialization step). Fit a parametric logistic model to obtain initial values $(\beta_0, \xi_0, \alpha_0)$ and set $\hat{\xi} = \xi_0/\|\xi_0\|$, $\hat{\beta} = \beta_0$ and $\hat{\alpha} = \alpha_0$.

**Step 1.** For a given point $c$ in a selected grid find $\varphi(c, \hat{\alpha}, \hat{\beta}) = \tilde{\eta}_0(c)$ and $\tilde{\eta}_1(c)$ by solving
the following equations
\[ \sum_{i=1}^{K} \hat{\Phi}_i(c) \hat{\Delta}_i \hat{K}_{ih}(c) \hat{V}_i^{-1} (Y_i - \hat{p}_i) = 0 \]  
(3.14)

with respect to \( \eta(c) \), where \( \hat{\Phi}_i(c) \) and \( \hat{K}_{ih}(c) \) are respectively \( \Phi_i(c) \) and \( K_{ih}(c) \) with \( \xi \) replaced by \( \hat{\xi} \), \( \hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{it}) \), \( \hat{\rho}_i = g\{x_{it}' \hat{\beta} + \phi_i(c, \hat{\xi}) \eta(c)\} \), \( g(a) = \text{expit}(a) = \exp(a) / \{1 + \exp(a)\} \), \( \hat{\Delta}_i \) is \( \Delta_i \) with \( p_{it} \) replaced by \( \hat{p}_{it} \), and \( \hat{V}_i \) is the working independence diagonal matrix with diagonal elements \( \text{diag}\{\hat{p}_{it}(1 - \hat{p}_{it})\} \).

**Step 2.** Given the estimate \( \hat{\phi}(c; \hat{\beta}, \hat{\xi}, \hat{\alpha}) = \hat{\eta}_0(c) \) and \( \hat{\eta}_1(c) \) for points \( c \) in the selected grid, update \((\hat{\beta}, \hat{\xi}, \hat{\alpha})\) by solving the following equations for \( \beta, \xi \) and \( \alpha \):

\[ \hat{U}_{\beta}(\beta, \xi, \alpha, \hat{\phi}(.) = \sum_{i=1}^{K} \hat{D}_i' \hat{V}_i^{-1} \{Y_i - \hat{p}_i(\beta, \xi, \alpha)\} = 0, \]  
(3.15)

\[ \hat{U}_{\alpha}(\beta, \xi, \alpha, \hat{\phi}(.) = \sum_{i=1}^{K} G_i' \hat{W}_i^{-1} (\hat{Z}_i - \hat{\rho}_i) = 0, \]  
(3.16)

where \( \hat{p}_i(\beta, \xi, \alpha) = (\hat{p}_{i1}(\beta, \xi, \alpha), \ldots, \hat{p}_{iT}(\beta, \xi, \alpha))^\prime \) with

\[ \hat{p}_{it}(\beta, \xi, \alpha) = g \left( x_{it}' \hat{\beta} + \hat{\varphi}(u_{it} \xi; \hat{\beta}, \hat{\xi}, \hat{\alpha}) \right), \]  
(3.17)

\( \hat{D}_i, \hat{V}_i, \hat{Z}_i \) and \( \hat{W}_i \) are the same as \( D_i, V_i, Z_i \) and \( W_i \) in (3.12) and (3.13) respectively with \( p_i \) replaced by \( \hat{p}_i(\beta, \xi, \alpha) \). Also \( \rho_i \) is the same as in Equation (3.13).

**Step 3.** Repeat steps 1 and 2 until the convergence of \((\hat{\beta}, \hat{\xi}, \hat{\alpha})\).
3.3. METHODS OF ESTIMATION

When implementing the foregoing algorithm it is often feasible to choose an initial value that is the estimate obtained from fitting a GEE model assuming a \( \varphi(\cdot) \) as the identity function and then estimating the model parameters based on a linear single index model or an ordinary logistic regression model. Our computational experience suggests that the algorithm is not severely but somewhat sensitive to the choice of initial values. The convergence criteria may be based on the difference between the estimates of two successive iterations or the absolute values of the estimating functions evaluated at the previous iteration.

We conclude this section with a discussion on the bandwidth selection. As bandwidth \( h \) affects both the bias and variance of an estimator, there is always a trade-off between these two inference criteria. Bias correction requires the choice of a relatively small bandwidth, whereas a smaller variance estimate needs a larger value of the bandwidth. In principle, the bandwidth selection is data driven, and traditional methods such as the cross-validation approach may be applied to select a proper bandwidth \( h \) based on available data. Carroll et al. (1997) noted that a sensible choice of the bandwidth \( h \) is generally difficult. Instead, they suggested an ad hoc bandwidth, given by \( \hat{h}_{\text{opt}} \times K^{-2/15} = O(K^{-1/3}) \), which satisfies \( Kh^4 \to 0 \) and \( Kh^2/(\log K)^2 \to \infty \).
### 3.3.3 Estimating the Response Model: Computational Details

In order to estimate the parameters $\beta, \xi, \alpha$ from our response model (3.6), we first need to follow “Step 1” of our algorithm. In this step, we estimate the single-index $\varphi(\cdot)$ by solving estimating equation (3.14). We can write,

$$
\tilde{\Phi}'(c)\tilde{\Delta}_i\tilde{K}_{ih}(c)\tilde{V}_i^{-1} = \begin{bmatrix}
K_h(C_{i1} - c) & \cdots & K_h(C_{iT} - c) \\
(C_{i1} - c)K_h(C_{i1} - c) & \cdots & (C_{i1} - c)K_h(C_{i1} - c)
\end{bmatrix},
$$

(3.18)

and

$$
\tilde{\Phi}'(c)\tilde{\Delta}_i\tilde{K}_{ih}(c)\tilde{V}_i^{-1} (Y_i - \tilde{p}_i) = \begin{bmatrix}
\sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - \tilde{p}_{it}) \\
\sum_{t=1}^{T} (C_{it} - c)K_h(C_{it} - c)(y_{it} - \tilde{p}_{it})
\end{bmatrix},
$$

(3.19)

where, $\tilde{p}_{it} = \frac{e^{x_i'\beta + \eta_0(c) + (C_{it} - c)\eta_1(c)}}{1 + e^{x_i'\beta + \eta_0(c) + (C_{it} - c)\eta_1(c)}}$, $K(\cdot)$ is the kernel density function (e.g. standard normal density function). We find,

$$
\frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_0(c)} = -\tilde{p}_{it}(1 - \tilde{p}_{it}),
$$

(3.20)

$$
\frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_1(c)} = -(C_{it} - c)\tilde{p}_{it}(1 - \tilde{p}_{it}),
$$

(3.21)
3.3. METHODS OF ESTIMATION

and

\[ \frac{\partial}{\partial \eta(c)} \{ \tilde{\Phi}_t(c) \tilde{\Delta}_t \tilde{K}_{ih}(c) \tilde{V}_t^{-1}(Y_t - \tilde{p}_t) \} \]

\[ = \left[ \sum_{t=1}^{T} K_h(C_{it} - c) \frac{\partial (y_{it} - \tilde{p}_{it})}{\partial \eta_0} \right] \left[ \sum_{t=1}^{T} (C_{it} - c) K_h(C_{it} - c) \frac{\partial (y_{it} - \tilde{p}_{it})}{\partial \eta_0} \right] \]

\[ = - \left[ \sum_{t=1}^{T} K_h(C_{i1} - c) \tilde{p}_{it}(1 - \tilde{p}_{it}) \sum_{t=1}^{T} (C_{it} - c) K_h(C_{it} - c) \tilde{p}_{it}(1 - \tilde{p}_{it}) \right] \]

\[ = - \left[ \begin{array}{ccc}
K_h(C_{i1} - c) & \cdots & K_h(C_{iT} - c) \\
(C_{i1} - c) K_h(C_{i1} - c) & \cdots & (C_{iT} - c) K_h(C_{iT} - c) \\
\tilde{p}_{i1}(1 - \tilde{p}_{i1}) & (C_{i1} - c) \tilde{p}_{i1}(1 - \tilde{p}_{i1}) & \\
\vdots & \vdots & \\
\tilde{p}_{iT}(1 - \tilde{p}_{iT}) & (C_{iT} - c) \tilde{p}_{iT}(1 - \tilde{p}_{iT}) & \\
\end{array} \right] \]

\[ \times \]

\[ = - \tilde{\Phi}_t(c) \tilde{\Delta}_t \tilde{K}_{ih}(c) \tilde{V}_t^{-1} H, \]  

\[ (3.22) \]

where \( H = \frac{\partial \tilde{p}_i}{\partial \eta(c)} \).

The iterative procedure for calculating the estimators \( \{ \hat{\eta}_o(c), \hat{\eta}_1(c) \} \) for a given \( c \) in a selected grid, begins with some starting values \( \{ \hat{\eta}_o(c)(0), \hat{\eta}_1(c)(0) \} \) and produces updated values \( \{ \hat{\eta}_o(c)(s+1), \hat{\eta}_1(c)(s+1) \} \) from interim values \( \{ \hat{\eta}_o(c)(s), \hat{\eta}_1(c)(s) \} \)
by means of the iterative equations

$$
\eta_{s+1} = \eta_s + \left( \sum_{i=1}^{K} \tilde{\Phi}_i(c) \tilde{\Delta}_i \tilde{K}_{ih}(c) \tilde{V}_i^{-1} H \right)^{-1} \sum_{i=1}^{K} \tilde{\Phi}_i(c) \tilde{\Delta}_i \tilde{K}_{ih}(c) \tilde{V}_i^{-1} (Y_i - \hat{p}_i),
$$

(3.23)

for $s = 0, 1, 2, \ldots$, where the second term on the right side of the above equation is evaluated at the current estimates $\{\hat{\eta}_0(c)_{(s)}, \hat{\eta}_1(c)_{(s)}\}$.

Given the estimate $\hat{\varphi}(c; \hat{\beta}, \hat{\xi}, \hat{\alpha}) = \hat{\eta}_0(c)$ and $\hat{\eta}_1(c)$ for points $c$ in the selected grid we can find

$$
\hat{p}_{it} = \frac{e^{X_{it} \beta + \hat{\eta}_0(c)}}{1 + e^{X_{it} \beta + \hat{\eta}_0(c)}},
$$

(3.24)

and

$$
\frac{\partial \hat{p}_{it}}{\partial \xi_j} = u_{it,j} \hat{\eta}_1(c) \hat{p}_{it} (1 - \hat{p}_{it}).
$$

(3.25)

The iterative procedure for calculating the estimators of the parameters $\left( \hat{\beta}, \hat{\xi}, \hat{\alpha} \right)$ begins with some starting values $\left( \hat{\beta}_0, \hat{\xi}_0, \hat{\alpha}_0 \right)$ from “Step 0” of our algorithm and produces updated values $\left( \hat{\beta}_{s+1}, \hat{\xi}_{s+1}, \hat{\alpha}_{s+1} \right)$ from interim values $\left( \hat{\beta}_s, \hat{\xi}_s, \hat{\alpha}_s \right)$ by means of the iterative equations

$$
\left( \hat{\beta}'_{s+1}, \hat{\xi}'_{s+1} \right)' = \left( \hat{\beta}'_s, \hat{\xi}'_s \right)' + \left( \sum_{i=1}^{K} \hat{D}'_i \hat{V}_i^{-1} \hat{D}_i \right)^{-1} \sum_{i=1}^{K} \hat{D}'_i \hat{V}_i^{-1} \{Y_i - \hat{p}_i(\beta, \xi, \alpha)\},
$$

(3.26)

and

$$
\hat{\alpha}_{s+1} = \hat{\alpha}_s + \left( \sum_{i=1}^{K} \tilde{G}_i' \tilde{W}_i^{-1} \tilde{G}_i \right)^{-1} \sum_{i=1}^{K} \tilde{G}_i' \tilde{W}_i^{-1} \left( \tilde{Z}_i - \rho_i \right),
$$

(3.27)
for $s = 0, 1, 2, \ldots$, where the second term on the right side of each estimating equation is evaluated at the current estimates $\left(\hat{\beta}_s, \hat{\xi}_s, \hat{\alpha}_s\right)$.

3.4 Asymptotic Properties

3.4.1 Asymptotics for $\hat{\varphi}(.)$

In order to show the asymptotic distribution of the single-index, we need the following key assumptions:

i.) The density function of $u_{it}$ has a continuous second derivative on its support.

ii.) The density function of $u_{0it} \xi$ is positive and uniformly continuous for $\xi$ in a neighborhood of its true value.

iii.) $\varphi^{(2)}(c)$ is continuous on its support.

iv.) The random vector $x_{it}$ is assumed to have a bounded support with $E(x'_{it}x_{it}) > 0$.

v.) $K(.)$ is a symmetric probability density function with bounded support.
From Equation (3.14) we can write,

\[
\sum_{i=1}^{K} 1' \tilde{\Delta}_i \tilde{K}_h(c) \tilde{V}_i^{-1} (Y_i - \tilde{p}_i) = 0
\]

\[
\Rightarrow \sum_{i=1}^{K} 1' \tilde{\Delta}_i \tilde{K}_h(c) \tilde{V}_i^{-1} (Y_i - p_i) - \sum_{i=1}^{K} 1' \tilde{\Delta}_i \tilde{K}_h(c) \tilde{V}_i^{-1} (\tilde{p}_i - p_i) = 0
\]

\[
\Rightarrow \sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - p_{it}) = \sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)(\tilde{p}_{it} - p_{it})
\]

(3.28)

Let the second term of Equation (3.28) be \( \sum_{i=1}^{K} \Phi_i^*(\tilde{\eta}_0(c)) \). By Taylor series expansion, we get

\[
\sum_{i=1}^{K} \Phi_i^*(\tilde{\eta}_0(c)) \approx \sum_{i=1}^{K} \Phi_i^*(\eta_0(c)) + \sum_{i=1}^{K} \frac{\partial}{\partial \tilde{\eta}_0(c)} \Phi_i^*(\tilde{\eta}_0(c))|_{\tilde{\eta}_0(c) = \eta_0(c)}
\]

\[
= \sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - p_{it}) (1 - p_{it}) (\tilde{\eta}_0(c) - \eta_0(c)).
\]

(3.29)

Thus from Equation (3.28) and (3.29) we get

\[
\hat{\varphi}(c) - \varphi(c) = \tilde{\eta}_0 - \eta_0
\]

\[
\approx \sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)(y_{it} - p_{it})
\]

\[
\sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)p_{it}(1 - p_{it}).
\]

(3.30)

Let the marginal density of \( u_i^T \boldsymbol{\xi} \) be denoted by \( f(.) \). Applying the asymptotic properties of the kernel estimators and Taylor series expansion we find the asymptotic expansion (see Carroll et al. (1997), Grace et al. (2009))

\[
\hat{\varphi}(c) - \varphi(c) = \int \frac{a^2 K(a) da}{2} \varphi^{(2)}(c)h^2 + \frac{1}{K f(c)} \sum_{i=1}^{K} \sum_{t=1}^{T} K_h(C_{it} - c)\Omega_{it}
\]

\[
+ o_p\{h^2 + (K h)^{-1/2}\},
\]

(3.31)
where $\Omega_{it}$ is the first element of the vector $(Y_{it} - p_{it})\Omega^{-1}(c)(1, x'_{it})'$ and

$$
\Omega(c) = E \begin{bmatrix} p_{it}(1 - p_{it}) & x'_{it} \\ x_{it} & x_{it}x'_{it} \end{bmatrix} | C_{it} = c \right) .
$$

As a consequence, the asymptotic distribution of the nonparametric function $\hat{\varphi}(c)$ is given by (Carroll et al. (1997), Grace et al. (2009))

$$(Kh)^{1/2} \left\{ \hat{\varphi}(c) - \varphi(c) - \int a^2 K(a) dt \frac{\varphi^{(2)}(c)h^2}{2} \right\} \rightarrow_d N \left( 0, \frac{d(c)}{f(c)} \int K^2(a) da \right) \quad (3.33)$$

where $d(c)$ is the first diagonal element of the matrix $\Omega^{-1}(c)$.

### 3.4.2 Asymptotics for $(\hat{\beta}, \hat{\xi}, \hat{\alpha})$

Applying the first order Taylor series approximation and following the same steps as in Equation (2.27), under some regularity conditions,

$$K^{1/2} \left[ (\hat{\beta} - \beta)', (\hat{\xi} - \xi)', (\hat{\alpha} - \alpha)' \right]' \text{ can be approximated by }$$

$$
\begin{bmatrix}
-K^{-1} \partial \hat{U}_\beta(\beta, \xi, \alpha, \varphi(.)) / \partial(\beta, \xi)'
& -K^{-1} \partial \hat{U}_\beta(\beta, \xi, \alpha, \varphi(.)) / \partial(\beta, \xi)'

-K^{-1} \partial \hat{U}_\alpha(\beta, \xi, \alpha, \varphi(.)) / \partial(\beta, \xi)'
& -K^{-1} \partial \hat{U}_\alpha(\beta, \xi, \alpha, \varphi(.)) / \partial(\beta, \xi)'

\end{bmatrix}^{-1} \begin{bmatrix}
K^{-1/2} \hat{U}_\beta(\beta, \xi, \alpha, \varphi(.))

K^{-1/2} \hat{U}_\alpha(\beta, \xi, \alpha, \varphi(.))

\end{bmatrix}
\equiv \hat{J}^{-1} \hat{L}
\quad (3.34)
$$

We can write

$$Var \left( \hat{J}^{-1} \hat{L} \right) = \hat{J}^{-1} Var(\hat{L})(\hat{J}^{-1})'.$$  \quad (3.35)
3.4. ASYMPTOTIC PROPERTIES

Now,

\[ \text{Var}(\dot{L}) = K^{-1} \begin{bmatrix} \text{Var}(\dot{\bar{\beta}}_{\xi}) & \text{Cov}(\dot{\bar{\beta}}_{\xi}, \dot{\bar{\alpha}}) \\ \text{Cov}(\dot{\bar{\beta}}_{\xi}, \dot{\bar{\alpha}}) & \text{Var}(\dot{\bar{\alpha}}) \end{bmatrix} \]

\[ = K^{-1} \left( \sum_{i=1}^{K} \dot{D}_i \dot{V}_i^{-1} \text{Var}(Y_i) \dot{V}_i^{-1} \dot{D}_i \quad \sum_{i=1}^{K} \dot{D}_i \dot{V}_i^{-1} \text{Cov}(Y_i, \dot{Z}_i) \dot{W}_i^{-1} G_i \right) \]

\[ = K^{-1} \begin{pmatrix} \dot{A}_{11} & \dot{A}_{12} \\ \dot{A}_{21} & \dot{A}_{22} \end{pmatrix}, \quad (3.36) \]

where the variance and covariance terms may be approximated by

\[ \text{Var}(Y_i) \approx (Y_i - \hat{p}_i)(Y_i - \hat{p}_i)' \]

\[ \text{Cov}(Y_i, \dot{Z}_i) \approx (Y_i - \hat{p}_i)(\dot{Z}_i - \rho_i)' \]

\[ \text{Var}(\dot{Z}_i) \approx (\dot{Z}_i - \rho_i)(\dot{Z}_i - \rho_i)' \quad (3.37) \]

Moreover, if \( \hat{\phi}(.) \) is consistent and fixed then the linear functions \( L \) generally have an asymptotic normal distribution with the mean vector zero and variance matrix

\[ \Sigma_L = \lim_{K \to \infty} K^{-1} \begin{pmatrix} \dot{A}_{11} & \dot{A}_{12} \\ \dot{A}_{21} & \dot{A}_{22} \end{pmatrix}. \quad (3.38) \]

Thus the joint asymptotic distribution of \( K^{1/2} \left[ (\hat{\beta} - \beta)', (\hat{\xi} - \xi), (\hat{\alpha} - \alpha) \right] = J^{-1} \dot{L} \) is Gaussian with the mean vector zero and variance matrix

\[ \Sigma = \lim_{K \to \infty} J^{-1} \Sigma_L J^{-1}' \quad \lim_{K \to \infty} (J^{-1})'. \quad (3.39) \]
Also as $K \to \infty$,

$$-K^{-1} \frac{\partial \tilde{U}_{\beta, \xi}(\beta, \xi, \alpha, \varphi(\cdot))}{\partial (\beta, \xi)'} = -K^{-1} \sum_{i=1}^{K} \hat{D}' \hat{V}_{i}^{-1} \frac{\partial (Y_i - \hat{p}_i)}{\partial (\beta, \xi)'}$$

$$- K^{-1} \sum_{i=1}^{K} \frac{\partial \hat{D}' \hat{V}_{i}^{-1}}{\partial (\beta, \xi)'} (Y_i - \hat{p}_i)$$

$$= K^{-1} \sum_{i=1}^{K} \hat{D}' \hat{V}_{i}^{-1} \hat{D}_i + o_p(1), \quad (3.40)$$

since $\frac{\partial (Y_i - \hat{p}_i)}{\partial (\beta, \xi)'} = (-1) \hat{D}_i$, and $\frac{\partial \hat{D}' \hat{V}_{i}^{-1}}{\partial (\beta, \xi)'}$ are fixed matrices that do not involve $Y_i$.

Similarly, as $\frac{\partial \hat{V}_{i}^{-1}}{\partial \alpha'}$, $\frac{\partial \hat{W}_{i}^{-1}}{\partial \alpha'}$ and $\frac{\partial \hat{E}_i' \hat{W}_{i}^{-1}}{\partial (\beta, \xi)'}$ are all free of $Y_i$, we can write as $K \to \infty$,

$$-K^{-1} \frac{\partial \tilde{U}_{\alpha}(\beta, \xi, \alpha, \varphi(\cdot))}{\partial \alpha'} = -K^{-1} \sum_{i=1}^{K} \hat{D}' \hat{W}_{i}^{-1} (Y_i - \hat{p}_i) = o_p(1), \quad (3.41)$$

$$- K^{-1} \sum_{i=1}^{K} \hat{D}' \hat{W}_{i}^{-1} \hat{Z}_i \hat{W}_{i}^{-1} (Y_i - \hat{p}_i)$$

$$= K^{-1} \sum_{i=1}^{K} \hat{D}' \hat{W}_{i}^{-1} \hat{D}_i + o_p(1), \quad (3.42)$$

and

$$-K^{-1} \frac{\partial \tilde{U}_{\alpha}(\beta, \xi, \alpha, \varphi(\cdot))}{\partial (\beta, \xi)'} = -K^{-1} \sum_{i=1}^{K} \hat{G}_i' \hat{W}_{i}^{-1} \frac{\partial (\hat{Z}_i - \rho_i)}{\partial (\beta, \xi)'}$$

$$- K^{-1} \sum_{i=1}^{K} \frac{\partial \hat{G}_i' \hat{W}_{i}^{-1}}{\partial (\beta, \xi)'} (\hat{Z}_i - \rho_i)$$

$$= - K^{-1} \sum_{i=1}^{K} \hat{G}_i' \hat{W}_{i}^{-1} \frac{\partial \hat{Z}_i}{\partial (\beta, \xi)'} + o_p(1). \quad (3.43)$$
\[ \frac{\partial \hat{Z}_{itu}}{\partial \beta_j} = -\left\{ \frac{\partial \hat{\rho}_{it}}{\partial \beta_j} (Y_{iu} - \hat{\rho}_{iu}) + \frac{\partial \hat{\rho}_{iu}}{\partial \beta_j} (Y_{it} - \hat{\rho}_{it}) + \frac{1}{2} (Y_{it} - \hat{\rho}_{it})(Y_{iu} - \hat{\rho}_{iu}) \right\} \times \left\{ (1 - 2\hat{\rho}_{it})\hat{q}_{it}^{-1} \frac{\partial \hat{\rho}_{it}}{\partial \beta_j} + (1 - 2\hat{\rho}_{iu})\hat{q}_{iu}^{-1} \frac{\partial \hat{\rho}_{iu}}{\partial \beta_j} \right\} (\hat{\rho}_{it}\hat{q}_{it}\hat{\rho}_{iu}\hat{q}_{iu})^{-1/2}. \]

(3.44)

To find \( \frac{\partial \hat{Z}_{itu}}{\partial \xi_j} \) we can replace \( \frac{\partial \hat{\rho}_{it}}{\partial \beta_j} \) and \( \frac{\partial \hat{\rho}_{iu}}{\partial \beta_j} \) by \( \frac{\partial \hat{\rho}_{it}}{\partial \xi_j} \) and \( \frac{\partial \hat{\rho}_{iu}}{\partial \xi_j} \) respectively in the above equation (3.44).

Hence, as \( K \to \infty \), we can write

\[ \dot{J} = K^{-1} \begin{bmatrix} \sum_{i=1}^{K} \hat{D}_i'\hat{V}_i^{-1}\hat{D}_i & 0 \\ -\sum_{i=1}^{K} G_i'\hat{W}_i^{-1} \frac{\partial \hat{Z}_i}{\partial (\beta', \xi')} & \sum_{i=1}^{K} G_i'\hat{W}_i^{-1}G_i \end{bmatrix} \]

(3.45)

and

\[ \dot{J}^{-1} = K \begin{bmatrix} \left( \sum_{i=1}^{K} \hat{D}_i'\hat{V}_i^{-1}\hat{D}_i \right)^{-1} & 0 \\ \hat{B}^* & \left( \sum_{i=1}^{K} G_i'\hat{W}_i^{-1}G_i \right)^{-1} \end{bmatrix} = \dot{J}^*, \]

(3.46)

where

\[ \hat{B}^* = \left( \sum_{i=1}^{K} G_i'\hat{W}_i^{-1}G_i \right)^{-1} \left( \sum_{i=1}^{K} G_i'\hat{W}_i^{-1} \frac{\partial \hat{Z}_i}{\partial (\beta', \xi')} \right) \left( \sum_{i=1}^{K} \hat{D}_i'\hat{V}_i^{-1}\hat{D}_i \right)^{-1}. \]

(3.47)

Finally, the asymptotic variance matrix of \( K^{1/2} [ (\hat{\beta} - \beta)', (\hat{\xi} - \xi), (\hat{\alpha} - \alpha) ] \) can be expressed as

\[ \Sigma = \dot{J}^* \Sigma_L \dot{J}^*. \]

(3.48)

where \( \dot{J}^* \) and \( \Sigma_L \) are defined in (3.46) and (3.38) respectively.
3.5 Simulation Study

To study and compare the empirical properties of the usual Prentice’s (1988) second order GEE (Method 1) with our proposed method (Method 2), we ran two sets of simulations. In the first set, the estimators were studied when the true model is a linear single-index model. In the second set, we considered the true model as a partially linear single-index model.

3.5.1 Models for Simulation Study

For the simulation study, we consider a two-group design configuration with a binary response measured on four occasions. The marginal model for the mean response $E(Y_{it}|x_{it}, u_{it}, \beta, \xi, \alpha)$ is given by

$$\text{logit} \{E(Y_{it}|x_{it}, u_{it}, \beta, \xi, \alpha)\} = \beta x_i + \varphi(\xi_1 u_{1it} + \xi_2 u_{2it} + \xi_3 u_{3it}), \quad t = 1, \ldots, 4,$$  

where $x_i$ is a dichotomous covariate indicating the group membership for the $i$th individual ($i = 1, \ldots, K$) observed over a fixed set of $T = 4$ time-points, $t = 1, 2, 3, 4$. Throughout the simulations, we consider $P(X_i = 1) = 0.5$.

We consider two different models to generate the data. In the first model, covariates $u_{ij}$’s were generated from the Uniform distribution $U[0, 1]$ and we fixed $\xi = (\xi_1, \xi_2, \xi_3)' = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})'$. In the second model, we chose covariates $u_i$’s as time point indicators, where $u_{jit} = 1$ if $t = j$ ($j = 1, 2, 3$) or 0 otherwise. We fixed $\xi = (\xi_1, \xi_2, \xi_3)' = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})'$ for this data set.
3.5. SIMULATION STUDY

To assess the performance of the proposed method, we compare it to the ordinary second order GEE approach of Prentice (1988) under two situations of data structures. First, we consider a scenario where a standard GEE model well fits the data with a linear single-index model. As a linear function is a special case of a nonlinear function, we expect that the proposed estimators would still be consistent, but there may be a possible efficiency loss incurred. In order to generate the data, \( \varphi(.) \) is specified as the identity function. Second, we consider a scenario when our proposed model well fits the data with a partially linear single-index model where we chose \( \varphi(a) = \sin(\pi(1 - a)) \).

Throughout the simulations, data were generated under exchangeable correlation structures among the responses and we chose \( \text{corr}(Y_{it}, Y_{i't'}) = \alpha = 0.3 \) for all \( t \neq t' = 1, \ldots, T \). Each simulation run was based on 1000 replications of data sets, with each data set containing \( K = 100 \) subjects and \( T = 4 \) observations per subject. We take the standard normal density as the kernel. The data-driven bandwidth \( h \) is used as discussed in Section 3.3.

3.5.2 Diagnostic methods

We compare three methods based on empirical biases and mean squared errors (MSE). The bias of an estimator \( \hat{\theta} \) of \( \theta \) is estimated by

\[
\text{bias}(\hat{\theta}) \approx \frac{1}{S} \sum_{s=1}^{S} (\hat{\theta}_s - \theta),
\]  

(3.50)
where \( \hat{\theta}_s \) is the estimate of \( \theta \) obtained from the \( s \)th simulated data set and \( S \) is the simulation size. The mean squared error (MSE) of \( \hat{\theta} \) is estimated by

\[
\text{MSE}(\hat{\theta}) \approx \frac{\sum_{s=1}^{S} (\hat{\theta}_s - \theta)^2}{S}.
\]

(3.51)

3.5.3 Results

Table 3.1 presents the empirical biases and mean squared errors (MSEs) of the estimators of the regression parameters \((\beta, \xi_1, \xi_2, \xi_3)\) and the correlation parameter \(\alpha\) for continuous covariates \(u_i\)'s under both methods (Method 1 and Method 2). Table 3.2 repeats these results for discrete covariates \(u_i\)'s.

Table 3.1: Comparison of the usual Prentice’s second order GEE (Method 1) with the proposed method (Method 2) when \(u_i\)'s are continuous. [True parameter values: \(\beta_1 = 1, \xi = (1/\sqrt{3}, -1/\sqrt{3}, 1/\sqrt{3})'\) and \(\alpha = 0.3\).]

<table>
<thead>
<tr>
<th>True Model</th>
<th>Parameters</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>(\beta_1)</td>
<td>-0.0056</td>
<td>0.0731</td>
</tr>
<tr>
<td></td>
<td>(\xi_1)</td>
<td>0.0125</td>
<td>0.1081</td>
</tr>
<tr>
<td></td>
<td>(\xi_2)</td>
<td>-0.0017</td>
<td>0.0921</td>
</tr>
<tr>
<td></td>
<td>(\xi_3)</td>
<td>0.0058</td>
<td>0.0924</td>
</tr>
<tr>
<td></td>
<td>(\alpha)</td>
<td>-0.0020</td>
<td>0.0027</td>
</tr>
<tr>
<td>Partially linear</td>
<td>(\beta_1)</td>
<td>-0.0273</td>
<td>0.0592</td>
</tr>
<tr>
<td></td>
<td>(\xi_1)</td>
<td>-0.2678</td>
<td>0.1591</td>
</tr>
<tr>
<td></td>
<td>(\xi_2)</td>
<td>0.0820</td>
<td>0.1147</td>
</tr>
<tr>
<td></td>
<td>(\xi_3)</td>
<td>-0.2590</td>
<td>0.1639</td>
</tr>
<tr>
<td></td>
<td>(\alpha)</td>
<td>-0.0056</td>
<td>0.0029</td>
</tr>
</tbody>
</table>
3.5. SIMULATION STUDY

Table 3.2: Comparison of the usual Prentice's second order GEE (Method 1) with the proposed method (Method 2) when $u_i$'s are discrete. [True parameter values: $\beta_1 = 1$, $\xi = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})'$ and $\alpha = 0.3$.]

<table>
<thead>
<tr>
<th>Fitted Model</th>
<th>Parameters</th>
<th>Method 1 Bias</th>
<th>Method 1 MSE</th>
<th>Method 2 Bias</th>
<th>Method 2 MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>$\beta_1$</td>
<td>-0.0525</td>
<td>0.0360</td>
<td>-0.0543</td>
<td>0.0432</td>
</tr>
<tr>
<td></td>
<td>$\xi_1$</td>
<td>0.0219</td>
<td>0.0396</td>
<td>0.0436</td>
<td>0.0504</td>
</tr>
<tr>
<td></td>
<td>$\xi_2$</td>
<td>0.0119</td>
<td>0.0534</td>
<td>-0.0382</td>
<td>0.0340</td>
</tr>
<tr>
<td></td>
<td>$\xi_3$</td>
<td>0.0044</td>
<td>0.0438</td>
<td>-0.0326</td>
<td>0.0285</td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
<td>-0.0027</td>
<td>0.0027</td>
<td>-0.0033</td>
<td>0.0026</td>
</tr>
<tr>
<td>Partially linear</td>
<td>$\beta_1$</td>
<td>-0.0321</td>
<td>0.0816</td>
<td>0.0107</td>
<td>0.0898</td>
</tr>
<tr>
<td></td>
<td>$\xi_1$</td>
<td>0.0076</td>
<td>0.0535</td>
<td>-0.0020</td>
<td>0.0499</td>
</tr>
<tr>
<td></td>
<td>$\xi_2$</td>
<td>-0.0175</td>
<td>0.0578</td>
<td>-0.0415</td>
<td>0.0389</td>
</tr>
<tr>
<td></td>
<td>$\xi_3$</td>
<td>-0.0435</td>
<td>0.0622</td>
<td>-0.0708</td>
<td>0.0432</td>
</tr>
<tr>
<td></td>
<td>$\alpha$</td>
<td>-0.0099</td>
<td>0.0034</td>
<td>-0.0026</td>
<td>0.0033</td>
</tr>
</tbody>
</table>

Both Table 3.1 and 3.2 show very similar results from the two methods when the true model is single-index linear. Although the finite sample biases from Method 1 tend to be smaller than those obtained from Method 2, the biases from Method 2 are still reasonably small. Also we notice that Method 2 tends to produce larger MSE as compared to Method 1, but the differences do not seem considerable. In summary, if a nonparametric function is included to the mean response, where the true model is actually characterized by an ordinary single-index linear model, the proposed method would still provide consistent estimates, though some efficiency loss may incur.

On the other hand, if the true underlying model is partially linear but we adopt a standard second order GEE (in this case Prentice’s (1988) approach) model, then
the resulting estimators could be biased. The biases for \( \xi \) estimates are apparent in Table 3.1 when covariates \( u_{ij} \)'s are generated from a continuous distribution. We also observe that Method 2 provides lower MSE for the estimators of the linear coefficient \( \beta \), but the difference is not profound. However, lower MSE from Method 2 is more apparent for the estimators of \( \xi \) and \( \alpha \), which is not surprising.

In Table 3.2, the two methods appear to provide similar results when the true underlying model is partially linear. This is perhaps due to the fact that the non-linear covariates are all discrete binary variables. As a consequence, the estimated single-index \( \hat{\varphi}(u'\xi) \) is not a smooth function. Hence our proposed method works simply as an ordinary GEE method for this type of data set.

## 3.6 Applications

### 3.6.1 Analysis of ICHS Data

Alfred Sommer and colleagues conducted a study (which we will refer to as the Indonesian Children’s Health Study or ICHS) in the Aceh province of Indonesia to determine the causes and effects of vitamin A deficiency in pre-school children (Sommer, 1982). We present an analysis of infectious disease data on 250 Indonesian children, a subset of the cohort studied by Somer, Katz, and Tarwojto (1983). The preschool children were examined up to six consecutive quarters for the presence of respiratory infection. There were 1,200 observations in total. We consider
complete data for the first four visits from 548 pre-school children with or without the respiratory infection.

We focus on the question of whether vitamin A deficient children are at increased risk of respiratory infection, which is one of the leading causes of morbidity and mortality in children from the developing world. Such a relationship is plausible because vitamin A is required for the integrity of epithelial cells, the first line of defence against infection in the respiratory tract. Here the goal is to draw inferences on the change in respiratory infection status in the presence of intraperson correlation based on our proposed method. The model parameters were estimated and compared using the second order GEE (Method 1) approach of Prentice (1988) and our proposed semiparametric approach (Method 2).

We set the binary response variable \( Y_{it} = 1 \) if the \( i \)th child suffers from reparatory infection at the \( t \)th visit, and 0 otherwise, for \( i = 1, \ldots, 137 \) and \( t = 1, \ldots, 4 \). The covariates of interest include “Xerop” which represents presence/absence (1 or 0) of xerophthalmia, an ocular manifestation of chronic vitamin A deficiency; “Time” represents time \( t \); “Age” represents the baseline age in months (centered at 36); height for age is represented by “Height”, as a percent of the National Center for Health Statistics (NCHS) standard(centered at 90%), which indicates long-term nutritional status.

The marginal mean response \( p_{it} = E(Y_{it}) \) is defined as a function of the covari-
ates in the form

\[
\logit(p_{it}) = \beta_0 + \beta_1 (\text{Xerop})_i + \varphi(\xi_1 \text{Time}_{it} + \xi_2 \text{Age}_i + \xi_3 \text{Height}_i)
\] (3.52)

where \(\varphi(.)\) is an unknown function. Exchangeable association structure is considered here. We take the standard normal density as the kernel. The data-driven bandwidth \(h\) is used as discussed in Section 3.3.

![Figure 3.1: Estimated nonlinear curves for ICHS study.](image)
Table 3.3: Estimates and Standard Errors of Regression and Association Parameters for the ICHS Study.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Method 1</th>
<th></th>
<th>Method 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>S.E</td>
<td>z-value</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.734</td>
<td>0.761</td>
<td>-0.96</td>
<td>-0.249</td>
</tr>
<tr>
<td>Xerophthalmia</td>
<td>1.448</td>
<td>0.634</td>
<td>2.29</td>
<td>1.451</td>
</tr>
<tr>
<td>Time</td>
<td>-1.452</td>
<td>0.641</td>
<td>-2.27</td>
<td>-0.525</td>
</tr>
<tr>
<td>Age</td>
<td>-2.114</td>
<td>0.722</td>
<td>-2.93</td>
<td>-0.735</td>
</tr>
<tr>
<td>Height for age</td>
<td>-1.133</td>
<td>0.788</td>
<td>-1.44</td>
<td>-0.429</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.049</td>
<td>0.034</td>
<td>1.43</td>
<td>0.048</td>
</tr>
</tbody>
</table>

Figure 3.1 displays the estimates of the single index $\varphi(u'\xi)$ against $u'\xi$. Here the estimated curve $\hat{\varphi}(.)$ hardly shows any evidence of a nonlinear trend. This suggests that the data might be well fitted by an usual GEE model. Our finding agrees with that of Zeger and Karim (1991). In their study, they noted that there is no evidence that the height-for-age relationship deviates from the logistic-linear model for the most undernourished children. However, the inclusion of a nonlinear function $\varphi(.)$ allows model (3.52) to be more flexible to capture curvature, although the interpretation of the nonparametric covariate effects differs from that in an ordinary GEE model. In principle, nonzero components of $\xi$ suggest a significant predictor of the response, as commented in Carroll et al. (1997).

Table 3.3 reports the estimates of the model parameters, their standard errors and the corresponding $z$-values. We observe that the estimates of the linear covariate effects from the two methods are generally close to each other. The xerophthalmia coefficient is 1.448, indicating that the odds of suffering from respi-
ratory infections is $\exp(1.448) = 4.25$ times higher in vitamin A deficient children. The odds of suffering from respiratory infections appear to decrease with increased time, age and height for age. The presence of intraperson correlation appears to be very small and insignificant under both methods.

### 3.6.2 Analysis of Smoking Data

We also present an analysis of data on cigarette smoking trends from the Coronary Artery Development in Young Adults (CARDIA) study, an epidemiological study that recorded cardiovascular risk factors on five occasions over a 10-year period in black and white males and females (Hughes et al., 1987). This study was conducted in four urban centres (Birmingham, AL; Chicago, IL, Minneapolis, MN; and Oakland, CA) across the United States in which a total of 5,115 young adults aged 18-30 years were followed prospectively and examined up to five times from 1986 to 1996. Recruitment, restricted to blacks and whites, was carried out to achieve approximate balance in sample size with respect to age, race, gender, and education. Study participants were scheduled for visits at years 0, 2, 5, 7, and 10. We consider complete data for the first four visits from 3,693 young adults with self reported smoking status (yes/no).

Here the goal is to draw inferences on the change in smoking prevalence of young adults in the presence of intraperson correlation based on our proposed method. The model parameters were estimated and compared using the second
order GEE (Method 1) approach of Prentice (1988) and the proposed semiparametric approach (Method 2).

Let the binary response variable \( Y_{it} = 1 \) if the \( i \)th individual is a smoker at the \( t \)th visit, and 0 if he/she is a nonsmoker. The marginal mean response \( p_{it} = E(Y_{it}) \) is defined as a function of the covariates in the form

\[
\text{logit}(p_{it}) = \beta_0 + \beta_1(\text{age}/10)_i + \beta_2x_t + \beta_3\text{eduh}_i + \beta_4\text{educ}_i \\
+ \varphi(\xi_1\text{racebf}_i + \xi_2\text{racewm}_i + \xi_3\text{racewf}_i),
\]

for \( i = 1, \ldots, 3693 \) and \( t = 1, \ldots, 4 \), where \( \text{age}_i = \text{age of individual } i \) in years at baseline time; \( x_t = \text{year since the baseline measurement} = 0,2,5,7 \); the binary indicators \( \text{eduh}_i = 1 \) if \( i \)th individual's education level is high school or less, and 0 otherwise; \( \text{educ}_i = 1 \) if education level is up to some college, and 0 otherwise; \( \text{racebf}_i = 1 \) if the person is a black female, and 0 otherwise; \( \text{racewm}_i = 1 \) if the person is a white male, and 0 otherwise; and \( \text{racewf}_i = 1 \) if the person is a white female, and 0 otherwise. Exchangeable association structure is modeled here. We take the standard normal density as the kernel. The data-driven bandwidth \( h \) is used as discussed in Section 3.3.

Figure 3.2 displays the estimates of the single index \( \varphi(u'\xi) \) against \( u'\xi \). Here the estimated curve \( \hat{\varphi}(u'\xi) \) does not show much evidence of a nonlinear trend, though a small curvature is visible. This suggests that the data might be well fitted by an usual GEE model.
3.6. APPLICATIONS

Figure 3.2: Estimated nonlinear curves for the smoking data from the CARDIA Study

Table 3.4: Estimates and Standard Errors of Regression and Association Parameters for the CARDIA Study.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Method 1</th>
<th>Method 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>S.E</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.937</td>
<td>0.716</td>
</tr>
<tr>
<td>Age/10</td>
<td>0.499</td>
<td>0.279</td>
</tr>
<tr>
<td>Year Followup</td>
<td>-0.011</td>
<td>0.012</td>
</tr>
<tr>
<td>Education (High School or less)</td>
<td>1.346</td>
<td>0.285</td>
</tr>
<tr>
<td>Education (Some college)</td>
<td>1.055</td>
<td>0.240</td>
</tr>
<tr>
<td>Race (Black Female)</td>
<td>-0.766</td>
<td>0.277</td>
</tr>
<tr>
<td>Race (White Male)</td>
<td>0.046</td>
<td>0.281</td>
</tr>
<tr>
<td>Race (White Female)</td>
<td>-0.332</td>
<td>0.308</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.741</td>
<td>0.030</td>
</tr>
</tbody>
</table>


3.7. CONCLUSIONS

Table 3.4 reports the estimates of the model parameters, their standard error and the corresponding \( z \)-values. We observe that the estimates for all the covariates from both methods are generally close to each other. The primary reason for similar parameter estimates for the covariates is that the non-linear covariates that are included in this model are all binary variables. As a result, our estimated single-index \( \hat{\phi}(u'\xi) \) is a step function as we can see from Figure 3.2. Hence our proposed method performed as an ordinary GEE method for this data set.

The results from Table 3.4 suggest that the level of education has strong influence on the smoking status of the subjects. For example, the young adults are estimated to have \( \exp(1.346) = 3.85 \) times higher odds to be a smoker if their level of education is up to high school or less than those who have a college degree or more. The presence of intraperson correlation appears to be high and significant under both methods.

3.7 Conclusions

The purpose of this research was to propose and explore a semiparametric approach to analyzing longitudinal binary data. Here the interest lies in the estimation of the association coefficients simultaneously with the marginal mean parameters. In some previous studies, for example, Grace et al. (2009), the independence working matrix is employed in the estimation algorithm when conducting estimation of
the mean parameters. In our proposed method, we incorporate the true correlation structure in our estimation algorithm.

Our simulation study demonstrates that if the true underlying model is partially linear, then our proposed method generally provides unbiased and efficient estimators. Even if a nonparametric function is included to fit data whose marginal mean model is actually characterized by an ordinary single-index linear model, the proposed method still leads to consistent estimates, though some efficiency loss may incur. Therefore, the method we describe here has applications in a wide variety of settings. It can also be generalized for accommodating data with more complex association structures.
Chapter 4

Weighted Semiparametric

Generalized Estimating Equations

4.1 Introduction

In this Chapter, we extend the proposed semiparametric method to the analysis of incomplete longitudinal data with dropouts. This research was motivated by the recent Genetic and Inflammatory Marker of Sepsis (GenIMS) study (Milbrandt et al., 2009), which was a longitudinal cohort study of patients with community acquired pneumonia who were recruited from 28 US academic and community hospitals. A goal of the GenIMS study was to understand the longitudinal response of coagulation activation in the progression of pneumonia to severe sepsis in patients. In this study, one of the important coagulation biomarkers for understanding the
mechanisms of pneumonia progression to sepsis is antithrombin. Longitudinal measurements on this antithrombin biomarker were obtained from the patients daily for the first seven days of hospitalization. But some patients responses were missing at the follow-up times due to the early discharge from the hospital or death during the first seven days. We assume that this drop-out process is informative, as the previous antithrombin biomarker level is thought to be a predictor of its missingness indicator. Also, an initial analysis of the data showed trends of non-linear relationship between response variable and two other covariates age and time. We adopt our proposed semiparametric approach to analyzing these data. This chapter presents theoretical developments of the proposed method. The results from the data analysis are presented in the next chapter.

To analyze incomplete data, we incorporate the inverse probability-weights of Robins et al. (1995) into our semiparametric method as proposed in chapter 3. Here we jointly estimate the marginal mean parameters and the association parameters by the second order generalized estimating equations in the context of incomplete longitudinal binary data with monotone dropouts.

This chapter is organized as follows. Section 4.2 introduces the model and notation to define partially linear single-index models for incomplete binary longitudinal data. Section 4.3 discusses weighted GEE estimation of partially linear single-index models and Section 4.4 gives the conclusions of the chapter.
4.2 Model and Notation

4.2.1 Response Model

Suppose $K$ subjects are observed at a fixed set of $T$ time points. Let $Y_{it}$ represent a binary response variable from subject $i$, $i = 1,\ldots,K$, at visit $t$, $t = 1,\ldots,T$. For the $i$th subject, we can form a $T \times 1$ vector, $Y_i = (Y_{i1},\ldots,Y_{iT})'$, of binary response variables. Let the lower case letters $y_{it}$ and $y_i$ denote the realizations of $Y_{it}$ and $Y_i$, respectively. Also, let $x_{it} = (x_{it,1},\ldots,x_{it,p})'$ be a $p \times 1$ and $u_{it} = (u_{it,1},\ldots,u_{it,q})'$ be another $q \times 1$ vectors of covariates from subject $i$ at time $t$. The covariates may be time-dependent or fixed across the entire observation times. Let $x_i = (x_{i1}',\ldots,x_{iT}')'$ and $u_i = (u_{i1}',\ldots,u_{iT}')'$.

Assume that the marginal distribution of $Y_{it}$ is Bernoulli:

$$Y_{it} \sim \text{Bernoulli}(p_{it}), \ i = 1,\ldots,K; \ t = 1,\ldots,T,$$  

(4.1)

with the probability of success,

$$p_{it} = E(Y_{it}|x_i,u_i) = E(Y_{it}|x_{it},u_{it}).$$  

(4.2)

Let $p_i = (p_{i1},\ldots,p_{iT})'$. Assuming that the mean of response variable $Y_{it}$ depends only on the covariate vector for subject $i$ at time $t$, i.e., $p_{it} = E(Y_{it}|x_i,u_i) = E(Y_{it}|x_{it},u_{it})$ (Pepe and Anderson, 1994), we consider modelling the mean response by the logistic regression:

$$\text{logit}(p_{it}) = \log \left( \frac{p_{it}}{1 - p_{it}} \right) = x_{it}'\beta + \varphi(u_{it}'\xi) \quad \text{with} \quad ||\xi|| = 1,$$  

(4.3)
where $\beta$ and $\xi$ are $p \times 1$ and $q \times 1$ unknown parameter vectors, respectively, and $\varphi(.)$ is an unknown smooth function. The restriction $||\xi|| = 1$ ensures identifiability of $\xi$. The marginal variance of the response variable $Y_{it}$ is specified as a function of the marginal mean as

$$v_{it} = \text{var}(Y_{it}|x_{it}, u_{it}) = p_{it}(1 - p_{it}).$$

(4.4)

We assume that $Y_{it}$ and $Y_{i't'}$ are uncorrelated when $i \neq i'$. Let

$$\text{corr}(Y_{it}, Y_{i't'}) = \alpha_{u't'}$$

(4.5)

represent the correlation between $Y_{it}$ and $Y_{i't'}$ for given $x_{it}$ and $u_{it}$, where

$\alpha = (\alpha_{12}, \ldots, \alpha_{1T}, \alpha_{23}, \ldots, \alpha_{T-1,T})'$ is the vector of correlation parameters. If $\varphi(.)$ is specified as the identity function, then model (4.3) becomes an ordinary logistic regression equation.

### 4.2.2 Dropout Model

To introduce a dropout model, let $R_i = (R_{i1}, \ldots, R_{iT})'$ denote the response indicators for the vector $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, i.e., $R_{it}$ be the indicator variable taking the value 1 if the response $Y_{it}$ is observed and 0 otherwise. We assume a monotone missing data pattern, where $R_{i1} \geq \ldots \geq R_{iT}$ and $R_{i1} = 1$ for all subjects.

In general, the dropout mechanism can depend on the full vector of responses $Y_i$ (including the unobserved components of $Y_i$) and the matrices of covariates $x_i$
and $u_i$. Let
\[\lambda_{it} = P(R_{it} = 1|R_{i1} = \ldots = R_{i,t-1} = 1, y_i, x_i, u_i, \tau) \]  
(4.6)
be the probability that the $i$th subject is observed at time $t$, given that the subject is observed at previous $t-1$ time points and given the response vector $y_i$ and covariate matrix $x_i$. Here the components of $\tau$ are referred to as the “nuisance parameters” of the missing data model. We denote the observed components of the response vector $Y_i$ by the vector $Y_i^o$ and unobserved components by the vector $Y_i^u$. In (4.6), as missingness depends on the unobserved values of the response variables, it is referred to as nonignorable (NI) missingness. Data are called missing at random (MAR) if
\[\lambda_{it} = P(R_{it} = 1|R_{i1} = \ldots = R_{i,t-1} = 1, x_i, u_i, \tau). \]  
(4.7)
Also, data are called missing completely at random (MCAR) if $\lambda_{it} = P(R_{it} = 1|R_{i1} = \ldots = R_{i,t-1} = 1, x_i, u_i, \tau)$.

Note that in the case of dropouts, the random vector $R_i = (R_{i1}, \ldots, R_{iT})'$ of binary indicators can be characterized by a single random variable
\[M_i = 1 + \sum_{t=1}^T R_{it}, \]  
(4.8)
which indicates the time of dropout. In this case, the missing data or dropout process can be defined by

\[\nu_{im_i} = f_{M_i}(m_i|x_i, \tau) = P(M_i = m_i|x_i, u_i, \tau). \]  
(4.9)
If we assume that all subjects are observed on the first occasion, then $M_i$ takes on values between 2 and $T + 1$, where the maximum value $(T + 1)$ corresponds to a complete measurement sequence. It can be shown that

$$P(M_i = m | y_i, x_i, \tau) = P(R_{i2} = \ldots = R_{im-1} = 1, R_{im} = 0 | y_{i1}, \ldots, y_{im}, x_i, u_i, \tau)$$

$$= \left\{ \prod_{t=2}^{m-1} P(R_{it} = 1 | R_{i1} = \ldots = R_{it-1} = 1, y_{i1}, \ldots, y_{it}, x_i, u_i, \tau) \right\} \times \left\{ P(R_{im} = 0 | R_{i1} = \ldots = R_{im-1} = 1, y_{i1}, \ldots, y_{im}, x_i, u_i, \tau) \right\}^{I \{m \leq T\}}$$

(4.10)

where $I\{\}$ denotes an indicator variable.

### 4.3 Methods of Estimation

#### 4.3.1 WGEE for Partially Linear Single-index Model

For a partially linear single index model, the estimating functions $U_{\beta, \xi}(\beta, \xi, \alpha, \varphi(\cdot))$ and $U_{\alpha}(\beta, \xi, \alpha, \varphi(\cdot))$, as defined in Chapter 3, involve an unknown smooth function $\varphi(\cdot)$. We need to estimate this function locally in order to estimate $\beta$, $\xi$ and $\alpha$. Assuming $\varphi(c)$ has the second derivative, we may approximate $\varphi(c)$ by a locally linear function within the neighborhood of $c_0$ via the Taylor series expansion $\varphi(c) \approx \varphi(c_0) + \varphi^{(1)}(c_0)(c - c_0)$ for a given point $c_0$, where $\varphi^{(1)}(c_0)$ is the first order differentiation of $\varphi(c)$ with respect to $c$ at $c_0$. Denote $\eta_0(c_0) = \varphi(c_0)$, $\eta_1(c_0) = \varphi^{(1)}(c_0)$, and $\eta(c_0) = (\eta_0(c_0), \eta_1(c_0))'$. 
4.3. METHODS OF ESTIMATION

We further introduce the following notation: \( C_{it} = \mathbf{u}'_t \xi, \phi_{it}(c, \xi) = (1, C_{it} - c), \)
\( \Phi_{it}(c, \xi) = \phi_{it}(c, \xi), \) and \( \Delta_i = \text{diag}(p_{it}^{(1)}, t = 1, \ldots, T), \)
where \( p_{it}^{(1)} \) is the first order derivative of \( p_{it} \) evaluated at \( x'_t \beta + \varphi(\mathbf{u}'_t \xi). \) Let \( K(c) \) be a
kernel function (or a symmetric density function) with a compact support and \( h \) be a
bandwidth. Denote \( K_h(a) = K(a/h)/h \) and \( K_{ih}(\cdot) = \text{diag}(K_h(C_{it} - c), t = 1, \ldots, T). \)

Recall that we partitioned \( Y_i \) into the observed components \( Y_{io} \) and the unob-
served components \( Y_{iu}. \) Similarly, we consider partitioning the mean vector \( p_i \) into
\( p_{io} \) and \( p_{iu}. \) Here we incorporate the inverse probability-weights of Robins et al.
(1995) into the corresponding estimating equations. We propose a weighted local
linear kernel method and weighted second order GEE approach for simultaneous
estimation of \( \varphi(\cdot), \beta, \xi \) and \( \alpha \) for binary longitudinal data with dropouts. Below
we describe an algorithm for simultaneous estimation of the mean parameters \( \beta, \xi \)
and the correlation parameters \( \alpha. \)

**Step 0.** (Initialization step). Fit a parametric logistic linear model to obtain initial
values \( (\beta_0, \xi_0, \alpha_0) \) and set \( \hat{\xi} = \xi_0/\|\xi_0\|, \hat{\beta} = \beta_0 \) and \( \hat{\alpha} = \alpha_0. \)

**Step 1.** For a given point \( c \) in a selected grid find \( \hat{\varphi}(c, \hat{\alpha}, \hat{\beta}) = \eta_0(c) \) and \( \eta_1(c) \) by solving
the following equations
\[
\sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{\Phi}'_{i}(c) \tilde{\Delta}^o_{i} \tilde{K}_{ih}(c) \tilde{V}_i^{-1} (Y_{io} - \tilde{p}_{io}) = 0 \quad (4.11)
\]
with respect to \( \eta(c), \) where \( \tilde{\Phi}_{i}(c) \) and \( \tilde{K}_{ih}(c) \) are respectively \( \Phi_{i}(c) \) and \( K_{ih}(c) \)
with \( \xi \) replaced by \( \hat{\xi}, \tilde{p}_{io} \) is the vector that consist of the observed components
4.3. METHODS OF ESTIMATION

of \((\tilde{p}_{11}, \ldots, \tilde{p}_{it})\), \(\tilde{p}_{it} = g(x'_i\hat{\beta} + \phi_{it}(c, \hat{\xi}(c)), g(a) = \text{expit}(a) = \exp(a)/(1 + \exp(a)), \Delta_{i}^o\) is \(\Delta_i\) with \(p_{it}\) replaced by observed \(\tilde{p}_{it}\), and \(\tilde{V}_i\) is the independence working matrix diag\{\(\tilde{p}_{it}(1 - \tilde{p}_{it})\)\} which contains only observed \(\tilde{p}_{it}\)'s.

**Step 2.** Given the estimate \(\hat{\varphi}(c; \hat{\beta}, \hat{\xi}, \hat{\alpha}) = \hat{\eta}_0(c)\) and \(\hat{\eta}_1(c)\) for points \(c\) in the selected grid, update \((\hat{\beta}, \hat{\xi}, \hat{\alpha})\) by solving the following equations for \(\beta, \xi\) and \(\alpha\):

\[
\tilde{U}_{\beta,\xi}(\beta, \xi, \alpha, \hat{\varphi}(.) = \sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{D}'_i \tilde{V}_i^{-1} \{Y_i^o - \hat{p}_i^o(\beta, \xi, \alpha)\} = 0, \quad (4.12)
\]

\[
\tilde{U}_\alpha(\beta, \xi, \alpha, \hat{\varphi}(.) = \sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{G}'_i \tilde{W}_i^{-1} \left(\hat{Z}_i^o - \rho_i^o\right) = 0, \quad (4.13)
\]

where \(\hat{p}_i^o(\beta, \xi, \alpha)\) is the vector that consists of the observed components of \((\hat{p}_{i1}(\beta, \xi, \alpha), \ldots, \hat{p}_{iT}(\beta, \xi, \alpha))'\) with

\[
\hat{p}_{it}(\beta, \xi, \alpha) = g\left(x'_i\beta + \hat{\varphi}(u'_i\xi; \hat{\beta}, \hat{\xi}, \hat{\alpha})\right); \quad (4.14)
\]

\(\tilde{D}_i, \tilde{V}_i, \hat{Z}_i^o\) and \(\tilde{W}_i\) are the same as \(D_i, V_i, Z_i\) and \(W_i\) in (3.12) and (3.13) respectively with \(p_i\) replaced by \(\hat{p}_i^o(\beta, \xi, \alpha)\).

**Step 3.** Repeat steps 1 and 2 until the convergence of \((\hat{\beta}, \hat{\xi}, \hat{\alpha})\).

When implementing the foregoing algorithm it is often feasible to choose an initial value that is the estimate obtained from fitting a GEE model assuming a \(\varphi(.)\) as the identity function and then estimating the model parameters based on a linear single-index model or an ordinary logistic regression model. Our numerical
experience indicates that the algorithm is not severely but somewhat sensitive to the choice of initial values. The convergence criteria may be based on the difference between the estimates of two successive iterations or the absolute values of the estimating functions evaluated at the previous iteration.

We conclude this section with a discussion on bandwidth selection. As bandwidth $h$ affects both bias and variance estimate, there is a trade-off between suitable bias and variance estimate. Bias correction requires the choice of a relatively small bandwidth, whereas variance estimate needs a larger value of the bandwidth. In principle, the bandwidth selection is data driven, and traditional methods such as the cross-validation approach may be applied to select a proper bandwidth $h$ based on available data. Carroll et al. (1997) noted that a sensible choice of the bandwidth $h$ is generally difficult. Instead, they suggested an ad hoc bandwidth, given by $\hat{h}_{\text{opt}} \times K^{-2/15} = O(K^{-1/3})$, which satisfies $K h^4 \to 0$ and $K h^2/(\log K)^2 \to \infty$.

### 4.3.2 Estimating Dropout Probabilities

Prior to solving the iterative equations (4.11), (4.12) and (4.13), we estimate the response probability weights $\nu_{im}$. From (4.10) we find the pseudo-likelihood function for $\tau$ as

$$L(\tau) = \prod_{i=1}^{K} P(M_i = m_i | y_{it}', \tau)$$

$$= \prod_{i=1}^{K} \left\{ \prod_{t=2}^{m_i-1} \frac{1}{1 + \exp(y_{it}')} \right\} \left\{ \frac{\exp(y_{im}')}{1 + \exp(y_{im}')} \right\}^{I\{m_i \leq T\}}. \quad (4.15)$$
4.3. METHODS OF ESTIMATION

where \( y_{it}^* = (1, y_{i,t-1}, y_{it}, x_{it}', u_{it}')' \) and \( \tau = (\tau_0, \tau_1, \tau_2, \ldots)' \). Let

\[
\logit\{p_{it}^*(\tau)\} = y_{it}^* \tau. \tag{4.16}
\]

We find the pseudo-ML estimator of \( \tau \) by maximizing the above likelihood function. From (4.15), the pseudo-score equations for \( \tau \) takes the form

\[
S(\tau) = \sum_{i=1}^{K} \left\{- \sum_{t=2}^{m_i-1} p_{it}^*(\tau) y_{it}^* + I(m_i \leq T) \{1 - p_{im}^*(\tau)\} y_{im}^* \right\} = 0. \tag{4.17}
\]

The approximate variance of the pseudo-ML estimator of \( \tau \) is obtained from the Information matrix

\[
I(\tau) = \sum_{i=1}^{K} \sum_{t=2}^{\min(m_i, T)} p_{it}^*(\tau)(1 - p_{it}^*(\tau)) y_{it}^* y_{it}'. \tag{4.18}
\]

We apply Newton-Rapson iterative algorithm to solve the estimating equations for the pseudo-ML estimator \( \hat{\tau} = (\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2, \ldots)' \). The predicted probabilities of response for individual \( i \) at time \( t \) is obtained as

\[
1 - \hat{p}_{it}^* = \frac{1}{1 + \exp(y_{it}^* \hat{\tau})}. \tag{4.19}
\]

Then we estimate the probability of dropping out for individual \( i \) at time \( m_i \) by

\[
\tilde{\nu}_{im_i} = P(M_i = m_i | y_{it}^*, \hat{\tau}) = \left\{ \prod_{i=2}^{m_i-1} (1 - \hat{p}_{it}^*) \right\} \times \{\hat{p}_{i,m_i}\} I\{m_i \leq T\}. \tag{4.20}
\]

We replace \( \nu_{im} \)'s in equations (4.11), (4.12) and (4.12) by \( \tilde{\nu}_{im} \) and then solve these equations.
4.3. METHODS OF ESTIMATION

4.3.3 Estimating the Response Model

In order to estimate the parameters \((\beta, \xi, \alpha)\) from our response model (4.3), first we need to follow “Step 1” of our algorithm. In this step we estimate the single-index \(\varphi(.)\) by solving estimating equation (4.11). For this, we have

\[
\tilde{\Phi}'(c) \tilde{\Delta} \tilde{\kappa}_h(c) \tilde{V}^{-1} \left( Y_i^\circ - \tilde{p}_i^\circ \right) = \begin{bmatrix}
\sum_{t=1}^{m_i-1} K_h(C_{it} - c)(y_{it} - \tilde{p}_{it}) \\
\sum_{t=1}^{m_i-1} (C_{it} - c)K_h(C_{it} - c)(y_{it} - \tilde{p}_{it})
\end{bmatrix}, \tag{4.21}
\]

where \(\tilde{p}_{it} = \frac{e^{x^\prime_i \beta + \eta_0(c) + (C_{it} - c)\eta_1(c)}}{1 + e^{x^\prime_i \beta + \eta_0(c) + (C_{it} - c)\eta_1(c)}}\) and \(K(.)\) is the kernel density function (e.g. standard normal density function). We have

\[
\frac{\partial (y_{it} - \tilde{p}_{it})}{\partial \eta_0(c)} = -\tilde{p}_{it}(1 - \tilde{p}_{it}), \tag{4.22}
\]

\[
\frac{\partial (y_{it} - \tilde{p}_{it})}{\partial \eta_1(c)} = -(C_{it} - c)\tilde{p}_{it}(1 - \tilde{p}_{it}), \tag{4.23}
\]
and

\[
\frac{\partial}{\partial \eta(c)} \left\{ \tilde{\Phi}'(c) \tilde{\Delta}^\circ \tilde{K}_{ih}(c) \tilde{V}^{-1}_i (Y_i^o - \tilde{p}_i^o) \right\} = - \left[ \begin{array}{c}
\sum_{t=1}^{m_1-1} K_h(C_{it} - c) \frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_0} \\
\sum_{t=1}^{m_1-1} K_h(C_{it} - c) \frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_1}
\end{array} \right]
\]

\[
\times \left[ \begin{array}{c}
\sum_{t=1}^{m_1-1} (C_{it} - c) K_h(C_{it} - c) \frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_0} \\
\sum_{t=1}^{m_1-1} (C_{it} - c) K_h(C_{it} - c) \frac{\partial(y_{it} - \tilde{p}_{it})}{\partial \eta_1}
\end{array} \right]
\]

\[
= - \tilde{\Phi}'(c) \tilde{\Delta}^\circ \tilde{K}_{ih}(c) \tilde{V}^{-1}_i H^o,
\]

(4.24)

where \( H^o = \frac{\partial \tilde{p}_i^o}{\partial \eta(c)} \).

The iterative procedure for calculating the estimators \((\tilde{\eta}_o(c), \tilde{\eta}_1(c))\) for a given \(c\) in a selected grid begins with some starting values \((\tilde{\eta}_o(c)_0, \tilde{\eta}_1(c)_0)\), and produces updated values \((\tilde{\eta}_o(c)_{(s+1)}, \tilde{\eta}_1(c)_{(s+1)})\) from interim values \((\tilde{\eta}_o(c)_s, \tilde{\eta}_1(c)_s)\)
4.3. METHODS OF ESTIMATION

by means of the iterative equations

\[ \eta_{s+1} = \eta_s + \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} \Phi_i'(c) \Delta_{i}^{\phi} K_{ih}(c) V_i^{-1} H^o \right) \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{\Phi}_i'(c) \Delta_{i}^{\phi} \tilde{K}_{ih}(c) \tilde{V}_i^{-1} (Y^o_i - \tilde{p}_i) \right), \]

(4.25)

for \( s = 0, 1, 2, \ldots \), where the second term on the right side of the above equation is evaluated at the current estimates \( (\hat{\eta}_o(c), \hat{\eta}_t(c)) \).

Given the estimate \( \hat{\varphi}(c; \hat{\beta}, \hat{\xi}, \hat{\alpha}) = \hat{\eta}_0(c) \) and \( \hat{\eta}_t(c) \) for points \( c \) in the selected grid we can find

\[ \hat{p}_{it} = \frac{e^{x_{it}^{\beta+\hat{\eta}_0(c)}}}{1 + e^{x_{it}^{\beta+\hat{\eta}_0(c)}}} \]

(4.26)

and

\[ \frac{\partial \hat{p}_{it}}{\partial \xi_j} = u_{it,j} \hat{\eta}_t(c) \hat{p}_{it}(1 - \hat{p}_{it}). \]

(4.27)

The iterative procedure for calculating the estimators of the regression parameters \( \left( \hat{\beta}, \hat{\xi}, \hat{\alpha} \right) \) begins with some starting values \( \left( \hat{\beta}_0, \hat{\xi}_0, \hat{\alpha}_0 \right) \) from “Step 0” of our algorithm and produces updated values \( \left( \hat{\beta}_{s+1}, \hat{\xi}_{s+1}, \hat{\alpha}_{s+1} \right) \) from interim values \( \left( \hat{\beta}_s, \hat{\xi}_s, \hat{\alpha}_s \right) \) by means of the iterative equations

\[ \left( \hat{\beta}'_{s+1}, \hat{\xi}'_{s+1} \right)' = \left( \hat{\beta}', \hat{\xi}'_s \right)' + \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{D}_i \tilde{V}_i^{-1} \tilde{D}_i \right)^{-1} \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} \tilde{D}_i \tilde{V}_i^{-1} \{ Y^o_i - \tilde{p}_i(\beta, \xi, \alpha) \} \right), \]

(4.28)

and

\[ \hat{\alpha}_{s+1} = \hat{\alpha}_s + \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' \tilde{W}_i^{-1} G_i \right)^{-1} \left( \sum_{i=1}^{K} \frac{1}{\nu_{im}} G_i' \tilde{W}_i^{-1} \left( \tilde{Z}_i^o - \tilde{\rho}_i \right) \right), \]

(4.29)
for \( s = 0, 1, 2, \ldots \), where the second term on the right side of each estimating equation is evaluated at the current estimates \( \hat{\beta}_s, \hat{\xi}_s, \hat{\alpha}_s \).

### 4.4 Conclusions

The purpose of this research was to propose and explore a semiparametric approach to analyzing longitudinal binary data with dropouts. In our proposed method, we incorporate the inverse probability-weights into the estimating equations in order to compensate for the missing observations. We also incorporate true correlation structures into our estimating equations. We present an application of this proposed method in the next chapter using the GenIMS data introduced earlier.
Chapter 5

Application: GenIMS Study

5.1 Introduction

In this chapter we present an analysis of the Genetic and Inflammatory Marker of Sepsis (GenIMS) data. The GenIMS study is a longitudinal cohort study of patients with community acquired pneumonia who were recruited from 28 US academic and community hospitals. A goal of the GenIMS study was to understand the longitudinal response of coagulation activation in the progression of pneumonia to severe sepsis in patients. In this study, one of the important coagulation biomarkers for understanding the mechanisms of pneumonia progression to sepsis is antithrombin. Longitudinal measurements on this antithrombin biomarker were obtained from the patients daily for the first seven days of hospitalization. But some patients’ responses were missing at the follow-up times due to the early discharge from the
hospital or death during the first seven days. Figure 5.1 presents boxplots of the antithrombin data over seven hospital days for each of the two groups, no severe sepsis and severe sepsis. Median levels and trends of the antithrombin biomarker differ in the two groups of patients over seven days.

The study involved 1080 patients. Among them, 339 completed the study, 718 dropped out and 23 died during the hospitalization. Roughly, two-third of the patients have missing observations due to drop-outs. We assume that this drop-out process is MAR. Data from person-exams occurring after a person’s first missed exam were omitted to create a data set with monotone missingness. Specifically,
324 person-exams of a total of 4577 were omitted to create a monotone data set.

Here our goal is to draw inferences on the change in antithrombin level of patients in the presence of missing data and intraperson correlation based on our proposed methods. The model parameters were estimated and compared using the ordinary second order GEE approach (Method 1) of Prentice (1988), the proposed unweighted semi-parametric method (Method 2), a weighted semiparametric method where weights are used only to estimate the regression and association parameters (Method 3) and our proposed weighted semiparametric method (Method 4).

### 5.2 Model and Notation

We set the binary response variable $Y_{it} = 1$ if the antithrombin biomarker level of the $i$th patient is above 89 (the overall median value of antithrombin biomarker level) at the $t$th visit, and 0 if the antithrombin biomarker level is less than or equal to 89 for $i = 1, \ldots, 1080$ and $t = 1, \ldots, 7$. Among the predictors considered in the study, “Sepsis” represents an indicator variable which is defined to be 1 if a patient has a severe sepsis at baseline, and 0 if not. Similarly, the binary covariates “Respdis”, “Carddis”, and “Diab” are used to indicate the disease status of a patient at baseline for respiratory disease, cardiac disease, and diabetes respectively. “Sex” is labeled as 1 for male and 2 for female, “Age” denotes the baseline age of a patient in
years and “Time” represents time $t$. Our goal is to investigate how these covariates effect the longitudinal response of antithrombin measurements.

The marginal mean response $p_{it} = E(Y_{it})$ is defined as a function of the covariates in the form

$$\text{logit}(p_{it}) = \beta_0 + \beta_1 \text{Sepsis}_i + \beta_2 \text{Respdis}_i + \beta_3 \text{Carddis}_i + \beta_4 \text{Diab}_i$$

$$+ \beta_5 \text{Sex}_i + \varphi(\xi_1 \text{Age}_i + \xi_2 \text{Time}_i),$$  \hspace{1cm} (5.1)

where $\varphi(\cdot)$ is an unknown function. Exchangeable association structure is modeled here. We take the standard normal density as the kernel. The data-driven bandwidth $h$ is used, as discussed in Chapter 3.

To estimate the model parameters in (5.1) by the WGEE methods, we first estimate the inverse probability weights based on the missing data model

$$\text{logit}(p_{it}^*) = \tau_0 + \tau_1 y_{i,t-1} + \tau_2 \text{Sepsis}_i + \tau_3 \text{Respdis}_i + \tau_4 \text{Carddis}_i + \tau_5 \text{Diab}_i$$

$$+ \tau_6 \text{Sex}_i + \tau_7 \text{Age}_i + \tau_8 \text{Time}_i,$$  \hspace{1cm} (5.2)

where $p_{it}^* = P(R_{it} = 0|R_{i1} = \ldots, R_{i,t-1} = 1, y_i, x_i, \tau)$ is the conditional probability that the $i$th individual drops out at time $t$.

### 5.3 Methods of Estimation

For calculating the GEE estimators, prior to solving the iterative equations, we estimate the response probability weights $\nu_{im}$ using (4.10). We find the pseudo-
5.3. METHODS OF ESTIMATION

likelihood function for $\tau$ as

$$L(\tau) = \prod_{i=1}^{K} P(M_i = m_i | y_{i1}, \ldots, y_{im_i-1}, x_i, \tau)$$

$$= \prod_{i=1}^{K} \left( \prod_{t=2}^{m_i-1} \frac{1}{1 + \exp(\tau_0 + \tau_1 y_{i,t-1} + x'_i \tau^*)} \right) \exp(\tau_0 + \tau_1 y_{i,t-1} + x'_i \tau^*) \left\{ \frac{1 - \hat{p}_{it}^*}{1 + \exp(\tau_0 + \tau_1 y_{i,t-1} + x'_i \hat{\tau}^*)} \right\}^{I\{m_i \leq T\}} \right) .$$ (5.3)

where $x_i = (\text{Sepsis}_i, \text{Respdis}_i, \text{Carddis}_i, \text{Diab}_i, \text{Sex}_i, \text{Age}_i, \text{Time}_{it})'$, $\tau = (\tau_0, \tau_1, \tau^*)'$ and $\tau^* = (\tau_2, \ldots, \tau_8)'$.

We find the pseudo-ML estimator $\hat{\tau}$ by maximizing the above likelihood function. The predicted probabilities of response for individual $i$ at time $t$ is obtained as

$$1 - \hat{p}_{it}^* = \frac{1}{1 + \exp(\tau_0 + \tau_1 y_{i,t-1} + x'_i \hat{\tau}^*)} .$$ (5.4)

Then we estimate the probability of dropping out for individual $i$ at time $m_i$ by

$$\hat{\nu}_{im_i} = P(M_i = m_i | y_{i1}, \ldots, y_{im_i}, \hat{\tau}) = \left\{ \prod_{i=2}^{m_i-1} (1 - \hat{p}_{it}^*) \right\} \times \left\{ \hat{p}_{im_i}^* \right\}^{I\{m_i \leq T\}} \right) .$$ (5.5)

The following four methods were compared in the GenIMS data analysis:

- **GEE:** Estimators of $\beta, \xi$ and $\alpha$ are obtained by solving the unweighted GEEs (2.16) and (2.17), following Prentice (1988). Here we consider $\varphi(.)$ as $I(.)$.

- **SGEE:** Here we consider $\varphi(.)$ as unknown function and estimate $\hat{\varphi}(.)$ by solving (3.14). Estimators of $\beta, \xi$ and $\alpha$ are obtained by solving the unweighted GEEs (3.15) and (3.16).
5.4. RESULTS

WSGEE1: Here we consider \( \varphi(.) \) as unknown function and obtain \( \hat{\varphi}(.) \) by solving (3.14). In order to estimate \( \beta, \xi \) and \( \alpha \) we solve following two weighted GEEs

\[
\sum_{i=1}^{K} \frac{1}{\hat{\nu}_{im}} \hat{D}_i' \hat{V}_i^{-1} \{ Y_i^o - \hat{p}_i^o(\beta, \xi, \alpha) \} = 0, \quad (5.6)
\]

\[
\sum_{i=1}^{K} \frac{1}{\hat{\nu}_{im}} \hat{G}_i' \hat{W}_i^{-1} \left( \hat{Z}_i^o - \rho_i^o \right) = 0, \quad (5.7)
\]

where \( \hat{\nu}_{im} \) is estimated by solving (5.5).

WSGEE2: We adopt our proposed method from chapter 4. Here we consider \( \varphi(.) \) as unknown function and obtain \( \hat{\varphi}(.) \) by solving (4.11). Estimators of \( \beta, \xi \) and \( \alpha \) are obtained by solving the weighted GEEs (4.12) and (4.13).

5.4 Results

Table 5.1 presents the pseudo-maximum likelihood estimates of the missing data model parameters \( \tau \), their standard errors, and the corresponding \( z \)-values. Results in this table suggest that the dropout probabilities vary across all the covariates as well as the antithrombin biomarker level of a patient at the previous time point. Patients are likely to have \( \exp(0.274) = 1.32 \) times higher odds to dropout if their biomarker level was greater than 89 at the previous time point. Also the study suggests that the patients who have any of the diseases (Sepsis, Respdis, Cardis or Diab) are less likely to dropout as compared to those who do not have those dis-
5.4. RESULTS

Table 5.1: ML Estimates and Standard Errors of Missing Data Model Parameters for the GenIMS Study.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>S.E</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.138</td>
<td>0.0514</td>
<td>-22.14</td>
</tr>
<tr>
<td>$Y_{t-1}$</td>
<td>0.274</td>
<td>0.0071</td>
<td>38.40</td>
</tr>
<tr>
<td>Sepsis</td>
<td>-0.389</td>
<td>0.0025</td>
<td>-157.51</td>
</tr>
<tr>
<td>Respdis</td>
<td>-0.135</td>
<td>0.0010</td>
<td>-135.32</td>
</tr>
<tr>
<td>Carddis</td>
<td>-0.126</td>
<td>0.0036</td>
<td>-34.88</td>
</tr>
<tr>
<td>Diab</td>
<td>-0.087</td>
<td>0.0004</td>
<td>-200.08</td>
</tr>
<tr>
<td>Sex</td>
<td>0.065</td>
<td>0.0111</td>
<td>5.85</td>
</tr>
<tr>
<td>Age</td>
<td>-0.142</td>
<td>0.0038</td>
<td>-37.00</td>
</tr>
<tr>
<td>Time</td>
<td>0.240</td>
<td>0.0015</td>
<td>160.35</td>
</tr>
</tbody>
</table>

eases. Male patients are likely to have somewhat higher odds to dropout than their female counterparts. Younger patients are more likely to miss an exam than older patients. We also find that time has a positive influence on the missingness of the data, i.e. the odds of dropout is higher at the later time points.

Figure 5.2 displays the estimates of the estimated curve $\hat{\varphi}(.)$ under SGEE, WSGEE1 and WSGEE2. Here $\hat{\varphi}(.)$ shows evidence of nonlinear trend under all three methods. Some strong curvature is visible for the estimates of single-indexes under all the methods. More local curvature is visible under the unweighted and partially weighted methods (SGEE and WSGEE1). WSGEE1 is showing more local curvatures than other two methods while WSGEE2 is showing the least local curvatures. In WSGEE2, we have incorporated the inverse probability weights into the estimating equations of the single-index in addition to the estimating equations for the regression and association parameters. Multiplying weights in general expanded our band width. As a result, we get less local curvature. In WSGEE1, we only incor-


5.4. RESULTS

porate inverse probability weights into the estimating equations of the regression and association parameters. As a consequence, the estimated curve shows signs of over fit. Our results suggests that the data are not well fitted by an ordinary GEE model. Inclusion of a nonlinear function $\varphi(.)$ allows model (5.1) to be more flexible to capture any curvature.

Table 5.2 reports the estimates of the model parameters, their standard error and the corresponding $z$-values. We observe that the estimates for the linear covariates from all four methods are generally close to each other for the regression parameters. We also note that for all the linear covariates, except for the covariate of “Sex”, the partially weighted and weighted semiparametric methods (WSGEE1 and WSGEE2) underestimate the coefficient and it is opposite for the “Sex” covariate.

It is evident from the results shown in table 5.2 that the nonlinear covariates “Age” and “Time” significantly affect the antithrombin levels of patients. The estimates and standard errors for these two methods are more or less similar for all semiparametric methods. WSGEE2 somewhat underestimates the “Time” coefficient from other two semiparametric methods. We also note that the $z$-values of nonlinear covariates for semiparametric methods are significantly larger than that of simple GEE method.
5.4. RESULTS

Figure 5.2: Estimated nonlinear curves for GenIMS study.
Table 5.2: Estimates and Standard Errors of Regression and Association Parameters for the GenIMS Study.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>GEE Estimate (SE)</th>
<th>z-value</th>
<th>SGEE Estimate (SE)</th>
<th>z-value</th>
<th>WSGEE1 Estimate (SE)</th>
<th>z-value</th>
<th>WSGEE2 Estimate (SE)</th>
<th>z-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.607 (0.276)</td>
<td>5.82</td>
<td>1.894 (0.177)</td>
<td>10.72</td>
<td>2.118 (0.189)</td>
<td>11.20</td>
<td>3.409 (0.188)</td>
<td>18.17</td>
</tr>
<tr>
<td>Sepsis</td>
<td>-0.496 (0.120)</td>
<td>-4.14</td>
<td>-0.506 (0.118)</td>
<td>-4.30</td>
<td>-0.471 (0.126)</td>
<td>-3.75</td>
<td>-0.479 (0.125)</td>
<td>-3.83</td>
</tr>
<tr>
<td>Respdis</td>
<td>0.458 (0.110)</td>
<td>4.16</td>
<td>0.456 (0.111)</td>
<td>4.12</td>
<td>0.428 (0.117)</td>
<td>3.67</td>
<td>0.428 (0.116)</td>
<td>3.70</td>
</tr>
<tr>
<td>Carddis</td>
<td>-0.302 (0.134)</td>
<td>-2.25</td>
<td>-0.323 (0.130)</td>
<td>-2.48</td>
<td>-0.272 (0.135)</td>
<td>-2.01</td>
<td>-0.288 (0.136)</td>
<td>-2.13</td>
</tr>
<tr>
<td>Diab</td>
<td>0.360 (0.148)</td>
<td>2.43</td>
<td>0.359 (0.147)</td>
<td>2.44</td>
<td>0.340 (0.153)</td>
<td>2.22</td>
<td>0.334 (0.153)</td>
<td>2.18</td>
</tr>
<tr>
<td>sex</td>
<td>-0.567 (0.105)</td>
<td>-5.38</td>
<td>-0.545 (0.105)</td>
<td>-5.17</td>
<td>-0.577 (0.112)</td>
<td>-5.17</td>
<td>-0.569 (0.111)</td>
<td>-5.12</td>
</tr>
<tr>
<td>Age</td>
<td>-0.182 (0.033)</td>
<td>-5.58</td>
<td>-0.719 (0.019)</td>
<td>-38.41</td>
<td>-0.721 (0.020)</td>
<td>-35.62</td>
<td>-0.725 (0.017)</td>
<td>-41.64</td>
</tr>
<tr>
<td>Time</td>
<td>0.135 (0.018)</td>
<td>7.61</td>
<td>0.695 (0.028)</td>
<td>25.10</td>
<td>0.693 (0.030)</td>
<td>23.02</td>
<td>0.688 (0.032)</td>
<td>21.75</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.473 (0.019)</td>
<td>24.64</td>
<td>0.478 (0.023)</td>
<td>20.60</td>
<td>0.474 (0.025)</td>
<td>18.88</td>
<td>0.472 (0.025)</td>
<td>19.11</td>
</tr>
</tbody>
</table>
5.4. RESULTS

The antithrombin levels are higher among patients with respiratory disease and/or diabetes, and lower among patients with cardiac disease and/or severe sepsis. Females have lower antithrombin levels as compared to male. The antithrombin measurements appear to decrease with increased age. The antithrombin levels also increase over time. The association coefficient $\alpha$ is also significant under all the methods.
Chapter 6

Conclusions

In this thesis, we have proposed and explored semiparametric marginal models for analyzing binary longitudinal data with or without dropouts. We specifically considered joint estimation of the marginal mean parameters and the association parameters by using second order generalized estimating equations when the marginal model is partially linear for binary longitudinal data.

We developed a set of weighted generalized estimating equations for fitting regression models to longitudinal binary responses. In the case of dropouts, we incorporate the inverse probability-weights of Robins et al. (1995) into the GEE approach of Prentice (1988) when analyzing longitudinal binary responses. Our simulation study demonstrates that the proposed WGEE2 approach generally provides unbiased and efficient estimators when the missing data mechanism follows a MAR model. Our analysis of longitudinal data on cigarette smoking trends shows
that this method improves the regression and association parameter estimates.

We also developed a semiparametric approach to analyzing longitudinal binary
data based on partially linear single-index models. We use the local polynomial
smoothing technique to estimate the single-index. We incorporated the true corre-
lation structure into our estimation algorithm. Our simulation study demonstrates
that if the true underlying model is partially linear then our proposed method gen-
erally provides unbiased and efficient estimators. It is evident from the simulations
and ICHS data analysis that, even if a nonparametric function is included to fit the
data whose marginal mean model is actually characterized by an ordinary general-
ized linear model, the proposed method still leads to consistent estimates, though
some efficiency loss may incur. In the analysis of the smoking trend data, we con-
sidered binary variables as the non-linear predictors which violated the underlying
smoothness assumption on \( \varphi(\cdot) \). In this application our estimates for the covariates
were more or less same as the estimates from the ordinary GEE model. Therefore,
the method we describe here is robust and has applications in a wide variety of
settings.

Finally, we have proposed a weighted local linear kernel method and weighted
second order GEE approach to conduct simultaneous estimation of \( \varphi(\cdot), \beta, \xi \) and \( \alpha \)
for binary longitudinal data with dropouts. We combine and compare all our pro-
posed methods to analyze the recent Genetic and Inflammatory Marker of Sepsis
(GenIMS) data. The estimates of the single-index showed strong evidence of a non-
linear trend. The weighted semiparametric method (WSGEE2) showed curvature in the mean response, and appeared to provide more efficient estimators as compared to the other methods. This application suggests that our proposed method can be used for improving the estimates of the model parameters where data are not well fitted by an ordinary GEE model for binary longitudinal response with dropouts.

6.1 Future Research

In this thesis, we considered only monotone missingness where subjects drop out prior to the end of the study and do not return. In practice, the pattern of missing data is rarely monotone, but it is close to monotone (Little and Rubin, 1987). Future research is warranted to explore and evaluate non-monotone missingness patterns in the data.

We use the local polynomial smoothing technique to estimate the single-index. Other semiparametric additive models (e.g., spline regression) can be used to estimate the single-index and compared with our proposed method.
Bibliography


[65] Wu, L. (2002). A joint model for nonlinear mixed-effects models with cen-
soring and covariates measured with error, with application to AIDS studies. 


Appendix A

R Codes for CARDIA Study Data Analysis

unlink(".RData")

library(stats)
library(boot)
library(glmmML)
#library(mvtnorm)
library(mvtBinaryEP)
library(geepack)
library(Matrix)

#-----------------------------Original Data-----------------------------#
smoke_data = read.table("/Users/.......
longitudinal_smoke_data.txt",header=T,sep="",quote="")

smoke_data = data.frame(smoke_data)
id=smoke_data[,1]
yr_fu_1 = smoke_data[,5]
#----------Values-----------------#

K=length(unique(id))
T=4
N=T*K
O=cbind(table(id))
#simple_id=rep(1:K,0)
u_id=rep(1:length(id))
#----------Omitting nonmonotone observations----------#
#yrsum=NULL
#yrsum acceptable values are 0,2,7,14 for monotone missingness
# unacceptable values are 5,9,12 and part of 7.

non_id=NULL
j=1
for (i in 1:K)
{
  yrsum= sum(yr_fu_1[j:(j+O[i]-1)])
  if (yrsum==5)
    non_id=c(non_id,u_id[(j+O[i]-1)])
  if (yrsum==9)
    non_id=c(non_id,u_id[(j+O[i]-1)])
  if (yrsum==12)
    non_id=c(non_id,u_id[(j+O[i]-2):(j+O[i]-1)])
  if (yrsum==7)
    if(length(yr_fu_1[j:(j+O[i]-1)])==2)
      non_id=c(non_id,u_id[(j+O[i]-1)])

  j=j+O[i]
}

smoke_data_mon=smoke_data[-non_id,]

id_mon=smoke_data_mon[,1]
Per_mis=((N-length(id_mon))/N)*100
M=cbind(table(id_mon))+1
#Generating I(R_{it}=1)
mis_R=NULL
for (i in 1:K)
{
  mis_R= c(mis_R,rep(1,M[i]-1),rep(0,T+1-M[i]))
}
#----------Completing data with missing observations-------#
smoke_data_comp=NULL
j=1
for (i in 1:K)
{
  if (M[i]==2)
    smoke_data_comp = rbind(smoke_data_comp,smoke_data_mon[j:(j+M[i]-2),],
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],2,NA,smoke_data_mon[j,7]),
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],5,NA,smoke_data_mon[j,7]),
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],7,NA,smoke_data_mon[j,7]))
  if (M[i]==3)
    smoke_data_comp = rbind(smoke_data_comp,smoke_data_mon[j:(j+M[i]-2),],
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],5,NA,smoke_data_mon[j,7]),
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],7,NA,smoke_data_mon[j,7]))
  if (M[i]==4)
    smoke_data_comp = rbind(smoke_data_comp,smoke_data_mon[j:(j+M[i]-2),],
      c(smoke_data_mon[j,1],smoke_data_mon[j,2],smoke_data_mon[j,3],
        smoke_data_mon[j,4],7,NA,smoke_data_mon[j,7]))
  if (M[i]==5)
    smoke_data_comp = rbind(smoke_data_comp,smoke_data_mon[j:(j+M[i]-2),])
  j=j+M[i]-1
}

#----------main variables----------#
id_comp = smoke_data_comp[,1]
age = smoke_data_comp[,2]
birth = smoke_data_comp[,3]
education = smoke_data_comp[,4]
yr_fu = smoke_data_comp[,5]
smoke = smoke_data_comp[,6]
racesex = smoke_data_comp[,7]
#-----------------------Creating indicator variables-------------#
#--------------race sex indicators-----------#
bf=(racesex ==2)*1  # Black female
wm=(racesex ==3)*1  # White male
wf=(racesex ==4)*1  # White female

#--------------education indicators------------#
hs=(education==1)*1  # High school or less
sc=(education==2)*1  # Some college

#--------------rescalling age------------#
age_10=age/10

#---------------Final Data----------------#
MData=data.frame(cbind(id_comp,bf,wm,wf,age_10,hs,sc,yr_fu,smoke))

#--------------Weight estimation--------------------#
#-----------estimation of v_0 and v_1 by the dropout MLE------------------#
# Calculating observed Lambda from the Logit model
v_hat=rep(0,6)
smoke_t_1=NULL
bf_t_1=NULL
wm_t_1=NULL
wf_t_1=NULL
yr_t_1=NULL
L_hat_s=NULL
j=1
for (i in 1:K) {
  smoke_t_1= c(smoke_t_1,smoke[j:(j+T-2)])
  bf_t_1 = c(bf_t_1,bf[(j+1):(j+T-1)])
  wm_t_1 = c(wm_t_1,wm[(j+1):(j+T-1)])
  wf_t_1 = c(wf_t_1,wf[(j+1):(j+T-1)])
  yr_t_1 = c(yr_t_1,yr_fu[(j+1):(j+T-1)])
  j=j+T
if (M[i]==2)
L_hat_s= c(L_hat_s,c(1,0,0))
if (M[i]==3)
L_hat_s= c(L_hat_s,c(0,1,0))
if (M[i]==4)
L_hat_s= c(L_hat_s,c(0,0,1))
if (M[i]==5)
L_hat_s= c(L_hat_s,c(0,0,0))
}

#-----------Iteration---------#
for (i in 1:25)
{
lgt_lmd_hat=c(rep(v_hat[1],(N-K)))+v_hat[2]*smoke_t_1+v_hat[3]*bf_t_1
+v_hat[4]*wm_t_1+v_hat[5]*wf_t_1+v_hat[6]*yr_t_1
L_hat=inv.logit(lgt_lmd_hat) #Working with P(R_it=0\....)

S_gamma_1= -sum(na.omit((L_hat)*(1-L_hat_s)))+sum(na.omit((1-L_hat)*L_hat_s))
S_gamma_2= -sum(na.omit((smoke_t_1*L_hat)*(1-L_hat_s)))+sum(na.omit((smoke_t_1*(1-L_hat))*L_hat_s))
S_gamma_3= -sum(na.omit((bf_t_1*L_hat)*(1-L_hat_s)))+sum(na.omit((bf_t_1*(1-L_hat))*L_hat_s))
S_gamma_4= -sum(na.omit((wm_t_1*L_hat)*(1-L_hat_s)))+sum(na.omit((wm_t_1*(1-L_hat))*L_hat_s))
S_gamma_5= -sum(na.omit((wf_t_1*L_hat)*(1-L_hat_s)))+sum(na.omit((wf_t_1*(1-L_hat))*L_hat_s))
S_gamma_6= -sum(na.omit((yr_t_1*L_hat)*(1-L_hat_s)))+sum(na.omit((yr_t_1*(1-L_hat))*L_hat_s))
S_gamma= c(S_gamma_1,S_gamma_2,S_gamma_3,S_gamma_4,S_gamma_5,S_gamma_6)
I_gamma_1= sum(na.omit(L_hat*(1-L_hat)))
I_{\gamma_2} = \text{sum(na.omit((smoke_t_1^2)*L_{hat}*(1-L_{hat})))}

I_{\gamma_{12}} = \text{sum(na.omit(smoke_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_3} = \text{sum(na.omit((bf_t_1^2)*L_{hat}*(1-L_{hat})))}

I_{\gamma_{13}} = \text{sum(na.omit(bf_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{23}} = \text{sum(na.omit(bf_t_1*smoke_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_4} = \text{sum(na.omit((wm_t_1^2)*L_{hat}*(1-L_{hat})))}

I_{\gamma_{14}} = \text{sum(na.omit(wm_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{24}} = \text{sum(na.omit(wm_t_1*smoke_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{34}} = \text{sum(na.omit(bf_t_1*wm_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_5} = \text{sum(na.omit((wf_t_1^2)*L_{hat}*(1-L_{hat})))}

I_{\gamma_{15}} = \text{sum(na.omit(wf_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{25}} = \text{sum(na.omit(wf_t_1*smoke_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{35}} = \text{sum(na.omit(bf_t_1*wf_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{45}} = \text{sum(na.omit(wm_t_1*wf_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_6} = \text{sum(na.omit((yr_t_1^2)*L_{hat}*(1-L_{hat})))}

I_{\gamma_{16}} = \text{sum(na.omit(yr_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{26}} = \text{sum(na.omit(yr_t_1*smoke_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{36}} = \text{sum(na.omit(bf_t_1*yr_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{46}} = \text{sum(na.omit(wm_t_1*yr_t_1*L_{hat}*(1-L_{hat})))}

I_{\gamma_{56}} = \text{sum(na.omit(wf_t_1*yr_t_1*L_{hat}*(1-L_{hat})))}
\[ I_{\gamma} = \text{matrix}(c(I_{\gamma 1}, I_{\gamma 12}, I_{\gamma 13}, I_{\gamma 14}, I_{\gamma 15}, I_{\gamma 16}, I_{\gamma 12}, I_{\gamma 2}, I_{\gamma 23}, I_{\gamma 24}, I_{\gamma 25}, I_{\gamma 26}, I_{\gamma 13}, I_{\gamma 23}, I_{\gamma 3}, I_{\gamma 34}, I_{\gamma 35}, I_{\gamma 36}, I_{\gamma 14}, I_{\gamma 24}, I_{\gamma 34}, I_{\gamma 4}, I_{\gamma 45}, I_{\gamma 46}, I_{\gamma 15}, I_{\gamma 25}, I_{\gamma 35}, I_{\gamma 45}, I_{\gamma 5}, I_{\gamma 56}, I_{\gamma 16}, I_{\gamma 26}, I_{\gamma 36}, I_{\gamma 46}, I_{\gamma 56}, I_{\gamma 6}), 6, 6) \]

\[ S_{\varhat} = S_{\varhat} + \text{solve}(I_{\gamma}) \times S_{\gamma} \times \text{t}(S_{\gamma}) \times \text{solve}(I_{\gamma}) \]

\[ v_{\varhat} = \text{solve}(I_{\gamma}) \times S_{\varhat} \times \text{solve}(I_{\gamma}) \]

\[ v_{\text{se}} = \sqrt{(v_{\varhat} \times \text{diag}(1, \text{length}(v_{\hat})) \times \text{rep}(1, \text{length}(v_{\hat})))} \]

---

#-------Estimating Probabilities of M--------#

\[ \text{lgt}_{lmd_{\hat}} = c(\text{rep}(v_{\hat}[1], (N-K)) + v_{\hat}[2] \times \text{smoke}_{t-1} + v_{\hat}[3] \times \text{bf}_{t-1} + v_{\hat}[4] \times \text{wm}_{t-1} + v_{\hat}[5] \times \text{wf}_{t-1} + v_{\hat}[6] \times \text{yr}_{t-1} \]

\[ \text{L}_{\hat} = \text{inv.logit}((\text{lgt}_{lmd_{\hat}})) \]

\[ \text{ProbM}_{\hat} = \text{NULL} \]

\[ j = 1 \]

\[ \text{for} \ (i \ \text{in} \ 1:K) \]

\[ \{ \]

\[ \text{ProbM}_{\hat} = \text{rbinding}(\text{ProbM}_{\hat}, \text{c}(\text{L}_{\hat}[j], (1-\text{L}_{\hat}[j]) \times \text{L}_{\hat}[j+1], (1-\text{L}_{\hat}[j]) \times (1-\text{L}_{\hat}[j+1]) \times \text{L}_{\hat}[j+2], (1-\text{L}_{\hat}[j]) \times (1-\text{L}_{\hat}[j+1]) \times (1-\text{L}_{\hat}[j+2]))) \]

\[ j = j + (T-1) \]

\[ \} \]

#ProbM_{\hat}
#--------Serial GEE------#

fml_mis= smoke~age_10+yr_fu+hs+sc+bf+wm+wf #bf+wm+wf+hs+sc+
Se_gee_mis <- geeglm(fml_mis, data=MData, id=id_comp,
family=binomial("logit"),scale.fix=TRUE, corstr="ar1")

beta_uw=rep(0,8)
alpha_uw=0

iter=25

for (k in 1:iter)
{
  R_alpha= matrix(c(1,alpha_uw,alpha_uw^2,alpha_uw^3,
                   alpha_uw,1,alpha_uw,alpha_uw^2,
                   alpha_uw^2,alpha_uw,1,alpha_uw,alpha_uw^3,
                   alpha_uw^2,alpha_uw,1),nrow=T, ncol=T)
  del_i=c(alpha_uw,alpha_uw^2,alpha_uw^3,alpha_uw,alpha_uw^2,alpha_uw)
  P=replace(inv.logit(lgt),(-mis_R+1)*c(1:N), values=NA)
  #B=diag(P*(1-P))
  D=cbind(P*(1-P), age_10*(P*(1-P)), yr_fu*(P*(1-P)),
          hs*(P*(1-P)), sc*(P*(1-P)), bf*(P*(1-P)), wm*(P*(1-P)),
          wf*(P*(1-P))))
  S= smoke-P

  I_term=0
  S_term=0
  S_var=0

  I_alpha= 0
  S_alpha= 0
  S_alpha_var=0
  j=1
  for (i in 1:K)
  {
    B_i=as.matrix(na.omit(diag(P[j:(j+T-1)]*(1-P[j:(j+T-1)])))[,1:(M[i]-1)])
#B_i=as.matrix(na.omit(B[j:(j+T-1),j:(j+T-1)])[,1:(M[i]-1)])
V_i= (B_i^(1/2))%*%R_alpha[1:dim(B_i)[1],1:dim(B_i)[2]]
%*%(B_i^\-(1/2))*(ProbM_hat[i,M[i]-1])
D_i= as.matrix(na.omit(D[j:(j+T-1),]))
S_i= na.omit(S[j:(j+T-1)])

I_term=I_term + t(D_i)%*%solve(V_i)%*%D_i
S_term=S_term + t(D_i)%*%solve(V_i)%*%S_i
S_var=S_var + t(D_i)%*%solve(V_i)%*%S_i%*%t(S_i)%*%solve(V_i)%*%D_i

---alpha estimation part-------#

if (M[i]>=3)
{
  z_i12 = ((smoke[j]-P[j])*(smoke[j+1]-P[j+1]))/(P[j]*(1-P[j])*P[j+1]*(1-P[j+1]))^(1/2)
z_i13 = ((smoke[j]-P[j])*(smoke[j+2]-P[j+2]))/(P[j]*(1-P[j])*P[j+2]*(1-P[j+2]))^(1/2)
z_i14 = ((smoke[j]-P[j])*(smoke[j+3]-P[j+3]))/(P[j]*(1-P[j])*P[j+3]*(1-P[j+3]))^(1/2)
z_i23 = ((smoke[j+1]-P[j+1])*(smoke[j+2]-P[j+2]))/(P[j+1]*(1-P[j+1])*P[j+2]*(1-P[j+2]))^(1/2)
z_i24 = ((smoke[j+1]-P[j+1])*(smoke[j+3]-P[j+3]))/(P[j+1]*(1-P[j+1])*P[j+3]*(1-P[j+3]))^(1/2)
z_i34 = ((smoke[j+2]-P[j+2])*(smoke[j+3]-P[j+3]))/(P[j+2]*(1-P[j+2])*P[j+3]*(1-P[j+3]))^(1/2)
Z_i=na.omit(c(z_i12,z_i13,z_i14,z_i23,z_i24,z_i34)- del_i)

W_i12 = 1 + (1-2*P[j])*(1-2*P[j+1])
*(P[j]*(1-P[j])*P[j+1]*(1-P[j+1]))^(1/2)*del_i[1] - del_i[1]^2
W_i13 = 1 + (1-2*P[j])*(1-2*P[j+2])
W_i14 = 1 + (1-2*P[j])*(1-2*P[j+3])
*(P[j]*(1-P[j])*P[j+3]*(1-P[j+3]))^(1/2)*del_i[3] - del_i[3]^2
W_i23 = 1 + (1-2*P[j+1])*(1-2*P[j+2])
*(P[j+1]*(1-P[j+1])*P[j+2]*(1-P[j+2]))^(1/2)*del_i[4] - del_i[4]^2
W_i24 = 1 + (1-2*P[j+1])*(1-2*P[j+3])
*(P[j+1]*(1-P[j+1])*P[j+3]*(1-P[j+3]))^(1/2)*del_i[5] - del_i[5]^2
W_i34 = 1 + (1-2*P[j+2])*(1-2*P[j+3])
*(P[j+2]*(1-P[j+2])*P[j+3]*(1-P[j+3]))^(1/2)*del_i[6] - del_i[6]^2
\[ W_i = \text{diag}(\text{na.omit}(c(W_{i12}, W_{i13}, W_{i14}, W_{i23}, W_{i24}, W_{i34})), \text{length}(Z_i)) \times (\text{ProbM}_{\hat{i}, M[i]-1}) \]

\[ E_i = c(1, 2\alpha_uw, 3\alpha_uw^2, 1, 2\alpha_uw, 1) \]
\[ \text{is.na}(E_i) = \text{is.na}(c(z_{i12}, z_{i13}, z_{i14}, z_{i23}, z_{i24}, z_{i34})) \times 1 \times c(1:6) \]
\[ E_i = \text{na.omit}(E_i) \]

\[ I_{\text{alpha}} = I_{\text{alpha}} + t(E_i)^{\text{solve}(W_i)} \times E_i \]
\[ S_{\text{alpha}} = S_{\text{alpha}} + t(E_i)^{\text{solve}(W_i)} \times Z_i \]
\[ S_{\text{alpha var}} = S_{\text{alpha var}} + t(E_i)^{\text{solve}(W_i)} \times Z_i \times t(Z_i) \times \text{solve}(W_i) \times E_i \]

\[ j = j + T \]

\[ \beta_uw = \beta_uw + \text{solve}(I_{\text{term}}) \times S_{\text{term}} \]
\[ \alpha_uw = \alpha_uw + \text{solve}(I_{\text{alpha}}) \times S_{\text{alpha}} \]

\[ \beta_{\text{var}}_uw = \text{solve}(I_{\text{term}}) \times S_{\text{var}} \times \text{solve}(I_{\text{term}}) \]
\[ \alpha_{\text{var}}_uw = \text{solve}(I_{\text{alpha}}) \times S_{\text{alpha var}} \times \text{solve}(I_{\text{alpha}}) \]

\# Se theta uw = c(\beta_uw, alpha_uw)
\# Se theta var uw = c((\beta_{\text{var}}_uw * \text{diag}(1, \text{length}(\beta_uw))))
\%*% rep(1, \text{length}(\beta_uw)), alpha_var_uw)
\# Se CI uw = cbind(Se theta uw + qnorm(0.025) * sqrt(Se theta var uw),
Se theta uw + qnorm(1 - 0.025) * sqrt(Se theta var uw))

\# Se theta 1w = c(\beta_uw, alpha_uw)
\# Se theta var 1w = c((\beta_{\text{var}}_uw * \text{diag}(1, \text{length}(\beta_uw))))
\%*% rep(1, \text{length}(\beta_uw)), alpha_var_uw)

Se theta rw = c(\beta_uw, alpha_uw)
Se theta var rw = c((\beta_{\text{var}}_uw * \text{diag}(1, \text{length}(\beta_uw))))
\%*% rep(1, \text{length}(\beta_uw)), alpha_var_uw)
#--------Exchangable GEE------#

fml_mis = smoke~age_10+yr_fu+hs+sc+bf+wm+wf  #bf+wm+wf+hs+sc+
gee_mis <- geeglm(fml_mis, data=MData, id=id_comp,
family=binomial("logit"),scale.fix=TRUE, corstr="exchangeable")

beta_uw=rep(0,8)
alpha_uw=0

iter=25

for (k in 1:iter)
{
R_alpha= matrix(rep(alpha_uw,T*T), ncol=T)
diag(R_alpha) = 1
del_i=rep(alpha_uw,6)
lgt=c(rep(beta_uw[1],N))+((beta_uw[2])*age_10)+ ((beta_uw[3])*yr_fu)
+((beta_uw[4])*hs)+((beta_uw[5])*sc)+((beta_uw[6])*bf)
+((beta_uw[7])*wm)+((beta_uw[8])*wf)
P=replace(inv.logit(lgt),(-mis_R+1)*c(1:N), values=NA)

#B=diag(P*(1-P))
D=cbind(P*(1-P), age_10*(P*(1-P)), yr_fu*(P*(1-P)), hs*(P*(1-P)),
sc*(P*(1-P)), bf*(P*(1-P)), wm*(P*(1-P)), wf*(P*(1-P)))
S= smoke-P

I_term=0
S_term=0
S_var=0

I_alpha= 0
S_alpha= 0
S_alpha_var=0
j=1
for (i in 1:K)
{
B_i=as.matrix(na.omit(diag(P[j:(j+T-1)]*(1-P[j:(j+T-1)])))[1:(M[i]-1)])
}
# B_i = as.matrix(na.omit(B[j:(j+T-1),j:(j+T-1)]),1:(M[i]-1))
V_i = (B_i^(1/2))%*%R_alpha[1:dim(B_i)[1],1:dim(B_i)[2]]
%*%(B_i^\top(1/2))*(ProbM_hat[i,M[i]-1])
D_i = as.matrix(na.omit(D[j:(j+T-1)]))
S_i = na.omit(S[j:(j+T-1)])

I_term = I_term + t(D_i)%*%solve(V_i)%*%D_i
S_term = S_term + t(D_i)%*%solve(V_i)%*%S_i
S_var = S_var + t(D_i)%*%solve(V_i)%*%S_i%*%t(S_i)%*%solve(V_i)%*%D_i

#--- alpha estimation part-------#

if (M[i]>=3) {
  z_i12 = ((smoke[j]-P[j])*(smoke[j+1]-P[j+1]))
           /(P[j] *(1-P[j]) *P[j+1] *(1-P[j+1]) )^(1/2)
  z_i13 = ((smoke[j]-P[j])*(smoke[j+2]-P[j+2]))
           /(P[j] *(1-P[j]) *P[j+2] *(1-P[j+2]) )^(1/2)
  z_i14 = ((smoke[j]-P[j])*(smoke[j+3]-P[j+3]))
           /(P[j] *(1-P[j]) *P[j+3] *(1-P[j+3]) )^(1/2)
  z_i23 = ((smoke[j+1]-P[j+1])*(smoke[j+2]-P[j+2]))
           /(P[j+1] *(1-P[j+1]) *P[j+2] *(1-P[j+2]) )^(1/2)
  z_i24 = ((smoke[j+1]-P[j+1])*(smoke[j+3]-P[j+3]))
           /(P[j+1] *(1-P[j+1]) *P[j+3] *(1-P[j+3]) )^(1/2)
  z_i34 = ((smoke[j+2]-P[j+2])*(smoke[j+3]-P[j+3]))
           /(P[j+2] *(1-P[j+2]) *P[j+3] *(1-P[j+3]) )^(1/2)

  Z_i = na.omit(c(z_i12,z_i13,z_i14,z_i23,z_i24,z_i34) - del_i)

  W_i12 = 1 + (1-2*P[j])*(1-2*P[j+1])
           *(P[j] *(1-P[j]) *P[j+1] *(1-P[j+1]) )^(1/2)*del_i[1] - del_i[1]^2
  W_i13 = 1 + (1-2*P[j])*(1-2*P[j+2])
  W_i14 = 1 + (1-2*P[j])*(1-2*P[j+3])
  W_i23 = 1 + (1-2*P[j+1])*(1-2*P[j+2])
  W_i24 = 1 + (1-2*P[j+1])*(1-2*P[j+3])
  W_i34 = 1 + (1-2*P[j+2])*(1-2*P[j+3])
\[ W_i = \text{diag}(\text{na.omit}(c(W_{i12}, W_{i13}, W_{i14}, W_{i23}, W_{i24}, W_{i34})), \text{length}(Z_i)) \times (\text{ProbM}_{\text{hat}[i,M[i]-1]}) \]

\[ E_i = \text{rep}(1, \text{length}(Z_i)) \]

\[ I_{\alpha} = I_{\alpha} + t(E_i) \times \text{solve}(W_i) \times E_i \]
\[ S_{\alpha} = S_{\alpha} + t(E_i) \times \text{solve}(W_i) \times Z_i \]
\[ S_{\alpha \text{ var}} = S_{\alpha \text{ var}} + t(E_i) \times \text{solve}(W_i) \times Z_i \times t(Z_i) \times \text{solve}(W_i) \times E_i \]

\[ j = j + T \]

\[ \beta_{uw} = \beta_{uw} + \text{solve}(I_{\text{term}}) \times S_{\text{term}} \]
\[ \alpha_{uw} = \alpha_{uw} + \text{solve}(I_{\alpha}) \times S_{\alpha} \]
\[ \beta_{\text{var}_{uw}} = \text{solve}(I_{\text{term}}) \times S_{\text{var}} \times \text{solve}(I_{\text{term}}) \]
\[ \alpha_{\text{var}_{uw}} = \text{solve}(I_{\alpha}) \times S_{\alpha \text{ var}} \times \text{solve}(I_{\alpha}) \]

\# \theta_{uw} = c(\beta_{uw}, \alpha_{uw})
\# \theta_{\text{var}_{uw}} = c((\beta_{\text{var}_{uw}} \times \text{diag}(1, \text{length}(\beta_{uw})))) \times \text{rep}(1, \text{length}(\beta_{uw})), \alpha_{\text{var}_{uw}}

\text{theta}_1w = c(\beta_{uw}, \alpha_{uw})
\text{theta}_\text{var}_1w = c((\beta_{\text{var}_{uw}} \times \text{diag}(1, \text{length}(\beta_{uw})))) \times \text{rep}(1, \text{length}(\beta_{uw})), \alpha_{\text{var}_{uw}}

\# \theta_{rw} = c(\beta_{uw}, \alpha_{uw})
\# \theta_{\text{var}_{rw}} = c((\beta_{\text{var}_{uw}} \times \text{diag}(1, \text{length}(\beta_{uw})))) \times \text{rep}(1, \text{length}(\beta_{uw})), \alpha_{\text{var}_{uw}}

\text{summary}(\text{gee_mis})
\text{print}(\theta_{uw})
\text{print}(\text{sqrt}(\theta_{\text{var}_{uw}}))
print(theta_1w)
print(sqrt(theta_var_1w))

print(theta_rw)
print(sqrt(theta_var_rw))

#---------------------Serial results-------------------#
summary(Se_gge_mis)

print(Se_theta_uw)
print(sqrt(Se_theta_var_uw))

print(Se_theta_1w)
print(sqrt(Se_theta_var_1w))

print(Se_theta_rw)
print(sqrt(Se_theta_var_rw))
Appendix B

R Codes for GenIMS Data Analysis

unlink(".RData")

library(stats)
l library(boot)
l library(glmML)
#library(mvtnorm)
l library(mvtBinaryEP)
l library(geepack)
l library(Matrix)
l library(splus2R)
l library(epicalc)
l library(pls)
l library(RODBC)
l library(gdata)
l library(lattice)
#-----------------------------Original Data------------------------#

GenIMS = read.xls("/Users/..../antithromb_salehin_aug_2013.xls", sheet = 1)

#-----------------------------Working Data------------------------#
Gen_data = GenIMS[,c(1,2,3,4,5,6,10,11,14,42,48)]

Sepsis = (Gen_data[,11]>2)*1
Athromb = Gen_data[,10]
Y = (Athromb>median(Athromb))*1
id = Gen_data[,1]
tme = Gen_data[,2]
last = Gen_data[,3]
mis = Gen_data[,4]
age = Gen_data[,5]/10
sex = Gen_data[,6]
respdis = Gen_data[,7]
carddis = Gen_data[,8]
diab = Gen_data[,9]

all_data=data.frame(cbind(id,tme,last,mis, Y,Sepsis,respdis,carddis,diab,sex,age))

#--------------Full Gee---------------#
fml= Y~Sepsis+respdis+carddis+diab+sex+age+tme
gee <- geeglm(fml, data=all_data, id=id,
family=binomial("logit"),scale.fix=TRUE, corstr="exchangeable")
summary(gee)

#--------Omitting Non-Monoton missingness----------#
O=cbind(table(id))
M=O+1
K=length(unique(id))
u_id=rep(1:length(id))

Gen_mon=NULL
M_ID=NULL
j=1
for (i in 1:K)
{
    r=j+O[i]-1

    M_id= (((c(tme[j]:tme[j]+O[i]-1)==tme[j]:r)*1)-1)*c(1:O[i])
    M_ID= c(M_ID,M_id)

    M_data=all_data[j:r,]

    j=j+O[i]
    if(sum(M_id)==0) Gen_mon=rbind(Gen_mon,M_data)
    else Gen_mon=rbind(Gen_mon,M_data[M_id,])
}

Gen_mon = data.frame(Gen_mon)

id = Gen_mon[,1]
tme = Gen_mon[,2]
last = Gen_mon[,3]
mis = Gen_mon[,4]
Y = Gen_mon[,5]
Sepsis = Gen_mon [,6]
respdis = Gen_mon[,7]
carddis = Gen_mon[,8]
diab = Gen_mon[,9]
sex= Gen_mon[,10]
age = Gen_mon[,11]

O=cbind(table(id))
K=length(unique(id))
T=7
N=length(id)

Z=cbind(age,tme)

#----------Monotone GEE----------

fml_mon= Y~Sepsis+respdis+carddis+diab+sex+age+tme
gee_mon <- geeglm(fml_mon, data=Gen_mon, id=id,
   family=binomial("logit"),scale.fix=TRUE, corstr="exchangeable")
summary(gee_mon)

#---------------------#
#----------------Weight estimation------------------------#
#-----------estimation of v_0 and v_1 by the dropout MLE-----------#

# Calculating observed Lambda from the Logit model

v_hat=rep(0,9)

Y_t_1=NULL
Sepsis_t_1=NULL
respdis_t_1=NULL
carddis_t_1=NULL
diab_t_1=NULL
sex_t_1=NULL
age_t_1=NULL
tme_t_1=NULL
L_hat_s=NULL
j=1
for (i in 1:K)
{
  #{if (O[i]==max(O))
  # Y_t_1= c(Y_t_1,Y[j:(j+O[i]-2)])
  # else
  # Y_t_1= c(Y_t_1,Y[j:(j+O[i]-1)])

  since all the covariates are baseline we can
  simply replace missing observation of any
  covariate by any corresponding observed
  covariate value for that individual.
  # this data set does not have any
information regarding covariates
for missing individuals

if (O[i]<T){
  Y_t_1= c(Y_t_1,Y[j:(j+O[i]-1)])
  Sepsis_t_1 = c(Sepsis_t_1,Sepsis[j:(j+O[i]-1)])
  respdis_t_1 = c(respdis_t_1,respdis[j:(j+O[i]-1)])
  carddis_t_1 = c(carddis_t_1,carddis[j:(j+O[i]-1)])
  diab_t_1 = c(diab_t_1,diab[j:(j+O[i]-1)])
  sex_t_1 = c(sex_t_1,sex[j:(j+O[i]-1)])
  age_t_1 = c(age_t_1,age[j:(j+O[i]-1)])
  tme_t_1 = c(tme_t_1,tme[j:(j+O[i]-1)])}

if (O[i]==T){
  Y_t_1= c(Y_t_1,Y[j:(j+O[i]-2)])
  Sepsis_t_1 = c(Sepsis_t_1,Sepsis[j:(j+O[i]-2)])
  respdis_t_1 = c(respdis_t_1,respdis[j:(j+O[i]-2)])
  carddis_t_1 = c(carddis_t_1,carddis[j:(j+O[i]-2)])
  diab_t_1 = c(diab_t_1,diab[j:(j+O[i]-2)])
  sex_t_1 = c(sex_t_1,sex[j:(j+O[i]-2)])
  age_t_1 = c(age_t_1,age[j:(j+O[i]-2)])
  tme_t_1 = c(tme_t_1,tme[j:(j+O[i]-2)])}

{j if (O[i]==T)
  L_hat_s= c(L_hat_s,rep(0,(T-1)))
  else
  L_hat_s= c(L_hat_s,c(rep(0,(O[i]-1))),1))}

j=j+O[i]
}

S_var_hat=0

#-----------Iteration---------#

for (i in 1:25)
{

lgt_lmd_hat=c(rep(v_hat[1],(length(Y_t_1))))+v_hat[2]*Y_t_1
+v_hat[3]*Sepsis_t_1+v_hat[4]*respdis_t_1+v_hat[5]*carddis_t_1
+v_hat[6]*diab_t_1+v_hat[7]*sex_t_1+v_hat[8]*age_t_1+v_hat[9]*tme_t_1
L_hat=inv.logit(lgt_lmd_hat) #Working with P(R_{it}=0\ldots)

S_gamma_1= -sum(na.omit((L_hat)
*(1-L_hat_s)))+sum(na.omit(((1-L_hat)*L_hat_s))
S_gamma_2= -sum(na.omit((Y_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((Y_t_1*(1-L_hat))*L_hat_s))
S_gamma_3= -sum(na.omit((Sepsis_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((Sepsis_t_1*(1-L_hat))*L_hat_s))
S_gamma_4= -sum(na.omit((respdis_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((respdis_t_1*(1-L_hat))*L_hat_s))
S_gamma_5= -sum(na.omit((carddis_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((carddis_t_1*(1-L_hat))*L_hat_s))
S_gamma_6= -sum(na.omit((diab_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((diab_t_1*(1-L_hat))*L_hat_s))
S_gamma_7= -sum(na.omit((sex_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((sex_t_1*(1-L_hat))*L_hat_s))
S_gamma_8= -sum(na.omit((age_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((age_t_1*(1-L_hat))*L_hat_s))
S_gamma_9= -sum(na.omit((tme_t_1*L_hat)
*(1-L_hat_s)))+sum(na.omit(((tme_t_1*(1-L_hat))*L_hat_s))
S_gamma= c(S_gamma_1,S_gamma_2,S_gamma_3,S_gamma_4,
S_gamma_5,S_gamma_6,S_gamma_7,S_gamma_8,S_gamma_9)

I_gamma_1= sum(na.omit(L_hat*(1-L_hat)))
I_gamma_2= sum(na.omit((Y_t_1^2)*L_hat*(1-L_hat)))
\[ I_{\gamma_{12}} = \sum(\text{na.omit}(Y_{t-1}L_{\hat{}}}*(1-L_{\hat{}})) \]

\[ I_{\gamma_{3}} = \sum(\text{na.omit}((\text{Sepsis}_{t-1}^2)L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{13}} = \sum(\text{na.omit}(\text{Sepsis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{23}} = \sum(\text{na.omit}(\text{Sepsis}_{t-1}Y_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{4}} = \sum(\text{na.omit}((\text{respdis}_{t-1}^2)L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{14}} = \sum(\text{na.omit}(\text{respdis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{24}} = \sum(\text{na.omit}(\text{respdis}_{t-1}Y_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{34}} = \sum(\text{na.omit}(\text{Sepsis}_{t-1}\text{respdis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{5}} = \sum(\text{na.omit}((\text{carddis}_{t-1}^2)L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{15}} = \sum(\text{na.omit}(\text{carddis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{25}} = \sum(\text{na.omit}(\text{carddis}_{t-1}Y_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{35}} = \sum(\text{na.omit}(\text{Sepsis}_{t-1}\text{carddis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{45}} = \sum(\text{na.omit}(\text{respdis}_{t-1}\text{carddis}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{6}} = \sum(\text{na.omit}((\text{diab}_{t-1}^2)L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{16}} = \sum(\text{na.omit}(\text{diab}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{26}} = \sum(\text{na.omit}(\text{diab}_{t-1}Y_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{36}} = \sum(\text{na.omit}(\text{Sepsis}_{t-1}\text{diab}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{46}} = \sum(\text{na.omit}(\text{respdis}_{t-1}\text{diab}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]

\[ I_{\gamma_{56}} = \sum(\text{na.omit}(\text{carddis}_{t-1}\text{diab}_{t-1}L_{\hat{}}*(1-L_{\hat{}}))) \]
I_{\gamma 7} = \text{sum(na.omit((sex_t_1^2)*L_{hat}*(1-L_{hat}))})

I_{\gamma 17} = \text{sum(na.omit(sex_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 27} = \text{sum(na.omit(sex_t_1*Y_{t_1}*L_{hat}*(1-L_{hat}))})

I_{\gamma 37} = \text{sum(na.omit(Sepsis_t_1*sex_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 47} = \text{sum(na.omit(respdis_t_1*sex_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 57} = \text{sum(na.omit(carddis_t_1*sex_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 67} = \text{sum(na.omit(diab_t_1*sex_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 8} = \text{sum(na.omit((age_t_1^2)*L_{hat}*(1-L_{hat}))})

I_{\gamma 18} = \text{sum(na.omit(age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 28} = \text{sum(na.omit(age_t_1*Y_{t_1}*L_{hat}*(1-L_{hat}))})

I_{\gamma 38} = \text{sum(na.omit(Sepsis_t_1*age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 48} = \text{sum(na.omit(respdis_t_1*age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 58} = \text{sum(na.omit(carddis_t_1*age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 68} = \text{sum(na.omit(diab_t_1*age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 78} = \text{sum(na.omit(sex_t_1*age_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 9} = \text{sum(na.omit((tme_t_1^2)*L_{hat}*(1-L_{hat}))})

I_{\gamma 19} = \text{sum(na.omit(tme_t_1*L_{hat}*(1-L_{hat}))})

I_{\gamma 29} = \text{sum(na.omit(tme_t_1*Y_{t_1}*L_{hat}*(1-L_{hat}))})

I_{\gamma 39} = \text{sum(na.omit(Sepsis_t_1*tme_t_1*L_{hat}*(1-L_{hat}))})
\[
I_{\gamma_{49}} = \text{sum(na.omit(respdis_t_1*tme_t_1*L_{hat}*(1-L_{hat}))}
\]
\[
I_{\gamma_{59}} = \text{sum(na.omit(carddis_t_1*tme_t_1*L_{hat}*(1-L_{hat}))}
\]
\[
I_{\gamma_{69}} = \text{sum(na.omit(diab_t_1*tme_t_1*L_{hat}*(1-L_{hat}))}
\]
\[
I_{\gamma_{79}} = \text{sum(na.omit(sex_t_1*tme_t_1*L_{hat}*(1-L_{hat}))}
\]
\[
I_{\gamma_{89}} = \text{sum(na.omit(age_t_1*tme_t_1*L_{hat}*(1-L_{hat}))}
\]
\[
I_{\gamma_{i}} = \text{matrix(c(I_{\gamma_{1}},I_{\gamma_{12}},I_{\gamma_{13}},I_{\gamma_{14}},}
\]
\[I_{\gamma_{15}},I_{\gamma_{16}},I_{\gamma_{17}},I_{\gamma_{18}},I_{\gamma_{19}},
\]
\[I_{\gamma_{12}},I_{\gamma_{2}},I_{\gamma_{23}},I_{\gamma_{24}},
\]
\[I_{\gamma_{25}},I_{\gamma_{26}},I_{\gamma_{27}},I_{\gamma_{28}},I_{\gamma_{29}},
\]
\[I_{\gamma_{13}},I_{\gamma_{23}},I_{\gamma_{3}},I_{\gamma_{34}},
\]
\[I_{\gamma_{35}},I_{\gamma_{36}},I_{\gamma_{37}},I_{\gamma_{38}},I_{\gamma_{39}},
\]
\[I_{\gamma_{14}},I_{\gamma_{24}},I_{\gamma_{34}},I_{\gamma_{4}},
\]
\[I_{\gamma_{25}},I_{\gamma_{46}},I_{\gamma_{47}},I_{\gamma_{48}},I_{\gamma_{49}},
\]
\[I_{\gamma_{15}},I_{\gamma_{25}},I_{\gamma_{35}},I_{\gamma_{45}},
\]
\[I_{\gamma_{5}},I_{\gamma_{56}},I_{\gamma_{57}},I_{\gamma_{58}},I_{\gamma_{59}},
\]
\[I_{\gamma_{16}},I_{\gamma_{26}},I_{\gamma_{36}},I_{\gamma_{46}},
\]
\[I_{\gamma_{66}},I_{\gamma_{67}},I_{\gamma_{68}},I_{\gamma_{69}},
\]
\[I_{\gamma_{17}},I_{\gamma_{27}},I_{\gamma_{37}},I_{\gamma_{47}},
\]
\[I_{\gamma_{57}},I_{\gamma_{67}},I_{\gamma_{7}},I_{\gamma_{78}},I_{\gamma_{79}},
\]
\[I_{\gamma_{18}},I_{\gamma_{28}},I_{\gamma_{38}},I_{\gamma_{48}},
\]
\[I_{\gamma_{58}},I_{\gamma_{68}},I_{\gamma_{78}},I_{\gamma_{8}},I_{\gamma_{89}},
\]
\[I_{\gamma_{19}},I_{\gamma_{29}},I_{\gamma_{39}},I_{\gamma_{49}},
\]
\[I_{\gamma_{59}},I_{\gamma_{69}},I_{\gamma_{79}},I_{\gamma_{89}},I_{\gamma_{9}})
\]
\[
v_{\hat{\gamma}} = v_{\hat{\gamma}} + \text{solve(I_{\gamma})}%*%S_{\gamma}
\]
\[
S_{\gamma\text{var}} = S_{\gamma\text{var}} + \text{solve(I_{\gamma})}%*%S_{\gamma}%*%t(S_{\gamma})%*%\text{solve(I_{\gamma})}
\]
\}
\]
\[
v_{\gamma\text{var}} = \text{solve(I_{\gamma})}%*%S_{\gamma\text{var}}%*%\text{solve(I_{\gamma})}
\]
\[ v_{se} = \sqrt{\left(v_{var\hat{}}\right) \cdot \text{diag}(1, \text{length}(v_{hat}))} \]

#--------Estimating Probabilities of M--------#

\[
lgt_{\lambda_{hat}} = c(\text{rep}(v_{hat}[1], (\text{length}(Y_{t_1})))) + v_{hat}[2] \cdot Y_{t_1} + v_{hat}[3] \cdot \text{Sepsis}_{t_1} + v_{hat}[4] \cdot \text{respd}_{t_1} + v_{hat}[5] \cdot \text{card}_{t_1} + v_{hat}[6] \cdot \text{diab}_{t_1} + v_{hat}[7] \cdot \text{sex}_{t_1} + v_{hat}[8] \cdot \text{age}_{t_1} + v_{hat}[9] \cdot \text{tme}_{t_1}
\]

\[
L_{hat} = \text{inv.logit}(lgt_{\lambda_{hat}})
\]

ProbM_{hat} = NULL
j=1
for (i in 1:K)
{
    prob_i=(\text{prod}(\{(1-L_{hat})[j:(j+0[i]-2)]\})^{((0[i]>1)\cdot 1)} \cdot \{L_{hat}[j+0[i]-1]\}^{((0[i]<T)\cdot 1)}
    ProbM_{hat}=\text{rbind}(ProbM_{hat},prob_i)
    j=(j+0[i])^{((0[i]<T)\cdot 1)} \cdot (j+0[i]-1)^{((0[i]==T)\cdot 1)}
}

#----------------Grace's Method-----------------#

\[
\tau_g = \text{gee_mon$coefficients[7:8]}
\]

\[
\tau_{norm} = (\tau_g/\text{vecnorm}(\tau_g)) \quad \text{#always length one}
\]

\[
\theta_g = c(\text{gee_mon$coefficients[1:6]}, \tau_{norm})
\]

\[
#\tau_{norm\_V}= (\theta_g[2:4]/\text{vecnorm}(\theta_g[2:4]))
\]

\[
#\eta_g = c(0,0)
\]

\[
\alpha_g=\text{gee_mon$geese$alpha}
\]

\[
\text{iter} = 15
\]
for (f in 1:iter)
{
    ETA_g = NULL
    tau_norm = theta_g[7:8]
    Ztau = Z%*%tau_norm
    qZ=quantile(Ztau,c(.01,.99))
    Ztau = replace(Ztau, ((Ztau<qZ[1])*1)*c(1:N), qZ[1])
    Ztau = replace(Ztau, ((Ztau>qZ[2])*1)*c(1:N), qZ[2])
    for (ss in 1:N) #For time varying Z’s
    {
        u = Ztau[ss]
        U = Ztau-u
        h = (K^(-1/5)) #bandwidth

        K_h= dnorm(U/h)/h
    
        eta_g = c(u,1)
    }

for (k in 1:25)
{
    #-------------------------eta estimation----------#

    lgt_g = ((theta_g[1])*rep(1,N))+((theta_g[2])*Sepsis) + ((theta_g[3])*respdis) + ((theta_g[4])*carddis)+ ((theta_g[5])*diab) + ((theta_g[6])*sex)+ c(rep(eta_g[1],N)) + eta_g[2]*U
    P_g=inv.logit(lgt_g)
    
    H = cbind(xi,U*xi)
\[
S_g = Y - P_g
\]

\[
\text{Score}_g = 0
\]

\[
I_g = 0
\]

\[
j = 1
\]

\[
\text{for (ie in 1:K)}
\]

\[
\{\]

\[
\text{CW}_i = \text{CW}[j:(j+O[ie]-1),] \]

\[
\text{H}_i = \text{H}[j:(j+O[ie]-1),] \]

\[
\text{S}_g_i = \text{S}_g[j:(j+O[ie]-1)] \]

\[
\{\text{if (O[ie]==1)}\]

\[
\text{I}_g = \text{I}_g + \text{CW}_i \% \% t(\text{H}_i) \]

\[
\text{else}\]

\[
\text{I}_g = \text{I}_g + t(\text{CW}_i) \% \% \text{H}_i \}

\[
\{\text{if (O[ie]==1)}\]

\[
\text{Score}_g = \text{Score}_g + \text{CW}_i \% \% t(\text{S}_g_i) \]

\[
\text{else}\]

\[
\text{Score}_g = \text{Score}_g + t(\text{CW}_i) \% \% \text{S}_g_i \}

\[
\}

\[
j = j+O[ie]
\]

\[
\}

\[
\eta_g = \eta_g + \text{solve}(I_g) \% \% \text{Score}_g 
\]

\[
\}

\[
\text{ETA}_g = \text{rbind(ETA}_g, t(\text{as.matrix(eta}_g))) \] #For time varying z’s

\[
\}

\[
\text{Eta}_g\_back1 = \text{ETA}_g
\]

\[
#------------------------theta estimation-----------------
\]

\[
\text{for (k in 1:iter)}
\]

\[
\{
\]

\[
\text{R}_\alpha = \text{matrix(rep(alpha}_g, T*T), ncol=T)
\]

\[
\text{diag(R}_\alpha) = 1
\]

\[
\text{del}_i = \text{rep(alpha}_g, 21)
\]
lgt = ((theta_g[1]) * rep(1, N)) + ((theta_g[2]) * Sepsis)  
+ ((theta_g[3]) * respdis) + ((theta_g[4]) * carddis) + ((theta_g[5]) * diab)  
+ ((theta_g[6]) * sex) + ETA_g[,1]  
+ ETA_g[,2] *(Z_t*theta_g[7:8]-Z_t*tau_norm)

P = inv.logit(lgt)

D = cbind((P*(1-P)), Sepsis*(P*(1-P)), respdis*(P*(1-P)),  
    carddis*(P*(1-P)), diab*(P*(1-P)), sex*(P*(1-P)),  
    age*ETA_g[,2]*(P*(1-P)), tme*ETA_g[,2]*(P*(1-P)))

S = Y - P

I_term = 0
S_term = 0
S_var = 0

I_alpha = 0
S_alpha = 0
S_alpha_var = 0

j = 1
for (i in 1:K)
{
    B_i = as.matrix(na.omit(diag(P[j:(j+O[i]-1)]  
        *(1-P[j:(j+O[i]-1)]]),O[i])))
    V_i = (B_i^-(1/2))
    %*%R_alpha[1:dim(B_i)[1],1:dim(B_i)[2]]%*%(B_i^-(1/2))*(ProbM_hat[i])
    D_i = as.matrix(na.omit(D[j:(j+O[i]-1)]))
    S_i = na.omit(S[j:(j+O[i]-1)])

    {if (O[i]==1)
        I_term = I_term + D_i%*%solve(V_i)%*%t(D_i)
    else
        I_term = I_term + t(D_i)%*%solve(V_i)%*%D_i}

    {if (O[i]==1)

S_term = S_term + D_i * solve(V_i) * t(S_i)
else
S_term = S_term + t(D_i) * solve(V_i) * S_i

{if (O[i] == 1)
S_var = S_var + D_i * solve(V_i) * t(S_i) * S_i * solve(V_i) * t(D_i)
else
S_var = S_var + t(D_i) * solve(V_i) * S_i * t(S_i) * solve(V_i) * D_i
}

#---alpha estimation part-------#

if (O[i] >= 2)
{
  z_i = NULL
  w_i = NULL
  for (l in 1:(O[i]-1))
  {
    r = j + l - 1
    for (s in 1:(O[i]-l))
    {
      z_ijl = ((Y[r] - P[r]) * (Y[r+s] - P[r+s])) / (P[r] * (1-P[r]) * P[r+s] * (1-P[r+s]))^(1/2)
      z_i = c(z_i, z_ijl)
      w_ijl = (1-2*P[r]) * (1-2*P[r+s]) * (P[r] * (1-P[r]) * P[r+s] * (1-P[r+s]))^(1/2)
      w_i = c(w_i, w_ijl)
    }
  }
  d = del_i[1:length(z_i)]
  Z_i = z_i - d
  vW_i = 1 + w_i * d - d^2
  W_i = diag(vW_i, length(z_i)) * (ProbM_hat[i])
  E_i = rep(1, length(Z_i))

  I_alpha = I_alpha + t(E_i) * solve(W_i) * E_i
  S_alpha = S_alpha + t(E_i) * solve(W_i) * Z_i
  S_alpha_var = S_alpha_var + t(E_i)
\%*%solve(W_i)%*%Z_i%*%t(Z_i)%*%solve(W_i)%*%E_i
}

j=j+0[i]
}

theta_g = theta_g + solve(I_term)%*%S_term
alpha_g = alpha_g + solve(I_alpha)%*%S_alpha

theta_var_g= solve(I_term)%*%S_var%*%solve(I_term)
alpha_var_g= solve(I_alpha)%*%S_alpha_var%*%solve(I_alpha)

}

}

print(theta_g)
print(alpha_g)
print(sqrt(theta_var_g))
print(sqrt(alpha_var_g))

plot(ETA_g[,1]~Ztau[,1],type="n")
lines(lowess(ETA_g[,1]~Ztau[,1],f = .001))