Identifiability and Estimation in Statistical Models with Nonignorable Missing Data

By

Jiwei Zhao

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2012

Date of final oral examination: 05/04/12

The dissertation is approved by the following members of the Final Oral Committee:

Jun Shao, Professor, Statistics
Kjell Doksum, Senior Research Scientist, Statistics
Kam-Wah Tsui, Professor, Statistics
Yazhen Wang, Professor, Statistics
Jun Zhu, Professor, Entomology
Acknowledgments

First and foremost, I would like to express my deepest gratitude and appreciation to my PhD advisor Professor Jun Shao. Throughout the past five years, Professor Shao has offered me invaluable technical advice as well as encouragement. I learn so much from him, not only on Statistics but also on personality. It is indeed a great honor to work closely with him during my graduate study.

My sincere thanks go to Professors Kjell Doksum, Kam-Wah Tsui and Yazhen Wang for serving on my preliminary exam and oral defense committees. I would like to thank them for their precious time and valuable suggestions on my thesis work and their generous support for my job application. I am also grateful to Professor Jun Zhu for serving on my oral defense committee, for her kind help and valuable time.

My gratitude is extended to all the faculty members, staff, and graduate students in the Department of Statistics.

Finally, very special thanks go to my wife, my parents, my parents-in-law and all of my other family members, for their endless love and support as always.
Abstract

The thesis is devoted to the identifiability and estimation in statistical models with nonignorable missing data. Missing data are common, especially in clinical trials, sample surveys, or longitudinal studies. When the missing mechanism is MCAR or MAR, mainly investigated in the literature, the model is identifiable, however, this is not generally true under MNAR, or, nonignorable missing, which is primarily studied in the current thesis. In real applications, actually there is a suspicion that the missing probability is indeed related to the values of the missing variable, i.e., the missing mechanism is nonignorable. In the thesis, we mainly investigate two different approaches with nonignorable missing data: pseudo likelihood approach and conditional likelihood approach. These two approaches are different from directly studying the observed likelihood of the model, and both of them have the robustness property to the missing mechanism.

The organization of the thesis is as follows. We first give an introduction of this topic in Chapter 1. In Chapter 2, we mainly study the identifiability and estimation in generalized linear models with nonignorable missing response using pseudo likelihood approach. Chapter 3 extends this idea to the model with both missing response and missing covariates. We derive the identifiability theorem and the asymptotic properties of the proposed estimators, and conduct simulations to compare their finite sample properties with or without missing covariates. In Chapter 4, we study the estimation in longitudinal studies with nonignorable
dropout. We develop a semiparametric pseudo likelihood method that produces consistent and asymptotically normal estimators under the assumption that there exists a dropout instrument, a covariate that is related to the response variable but not related to the dropout conditioned on the responses. Finally, in Chapter 5, we introduce conditional likelihood and approximate conditional likelihood in multivariate data with missing values. Similar to the above, we also derive the identifiability and asymptotic properties of the proposed estimators. The conditional likelihood approach is applicable under a more general missing mechanism assumption, but it would only identify fewer parameters than pseudo likelihood approach under some scenarios.
Contents

Acknowledgments  

Abstract  

1 Introduction  

2 Identifiability and Estimation with Nonignorable Nonresponse  
   2.1 Introduction  
   2.2 Generalized Linear Models  
   2.3 Missing Mechanisms  
   2.4 Identifiability in Model A  
   2.5 Identifiability in Model B  
   2.6 Asymptotic Theory  
   2.7 Proofs of Theorems  

3 Estimation with Nonignorable Missing Response and Covariates  
   3.1 Nonignorable Missing Covariates and the Estimation Procedure
3.2 Theoretical Results .......................................................... 33
3.3 Non-ignorable missing covariate and response ...................... 36
3.4 More than one missing covariates, and missing response ........... 38
3.5 Numerical Results .............................................................. 39
3.6 Proofs of Theorems ............................................................ 48

4 Estimation in Longitudinal Studies with Nonignorable Dropout 52
4.1 Introduction ................................................................. 52
4.2 Estimation Based on Pseudo Likelihoods ......................... 55
  4.2.1 The case where X is a dropout instrument .............. 55
  4.2.2 The case where a sub-vector of X is a dropout instrument . 58
4.3 Asymptotic Properties ..................................................... 59
4.4 Some Empirical Results .................................................. 62
  4.4.1 Estimation based on HIV-CD4 Data .................. 62
  4.4.2 A Simulation Study ............................................. 64
4.5 Discussion on Assumptions ............................................. 65
4.6 Proofs of Theorems ....................................................... 67

5 Approximate Conditional Likelihoods in Multivariate Missing Data Prob-
lems ................................................................. 77
5.1 Introduction ................................................................. 77
List of Tables

3.1 Simulation results: Y is binary, U is continuous, Y missing .......................... 40
3.2 Simulation results: Y is binary, U is discrete, Y missing ............................ 41
3.3 Simulation results: Y is binary, U is continuous, Y and U missing ................. 42
3.4 Simulation results: Y is binary, U is discrete, Y and U missing ....................... 43
3.5 Simulation results: Y is Poisson, U is continuous, Y missing .......................... 44
3.6 Simulation results: Y is Poisson, U is discrete, Y missing ............................ 45
3.7 Simulation results: Y is Poisson, U is continuous, Y and U missing ................. 46
3.8 Simulation results: Y is Poisson, U is discrete, Y and U missing ....................... 47
4.1 Estimates and SE’s under treatment 1 of the HIV-AIDS study ......................... 72
4.2 Estimates and SE’s under treatment 2 of the HIV-AIDS study ......................... 73
4.3 Estimates and SE’s under treatment 3 of the HIV-AIDS study ......................... 74
4.4 Estimates and SE’s under treatment 4 of the HIV-AIDS study ......................... 75
4.5 Simulation results under dropout mechanism \([9], n = 300\) ......................... 76
Chapter 1

Introduction

Consider a general statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta$ denotes the parameter space. Before one conducts statistical inference, a fundamental problem is whether the model, or the parametrization, is identifiable, that is, $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$. For example, when $P_\theta = N(\mu, \sigma^2)$, the i.i.d. normal family with unknown mean $\mu$ and variance $\sigma^2$, the model is identifiable. On the other hand, when $P_\theta = pN(\mu_1, \sigma_1^2) + (1 - p)N(\mu_2, \sigma_2^2)$, the i.i.d. normal mixture family with $(p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ all unknown, the model is not identifiable unless some constraints on the parameter space are imposed.

Missing data often arise in various experimental settings, including surveys, clinical trials, and longitudinal studies. Throughout the thesis, Let $[\cdot|\cdot]$ or $[\cdot]$ be a generic notation for the conditional or unconditional probability density. In problems with missing data, we usually denote the model as $[R, W]$, where the indicator $R = 1$ means the corresponding $W$ is observed, and $W = (Y, X)$, having $Y$ as the response and $X$ as its covariates. There are several approaches to formulate statistical models with missing data. In this thesis, we focus on selection model (Little and Rubin, 2002), which factorizes $[R, W] = [R|W][W]$, in which $[W]$ is the model for the data generating process, while $[R|W]$ is the one for the missing
mechanism. We assume that the two models are identifiable respectively.

In statistical models with missing data, we conduct inference mainly based on the observed likelihood (Little and Rubin, 2002). Although the models \([R|W]\) and \([W]\) are identifiable respectively, the observed likelihood may not be and we may not fully recover the whole information of \([R|W]\) and \([W]\). When the missing mechanism is missing at random (MAR, i.e., the missing probability is unrelated to the missing values of the variable itself) or missing completely at random (MCAR, i.e, the missing probability is a constant), the observed likelihood is always identifiable, therefore, there is no worry on this point. In the missing data literature, most work is based on MAR or MCAR assumption.

However, there is a suspicion that the missing probability is indeed related to the values of the missing variable. This happens, for example, in clinical studies evaluating a treatment, where the effects of the treatment may affect the participation of patients. When the missing probability is related to values of the missing variable, the mechanism is called missing not at random (MNAR), or, nonignorable (Little and Rubin 2002). Under the nonignorable missing mechanism, which is the focus of the thesis, the model identifiability becomes a serious and challenging problem and the observed likelihood may not be identifiable even in a most simple scenario (Wang, Shao and Kim, 2012).

In the literature, Greenless, Reece and Zieschang (1982) proposed a maximum likelihood method under parametric assumptions on both \([R|W]\) and \([W]\). But a fully parametric approach is sensitive to the model assumptions. Since the population is not identifiable when both \([R|W]\) and \([W]\) are nonparametric (Robins and Ritov, 1997), efforts have been made in situations where one of the two is parametric. Qin, Leung and Shao (2002) focused on the case where \([R|W]\) is parametric but \([W]\) is nonparametric. Wang, Shao and Kim (2012) studied the observed likelihood with nonignorable nonresponse, and gave some tech-
nical conditions under which the observed likelihood is identifiable. In real applications, the model for the data generating process is often statisticians’ main interest. Therefore, Tang, Little and Raghunathan (2003) considered the situation where $[W]$ is parametric but $[R|W]$ is nonparametric. They considered nonignorable nonresponse while the covariates are fully observed. They assumed that the missingness of the response only depends on the response, not on the covariates, i.e., $[R = 1|Y, X] = [R = 1|Y]$.

The main purpose of the thesis is to study model identifiability and propose estimation procedures with nonignorable missing data. In Tang et al. (2003), they focused on the model with continuous response variables, especially normal distributed. However, in practice, the response variable may be categorical in many cases. In Chapter 2, we consider the problem in generalized linear models (GLM, McCullagh and Nelder, 1989). There are mainly three contributions in this chapter. Firstly, we consider generalized linear models for the data generating process, which includes continuous outcome variable as a special case, and are frequently used when modeling categorical outcomes variables. Secondly, based on generalized linear models on the data, under the same missing mechanism as Tang et al. (2003), which assumes the missing probability can only depend on the values of the response, we give some conditions under which the whole model is identifiable. Thirdly, we propose a more practical missing mechanism, which assumes the missing probability can not only depend on the values of the response, but also on the values of some components of the covariates. We also study the identifiability theorem under this assumption. In Chapter 3, we extend to the model with both missing response and missing covariates. We compare the behaviors with or without missing covariates.

In Chapter 4, we study the estimation in longitudinal studies with nonignorable dropout. Without any further assumption, unknown parameters may not be identifiable when dropout
is nonignorable. We develop a semiparametric pseudo likelihood method that produces consistent and asymptotically normal estimators under the assumption that there exists a dropout instrument, a covariate that is related to the response variable but not related to the dropout conditioned on the responses. Although consistency and asymptotic normality for the proposed estimators can be established using a standard argument, their asymptotic covariance matrices are very complicated because the estimation at $t$ uses estimators from all time prior to $t$. Our main effort is to derive easy-to-compute consistent estimators of the asymptotic covariance matrices for assessing variability or inference.

In Chapter 5, we propose conditional likelihood and approximate conditional likelihood in multivariate data with missing values. This is the other different angle from studying the regular observed likelihood. We impose a parametric model on the data generating process $[W]$, while a nonparametric one on the missing mechanism $[R|W]$. Different from Tang, Little and Raghunathan (2003), our missing mechanism $[R|W]$ can depend on both response and covariates. We mainly consider two missing mechanisms: $[R = 1|Y, X] = s(Y)t(X)$ and $[R = 1|Y, U, Z] = s(Y, U)t(Z)$, where $s, t$ are two unknown functions. This approach allows us focusing on the inference of $[W]$ while giving a flexible and general assumption on $[R|W]$. Although this approach is originally designed for nonignorable missing data, it also applies for certain MAR missing mechanisms.
Chapter 2

Identifiability and Estimation with Nonignorable Nonresponse

2.1 Introduction

Missing data often arise in various experimental settings, including surveys, clinical trials, and longitudinal studies. When missing response is unrelated to the missing values of the response itself, it is called ignorable, or missing at random (MAR). Commonly, however, there is a suspicion that nonresponse is indeed related to the values of the missing response. This happens, for example, in clinical studies evaluating a treatment, where the effects of the treatment may affect the participation of patients. When nonresponse is related to values of the missing response, the nonresponse is called nonignorable (Little and Rubin 2002). When there is nonignorable nonresponse, serious biases in the estimates of the parameters may result if we do not model the distribution of the missing data, often called the missing mechanism. Little and Rubin (2002) has discussed this in detail and provided excellent examples illustrating this point. In this chapter, we focus on the case with nonignorable nonresponse.
The literature on estimation for nonignorable nonresponse is somewhat sparse. Greenlees, Reece, and Zieschang (1982) studied maximum likelihood estimators for survey data with nonignorable nonresponse, based on parametric models on both missing mechanism and the data generating process. However, a fully parametric approach is sensitive to the parametric model assumptions. On the other hand, it is almost impossible to develop a pure nonparametric method with nonignorable nonresponse because the population may not be identifiable (Robins and Ritov, 1997). Hence, in the literature efforts have been made in semi-parametric situations where one of the missing mechanism and data generating process is parametric and the other is nonparametric. For example, Qin, Leung and Shao (2002) considered a nonignorable nonresponse problem using empirical likelihood approach, where they assumed a parametric model on the missing mechanism and allowed nonparametric on the data. Wang, Shao and Kim (2012) studied the observed likelihood with nonignorable nonresponse, and gave some technical conditions under which the observed likelihood is identifiable. On the other side, Tang, Little and Raghunathan (2003) also considered a nonignorable nonresponse problem and they imposed a parametric model on the data, but allowed nonparametric on the missing mechanism. In real applications, it is more difficult to assume a suitable parametric model for the missing mechanism, additionally, people are more interested in the statistical inference on the data generating process.

In Tang et al. (2003), they focused on the model with continuous response variables, especially normal distributed. However, in practice, the response variable may be categorical in many cases. There are mainly three contributions in this chapter. Firstly, we consider generalized linear models for the data generating process, which includes continuous outcome variable as a special case, and are frequently used when modeling categorical outcomes variables. Secondly, based on generalized linear models on the data, under the same missing mechanism as Tang et al. (2003), which assumes the missing probability can only depend
on the values of the response, we give some conditions under which the whole model is identifiable. Thirdly, we propose a more practical missing mechanism, which assumes the missing probability can not only depend on the values of the response, but also on the values of some components of the covariates. We also study the identifiability theorem under this assumption.

The remaining of the chapter is organized as follows: in section 2, we give a brief review of the generalized linear models (GLM, McCullagh and Nelder, 1989). Section 3 mainly illustrates the missing mechanisms. In sections 4 and 5, we establish the results on model identifiability, which is crucial to construct valid estimators with nonignorable nonresponse. Based on the conditions when the model is identifiable, the asymptotic theory of the constructed estimators is developed in section 6. Some numerical results and the concluding remarks are provided afterwards. All the technical proofs are given in the appendix.

2.2 Generalized Linear Models

Suppose we have independent and identically distributed samples \( \{(r_i, y_i, x_i), i = 1, \ldots, N\} \) from the population \((R, Y, X)\), where \(N\) is the sample size. We assume the \(p\)-dimensional covariate \(X = (X_1, \ldots, X_p)^\tau\) is fully observed, while the scalar response \(Y\) is observed if and only if the missing indicator \(R = 1\). Also, without loss of generality, we assume \(r_i = 1, i = 1, \ldots, n\) and \(r_i = 0, i = n + 1, \ldots, N\), hence, \(n\) is the sample size of the fully observed subjects.

In this chapter, we assume the response variable \(Y\) follows the following p.d.f.

\[
p_Y(y; \eta, \phi) = \exp \left\{ \frac{y\eta - b(\eta)}{\phi} + c(y; \phi) \right\},
\]

w.r.t. some \(\sigma\)-finite measure \(\nu\), where \(b\) and \(c\) are some specific functions, \(\phi > 0\) denotes the dispersion parameter. For simplicity, we use \(\phi\) to denote the whole dispersion parameter,
instead of original function $a(\phi)$. Notice that if $\phi$ is known, this is an exponential family model with canonical parameter $\eta$. From the properties of GLM, we have

$$E(Y) = b'(\eta) \text{ and } \text{Var}(Y) = \phi b''(\eta),$$

where $b'$ and $b''$ are the first and second order derivatives of $b$, respectively. Define $\mu(\eta) = b'(\eta)$. It is assumed that $\eta$ is related to the covariate $x$ through the relationship:

$$g(\mu(\eta)) = \alpha + \beta^\top x,$$  \hspace{1cm} (2)

where $\beta = (\beta_1, \ldots, \beta_p)^\top$, and $g$, called a link function, is a known one-to-one, third-order continuously differentiable function. Define $\psi = (g \circ \mu)^{-1}$. If $\mu = g^{-1}$, then $g$ is called the canonical link function, denoted as $g_c$. If $g$ is not canonical, we assume that $\frac{d}{d\eta} \psi^{-1}(\eta) \neq 0$ for all $\eta$. Denote $\zeta = b \circ \psi$. Thus, the probability density function of response variable $Y$ given the covariate $X$, $p(y|x; \theta)$, can be expressed as

$$p(y|x; \theta) = \exp\left\{ \frac{y \psi(\alpha + \beta^\top x) - \zeta(\alpha + \beta^\top x)}{\phi} + c(y; \phi) \right\}, \hspace{1cm} (3)$$

with respect to the measure $\nu$, where $\theta = (\alpha, \beta^\top, \phi)^\top$. The parameter space $\Theta \subset R \otimes R^p \otimes (0, \infty)$. Denote the true value of $\theta$ as $\theta_0 = (\alpha_0, \beta_0^\top, \phi_0)$, where $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^\top$ denotes the true value of $\beta$. Suppose the parameter space $\Theta$ is compact, and the true value $\theta_0 \in \Theta^0$, the interior of $\Theta$. Hence, generally $\psi = g_c \circ g^{-1}$, $\zeta = b \circ g_c \circ g^{-1}$,

$$E(Y|X) = b'(\psi(\alpha + \beta^\top x)) \text{ and } \text{Var}(Y|X) = \phi b''(\psi(\alpha + \beta^\top x)).$$

Generalized linear models (GLM) are widely used in practice. Besides the linear regression with normal errors, GLM also includes logistic regression, Poisson regression, etc., which are frequently used when modeling categorical data. In practice, canonical link is more preferred and the function $\psi(t) = t$ when the canonical link case, hence $p(y|x; \theta)$ can
be written as
\[
p(y|x; \theta) = \exp \left\{ \frac{y(\alpha + \beta^\tau x) - b(\alpha + \beta^\tau x)}{\phi} + c(y; \phi) \right\}.
\] (4)

However, in GLM, canonical links do not always provide the best fit. Generally, there is no reason apriori why a canonical link should be used, and in many cases a noncanonical link is more preferable, see McCullagh and Nelder (1989) and Czado and Munk (2000). In this chapter, we focus on both canonical and noncanonical link functions (Wedderburn 1976).

We firstly list several frequently used GLMs below (McCullagh and Nelder, 1989).

Example 1 (Normal).

\[
b(\eta) = \eta^2/2, b'(\eta) = \eta, b''(\eta) = 1,
\]
\[
p(y|x; \theta) = \exp \left\{ \frac{y\psi(\alpha + \beta^\tau x) - \frac{1}{2}(\psi(\alpha + \beta^\tau x))^2}{\phi} - \frac{1}{2} \left( \frac{y^2}{\phi} + \log(2\pi\phi) \right) \right\}, y \in R,
\]

where
\[
g_c(t) = t, \psi(t) = g(t)^{-1}, E(Y|X) = \psi(\alpha + \beta^\tau x), \text{ and } Var(Y|X) = \phi.
\]

Example 2 (Poisson).

\[
b(\eta) = \exp\{\eta\}, b'(\eta) = \exp\{\eta\}, b''(\eta) = \exp\{\eta\},
\]
\[
p(y|x; \theta) = \exp \left\{ y\psi(\alpha + \beta^\tau x) - \exp(\psi(\alpha + \beta^\tau x)) - \log y! \right\}, y = 0, 1, \cdots,
\]

where
\[
g_c(t) = \log(t), \psi(t) = \log(g(t)^{-1}), E(Y|X) = Var(Y|X) = \exp\{\psi(\alpha + \beta^\tau x)\}.
\]

Notice that dispersion parameter \(\phi = 1\). The noncanonical link \(g(t) = t^\gamma, 0 < \gamma < 1\) leads to \(\psi(t) = \frac{1}{\gamma} \log(t)\). The link function \(g(t) = 2\sqrt{t}\), making \([\psi'(t)]^2 b''(\psi(t)) = 1\), leads to \(\psi(t) = 2 \log(t/2)\) (Shao, 2003, pp.283).
Example 3 (Binomial).

\[ b(\eta) = \log(1 + e^\eta), \quad b'(\eta) = \frac{e^\eta}{1 + e^\eta}, \quad b''(\eta) = \frac{e^\eta}{(1 + e^\eta)^2}, \]
\[ p(y|x; \theta) = \exp \left\{ \frac{y \psi(\alpha + \beta^\tau x) - \log(1 + \exp\{\psi(\alpha + \beta^\tau x)\})}{1/m} + \log \left( \frac{m}{my} \right) \right\}, \quad y = \frac{0, 1, \ldots, m}{m}, \]

where
\[ g_c(t) = \log \frac{t}{1 - t}, \quad \psi(t) = \log \frac{g(t)^{-1}}{1 - g(t)^{-1}}, E(Y|X) = \frac{\exp\{\psi(\alpha + \beta^\tau x)\}}{1 + \exp\{\psi(\alpha + \beta^\tau x)\}}, \]
\[ \text{Var}(Y|X) = \frac{\exp\{\psi(\alpha + \beta^\tau x)\}}{m(1 + \exp\{\psi(\alpha + \beta^\tau x)\})^2}. \]

Notice that when \( m = 1 \), the dispersion parameter \( \phi = 1 \). The noncanonical links

\[ g(t) = t, \quad \text{arcsin} \sqrt{t}, \quad \log(-\log(1 - t)), \quad \Phi^{-1}(t) \]

leads to

\[ \psi(t) = \log \frac{t}{1 - t}, \quad 2 \log|\tan(t)|, \quad \log(\exp(e^t) - 1), \quad \log \frac{\Phi(t)}{1 - \Phi(t)}, \]

respectively.

Example 4 (Gamma).

\[ b(\eta) = -\log(-\eta), \quad b'(\eta) = -\frac{1}{\eta}, \quad b''(\eta) = \frac{1}{\eta^2}, \quad \eta < 0, \]
\[ p(y|x; \theta) = \exp \left\{ \frac{y \psi(\alpha + \beta^\tau x) + \log(-\psi(\alpha + \beta^\tau x))}{1/\nu} + \nu \log(\nu y) - \log y - \log \Gamma(\nu) \right\}, \quad y > 0, \]

where
\[ g_c(t) = -\frac{1}{t}, \quad \psi(t) = -\frac{1}{g(t)^{-1}}, \quad E(Y|X) = -\frac{1}{\psi(\alpha + \beta^\tau x)}, \quad \text{Var}(Y|X) = \frac{1}{\nu(\psi(\alpha + \beta^\tau x))^2}. \]

The noncanonical links \( g(t) = \log(t), \quad t^\gamma, \quad -1 \leq \gamma < 0 \) leads to \( \psi(t) = -e^{-t}, \quad -t^{-\frac{1}{\gamma}} \).

Example 5 (Inverse Gaussian).

\[ b(\eta) = -(2\eta)^{1/2}, \quad b'(\eta) = -(2\eta)^{-1/2}, \quad b''(\eta) = (2\eta)^{-3/2}, \quad \eta < 0, \]
\[ p(y|x; \theta) = \exp \left\{ \frac{y \psi(\alpha + \beta^\tau x) + (2\psi(\alpha + \beta^\tau x))^{1/2}}{\phi} - \frac{1}{2} \left( \frac{1}{\phi y} + \log(2\pi\phi y^3) \right) \right\}, \quad y > 0, \]
where

\[ g_c(t) = -\frac{1}{2t^2}, \psi(t) = -\frac{1}{2(g(t)^{-1})^2}, E(Y|X) = (-2\psi(\alpha + \beta^* x))^{-1/2}, \]

\[ \text{Var}(Y|X) = \phi(-2\psi(\alpha + \beta^* x))^{-3/2}. \]

### 2.3 Missing Mechanisms

Throughout this chapter, let \([\cdot|\cdot]\) or \([\cdot]\) be a generic notation for the conditional or unconditional probability density. For the missing indicator \(R\), we assume

\[ [R = 1|Y, X] = [R = 1|Y], \]

which means, the missing probability of \(Y\) only depends on \(Y\) itself, but not on its covariate \(X\). We define (11), (4) under the missing mechanism (5) as model A.

Under (5), \([X|Y, R = 1] = [X|Y]\), hence, the completely observed subjects are a random sample from \([X|Y]\) (Tang et al., 2003), and a natural approach to estimate the parameter of interest \(\theta\) is to base inference on the likelihood of \([X|Y]\) based on the completely observed subjects, i.e., by maximizing the following likelihood function

\[
L_A(\theta) = \prod_{r_i=1} p(x_i|y_i; \theta) \propto \prod_{i=1}^n \frac{p(y_i|x_i; \theta)}{\sum_{j=1}^N p(y_i|x_j; \theta)},
\]

or the log-likelihood,

\[
l_A(\theta) = \sum_{i=1}^N r_i \log p(y_i|x_i; \theta) - \sum_{i=1}^N r_i \log \left\{ \sum_{j=1}^N p(y_i|x_j; \theta) \right\}. \tag{6}
\]

However, model A has a strong assumption that, given \(Y\), the missing indicator \(R\) is conditionally independent with the whole covariate \(X\). In practice, \(R\) may have some correlation with \(X\), or, at least, some components of \(X\). Motivated by this fact, we decompose
as \( p_1 \)-dimensional covariate \( U \) and \( p_2 \)-dimensional covariate \( Z \), where \( p_1 + p_2 = p \), and we assume the more general and more practical missing mechanism

\[
[R = 1|Y, U, Z] = [R = 1|Y, U].
\]  

(7)

Notice that, under (7), the distribution of missing indicator can depend on response \( Y \) and part, not whole, of covariate \( X \). For the data generating process, we still focus on the following GLM:

\[
p(y|u, z; \theta) = \exp \left\{ \frac{y\psi(\alpha + \beta_1^T u + \beta_2^T z)}{\phi} - \frac{\zeta(\alpha + \beta_1^T u + \beta_2^T z)}{\phi} + c(y; \phi) \right\},
\]

(8)

and its version with canonical link:

\[
p(y|u, z; \theta) = \exp \left\{ \frac{y(\alpha + \beta_1^T u + \beta_2^T z)}{\phi} - \frac{b(\alpha + \beta_1^T u + \beta_2^T z)}{\phi} + c(y; \phi) \right\},
\]

(9)

where \( \beta_1 \) and \( \beta_2 \) are \( p_1 \) and \( p_2 \) dimensional parameters, respectively. We call (8), (9) under the missing mechanism assumption (7) as model B.

Apparently, model B is a generalization of model A. If, specially, \( U \) is a constant scalar, then model B will shrink to model A. Under model B, we can impose a parametric model \( p(u|z; \gamma) \) on \( U \) and \( Z \), where \( \gamma \) is a \( q \)-dimensional parameter. Since \( U \) and \( Z \) are both fully observed, we can firstly obtain an efficient estimator of \( \gamma, \hat{\gamma} \). Then, based on the idea that \([Z|Y, U, R = 1] = [Z|Y, U]\), we can maximize the following pseudo-log-likelihood to estimate \( \theta \):

\[
l_B(\theta) = \sum_{i=1}^{N} r_i \log p(y_i|u_i, z_i; \theta) - \sum_{i=1}^{N} r_i \log \left\{ \sum_{j=1}^{N} p(y_i|u_i, z_j; \theta)p(u_i|z_j; \hat{\gamma}) \right\}.
\]

(10)

Notice that in this estimation procedure, the data \( \{u_i, i = n + 1, \ldots, N\} \) are only utilized to obtain \( \hat{\gamma} \).
2.4 Identifiability in Model A

Tang et al. (2003) studied the missing mechanism (5) and focused on the identifiability when the response is distributed as normal. In this section, we provide some conditions under which the parameter of interest $\theta$ is identifiable based on Model A by the estimation procedure discussed in last section. We consider two scenarios: the dispersion parameter $\phi$ is one component of the unknown parameter, or when $\phi$ is known. The latter case is interesting, especially in binary regression or Poisson regression. The following theorem illustrates that, under some conditions of the support space of the covariate $X$ and the function $b$ defined in the GLM, the parameter $\theta$ is identifiable.

Theorem 1. For GLM with canonical link (4), define the space of the support values of $X_i$ as $X_i, i = 1, 2, \ldots, p$, and the space of the support values of $X = (X_1, \ldots, X_p)^\tau$ as $X = X_1 \otimes \ldots \otimes X_p$.

(i) Under the following condition (C1), the parameter $\theta = (\alpha, \beta^\tau, \phi)^\tau$ is fully identifiable;

(ii) If $\phi > 0$ is known, under the following condition (C2), the parameters $\alpha$ and $\beta$ are identifiable.

(C1). There exists $X_{01} = \{x_0, x_1, \ldots, x_q\} \subset X$, where $q$ depends on $p$, such that $2q \geq p + 2$, and the following $(2q) \times (p + 2)$ matrix is of full rank when evaluated at the true values
\[ \alpha = \alpha_0, \beta = \beta_0: \]
\[
\begin{pmatrix}
0 & x_1^r - x_0^r & \beta^r(x_1 - x_0) \\
\vdots & \vdots & \vdots \\
0 & x_q^r - x_0^r & \beta^r(x_q - x_0)
\end{pmatrix};
\]
\[
b'(\alpha + \beta^r x_1) - b'(\alpha + \beta^r x_0) \quad b'(\alpha + \beta^r x_1) x_1^r - b'(\alpha + \beta^r x_0) x_0^r \quad b(\alpha + \beta^r x_1) - b(\alpha + \beta^r x_0) \\
\vdots & \vdots & \vdots \\
b'(\alpha + \beta^r x_q) - b'(\alpha + \beta^r x_0) \quad b'(\alpha + \beta^r x_q) x_q^r - b'(\alpha + \beta^r x_0) x_0^r \quad b(\alpha + \beta^r x_q) - b(\alpha + \beta^r x_0)
\]

(C2). There exists \( X_{02} = \{x_0, x_1, \ldots, x_q\} \subset X \), where \( q \) depends on \( p \), such that \( 2q \geq p + 1 \), and the following \( (2q) \times (p + 1) \) matrix is of full rank when evaluated at the true values \( \alpha = \alpha_0, \beta = \beta_0: \)
\[
\begin{pmatrix}
0 & x_1^r - x_0^r \\
\vdots & \vdots \\
0 & x_q^r - x_0^r \\
b'(\alpha + \beta^r x_1) - b'(\alpha + \beta^r x_0) \quad b'(\alpha + \beta^r x_1) x_1^r - b'(\alpha + \beta^r x_0) x_0^r \\
\vdots & \vdots \\
b'(\alpha + \beta^r x_q) - b'(\alpha + \beta^r x_0) \quad b'(\alpha + \beta^r x_q) x_q^r - b'(\alpha + \beta^r x_0) x_0^r
\end{pmatrix}.
\]

Remark 1. The condition \( b'' > 0 \) is assumed as an underlying assumption in the GLM definition. Also, the true value \( \beta_0 \neq 0 \) is a necessary requirement in condition (C1) and (C2).

Remark 2. In Theorem 1 (i) with \( p = 1 \), we need at least 3 different support values to identify the parameter of interest \( \theta \); in Theorem 1 (ii) with \( p = 1 \), the parameters \( \alpha \) and \( \beta \) are identifiable if and only if \( X \) contains at least two different support values.
Corollary 1. In Theorem 1 (i) with \( p = 1 \), if \( \mathcal{X} \) only contains two support values, then the parameter \( \theta \) can not be fully identifiable no matter what the function \( b \) is.

Example 6 (Normal Continued). Consider normal distribution, discussed in example 1, with canonical link and \( p = 1 \). When the dispersion parameter is unknown, since \( b(t) = t^2 / 2 \), and \( b'(t) = t \), then all parameters are identifiable if \( X \) has at least 3 support points. This result is consistent with Tang et al. (2003). Especially, if the dispersion parameter is known, then all parameters are identifiable if \( X \) has at least 2 support points.

Example 7 (Poisson continued). Consider Poisson distribution, discussed in example 2 with canonical link and \( p = 1 \). From Theorem 1 (ii), the parameters \( \alpha \) and \( \beta \) are both identifiable if \( X \) has at least 2 support points.

Example 8 (Binomial Continued). Consider binomial distribution, discussed in example 3, with canonical link, \( m = 1 \), \( p = 1 \). From Theorem 1 (ii), the parameters \( \alpha \) and \( \beta \) are both identifiable if \( X \) has at least 2 support points.

Example 9 (Gamma continued). Consider Gamma distribution, discussed in example 4, with canonical link and \( p = 1 \). It can be derived that the parameter \( \theta \) is identifiable if \( \mathcal{X} \) contains three different support values \( \{x_0, x_1, x_2\} \) and

\[
(t_0^o)^{c-1}(t_0 + c\xi_1)^{t_0 + c\xi_1} \neq [(t_0 + \xi_1)^{t_0 + \xi_1}]^c,
\]

where \( t_0 = \alpha_0 + \beta_0 x_0 \), \( \xi_1 = \beta_0 (x_1 - x_0) \), \( c = \frac{x_2 - x_0}{x_1 - x_0} \), \( t_0 < 0 \), \( t_0 + \xi_1 < 0 \), \( t_0 + c \xi_1 < 0 \).

The next result is based on model A with noncanonical link \([11, 5]\).

Theorem 2. For GLM with noncanonical link \([11]\), define the space of the support values of \( X_i \) as \( \mathcal{X}_i \), \( i = 1, 2, \ldots, p \), and the space of the support values of \( X = (X_1, \ldots, X_p)^\tau \) as \( \mathcal{X} = \mathcal{X}_1 \otimes \ldots \otimes \mathcal{X}_p \).
(i) Under the following condition $(C1)$, the parameter $\theta = (\alpha, \beta^\tau, \phi)^\tau$ is fully identifiable;

(ii) If $\phi > 0$ is known, under the following condition $(C2)$, the parameters $\alpha$ and $\beta$ are identifiable.

$(C1)$. There exists $X_{01} = \{x_0, x_1, \ldots, x_q\} \subset \mathcal{X}$, where $q$ depends on $p$, such that $2q \geq p + 2$, and the following $(2q) \times (p + 2)$ matrix is of full rank when evaluated at the true values $\alpha = \alpha_0, \beta = \beta_0$:

$$
\begin{pmatrix}
\psi'(\alpha + \beta^\tau x_1) - \psi'(\alpha + \beta^\tau x_0) & \psi'(\alpha + \beta^\tau x_1)x_1^\tau - \psi'(\alpha + \beta^\tau x_0)x_0^\tau \\
\vdots & \vdots \\
\psi'(\alpha + \beta^\tau x_q) - \psi'(\alpha + \beta^\tau x_0) & \psi'(\alpha + \beta^\tau x_q)x_q^\tau - \psi'(\alpha + \beta^\tau x_0)x_0^\tau \\
\frac{\partial}{\partial x_1} \zeta'\left((\alpha + \beta^\tau x_1) - \zeta'(\alpha + \beta^\tau x_0)\right) & \frac{\partial}{\partial x_1} \zeta'(\alpha + \beta^\tau x_1)x_1^\tau - \zeta'(\alpha + \beta^\tau x_0)x_0^\tau \\
\vdots & \vdots \\
\frac{\partial}{\partial x_q} \zeta'\left((\alpha + \beta^\tau x_q) - \zeta'(\alpha + \beta^\tau x_0)\right) & \frac{\partial}{\partial x_q} \zeta'(\alpha + \beta^\tau x_q)x_q^\tau - \zeta'(\alpha + \beta^\tau x_0)x_0^\tau
\end{pmatrix}.
$$

$(C2)$. There exists $X_{02} = \{x_0, x_1, \ldots, x_q\} \subset \mathcal{X}$, where $q$ depends on $p$, such that $2q \geq p + 1$, and the following $(2q) \times (p + 1)$ matrix is of full rank when evaluated at the true values $\alpha = \alpha_0, \beta = \beta_0$:

$$
\begin{pmatrix}
\psi'(\alpha + \beta^\tau x_1) - \psi'(\alpha + \beta^\tau x_0) & \psi'(\alpha + \beta^\tau x_1)x_1^\tau - \psi'(\alpha + \beta^\tau x_0)x_0^\tau \\
\vdots & \vdots \\
\psi'(\alpha + \beta^\tau x_q) - \psi'(\alpha + \beta^\tau x_0) & \psi'(\alpha + \beta^\tau x_q)x_q^\tau - \psi'(\alpha + \beta^\tau x_0)x_0^\tau \\
\frac{\partial}{\partial x_1} \zeta'\left((\alpha + \beta^\tau x_1) - \zeta'(\alpha + \beta^\tau x_0)\right) & \frac{\partial}{\partial x_1} \zeta'(\alpha + \beta^\tau x_1)x_1^\tau - \zeta'(\alpha + \beta^\tau x_0)x_0^\tau \\
\vdots & \vdots \\
\frac{\partial}{\partial x_q} \zeta'\left((\alpha + \beta^\tau x_q) - \zeta'(\alpha + \beta^\tau x_0)\right) & \frac{\partial}{\partial x_q} \zeta'(\alpha + \beta^\tau x_q)x_q^\tau - \zeta'(\alpha + \beta^\tau x_0)x_0^\tau
\end{pmatrix}.
$$

**Corollary 2.** Theorem 1 is a special case of Theorem 2 with $\psi(t) = t$. Consider Theorem 2 (i) with $p = 1$. If $\mathcal{X}$ only contains two support values $\{x_0, x_1\}$, then not all parameters...
α, β, φ are identifiable no matter what functions ψ and b are. Consider Theorem 2 (ii) with p = 1. When X only contains two support values \{x_0, x_1\}, if \(\psi(\alpha_0 + \beta_0 x_0) \neq \psi(\alpha_0 + \beta_0 x_1)\), since \(b'\) is strictly increasing, then \(\alpha\) and \(\beta\) are identifiable.

2.5 Identifiability in Model B

In this section, we study the identifiability theorem with missing mechanism \([R = 1|Y, U, Z] = [R = 1|Y, U]\), where U and Z are two components of covariate X. We first consider the general regression model \(p(y|u, z; \theta)\). Similar to Tang et al. (2003), we have the following results:

**Theorem 3.** Suppose \(\gamma_0\) is the true regression parameter of \([U|Z]\) and \(F_0(z)\) is the true cumulative distribution function of Z. For any \(\theta \in \Theta, \theta \neq \theta_0\) and any arbitrary function \(c(y, u)\), define

\[
D_\theta = \{(y, u) : p(y|u, z; \theta) = c(y, u)p(y|u, z; \theta_0) \text{ for any } z\}.
\]

If \(P(R = 1) > 0\) and \(P(D_\theta) < 1\) for any \(\theta \neq \theta_0\), then

\[
E_0[1_{(R=1)} \log p(z|y, u; \tilde{\theta}_1, \gamma_0, F_0)] < E_0[1_{(R=1)} \log p(z|y, u; \tilde{\theta}_{10}, \gamma_0, F_0)], \tilde{\theta}_1 \neq \tilde{\theta}_{10},
\]

where \(E_0\) denotes expectation with respect to the true value \(\theta_0\), \(\tilde{\theta}_1\) denotes the parameter identified from \([Z|Y, U]\), and \(\tilde{\theta}_{10}\) denotes its true value.

**Corollary 3.** Suppose \(\theta\) can be reparameterized as \(\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)\), with true value \(\tilde{\theta}_0 = (\tilde{\theta}_{10}, \tilde{\theta}_{20})\), and the density \(p(y|u, z; \theta)\) can be written as

\[
p(y|u, z; \theta) = \exp\{h_1(y, u, z; \tilde{\theta}_1) + h_2(y, u; \tilde{\theta})\},
\]

where \(h_1\) is known up to \(\tilde{\theta}_1\), and \(h_2\) may not be specified. Also, assume that, for any \(k \neq \tilde{\theta}_{10}\) in the domain of \(\tilde{\theta}_1\), the function \(h_1(y, u, z; k) - h_1(y, u, z; \tilde{\theta}_{10})\) is not a function of \(y, u\) alone. Then \(\tilde{\theta}_1\) is identifiable.
In the following, we focus on the GLM with covariates $U$ and $Z$:

$$p(y|u,z;\theta) = \exp\left\{ \frac{y\psi(\alpha + \beta_u^T u + \beta_z^T z)}{\phi} - \frac{b\psi(\alpha + \beta_u^T u + \beta_z^T z)}{\phi} + c(y;\phi) \right\}, \quad (11)$$

where $\theta = (\alpha, \beta_u^T, \beta_z^T, \phi)^T$. Suppose $\beta_u$ is $p_u$-dimensional and $\beta_z$ is $p_z$-dimensional. Denote $b \circ \psi = \zeta$.

**Theorem 4.** For GLM with covariates $U$ and $Z$, define the space of the support values of $U_i$ as $U_i$, $i = 1, 2, \ldots, p_u$, and the space of the support values of $U$ as $U = U_1 \otimes \cdots \otimes U_{p_u}$. Similarly define the space of the support values of $Z$ as $Z = Z_1 \otimes \cdots \otimes Z_{p_z}$.

- Under the following condition (C1), the parameter $\theta = (\alpha, \beta_u^T, \beta_z^T, \phi)^T$ is fully identifiable;

- If $\phi > 0$ is known, under the following condition (C2), the parameters $\alpha$, $\beta_u$ and $\beta_z$ are identifiable.

(C1). There exists $U_{01} = \{u_0, u_1, \ldots, u_q\} \subset U$, and $Z_{01} = \{z_0, z_1, \ldots, z_r\} \subset Z$, such that $2r(q + 1) \geq p_u + p_z + 2$, and there exists $p_u + p_z + 2$ functions from the following $2r(q + 1)$ functions:

$$\frac{\psi(\alpha + \beta_u^T u_j + \beta_z^T z_i)}{\phi} - \frac{\psi(\alpha + \beta_u^T u_j + \beta_z^T z_0)}{\phi},$$

$$\frac{\zeta(\alpha + \beta_u^T u_j + \beta_z^T z_i)}{\phi} - \frac{\zeta(\alpha + \beta_u^T u_j + \beta_z^T z_0)}{\phi},$$

$i = 1, 2, \ldots, r; j = 0, 1, \ldots, q$,

such that the $p_u + p_z + 2$ functions are independent at the true value $\theta_0 = (\alpha_0, \beta_{u0}^T, \beta_{z0}^T, \phi)^T$;

(C2). There exists $U_{02} = \{u_0, u_1, \ldots, u_q\} \subset U$, and $Z_{02} = \{z_0, z_1, \ldots, z_r\} \subset Z$, such that $2r(q + 1) \geq p_u + p_z + 1$, and there exists $p_u + p_z + 1$ functions from the following $2r(q + 1)$
functions:

$$\psi(\alpha + \beta_u u_j + \beta_z z_i) - \psi(\alpha + \beta_u u_j + \beta_z z_0),$$

$$\zeta(\alpha + \beta_u u_j + \beta_z z_i) - \zeta(\alpha + \beta_u u_j + \beta_z z_0),$$

$$i = 1, 2, \ldots, r; j = 0, 1, \ldots, q,$$

such that the $p_u + p_z + 1$ functions are independent at the true values $\alpha_0, \beta_{u0}, \beta_{z0}$.

For simplicity, we consider a special case of GLM with canonical link with both $U$ and $Z$ are 1-dimensional, also, $Z$ only has two different support values, i.e., $r = 1$.

**Corollary 4.** For Model B with canonical link, we have

$$p(y|u, z; \theta) = \exp \left\{ \frac{y(\alpha + \beta_u u + \beta_z z) - b(\alpha + \beta_u u + \beta_z z)}{\phi} + c(y; \phi) \right\},$$

and the following results:

(i) If $q \geq 2$ ($U$ is continuous or discrete with at least 3 support values) and $d''(x) \neq 0$, then all parameters $\alpha, \beta_u, \beta_z, \phi$ are identifiable, where, for any fixed $t \neq 0$,

$$d(x) = \frac{b(x + t) - b(x) - tb'(x + t)}{b'(x + t) - b'(x)}.$$

(ii) If $q = 1$, then no matter what the function $b$ is, the parameters are not fully identifiable.

(iii) If the dispersion parameter $\phi > 0$ is known. If $q = 1$, then the parameters $\alpha, \beta_u, \beta_z$ are identifiable.

**Remark 3.** The whole solution of the function $b$ satisfying $d'' = 0$ is not fully obtained. Two known solutions are $b(t) = t^2/2$ and $b(t) = e^t$.

**Example 10** (Normal Continued). Consider the linear model where $b(x) = \frac{x^2}{2}$, then $b''(x) = 1$, but $d''(x) = 0$, hence the parameters are not all identifiable when $Z$ only has two support
points. Previous results show that, if there is no $U$, to identify all parameters, we also need the covariate to have at least three support points. Therefore, for linear model, the covariate $U$ does not help to identify the parameters.

**Example 11** (Binomial Continued). Consider the binomial distribution ($m > 1$) with canonical link where $b(x) = \log(1 + \exp(x))$, then $b''(x) \neq 0$, and $d''(x) \neq 0$, hence, the parameters are all identifiable when $Z$ only have two support points. Previous results show that, if there is no $U$, to identify all parameters, we need the covariate to have at least three support points. Therefore, for logistic regression with canonical link (with unknown dispersion parameter), the covariate $U$ does help to identify all parameters.

**Example 12** (Poisson Continued). Consider Poisson distribution with no dispersion parameter. If $Z$ only has two support values, the parameters are all identifiable, with or without the emergence of $U$. The same conclusion as Normal distribution. The instrument $U$ does not help to identify parameters.

**Example 13** (Gamma continued). Consider Gamma distribution with canonical link where $b(x) = -\log(-x), x < 0$, hence, $b'(x) = -x^{-1}$ and $b''(x) = x^{-2}$. Therefore, $d''(x) \neq 0$, the parameters are all identifiable when $Z$ only has two support points. Previous results show that, if there is no $U$, to identify all parameters, we need the covariate to have at least three support points. Therefore, for Gamma regression with canonical link, the covariate $U$ does help to identify all parameters.

**Remark 4.** If $U$ is discrete, we could consider the following more flexible model:

$$p(y|u, z; \theta) = \exp \left\{ \frac{y(\alpha_0 u + \beta_1 u z) - b(\alpha_0 u + \beta_1 u z)}{\phi} + c(y; \phi) \right\},$$

then for any fixed $U = u$, the model becomes Model A, hence if $Z$ has at least three support points, we can identify all parameters in this more flexible model.
2.6 Asymptotic Theory

In this section, we show that the proposed estimator is consistent and asymptotically normal as the sample size $N \to \infty$. For simplicity, we only prove the situation in model B, i.e., the maximizer of the following pseudo-log-likelihood:

$$l_B(\theta) = \sum_{i=1}^{N} r_i \log p(y_i|u_i, z_i; \theta) - \sum_{i=1}^{N} r_i \log \left\{ \sum_{j=1}^{N} p(y_i|u_i, z_j; \theta) p(u_i|z_j; \hat{\gamma}) \right\}.$$

For convenience, denote $F(z)$ as the distribution of covariate $Z$ and $F_0(z)$ as its truth. Denote

$$H(\theta, \gamma, F) = \frac{1}{N} \sum_{i=1}^{N} H_i(\theta, \gamma, F) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \log p(y_i|u_i, z_i; \theta) - \log \left\{ \int p(y_i|u_i, z; \theta) p(u_i|z; \gamma) dF(z) \right\} \right\},$$

and $H_i(\theta, \gamma, F)$ be the one based on the $i$th subject, i.e.,

$$H_i(\theta, \gamma, F) = r_i \left\{ \log p(y_i|u_i, z_i; \theta) - \log \left\{ \int p(y_i|u_i, z; \theta) p(u_i|z; \gamma) dF(z) \right\} \right\}.$$

It is clear that to maximize $l_B(\theta)$ is the same as to maximize

$$l(\theta, \hat{\gamma}, \hat{F}) = \frac{1}{N} \sum_{i=1}^{N} H_i(\theta, \hat{\gamma}, \hat{F}),$$

where $\hat{\gamma}$ denotes an estimate of nuisance parameter $\gamma$ and $\hat{F}$ denotes the empirical distribution of $F$. Assume that under some regularity conditions, $\hat{\gamma}$ is consistent, i.e., $\hat{\gamma} \xrightarrow{p} \gamma_0$, as $N \to \infty$, and $\hat{\gamma}$ is asymptotically normal, i.e.,

$$\sqrt{N}(\hat{\gamma} - \gamma_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \tau(u_i, z_i; \gamma_0) + o_p(1),$$

where $\tau(u, z; \gamma_0)$ as its influence function.

**Theorem 5.** Let $\hat{\theta}$ be the proposed estimator through maximizing $l(\theta, \hat{\gamma}, \hat{F})$. Assume the following regularity conditions hold:

(a) $E\{H(\theta, \hat{\gamma}, \hat{F}) - H(\theta, \gamma_0, F_0)\} \to 0$, for any $\gamma, \hat{F}$ such that $\|\hat{\gamma} - \gamma_0\| \to 0$, $\|\hat{F} - F_0\| \to 0$, as $N \to \infty$;
(b) there exist \( \epsilon_1 > 0, \epsilon_2 > 0 \), such that

\[
\sup_{\theta, \|\gamma - \gamma_0\| < \epsilon_1, \|F - F_0\| < \epsilon_2} \left| \frac{1}{N} \sum_{i=1}^{N} H_i(\theta, \gamma, F) - EH(\theta, \gamma, F) \right| \rightarrow 0, \text{ as } N \rightarrow \infty,
\]

Then, \( \hat{\theta} \rightarrow \theta_0 \) in probability as \( N \rightarrow \infty \).

Next, we establish the asymptotical normality of \( \hat{\theta} \), which is crucial for large sample inference. We derive an asymptotic representation of \( \sqrt{N}(\hat{\theta} - \theta_0) \), which allows us to obtain an easy-to-compute consistent estimator of the asymptotic covariance matrix of \( \sqrt{N}(\hat{\theta} - \theta_0) \) without knowing its actual form.

**Theorem 6.** Assume the following regularity conditions:

(a) The function \( H(\theta, \gamma, F) \) is continuously twice differentiable with respect to \( \theta \). Denote

\[
A_1 = E \left[ \frac{\partial^2 H(\theta_0, \gamma_0, F_0)}{\partial \theta \partial \theta} \right] \quad \text{and} \quad A_2 = E \left[ \frac{\partial^2 H(\theta_0, \gamma_0, F_0)}{\partial \theta \partial \gamma} \right].
\]

\( A_1 \) is positive definite.

(b) There exists an open subset \( \Omega \) containing \( \theta_0 \) such that

\[
\sup_{\theta \in \Omega} \left\| \frac{\partial^2 H(\theta, \gamma_0, F_0)}{\partial \theta \partial \theta} \right\| < M,
\]

where \( M \) is an integrable function and \( \|A\|^2 = \text{trace}(A'A) \) for a matrix \( A \).

Then, as \( N \rightarrow \infty \),

\[
\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(u_i, z_i; \theta_0, \gamma_0, F_0, A_1, A_2) + o_p(1) \xrightarrow{d} N(0, \Sigma),
\]

where \( \Sigma \) is the covariance matrix of \( \psi \), and \( \psi \) is a known function defined in the proof.

Finally, we propose the following estimator of \( \Sigma \). Let \( D_i = \psi(u_i, z_i; \theta_0, \gamma_0, F_0, A_1, A_2) \). Since \( \Sigma = \text{Var}(D_i) \), the sample covariance matrix based on \( D_i, \ldots, D_N \) is a consistent estimator of \( \Sigma \). However, \( D_i \) contains some unknown quantities \( \theta_0, \gamma_0, F_0, A_1, A_2 \). Substituting
Di by $\hat{D}_i = \psi(u_i, z_i; \hat{\theta}, \hat{\gamma}, \hat{F}, \hat{A}_1, \hat{A}_2)$, where

$$
\hat{A}_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 H_i(\theta, \gamma, F)}{\partial \theta \partial \theta'} \bigg|_{\theta = \hat{\theta}, \gamma = \hat{\gamma}, F = \hat{F}} \quad \hat{A}_2 = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 H_i(\theta, \gamma, F)}{\partial \theta \partial \gamma'} \bigg|_{\theta = \hat{\theta}, \gamma = \hat{\gamma}, F = \hat{F}}.
$$

we define the sample covariance matrix based on $\hat{D}_i, i = 1, \ldots, N$ as our estimator $\hat{\Sigma}$.

**Theorem 7.** Assume that the conditions in Theorem 6 hold and that

(a) $\sup_{\|u\| \leq c_1, \|z\| \leq c_2} \| \psi(u, z, \hat{\theta}, \hat{\gamma}, \hat{F}, \hat{A}_1, \hat{A}_2) - \psi(u, z, \theta_0, \gamma_0, F_0, A_1, A_2) \| = o_p(1)$ for any $c_1, c_2 > 0$.

(b) There exist a constant $c_0 > 0$ and a function $h(u, z) \geq 0$ such that $E[h(U, Z)] < \infty$

and $P(\| \psi(u, z, \hat{\theta}, \hat{\gamma}, \hat{F}, \hat{A}_1, \hat{A}_2) \|^2 \leq h(u, z)$ for all $\|u\| \geq c_0, \|z\| \geq c_0) \to 1$.

Then, $\hat{\Sigma}$ is consistent for $\Sigma$.

### 2.7 Proofs of Theorems

**Proof of Theorem 1.** (i) For simplicity we only prove the case $p = 1$. Suppose $F_0(x)$ is the true cumulative distribution function of $X$. Consider

$$
p(x|y, \theta_0, F_0) = p(x|y, \theta, F_0),
$$

which happens if and only if

$$
\frac{p(y|x, \theta_0)}{\int p(y|x, \theta_0) dF_0(x)} = \frac{p(y|x, \theta)}{\int p(y|x, \theta) dF_0(x)}, a.s.
$$

if and only if

$$
\frac{p(y|x, \theta_0)}{p(y|x, \theta)} = c(y, \theta_0, \theta), a.s.
$$

It means, under the assumption given in Lemma 1 of Tang et al. (2003), the objective function to be maximized has a unique maximizer.
Next, we only need to prove that if \( X \) has three support values, e.g., \( \{0, 1, 2\} \), we can identify parameter \( \theta \). Firstly we can identify the following functions:

\[
\frac{\beta}{\phi}, b_1(\theta) = \frac{b(\alpha + \beta) - b(\alpha)}{\phi}, b_2(\theta) = \frac{b(\alpha + 2\beta) - b(\alpha)}{\phi}.
\]

Under the assumption that the following Jacobian determinant, evaluated at the true value, is not zero:

\[
\begin{vmatrix}
0 & 1 & \beta_0 \\
\beta_0 & b'(\alpha_0 + \beta_0) - b'(\alpha_0) & b(\alpha_0 + \beta_0) - b(\alpha_0) \\
\beta_0 & b'(\alpha_0 + 2\beta_0) - b'(\alpha_0) & 2b'(\alpha_0 + 2\beta_0) - b(\alpha_0) - b(\alpha_0)
\end{vmatrix} \neq 0,
\]

from the implicit function theorem, there exist neighborhoods

\[
U = B(\theta_0, \epsilon) \subset \Theta, V = B((\beta_0/\phi_0, b_1(\theta_0), b_2(\theta_0))^{\tau}, \eta) \subset \mathbb{R}^3, \epsilon, \eta > 0,
\]

and uniquely defined function \( g = (g_1, g_2, g_3)^{\tau} \) on \( V \), such that every component \( g_i \) is first order continuously differentiable, and

\[
\theta = g(\beta/\phi, b_1(\theta), b_2(\theta)),
\]

where \((\beta/\phi, b_1(\theta), b_2(\theta))^{\tau} \in V \) and \( \theta \in U \). Therefore,

\[
\theta_0 = g(\beta_0/\phi_0, b_1(\theta_0), b_2(\theta_0)),
\]

and the parameter \( \theta \) is identifiable.

(ii) When \( \phi \) is known, the condition shrinks to \( b'(\alpha_0 + \beta_0) \neq b'(\alpha_0) \). Since \( b' \) is strictly increasing, this condition is automatically satisfied if and only if \( \beta_0 \neq 0 \).

\( \square \)

**Proof of Corollary 1.** Suppose we only have two support values \( \{0, 1\} \), then we can identify the following two functions:

\[
u_1 = \frac{\beta}{\phi}, \text{ and } u_2 = \frac{b(\alpha + \beta) - b(\alpha)}{\phi}.
\]
To identify $\alpha, \beta, \phi$, we need to represent them from $u_1$ and $u_2$. Clearly, $\alpha$ and $\beta$ must satisfy
\[
\frac{b(\alpha + \beta) - b(\alpha)}{\beta} = \frac{u_2}{u_1}.
\]
(12)

Firstly, if $b(x) = mx + n$, i.e., $b'$ is a constant, then the above equation becomes $m = \frac{u_2}{u_1}$, obviously we can not identify $\alpha$ and $\beta$. If $b'$ is not a constant, from implicit function theorem, we have
\[
\alpha = h_1(u_2/u_1, \beta),
\]
i.e., there are infinite many pairs $(\alpha, \beta)$ satisfying equation (12). Actually, if $(\tilde{\alpha}, \tilde{\beta})$ satisfies equation (12), then $(\tilde{\alpha} + \tilde{\beta}, -\tilde{\beta})$ also satisfies (12), hence, $\alpha, \beta$ are not identifiable. 

**Proof of Theorem 3.** For density function $p(z|y, u; \tilde{\theta}_1, \gamma_0, F_0)$, from Jensen’s inequality, we have
\[
E_0[- \log p(z|y, u; \tilde{\theta}_{10}, \gamma_0, F_0)] \leq E_0[- \log p(z|y, u; \tilde{\theta}_1, \gamma_0, F_0)],
\]
and ”=” holds if and only if $p(z|y, u; \tilde{\theta}_{10}, \gamma_0, F_0) = p(z|y, u; \tilde{\theta}_1, \gamma_0, F_0)$ a.s. under the truth $\theta_0$, 
\[
\begin{align*}
\Leftrightarrow \quad & \frac{p(y|u, z; \theta_0)p(u|z; \gamma_0)f_0(z)}{\int p(y|u, z; \theta_0)p(u|z; \gamma_0)dF_0(z)} = \frac{p(y|u, z; \theta)p(u|z; \gamma_0)f_0(z)}{\int p(y|u, z; \theta)p(u|z; \gamma_0)dF_0(z)} \\
\Leftrightarrow \quad & \frac{p(y|u, z; \theta_0)}{p(y|u, z; \theta)} = \frac{\int p(y|u, z; \theta_0)p(u|z; \gamma_0)dF_0(z)}{\int p(y|u, z; \theta)p(u|z; \gamma_0)dF_0(z)} = c(y, u, \theta_0, \theta, \gamma_0).
\end{align*}
\]
By assumption, $P(D_0) < 1$ for any $\theta \neq \theta_0$, $P(R = 1) > 0$, hence,
\[
E_0[1_{(R=1)} \log p(z|y, u; \tilde{\theta}_1, \gamma_0, F_0)] < E_0[1_{(R=1)} \log p(z|y, u; \tilde{\theta}_{10}, \gamma_0, F_0)], \tilde{\theta}_1 \neq \tilde{\theta}_{10}.
\]

**Proof of Corollary 3.** Under the given assumption, it is straightforward that $h_1(y, u, z; \tilde{\theta}_1)$ can be identified through $[Z|Y, U]$. Therefore, $\tilde{\theta}_1$ is identifiable.
Proof of Corollary 4. For simplicity, we only prove (i). When $U$ is continuous, we can identify $\frac{\partial z}{\phi}$ and $\frac{b(\alpha + \beta_u + \beta_z)}{\phi} - \frac{b(\alpha + \beta_u)}{\phi}$, $\forall u$. For simplicity, we pick $u = 0, 1, 2$, and we consider when the following Jacobian matrix will not be of full rank.

$$
\begin{pmatrix}
0 & 0 & 1 \\
\beta_z & b'(\alpha + \beta_z) - b'(\alpha) & 0 \\
b'\left(\alpha + \beta_u + \beta_z\right) - b'(\alpha + \beta_u) & b'(\alpha + \beta_u + \beta_z) - b'(\alpha + \beta_u) & b'(\alpha + \beta_u + \beta_z)
\end{pmatrix}
$$

Scenario 1: suppose there are non zero constants $A,B,C,D$, such that

$$
C + D\beta_z = 0,
$$

$$
A\left(b'(\alpha + \beta_z) - b'(\alpha)\right) + Cb'(\alpha + \beta_z) + D(b(\alpha + \beta_z) - b(\alpha)) = 0,
$$

$$
A\left(b'(\alpha + \beta_u + \beta_z) - b'(\alpha + \beta_u)\right) + B(b'(\alpha + \beta_u + \beta_z) - b'(\alpha + \beta_u)) + Cb'(\alpha + \beta_u + \beta_z) + D(b(\alpha + \beta_u + \beta_z) - b(\alpha + \beta_u)) = 0,
$$

$$
A\left(b'(\alpha + 2\beta_u + \beta_z) - b'(\alpha + 2\beta_u)\right) + 2B(b'(\alpha + 2\beta_u + \beta_z) - b'(\alpha + 2\beta_u)) + Cb'(\alpha + 2\beta_u + \beta_z) + D(b(\alpha + 2\beta_u + \beta_z) - b(\alpha + 2\beta_u)) = 0,
$$

then, we have the following

$$
\frac{b(\alpha + \beta_z) - b(\alpha) - \beta_z b'(\alpha + \beta_z)}{b'(\alpha + \beta_z) - b'(\alpha)} + \frac{b(\alpha + 2\beta_u + \beta_z) - b(\alpha + 2\beta_u) - \beta_z b'(\alpha + 2\beta_u + \beta_z)}{b'(\alpha + 2\beta_u + \beta_z) - b'(\alpha + 2\beta_u)} = 2 \frac{b(\alpha + 2\beta_u + \beta_z) - b(\alpha + \beta_u) - \beta_z b'(\alpha + 2\beta_u + \beta_z)}{b'(\alpha + 2\beta_u + \beta_z) - b'(\alpha + \beta_u)},
$$
hence, from Cauchy’s functional equation, for any fixed $\beta_z \neq 0$, if we define

$$d(x) = \frac{b(x + \beta_z) - b(x) - \beta_z b'(x + \beta_z)}{b'(x + \beta_z) - b'(x)}, \quad (13)$$

then $d''(x) = 0$.

Scenario 2: if $A, B, C, D$ are not all non-zero, then we will have the following

$$\frac{b(\alpha + y) - b(\alpha) - yb'(\alpha + y)}{b'(\alpha + y) - b'(\alpha)}$$

to be a constant, which is equivalent to $b'' = 0$.

This completes the proof.

Proof of Theorem 5. Under the regularity conditions,

$$\hat{\theta} = \arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} H(\theta, \gamma, \hat{F})$$

$$= \arg\max_\theta \left[ \left\{ \frac{1}{N} \sum_{i=1}^{N} H(\theta, \gamma, \hat{F}) - EH(\theta, \gamma, F)|_{\gamma=\hat{\gamma}, F=\hat{F}} \right\} + EH(\theta, \gamma, F)|_{\gamma=\hat{\gamma}, F=\hat{F}} \right]$$

$$= o_p(1) + \arg\max_\theta \{ EH(\theta, \gamma, F)|_{\gamma=\hat{\gamma}, F=\hat{F}} - EH(\theta, \gamma_0, F_0) + EH(\theta, \gamma_0, F_0) \}$$

$$= o_p(1) + \arg\max_\theta E \{ R \log p(z|y, u, \theta, \gamma_0, F_0) - R \log p(u, z|\gamma_0, F_0) \}$$

$$= o_p(1) + \theta_0.$$

Proof of Theorem 6. Denote $\nabla_\theta l(\theta, \gamma, F)$ to be the derivative with respect to $\theta$. From Taylor’s expansion and $\nabla_\theta l(\hat{\theta}, \hat{\gamma}, \hat{F}) = 0$, we have

$$-\nabla_\theta l(\theta_0, \gamma_0, F_0) = \nabla_\theta l(\hat{\theta}, \hat{\gamma}, \hat{F}) - \nabla_\theta l(\theta_0, \hat{\gamma}, \hat{F}) + \nabla_\theta l(\theta_0, \hat{\gamma}, \hat{F}) - \nabla_\theta l(\theta_0, \gamma_0, \hat{F})$$

$$+ \nabla_\theta l(\theta_0, \gamma_0, \hat{F}) - \nabla_\theta l(\theta_0, \gamma_0, F_0)$$

$$= \nabla^2_{\theta\theta} l(\theta_0, \hat{\gamma}, \hat{F})(\hat{\theta} - \theta_0) + \nabla^2_{\theta\gamma} l(\theta_0, \gamma_0, \hat{F})(\hat{\gamma} - \gamma_0)$$

$$+ \nabla_\theta l(\theta_0, \gamma_0, \hat{F}) - \nabla_\theta l(\theta_0, \gamma_0, F_0) + o_p(N^{-1/2}).$$
From the theory of V-statistics, we have

\[
\nabla_{\theta} l(\theta_0, \gamma_0, \hat{F}) - \nabla_{\theta} l(\theta_0, \gamma_0, F_0) = V_N + o_p(N^{-1/2})
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} 2v(z_i; \theta_0, \gamma_0, F_0) + o_p(N^{-1/2}),
\]

where

\[
v(z_i; \theta_0, \gamma_0) = \frac{P(R = 1)}{2} E \left\{ \frac{\int \nabla_{\theta} p(y_j | u_j, z; \theta_0) p(u_j | z; \gamma_0) dF_0 p(y_j | u_j, z_i; \theta_0) p(u_j | z_i; \gamma_0)}{(\int p(y_j | u_j, z; \theta_0) p(u_j | z; \gamma_0) dF_0)^2} \right.
\]

\[\left. - \nabla_{\theta} p(y_j | u_j, z_i; \theta_0) p(u_j | z_i; \gamma_0) \right| r_i = 1, z_i \}.
\]

Under the regularity conditions, we have

\[
\sqrt{N}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} -A_i^{-1} \left[ A_2 \tau(u_i, z_i; \gamma_0, F_0) + \frac{\partial H_i(\theta_0, \gamma_0, F_0)}{\partial \theta} + 2v(z_i; \theta_0, \gamma_0, F_0) \right] + o_p(1),
\]

and this completes the proof.
Chapter 3

Estimation with Nonignorable Missing Response and Covariates

3.1 Nonignorable Missing Covariates and the Estimation Procedure

In some response-biased sampling problems, the observed data are a sample from $[X, Y]$ with the sampling probability depending only on a univariate response $Y$. (Here $Y$ is treated as a response and $X$ is the covariate variable. Throughout the chapter, we use $[\cdot]$ and $[\cdot|\cdot]$ to denote a generic distribution or conditional distribution.) Hence, the observed data are a random sample from $[X|Y]$, and some statistical inference procedure can be conducted from the regression $[X|Y]$, instead of $[Y|X]$. Chen (2001), using this idea, firstly proposed semi-parametric maximum likelihood estimators of the identifiable regression coefficients. They showed that, part of the regression parameters of $[Y|X]$, can be identified by the likelihood of $[X|Y]$ defined on the subjects of complete cases. Tang et al. (2003) extended this idea to a non-ignorable non-response missing data problem, and showed that under some conditions,
the regression parameters are partially identifiable and consistent estimators of identified
differentiable and consistent estimators of identified
parameters are provided. Especially, Tang et al. (2003) focused on the case where the
response $Y$ is normal, and the formula of $[X|Y]$ can be derived and written explicitly.

However, Tang et al. (2003) assumed that the covariates are fully observed, which is
not practical most of the time. In this section, we allow both the response and the covariates
may have non-ignorable missing values, and we derive some valid estimation procedures.

Let $Y$ be the binary variable taking values 0 or 1. (This is practically useful in some case-
control studies and clinical trials. Possible generalizations of the whole procedures will be
discussed in later sections.) We assume there are two components for the covariates: $X$
and $Z$, where $X$ has missing values and $Z$ is fully observed. Suppose the regression model
$[Y|X,Z]$ is $f(y|x,z;\theta)$, where $f$ is a known function and $\theta$ is the parameter of interest. We
also impose a regression model on $[X|Z]$, which is $g(x|z;\alpha)$, where $g$ is a known function
and $\alpha$ is a nuisance parameter. For the distribution of $Z$, we consider two alternatives: one
is to assume a parametric model $h(z;\beta)$, where $h$ is a known function and $\beta$ is an unknown
nuisance parameter, the other is to assume a non-parametric empirical distribution $H_n(z)$,
where $H_n(z) = \frac{1}{n} \sum_{i=1}^{n} I(z \leq z_i)$ is the empirical distribution of $Z$. More specifically, we
consider the independent and identical distributed data set $\{y_i, x_i, z_i\}$, for $i = 1, \cdots, N$ and
we assume the first $n$ samples are complete cases without loss of generality.

For the missing data mechanism, we suppose $R_\xi$ to be the missing indicator of the
random variable $\xi$, which means $\xi$ is observed if $R_\xi = 1$, and $\xi$ is missing if $R_\xi = 0$. For
example, we denote $R_Y$ to be the missing indicator of the response $Y$ and $R_X$ be the missing
indicator of the incomplete covariate $X$. In the following derivations, we firstly investigate
the case where only $X$ has missing values and we denote $R_X$ as $R$ for simplicity. The case
when both $Y$ and $X$ have non-ignorable missing values will be derived in later parts.
In this part, we firstly consider the case with missing covariate $X$ only and we simply use $R$ to denote missing indicator. In reality, the missing data mechanism is often more difficult to understand and the parametric form of $[R|Y,X,Z]$ is often not well known, even not 'testable'. Also, the parametric assumptions on missing data mechanism may be sensitive. Therefore, it is of interest to consider statistical methods that do not require specification of the missing mechanism. But, on the other hand, it is hopeless to do statistical inference if we do not make any assumptions on the missing mechanism. In this part, we consider the assumption

$$[R|Y,X,Z] = [R|Y,X].$$

(1)

Under this assumption, the complete cases are a random sample from $[Z|Y,X]$, therefore a natural way is to base inference on the likelihood of parameters of $[Z|Y,X]$ based on the complete cases, which is defined as $L_2(\theta, \alpha; \beta)$ below.

Define

$$L_1(\beta) = \prod_{i=1}^{N} h(z_i; \beta),$$

$$L_2(\theta, \alpha; \beta) = \prod_{r_i=1} \left[ \frac{f(y_i|x_i, z_i; \theta)g(x_i|z_i; \alpha)h(z_i; \beta)}{\int f(y_i|x_i, z; \theta)g(x_i|z; \alpha)h(z; \beta)dz} \right]$$

for parametric case, or

$$L_2(\theta, \alpha) = \prod_{r_i=1} \left[ \frac{f(y_i|x_i, z_i; \theta)g(x_i|z_i; \alpha)}{\sum_{j=1}^{N} f(y_i|x_i, z_j; \theta)g(x_i|z_j; \alpha)} \right]$$

for nonparametric case,

then we propose the following two-step pseudo-likelihood estimation procedure to derive the estimator of $\theta$:

Step 1: estimate $\beta$ by maximizing the likelihood $L_1(\beta)$, and denote it as $\hat{\beta}$. If $[Z]$ is of nonparametric form, let $\hat{\beta} = H_n(z)$. This step utilizes all observed $Z$, and hence $\hat{\beta}$ is efficient.

Step 2: estimate $(\hat{\theta}, \hat{\alpha})$ simultaneously by maximizing the pseudo-likelihood $L_2(\theta, \alpha; \hat{\beta})$,
where \( \hat{\beta} \) is derived from step 1. Then \( \hat{\theta} \) is the final valid estimator of \( \theta \).

To illustrate the intuition behind \( L_2(\theta, \alpha; \beta) \), let’s take the nonparametric version for an example. Then,

\[
\log L_2(\theta, \alpha) = \sum_{r_i = 1} \left\{ \log f(y_i | x_i, z_i; \theta) + \log g(x_i | z_i; \alpha) - \log \sum_{j=1}^N f(y_i | x_i, z_j; \theta) g(x_i | z_j; \alpha) \right\},
\]

which utilizes the complete case \( Y \) and \( X \), although it utilizes all \( Z \). The last term, \( \log \sum_{j=1}^N f(y_i | x_i, z_j; \theta) g(x_i | z_j; \alpha) \), is similar to a penalty function in high dimensional data penalized inference.

We can see that, in step 2, we need to maximize the pseudo-likelihood function over \( \theta \) and \( \alpha \) simultaneously, where \( \theta \) is the parameter of interest, but \( \alpha \) is a nuisance parameter. This motivates us to derive the following three-step pseudo-likelihood estimation procedure under an additional assumption

\[
[R|X, Z] = [R|X]. \tag{2}
\]

Similarly, define

\[
L_3(\alpha; \beta) = \prod_{r_i = 1} \frac{g(x_i | z_i; \alpha) h(z_i; \beta)}{\int g(x_i | z_i; \alpha) h(z; \beta) dz} \text{ for parametric case } ,
\]

\[
L_3(\alpha) = \prod_{r_i = 1} \frac{g(x_i | z_i; \alpha)}{\sum_{j=1}^N g(x_i | z_j; \alpha)} \text{ for nonparametric case } ,
\]

\[
L_4(\theta; \alpha, \beta) = \prod_{r_i = 1} \frac{f(y_i | x_i, z_i; \theta) g(x_i | z_i; \alpha) h(z_i; \beta)}{\int f(y_i | x_i, z; \theta) g(x_i | z; \alpha) h(z; \beta) dz} \text{ for parametric case } ,
\]

\[
L_4(\theta; \alpha) = \prod_{r_i = 1} \frac{f(y_i | x_i, z_i; \theta) g(x_i | z_i; \alpha)}{\sum_{j=1}^N f(y_i | x_i, z_j; \theta) g(x_i | z_j; \alpha)} \text{ for nonparametric case } ,
\]

and the estimation procedure is as follows.

Step 1’: same as step 1.

Step 2’: estimate \( \alpha \) by maximizing the pseudo-likelihood \( L_3(\alpha; \hat{\beta}) \). This estimator is valid under the additional assumption, which means the complete cases are a random sample
from \([Z|X]\).

Step 3': estimate \(\theta\) by maximizing the pseudo-likelihood \(L_4(\theta; \hat{\alpha}, \hat{\beta})\).

However, we need assumptions both (1) and (2) to make this three-step procedure work. A natural question will be when and how assumptions (1) and (2) can be both satisfied. If \([Y|X, Z] = [Y|X]\), which means \(Y \perp Z|X\), then the two assumptions \([R|Y, X, Z] = [R|Y, X]\) and \([R|X, Z] = [R|X]\) are equivalent. This is obvious from Dawid (1979). However, our proposed estimation procedure won’t work under assumption \([Y|X, Z] = [Y|X]\), since the parameters won’t be identifiable. It is not surprising, since the complete observed covariate \(Z\) plays an important role in our estimation procedure, and if we make some assumptions under which \(Y\) and \(Z\) are ‘separated’, then \(Z\) is useless in some sense.

It is easy to see that, if \([R|Y, X, Z] = [R|X]\), then both assumptions (1) and (2) will be automatically satisfied. Under this assumption \([R|Y, X, Z] = [R|X]\), our estimation procedure still works, however, it can be proved that in this case, it is equivalent to complete case analysis.

This part states that under some appropriate assumptions, we can propose a set of estimation procedures. The theoretical properties of the corresponding estimators, including identifiability, will be addressed in the following part.

### 3.2 Theoretical Results

In this part, we show some theoretical results by using the two-step pseudo-likelihood estimation procedure stated above (steps 1,2). The analogous results for three-step pseudo-likelihood estimation procedure (steps 1',2',3') are similar to obtain. The following lemma and theorem give conditions under which \(\theta\) is identifiable.
**Lemma.** Suppose the joint distribution \([Y, X, Z], P_0\), admits the density function \(p(y, x, z; \theta_0, G_0)\), where \(\theta_0\) is the true regression parameter of \([Y|X, Z]\), and \(G_0\) is the true joint distribution function \([X, Z]\). For any given \(\theta \neq \theta_0\), and an arbitrary function \(c(y, x)\), define

\[
D_\theta = \{(y, x) : f(y|x, z; \theta) = c(y, x) f(y|x, z; \theta_0), \text{ for any } z\}.
\]

If \(P(R = 1) > 0\) and \(P(D_\theta) < 1\) for any \(\theta \neq \theta_0\), then

\[
E_0\{I(R = 1) \log p(z|y, x, \theta, G_0)\} < E_0\{I(R = 1) \log p(z|y, x, \theta_0, G_0)\}, \theta \neq \theta_0,
\]

where \(E_0(\cdot)\) denotes the expectation with respect to \(P_0\).

The proof of the lemma is given in the appendix.

Notice that \(f(y|x, z; \theta)g(x, z; G) = f(y|x, z; \theta')g(x, z; G')\) if and only if \(f(y|x, z; \theta) = f(y|x, z; \theta')\) and \(g(x, z; G) = g(x, z; G')\), which means whether the true distribution \(G_0\) is known or not does not affect the identifiability of \(\theta\). Given the result in the above lemma, the following theorem is obvious.

**Theorem 1.** Suppose \(\theta\) can be reparameterized as \(\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)\) with true value \(\tilde{\theta}_0 = (\tilde{\theta}_{01}, \tilde{\theta}_{02})\), and the density function can be written as

\[
f(y|x, z, \theta) = \exp\{h_1(x, y, z; \tilde{\theta}_1) + h_2(x, y; \tilde{\theta})\},
\]

such that, for any \(b \neq \tilde{\theta}_{01}\) in the domain of \(\tilde{\theta}_1\), the function \(h_1(x, y, z; b) - h_1(x, y, z; \tilde{\theta}_{01})\) is not a function of \((x, y)\) alone. Then \(\tilde{\theta}_1\) is identifiable.

In practical situations, if \([Y]\) is binary distributed with canonical link, for example,

\[
P(Y = 1|X, Z) = \frac{\exp(\theta_0 + \theta_1 X + \theta_2 Z)}{1 + \exp(\theta_0 + \theta_1 X + \theta_2 Z)}, \text{ then } \theta = (\theta_0, \theta_1, \theta_2) \text{ is fully identifiable. Furthermore, if } [Y] \text{ is binary distributed with non-canonical link, then } \theta \text{ is also fully identifiable. From this theorem, we can see that our results are not only restricted to binary response case. In situations where response is general discrete or continuous random variable, our procedure also works only if it satisfies the conditions stated above.}
Denote P1 as the pseudo-likelihood estimator when $H(\cdot)$ is parametric with estimated $\hat{\lambda}$, P2 the pseudo-likelihood estimator when $H(\cdot)$ is non-parametric. Similar to Tang et al. (2003), as in the asymptotic theory of the traditional pseudo-likelihood method, which are stated in Gong et al. (1981), Parke (1986), and Pierce (1982), the variability of the estimators of the nuisance parameters should be incorporated in deriving the asymptotic properties of the pseudo-likelihood estimator of $\theta$. The following theorems establish the consistency and asymptotic normality of the P1 and P2 estimators respectively.

**Theorem 2.** Let $\hat{\theta}$ be a P1 or P2 estimator of $\theta$. Define

$$k(w; \theta, \alpha, H) = I(R = 1) \left\{ \log f(y|x, z; \theta) - \log \int f(y|x, z; \theta) g(x|z; \alpha) dH(z) \right\},$$

where $w = (y, x, z, R)$. In addition to the conditions given in the above lemma, assume that the following regularity conditions hold:

(a) $E\{k(w; \theta, \hat{\alpha}, \hat{H}) - k(w; \theta, \alpha_0, H_0)\} \to 0$, for any $\hat{\alpha}, \hat{H}$ such that $\|\hat{\alpha} - \alpha_0\| \to 0$, $\|\hat{H} - H_0\| \to 0$, as $n \to \infty$;

(b) there exist $\epsilon_1 > 0$, $\epsilon_2 > 0$, such that

$$\sup_{\theta, \|\alpha - \alpha_0\| < \epsilon_1, \|H - H_0\| < \epsilon_2} \left| \frac{1}{n} \sum_{i=1}^{n} k(w_i; \theta, \alpha, H) - Ek(w; \theta, \alpha, H) \right| \to 0, \text{ as } n \to \infty,$$

Then, $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$.

**Theorem 3.** Denote

$$s(\beta) = \log h(z; \beta),$$

$$t(\alpha, \beta) = I(R = 1) \left\{ \log g(x|z; \alpha) - \log \int g(x|z; \alpha) h(z; \beta) dz \right\},$$

$$l(\theta, \alpha, \beta) = I(R = 1) \left\{ \log f(y|x, z; \theta) - \log \int f(y|x, z; \theta) g(x|z; \alpha) h(z; \beta) dz \right\}.$$
Under the regularity conditions given in the identifiability lemma and consistency theorem, the P1 estimator $\hat{\theta}$ is asymptotically normal:

\[
\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma_1),
\]
as $n \to \infty$, where $\Sigma_1$ is presented in the appendix.

**Theorem 4.** Denote

\[
t(\alpha) = I(R = 1) \left\{ \log g(x|z; \alpha) - \log \int g(x|z; \alpha) dH(z) \right\},
\]

\[
l(\theta, \alpha) = I(R = 1) \left\{ \log f(y|x, z; \theta) - \log \int f(y|x, z; \theta) g(x|z; \alpha) dH(z) \right\}.
\]

Under the regularity conditions given in the identifiability lemma and consistency theorem, the P2 estimator $\hat{\theta}$ is asymptotically normal:

\[
\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \Sigma_2),
\]
as $n \to \infty$, where $\Sigma_2$ is presented in the appendix.

### 3.3 Non-ignorable missing covariate and response

The last two parts state the detailed results with non-ignorable missing covariate. In practice, we may also encounter situations where both covariate and response are subject to missing values. This case will be addressed in this part. Similarly, under appropriate assumptions, we propose pseudo-likelihood estimation procedures, and obtain the identifiability results, and the corresponding asymptotic results.

Since we have two missing components in this situation, we denote $R_Y$ and $R_X$ as the missing indicators of response $Y$ and covariate $X$, respectively. The assumption of the missing mechanism is

\[
[R_Y, R_X|Y, X, Z] = [R_Y, R_X|Y, X].
\]
Under this assumption, the complete case data are a random sample from \([Z|Y, X]\). In this part, we propose similar three-step pseudo-likelihood estimation procedure under assumptions (3) and (4):

\[
[R_X|X, Z] = [R_X|X].
\] (4)

Notice that (3) is an assumption imposed on the joint distribution of \(R_Y\) and \(R_X\). It allows the correlation between \(R_Y\) and \(R_X\). The assumptions are reasonable in practice. For example, if we assume \([R_Y|R_X, Y, X, Z] = [R_Y|R_X, Y, X]\) and \([R_X|Y, X, Z] = [R_X|X]\), then it is easy to see that the assumptions (3) and (4) are satisfied.

Define

\[
L_3(\alpha; \beta) = \prod_{r_{x_i}=1} g(x_i|z_i; \alpha)h(z_i; \beta) \int g(x_i|z; \alpha)h(z; \beta) dz
\] for parametric case ,

\[
L_3(\alpha) = \prod_{r_{z_i}=1} g(x_i|z_i; \alpha) \sum_{j=1}^{N} g(x_i|z_j; \alpha)
\] for nonparametric case ;

\[
L_4(\theta; \alpha, \beta) = \prod_{r_{y_i}=1, r_{x_i}=1} \frac{f(y_i|x_i, z_i; \theta)g(x_i|z_i; \alpha)h(z_i; \beta)}{\int f(y_i|x_i, z; \theta)g(x_i|z; \alpha)h(z; \beta) dz}
\] for parametric case ,

\[
L_4(\theta; \alpha) = \prod_{r_{y_i}=1, r_{x_i}=1} \frac{f(y_i|x_i, z_i; \theta)g(x_i|z_i; \alpha)}{\sum_{j=1}^{N} f(y_i|x_i, z_j; \theta)g(x_i|z_j; \alpha)}
\] for nonparametric case ,

and our estimation procedure is as follows:

Step A: same as step 1.

Step B: estimate \(\alpha\) by maximizing the pseudo-likelihood \(L_3(\alpha; \hat{\beta})\), or \(L_3(\alpha)\), for parametric and nonparametric cases separately.

Step C: estimate \(\theta\) by maximizing the pseudo-likelihood \(L_4(\theta; \hat{\alpha}, \hat{\beta})\), or \(L_4(\theta; \hat{\alpha})\).

The idea is that, after we impose one more missing indicator \(R_Y\), we can solve the awkward case that we need to maximize the parameter of interest and the nuisance parameter simultaneously. The identifiability theorem, the consistency, and the asymptotic normality of
the parameters can be derived similarly. To make parameters identifiable, we need complete observed covariate $Z$ to be continuous or discrete with point mass at at least three points. (Tang et al. 2003)

### 3.4 More than one missing covariates, and missing response

If there are more than one missing covariates, say, $X_1$, $X_2$, and their missing indicators are respectively $R_{X_1}$ and $R_{X_2}$. For simplicity we assume $R_{X_2} = 1$ if $R_{X_1} = 1$ (monotone missing). Similarly, if we make assumptions

$$[R_{X_2}|X_2, Z] = [R_{X_2}|X_2],$$

$$[R_{X_1}, R_{X_2}|Y, X_1, X_2, Z] = [R_{X_1}, R_{X_2}|Y, X_1, X_2],$$

then similar to 3.1, estimation procedure can be constructed, but as explained in 3.1, we also need to maximize the parameter of interest and part of the nuisance parameters simultaneously. Similarly, if we impose response missing indicator $R_Y$, then we can make the following assumptions

$$[R_{X_2}|X_2, Z] = [R_{X_2}|X_2],$$

$$[R_{X_1}, R_{X_2}|X_1, X_2, Z] = [R_{X_1}, R_{X_2}|X_1, X_2],$$

$$[R_Y, R_{X_1}, R_{X_2}|Y, X_1, X_2, Z] = [R_Y, R_{X_1}, R_{X_2}|Y, X_1, X_2],$$

which can be satisfied by the following practical assumptions

$$[R_{X_2}|X_2, Z] = [R_{X_2}|X_2],$$

$$[R_{X_1}|X_1, R_{X_2}, X_2, Z] = [R_{X_1}|X_1, R_{X_2}, X_2],$$

$$[R_Y|Y, R_{X_1}, X_1, R_{X_2}, X_2, Z] = [R_Y|Y, R_{X_1}, X_1, R_{X_2}, X_2],$$

then similar estimation procedure can be proposed as 3.3.
3.5 Numerical Results

To examine the finite sample behaviors of the proposed estimators, in this section, we compare the performance with the estimators without missing values, and the estimators providing MAR assumption in the following four scenarios. In each case, the total sample size $N = 500$ and the observed percentage is about 75-85%.

1. $[Y|U, Z]$ follows from logistic regression with canonical link, and

$$P[Y = 1|U, Z] = [1 + \exp \{-(-1 - 2U + 4Z)\}]^{-1};$$

$[U|Z] \sim N(1 - Z, 1); [Z] \sim N(0, 1); P[R = 1|Y, U] = 1 - \Phi(2Y + U).$

2. Same as scenario 1 except $P[U = 1|Z] = [1 + \exp \{-(-1 - Z)\}]^{-1}.$

3. $[Y|U, Z]$ follows from Poisson regression with canonical link, and

$$E[Y|U, Z] = \exp \{-0.5 - U + Z\};$$

$[U|Z] \sim N(1 - Z, 1); [Z] \sim N(0, 1); P[R = 1|Y, U] = 1 - \Phi(0.5 \times Y + U).$

4. Same as scenario 3 except $P[U = 1|Z] = [1 + \exp \{-(-1 - Z)\}]^{-1}.$

Based on 1000 simulation runs, the following tables report the bias for parameter estimation, standard deviation (SD) of the parameter estimate, standard error (SE), which is an estimate of SD, and the coverage probability (CP) of the approximately 95% confidence intervals of the parameter, using estimate±1.96SE, for each scenario. For scenarios 1-4, logistic regression or Poisson regression with theoretical dispersion parameter as 1, standard errors of the estimators providing no missing values, or MAR assumption, are obtained through formulas; while for scenarios with unknown dispersion parameter, standard errors of the estimators providing no missing values, or MAR assumption, are computed by bootstrapping based on 200 bootstrap samples. For all scenarios, the standard errors of the proposed
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>true=-1</td>
<td>true=-2</td>
<td>true=4</td>
<td></td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0414</td>
<td>-0.0675</td>
<td>0.1206</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3012</td>
<td>0.3206</td>
<td>0.5530</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.2873</td>
<td>0.3016</td>
<td>0.5312</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.1</td>
<td>94.7</td>
<td>95.7</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.3166</td>
<td>-0.1490</td>
<td>0.1535</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3436</td>
<td>0.3676</td>
<td>0.6220</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3224</td>
<td>0.3368</td>
<td>0.5925</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>78.2</td>
<td>95.2</td>
<td>95.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0335</td>
<td>-0.0937</td>
<td>0.1555</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3713</td>
<td>0.4776</td>
<td>0.6245</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3617</td>
<td>0.4590</td>
<td>0.5733</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.9</td>
<td>94.8</td>
<td>94.6</td>
</tr>
</tbody>
</table>

Table 3.1: Simulation results: $Y$ is binary, $U$ is continuous, $Y$ missing

Similarly, simulations under cases with both $Y$ and $U$ are missing are also conducted. The observed percentage is about 65-75%. Tables 1-4 summarize the results for binary $Y$ with only $Y$ missing, or both $Y$ and $U$ are missing, respectively. We consider continuous $U$ (normal distributed) and discrete $U$ (binary distributed). Tables 5-8 summarize the results for Poisson distributed $Y$. 

estimators are obtained through formulas.
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$ true=-1</th>
<th>$\beta_u$ true=-2</th>
<th>$\beta_z$ true=4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0112</td>
<td>-0.0591</td>
<td>0.0937</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3174</td>
<td>0.3854</td>
<td>0.4595</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3034</td>
<td>0.3692</td>
<td>0.4388</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.1</td>
<td>94.8</td>
<td>94.9</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.4506</td>
<td>-0.2580</td>
<td>0.1068</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3535</td>
<td>0.4352</td>
<td>0.5085</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3435</td>
<td>0.4199</td>
<td>0.4838</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>71.6</td>
<td>93.0</td>
<td>94.5</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0083</td>
<td>-0.0703</td>
<td>0.1070</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3512</td>
<td>0.4443</td>
<td>0.5065</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3455</td>
<td>0.4317</td>
<td>0.4893</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.6</td>
<td>94.8</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Table 3.2: Simulation results: Y is binary, U is discrete, Y missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>true=-1</td>
<td>true=-2</td>
<td>true=4</td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0299</td>
<td>-0.0668</td>
<td>0.1162</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.2898</td>
<td>0.3175</td>
<td>0.5462</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.2868</td>
<td>0.3013</td>
<td>0.5302</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>96.0</td>
<td>94.4</td>
<td>94.9</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.2132</td>
<td>-0.2910</td>
<td>0.2137</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.4065</td>
<td>0.4338</td>
<td>0.7177</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3804</td>
<td>0.3899</td>
<td>0.6862</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>89.2</td>
<td>92.5</td>
<td>95.3</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0473</td>
<td>-0.1242</td>
<td>0.2217</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.4553</td>
<td>0.5592</td>
<td>0.7195</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.4467</td>
<td>0.5488</td>
<td>0.7011</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.4</td>
<td>94.8</td>
<td>94.2</td>
</tr>
</tbody>
</table>

Table 3.3: Simulation results: Y is binary, U is continuous, Y and U missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>true=-1</td>
<td>true=-2</td>
<td>true=4</td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0130</td>
<td>-0.0417</td>
<td>0.0864</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.3062</td>
<td>0.3809</td>
<td>0.4610</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.3029</td>
<td>0.3681</td>
<td>0.4374</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.2</td>
<td>95.1</td>
<td>95.0</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.2555</td>
<td>-0.3012</td>
<td>0.1565</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.4384</td>
<td>0.5206</td>
<td>0.5934</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.4137</td>
<td>0.4910</td>
<td>0.5544</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>88.2</td>
<td>92.4</td>
<td>94.3</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0049</td>
<td>-0.1011</td>
<td>0.1625</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.4641</td>
<td>0.5548</td>
<td>0.5950</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.4613</td>
<td>0.5485</td>
<td>0.5783</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.3</td>
<td>95.2</td>
<td>95.0</td>
</tr>
</tbody>
</table>

Table 3.4: Simulation results: Y is binary, U is discrete, Y and U missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>true=-0.5</td>
<td>true=-1</td>
<td>true=1</td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>0.0010</td>
<td>0.0004</td>
<td>-0.0008</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.0707</td>
<td>0.0338</td>
<td>0.0479</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0692</td>
<td>0.0326</td>
<td>0.0456</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.5</td>
<td>95.1</td>
<td>94.3</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.0849</td>
<td>0.0184</td>
<td>-0.0181</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.0783</td>
<td>0.0351</td>
<td>0.0509</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0750</td>
<td>0.0333</td>
<td>0.0476</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>76.7</td>
<td>90.8</td>
<td>92.4</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0055</td>
<td>0.0040</td>
<td>0.0062</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.1008</td>
<td>0.0454</td>
<td>0.0626</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.9945</td>
<td>0.0450</td>
<td>0.0588</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.4</td>
<td>94.3</td>
<td>95.0</td>
</tr>
</tbody>
</table>

Table 3.5: Simulation results: Y is Poisson, U is continuous, Y missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$ (true=-0.5)</th>
<th>$\beta_u$ (true=-1)</th>
<th>$\beta_z$ (true=1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bias</td>
<td>-0.0019</td>
<td>-0.0030</td>
<td>0.0010</td>
</tr>
<tr>
<td>No missing</td>
<td>SD</td>
<td>0.1068</td>
<td>0.1317</td>
<td>0.0621</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1065</td>
<td>0.1329</td>
<td>0.0617</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.2</td>
<td>95.3</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>0.1232</td>
<td>-0.0012</td>
<td>-0.0349</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>SD</td>
<td>0.1179</td>
<td>0.1402</td>
<td>0.0653</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1196</td>
<td>0.1433</td>
<td>0.0662</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>81.1</td>
<td>96.7</td>
<td>91.8</td>
</tr>
<tr>
<td></td>
<td>bias</td>
<td>-0.0173</td>
<td>-0.0016</td>
<td>0.0088</td>
</tr>
<tr>
<td>Proposed</td>
<td>SD</td>
<td>0.1443</td>
<td>0.1821</td>
<td>0.0835</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1458</td>
<td>0.1765</td>
<td>0.0782</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.7</td>
<td>94.6</td>
<td>95.0</td>
</tr>
</tbody>
</table>

Table 3.6: Simulation results: $Y$ is Poisson, $U$ is discrete, $Y$ missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$ true=-0.5</th>
<th>$\beta_u$ true=-1</th>
<th>$\beta_z$ true=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0030</td>
<td>-0.0001</td>
<td>0.0016</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.0698</td>
<td>0.0323</td>
<td>0.0480</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0693</td>
<td>0.0327</td>
<td>0.0457</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.4</td>
<td>95.1</td>
<td>93.8</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.0480</td>
<td>0.0037</td>
<td>-0.0163</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.0822</td>
<td>0.0363</td>
<td>0.0553</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.0812</td>
<td>0.0362</td>
<td>0.0514</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>89.8</td>
<td>95.0</td>
<td>92.8</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0069</td>
<td>0.0034</td>
<td>0.0074</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.1056</td>
<td>0.0448</td>
<td>0.0676</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1044</td>
<td>0.0451</td>
<td>0.0630</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>94.5</td>
<td>94.6</td>
<td>94.5</td>
</tr>
</tbody>
</table>

Table 3.7: Simulation results: Y is Poisson, U is continuous, Y and U missing
<table>
<thead>
<tr>
<th>Method</th>
<th>Measure</th>
<th>$\alpha$</th>
<th>$\beta_u$</th>
<th>$\beta_z$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>true=-0.5</td>
<td>true=-1</td>
<td>true=1</td>
</tr>
<tr>
<td>No missing</td>
<td>bias</td>
<td>-0.0030</td>
<td>-0.0050</td>
<td>0.0011</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.1036</td>
<td>0.1353</td>
<td>0.0625</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1064</td>
<td>0.1327</td>
<td>0.0617</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>96.0</td>
<td>95.0</td>
<td>95.2</td>
</tr>
<tr>
<td>Assuming MAR</td>
<td>bias</td>
<td>0.0921</td>
<td>-0.0456</td>
<td>-0.0378</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.1340</td>
<td>0.1651</td>
<td>0.0798</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1360</td>
<td>0.1604</td>
<td>0.0784</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>89.6</td>
<td>94.2</td>
<td>92.1</td>
</tr>
<tr>
<td>Proposed</td>
<td>bias</td>
<td>-0.0172</td>
<td>-0.0138</td>
<td>0.0126</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>0.1651</td>
<td>0.2071</td>
<td>0.1011</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.1692</td>
<td>0.2033</td>
<td>0.1005</td>
</tr>
<tr>
<td></td>
<td>CP(%)</td>
<td>95.5</td>
<td>95.0</td>
<td>95.1</td>
</tr>
</tbody>
</table>

Table 3.8: Simulation results: Y is Poisson, U is discrete, Y and U missing
3.6 Proofs of Theorems

Proof of Identifiability Lemma. For any \( \theta \),

\[
E \{ I(R = 1) \log p(z|y, x, \theta, G_0) \} = E[ E \{ I(R = 1) \log p(z|y, x, \theta, G_0) | y, x \} ] \\
= -E[I(R = 1)E \{ - \log p(z|y, x, \theta, G_0) | y, x \} ].
\]

For any fixed \((y, x)\), since \(- \log(\cdot)\) is a strictly convex function, using Jensen’s inequality and information theory,

\[
E \{ - \log p(z|y, x, \theta_0, G_0) | y, x \} \leq E \{ - \log p(z|y, x, \theta, G_0) | y, x \},
\]

and the equality holds if and only if \( p(z|y, x, \theta_0, G_0) = p(z|y, x, \theta, G_0) \),

\[
\iff \frac{f(y|x, z; \theta_0)g_0(x,z)}{\int f(y|x, z; \theta_0)g_0(x,z)dz} = \frac{f(y|x, z; \theta)g_0(x,z)}{\int f(y|x, z; \theta)g_0(x,z)dz} = c(y, x, \theta, \theta_0).
\]

By assumption, \( P(R = 1) > 0, P(D_\theta) < 1 \) for any \( \theta \neq \theta_0 \), then

\[
E[I(R = 1)E \{ \log p(z|y, x, \theta_0, G_0) \}] > E[I(R = 1)E \{ \log p(z|y, x, \theta, G_0) \}], \text{ for any } \theta \neq \theta_0.
\]

Proof of Consistency Theorem. Under the regularity conditions,

\[
\hat{\theta} = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} k(w_i; \theta, \hat{\alpha}, \hat{\beta}) \\
= \arg \max_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} k(w_i; \theta, \hat{\alpha}, \hat{\beta}) - Ek(w; \theta, \alpha, \beta)|_{\alpha=\hat{\alpha}, \beta=\hat{\beta}} \right. + \left. Ek(w; \theta, \alpha, \beta)|_{\alpha=\hat{\alpha}, \beta=\hat{\beta}} \right\} \\
= o_p(1) + \arg \max_{\theta} \left\{ Ek(w; \theta, \alpha, \beta)|_{\alpha=\hat{\alpha}, \beta=\hat{\beta}} - Ek(w; \theta, \alpha_0, \beta_0) \right\} + Ek(w; \theta, \alpha_0, \beta_0) \\
= o_p(1) + \arg \max_{\theta} E \{ R \log p(z|y, x, \theta, \alpha_0, \beta_0) - R \log p(x, z|\alpha_0, \beta_0) \} \\
= o_p(1) + \theta_0.
\]

Proof of parametric Normality Theorem. From the first step, we know that

\[
\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, (E_{\hat{\beta}} s^2_{\beta})^{-1}).
\]
From the second step, we have

\[ 0 = \mathbb{P}_n t_\alpha(\hat{\alpha}, \hat{\beta}) \]

\[ = \mathbb{P}_n t_\alpha(\alpha_0, \hat{\beta}) + (\hat{\alpha} - \alpha_0) \mathbb{P}_n t_{\alpha \alpha}(\alpha_0, \hat{\beta}) \]

\[ = \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) + (\hat{\beta} - \beta_0) \mathbb{P}_n t_{\alpha \beta}(\alpha_0, \beta_0) + (\hat{\alpha} - \alpha_0) \mathbb{P}_n t_{\alpha \alpha}(\alpha_0, \hat{\beta}) \]

\[ = \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) - Et_\alpha t_\beta(\hat{\beta} - \beta_0) - Et_\alpha^2(\hat{\alpha} - \alpha_0) \]

Hence,

\[ \sqrt{n}(\hat{\alpha} - \alpha_0) = (Et_\alpha^2)^{-1} [\sqrt{n} \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) - Et_\alpha t_\beta \cdot \sqrt{n}(\hat{\beta} - \beta_0)] \]

\[ = (Et_\alpha^2)^{-1} [\sqrt{n} \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) - Et_\alpha t_\beta (Es_\beta^2)^{-1} \cdot \sqrt{n} \mathbb{P}_n s_\beta(\beta_0)] \]

\[ = (Et_\alpha^2)^{-1} [\sqrt{n} \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) - Et_\alpha s_\beta (Es_\beta^2)^{-1} \cdot \sqrt{n} \mathbb{P}_n s_\beta(\beta_0)], \]

where the last equality follows from \( Et_\alpha t_\beta = Et_\alpha s_\beta \), and this is because

\[ 0 = \frac{\partial}{\partial \beta}(Et_\alpha) = \frac{\partial}{\partial \beta} \int t_\alpha L_{\text{full}} dP = \int t_\alpha \frac{\partial L_{\text{full}}}{\partial \beta} dP + \int t_\alpha \frac{\partial L_{\text{full}}}{\partial \beta} dP \]

\[ = Et_\alpha + Et_\alpha s_\beta = -Et_\alpha t_\beta + Et_\alpha s_\beta. \]

Since

\[ \sqrt{n} \mathbb{P}_n t_\alpha(\alpha_0, \beta_0) \to_d N(0, Et_\alpha^2), \sqrt{n} \mathbb{P}_n s_\beta(\beta_0) \to_d N(0, Es_\beta^2), \]

then

\[ \sqrt{n}(\hat{\alpha} - \alpha_0) \to_d N(0, \Sigma_\alpha), \]

where

\[ \Sigma_\alpha = (Et_\alpha^2)^{-1} [Et_\alpha^2 - Et_\alpha s_\beta (Es_\beta^2)^{-1} Es_\beta t_\alpha] (Et_\alpha^2)^{-1}. \]
From the third step,

\[
0 = \mathbb{P}_n l_\theta(\hat{\theta}, \hat{\alpha}, \hat{\beta})
\]

\[
= \mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) + (\hat{\theta} - \theta_0)\mathbb{P}_n l_{\theta\theta}(\theta_0, \alpha_0, \beta_0)
\]

\[
= \mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) + (\hat{\alpha} - \alpha_0)\mathbb{P}_n l_{\alpha\alpha}(\theta_0, \alpha_0, \beta_0) + (\hat{\beta} - \beta_0)\mathbb{P}_n l_{\beta\beta}(\theta_0, \alpha_0, \beta_0) + (\hat{\beta} - \beta_0)\mathbb{P}_n l_{\beta\theta}(\theta_0, \alpha_0, \hat{\beta}) + (\hat{\theta} - \theta_0)\mathbb{P}_n l_{\theta\theta}(\theta_0, \hat{\alpha}, \beta)
\]

\[
= \mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) - El_\beta l_\beta \cdot (\hat{\beta} - \beta_0) - El_\alpha l_\alpha \cdot (\hat{\alpha} - \alpha_0) - El_\theta^2 \cdot (\hat{\theta} - \theta_0).
\]

Hence,

\[
\sqrt{n}(\hat{\theta} - \theta_0) = (El_\theta^2)^{-1} \left[ \sqrt{n}\mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) - El_\beta l_\beta \cdot \sqrt{n}(\hat{\beta} - \beta_0) - El_\alpha l_\alpha \cdot \sqrt{n}(\hat{\alpha} - \alpha_0) \right]
\]

\[
= (El_\theta^2)^{-1} \left[ \sqrt{n}\mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) - El_\beta s_\beta \cdot \sqrt{n}(\hat{\beta} - \beta_0) - El_\alpha l_\alpha \cdot \sqrt{n}(\hat{\alpha} - \alpha_0) \right]
\]

\[
= (El_\theta^2)^{-1} \left[ \sqrt{n}\mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) - El_\beta s_\beta (E s_\beta^2)^{-1} \cdot \sqrt{n}\mathbb{P}_n s_\beta(\beta_0) - El_\alpha l_\alpha \cdot \sqrt{n}(\hat{\alpha} - \alpha_0) \right]
\]

where the first two terms

\[
\sqrt{n}\mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) - El_\beta s_\beta \cdot \sqrt{n}(\hat{\beta} - \beta_0) \rightarrow_d N(0, \Sigma_1),
\]

and

\[
\Sigma_1 = El_\theta^2 - El_\beta s_\beta (E s_\beta^2)^{-1} E s_\beta l_\theta,
\]

which is true because of the similar arguments as in the second step estimation. Furthermore, if we assume

\[
\sqrt{n} \begin{pmatrix}
\mathbb{P}_n l_\theta(\theta_0, \alpha_0, \beta_0) \\
\mathbb{P}_n s_\beta(\beta_0) \\
\hat{\alpha} - \alpha_0
\end{pmatrix} \rightarrow_d MVN \begin{pmatrix}
El_\theta^2 & El_\beta s_\beta & \Sigma_{13} \\
E s_\beta l_\theta & E s_\beta^2 & \Sigma_{23} \\
\Sigma_{13}^T & \Sigma_{23}^T & \Sigma_{\alpha}
\end{pmatrix},
\]

then

\[
\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Sigma_\theta),
\]
where

\[
\Sigma_\theta = (El_\theta^2)^{-1} \{ El_\theta^2 - El_\theta s_\beta (Es_\beta^2)^{-1} Es_\beta l_\theta + El_\theta l_\alpha \Sigma_\alpha El_\alpha l_\theta - El_\theta l_\alpha \Sigma_{13}^T - \Sigma_{13} El_\alpha l_\theta \\
+ El_\theta l_\alpha \Sigma_{23}^T (Es_\beta^2)^{-1} Es_\beta l_\theta + El_\theta s_\beta (Es_\beta^2)^{-1} \Sigma_{23} El_\alpha l_\theta \} (El_\theta^2)^{-1}.
\]

**Proof of non-parametric Normality Theorem.** This proof is similar to Theorem 6 in Chapter 2.
Chapter 4

Estimation in Longitudinal Studies with Nonignorable Dropout

4.1 Introduction

Longitudinal data or repeated measurements are often encountered in medical, health, economical, and social studies. For a sampled subject, let $Y = (Y_1, ..., Y_T)$ be a $T$-dimensional longitudinal response vector and $X$ be a cross-sectional or longitudinal covariate vector associated with $Y$. We focus on the estimation or inference on some unknown parameters in $p(Y|X)$ or $p(Y)$, where $p(\cdot|\cdot)$ is a generic notation for the conditional density and $p(\cdot)$ is for the unconditional density. In many studies values of $X$ are completely observed but a subject may dropout at $t < T$ so that $(Y_t, ..., Y_T)$ is missing. Let $R = (R_1, ..., R_T)$, where $R_t = 1$ if $Y_t$ is observed and $R_t = 0$ if $Y_t$ is missing, $t = 1, ..., T$. If the subject drops out at time $t$, then $R_1 = \cdots = R_{t-1} = 1$ and $R_t = \cdots = R_T = 0$. Also, if $R_{t-1} = 0$, then $R_t = 0$ with certainty, $t = 1, ..., T$.

For longitudinal data, it is natural that dropout at time $t$ is not related to future values
\( Y_{t+1}, \ldots, Y_T \) \cite[e.g.,]{DiggleKenward1994} and, thus,

\[
P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_t, X, R_{t-1} = 1), \quad t = 1, \ldots, T. \tag{1}
\]

If dropout is not related to the current value \( Y_t \), i.e.,

\[
P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_{t-1}, X, R_{t-1} = 1), \quad t = 1, \ldots, T,
\]

then the dropout is ignorable \cite{Little1995, LittleRubin2002}, which is a much stronger assumption than (1) because the dropout propensity depends on \((Y_1, \ldots, Y_{t-1}, X)\) that is observed. Estimation methods under ignorable dropout are well developed \cite[e.g.,]{LittleRubin2002, Paik1997}. However, in many longitudinal studies dropout depends on not only \((Y_1, \ldots, Y_{t-1}, X)\) but also \( Y_t \) that may be missing and, hence, is nonignorable. Nonignorable dropout presents a great challenge in the estimation of unknown parameters in \( p(Y|X) \) or \( p(Y) \) \cite[see, e.g.,]{RobinsRotnitzkyZhao1995, TroxelHarringtonLipsitz1998, TroxelLipsitzHarrington1998}.

The purpose of our study is to develop an estimation method under nonignorable dropout. Without any further assumption, however, some unknown parameters in \( p(Y|X) \) or \( p(Y) \) are not identifiable. To identify the unknown parameters, we need to assume that some component of \( V_t = (Y_1, \ldots, Y_t, X) \) is not related to dropout, conditioned on the other components. Note that the ignorable dropout assumption assumes that \( Y_t \) is not related to dropout, conditional on the rest of the components of \( V_t \). For nonignorable dropout, we assume that \( X = (U, Z) \) and the component \( Z \) is not related to dropout conditional on other components of \( V_t \), i.e.,

\[
P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_1, \ldots, Y_{t-1}, U, R_{t-1} = 1), \quad t = 1, \ldots, T. \tag{2}
\]

The difference between (1) and (2) is that the covariate \( Z \) is not present on the right hand side of (2), which makes it possible for us to identify and estimate unknown parameters,
provided that Y and Z are dependent conditioned on U, i.e., Z is a useful covariate. Such a covariate Z is referred to as an instrument for dropout. Furthermore, we need to assume that at least one of \( p(Y|X) \) and \( P(R_t = 1|Y_1, \ldots, Y_t, U, R_{t-1} = 1) \) is parametric. Otherwise some unknown parameters are not identifiable (Robins and Ritov, 1997). In this chapter, we follow Tang, Little, and Raghunathan (2003) and assume a parametric model on \( p(Y|X) \):

\[
p(Y|X) = \prod_{t=1}^{T} f_t(Y_t|V_{t-1}, \theta_t),
\]

where \( f_t(Y_t|V_{t-1}, \theta_t) \) is the probability density of \( Y_t \) given \( V_{t-1} = (Y_1, \ldots, Y_{t-1}, X) \), \( f_t \)'s are known functions, and \( \theta_t \)'s are distinct unknown parameter vectors.

The approach in Tang et al. (2003) is for a general multivariate \( Y \) under the assumption \( p(R|Y, X) = p(R|Y) \) that allows us to estimate \( p(X|Y) \) using the observed \( (X, Y) \), and hence the parameters in \( p(Y|X) \) through the Bayes formula \( p(X|Y) = p(Y|X)p(X)/\int p(Y|x)p(x)dx \). However, this approach has the following two problems. First, it discards observed but incomplete data from dropped out subjects. Second, the dimension of \( X \) is required to be as large as the dimension of \( Y \), which limits the application scope. For longitudinal \( Y \), Tang et al. (2003) actually improved their approach regarding the previously discussed problems, but under the following assumption much stronger than (2):

\[
P(R_t = 1|Y, X, R_{t-1} = 1) = P(R_t = 1|Y_t, R_{t-1} = 1), \quad t = 1, \ldots, T,
\]

that is, conditioned on \( Y_t \), the dropout propensity depends on neither past responses \( Y_1, \ldots, Y_{t-1} \) nor the entire covariate vector \( X \).

Assuming (2) with no model on \( P(R_t = 1|Y_1, \ldots, Y_t, U, R_{t-1} = 1) \) and assuming (3), we derive a semiparametric pseudo likelihood for estimating parameters in \( p(Y|X) \) or \( p(Y) \).

We are able to utilize all observed data. Since our method is based on pseudo likelihoods constructed sequentially as \( t = 1, \ldots, T \), we do not require a high-dimensional covariate \( X \) to
identify parameters. Also, at each step the maximization in our method is carried out with a low dimensional vector of parameters and, hence, the computation is sensible.

The methodology is developed in Section 2. Consistency and asymptotic normality of the proposed estimators are shown in Section 3. Although the asymptotic normality follows from a standard argument, the asymptotic covariance matrices of the proposed estimators are very complicated, because of the use of previously estimated parameters in the pseudo likelihoods. We establish an asymptotic representation that allows us to obtain easy-to-compute consistent estimators of the asymptotic covariance matrices. Section 4 contains some empirical results. A discussion on assumptions is given in Section 5. The Appendix contains technical details.

4.2 Estimation Based on Pseudo Likelihoods

Under assumptions (2) and (3), we consider the estimation of \( \theta = (\theta_1, \ldots, \theta_T) \) based on an independent and identically distributed sample \( (Y_1^{(i)}, \ldots, Y_T^{(i)}, R_1^{(i)}, \ldots, R_T^{(i)}, X^{(i)}), \ i = 1, \ldots, n, \) from the population \( p(Y, R, X) \), where \( Y_t^{(i)} \) is observed if and only if \( R_t^{(i)} = 1 \).

4.2.1 The case where \( X \) is a dropout instrument

We first consider the case of \( X = Z \) and \( U = 0 \) in (2), i.e., conditioned on \( (Y_1, \ldots, Y_t) \), the dropout propensity does not depend on the entire covariate vector \( X \) so that \( X \) is a
dropout instrument. When \( t = 1 \), we consider the likelihood

\[
\prod_{R_1^{(i)} = 1} p(X^{(i)}|Y_1^{(i)}, R_1^{(i)} = 1) = \prod_{R_1^{(i)} = 1} p(X^{(i)}|Y_1^{(i)})
\]

\[
= \prod_{R_1^{(i)} = 1} \frac{p(Y_1^{(i)}|X^{(i)}) p(X^{(i)})}{\int p(Y_1^{(i)}|x)p(x)dx}
\]

\[
= \prod_{R_1^{(i)} = 1} \frac{f_1(Y_1^{(i)}|X^{(i)}) p(X^{(i)})}{\int f_1(Y_1^{(i)}|x)p(x)dx},
\]

where the first equality follows from assumption (2) with \( U = 0 \), the second equality follows from the Bayes formula, and the last equality follows from assumption (3). Substituting \( p(X) \) by the nonparametric empirical distribution of \( X \) by the nonparametric empirical distribution of \( (X, Y_1) \), we obtain an estimator \( \hat{\theta}_1 \) by maximizing the pseudo likelihood

\[
\prod_{R_1^{(i)} = 1} \frac{f_1(Y_1^{(i)}|X^{(i)}, \theta_1)}{\sum_{j=1}^n f_1(Y_1^{(i)}|X^{(i)}, \theta_1)}.
\]

For \( t = 2, \ldots, T \), suppose that \( \hat{\theta}_1, \ldots, \hat{\theta}_{t-1} \) have been obtained. Consider the likelihood

\[
\prod_{R_t^{(i)} = 1} p(X^{(i)}|Y_1^{(i)}, \ldots, Y_t^{(i)}, R_t^{(i)} = 1) = \prod_{R_t^{(i)} = 1} \frac{p(Y_1^{(i)}, \ldots, Y_t^{(i)}|X^{(i)}) p(X^{(i)})}{\int p(Y_1^{(i)}, \ldots, Y_t^{(i)}|x)p(x)dx}.
\]

Under (3),

\[
p(Y_1^{(i)}, \ldots, Y_t^{(i)}|X^{(i)}) = f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \theta_s),
\]

where \( V_s^{(i)} = (Y_1^{(i)}, \ldots, Y_s^{(i)}, X^{(i)}) \). Replacing each \( \theta_s \) by the previously obtained \( \hat{\theta}_s \) and \( p(X^{(i)}) \) by the nonparametric empirical distribution of \( X \), we estimate \( \theta_t \) by maximizing the pseudo likelihood

\[
\prod_{R_t^{(i)} = 1} \frac{f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \hat{\theta}_s)}{\sum_{j=1}^n f_t(Y_t^{(i)}|X^{(j)}, Y_1^{(i)}, \ldots, Y_{t-1}^{(i)}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|X^{(j)}, Y_1^{(i)}, \ldots, Y_{s-1}^{(i)}, \hat{\theta}_s)}.
\]

Note that all observed values up to time \( t \) are included in this likelihood.
It is implicitly assumed that \( f_t(Y_t|V_{t-1}, \theta_t) \) depends on \( X \), i.e., \( X \) is a useful covariate. Otherwise, \( f_t(Y_t(i)|V_{t-1}, \theta_t) \) can be canceled in (4) and the likelihood does not contain \( \theta_t \).

Let the number of covariates be \( K \geq 1 \). If \( X \) is cross-sectional, then \( X \) is \( K \)-dimensional. If \( X \) is longitudinal, then \( X = (X_1, ..., X_T) \) and each \( X_t \) is \( K \)-dimensional so that the dimension of \( X \) is \( KT \). For many longitudinal studies, \( Y_t \) is statistically related to \( X_1, ..., X_t \) only, \( t = 1, ..., T \). In such cases,

\[
p(Y_t|V_{t-1}) = p(Y_t|X_1, ..., X_t, Y_1, ..., Y_{t-1}) = f_t(Y_t|X_1, ..., X_t, Y_1, ..., Y_{t-1}, \theta_t)
\]

and the proposed pseudo likelihood (4) can be used with \( f_t(Y_t(i)|V_{t-1}, \theta_t) \) replaced by \( f_t(Y_t(i)|X_1(i), ..., X_t(i), Y_1(i), ..., Y_{t-1}(i), \theta_t) \), \( t = 1, ..., T \).

If we do not substitute \( \theta_1, ..., \theta_{t-1} \) by their estimates, in theory we can estimate \( (\theta_1, ..., \theta_t) \) by maximizing (4) with \( \hat{\theta}_s \) replaced by \( \theta_s \), \( s = 1, ..., t - 1 \). However, the computation may not be feasible because the dimension of \( (\theta_1, ..., \theta_t) \) is much higher than the dimension of \( \theta_t \).

The original approach in Tang et al. (2003) requires a check on whether we can identify \( \theta \) from the parameters in \( p(X|Y) \) because we estimate parameters in \( p(Y|X) \) through estimating parameters in \( p(X|Y) \) and the Bayes formula. When \( (Y, X) \) is multivariate normal, the requirement is that the dimension of \( X \) has to be at least \( T \). This restrictive requirement is not needed in our proposed approach under assumption (2), because we estimate \( \theta_t \)'s one at a time. In fact, when \( p(Y|X) \) is normal, a one-dimensional continuous \( X \) or discrete \( X \) taking at least 3 values is sufficient for estimating \( \theta \). This can be shown using induction. For \( t = 1 \), \( Y_1 \) is one-dimensional and the result in Tang et al. (2003) showed that \( \theta_1 \) in

\[
p(Y_1|X) = f_1(Y_1|X, \theta_1)
\]

can be estimated with a one-dimensional continuous \( X \) or discrete \( X \) taking at least 3 values when \( X \) is related to \( Y_1 \). Assuming that \( \theta_1, ..., \theta_{t-1} \) are estimated, we want to show that \( \theta_t \) in

\[
p(Y_t|V_{t-1}) = f_t(Y_t|V_{t-1}, \theta_t)
\]

can be estimated. Since parameters in \( p(Y_1, ..., Y_{t-1}|X) \) have been estimated, we can treat \( (Y_1, ..., Y_{t-1}, X) \) as a covariate vector.
and, thus, $\theta_t$ can be estimated based on the result in Tang et al. (2003).

### 4.2.2 The case where a sub-vector of $X$ is a dropout instrument

Let $X = (U, Z)$ as in [2]. Note that

$$p(Z|Y_1, ..., Y_t, U, R_t = 1) = p(Z|Y_1, ..., Y_t, U)$$

$$= \frac{p(Y_1, ..., Y_t, U|Z)p(Z)}{\int p(Y_1, ..., Y_t, U|Z)p(Z)dz}$$

$$= \frac{p(Y_1, ..., Y_t|U, Z)p(U|Z)p(Z)}{\int p(Y_1, ..., Y_t|U, z)p(U|Z)p(z)dz}$$

$$= \frac{p(Y_1, ..., Y_t|U, Z)p(Z|U)}{\int p(Y_1, ..., Y_t|U, z)p(z|U)p(U)dz}$$

First, if $U$ is a discrete covariate, then we can substitute $p(Z|U = u)$ by the empirical distribution of $Z$ conditioned on $U = u$, which results in the following likelihood for the estimation of $\theta_t$:

$$\prod_u \prod_{R_t^{(i)} = 1, U^{(i)} = u} \sum_{U^{(i)} = u} f_t(Y_t^{(i)}|V_{t-1}^{(i)}, \hat{\theta}_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}^{(i)}, \hat{\theta}_s)$$

where $\hat{\theta}_1, ..., \hat{\theta}_{t-1}$ are estimators from the previous steps. Next, consider the case where $U$ is continuous and a parametric model on $p(Z|U) = g_\xi(Z|U)$ is assumed, where $\xi$ is an unknown parameter vector. Since $U$ and $Z$ have no missing data, $\xi$ can be estimated by $\hat{\xi}$ using the likelihood based on $X^{(1)}, ..., X^{(n)}$, which leads to the following likelihood for the estimation of $\theta_t$:

$$\prod_{R_t^{(i)} = 1} \int f_t(Y_t^{(i)}|U^{(i)}, z, Y_1^{(i)}, ..., Y_{t-1}^{(i)}, \hat{\theta}_t) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|U^{(i)}, z, Y_1^{(i)}, ..., Y_{s-1}^{(i)}, \hat{\theta}_s)g_\xi(Z^{(i)}|U^{(i)})dz$$

Finally, consider the case where $U$ is continuous, a parametric model on $p(U|Z) = h_\zeta(U|Z)$ is assumed, where $\zeta$ is an unknown parameter vector, and $\zeta$ is estimated by $\hat{\zeta}$ using the likelihood based on $X^{(1)}, ..., X^{(n)}$. Then, the following likelihood can be used for the estimation
of $\theta_t$:

$$
\prod_{R^{(i)}_t=1} \sum_{j=1}^n \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|V_{s-1}, \hat{\theta}_s) h_z(U^{(i)}|Z^{(i)}) \prod_{s=1}^{t-1} f_s(Y_s^{(i)}|U^{(i)}, Z^{(i)}, Y_1^{(i)}, ..., Y_{s-1}^{(i)}, \hat{\theta}_s) h_z(U^{(i)}|Z^{(i)})
$$

In any case it is assumed that $f_t(Y_t|V_{t-1}, \theta_t)$ depends on $Z$, i.e., $Z$ is a useful covariate, although $f_t(Y_t|V_{t-1}, \theta_t)$ may not depend on $U$.

### 4.3 Asymptotic Properties

Under some regularity conditions, we now show that $\hat{\theta}_t$, $t = 1, ..., T$, are consistent and asymptotically normal as $n \to \infty$. For simplicity, we focus on the situation where $X = Z$ (Section 2.1). Results for the situations described in Section 2.2 can be similarly derived. In addition to (2) and (3), the following are two key conditions for the consistency of $\hat{\theta}_t$:

$$
\pi_t = P(R_t = 1) > 0, \quad t = 1, ..., T, \quad \text{(5)}
$$

and, for any $\theta_t$ in the parameter space that is not the same as the true parameter value $\theta^0_t$ and any function $\phi$ of $(Y_1, ..., Y_t, \theta_t)$,

$$
P \left( (Y_1, ..., Y_t) : \frac{f_t(Y_t|V_{t-1}, \theta_t)}{f_t(Y_t|V_{t-1}, \theta^0_t)} = \phi(Y_1, ..., Y_t, \theta_t) \text{ for any } X \right) < 1. \quad \text{(6)}
$$

We now explain why $\hat{\theta}_t$ is consistent under (5)-(6). Let $F$ denote the distribution of $X$ and, for any $t$, $\varphi_t = (\theta_1, ..., \theta_t, F)$ ($\varphi_0 = F$),

$$
G_t(\varphi_t) = \frac{f_t(Y_t|V_{t-1}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s|V_{s-1}, \theta_s) dF^0(X)}{\int f_t(Y_t|x, Y_1, ..., Y_{t-1}, \theta_t) \prod_{s=1}^{t-1} f_s(Y_s|x, Y_1, ..., Y_{s-1}, \theta_s) dF^0(x)},
$$
and $H_t(\varphi_t) = R_t \log G_t(\varphi_t)$. Let $\varphi_t^0 = (\theta_1^0, ..., \theta_{t-1}^0, F^0)$ be the true value of $\varphi_t$. Then

$$E[H_t(\theta_t, \varphi_{t-1}^0)] - E[H_t(\varphi_t^0)] = E \left\{ R_t \log \frac{G_t(\theta_t, \varphi_{t-1}^0)}{G_t(\varphi_t^0)} \right\}$$

$$= \pi_t E \left\{ \log \frac{G_t(\theta_t, \varphi_{t-1}^0)}{G_t(\varphi_t^0)} \right\}$$

$$\leq \pi_t \log E \left\{ \frac{G_t(\theta_t, \varphi_{t-1}^0)}{G_t(\varphi_t^0)} \right\}$$

$$= 0$$

with the equality holds if and only if (since $\pi_t > 0$)

$$\frac{G_t(\theta_t, \varphi_{t-1}^0)}{G_t(\varphi_t^0)} = 1 \text{ a.s.,}$$

which is, by the definition of $G_t$ function, equivalent to that, almost surely,

$$\frac{f_t(Y_t|V_{t-1}, \theta_t)}{f_t(Y_t|V_{t-1}, \varphi_t^0)} = \frac{\int f_t(Y_t|x, Y_1, ..., Y_{t-1}, \theta_t^0) \prod_{s=1}^{t-1} f_s(Y_s|x, Y_1, ..., Y_{s-1}, \theta_s^0)dF^0(x)}{\int f_t(Y_t|x, Y_1, ..., Y_{t-1}, \varphi_t^0) \prod_{s=1}^{t-1} f_s(Y_s|x, Y_1, ..., Y_{s-1}, \theta_s^0)dF^0(x)},$$

a function of $(Y_1, ..., Y_t, \theta_t)$. Therefore, conditions (5) and (6) ensure that

$$E[H_t(\theta_t, \varphi_{t-1}^0)] < E[H_t(\varphi_t^0)].$$

This means that the expectation of the log of the likelihood function in (4) has a unique maximum at $\theta_t = \theta_t^0$. Let $H_t^{(i)}(\varphi_t)$ be defined as $H_t(\varphi_t)$ but based on data from the $i$th subject. Then, the estimator $\hat{\theta}_t$ obtained by maximizing (4) satisfies

$$\sum_{i=1}^{n} H_t^{(i)}(\hat{\theta}_t, \hat{\varphi}_{t-1}) = \max_{\theta_t} \sum_{i=1}^{n} H_t^{(i)}(\theta_t, \hat{\varphi}_{t-1}),$$

where $\hat{\varphi}_{t-1} = (\hat{\theta}_1, ..., \hat{\theta}_{t-1}, \hat{F})$, $\hat{\theta}_1, ..., \hat{\theta}_{t-1}$ are estimators from the previous steps, and $\hat{F}$ is the empirical distribution based on $X^{(j)}$, $j = 1, ..., n$. Under some regularity conditions (such as those given in Theorem 1 of Tang et al., 2003), $\hat{\theta}_t$ converges in probability to the unique maximum point $\theta_t^0$. 
Asymptotic normality of $\hat{\theta}_t$, which is crucial for large sample inference, can be established using a standard argument. Our contribution is to derive an asymptotic representation of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$, which allows us to obtain an easy-to-compute consistent estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$ without knowing its actual form. The asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_t - \theta_t^0)$ is very complicated because of the fact that $\hat{\theta}_t$ is defined in terms of previous estimators $\hat{\theta}_1,...,\hat{\theta}_{t-1}$ and $\hat{F}$. As we discussed in Section 2.1, without using $\hat{\theta}_1,...,\hat{\theta}_{t-1}$ in the estimation of $\theta_t$ may not be computationally feasible.

**Theorem 1.** Assume (2), (3), (5), (6), and the following two conditions.

(C1) The functions $f_t$’s in (3) are continuously twice differentiable with respect to $\theta_t$ and $E \left[ \frac{\partial^2 H_t(\varphi^0_t)}{\partial \theta_t \partial \theta_j'} \right]$ is positive definite.

(C2) There exists an open subset $\Omega_t$ containing $\theta^0_t$ such that

$$\sup_{\theta_t \in \Omega_t} \left\| \frac{\partial^2 H_t(\theta_t, \varphi^0_{t-1})}{\partial \theta_t \partial \theta_j'} \right\| < M_{tj}, \quad j = 1, ..., t,$$

where $M_{tj}$ are integrable functions and $\|A\|^2 = \text{trace}(A'A)$ for a matrix $A$.

Then, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta}_t - \theta^0_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_t(W_t^{(i)}, A_t, \varphi^0_t) + o_p(1) \to_d N(0, \Sigma_t),$$

where $\to_d$ denotes convergence in distribution, $o_p(1)$ denotes a quantity converging to 0 in probability, $\Sigma_t$ is the covariance matrix of $\psi_t(W_t^{(i)}, A_t, \varphi^0_t)$, $W_t^{(i)} = (V_t^{(i)}, R_t^{(i)})$, $i = 1, ..., n$, $A_1 = A_{11}$, $A_t = (A_{t-1}, A_{t1}, ..., A_{tt})$, $t \geq 2$,

$$A_{tj} = E \left[ \frac{\partial^2 H_t(\varphi^0_t)}{\partial \theta_j \partial \theta_j'} \right], \quad j = 1, ..., t,$$

and $\psi_t$ is a known function defined in (10)-(11) of the Appendix, $t = 1, ..., T$.

The functions $\psi_t$, $t = 1, ..., T$, are defined iteratively according to (10)-(11) and, hence, their covariance matrices are very complicated. The explicit forms of $\psi_t$, when $t = 1, 2, 3, 4,$
are given in the Appendix. One may apply the bootstrap method to obtain estimators of \( \Sigma_t \)'s, but in each bootstrap replication, maximizing a bootstrap analog of (1) is required, which results in a very large amount of computation. Instead, we propose the following estimator of \( \Sigma_t \), utilizing the representation in (7). Let \( D_t^{(i)} = \psi_t(W_t^{(i)}, A_t, \varphi_t^0) \). Since \( \Sigma_t = \text{Var}(D_t^{(i)}) \), the sample covariance matrix based on \( D_t^{(1)}, ..., D_t^{(n)} \) is a consistent estimator of \( \Sigma_t \). However, \( D_t^{(i)} \) contains the unknown \( \varphi_t^0 \) and \( A_t \). Substituting \( D_t^{(i)} \) by \( \hat{D}_t^{(i)} = \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t) \), \( i = 1, ..., n \), where \( \hat{A}_t = (\hat{A}_{t-1}, \hat{A}_{t1}, ..., \hat{A}_{tt}) \) and

\[
\hat{A}_{tj} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 H_t^{(i)}(\varphi_t)}{\partial \theta_i \partial \theta_j} \bigg|_{\varphi_t = \hat{\varphi}_t}, \quad j = 1, ..., t,
\]

we define the sample covariance matrix based on \( \hat{D}_t^{(1)}, ..., \hat{D}_t^{(n)} \) as our estimator \( \hat{\Sigma}_t \). This estimator is easy to compute, using (10)-(11) in the Appendix. Under some conditions, \( \hat{\Sigma}_t \) is consistent, which is proved in the Appendix.

**Theorem 2.** Assume that the conditions in Theorem 1 hold and that

(C3) \( \sup_{\|w\| \leq c} \| \psi_t(w, \hat{A}_t, \hat{\varphi}_t) - \psi_t(w, A_t, \varphi_t^0) \| = o_p(1) \) for any \( c > 0 \).

(C4) There exist a constant \( c_0 > 0 \) and a function \( h(w) \geq 0 \) such that \( E[h(W_t^{(1)})] < \infty \) and

\[
P(\| \psi_t(w, \hat{A}_t, \hat{\varphi}_t) \|^2 \leq h(w) \text{ for all } \|w\| \geq c_0) \rightarrow 1.
\]

Then, as \( n \rightarrow \infty \), \( \| \hat{\Sigma}_t - \Sigma_t \| = o_p(1) \).

### 4.4 Some Empirical Results

In this section, we present some results based on a real data set and a simulation study.

#### 4.4.1 Estimation based on HIV-CD4 Data

We applied the proposed method to a longitudinal data set from the study of HIV-AIDS patients with advanced immune suppression, conducted by the AIDS Clinical Trial Group...
Patients were randomized to one of the four daily regimens containing 600mg of zidovudine: zidovudine alternating monthly with 400mg didanosine (Treatment 1), zidovudine plus 2.25mg of zalcitabine (Treatment 2), zidovudine plus 400mg of didanosine (Treatment 3), and zidovudine plus 400mg of didanosine and 400mg of nevirapine (Treatment 4). The data set can be accessed at the following website: “http://biosun1.harvard.edu/~fitzmaur/ala/cd4.txt”.

For the HIV study, the CD4 cell count, which decreases as HIV progresses, is of prime interest. CD4 counts were collected from patients before the treatments were applied (baseline measurements). After the treatments were applied, CD4 counts were collected from each patient every 8 weeks. We considered the first $T = 4$ follow-up time points. For each patient, the $t$th observation is the one closest to week $8t$ in the interval $(8t - 4, 8t + 4]$, $t = 1, 2, 3, 4$. The following is a summary of the number of observed values by time points and treatment.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>$t = 0$</th>
<th>$t = 1$</th>
<th>$t = 2$</th>
<th>$t = 3$</th>
<th>$t = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>320</td>
<td>223</td>
<td>174</td>
<td>127</td>
<td>110</td>
</tr>
<tr>
<td>2</td>
<td>322</td>
<td>218</td>
<td>184</td>
<td>143</td>
<td>110</td>
</tr>
<tr>
<td>3</td>
<td>327</td>
<td>221</td>
<td>184</td>
<td>135</td>
<td>108</td>
</tr>
<tr>
<td>4</td>
<td>330</td>
<td>235</td>
<td>187</td>
<td>136</td>
<td>116</td>
</tr>
<tr>
<td>Total</td>
<td>1299</td>
<td>897</td>
<td>729</td>
<td>541</td>
<td>444</td>
</tr>
</tbody>
</table>

The average dropout proportion for 4 time points after baseline ($t - 0$) are 31.9%, 43.9%, 58.4%, and 66.8%, respectively.

To apply the proposed method, we considered log(CD4+1) at time point $t$ as $Y_t$ and log(baseline measurement +1) as the dropout instrument $Z$. Because the baseline measurements were taken before the treatments were applied, it is reasonable to assume that the dropout propensity at time $t$ does not depend on $Z$ given the CD4 counts at time $1, ..., t$.  

We assumed that

\[
Y_1 = \beta_{10} + \beta_{11}Z + \varepsilon_1 \\
Y_2 = \beta_{20} + \beta_{21}Z + \beta_{22}Y_1 + \varepsilon_2 \\
Y_3 = \beta_{30} + \beta_{31}Z + \beta_{32}Y_1 + \beta_{33}Y_2 + \varepsilon_3 \\
Y_4 = \beta_{40} + \beta_{41}Z + \beta_{42}Y_1 + \beta_{43}Y_2 + \beta_{44}Y_3 + \varepsilon_4,
\]

(8)

where \( \varepsilon_t \sim N(0, \sigma_t^2) \), \( t = 1, \ldots, 4 \), \( \varepsilon_t \)'s are independent, and \( \beta_{ij} \)'s and \( \sigma_t \)'s are unknown parameters.

Tables 1-4 display estimates of parameters based on the HIV-CD4 data under treatments 1-4, respectively, and their standard errors (SE). For each parameter, we computed two estimates, the proposed estimate and the MAR estimate, which was obtained by regression under the missing at random (MAR) assumption. The SE’s of the proposed estimates were calculated using the results in Theorem 2 in Section 3. To compare, we also computed the difference of the MAR estimate and the proposed estimate, its SE, and the two-sided p-value of testing whether two estimates are the same. The SE’s of the MAR estimates and the differences were computed by bootstrapping.

It can be seen from Tables 1-4 that the differences between the MAR and proposed estimates are not negligible in some cases (p-values less or nearly equal to 5%) while in some cases the two estimates are about the same.

### 4.4.2 A Simulation Study

A simulation study was conducted under model (8) with \( n = 300 \) and parameters equal to the estimated values under Treatment 3 in the HIV-CD4 example. These values are shown in Table 5. The covariate \( Z \) was generated from \( N(2.9065, 0.9544^2) \), where the parameters are the estimates of the baseline CD4. The dropout indicators at time points \( t = 1, 2, 3, 4 \)
were generated from the following logistic model:

\[
P(R_1 = 1 | Z, Y_1) = 1 - \frac{1}{1 + \exp(-9 + 4Y_1)},
\]
\[
P(R_2 = 1 | Z, Y_1, Y_2, R_1 = 1) = 1 - \frac{1}{1 + \exp(-14 + Y_1 + 5Y_2)},
\]
\[
P(R_3 = 1 | Z, Y_1, Y_2, Y_3, R_2 = 1) = 1 - \frac{1}{1 + \exp(-18 + 2Y_2 + 4Y_3)},
\]
\[
P(R_4 = 1 | Z, Y_1, Y_2, Y_3, Y_4, R_3 = 1) = 1 - \frac{1}{1 + \exp(-15 + 2Y_3 + 3Y_4)}.
\]

The parameters in (9) were chosen so that the unconditional dropout rates are similar to the observed dropout proportions under Treatment 3 of the HIV-CD4 data, approximately 30%, 40%, 60%, and 70% for \( t = 1, 2, 3, \) and 4, respectively.

We studied the method based on the MAR assumption and the proposed method. As a standard, we also include the standard regression method when there is no dropout.

Based on 1000 simulation runs, Table 5 reports the bias for parameter estimation, standard deviation (SD) of the parameter estimate, standard error (SE), which is an estimate of SD, and the coverage probability (CP) of the approximate 95% confidence intervals of the parameter, using estimate ±1.96SE. Again, the SE’s for the proposed method are obtained using Theorem 2 and the SE’s for the MAR method are computed by bootstrapping. The results in Table 5 show that the proposed estimators and their SE’s work well, and the method based on the MAR assumption produces biased estimators and the biases are large enough to result in poor CP. The SD’s of the MAR estimators, however, may be smaller than those of the proposed estimators. Hence, the MAR estimators may be more efficient when they are nearly unbiased, e.g., when the MAR assumption holds.

### 4.5 Discussion on Assumptions

The key assumptions for our approach are (2) and (3). As we discussed in Section 1, to identify the unknown parameters, it is necessary that at least one component of \( V_I = \)
(Y_1, ..., Y_t, X) is not related to dropout at time point t, conditioned on the other components. This component is Y_t under the MAR dropout assumption, whereas it is a component Z of X under our assumption \(^2\). Unfortunately, none of these assumptions on the dropout mechanism can be checked using data due to the presence of missing values. We have to carefully study each particular problem and decide which assumption is reasonable or approximately holds. In the HIV-CD4 problem, for example, the difference between the MAR method and our approach is whether the current response Y_t or the baseline response Z is related to dropout at time point t, given the other values. Since Y_t is a more recent value for each patient at time point t, we think that our assumption is more reasonable.

It is important to develop estimation methods under various assumptions on the dropout mechanism. The results will be useful as different tools for application and/or for a sensitivity analysis under different assumptions.

We also need to assume at least one of p(Y|X) and p(R|Y, X) is parametric to be able to identify parameters. Again, with missing data, we are not able to verify a parametric model such \(^3\) using observed data. This is because the parametric model is imposed on the density p(Y_t|X, Y_1, ..., Y_{t-1}), which is a mixture of p(Y_t|X, Y_1, ..., Y_{t-1}, R_t = 1) and p(Y_t|X, Y_1, ..., Y_{t-1}, R_t = 0), and we are not able to check a parametric model assumption on p(Y_t|X, Y_1, ..., Y_{t-1}, R_t = 0) since no Y_t-observation comes from it. Parametric models may be sensitive to model violations. The same issue exists for the likelihood approach in Little and Rubin (2002) under ignorable nonresponse. The robustness of the proposed method against violation of assumption \(^3\) is under further investigation.
4.6 Proofs of Theorems

Proof of Theorem 1. Let

\[ l_t(\theta_t, \hat{\theta}_{t-1}) = \frac{1}{n} \sum_{i=1}^{n} H_{t}^{(i)}(\theta_t, \hat{\theta}_{t-1}) \]

and \( \nabla l_t(\theta_t, \hat{\theta}_{t-1}) \) be its derivative with respect to \( \theta_t \). We first prove the case of \( t = 1 \). By Taylor’s expansion and the fact that \( \hat{\theta}_0 = \hat{F} \) and \( \nabla l_1(\hat{\theta}_1, \hat{F}) = 0 \), we have

\[-\nabla l_1(\theta_1, F^0) = \nabla l_1(\theta_1, \hat{F}) - \nabla l_1(\theta_1, F^0) + \nabla l_1(\hat{\theta}_1, \hat{F}) - \nabla l_1(\theta_1, \hat{F}) = \nabla l_1(\theta_1, \hat{F}) - \nabla l_1(\theta_1, F^0) + \nabla^2 l_1(\theta_1, \hat{F})(\hat{\theta}_1 - \theta_1) + o_p(n^{-1/2}),\]

where \( \nabla^2 \) is the second order derivative with respect to \( \theta_1 \) and \( o_p(n^{-1/2}) = n^{-1/2}o_p(1) \). Note that

\[ \nabla l_1(\theta_1, \hat{F}) - \nabla l_1(\theta_1, F^0) = \frac{1}{n} \sum_{i=1}^{n} R_1^{(i)} \left\{ \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d\hat{F} \int f_1(Y_1^{(i)}|x, \theta_1^0)d(\hat{F} - F^0)}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF} \right\} \]

\[ \quad - \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d(\hat{F} - F^0)}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} R_1^{(i)} \left\{ \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d\hat{F} \int f_1(Y_1^{(i)}|x, \theta_1^0)d(\hat{F} - F^0)}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0} \right\} \]

\[ \quad - \frac{\int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)d(\hat{F} - F^0)}{\int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0} \]

\[ + o_p(n^{-1/2}). \]

Let

\[ g_1(Y_1^{(i)}, \varphi_1^0) = \int f_1(Y_1^{(i)}|x, \theta_1^0)dF^0 \quad \text{and} \quad \nabla g_1(Y_1^{(i)}, \varphi_1^0) = \int \nabla f_1(Y_1^{(i)}|x, \theta_1^0)dF^0. \]

Then

\[ \nabla l_1(\theta_1, \hat{F}) - \nabla l_1(\theta_1, F^0) = B_{n1} + o_p(n^{-1/2}), \]

where

\[ B_{n1} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \frac{R_1^{(i)} \nabla g_1(Y_1^{(i)}, \varphi_1^0) f_1(Y_1^{(i)}|X^{(j)}, \theta_1^0)}{[g_1(Y_1^{(i)}, \varphi_1^0)]^2} - \frac{R_1^{(i)} \nabla f_1(Y_1^{(i)}|X^{(j)}, \theta_1^0)}{g_1(Y_1^{(i)}, \varphi_1^0)} \right\} \]

\[ = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h_1(W_1^{(i)}, W_1^{(j)}) \]
is a V-statistic with the following kernel:

\[
\begin{align*}
    h_1(W_1^{(i)}, W_1^{(j)}) &= \frac{1}{2} \left\{ \frac{R_1^{(i)} \nabla g_1(Y_1^{(i)}, \varphi_0) f_1(Y_1^{(i)} | X^{(j)}, \theta_1^0)}{[g_1(Y_1^{(i)}, \varphi_0)]^2} - \frac{R_1^{(j)} \nabla f_1(Y_1^{(i)} | X^{(j)}, \theta_1^0)}{g_1(Y_1^{(i)}, \varphi_0)} \right. \\
    &\quad + \left. \frac{R_1^{(j)} \nabla g_1(Y_1^{(j)}, \varphi_0) f_1(Y_1^{(j)} | X^{(i)}, \theta_1^0)}{[g_1(Y_1^{(j)}, \varphi_0)]^2} - \frac{R_1^{(i)} \nabla f_1(Y_1^{(j)} | X^{(i)}, \theta_1^0)}{g_1(Y_1^{(j)}, \varphi_0)} \right\}.
\end{align*}
\]

Let

\[
    h_{11}(W_1^{(i)}) = E[h_1(W_1^{(i)}, W_1^{(j)} | W_1^{(i)})] = \frac{\pi_1}{2} E \left\{ \frac{\nabla g_1(Y_1^{(j)}, \varphi_1) f_1(Y_1^{(j)} | X^{(i)}, \theta_1^0)}{[g_1(Y_1^{(j)}, \varphi_1)]^2} - \frac{\nabla f_1(Y_1^{(j)} | X^{(i)}, \theta_1^0)}{g_1(Y_1^{(j)}, \varphi_1)} \right| R_1^{(i)} = 1, X^{(i)} \right\},
\]

which is a function of \( \varphi_1^0 \) and \( X^{(i)} \), not depending on \( Y_1^{(i)} \) or \( R_1^{(i)} \). We denote this function as \( h_{11}(X^{(i)}, \varphi_1^0) \). From the theory of V-statistics,

\[
    B_{n1} = \frac{1}{n} \sum_{i=1}^{n} 2h_{11}(X^{(i)}, \theta_1^0) + o_p(n^{-1/2}).
\]

Under the given regularity conditions, \( \nabla^2 l_1(\theta_1^0, \hat{F}) - A_{11} = o_p(1) \). Therefore,

\[
    \sqrt{n}(\hat{\theta}_1 - \theta_1^0) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_{11}^{-1} \left\{ \frac{\partial H_1^{(i)}(\theta_1^0, F^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\} + o_p(1).
\]

Thus, result (7) with \( t = 1 \) follows by letting

\[
    \psi_1(W_1^{(i)}, A_1, \varphi_1^0) = -A_{11}^{-1} \left\{ \frac{\partial H_1^{(i)}(\theta_1^0, F^0)}{\partial \theta_1} + 2h_{11}(X^{(i)}, \varphi_1^0) \right\}.
\]
Now, suppose that we have obtained result (7) for \( \hat{\theta}_1, \ldots, \hat{\theta}_{t-1} \). Let’s prove (7) for \( \hat{\theta}_t \). Note that

\[
-\nabla l_t(\phi_t^0) = \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \theta_1^0, \hat{F}) - \nabla l_t(\phi_t^0) \\
+ \nabla l_t(\theta_t^0, \hat{\phi}_{t-1}) - \nabla l_t(\theta_t^0, \theta_{t-1}^0, \hat{\phi}_{t-2}) \\
+ \cdots \\
+ \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \hat{\theta}_1, \hat{F}) - \nabla l_t(\theta_t^0, \ldots, \theta_2^0, \theta_1^0, \hat{F}) \\
+ \nabla l_t(\hat{\phi}_t) - \nabla l_t(\theta_t^0, \hat{\phi}_{t-1}) \\
= B_{nt} + \nabla^2 l_t(\theta_t^0, \hat{\phi}_{t-1})(\hat{\theta}_t - \theta_t^0) \\
+ \nabla^2 l_{t(t-1)}(\theta_t^0, \theta_{t-1}^0, \hat{\phi}_{t-2})(\hat{\theta}_{t-1} - \theta_{t-1}^0) \\
+ \cdots \\
+ \nabla^2 l_{1t}(\theta_t^0, \ldots, \theta_1^0, \hat{F})(\hat{\theta}_1 - \theta_1^0) + o_p(n^{-1/2}),
\]

where \( \nabla^2 l_{ij} \) is the second order derivative with respect to \( \theta_t \) and \( \theta_j, j = 1, \ldots, t \). Similar to the case of \( t = 1 \), we can show that

\[
B_{nt} = \frac{1}{n} \sum_{i=1}^n 2h_{1t}(X^{(i)}, \phi_t^0) + o_p(n^{-1/2}),
\]

where

\[
h_{1t}(X^{(i)}, \phi_t^0) = \frac{\pi_t}{2} E \left\{ \frac{\nabla g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0) \prod_{s=1}^t f_s(Y_s^{(j)}|Y_{s-1}^{(j)}, \theta_s^0)}{[g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0)]^2} \left| R_1^{(i)} = 1, X^{(i)} \right\} \\
- \frac{\pi_t}{2} E \left\{ \frac{\nabla f_t(Y_t^{(j)}|Y_{t-1}^{(j)}, \theta_t^0) \prod_{s=1}^{t-1} f_s(Y_s^{(j)}|Y_{s-1}^{(j)}, \theta_s^0)}{g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0)} \left| R_1^{(i)} = 1, X^{(i)} \right\},
\]

\[
g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0) = \int \prod_{s=1}^t f_s(Y_s^{(j)}|x, Y_1^{(j)}, \ldots, Y_{s-1}^{(j)}, \theta_s^0) dF^0,
\]

and

\[
\nabla g_t(Y_t^{(j)}, \ldots, Y_1^{(j)}, \varphi_t^0) = \int \nabla f_t(Y_t^{(j)}|x, Y_1^{(j)}, \ldots, Y_{t-1}^{(j)}, \theta_t^0) \prod_{s=1}^{t-1} f_s(Y_s^{(j)}|x, Y_1^{(j)}, \ldots, Y_{s-1}^{(j)}, \theta_s^0) dF^0.
\]
Under the given regularity conditions, $\nabla^2 \eta(t_1, \ldots, t_j, \varphi_{j-1}) - A_{ij} = o_p(1)$. Then

$$\sqrt{n}(\hat{\theta}_t - \theta^0_t) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_{it}^{-1} \braces{\frac{\partial H^{(i)}(\varphi^0_t)}{\partial \theta_t} + 2h_{1t}(X^i, \varphi^0_t) + \sum_{j=1}^{t-1} A_{tj} \psi_j(W^{(i)}_j, A_j, \varphi^0_j)} + o_p(1)$$

and result (7) holds with the following iteratively defined $\psi_t$: $\psi_1$ is given by (10); having

$\psi_1, \ldots, \psi_{t-1}$, $\psi_t$ is defined as

$$\psi_t(W^{(i)}_t, A_t, \varphi^0_t) = -A_{tt}^{-1} \braces{\frac{\partial H^{(i)}(\varphi^0_t)}{\partial \theta_t} + 2h_{1t}(X^i, \varphi^0_t) + \sum_{j=1}^{t-1} A_{tj} \psi_j(W^{(i)}_j, A_j, \varphi^0_j)}.$$  (11)

The explicit forms of $\psi_t$, $t = 1, 2, 3, 4$, are shown as follows:

$$\psi_1(W^{(i)}_1, A_1, \varphi^0_1) = -A_{11}^{-1} \braces{\frac{\partial H_1^{(i)}(\varphi^0_1)}{\partial \theta_1} + 2h_{11}(X^i, \varphi^0_1)},$$

$$\psi_2(W^{(i)}_2, A_2, \varphi^0_2) = -A_{22}^{-1} \braces{\frac{\partial H_2^{(i)}(\varphi^0_2)}{\partial \theta_2} + 2h_{12}(X^i, \varphi^0_2)},$$

$$\psi_3(W^{(i)}_3, A_3, \varphi^0_3) = -A_{33}^{-1} \braces{\frac{\partial H_3^{(i)}(\varphi^0_3)}{\partial \theta_3} + 2h_{13}(X^i, \varphi^0_3)},$$

$$\psi_4(W^{(i)}_4, A_4, \varphi^0_4) = -A_{44}^{-1} \braces{\frac{\partial H_4^{(i)}(\varphi^0_4)}{\partial \theta_4} + 2h_{14}(X^i, \varphi^0_4)},$$

$$+A_{22}^{-1} A_{21} A_{11}^{-1} \braces{\frac{\partial H_1^{(i)}(\varphi^0_1)}{\partial \theta_1} + 2h_{11}(X^i, \varphi^0_1)},$$

$$+A_{33}^{-1} A_{32} A_{22}^{-1} \braces{\frac{\partial H_2^{(i)}(\varphi^0_2)}{\partial \theta_2} + 2h_{12}(X^i, \varphi^0_2)},$$

$$+(-A_{33}^{-1} A_{32} A_{22}^{-1} A_{21} A_{11}^{-1} + A_{33}^{-1} A_{32} A_{11}^{-1}) \braces{\frac{\partial H_1^{(i)}(\varphi^0_1)}{\partial \theta_1} + 2h_{11}(X^i, \varphi^0_1)},$$

$$+(-A_{44}^{-1} A_{43} A_{33}^{-1} A_{32} A_{22}^{-1} + A_{44}^{-1} A_{42} A_{22}^{-1} \braces{\frac{\partial H_2^{(i)}(\varphi^0_2)}{\partial \theta_2} + 2h_{12}(X^i, \varphi^0_2)},$$

$$+(A_{44}^{-1} A_{43} A_{33}^{-1} A_{32} A_{22}^{-1} A_{21} A_{11}^{-1} - A_{44}^{-1} A_{42} A_{22}^{-1} A_{21} A_{11}^{-1} + A_{44}^{-1} A_{41} A_{11}^{-1}) \braces{\frac{\partial H_1^{(i)}(\varphi^0_1)}{\partial \theta_1} + 2h_{11}(X^i, \varphi^0_1)}.$$
Proof of Theorem 2. Note that

\[ \Sigma_t = \text{Var}(D_t^{(i)}) = \int \psi_t(w, A_t, \varphi_t^0) \psi_t(w, A_t, \varphi_t^0)^\tau dP(w) \]

and

\[ \hat{\Sigma}_t = \int \psi_t(w, \hat{A}_t, \hat{\varphi}_t) \psi_t(w, \hat{A}_t, \hat{\varphi}_t)^\tau dP_n(w), \]

where \( P(w) \) denotes the underlying true distribution of \( W_t \) and \( P_n(w) \) denotes its empirical distribution based on data \( W_t^{(j)}, j = 1, ..., n \). Note that \( \| \hat{\Sigma}_t - \Sigma_t \| \) is bounded by

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t) \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t)^\tau - \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, A_t, \varphi_t^0) \psi_t(W_t^{(i)}, A_t, \varphi_t^0)^\tau \\
+ \frac{1}{n} \sum_{i=1}^{n} \psi_t(W_t^{(i)}, A_t, \varphi_t^0) \psi_t(W_t^{(i)}, A_t, \varphi_t^0)^\tau - \int \psi_t(w, A_t, \varphi_t^0) \psi_t(w, A_t, \varphi_t^0)^\tau dP(w) \right\|
\]

\[
\leq \frac{1}{n} \sum_{i=1}^{n} \| Q_{n,t} \| I_{[0,c]}(\| W_t^{(i)} \|) + \frac{1}{n} \sum_{i=1}^{n} \| Q_{n,t} \| I_{(c,\infty)}(\| W_t^{(i)} \|) + o_p(1),
\]

where the inequality follows from the triangular inequality and law of large numbers, and

\[ Q_{n,t} = \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t) \psi_t(W_t^{(i)}, \hat{A}_t, \hat{\varphi}_t)^\tau - \psi_t(W_t^{(i)}, A_t, \varphi_t^0) \psi_t(W_t^{(i)}, A_t, \varphi_t^0)^\tau. \]

By condition (C3), for any \( \epsilon > 0 \),

\[
\left( \frac{1}{n} \sum_{i=1}^{n} \| Q_{n,t} \| I_{[0,c]}(\| W_t^{(i)} \|) \right) < \epsilon/2
\]

when \( n \) is sufficiently large. For any \( \tilde{\epsilon} > 0 \), we can choose \( c \) such that \( E[h(w)I_{(c,\infty)}(\|w\|)] < \epsilon \tilde{\epsilon}/4 \). By Chebyshev’s inequality and condition (C4), we have

\[
P \left( \frac{1}{n} \sum_{i=1}^{n} \| Q_{n,t} \| I_{(c,\infty)}(\| W_t^{(i)} \|) > \epsilon/2 \right) < \tilde{\epsilon}.
\]

Then,

\[
P \left( \frac{1}{n} \sum_{i=1}^{n} \| Q_{n,t} \| > \epsilon \right) \to 0.
\]

This proves that \( \| \hat{\Sigma}_t - \Sigma_t \| = o_p(1) \).
<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\sigma_{1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.1265</td>
<td>0.9128</td>
<td>0.6955</td>
</tr>
<tr>
<td>SE</td>
<td>0.2028</td>
<td>0.0618</td>
<td>0.0500</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>-0.1089</td>
<td>0.9936</td>
<td>0.7406</td>
</tr>
<tr>
<td>SE</td>
<td>0.2483</td>
<td>0.0792</td>
<td>0.0593</td>
</tr>
<tr>
<td>Difference</td>
<td>0.2354</td>
<td>-0.0808</td>
<td>-0.0451</td>
</tr>
<tr>
<td>SE</td>
<td>0.1322</td>
<td>0.0476</td>
<td>0.0229</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0750</td>
<td>0.0896</td>
<td>0.0489</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
<th>$\beta_{22}$</th>
<th>$\sigma_{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.1705</td>
<td>0.5198</td>
<td>0.3525</td>
<td>0.6767</td>
</tr>
<tr>
<td>SE</td>
<td>0.2430</td>
<td>0.1012</td>
<td>0.0886</td>
<td>0.0478</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.0007</td>
<td>0.7605</td>
<td>0.1304</td>
<td>0.7979</td>
</tr>
<tr>
<td>SE</td>
<td>0.3047</td>
<td>0.1529</td>
<td>0.1322</td>
<td>0.0938</td>
</tr>
<tr>
<td>Difference</td>
<td>0.1698</td>
<td>-0.2407</td>
<td>0.2219</td>
<td>-0.1212</td>
</tr>
<tr>
<td>SE</td>
<td>0.1860</td>
<td>0.0855</td>
<td>0.0742</td>
<td>0.0394</td>
</tr>
<tr>
<td>p-value</td>
<td>0.3613</td>
<td>0.0049</td>
<td>0.0028</td>
<td>0.0021</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{30}$</th>
<th>$\beta_{31}$</th>
<th>$\beta_{32}$</th>
<th>$\beta_{33}$</th>
<th>$\sigma_{3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.2536</td>
<td>0.2029</td>
<td>0.2896</td>
<td>0.3235</td>
<td>0.5701</td>
</tr>
<tr>
<td>SE</td>
<td>0.2230</td>
<td>0.0928</td>
<td>0.1105</td>
<td>0.0863</td>
<td>0.0500</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.1638</td>
<td>0.0409</td>
<td>0.4866</td>
<td>0.3644</td>
<td>0.1987</td>
</tr>
<tr>
<td>SE</td>
<td>0.3061</td>
<td>0.1027</td>
<td>0.1882</td>
<td>0.1326</td>
<td>0.1008</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0898</td>
<td>0.1620</td>
<td>-0.1970</td>
<td>-0.0409</td>
<td>0.3714</td>
</tr>
<tr>
<td>SE</td>
<td>0.1645</td>
<td>0.0707</td>
<td>0.1066</td>
<td>0.0937</td>
<td>0.0622</td>
</tr>
<tr>
<td>p-value</td>
<td>0.5851</td>
<td>0.0219</td>
<td>0.0646</td>
<td>0.6625</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{40}$</th>
<th>$\beta_{41}$</th>
<th>$\beta_{42}$</th>
<th>$\beta_{43}$</th>
<th>$\beta_{44}$</th>
<th>$\sigma_{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.4823</td>
<td>-0.0839</td>
<td>-0.0476</td>
<td>0.4890</td>
<td>0.3799</td>
<td>0.6482</td>
</tr>
<tr>
<td>SE</td>
<td>0.2555</td>
<td>0.1355</td>
<td>0.1029</td>
<td>0.1570</td>
<td>0.1720</td>
<td>0.0670</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.5038</td>
<td>-0.1091</td>
<td>-0.0521</td>
<td>0.5406</td>
<td>0.3201</td>
<td>0.6330</td>
</tr>
<tr>
<td>SE</td>
<td>0.3119</td>
<td>0.1235</td>
<td>0.1147</td>
<td>0.1764</td>
<td>0.1539</td>
<td>0.1064</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.0215</td>
<td>0.0252</td>
<td>0.0045</td>
<td>-0.0516</td>
<td>0.0598</td>
<td>0.0152</td>
</tr>
<tr>
<td>SE</td>
<td>0.1718</td>
<td>0.0949</td>
<td>0.0743</td>
<td>0.1175</td>
<td>0.1192</td>
<td>0.0474</td>
</tr>
<tr>
<td>p-value</td>
<td>0.9004</td>
<td>0.7906</td>
<td>0.9517</td>
<td>0.6606</td>
<td>0.6159</td>
<td>0.7485</td>
</tr>
</tbody>
</table>

Table 4.1: Estimates and SE’s under treatment 1 of the HIV-AIDS study
Table 4.2: Estimates and SE’s under treatment 2 of the HIV-AIDS study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\sigma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.5862</td>
<td>0.8088</td>
<td>0.8121</td>
</tr>
<tr>
<td>SE</td>
<td>0.2372</td>
<td>0.0755</td>
<td>0.0615</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.2782</td>
<td>0.9275</td>
<td>0.8737</td>
</tr>
<tr>
<td>SE</td>
<td>0.2847</td>
<td>0.0887</td>
<td>0.0642</td>
</tr>
<tr>
<td>Difference</td>
<td>0.3080</td>
<td>-0.1187</td>
<td>-0.0616</td>
</tr>
<tr>
<td>SE</td>
<td>0.1777</td>
<td>0.0658</td>
<td>0.0391</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0830</td>
<td>0.0712</td>
<td>0.1152</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
<th>$\beta_{22}$</th>
<th>$\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.1752</td>
<td>0.4091</td>
<td>0.4610</td>
<td>0.7026</td>
</tr>
<tr>
<td>SE</td>
<td>0.2145</td>
<td>0.0944</td>
<td>0.0834</td>
<td>0.0518</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.5213</td>
<td>0.3212</td>
<td>0.4283</td>
<td>0.5784</td>
</tr>
<tr>
<td>SE</td>
<td>0.2552</td>
<td>0.1647</td>
<td>0.1184</td>
<td>0.0516</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.3461</td>
<td>0.0879</td>
<td>0.0327</td>
<td>0.1242</td>
</tr>
<tr>
<td>SE</td>
<td>0.1515</td>
<td>0.0724</td>
<td>0.0691</td>
<td>0.0343</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0223</td>
<td>0.2247</td>
<td>0.6361</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{30}$</th>
<th>$\beta_{31}$</th>
<th>$\beta_{32}$</th>
<th>$\beta_{33}$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>-0.0833</td>
<td>0.0875</td>
<td>0.3437</td>
<td>0.4820</td>
<td>0.7429</td>
</tr>
<tr>
<td>SE</td>
<td>0.2387</td>
<td>0.1326</td>
<td>0.1225</td>
<td>0.1099</td>
<td>0.0604</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>-0.0974</td>
<td>-0.1068</td>
<td>0.2943</td>
<td>0.6861</td>
<td>0.1524</td>
</tr>
<tr>
<td>SE</td>
<td>0.3191</td>
<td>0.1749</td>
<td>0.0911</td>
<td>0.1148</td>
<td>0.0736</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0141</td>
<td>0.1943</td>
<td>0.0494</td>
<td>-0.2041</td>
<td>0.5905</td>
</tr>
<tr>
<td>SE</td>
<td>0.1684</td>
<td>0.1066</td>
<td>0.0816</td>
<td>0.0933</td>
<td>0.0572</td>
</tr>
<tr>
<td>p-value</td>
<td>0.9333</td>
<td>0.0683</td>
<td>0.0449</td>
<td>0.0287</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{40}$</th>
<th>$\beta_{41}$</th>
<th>$\beta_{42}$</th>
<th>$\beta_{43}$</th>
<th>$\beta_{44}$</th>
<th>$\sigma_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>-0.0855</td>
<td>0.2793</td>
<td>-0.1456</td>
<td>0.3189</td>
<td>0.5010</td>
<td>0.6635</td>
</tr>
<tr>
<td>SE</td>
<td>0.2847</td>
<td>0.1205</td>
<td>0.0890</td>
<td>0.1128</td>
<td>0.1414</td>
<td>0.0725</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>-0.0938</td>
<td>0.3149</td>
<td>-0.2342</td>
<td>0.2877</td>
<td>0.6521</td>
<td>0.7789</td>
</tr>
<tr>
<td>SE</td>
<td>0.3113</td>
<td>0.1139</td>
<td>0.1053</td>
<td>0.1112</td>
<td>0.1382</td>
<td>0.0837</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0083</td>
<td>-0.0356</td>
<td>0.0886</td>
<td>0.0312</td>
<td>-0.1511</td>
<td>-0.1154</td>
</tr>
<tr>
<td>SE</td>
<td>0.1216</td>
<td>0.0853</td>
<td>0.0683</td>
<td>0.0957</td>
<td>0.1313</td>
<td>0.0612</td>
</tr>
<tr>
<td>p-value</td>
<td>0.9456</td>
<td>0.6764</td>
<td>0.1946</td>
<td>0.7444</td>
<td>0.2498</td>
<td>0.0593</td>
</tr>
</tbody>
</table>
Table 4.3: Estimates and SE’s under treatment 3 of the HIV-AIDS study

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{10}$</th>
<th>$\beta_{11}$</th>
<th>$\sigma_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.8357</td>
<td>0.7709</td>
<td>0.9397</td>
</tr>
<tr>
<td>SE</td>
<td>0.3050</td>
<td>0.0945</td>
<td>0.0579</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.5549</td>
<td>0.8588</td>
<td>1.0146</td>
</tr>
<tr>
<td>SE</td>
<td>0.4009</td>
<td>0.1256</td>
<td>0.0831</td>
</tr>
<tr>
<td>Difference</td>
<td>0.2808</td>
<td>-0.0879</td>
<td>-0.0749</td>
</tr>
<tr>
<td>SE</td>
<td>0.2323</td>
<td>0.0807</td>
<td>0.0535</td>
</tr>
<tr>
<td>p-value</td>
<td>0.2267</td>
<td>0.2761</td>
<td>0.1615</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{20}$</th>
<th>$\beta_{21}$</th>
<th>$\beta_{22}$</th>
<th>$\sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.0242</td>
<td>0.3468</td>
<td>0.6094</td>
<td>0.7503</td>
</tr>
<tr>
<td>SE</td>
<td>0.2841</td>
<td>0.1068</td>
<td>0.0791</td>
<td>0.0571</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.0065</td>
<td>0.4208</td>
<td>0.5472</td>
<td>0.8023</td>
</tr>
<tr>
<td>SE</td>
<td>0.3257</td>
<td>0.1548</td>
<td>0.1503</td>
<td>0.1821</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0177</td>
<td>-0.0740</td>
<td>-0.0520</td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.1515</td>
<td>0.0961</td>
<td>0.0582</td>
<td>0.0563</td>
</tr>
<tr>
<td>p-value</td>
<td>0.9070</td>
<td>0.4413</td>
<td>0.2852</td>
<td>0.3557</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{30}$</th>
<th>$\beta_{31}$</th>
<th>$\beta_{32}$</th>
<th>$\beta_{33}$</th>
<th>$\sigma_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>0.0440</td>
<td>0.0957</td>
<td>0.1501</td>
<td>0.6865</td>
<td>0.6315</td>
</tr>
<tr>
<td>SE</td>
<td>0.2045</td>
<td>0.0924</td>
<td>0.0870</td>
<td>0.1425</td>
<td>0.0670</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>0.3871</td>
<td>0.1722</td>
<td>0.0785</td>
<td>0.5883</td>
<td>0.9411</td>
</tr>
<tr>
<td>SE</td>
<td>0.3584</td>
<td>0.1342</td>
<td>0.1181</td>
<td>0.1329</td>
<td>0.0731</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.3431</td>
<td>-0.0765</td>
<td>-0.0716</td>
<td>-0.0982</td>
<td>-0.3096</td>
</tr>
<tr>
<td>SE</td>
<td>0.1622</td>
<td>0.0865</td>
<td>0.0683</td>
<td>0.1237</td>
<td>0.0320</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0344</td>
<td>0.3765</td>
<td>0.2945</td>
<td>0.4273</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\beta_{40}$</th>
<th>$\beta_{41}$</th>
<th>$\beta_{42}$</th>
<th>$\beta_{43}$</th>
<th>$\beta_{44}$</th>
<th>$\sigma_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR estimate</td>
<td>-0.3313</td>
<td>0.0966</td>
<td>0.0954</td>
<td>0.3468</td>
<td>0.4614</td>
<td>0.6296</td>
</tr>
<tr>
<td>SE</td>
<td>0.2239</td>
<td>0.1288</td>
<td>0.1288</td>
<td>0.1204</td>
<td>0.1126</td>
<td>0.0680</td>
</tr>
<tr>
<td>Proposed estimate</td>
<td>-0.3493</td>
<td>0.1104</td>
<td>0.0891</td>
<td>0.3546</td>
<td>0.4505</td>
<td>0.6007</td>
</tr>
<tr>
<td>SE</td>
<td>0.2488</td>
<td>0.1276</td>
<td>0.1295</td>
<td>0.0919</td>
<td>0.1284</td>
<td>0.0641</td>
</tr>
<tr>
<td>Difference</td>
<td>0.0180</td>
<td>-0.0138</td>
<td>0.0063</td>
<td>-0.0078</td>
<td>0.0109</td>
<td>0.0229</td>
</tr>
<tr>
<td>SE</td>
<td>0.1686</td>
<td>0.1094</td>
<td>0.1135</td>
<td>0.0677</td>
<td>0.1092</td>
<td>0.0581</td>
</tr>
<tr>
<td>p-value</td>
<td>0.9150</td>
<td>0.8996</td>
<td>0.9557</td>
<td>0.9083</td>
<td>0.9205</td>
<td>0.6935</td>
</tr>
</tbody>
</table>
Table 4.4: Estimates and SE’s under treatment 4 of the HIV-AIDS study
| β_{10} = 0.5549 | -0.0058 | 0.1858 | 0.1870 | 94.7 | 1.2224 | 0.2158 | 0.2087 | 0.0 | -0.0294 | 0.3615 | 0.3532 | 95.6 |
| β_{11} = 0.8588 | 0.0018 | 0.0613 | 0.0612 | 94.2 | -0.2751 | 0.0690 | 0.0670 | 2.0 | 0.0076 | 0.1012 | 0.1011 | 96.2 |
| σ_1 = 1.0416 | -0.0011 | 0.0413 | 0.0408 | 94.0 | -1.1808 | 0.0405 | 0.0406 | 1.3 | -0.0003 | 0.0979 | 0.0924 | 93.7 |
| β_{20} = 0.0065 | 0.0026 | 0.1495 | 0.1502 | 94.4 | 0.8804 | 0.2487 | 0.2525 | 7.9 | -0.0028 | 0.3495 | 0.3544 | 96.5 |
| β_{21} = 0.4208 | -0.0014 | 0.0635 | 0.0621 | 94.4 | -0.0814 | 0.0750 | 0.0739 | 78.3 | 0.0034 | 0.1382 | 0.1540 | 97.0 |
| β_{22} = 0.5472 | 0.0009 | 0.0475 | 0.0456 | 93.9 | -0.1184 | 0.0676 | 0.0659 | 55.7 | -0.0023 | 0.1134 | 0.1233 | 97.1 |
| σ_2 = 0.8023 | -0.0019 | 0.0326 | 0.0323 | 93.6 | -0.0834 | 0.0392 | 0.0376 | 41.3 | -0.0126 | 0.1468 | 0.1435 | 95.9 |
| β_{30} = 0.3871 | 0.0004 | 0.1791 | 0.1771 | 94.2 | 1.6413 | 0.4108 | 0.4100 | 3.5 | 0.0031 | 0.3228 | 0.3371 | 95.4 |
| β_{31} = 0.1722 | -0.0007 | 0.0768 | 0.0786 | 94.8 | -0.0476 | 0.1078 | 0.1075 | 91.3 | -0.0101 | 0.1098 | 0.1072 | 94.9 |
| β_{32} = 0.0785 | 0.0007 | 0.0637 | 0.0654 | 94.9 | -0.0071 | 0.1016 | 0.0978 | 93.5 | -0.0073 | 0.1177 | 0.1163 | 93.8 |
| β_{33} = 0.5883 | 0.0000 | 0.0662 | 0.0678 | 95.3 | -0.2809 | 0.1140 | 0.1093 | 29.5 | -0.0064 | 0.1313 | 0.1329 | 96.2 |
| σ_3 = 0.9411 | 0.0001 | 0.0377 | 0.0380 | 94.6 | -0.1416 | 0.0538 | 0.0521 | 26.4 | 0.0121 | 0.0771 | 0.0808 | 94.6 |
| β_{40} = -0.3493 | 0.0024 | 0.1109 | 0.1135 | 95.7 | 0.6963 | 0.3892 | 0.3874 | 55.5 | -0.0123 | 0.2425 | 0.2527 | 95.7 |
| β_{41} = 0.1104 | 0.0004 | 0.0516 | 0.0508 | 94.4 | -0.0150 | 0.0882 | 0.0874 | 93.0 | -0.0078 | 0.1313 | 0.1338 | 94.5 |
| β_{42} = 0.0891 | -0.0003 | 0.0429 | 0.0418 | 94.0 | -0.0097 | 0.0744 | 0.0768 | 95.4 | 0.0082 | 0.1214 | 0.1313 | 94.8 |
| β_{43} = 0.3546 | -0.0001 | 0.0485 | 0.0485 | 94.8 | -0.0217 | 0.0834 | 0.0872 | 95.0 | -0.0076 | 0.0790 | 0.0809 | 95.2 |
| σ_4 = 0.6007 | -0.0016 | 0.0244 | 0.0242 | 93.8 | -0.0366 | 0.0415 | 0.0404 | 81.9 | -0.0127 | 0.0707 | 0.0678 | 94.8 |

Table 4.5: Simulation results under dropout mechanism [9], n = 300
Chapter 5

Approximate Conditional Likelihoods in
Multivariate Missing Data Problems

5.1 Introduction

Consider a general statistical model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \), where \( \Theta \) denotes the parameter space. Before one conducts statistical inference, a fundamental problem is whether the model, or the parametrization, is identifiable, that is, \( \theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2} \). For example, when \( P_{\theta} = N(\mu, \sigma^2) \), the i.i.d. normal family with unknown mean \( \mu \) and variance \( \sigma^2 \), the model is identifiable. On the other hand, when \( P_{\theta} = pN(\mu_1, \sigma^2_1) + (1-p)N(\mu_2, \sigma^2_2) \), the i.i.d. mixture normal family with \((p, \mu_1, \mu_2, \sigma^2_1, \sigma^2_2)\) all unknown, the model is not identifiable unless some constraints on the parameter space are imposed.

Let \([-\cdot\cdot]\) or \([-\cdot\cdot\cdot]\) be a generic notation for the conditional or unconditional probability density. In problems with missing data, we usually denote the model as \([R, W]\), where the indicator \( R = 1 \) means the corresponding \( W \) is observed, and \( W = (Y, X) \), having \( Y \) as the response and \( X \) as its covariates. There are several approaches to formulate statistical
models with missing data. In this chapter, we focus on selection model (Little and Rubin, 2002), which factorizes \([R, W] = [R|W][W]\), in which \([W]\) is the model for the data generating process, while \([R|W]\) is the one for the missing mechanism. We assume that the two models are identifiable respectively.

In statistical models with missing data, we conduct inference mainly based on the observed likelihood (Little and Rubin, 2002). Although the models \([R|W]\) and \([W]\) are identifiable respectively, the observed likelihood may not be and we may not fully recover the whole information of \([R|W]\) and \([W]\). When the missing mechanism is missing at random (MAR) or missing completely at random (MCAR), the observed likelihood is always identifiable, therefore, there is no worry on this point. In the missing data literature, most work is based on MAR or MCAR assumption.

However, when the missing mechanism is missing not at random (MNAR), or, non-ignorable missing, which is the focus of the current chapter, this becomes a serious and challenging problem and the observed likelihood may not be identifiable even in a most simple scenario (Wang, Shao and Kim, 2012). In the literature, Greenless, Reece and Zieschang (1982) proposed a maximum likelihood method under parametric assumptions on both \([R|W]\) and \([W]\). But a fully parametric approach is sensitive to the model assumptions. Since the population is not identifiable when both \([R|W]\) and \([W]\) are nonparametric (Robins and Ritov, 1997), efforts have been made in situations where one of the two is parametric. Qin, Leung and Shao (2002) focused on the case where \([R|W]\) is parametric but \([W]\) is non-parametric. Wang, Shao and Kim (2012) studied the observed likelihood with nonignorable nonresponse, and gave some technical conditions under which the observed likelihood is identifiable. Their conditions need the covariate \(X\) consisting of at least two components \(U\) and \(Z\). In real applications, the model for the data generating process is often statisticians’ main
interest. Therefore, Tang, Little and Raghunathan (2003) considered the situation where $[W]$ is parametric but $[R|W]$ is nonparametric. They considered nonignorable nonresponse while the covariates are fully observed. They assumed that the missingness of the response only depends on the response, not on the covariates, i.e., $[R = 1|Y, X] = [R = 1|Y]$.

In this chapter, we propose conditional likelihood and approximate conditional likelihood in multivariate data with missing values. This is a different angle from studying the regular observed likelihood. We impose a parametric model on the data generating process $[W]$, while a nonparametric one on the missing mechanism $[R|W]$. Different from Tang, Little and Raghunathan (2003), our missing mechanism $[R|W]$ can depend on both response and covariates. We mainly consider two missing mechanisms: $[R = 1|Y, X] = s(Y)t(X)$ and $[R = 1|Y, U, Z] = s(Y, U)t(Z)$, where $s, t$ are two unknown functions. This approach allows us focusing on the inference of $[W]$ while giving a flexible and general assumption on $[R|W]$. Although this approach is originally designed for nonignorable missing data, it also applies for certain MAR missing mechanisms. The idea of conditional likelihood is originally motivated from biased sampling problems (Kalbfleisch, 1978; Liang and Qin, 2000).

The remaining of the chapter is organized as follows: in section 2, we briefly describe the idea of conditional likelihood and approximate conditional likelihood. Section 3 contains a summary of univariate generalized linear models (GLM, McCullagh and Nelder, 1989) and some models for multivariate data, which are frequently used for the data generating process in real applications and hence the main target. Sections 4 and 5 establish the identifiability theorems and the asymptotic behaviors of the proposed estimators, the main results of this chapter. Some numerical results and the concluding remarks are provided afterwards. All the technical proofs are given in the appendix.
5.2 Approximate Conditional Likelihood

Suppose we have independent and identically distributed samples \( \{(r_i, y_i, x_i), i = 1, \cdots, N\} \) from the population \([R, Y, X]\), where \( Y \) is a \( q \)-dimensional response variable, \( X \) is a \( p \)-dimensional covariate, \( N \) is the sample size. We define the missing indicator \( R = 1 \) if and only if the corresponding response \( Y \) and its covariate \( X \) are both fully observed. Also, without loss of generality, we assume \( r_i = 1, i = 1, \cdots, n \) and \( r_i = 0, i = n + 1, \cdots, N \), hence, \( n \) is the sample size of the fully observed subjects.

The missing indicator \( R \) as we defined may have different meanings under various scenarios: \( R \) may indicate the missingness of \( Y \) if \( X \) is fully observed; it may indicate the missingness of \( X \) if \( Y \) is fully observed; it may also indicate the joint missingness of \( Y \) and \( X \) when both response and covariate have missing values. In this chapter, we consider a general missing mechanism as follows:

\[
[R = 1|Y, X] = s(Y)t(X), \tag{1}
\]
i.e., the missingness depends on \( s(Y) \), an unknown function of \( Y \) only, multiplies \( t(X) \), an unknown function of \( X \) only. Throughout this chapter, we do not need to specify the forms of \( s \) and \( t \). This missing data mechanism includes many special cases investigated in the literature. For example, when \( X \) is fully observed and only \( Y \) has missing values, the missing at random (MAR) case and nonignorable nonresponse case (Tang, Little and Raghunathan, 2003) are both special cases of (1). Liang and Qin (2000) considered the assumption \([R|Y, X] = [R|Y]\) for univariate \( Y \), which is also a special case of (1). It also includes MAR missing covariate case. In the following sections, we also consider

\[
[R = 1|Y, U, Z] = s(Y, U)t(Z), \tag{2}
\]
if the covariate \( X = (U, Z) \). We only focus on (1) in this section for the purpose of introducing
the conditional likelihood.

To describe the idea of conditional likelihood and approximate conditional likelihood, 
we first assume the response variable $Y$ is univariate. We assume $Y$, given the $p$-dimensional 
covariate $X = (X_1, \cdots, X_p)^\top$, follows the model with probability density function $p(y|x; \theta)$ 
with respect to some $\sigma$-finite measure $\nu$, where $\theta$ indicates the parameter of interest. Suppose 
the parameter space $\Theta$ is compact, $\Theta \subset R^d$, where $d$ is the dimension of $\theta$. Denote $\theta_0$ as the 
true value of $\theta$, $\theta_0 \in \Theta^0$, the interior of $\Theta$.

If there is no missing values in the data set, to estimate $\theta$, we usually consider the 
regular maximum likelihood estimator (MLE), which maximizes the following likelihood 
function:

$$
\prod_{i=1}^{N} p(y_i|x_i; \theta),
$$

however, in missing data problems, this approach doesn’t work because of the incompleteness 
of the data set. One alternative is to consider the MLE based on completely observed 
subjects, i.e., to maximize the following objective function:

$$
\prod_{i=1}^{n} p(y_i|x_i; \theta).
$$

This naive way may produce less efficient, but still valid estimators only under some special 
cases of (1), for instance, MAR response case. In general, it is unknown how to construct 
valid estimators under assumption (1).

Let’s consider the completely observed subjects, $i = 1, \cdots, n$, which results the infer-
ence based on $[Y|X, R = 1]$. Since

$$
[Y|X, R = 1] = \frac{[R = 1|Y, X]}{[R = 1|X]} [Y|X], \quad (3)
$$

this way will result in correct inference of $\theta$ if $[R = 1|Y, X] = [R = 1|X]$, or both of them 
could be correctly specified.
Based on (1), the ratio term on the right hand side of (3), the weight function, becomes
\[
\frac{[R = 1|Y, X]}{[R = 1|X]} = s(Y) \frac{t(X)}{[R = 1|X]},
\]
and by further conditioning on the order statistics \(y_{(1)}, \ldots, y_{(n)}\) of \(Y_1, \ldots, Y_n\), we have the following conditional likelihood for \(\theta\):
\[
p(y_1, \ldots, y_n|r = 1, x_1, \ldots, x_n, y_{(1)}, \ldots, y_{(n)}) = \frac{\prod_{i=1}^n p(y_i|x_i; \theta)}{\sum_c \prod_{i=1}^n p(y_{(i)}|x_i; \theta)},
\]
where the summation corresponds to all possible permutations of \(\{1, 2, \ldots, n\}\). Notice that all weight functions are eliminated through conditioning, hence we don’t need any specifications of the weight functions, or the missing mechanism.

This kind of conditional likelihood was first proposed by Kalbfleisch (1978), but in practice, (5) encounters a tremendous computational burden with an order of \(n!\) (Liang and Qin 2000). To reduce the computational burden, the following \(m\)-wise approximate conditional likelihood can be considered:
\[
\prod_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} \frac{\prod_{j=1}^m p(y_{i_j}|x_{i_j}; \theta)}{\sum_{c_m} \prod_{j=1}^m p(y_{(i_j)}|x_{i_j}; \theta)},
\]
where \(c_m\) indicates all possible permutations of \(\{i_1, \ldots, i_m\}\), \(2 \leq m \leq n\). When \(m = n\), (6) shrinks to (5); while when \(m = 2\), (6) becomes the pairwise approximate conditional likelihood, considered in Liang and Qin (2000):
\[
L_v(\theta) = \prod_{1 \leq i < j \leq n} \frac{p(y_i|x_i; \theta)p(y_j|x_j; \theta)}{p(y_i|x_i; \theta)p(y_j|x_j; \theta) + p(y_j|x_i; \theta)p(y_i|x_j; \theta)} = \prod_{1 \leq i < j \leq n} \frac{1}{1 + R_{ij}(\theta)},
\]
where
\[
R_{ij}(\theta) = \frac{p(y_j|x_i; \theta)p(y_i|x_j; \theta)}{p(y_i|x_i; \theta)p(y_j|x_j; \theta)}
\]
is termed as generalized odds ratio in Qin and Liang (1999).
In this chapter, we focus on the pairwise approximate conditional likelihood \( L_c(\theta) \). The results for general \( m \)-wise approximate version can be similarly derived. The \( m \)-wise version may increase the estimation efficiency, however, the computational burden increases as \( m \) gets larger. Also, it turns out that general \( m \)-wise approximate version doesn’t help for the model identifiability.

For multivariate \( q \)-dimensional \( Y \), we can similarly define the corresponding approximate conditional likelihood \( L_c(\theta) \).

5.3 Models for the Data Generating Process

In this section, we mainly give a brief review of the models for the data generating process: univariate GLM, a multivariate extension of GLM for multicategorical responses, and multivariate linear regression model.

5.3.1 Univariate GLM

We first assume the response variable \( Y \) is one dimensional and follows the p.d.f.

\[
p_Y(y; \eta, \phi) = \exp \left\{ \frac{y \eta - b(\eta)}{\phi} + c(y; \phi) \right\},
\]

w.r.t. some \( \sigma \)-finite measure \( \nu \), where \( b \) and \( c \) are some specific functions, \( \phi > 0 \) denotes the dispersion parameter. For simplicity, we use \( \phi \) to denote the whole dispersion parameter, instead of original function \( a(\phi) \). Notice that if \( \phi \) is known, this is an exponential family model with canonical parameter \( \eta \). From the properties of GLM, we have

\[
E(Y) = b'(\eta) \text{ and } \text{Var}(Y) = \phi b''(\eta),
\]
where $b'$ and $b''$ are the first and second order derivatives of $b$, respectively. Define $\mu(\eta) = b'(\eta)$. It is assumed that $\eta$ is related to the covariate $x$ through the relationship:

$$g(\mu(\eta)) = \alpha + \beta^\top x,$$

where $\beta = (\beta_1, \cdots, \beta_p)^\top$, and $g$, called a link function, is a known one-to-one, third-order continuously differentiable function. Define $\psi = (g \circ \mu)^{-1}$. If $\mu = g^{-1}$, then $g$ is called the canonical link function, denoted as $g_c$. If $g$ is not canonical, we assume that $\frac{d}{d\eta} \psi^{-1}(\eta) \neq 0$ for all $\eta$. Thus, $p(y|x; \theta)$ can be expressed as

$$p(y|x; \theta) = \exp \left\{ \frac{y\psi(\alpha + \beta^\top x) - b(\psi(\alpha + \beta^\top x))}{\phi} + c(y; \phi) \right\},$$

where $\theta = (\alpha, \beta^\top, \phi)^\top$. The parameter space $\Theta \subset R \otimes R^p \otimes (0, \infty)$. Denote the true value of $\theta$ as $\theta_0 = (\alpha_0, \beta_0^\top, \phi_0)$, where $\beta_0 = (\beta_{10}, \cdots, \beta_{p0})^\top$ denotes the true value of $\beta$. Hence, generally $\psi = g_c \circ g^{-1}$,

$$E(Y|X) = b'(\psi(\alpha + \beta^\top x)),$$

and $\text{Var}(Y|X) = \phi b''(\psi(\alpha + \beta^\top x))$.

Generalized linear models (GLM) are widely used in practice. Besides the linear regression with normal errors, GLM also includes logistic regression, Poisson regression, etc., which are frequently used when modeling categorical data. In practice, canonical link is more preferred and the function $\psi(t) = t$ in the canonical link case. In GLM, canonical links do not always provide the best fit. Generally, there is no reason apriori why a canonical link should be used, and in many cases a noncanonical link is more preferable, see McCullagh and Nelder (1989) and Czado and Munk (2000). In this chapter, we focus on both canonical and noncanonical link functions (Wedderburn 1976). We list several frequently used GLMs below (McCullagh and Nelder, 1989).
Example 14 (Normal).

\[ b(\eta) = \eta^2, b'(\eta) = \eta, b''(\eta) = 1, \]
\[ p(y|x; \theta) = \exp \left\{ \frac{y\psi(\alpha + \beta^T x) - \frac{1}{2}(\psi(\alpha + \beta^T x))^2}{\phi} - \frac{1}{2} \left( \frac{y^2}{\phi} + \log(2\pi\phi) \right) \right\}, y \in \mathbb{R}, \]

where
\[ g_c(t) = t, \psi(t) = g(t)^{-1}, E(Y|X) = \psi(\alpha + \beta^T x), \text{ and } \text{Var}(Y|X) = \phi. \]

Example 15 (Poisson).

\[ b(\eta) = \exp\{\eta\}, b'(\eta) = \exp\{\eta\}, b''(\eta) = \exp\{\eta\}, \]
\[ p(y|x; \theta) = \exp \left\{ y\psi(\alpha + \beta^T x) - \exp(\psi(\alpha + \beta^T x)) - \log y! \right\}, y = 0, 1, \ldots, \]

where
\[ g_c(t) = \log(t), \psi(t) = \log(g(t)^{-1}), E(Y|X) = \text{Var}(Y|X) = \exp\{\psi(\alpha + \beta^T x)\}. \]

Notice that dispersion parameter \( \phi = 1 \). The noncanonical link \( g(t) = t^\gamma, 0 < \gamma < 1 \) leads to \( \psi(t) = \frac{1}{\gamma} \log(t) \). The link function \( g(t) = 2\sqrt{t} \), making \([\psi'(t)]^2b''(\psi(t)) = 1\), leads to \( \psi(t) = 2\log(t/2) \) (Shao, 2003, pp.283).

Example 16 (Binomial).

\[ b(\eta) = \log(1 + e^\eta), b'(\eta) = \frac{e^\eta}{1 + e^\eta}, b''(\eta) = \frac{e^\eta}{(1 + e^\eta)^2}, \]
\[ p(y|x; \theta) = \exp \left\{ \frac{y\psi(\alpha + \beta^T x) - \log(1 + \exp\{\psi(\alpha + \beta^T x)\})}{1/m} + \log \left( \frac{m}{my} \right) \right\}, y = \frac{0, 1, \ldots, m}{m}, \]

where
\[ g_c(t) = \log \frac{t}{1-t}, \psi(t) = \log \frac{g(t)^{-1}}{1 - g(t)^{-1}}, E(Y|X) = \frac{\exp\{\psi(\alpha + \beta^T x)\}}{1 + \exp\{\psi(\alpha + \beta^T x)\}}, \]
\[ \text{Var}(Y|X) = \frac{\exp\{\psi(\alpha + \beta^T x)\}}{m(1 + \exp\{\psi(\alpha + \beta^T x)\})^2}. \]
Notice that when \( m = 1 \), the dispersion parameter \( \phi = 1 \). The noncanonical links

\[
g(t) = t, \arcsin \sqrt{t}, \log(-\log(1-t)), \Phi^{-1}(t)
\]

leads to

\[
\psi(t) = \log \frac{t}{1 - t}, 2 \log |\tan(t)|, \log(\exp(e') - 1), \log \frac{\Phi(t)}{1 - \Phi(t)},
\]

respectively.

**Example 17** (Gamma).

\[
\begin{align*}
b(\eta) &= -\log(-\eta), b'(\eta) = -\frac{1}{\eta}, b''(\eta) = \frac{1}{\eta^2}, \eta < 0, \\
p(y|x; \theta) &= \exp \left\{ \frac{y\psi(\alpha + \beta^\tau x) + \log(-\psi(\alpha + \beta^\tau x))}{1/\nu} + \nu \log(\nu y) - \log y - \log \Gamma(\nu) \right\}, y > 0,
\end{align*}
\]

where

\[
g_c(t) = -\frac{1}{t}, \psi(t) = -\frac{1}{g(t)^{-1}}, E(Y|X) = -\frac{1}{\psi(\alpha + \beta^\tau x)}, \text{Var}(Y|X) = \frac{1}{\nu(\psi(\alpha + \beta^\tau x))^2}.
\]

The noncanonical links \( g(t) = \log(t), t^\gamma, -1 \leq \gamma < 0 \) leads to \( \psi(t) = -e^{-t}, -t^{-\gamma} \).

**Example 18** (Inverse Gaussian).

\[
\begin{align*}
b(\eta) &= -(2\eta)^{1/2}, b'(\eta) = -(2\eta)^{-1/2}, b''(\eta) = -(2\eta)^{-3/2}, \eta < 0, \\
p(y|x; \theta) &= \exp \left\{ \frac{y\psi(\alpha + \beta^\tau x) + (-2\psi(\alpha + \beta^\tau x))^{1/2}}{\phi} - \frac{1}{2} \left( \frac{1}{\phi y} + \log(2\pi \phi y^3) \right) \right\}, y > 0,
\end{align*}
\]

where

\[
g_c(t) = -\frac{1}{2t^2}, \psi(t) = -\frac{1}{2(g(t))^{-1}}, E(Y|X) = -(2\psi(\alpha + \beta^\tau x))^{-1/2}, \\
\text{Var}(Y|X) = \phi(-2\psi(\alpha + \beta^\tau x))^{-3/2}.
\]
5.3.2 Models for Multivariate Responses

When the response variable $Y$ is multicategorical, we mainly focus on the following multivariate extension of GLM (Fahrmeir and Tutz, 2001). Similar to univariate GLM, we have

$$p(Y|X; \theta) = \exp \left\{ \frac{Y^\tau \psi(A + B^\tau X) - b(\psi(A + B^\tau X))}{\phi} + c(Y; \phi) \right\},$$

where $A$ is a $q \times 1$ coefficient vector, $B$ is a $p \times q$ coefficient matrix. The definitions of functions $b$, $c$, $\psi$ and scalar dispersion parameter $\phi > 0$ are similar to univariate case. Notice that $\psi$ is defined from $\mathbb{R}^q$ to $\mathbb{R}^q$ and we denote $\psi_i$ as its $i$-th component. We have the following:

**Example 19** (Multinomial). Suppose we originally have $\tilde{Y} = (\tilde{Y}_1, \cdots, \tilde{Y}_{q+1})^\tau$, and $\tilde{Y}^\tau J_{q+1} = m$, where $J_{q+1}$ is the $(q + 1) \times 1$ column vector of all 1’s, and $\tilde{Y}_i$’s and $m$ are all integer values. Then $Y = (Y_1, \cdots, Y_q)^\tau$, where $Y_i = \tilde{Y}_i/m$ follows the multinomial distribution:

$$p(Y|X; \theta) = \exp \left\{ \frac{\sum_{i=1}^q Y_i \psi_i(A + B^\tau X) - \log(1 + \sum_{i=1}^q e^{\psi_i(A + B^\tau X)})}{1/m} ight. \left. + \log \frac{m!}{(mY_1)! (mY_2)! \cdots (mY_q)! (m - mY_1 - \cdots - mY_q)!} \right\}.$$

Notice that binomial distribution (Example 3) is a special case of multinomial distribution with $q = 1$. Generally we have

$$b(\eta) = \log \left( 1 + \sum_{i=1}^q \exp\{\eta_i\} \right), b'(\eta) = \frac{\exp\{\eta_i\}}{1 + \sum_{i=1}^q \exp\{\eta_i\}} \text{ as its } i\text{-th component.}$$

The canonical link $g_c(t) = \log \frac{t_i}{1 - \sum_{i=1}^q t_i}$ as its $i$-th component, and the $i$-th component of $\psi(t)$,

$$\psi_i(t) = \log \frac{g^{-1}_i(t)}{1 - \sum_{i=1}^q g^{-1}_i(t)}.$$

The noncanonical link functions

$$g(t) = t, \arcsin \sqrt{t}, \log(-\log(1 - t)), \Psi^{-1}(t)$$
leads to

\[ \psi_i(t) = \log \frac{t_i}{1 - \sum_{i=1}^{q} t_i}, \log \frac{\sin^2 t_i}{1 - \sum_{i=1}^{q} \sin^2 t_i}, \log \frac{1 - \exp(-e^{t_i})}{1 - \sum_{i=1}^{q} (1 - \exp(-e^{t_i}))}, \log \frac{\Psi(t_i)}{1 - \sum_{i=1}^{q} \Psi(t_i)} \]

respectively.

As the last example, we consider the following multivariate linear regression model:

**Example 20 (Multivariate Normal).** Suppose we have

\[ Y = A + B^\tau X + E, \]

where \( A \) is a \( q \times 1 \) coefficient vector, \( B \) is a \( p \times q \) coefficient matrix, and \( E \sim N(0, \Sigma) \), \( \Sigma \) is a \( q \times q \) positive definite matrix. Denote \( \Lambda = \Sigma^{-1} = \{\sigma_{ij}^{-1}\}_{1\leq i,j\leq q} \), the inverse of \( \Sigma \). Therefore, we have

\[
    p(Y|X; \theta) = (2\pi)^{-\frac{q}{2}}|\Sigma|^{-\frac{1}{2}} \exp \left\{ Y^\tau \Lambda B^\tau X + Y^\tau \Lambda A - A^\tau \Lambda B^\tau X - \frac{1}{2} (Y^\tau \Lambda Y + X^\tau \Lambda B^\tau X + A^\tau \Lambda A) \right\}.
\]

### 5.4 Identifiability

Through approximate conditional likelihood \([8]\), the parameter of interest \( \theta \) may not be fully identifiable. In subsection 1, we focus on regression models with missing mechanism \([1]\):

\[ [R = 1|Y, X] = s(Y)t(X). \]

In subsection 2, we devote to missing mechanism \([2]\):

\[ [R = 1|Y, U, Z] = s(Y, U)t(Z), \]

where \( U \) and \( Z \) are two components of covariate \( X \).
5.4.1 Missing Mechanism \( R = 1 | Y, X \) \( = s(Y)t(X) \)

Firstly, we consider the general regression model \( p(y|x; \theta) \), where the response variable \( Y \) could be multidimensional. We have the following results:

**Theorem 8.** For any \( \theta \in \Theta, \theta \neq \theta_0 \) and any arbitrary functions \( f(x), g(y) \), define

\[
D_\theta = \{ y : p(y|x; \theta) = \exp\{f(x) + g(y)\}p(y|x; \theta_0) \text{ for any } x \}.
\]

If \( P(R = 1) > 0 \) and \( P(D_\theta) < 1 \) for any \( \theta \neq \theta_0 \), then

\[
E_0 \left[ -1_{(R=1)} \log(1 + R_{ij}(\theta)) \right] < E_0 \left[ -1_{(R=1)} \log(1 + R_{ij}(\theta_0)) \right], \theta \neq \theta_0,
\]

where \( E_0 \) denotes expectation with respect to the true value \( \theta_0 \).

To prove this result, we need the following small lemma.

**Lemma 1.** For bivariate continuous differentiable function \( e(x, y) \),

\[
e(x_1, y_1) + e(x_2, y_2) = e(x_1, y_2) + e(x_2, y_1), \text{ for any } x_1 \neq x_2, y_1 \neq y_2,
\]

if and only if \( e(x, y) = f(x) + g(y) \), i.e., \( \frac{\partial^2}{\partial x \partial y} e(x, y) = 0 \).

**Corollary 5.** Suppose \( \theta \) can be reparameterized as \( \tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) \), with true value \( \tilde{\theta}_0 = (\tilde{\theta}_{10}, \tilde{\theta}_{20}) \), and the density \( p(y|x; \theta) \) can be written as

\[
p(y|x; \theta) = \exp\{h_1(x, y; \tilde{\theta}) + h_2(x; \tilde{\theta}) + h_3(y; \tilde{\theta})\},
\]

where \( h_1 \) is known up to \( \tilde{\theta}_1 \), and \( h_2, h_3 \) may not be specified. Also, assume that, for any \( k \neq \tilde{\theta}_{10} \) in the domain of \( \tilde{\theta}_1 \), the function \( h_1(x, y; k) - h_1(x, y; \tilde{\theta}_{10}) \) is neither a function of \( y \) alone nor a function of \( x \) alone. Then \( \tilde{\theta}_1 \) is identifiable.

**Remark 5.** Compared with the proposition in Tang, Little and Raghunathan (2003), one could see that they may identify more parameters than we do. The reason is that their
method only applies to the missing mechanism \( [R = 1|Y, X] = [R = 1|Y] \), and they can take advantage of the completeness of covariate \( X \), while our assumption \([1]\) is more general, and our \( X \) may also have missing values.

**Remark 6.** In the result above, the functions \( h_2 \) and \( h_3 \) need not to be specified or could be misspecified.

Next, we focus on the univariate GLM case \([11]\). Under \([11]\), the generalized odds ratio \( R_{ij}(\theta) \) can be simplified as follows:

\[
R_{ij}(\theta) = \frac{p(y_j|x_i;\theta)p(y_i|x_j;\theta)}{p(y_i|x_i;\theta)p(y_j|x_j;\theta)}
\]

\[
= \exp \left\{ \frac{(y_i - y_j)(\psi(\alpha + \beta^\tau x_j) - \psi(\alpha + \beta^\tau x_i))}{\phi} \right\},
\]

hence, the parameters identified from \([13]\) will be identifiable through \( L_c(\theta) \), \([8]\).

In the theorem below, we state the identifiability theorem for univariate GLM case. It mainly tells us, the parameter of interest will be identifiable under some conditions of function \( \psi = g_c \circ g^{-1} \) and the space of the support values of covariate \( X \). We consider two scenarios: the dispersion parameter \( \phi \) is one component of the unknown parameter, or \( \phi \) is known. The latter case is more interesting, especially in binary regression or Poisson regression.

**Theorem 9.** For univariate GLM \([11]\), define the space of the support values of \( X_i \) as \( \mathcal{X}_i \), \( i = 1, 2, \cdots, p \), and the space of the support values of \( X = (X_1, \cdots, X_p)^\tau \) as \( \mathcal{X} = \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_p \). Assume the function \( \psi \) is first order continuously differentiable.

(i) Under the following condition \((C1)\), the parameter \( \theta = (\alpha, \beta^\tau, \phi)^\tau \) is fully identifiable through \( L_c(\theta) \);

(ii) If \( \phi > 0 \) is known, under the following condition \((C2)\), the parameters \( \alpha \) and \( \beta \) are identifiable through \( L_c(\theta) \).
(C1). There exists \( X_{01} = \{x_0, x_1, \cdots, x_{p+2} \} \subset X \), such that functions
\[
\psi_i(\theta) = \frac{\psi(\alpha + \beta^\top x_i)}{\phi} - \frac{\psi(\alpha + \beta^\top x_0)}{\phi}, \quad i = 1, 2, \cdots, p + 2,
\]
are independent at \( \theta_0 = (\alpha_0, \beta_0^\top, \phi_0)^\top \), i.e., the following \((p + 2) \times (p + 2)\) matrix is of full rank when evaluated at the true values \( \alpha = \alpha_0, \beta = \beta_0 \):
\[
M_1 = \begin{pmatrix}
\psi'(\alpha + \beta^\top x_1) - \psi'(\alpha + \beta^\top x_0) & \psi'(\alpha + \beta^\top x_1)x_1^\top - \psi'(\alpha + \beta^\top x_0)x_0^\top \\
\vdots & \vdots \\
\psi'(\alpha + \beta^\top x_{p+2}) - \psi'(\alpha + \beta^\top x_0) & \psi'(\alpha + \beta^\top x_{p+2})x_{p+2}^\top - \psi'(\alpha + \beta^\top x_0)x_0^\top \\
\psi(\alpha + \beta^\top x_1) - \psi(\alpha + \beta^\top x_0) & \\
\vdots & \\
\psi(\alpha + \beta^\top x_{p+2}) - \psi(\alpha + \beta^\top x_0)
\end{pmatrix}
\]

(C2). There exists \( X_{02} = \{x_0, x_1, \cdots, x_{p+1} \} \subset X \), such that functions
\[
\psi_i(\alpha, \beta) = \psi(\alpha + \beta^\top x_i) - \psi(\alpha + \beta^\top x_0), \quad i = 1, 2, \cdots, p + 1,
\]
are independent at the true values \( \alpha_0 \) and \( \beta_0 \), i.e., the following \((p + 1) \times (p + 1)\) matrix is of full rank when evaluated at the true values \( \alpha = \alpha_0, \beta = \beta_0 \):
\[
M_2 = \begin{pmatrix}
\psi'(\alpha + \beta^\top x_1) - \psi'(\alpha + \beta^\top x_0) & \psi'(\alpha + \beta^\top x_1)x_1^\top - \psi'(\alpha + \beta^\top x_0)x_0^\top \\
\vdots & \vdots \\
\psi'(\alpha + \beta^\top x_{p+1}) - \psi'(\alpha + \beta^\top x_0) & \psi'(\alpha + \beta^\top x_{p+1})x_{p+1}^\top - \psi'(\alpha + \beta^\top x_0)x_0^\top
\end{pmatrix}
\]

Remark 7. The assumption \( \beta_0 \neq 0 \) is already implied in condition (C1) and (C2). The property that matrices \( M_1 \) or \( M_2 \) is of full rank also implies some necessary conditions on \( X_{01} \) or \( X_{02} \). For example, in condition (C1), the following requirement is implied: for
any \( j = 1, 2, \cdots, p \), there exists \( p_1, p_2 \in \{0, 1, \cdots, p + 2\} \), such that \( x_{p_1 j} \neq x_{p_2 j} \), where \( x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\tau \).

**Remark 8.** Under canonical link case, \( \psi(t) = t \), we can only identify \( \frac{\beta}{\phi} \) through approximate conditional likelihood (Liang and Qin, 2000).

**Remark 9.** We could notice that whether the condition \((C1)\) is satisfied or not does not depend on \( \phi_0 \), the true value of the dispersion parameter. The satisfaction of condition \((C1)\) and \((C2)\) only depend on the true values \( \alpha_0 \) and \( \beta_0 \).

In the special case of \( p = 1 \), the following results give us the explicit forms of function \( \psi \), such that \((C1)\) or \((C2)\) is satisfied.

**Corollary 6** (Theorem above with \( p = 1 \)). For univariate GLM \([11]\) with \( p = 1 \), define the space of support values of \( X \) as \( \mathcal{X} \). Assume the function \( \psi \) is first order continuously differentiable.

(i) Under the following condition \((C3)\), the parameter \( \theta = (\alpha, \beta, \phi)^\tau \) is fully identifiable through \( L_c(\theta) \);

(ii) If \( \phi > 0 \) is known, under the following condition \((C4)\), the parameters \( \alpha \) and \( \beta \) are identifiable through \( L_c(\theta) \).

(iii) The matrix \( M3 \) in \((C3)\) is singular at any values of \( \alpha \) and \( \beta \) if \( \psi'(t) = \pm \frac{1}{l} \exp\{kt + c\} \), \( l \neq 0 \), or \( \psi'(t) = \frac{1}{k + c} \), \( k \neq 0 \), or \( \psi'(t) = c \), where \( c \) is a constant.

(iv) The matrix \( M4 \) in \((C4)\) is singular at any values of \( \alpha \) and \( \beta \) if and only if \( \psi'(t) = \frac{1}{kt + c} \), \( k \neq 0 \), or \( \psi'(t) = c \), where \( c \) is a constant.

**\((C3)\).** There exists \( \mathcal{X}_{03} = \{x_0, x_1, x_2, x_3\} \subset \mathcal{X} \), i.e., the support space of \( X \) has at least 4
values. The functions
\[ \psi_i(\theta) = \frac{\psi(\alpha + \beta x_i)}{\phi} - \frac{\psi(\alpha + \beta x_0)}{\phi}, \quad i = 1, 2, 3, \]
are independent at \( \theta_0 = (\alpha_0, \beta_0, \phi_0)^T \), i.e., the following \( 3 \times 3 \) matrix is of full rank when evaluated at the true values \( \alpha = \alpha_0, \beta = \beta_0 \):
\[
M_3 = \begin{pmatrix}
\psi'(\alpha + \beta x_1) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_1)x_1 - \psi'(\alpha + \beta x_0)x_0 & \psi(\alpha + \beta x_1) - \psi(\alpha + \beta x_0) \\
\psi'(\alpha + \beta x_2) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_2)x_2 - \psi'(\alpha + \beta x_0)x_0 & \psi(\alpha + \beta x_2) - \psi(\alpha + \beta x_0) \\
\psi'(\alpha + \beta x_3) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_3)x_3 - \psi'(\alpha + \beta x_0)x_0 & \psi(\alpha + \beta x_3) - \psi(\alpha + \beta x_0)
\end{pmatrix};
\]

(C4). There exists \( X_{04} = \{x_0, x_1, x_2\} \subset X \), i.e., the support space of \( X \) has at least 3 values.

The functions
\[ \psi_i(\theta) = \psi(\alpha + \beta x_i) - \psi(\alpha + \beta x_0), \quad i = 1, 2, \]
are independent at the true values \( \alpha_0 \) and \( \beta_0 \), i.e., the following \( 2 \times 2 \) matrix is of full rank when evaluated at the true values \( \alpha = \alpha_0, \beta = \beta_0 \):
\[
M_4 = \begin{pmatrix}
\psi'(\alpha + \beta x_1) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_1)x_1 - \psi'(\alpha + \beta x_0)x_0 \\
\psi'(\alpha + \beta x_2) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_2)x_2 - \psi'(\alpha + \beta x_0)x_0
\end{pmatrix}.
\]

To prove Corollary 2, we need the following lemma from knowledge of functional equations.

**Lemma 2** (Cauchy’s functional equation). Cauchy’s functional equation is the functional equation
\[ f(x + y) = f(x) + f(y). \]
Solutions to this are called additive functions. Over the rational numbers, it can be shown using elementary algebra that there is a single family of solutions, namely \( f(x) = cx \) for any arbitrary rational number \( c \). Over the real numbers, this is still a family of solutions;
however there can exist other solutions that are extremely complicated. Further constraints on \( f \) sometimes preclude other solutions, for example:

1. if \( f \) is continuous. This condition was weakened in 1875 by Darboux who showed that it was only necessary for the function to be continuous at one point.

2. if \( f \) is monotonic on any interval.

3. if \( f \) is bounded on any interval.

Example 21 (Poisson continued). Consider Poisson distribution discussed in Example 2, with \( p = 1 \) and dispersion parameter \( \phi = 1 \).

- Canonical link \( g_c(t) = \log(t) \), hence \( \psi(t) = t \). Parameter \( \beta \) is identifiable, but \( \alpha \) is not identifiable.

- Noncanonical link \( g(t) = t^\gamma, 0 < \gamma < 1 \), hence \( \psi(t) = \frac{1}{\gamma} \log(t) \) and \( \psi'(t) = \frac{1}{\gamma t} \). Therefore, from Corollary 2, we can not identify the whole parameter. On the other hand, since

\[
\frac{\log(\alpha + \beta x)}{\gamma} = \frac{\log(\zeta) + \log(\frac{\alpha}{\zeta} + \frac{\beta}{\zeta} x)}{\gamma},
\]

for any \( \zeta > 0 \), neither \( \alpha \) nor \( \beta \) is identifiable through \( L_c(\theta) \). Actually we can only identify the ratio between \( \alpha \) and \( \beta \) in this case.

Example 22 (Binomial continued). Consider logistic regression, a special case of example 3 with \( m = 1, p = 1 \).

- Canonical link \( g_c(t) = \log \left( \frac{1}{1-t} \right) \), hence \( \psi(t) = t \). Parameter \( \beta \) is identifiable, but \( \alpha \) is not identifiable.
• Noncanonical link \( g(t) = t \), hence \( \psi(t) = \log \frac{t}{1-t} \). Therefore, \( \frac{1}{\psi'(t)} = t - t^2 \). It can be easily checked that both \( \alpha \) and \( \beta \) are identifiable if and only if the true parameter \( \beta_0 \neq 0 \).

**Example 23 (Gamma continued).** Consider Gamma distribution, discussed in Example 4 with \( p = 1 \).

• Canonical link \( g_c(t) = -\frac{1}{t} \), hence \( \psi(t) = t \). Only \( \beta \nu \) is identifiable. If the dispersion parameter \( 1/\nu \) is known, \( \beta \) is identifiable, but \( \alpha \) is still not identifiable.

• Noncanonical link \( g(t) = \log(t) \), hence \( \psi(t) = -e^{-t} \), \( \psi'(t) = e^{-t} \). Therefore, the summation of the first column and the third column of the matrix in condition (C3) equals zero, and we can not identify the whole parameter. On the other hand, from direct derivation, we can only identify \( \alpha - \log(\nu) \) and \( \beta \).

If the dispersion parameter \( 1/\nu \) is known, \( \alpha \) and \( \beta \) are identifiable if and only if

\[
(e^{\beta_0(x_1-x_0)})^c - c e^{\beta_0(x_1-x_0)} + c - 1 \neq 0,
\]

which is equivalent to \( e^{\beta_0(x_1-x_0)} \neq 1 \), if and only if \( \beta_0 \neq 0 \). Here, \( c = \frac{x_2-x_0}{x_1-x_0} \).

• Noncanonical link \( g(t) = t^{\gamma} \), \( -1 < \gamma < 0 \), hence, \( \psi(t) = -t^{-\frac{1}{\gamma}} \), \( \psi'(t) = \frac{1}{\gamma} t^{-\frac{1+\gamma}{\gamma}} \). Since

\[
\nu(\alpha + \beta x)^{-1/\gamma} = (\nu^{-\gamma} \alpha + \nu^{-\gamma} \beta x)^{-1/\gamma},
\]

we can only identify the ratio between \( \alpha \) and \( \beta \).

If the dispersion parameter \( 1/\nu \) is known, \( \alpha \) and \( \beta \) are identifiable if and only if

\[
(c - 1)(\alpha_0 + \beta_0 x_0)^{\frac{1+\gamma}{\gamma}} - c(\alpha_0 + \beta_0 x_1)^{\frac{1+\gamma}{\gamma}} + (\alpha_0 + \beta_0 x_2)^{\frac{1+\gamma}{\gamma}} \neq 0,
\]

where \( c = \frac{x_2-x_0}{x_1-x_0} \). Especially, when \( \gamma = -\frac{1}{2} \), the condition shrinks to \( \beta_0 \neq 0 \).
Example 24 (Multinomial continued). Consider the multinomial distribution, discussed in Example 6 with \( q = 2, p = 1 \) and known dispersion parameter \( 1/m \).

- Canonical link \( g_c(t) = \log \frac{t}{1-t J_q} \) and \( \psi(t) = t \). Similarly to the univariate case, only the coefficient matrix \( B \) is identifiable.

- Noncanonical link \( g(t) = t \) and \( \psi(t) = \log \frac{t}{1-t J_q} \).

Example 25 (Multivariate Normal continued). Consider the multivariate normal distribution, discussed in Example 7. It can be easily seen that only the parameter \( \Lambda B^T \), which is \( q \times p \), is identifiable.

5.4.2 Missing Mechanism \( [R = 1|Y, U, Z] = s(Y, U)t(Z) \)

In this subsection, we consider the missing mechanism \( [R = 1|Y, U, Z] = s(Y, U)t(Z) \). This mechanism contains \( [R = 1|Y, U, Z] = (1 + \exp\{\zeta + \eta Y + \tau U\})^{-1} \) as a special case, which depends on both the response variable \( Y \), and \( U \), part of the covariate \( X \). This is an advantage compared to the missing mechanism \( [R = 1|Y, X] = s(Y)t(X) \).

Consider a general regression model \( p(y|u, z; \theta) \), where \( \theta \) is the parameter of interest, similar as previous sections. We impose a general regression model between \( U \) and \( Z \), \( p_u(u|z; \xi) \), where \( \xi \) is the nuisance parameter, \( \xi \in \Xi \), and denote the true value as \( \xi_0 \). Since \( [Y, U|Z] = [Y|U, Z][U|Z] \), we have the regression model between \( Y, U \) and \( Z \) as
\[
p_{yu}(y, u|z; \theta, \xi) = p(y|u, z; \theta)p_u(u|z; \xi).
\]

Based on missing mechanism (2), similar as previous sections, we consider the following
approximate conditional likelihood:

\[
L_c(\theta, \xi) = \prod_{1 \leq i < j \leq n} \frac{p_{uy}(y_i, u_i | z_i; \theta, \xi)p_{uy}(y_j, u_j | z_j; \theta, \xi)}{p_{uy}(y_i, u_i | z_i; \theta, \xi)p_{uy}(y_j, u_j | z_j; \theta, \xi) + p_{uy}(y_j, u_j | z_j; \theta, \xi)p_{uy}(y_i, u_i | z_i; \theta, \xi)} \tag{14}
\]

\[
= \prod_{1 \leq i < j \leq n} \frac{1}{1 + R_{ij}(\theta, \xi)}, \tag{15}
\]

where

\[
R_{ij}(\theta, \xi) = \frac{p_{uy}(y_j, u_j | z_j; \theta, \xi)p_{uy}(y_i, u_i | z_i; \theta, \xi)}{p_{uy}(y_i, u_i | z_i; \theta, \xi)p_{uy}(y_j, u_j | z_j; \theta, \xi)} \tag{16}
\]

\[
= \frac{p(y_j | u_j, z_j; \theta)p(y_i | u_i, z_i; \theta)}{p(y_i | u_i, z_i; \theta)p(y_j | u_j, z_j; \theta)} \cdot \frac{p_u(u_j | z_j; \xi)p_u(u_i | z_i; \xi)}{p_u(u_i | z_i; \xi)p_u(u_j | z_j; \xi)} \tag{17}
\]

Generally we have the following results:

**Theorem 10.** For any \( \theta \in \Theta, \theta \neq \theta_0, \xi \in \Xi, \xi \neq \xi_0 \) and any arbitrary functions \( f_1(y, u), f_2(u), g_1(z), g_2(z) \), define

\[
D_{\theta, \xi} = \{(y, u) : p(y | u, z; \theta) = [\exp\{f_1(y, u) + g_1(z)\}\]p(y | u, z; \theta_0), \]

\[
p_u(u | z; \xi) = [\exp\{f_2(u) + g_2(z)\}\]p_u(u | z; \xi_0) \text{ for any } z\}.
\]

If \( P(R = 1) > 0 \) and \( P(D_{\theta, \xi}) < 1 \) for any \( \theta \neq \theta_0, \xi \neq \xi_0 \), then

\[
E_0 \left[-1_{(R = 1)} \log(1 + R_{ij}(\theta, \xi))\right] < E_0 \left[-1_{(R = 1)} \log(1 + R_{ij}(\theta_0, \xi_0))\right], \theta \neq \theta_0, \xi \neq \xi_0,
\]

where \( E_0 \) denotes expectation with respect to the true value \( \theta_0, \xi_0 \).

Similar as Theorem 1, we need the following lemma to prove this result.

**Lemma 3.** For continuously differentiable function \( e(y, u, z) \),

\[
e(y_1, u_1, z_1) + e(y_2, u_2, z_2) = e(y_1, u_1, z_2) + e(y_2, u_2, z_1), \text{ for any } (y_1, u_1) \neq (y_2, u_2), z_1 \neq z_2,
\]

if and only if \( e(y, u, z) = f(y, u) + g(z) \), i.e., \( \frac{\partial^2}{\partial y \partial z} e(y, u, z) = \frac{\partial^2}{\partial u \partial z} e(y, u, z) = 0. \)
Corollary 7. Suppose θ can be reparameterized as \( \tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2) \), with true value \( \tilde{\theta}_0 = (\tilde{\theta}_{10}, \tilde{\theta}_{20}) \), and the density \( p(y|u, z; \theta) \) can be written as

\[
p(y|u, z; \theta) = \exp\{h_1(y, u, z; \tilde{\theta}_1) + h_2(y, u; \tilde{\theta}) + h_3(z; \tilde{\theta})\},
\]

where \( h_1 \) is known up to \( \tilde{\theta}_1 \), and \( h_2, h_3 \) may not be specified. Also, assume that, for any \( k \neq \tilde{\theta}_{10} \) in the domain of \( \tilde{\theta}_1 \), the function \( h_1(y, u, z; k) - h_1(y, u, z; \tilde{\theta}_{10}) \) is neither a function of \( y, u \) alone nor a function of \( z \) alone. Then \( \tilde{\theta}_1 \) is identifiable.

Remark 10. In the results shown above, we actually treat \( \theta \) and \( \xi \) in the same manner. If, for example, both \( U \) and \( Z \) are completely observed, we could firstly obtain an estimator of \( \xi \), and then concentrate on the estimation of \( \theta \). For the parameter identification, the results will be similar. For the estimation performance, it may increase the efficiency.

In the following, we focus on the univariate GLM with covariates \( U \) and \( Z \):

\[
p(y|u, z; \theta) = \exp\left\{ \frac{y\psi(\alpha + \beta_u^u u + \beta_z^z z)}{\phi} - \frac{b\psi(\alpha + \beta_u^u u + \beta_z^z z)}{\phi} + c(y; \phi) \right\}, \quad (18)
\]

where \( \theta = (\alpha, \beta_u^u, \beta_z^z, \phi)^\top \). Suppose \( \beta_u \) is \( p_u \)-dimensional and \( \beta_z \) is \( p_z \)-dimensional. Denote \( b \circ \psi = \tau \).

Similar to Theorem 2, we establish the identifiability theorem for (18), which tells us, the parameter of interest will be identifiable under some conditions of functions \( \psi = g_c \circ g^{-1} \) and \( \tau = b \circ \psi = b \circ g_c \circ g^{-1} \) and the space of the support values of covariates \( U \) and \( Z \). Similar as before, we also consider two scenarios: the dispersion parameter \( \phi \) is one component of the unknown parameter, or \( \phi \) is known.

Theorem 11. For univariate GLM with covariates \( U \) and \( Z \) (18), define the space of the support values of \( U_i \) as \( \mathcal{U}_i \), \( i = 1, 2, \ldots, p_u \), and the space of the support values of \( U \) as \( \mathcal{U} = \mathcal{U}_1 \otimes \cdots \otimes \mathcal{U}_{p_u} \). Similarly define the space of the support values of \( Z \) as \( \mathcal{Z} = \mathcal{Z}_1 \otimes \cdots \otimes \mathcal{Z}_{p_z} \). Assume the function \( \psi \) and \( \tau \) are first order continuously differentiable.
• Under the following condition (C5), the parameter \( \theta = (\alpha, \beta_u^*, \beta_z^* , \phi)^\tau \) is fully identifiable through \( L_c(\theta, \xi) \):

• If \( \phi > 0 \) is known, under the following condition (C6), the parameters \( \alpha, \beta_u \) and \( \beta_z \) are identifiable through \( L_c(\theta, \xi) \).

(C5). There exists \( U_{01} = \{u_0, u_1, \ldots , u_q\} \subset U \), and \( Z_{01} = \{z_0, z_1, \ldots , z_r\} \subset Z \), such that

\[(2q + 1)r \geq p_u + p_z + 2 , \text{ and there exists } p_u + p_z + 2 \text{ functions from the following } (2q + 1)r \text{ functions:}

\[
\frac{\psi(\alpha + \beta_u^* u_j + \beta_z^* z_i)^\phi}{\phi} - \frac{\psi(\alpha + \beta_u^* u_j + \beta_z^* z_0)^\phi}{\phi} , i = 1, 2, \ldots, r; j = 0, 1, \ldots, q ,
\]

\[
\frac{\tau(\alpha + \beta_u^* u_j + \beta_z^* z_i)^\phi}{\phi} - \frac{\tau(\alpha + \beta_u^* u_j + \beta_z^* z_0)^\phi}{\phi} - \frac{\tau(\alpha + \beta_u^* u_0 + \beta_z^* z_i)^\phi}{\phi} + \frac{\tau(\alpha + \beta_u^* u_0 + \beta_z^* z_0)^\phi}{\phi} ,
\]

\[i = 1, \ldots, r; j = 1, \ldots, q ,\]

such that the \( p_u + p_z + 2 \) functions are independent at the true value \( \theta_0 = (\alpha_0, \beta_{u0}^*, \beta_{z0}^* , \phi)^\tau \);

(C6). There exists \( U_{02} = \{u_0, u_1, \ldots , u_q\} \subset U \), and \( Z_{02} = \{z_0, z_1, \ldots , z_r\} \subset Z \), such that

\[(2q + 1)r \geq p_u + p_z + 1, \text{ and there exists } p_u + p_z + 1 \text{ functions from the following } (2q + 1)r \text{ functions:}

\[
\psi(\alpha + \beta_u^* u_j + \beta_z^* z_i) - \psi(\alpha + \beta_u^* u_j + \beta_z^* z_0) , i = 1, 2, \ldots, r; j = 0, 1, \ldots, q ,
\]

\[
\tau(\alpha + \beta_u^* u_j + \beta_z^* z_i) - \tau(\alpha + \beta_u^* u_j + \beta_z^* z_0) - \tau(\alpha + \beta_u^* u_0 + \beta_z^* z_i) + \tau(\alpha + \beta_u^* u_0 + \beta_z^* z_0) ,
\]

\[i = 1, \ldots, r; j = 1, \ldots, q ,\]

such that the \( p_u + p_z + 1 \) functions are independent at the true values \( \alpha_0, \beta_{u0}, \beta_{z0} \).

Remark 11. When the true value \( \beta_{z0} = 0 \), it is impossible to identify anything, similar as Theorem 2. While when the true value \( \beta_{u0} = 0 \), it is still possible to identify part of the parameters.
Remark 12. In Theorem 4, both $\psi$ and $b$ are involved in the conditions to identify the parameters, while in Theorem 2, only function $\psi$ is involved.

Remark 13. Under the canonical link case, it is possible to identify all parameters.

In the next, we consider the special case of $p_u = 1$, $p_z = 1$, and $\phi > 0$ is known. We study under what conditions the parameters $\alpha$, $\beta_u$ and $\beta_z$ are all identifiable if $U$ only has two support values $u_0, u_1$, $Z$ also only has two support values $z_0, z_1$. For simplicity, we assume $u_0 = z_0 = 0, u_1 = z_1 = 1$. It can be derived that condition (C6) is not satisfied if

$$
\tau'(\alpha_0)\psi'(\alpha_0 + \beta_{z0})[\psi'\left(\alpha_0 + \beta_{u0} + \beta_{z0}\right) - \psi'\left(\alpha_0 + \beta_{u0}\right)] + \tau'(\alpha_0 + \beta_{u0} + \beta_{z0})\psi'(\alpha_0 + \beta_{u0})[\psi'(\alpha_0) - \psi'(\alpha_0 + \beta_{z0})] = \tau'(\alpha_0 + \beta_{z0})\psi'(\alpha_0)[\psi'(\alpha_0 + \beta_{u0}) - \psi'(\alpha_0 + \beta_{u0} + \beta_{z0})].
$$

Example 26 (Normal continued). Consider the normal distribution, studied in Example 1.

- Canonical link $g_c(t) = t$, and $b(t) = t^2/2$, hence, $\tau(t) = t^2/2$. Only $\beta_u$ and $\beta_z$ are identifiable.

Example 27 (Poisson continued). Consider the Poisson distribution, studied in Example 2 and Example 8.

- Canonical link $g_c(t) = \log(t)$, and $b(t) = e^t$. Hence, $\psi(t) = t$, $\tau(t) = e^t$. Only $\beta_z$ and $e^{\alpha + \beta_u} - e^\alpha$ are identifiable.

- Noncanonical link $g(t) = t^\gamma, 0 < \gamma < 1$. Hence, $\psi(t) = \frac{1}{\gamma} \log(t)$ and $\tau(t) = t^{-\gamma}$. It can be derived that if $\beta_{z0} \neq 0$, and

$$
\frac{1}{\gamma} (\alpha_0 + \beta_{u0} + \beta_{z0})^{\frac{1}{\gamma}} \neq (\alpha_0 + \beta_{u0})^{\frac{1}{\gamma}} + (\alpha_0 + \beta_{z0})^{\frac{1}{\gamma}},
$$

then $\alpha$, $\beta_u$ and $\beta_z$ are all identifiable.
Example 28 (Binomial continued). Consider logistic regression, studied in Example 3 and Example 9.

- Canonical link \( g_c(t) = \log \frac{t}{1-t} \), and \( b(t) = \log(1 + e^t) \). Hence, \( \psi(t) = t \), \( \tau(t) = \log(1 + e^t) \). Only \( \beta_z \) and \( \log(1 + e^{\alpha + \beta_u + \beta_z}) - \log(1 + e^{\alpha + \beta_u}) - \log(1 + e^{\alpha + \beta_z}) + \log(1 + e^{\alpha}) \) are identifiable.

- Noncanonical link \( g(t) = t \). Hence, \( \psi(t) = \log \frac{t}{1-t} \) and \( \tau(t) = -\log(1 - t) \). If condition (C6) is satisfied, then \( \alpha \), \( \beta_u \) and \( \beta_z \) are all identifiable.

Example 29 (Gamma continued). Consider Gamma distribution, studied in Example 4 and Example 10.

- Canonical link \( g_c(t) = -\frac{1}{t} \), and \( b(t) = -\log(-t) \). Hence, \( \psi(t) = t \), \( \tau(t) = -\log(-t) \). Only \( \beta_z \) and \( \log(-\alpha - \beta_u - \beta_z) - \log(-\alpha - \beta_u) - \log(-\alpha - \beta_z) + \log(-\alpha) \) are identifiable.

- Noncanonical link \( g(t) = \log(t) \). Hence, \( \psi(t) = -e^{-t} \) and \( \tau(t) = t \). Only \( \beta_u \) and \( e^{-\alpha - \beta_z} - e^{-\alpha} \) are identifiable.

- Noncanonical link \( g(t) = t^\gamma, -1 < \gamma < 0 \). Hence, \( \psi(t) = -t^{-\frac{1}{\gamma}} \) and \( \tau(t) = \frac{1}{\gamma} \log(t) \). If condition (C6) is satisfied, then \( \alpha \), \( \beta_u \) and \( \beta_z \) are all identifiable.

5.5 Asymptotic Theory

In this section, we provide the consistency and asymptotic normality for the maximum approximate conditional likelihood estimator. For simplicity, we only consider the missing mechanism \([R = 1|Y, X] = s(Y)t(X)\).

The maximum approximate conditional likelihood estimator \( \hat{\theta}_n \) is obtained through maximizing \( L_c(\theta) \), which is the same as maximizing the following approximate condi-
tional log-likelihood

\[ l_c(\theta) = \sum_{1 \leq i < j \leq n} -\log(1 + R_{ij}(\theta)). \]

It can be seen that \( (n^2 - 1)l_c(\theta) \) is a U-statistic of order \( m = 2 \) for any \( \theta \in \Theta \). Also, \( \hat{\theta}_n \) is the solution of the following score equation

\[ \frac{\partial l_c(\theta)}{\partial \theta} = \sum_{1 \leq i < j \leq n} -\frac{1}{1 + R_{ij}(\theta)} \frac{\partial R_{ij}(\theta)}{\partial \theta} = 0. \]

**Theorem 12.** Under the identifiability conditions stated in Theorem 1 and some other regularity conditions as follows: the parameter space \( \Theta \) is compact, the true value \( \theta_0 \in \Theta^0 \), the interior of \( \Theta \), \( E_{\theta_0} |\log(1 + R_{12}(\theta))| < \infty \) for any \( \theta \in \Theta \). Then \( \hat{\theta}_n \) is consistent, i.e., \( \hat{\theta}_n \overset{p}{\to} \theta_0 \), as \( n \to \infty \).

**Theorem 13.** Under the identifiability conditions stated in Theorem 1 and some other regularity conditions as follows: the parameter space \( \Theta \) is compact, the true value \( \theta_0 \in \Theta^0 \), the interior of \( \Theta \), \( E_{\theta_0} |\log(1 + R_{12}(\theta))|^2 < \infty \) for any \( \theta \in \Theta \). Then, \( \hat{\theta}_n \) is asymptotically normal, i.e., \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, \Sigma) \), as \( n \to \infty \), where \( \Sigma \) is defined in the appendix.

**Remark 14.** In this section, we study the asymptotic results based on the conditions under which the parameter \( \theta \) is fully identifiable. When the parameter is only partially identifiable, for example, in the GLM canonical link case, we can only identify \( \frac{\partial}{\partial \theta} \), the asymptotic results are still valid, but for the parameters that can be identified.

### 5.6 Proofs of Theorems

**Proof of Lemma 1.** From \( e(x_1, y_1) + e(x_2, y_2) = e(x_1, y_2) + e(x_2, y_1) \), we have

\[ e(x_1, y_1) - e(x_1, y_2) = e(x_2, y_1) - e(x_2, y_2), x_1 \neq x_2, y_1 \neq y_2. \]

Define \( h(x, y_1, y_2) = e(x, y_1) - e(x, y_2) \), we know that \( h \) is constant to \( x \), i.e.,

\[ \frac{\partial}{\partial x} e(x, y_1) = \frac{\partial}{\partial x} e(x, y_2), y_1 \neq y_2, \]
similarly,
\[ \frac{\partial^2}{\partial x \partial y} e(x, y) = 0. \]

Proof of Theorem 1. For density function \( \frac{1}{1+R_{ij}(\theta)} \) (up to a constant), from Jensen’s inequality, we have
\[ E_0 \left[ -\log \frac{1}{1 + R_{ij}(\theta_0)} \right] \leq E_0 \left[ -\log \frac{1}{1 + R_{ij}(\theta)} \right], \]
and “=” holds if and only if \( R_{ij}(\theta_0) = R_{ij}(\theta) \) a.s. under the truth \( \theta_0 \),
\[ \Leftrightarrow \frac{p(y_j|x_i;\theta_0)p(y_i|x_j;\theta_0)}{p(y_i|x_i;\theta_0)p(y_j|x_j;\theta_0)} = \frac{p(y_j|x_i;\theta)p(y_i|x_j;\theta)}{p(y_i|x_i;\theta)p(y_j|x_j;\theta_0)} \]
\[ \Leftrightarrow \frac{p(y_j|x_i;\theta_0)}{p(y|x;\theta)} = \exp\{f(x) + g(y)\}. \]

By assumption, \( P(D_\theta) < 1 \) for any \( \theta \neq \theta_0 \), \( P(R = 1) > 0 \), hence,
\[ E_0 \left[ 1_{(R=1)} \log \frac{1}{1 + R_{ij}(\theta_0)} \right] > E_0 \left[ 1_{(R=1)} \log \frac{1}{1 + R_{ij}(\theta)} \right], \theta \neq \theta_0. \]

Proof of Corollary 1. Under the given assumption, it is straightforward that \( h_1(x, y; \tilde{\theta}_1) \) can be identified through pairwise condition likelihood (8). Therefore, \( \tilde{\theta}_1 \) is identifiable.

Proof of Theorem 2. We only need to prove (i). Firstly, through pairwise approximate conditional likelihood (8), we can identify \( p + 2 \) real-valued functions: \( \psi_i(\theta), i = 1, \cdots, p + 2 \).

Under the given assumption, from implicit function theorem, there exist neighborhoods
\[ U = B(\theta_0, \epsilon) \subset \Theta, V = B((\psi_1(\theta_0), \cdots, \psi_{p+2}(\theta_0))^\top, \eta) \subset \mathbb{R}^{p+2}, \epsilon, \eta > 0, \]
and uniquely defined function $g = (g_1, \ldots, g_{p+2})^\tau$ on $V$, such that every component $g_i$ is first order continuously differentiable, and

$$\theta = g(\psi_1(\theta), \ldots, \psi_{p+2}(\theta)),$$

where $(\psi_1(\theta), \ldots, \psi_{p+2}(\theta))^\tau \in V$ and $\theta \in U$. Therefore,

$$\theta_0 = g(\psi_1(\theta_0), \ldots, \psi_{p+2}(\theta_0)),$$

and the parameter $\theta$ is identifiable.

**Proof of Corollary 2.** We only prove (iv). We show that, the solution for the matrix $M_4$ to be singular at any values of $\alpha$ and $\beta$ can be fully derived. This is a sufficient condition for the matrix $M_4$ to be singular only at the true values $\alpha_0, \beta_0$. Denote $t_0 = \alpha + \beta x_0$, $\xi_1 = \beta(x_1 - x_0)$, and $c = \frac{x_2 - x_0}{x_1 - x_0}, c \neq 1, \neq 0$, then

$$0 = \begin{vmatrix} \psi'(\alpha + \beta x_1) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_1)x_1 - \psi'(\alpha + \beta x_0)x_0 \\ \psi'(\alpha + \beta x_2) - \psi'(\alpha + \beta x_0) & \psi'(\alpha + \beta x_2)x_2 - \psi'(\alpha + \beta x_0)x_0 \\ \psi'(t_0 + \xi_1) - \psi'(t_0) & \psi'(t_0 + \xi_1)(x_1 - x_0) + (\psi'(t_0 + \xi_1) - \psi'(t_0))x_0 \\ \psi'(t_0 + c\xi_1) - \psi'(t_0) & \psi'(t_0 + c\xi_1)(x_2 - x_0) + (\psi'(t_0 + c\xi_1) - \psi'(t_0))x_0 \\ \psi'(t_0 + \xi_1) - \psi'(t_0) & \psi'(t_0 + \xi_1) \\ \psi'(t_0 + c\xi_1) - \psi'(t_0) & c\psi'(t_0 + c\xi_1) \end{vmatrix},$$

which is equivalent to

$$(c - 1)f(t_0) + f(t_0 + c\xi_1) = cf(t_0 + \xi_1),$$

where $\frac{1}{\psi'(\cdot)} = f(\cdot)$. Firstly, let $t_0 = 0$, it leads to $cf(\xi_1) = (c - 1)f(0) + f(c\xi_1)$, and

$$(c - 1)g(t_0) + g(t_0 + c\xi_1) = g(ct_0 + c\xi_1),$$
where $g(\cdot) = f(\cdot) - f(0)$. Then, let $\xi_1 = -t_0/c$, it leads to $(c - 1)g(t_0) = g((c - 1)t_0)$, and finally we have

$$g((c - 1)t_0) + g(t_0 + c\xi_1) = g(ct_0 + c\xi_1),$$

which is Cauchy’s functional equation. From Lemma 2, over the rational numbers, the solution is $g(u) = g(1)u$, the linear function with zero intercept. Over the real numbers, under some assumptions, e.g., $g$ is continuous, or monotone, or bounded, the solution is still the linear function with zero intercept $g(u) = g(1)u$. Go back to function $\psi'$, the solution is

$$\psi'(x) = \frac{1}{kx + c}, k \neq 0, \text{ or } \psi'(x) = c.$$

This completes the proof. \qed

Proof of Theorem 3. For density function $\frac{1}{1 + R_{ij}(\theta, \xi)}$ (up to a constant), from Jensen’s inequality, we have

$$E_0 \left[ -\log \frac{1}{1 + R_{ij}(\theta_0, \xi_0)} \right] \leq E_0 \left[ -\log \frac{1}{1 + R_{ij}(\theta, \xi)} \right],$$

and "=” holds if and only if $R_{ij}(\theta_0, \xi) = R_{ij}(\theta, \xi)$ a.s. under the truth $\theta_0, \xi_0$,

$$\Leftrightarrow \frac{p_{yu}(y_j, u_j|z_i; \theta_0, \xi_0)p_{yu}(y_i, u_i|z_j; \theta_0, \xi_0)}{p_{yu}(y_i, u_i|z_j; \theta, \xi)p_{yu}(y_j, u_j|z_i; \theta, \xi)} = \frac{p_{yu}(y_j, u_j|z_i; \theta, \xi)p_{yu}(y_i, u_i|z_j; \theta, \xi)}{p_{yu}(y_i, u_i|z_j; \theta, \xi)p_{yu}(y_j, u_j|z_i; \theta, \xi)},$$

$$\Leftrightarrow \frac{p_{yu}(y, u|z; \theta_0)}{p_{yu}(y, u|z; \theta)} = \exp\{f(y, u) + g(z)\},$$

$$\Leftrightarrow \frac{p(y|u, z; \theta_0)}{p(y|u, z; \theta)} = \exp\{f_1(y, u) + g_1(z)\}, \frac{p_u(u|z; \xi_0)}{p_u(u|z; \xi)} = \exp\{f_2(u) + g_2(z)\},$$

where $f(y, u) = f_1(y, u) + f_2(u)$, $g(z) = g_1(z) + g_2(z)$. By assumption, $P(D_{\theta, \xi}) < 1$ for any $\theta \neq \theta_0, \xi \neq \xi_0, P(R = 1) > 0$, hence,

$$E_0 \left[ 1_{(R=1)} \log \frac{1}{1 + R_{ij}(\theta_0, \xi_0)} \right] > E_0 \left[ 1_{(R=1)} \log \frac{1}{1 + R_{ij}(\theta, \xi)} \right], \theta \neq \theta_0, \xi \neq \xi_0.$$
Proof of Consistency. Firstly, by the theory of U-statistics, we have
\[
\binom{n}{2}^{-1} l_{c}(\theta) \xrightarrow{p} E[-\log\{1 + R_{12}(\theta)\}], \forall \theta \in \Theta.
\]
Notice that
\[
E[-\log\{1 + R_{12}(\theta)\}] = E[E(-\log(1 + R_{12}(\theta))|y(1), y(2), x_1, x_2)].
\]
By the conditional Kullback-Leibler information inequality (Anderson 1970),
\[
E\{-\log(1 + R_{12}(\theta))|y(1), y(2), x_1, x_2}\}
\]
achieves the maximum value when \(\theta = \theta_0\), the true value of \(\theta\). By identifiability theorem, \(\theta_0\) is the unique maximizer.

Notice that
\[
\binom{n}{2}^{-1} \frac{\partial l_{c}}{\partial \theta}(\theta) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} -\frac{1}{1 + R_{ij}(\theta)} \frac{\partial R_{ij}}{\partial \theta}(\theta).
\]
Since \(-\frac{1}{1 + R_{ij}(\theta)} \frac{\partial R_{ij}}{\partial \theta}(\theta)\) is continuous on \(\Theta\), it is bounded. Suppose \(|-\frac{1}{1 + R_{ij}(\theta)} \frac{\partial R_{ij}}{\partial \theta}(\theta)| < M\), then we have
\[
| -\log(1 + R_{ij}(\theta)) - (-\log(1 + R_{ij}(\tilde{\theta})))| \leq M\|\theta - \tilde{\theta}\|, \forall \theta, \tilde{\theta} \in \Theta,
\]
where \(\|\cdot\|\) indicates the Euclidean norm. Thus we have the following
\[
\left| \binom{n}{2}^{-1} l_{c}(\theta) - \binom{n}{2}^{-1} l_{c}(\tilde{\theta}) \right| \leq \binom{n}{2}^{-1} M\|\theta - \tilde{\theta}\|,
\]
and the following stochastic equicontinuity:
\[
\lim_{n \to \infty} P \left( \sup_{\theta \in \Theta} \sup_{\tilde{\theta} \in B(\theta, \delta)} \left| \binom{n}{2}^{-1} l_{c}(\theta) - \binom{n}{2}^{-1} l_{c}(\tilde{\theta}) \right| > \epsilon \right) < \delta, \forall \epsilon, \delta > 0.
\]
Hence, by Corollary 2.2 in Newey (1991), we have the following uniform convergence:
\[
\sup_{\theta \in \Theta} \left| \binom{n}{2}^{-1} l_{c}(\theta) - E\{-\log(1 + R_{12}(\theta))\} \right| \xrightarrow{p} 0.
\]
Therefore, we have \(\hat{\theta}_n \xrightarrow{p} \theta_0\), as \(n \to \infty\). \(\square\)
Proof of Asymptotic Normality. Firstly Taylor expansion gives us
\[
0 = \frac{\partial l_c(\hat{\theta}_n)}{\partial \theta} = \frac{\partial l_c(\theta_0)}{\partial \theta} + (\hat{\theta}_n - \theta_0) \frac{\partial^2 l_c}{\partial \theta^2}(\theta_0) + o_p(n^{-3/2}),
\]
therefore, we have
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left[ \left( \frac{n}{2} \right)^{-1} \frac{\partial^2 l_c}{\partial \theta^2}(\theta_0) \right]^{-1} \sqrt{n} \left( \frac{n}{2} \right)^{-1} \frac{\partial l_c(\theta_0)}{\partial \theta} + o_p(1).
\]
Since \( \left( \frac{n}{2} \right)^{-1} \frac{\partial^2 l_c}{\partial \theta^2}(\theta_0) \) and \( \left( \frac{n}{2} \right)^{-1} \frac{\partial l_c(\theta_0)}{\partial \theta} \) are both U-statistics, from the theory of U-statistics, we have the following
\[
\left( \frac{n}{2} \right)^{-1} \frac{\partial^2 l_c}{\partial \theta^2}(\theta_0) \xrightarrow{p} - E \left[ \frac{\partial}{\partial \theta} \left( \frac{1}{1 + R_{12}} \frac{\partial R_{12}}{\partial \theta} \right) \right],
\]
\[
\sqrt{n} \left( \frac{n}{2} \right)^{-1} \frac{\partial l_c(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, 4\zeta_1).
\]
Therefore,
\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma),
\]
where \( \Sigma \) is defined as above. \( \square \)
Bibliography


