TOPICS IN SEMIPARAMETRIC REGRESSION ESTIMATION WITH MISSING COVARIATES USING SINGLE-INDEX MODELS

A Dissertation

by

ZHUOER SUN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Suojin Wang
Committee Members,	Jim Ji
	Samiran Sinha
	Lan Zhou
Head of Department,	Valen Johnson

August 2018

Major Subject: Statistics

ABSTRACT

Missing data are very common in many areas such as sociology, biomedical sciences and clinical trials. Simply ignoring the incomplete cases may cause bias in estimation procedures. In this dissertation we investigate semiparametric estimation of linear regression coefficients through generalized estimating equations with single-index models when some covariates are missing at random for both independent and identically distributed (i.i.d.) data and longitudinal data. Existing popular semiparametric estimators by weighted estimating equations may run into difficulties when some selection probabilities are small or the dimension of the covariates is not low.

For i.i.d. data, we propose a new simple parameter estimator using a kernel assisted estimator for the augmentation by a single-index model without using the inverse of selection probabilities. We explore the asymptotic efficiency of the proposed estimator and its relationships with existing estimators. In particular, we show that under certain conditions the proposed estimator is as efficient as the existing methods based on standard kernel smoothing, which are often practically infeasible in the case of multiple covariates.

For incomplete longitudinal data, we propose a similar estimator when the covariate is non-monotone missing at random. Heteroscedasticity is considered and working independence correlation structure is applied to simplify the estimation procedure. Asymptotic consistency and normality are derived along with sandwich formulas for asymptotic covariances.

The above methods are supported by simulation studies and real data examples. The numerical results show that the proposed estimators avoid some numerical issues caused by estimated small selection probabilities that are needed in other estimators.

DEDICATION

To my parents, my grandmother, and to the memory of my grandfather.

ACKNOWLEDGMENTS

I would like to sincerely thank my research advisor, Dr. Suojin Wang, for his guidance and patience on my research projects during the past five years. Dr. Wang has provided me with so much help and support whenever I ran into difficulties. I learned a lot from him, not only the knowledge for research in statistics, but also the attitude towards academics and life. Without his persistent help and guidance, this dissertation would not have been possible.

I would like to thank my committee members, Dr. Jim Ji, Dr. Samiran Sinha and Dr. Lan Zhou, for their valuable advice on my research projects and presentation skills.

I would like to thank my parents and my grandparents, who supported me all the time and gave me all the power I needed. Nobody means more to me than my families in the pursuit of my doctorate degree.

Finally I would like to thank all my friends for their listening and help all the way.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Suojin Wang (advisor), Samiran Sinha and Lan Zhou of the Department of Statistics and Professor Jim Ji of the Department of Electrical and Computer Engineering.

The methodologies and proofs in Chapter 3 and 4 were coauthored with Professor Suojin Wang.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a graduate assistantship from Department of Statistics, Texas A&M University.

NOMENCLATURE

AC	Available Case Analysis
AIPW	Augmented Inverse-Probability Weighted Estimator
AMSE	Asymptotic Mean Square Error
CC	Complete Case Analysis
DR	Doubly Robust
EM	Expectation-Maximization
GEE	Generalized Estimating Equation
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
I.I.D.	Independent and Identically Distributed
IPW	Inverse-Probability Weighted Estimator
MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Multiple Imputation
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
MS	Mean-Score Estimator
NW	Nadaraya-Watson
OLS	Ordinary Least Square
SE	Standard Error
SIM	Single-Index Model
WEE	Weighted Estimating Equation

WI	Working Independence
WLS	Weighted Least Square

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Regression analysis is a very wide and classic topic in statistics. It explores the relationship between the independent variable (response) and the dependent variables (covariates). It is useful for prediction and forecasting. There have been many parametric or nonparametric models and methods developed for different kinds of regression problems for the purpose of estimation, prediction, statistical inference and so on. However, when some of the data are missing due to some reasons, the commonly used methods and theory for complete data may not work well if you just simply ignore the incomplete cases. Unfortunately, the missing data problem is not uncommon in many fields, such as sociology, epidemiology, biomedical science, clinical trials and public health research. The reasons for missingness can be, but not limited to unavailability of measurements, physical loss of data, survey nonresponse, study subjects' refusal to answer the questions or to continue the participation, or even the patients' deaths. And it is even easier for us to have data partially missing when we access large volumes of data nowadays.

One example of independent and identically distributed (i.i.d.) data is the Canada 2010/2011 Youth Smoking Survey (YSS) data. The 2010/2011 YSS is a Health Canada sponsored pan-Canadian, classroom-based survey of a representative sample of students in grades 6 through 12. It was implemented in schools between October 2010 and June 2011 by provincial level teams located in the 9 participating provinces in Canada. The original dataset has many attributes, including smoking status, a score evaluating the student's self-esteem, age, sex, body mass index (BMI) and other features. An important research interest is to explore whether smoking will have influence on the student's self-esteem, controlling other variables. One possible way is to establish a regression model between the self-esteem score and the predictors, including the smoking status. However, we find the data for BMI may be missing for some students and the missing proportion is about 30%. Throwing away those incomplete cases definitely will result in a big loss of information and the results from standard regression methods may be incorrect.

In the case of longitudinal data, it is more often to encounter the situation of missing data be-

cause repeated measurements need to be made on each subject during the follow-up. For example, a double-blinded randomized trial in primary biliary cirrhosis of the liver (PBC) for comparing the drug D-penicillamine (DPCA) with a placebo was conducted by the Mayo Clinic between January, 1974 and May, 1984. PBC is a rare but fatal liver disease with a prevalence of about 50-cases-per-million population. There were 312 patients involved in this trial with information gathered routinely during the follow-up. The original dataset has 19 attributes including some demographic variables like age and sex, clinical measurements like the presence/absence of ascites, and biochemical measurements such as the levels of bilirubin, albumin. Since bilirubin is a very important prognostic factor in PBC (Shapiro et al. (1979)), we would like to explore the difference between the two treatment groups on bilirubin levels, controlling other variables such as cholesterol level and age. But in the original data, almost 40% of total measurements on cholesterol level are missing. For each patient, the data is partially missing, which means the availability of cholesterol level varies from the current visit to the next. This arbitrary pattern makes the missing data problem even more complicated.

Many research works have been done for linear regression models and generalized linear models (GLM). Fuchs (1982), Schluchter & Jackson (1989), Horton & Laird (1999) and Ibrahim (1990) proposed estimation procedure through maximum likelihood. These model-based methods are flexible and clear for inference, and the asymptotic properties can be obtained via the second derivatives of the log-likelihood.

Bayesian methods can be considered as another approach based on likelihood. One can find details of estimation in Ibrahim et al. (2002) and Daniels & Hogan (2008). But it can be challenging to correctly specify the conditional covariate distribution and the joint priors over parameters.

Multiple imputation (MI) might be the most popular approach in industry and software packages. The basic idea of MI is to impute missing data to create a new "complete" sample, and then analyze it as if it were a complete data set for multiple times. MI methods for linear regression models are discussed in Rubin (2004) and Little & Rubin (2014).

However, in most situations, all the above three methods depend on the specification of the

likelihood. They can be very sensitive to misspecification of the likelihood. To have more robust estimation with less likelihood assumptions, Robins et al. (1994) proposed a class of semiparametric estimators based on weighted estimating equations (WEE) under the missing at random (MAR) mechanism. The weights are the reciprocals of the conditional probabilities of the data being observed, so we call them inverse-probability weights. These probabilities can be modeled parametrically, such as a logistic model. Wang et al. (1997) proposed a nonparametric kernel smoother for these probabilities and developed asymptotic theory for the estimator. The WEE can also include the augmentation term, which makes use of the data in the incomplete cases and can always improve the efficiency. Again parametric models can be assumed on this augmentation, and Wang & Wang (2001) showed a nonparametric kernel estimator for it.

These semiparametric estimators look reasonable and more flexible than likelihood-based methods, but there are two main problems for them. The first one is that the inverse-probability weights can be highly variable and skewed when the probabilities are positive but near zero. The second problem is that we may suffer from "curse of dimensionality" when we use multivariate kernel functions. Therefore, in this dissertation we will investigate the following questions for both i.i.d. data and longitudinal data:

1. How to effectively estimate the parameters in the linear model semiparametrically without inverse-probabilities when covariates are MAR.

2. How to estimate the augmentation nonparametrically on a low-dimension basis.

3. How efficient the parameter estimators are compared to the existing methods.

The following chapters are organized as follows. In Chapter 2, we review the basic missing data concepts, useful longitudinal models and existing methods for handling missing data. In Chapter 3, we propose a new semiparametric estimator for i.i.d. data without using inverse-probability weights. We study the properties of the proposed estimator through theories and numerical studies. In Chapter 4, we extend the method to longitudinal data. The methodology is supported by

theorems, simulations and a real data example. We make conclusion remarks for this dissertation in Chapter 5.

2. LITERATURE REVIEW

In this chapter, we review the main concepts and models in missing data analysis, longitudinal data analysis, nonparametric regression and semiparametric models. In Section 2.1, we introduce the concepts of different missing mechanisms and missing patterns. In Section 2.2, we briefly introduce three commonly used models in longitudinal data analysis: marginal models, random effects models and transition models. In Section 2.3, we discuss some particular nonparametric kernel smoothers and their asymptotic properties. In Section 2.4, we provide a brief review of the literature of existing methods for handling missing data in both i.i.d. data and longitudinal data, especially the semiparametric methods based on estimating equations.

2.1 Missing Data Concepts

2.1.1 Missing Mechanisms

Missing mechanisms are important to the missing data analysis. Different missing mechanisms lead to different methodologies for handling the incomplete data problem. The concept was formalized first by Rubin (1976), and explained in details by Little & Rubin (2014), among other authors.

Missing mechanisms are first discussed on missing response in the data. Let Y denote the complete response data and R be the missing data indicator. Let Y_{obs} denote the observed components of Y, and Y_{mis} the missing components. The missing mechanism concepts are established on the conditional distribution of R given Y as $f(R|Y,\phi)$, where ϕ denotes unknown parameters. If

$$f(R|Y,\phi) = f(R|\phi) \ \text{ for all } \ Y,\ \phi,$$

that is, the missingness does not depend on any values of the response data Y, no matter Y_{obs} or Y_{mis}. Then this missing mechanism is called missing completely at random (MCAR). We can have a more general assumption that the missingness can depend on the observed part Y_{obs}, but not on

the missing part Y_{mis}, that is,

$$f(R|Y,\phi) = f(R|Y_{\text{obs}},\phi) \text{ for all } Y_{\text{mis}}, \phi.$$

This missing mechanism is called missing at random (MAR). Then the last missing mechanism is called missing not at random (MNAR) if the distribution of R can also depend on the missing values Y_{mis}, which violates the assumptions of MCAR and MAR.

If we consider i.i.d. univariate data as $Y = (y_1, ..., y_n)^\top$, $R = (R_1, ..., R_n)^\top$, then we have

$$f(Y,R|\theta,\phi) = f(Y|\theta)f(R|Y,\phi) = \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} f(R_i|y_i,\phi),$$

where θ denotes the unknown parameters for the density of y_i as $f(y_i|\theta)$. In this situation, MCAR is equivalent to MAR because both of them imply $f(R_i|y_i,\phi) = f(R_i|\phi)$.

However, we will have a different conclusion when we consider missing covariate data. Consider the regression analysis of a response, Y, on a set of covariates (X, Z) with covariate X missing for some subjects. Then MCAR means $Pr(R_i = 1|y_i, X_i, Z_i) \equiv p$ with a constant p, while MAR implies $Pr(R_i = 1|y_i, X_i, Z_i) = Pr(R_i = 1|y_i, Z_i) = \pi(Q_i)$ with a function $\pi(\cdot)$ and $Q_i = (y_i, Z_i)^\top$. For longitudinal data, let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{im_i})^\top$ denote the $m_i \times 1$ completely observed response vector for subject i, $\mathbf{X}_i = (X_{i1}, ..., X_{im_i})^\top$ be the covariate vector that may be missing at some time point t_{ij}, $\mathbf{Z}_i = (\boldsymbol{Z}_{i1}, ..., \boldsymbol{Z}_{im_i})^\top$ be the covariate matrix that is always observed and $\mathbf{R}_i = (R_{i1}, ..., R_{im_i})^\top$ be the indicator vector of subject i. Then MAR means

$$f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{\gamma}) = f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i^{(o)}, \mathbf{Z}_i, \boldsymbol{\gamma}),$$

where $\mathbf{X}_i^o$ denotes the observed part of $\mathbf{X}_i$ and $\boldsymbol{\gamma}$ denotes the unknown parameters.

2.1.2 Missing Pattern

Since we consider both cases of a single covariate being missing and several covariates being missing at the same time in this dissertation, to distinguish from the missing pattern in the situation

of several covariates missing separately in i.i.d. data, the following concepts of missing pattern are only applied to incomplete longitudinal data, mainly described in Diggle (2002).

The missing pattern is monotone (or *dropouts*) if whenever X_{ij} is missing, so are X_{ik} for all $k \geq j$; otherwise we say the missing pattern is non-monotone (or *intermittent*). Dropouts are common in clinical trials data because the subject leaves the study and never comes back. This can be caused by the subject's worse health condition or even death. In the current study, it is not reasonable to assume that the covariate X_{ij} has the monotone missing pattern while the responses Y_{ij}'s are fully observed. This motivates us to consider non-monotone missingness, which is always more difficult than monotone missingness to make the factorization of the likelihood.

2.2 Models for Longitudinal Data Analysis

In this section, we briefly introduce the three most commonly used models for longitudinal data analysis. For this dissertation we focus on linear models, so the models will be introduced based on generalized linear models (GLM). Details about these models are introduced systematically in Diggle (2002).

2.2.1 Marginal Models

The basic assumption for marginal models of longitudinal data is that the regression on predictors is modeled separately from within-subject correlation. That is,

$$E(Y_{ij}|\mathbf{X}_i) = E(Y_{ij}|\boldsymbol{X}_{ij}) = \mu_{ij},$$

where Y_{ij} is the response variable for the jth measurement of subject i, $\boldsymbol{X}_{ij}$ is the set of predictors and $\mathbf{X}_i = (\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{im_i})$, μ_{ij} is the marginal mean of the response given the predictors, $i = 1, ..., n$, $j = 1, ..., m_i$. With GLM, we have $h(u_{ij}) = \boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}$, $Var(Y_{ij}) = \nu(\mu_{ij})\phi$, where $h(\cdot)$ is a known link function, $\nu(\cdot)$ is a known variance function and ϕ is an unknown scale parameter. The within-subject correlation is a function of marginal means with additional parameters $\boldsymbol{\alpha}$, that is, $\mathrm{Cor}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha})$, where $\rho(\cdot)$ is a known function. Note that for linear model, the variance is independent of the mean.

The marginal model is the approach we take for analyzing incomplete longitudinal data in Chapter 4. When the probability distribution of the response is available, the above necessary functions are known and normal maximum likelihood methods can be applied to obtain the estimates of $\boldsymbol{\beta}$. The parameterization of the likelihood for binary and count data can be found in Bishop et al. (1977) and Fitzmaurice & Laird (1993). However, that information is not always available or can be correctly specified. Liang & Zeger (1986) proposed to estimate $\boldsymbol{\beta}$ through the following generalized estimating equations (GEE) as

$$n^{-1/2}\sum_{i=1}^{n}\frac{\partial \boldsymbol{\mu}_i^{\top}}{\partial \boldsymbol{\beta}}\mathrm{Var}(\mathbf{Y}_i)^{-1}(\mathbf{Y}_i-\boldsymbol{\mu}_i)=0,$$

where $\boldsymbol{\mu}_i=(\mu_{i1},...,\mu_{im_i})^{\top}$. Liang & Zeger (1986) showed that the solution of GEE is consistent for $\boldsymbol{\beta}$ provided that the model for marginal mean μ_{ij} is correctly specified even if the covariance $\mathrm{Var}(\mathbf{Y}_i)$ is misspecified. Write

$$\mathrm{Var}(\mathbf{Y}_i)=\mathbf{F}_i^{1/2}\mathbf{C}_i(\boldsymbol{\rho})\mathbf{F}_i^{1/2},$$

where $\mathbf{F}_i=\mathrm{diag}(\sigma_{ij}^2)$, $\mathbf{C}_i(\boldsymbol{\rho})$ is called the working correlation matrix with parameters $\boldsymbol{\rho}$ and a particular form such as uniform correlation or AR(1).

2.2.2 Random Effects Models

Random effects models consider that the response is assumed to be a linear function of predictors with regression coefficients varying from one individual to the next. Using the GLM framework, we assume that conditional on unobservable variables $\boldsymbol{U}_i$, Y_{ij}'s are independently drawn from the exponential family $f(Y_{ij}|\boldsymbol{U}_i;\boldsymbol{\beta})$ such that

$$h\{E(Y_{ij}|\boldsymbol{U}_i)\}=\boldsymbol{X}_{ij}^{\top}\boldsymbol{\beta}+\boldsymbol{d}_{ij}^{\top}\boldsymbol{U}_i,$$

where $\boldsymbol{U}_i$ are $q\times 1$ i.i.d. random variables from certain density $f(\boldsymbol{U}_i)$, usually a Gaussian distribution with mean 0 and variance $\mathbf{G}$; $\boldsymbol{d}_{ij}$ are q-element vectors of predictors attached to each

measurement. If we ignore the serial correlation and measurement errors in the responses, then $\boldsymbol{U}_i$ is the main source of within-subject correlation. Consider a linear model,

$$Y_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{d}_{ij}^\top \boldsymbol{U}_i + \epsilon_{ij},$$

where ϵ_{ij} are independent Gaussian random noise with mean 0 and variance σ^2. Then

$$\mathrm{Var}(\mathbf{Y}_i) = \mathbf{D}_i \mathbf{G} \mathbf{D}_i^\top + \sigma^2 \mathbf{I}_i$$

with $\mathbf{D}_i = (\boldsymbol{d}_{i1}, ..., \boldsymbol{d}_{im_i})^\top$. There are some particular examples for the random effects model. If we have $\boldsymbol{d}_{ij} = \boldsymbol{X}_{ij}$, then each individual can be thought to have their own regression coefficient $\boldsymbol{\beta} + \boldsymbol{U}_i$. If $q = 1$ with $d_{ij} = 1$, then the model actually means a random intercept regression model. Based on the exponential family distribution of response and the density of $f(\boldsymbol{U}_i; \mathbf{G})$, the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ can be obtained from the marginal distribution of $\mathbf{Y}$ as

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{n} \int \prod_{j=1}^{m_i} f(Y_{ij} | \boldsymbol{U}_i; \boldsymbol{\beta}) f(\boldsymbol{U}_i; \mathbf{G}) d\boldsymbol{U}_i,$$

where $\boldsymbol{\theta}$ includes $\boldsymbol{\beta}$ and parameters in $\mathbf{G}$. If we do not have a closed form for the integral, the expectation-maximization (EM) algorithm can be used to get the MLE of $\boldsymbol{\theta}$. In this sense, $\boldsymbol{U}_i$ can be regarded as a latent variable and $L(\boldsymbol{\theta}; \mathbf{Y})$ is the observed likelihood.

2.2.3 Transition Models

Under a transition model, we consider the conditional distribution of each response Y_{ij} given the past responses $\mathcal{H}_{ij} = (Y_{i1}, ..., Y_{i,j-1})$ and covariates $\boldsymbol{X}_{ij}$. These conditional distributions also intriduce the correlation between Y_{ij}'s. Under GLM,

$$h(\mu_{ij}^C) = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \sum_{r=1}^{s} f_r(\mathcal{H}_{ij}; \boldsymbol{\alpha}),$$

where $\mu_{ij}^C = E(Y_{ij}|\mathcal{H}_{ij})$. For a linear regression with autoregressive errors for Gaussian data, the transition model actually means a Markov model, as

$$Y_{ij} = \boldsymbol{X}_{ij}^\top \boldsymbol{\beta} + \sum_{r=1}^{s} \alpha_r (Y_{i,j-r} - \boldsymbol{X}_{i,j-r}^\top \boldsymbol{\beta}) + \epsilon_{ij}.$$

In this way, the joint distribution of $\mathbf{Y}$ can be easily factored as

$$f(\mathbf{Y}) = f(Y_{i1}) \prod_{j=2}^{m_i} f(Y_{ij}|\mathcal{H}_{ij}).$$

More materials about transition models for categorical data and the methods of estimation can be found in Diggle (2002).

2.3 Kernel Smoother

In this section we briefly review the nonparametric kernel regression techniques which is used later for estimating the conditional expectations in this dissertation. There is abundant literature on this topic and we only focus on the local constant weighted smoother here.

2.3.1 Nadaraya-Watson Estimator

Consider the regression problem with dependent variable Y and independent variable X. Generally, we assume

$$y_i = m(x_i) + \varepsilon_i, \;\; i = 1, ..., n,$$

where ε is the white noise, and $m(\cdot)$ is usually a smooth function. Then $E(Y|X = x) = m(x)$. To reveal the relationship between Y and X flexibly, we want to estimate the function $m(\cdot)$ nonparametrically. Based on the truth that

$$m(x) = E(Y|X = x) = \int y \frac{f(x, y)}{f(x)} dy = \frac{\int y f(x, y) dy}{f(x)}$$

and the kernel density estimation for the joint density $f(x, y)$ and $f(x)$ with a kernel K, Nadaraya (1964) and Watson (1964) proposed the Nadaraya-Watson (NW) estimator as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i K_h(x - x_i)}{\sum_{i=1}^{n} K_h(x - x_i)},$$

where $K_h(\cdot) = K(\cdot/h)$ with h the bandwidth. It is essentially a local constant estimator. The convergence of $\hat{m}(x)$ has been shown by Härdle (1990) and Gasser & Müller (1984) derived the asymptotic mean square error (AMSE) for this estimator. For simplicity, with a regular second-order kernel function $K(u)$ and $h \to 0$, $nh \to \infty$, we have

$$AMSE(\hat{m}(x)) = \frac{1}{nh} C_1 + h^4 C_2,$$

where C_1 and C_2 denote some finite constants. Then minimizing AMSE with respect to the h leads to the optimal bandwidth rate as $h_{opt} = O(n^{-1/5})$ and the corresponding AMSE of order $O(n^{-4/5})$.

2.3.2 Single-index Model

For the above NW estimator, we will have the boundary issue as the metric neighborhoods tend to contain less points on boundaries, which may lead to bias. When X is multi-dimensional, the boundary effects are even more severe (Friedman et al. (2001)). We will need to use some proper multivariate kernels, but still suffer from "curse of dimensionality". One possible approach to overcome the difficulty of high-dimensional X is to use the single-index model (SIM) for dimension reduction by Ichimura (1993). Assume the smooth function $m : \mathbb{R}^p \to \mathbb{R}$ of $X \in \mathbb{R}^p$ has a particular form as a function of the linear combination of X, that is,

$$y_i = g(\theta^\top x_i) + \varepsilon_i,$$

where $g : \mathbb{R} \to \mathbb{R}$ is a univariate smooth function, θ is a p-elements unknown index vector. For identifiability, we assume the first non-zero element of θ is positive 1. Following the idea of the

NW estimator, one estimator of g can be

$$\hat{g}(u|\theta) = \frac{\sum_{j=1}^{n} y_j K_h(u - \theta^\top x_j)}{\sum_{j=1}^{n} K_h(u - \theta^\top x_j)}.$$

Härdle et al. (1993) discussed the estimation of θ and the choice of the bandwidth h. A leave-one-out estimator of g is

$$\hat{g}_i(u|\theta) = \frac{\sum_{j \neq i} y_j K_h(u - \theta^\top x_j)}{\sum_{j \neq i} K_h(u - \theta^\top x_j)}.$$

Then define

$$\hat{S}(\theta, h) = \sum_{i=1}^{n} \{y_i - \hat{g}_i(\theta^\top x_i|\theta)\}^2,$$

so that one can use cross-validation to minimize $\hat{S}(\theta, h)$ with respect to (θ, h). Härdle et al. (1993) showed that the resulting estimator $\hat{\theta}$ is a root-n consistent estimate of θ, and the optimal rate for h is still $O(n^{-1/5})$.

2.4 Existing Methods for Handling Missing Covariate Data

In this section, we briefly introduce the existing methods for missing covariate data in both univariate i.i.d. data and longitudinal data. The main focus will be on the semiparametric methods based on GEE under MAR mechanism. These reviews depend on Ibrahim et al. (2005) and Ibrahim & Molenberghs (2009).

2.4.1 Methods for i.i.d Data

Consider the regression analysis of a response, Y, on a set of covariates $(X, Z^\top)^\top$ as a linear regression model $Y = W\beta + \varepsilon$, where $W = (1, X, Z^\top)^\top$ is the covariate vector, $E(\varepsilon|W) = 0$, and the covariate X may be missing for some subjects.

Much work using the maximum likelihood has been developed in the area of missing covariate regression analysis. This model-based method is flexible and clear for inference, and the asymptotic properties can be obtained via the second derivatives of the log-likelihood. When the observed likelihood is not available in closed form because of difficulty of the multi-dimensional integral, the EM algorithm is a popular technique for obtaining the maximum likelihood estimate with ig-

norable missing categorical or continuous covariates (Fuchs (1982), Schluchter & Jackson (1989) and Ibrahim (1990)).

Another likelihood-based approach is Bayesian methods, which are straightforward in terms of concepts and inferences. On the other hand, it can be challenging to correctly specify the conditional covariate distribution $f(X_i|y_i, Z_i; \gamma)$ and the joint priors over parameters (β, γ). Ibrahim et al. (2002) considered Bayesian methods for MAR covariates in GLM with informative prior based on historical data.

The most popular approach in industry and software packages might be multiple imputation (MI). The basic idea of MI is to impute missing data multiple times to create M "complete" datasets, then analyze them and summarize the results to have a final estimate (usually by taking average of the results from the M datasets). MI methods for MAR covariates in linear regression models are discussed in Rubin (2004) and Little & Rubin (2014). In terms of specifying the conditional covariate distribution, MI works similarly to the EM algorithm and the motivation is Bayesian, but the idea of MI itself is quite general and can be applied to other methods (Ibrahim et al. (2005)). Hsu et al. (2014) proposed a nearest neighbor-based nonparametric multiple imputation approach for missing covariate data by using the distance calculated from a two-dimensional summary score of information about the missing covariate and the missingness indicator.

In most situations, all the above three methods depend on the specification of the likelihood. When the distributional assumptions are correct, they are optimal. However, they can give biased estimators when the assumptions are violated. To have more robust estimation with less likelihood assumptions, some semiparametric methods such as weighted estimating equations (WEE) are proposed. Robins et al. (1994) proposed a class of semiparametric estimators based on inverse-probability WEE

$$\Delta(\beta, \psi) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi_i} D_i(W_i)(y_i - W_i^\top \beta) + \left(1 - \frac{R_i}{\pi_i}\right) \psi_i(Q_i) \right\} = 0,$$

where n is the total sample size (including incomplete cases), $Q_i = (y_i, Z_i^\top)^\top$, $D_i(W_i)$ is a func-

tion satisfying a local identification condition as nonsingular $E\{D_i(W_i)W_i^\top\}$, $\psi_i(Q_i)$ is an arbitrary function of Q_i, R_i is the binary indicator such that $R_i = 1$ if the covariate X_i is observed and $R_i = 0$ otherwise, $\pi_i = E(R_i|Q_i)$ is the probability of observing X_i. Robins et al. (1994) discussed the choices of D_i and ψ_i, and showed that there exist unique $D_i^{(eff)}$ and corresponding $\psi_i^{(eff)}(Q_i) = E\{D_i^{(eff)}(W_i)(y_i - W_i^\top\beta)|Q_i\}$ that can achieve the semiparametric efficiency bound of $\hat{\beta}$. However, $D_i^{(eff)}$ does not always have a closed form, so we choose a convenient one as $D_i = \partial\mu_i/\partial\beta = W_i$, where $\mu_i = E(y_i|X_i, Z_i)$. Thus the estimating equation is the GEE for this regression problem. Then $\psi_i(Q_i) = E(T_i|Q_i)$ with $T_i = W_i(y_i - W_i^\top\beta)$, the regular score function.

If we let $\psi_i \equiv 0$, the resulting estimator is the inverse-probability weighted estimator (IPW). We usually fit a logistic regression model of R_i on Q_i to estimate π_i's in the equations. When the logistic regression model is incorrect for π_i, the estimation can be biased. To overcome this difficulty, Wang et al. (1997) proposed a nonparametric kernel smoother for the selection probabilities as

$$\hat{\pi}(q) = \frac{\sum_{i=1}^{n} R_i K_{h_1}(q - Q_i)}{\sum_{i=1}^{n} K_{h_1}(q - Q_i)}, \tag{2.1}$$

where K is an rth-order kernel function, h_1 is the bandwidth parameter, and $K_{h_1}(\cdot) = K(\cdot/h_1)$. Wang et al. (1997) also developed asymptotic theory for the estimator, including the optimal bandwidth rate.

When $\psi_i \neq 0$, the estimator is the augmented inverse-probability weighted estimator (AIPW). This estimator can be more efficient than IPW because it also incorporates the incomplete cases. The most important advantage of AIPW is that it is doubly robust (DR) in the sense that the estimator will be consistent when either the selection probability model (π_i) or the augmentation model ($\psi_i(Q_i)$) is correctly specified. There are many works discussing the DR estimator and extensions, such as Bang & Robins (2005), Kang & Schafer (2007) and Robins & Ritov (1997). Extending Wang et al. (1997), Wang & Wang (2001) considered kernel estimation for both π_i and $\psi_i(Q_i)$, which avoids modeling them parametrically. The kernel estimator for ψ_i can be expressed

as

$$\hat{\psi}(q) = \frac{\sum_{i=1}^{n} R_i T_i K_{h_2}(q - Q_i)}{\sum_{i=1}^{n} R_i K_{h_2}(q - Q_i)} \tag{2.2}$$

with another kernel bandwidth h_2. It can share the same order of kernel function and same rate of bandwidth with $\hat{\pi}(q)$ in (2.1).

When we still have $\psi_i \neq 0$ but do not include the inverse-probability weights, we obtain the mean-score estimator (MS). Reilly & Pepe (1995) proposed this estimator when all the components of Q are discrete. Wang & Wang (2001) extended it to the setting where some components are continuous. When both the selection probability π_i and the augmentation ψ_i are estimated nonparametrically by the standard kernel smoother, Wang & Wang (2001) showed the asymptotic equivalence among IPW, AIPW and MS.

Although Robins et al. (1994) restricted π_i to be bounded away from 0, in practice you may still get some positive but near-zero values for the estimates $\hat{\pi}_i$, which can cause the inverse-probabilities to be highly variable and skewed (Kang & Schafer (2007), Robins et al. (2007)). Han & Wang (2013), Han (2014) and Han (2016) recently developed some methods to improve the robustness of AIPW estimators by allowing multiple working models for both the selection probability and the augmentation. Multiple robustness can then be gained, which means that the estimators are consistent if any of the working models is correctly specified. In addition, these estimators are not sensitive to near-zero selection probabilities.

2.4.2 Methods for Incomplete Longitudinal Data

Consider the following longitudinal linear model:

$$Y_{ij} = \boldsymbol{W}_{ij}^{\top}\boldsymbol{\beta} + \varepsilon_{ij} = \beta_0 + X_{ij}\beta_1 + \boldsymbol{Z}_{ij}^{\top}\beta_2 + \varepsilon_{ij} \tag{2.3}$$

for $i = 1, ..., n$ and $j = 1, ..., m_i$, where $\boldsymbol{W}_{ij} = (1, X_{ij}, \boldsymbol{Z}_{ij}^{\top})^{\top}$, $\boldsymbol{\beta}$ is the vector of the regression coefficients, Y_{ij} is a continuous response and $(X_{ij}, \boldsymbol{Z}_{ij}^{\top})^{\top}$ are covariates of subject i observed at time t_{ij} with corresponding random error ε_{ij}. Here we consider the sparse longitudinal case,

which means that m_i is bounded when $n \to \infty$. Different subjects are mutually independent, but generally there is within-subject correlation for observations measured at different time points. Let R_{ij} denote the indicator of the availability of X_{ij}. That is, let $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ if X_{ij} is missing. Let $\mathbf{R}_i = (R_{i1}, ..., R_{im_i})^\top$ be the indicator vector of subject i.

As a regression problem, it is always the basic goal to get an unbiased estimator of $\boldsymbol{\beta}$. An extensive literature discussed the situation of dropouts (Diggle (2002)) under the MAR mechanism (Little & Rubin (2014)). When the joint likelihood of the response and covariates is available through normal random effect model or generalized linear mixed model (GLMM) with the data MAR, regular maximum likelihood methods such as EM algorithm give consistent estimates (Horton & Laird (1999), Fuchs (1982), Schluchter & Jackson (1989) and Ibrahim (1990)). Specifically, by modeling the dropout process in addition, one can use selection models or pattern-mixture models based on two different factorizations of the joint density of the responses, covariates and missingness indicators (Little (1993), Little (1995)). As a simple example, if you have dropout response data $\boldsymbol{y}$ with covariates $\boldsymbol{x}$ and missingness indicators $\boldsymbol{r}$ are given, the selection model factors the complete data distribution as

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{r}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ are the parameters, while a pattern-mixture model factors the distribution as

$$f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\theta}).$$

Although from a theoretical point of view, the two models are just two different ways factoring the same full-data distribution, practically these two approaches lead to different kinds of simplifying assumptions and different analyses, as indicated in Diggle (2002). These two models can also extend to the MNAR mechanism (Ibrahim & Molenberghs (2009)). Bayesian methods (Daniels & Hogan (2008)) and multiple imputation (Schafer (1997), Rubin (2004)) can also be considered to handle the missing data problem. However, all the above methods usually require likelihood

assumptions and can be sensitive to model misspecification.

For complete longitudinal data, Liang & Zeger (1986) proposed to perform the analysis based on GEE as

$$n^{-1/2}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}_i^{\top}}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}\{\mathbf{Y}_i-\boldsymbol{\mu}_i(\boldsymbol{\beta})\}=n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_i^{\top}\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\beta})=0,$$

where $\boldsymbol{\mu}_i=(\mu_{i1},...,\mu_{im_i})^{\top}$ is the vector of conditional mean with $\mu_{ij}=E(Y_{ij}|\mathbf{X}_i,\mathbf{Z}_i)$. With a working covariance structure, this semiparametric method gives consistent results once the marginal mean of the outcomes at each time is correctly specified. When we have incomplete data, GEE generally produces asymptotically unbiased estimates only under the MCAR mechanism. As reviewed in Section 2.4, Robins et al. (1994) first introduced a class of inverse-probability weighted estimators and augmented inverse-probability weighted estimators for i.i.d. data based on GEE when data are MAR. The weights are obtained from the parametric models for the selection probabilities. These models need to be correctly specified to guarantee the consistency of the estimation for IPW. By choosing the augmentation to be the conditional expectation of the score function in the first part of the estimating equations, AIPW is doubly robust. That is to say, AIPW is consistent when either the selection probability model or the missing covariate model conditional on the observed data is correctly specified. Robins et al. (1995) extended the idea of IPW to longitudinal data with monotone missing response.

Recently, more research works have been done on DR estimators for incomplete longitudinal data. Lipsitz et al. (1999) first introduced the DR estimator for cross-sectional studies with a missing covariate and properties similar to maximum likelihood. Other literature includes Van der Laan & Robins (2003), Bang & Robins (2005) and Seaman & Copas (2009). But this literature mainly solves the problem for monotone missing incomplete response. Dealing with non-monotone missing values is generally more difficult because the variety of patterns makes the factorization of the likelihood very challenging. Chen et al. (2010) and Chen & Zhou (2011) discussed the DR estimator for both response and covariates non-monotone missing at random. These works are im-

pressive for a more complicated situation, but many assumptions are needed for the identifiability of the models. It is not easy to calculate the marginal selection probabilities even under the correct parametric models and the augmentation may involve nontrivial integration.

3. SEMIPARAMETRIC ESTIMATION IN REGRESSION WITH MISSING COVARIATES FOR I.I.D. DATA

In this chapter we investigate semiparametric estimation of regression coefficients through generalized estimating equations with single-index models when some covariates are missing at random. Existing popular semiparametric estimators by weighted estimating equations may run into difficulties when some selection probabilities are small or the dimension of the covariates is not low. We propose a new simple parameter estimator using a kernel assisted estimator for the augmentation by a single-index model without using the inverse of selection probabilities. We explore the asymptotic efficiency of the proposed estimator and its relationships with existing estimators. In particular, we show that under certain conditions the proposed estimator is as efficient as the existing methods based on standard kernel smoothing, which are often practically infeasible in the case of multiple covariates. A simulation study and a real data example are presented to illustrate the proposed method. The numerical results show that the proposed estimator avoids some numerical issues caused by estimated small selection probabilities that are needed in other estimators.

3.1 Introduction

Standard methods for regression generally require fully observed data. In practice, however, for the regression analysis of a response, Y, on a set of covariates (X, Z), the covariate X may be missing for some subjects. This is common in many areas such as biomedical sciences and clinical trials due to different reasons, including unavailability of covariate measurements, loss of data, and survey nonresponse. For example, in a subset of Canada 2010/2011 Youth Smoking Survey (YSS) data as described in Section 3.7, 144 students (total $n = 493$) had their BMI (Body Mass Index) missing. Here we consider the linear regression model $Y = W\beta + \varepsilon$, where $W = (1, X, Z^\top)^\top$ is the covariate vector, and $E(\varepsilon|W) = 0$. The objective of this regression analysis is to estimate the regression coefficients β when the scalar covariate X is assumed to be missing at random (MAR) in the sense of Rubin (1976).

Much work using the maximum likelihood has been developed in the area of missing covariate regression analysis. This model-based method is flexible and clear for inference, and the asymptotic properties can be obtained via the second derivatives of the log-likelihood. However, the observed likelihood is generally difficult to get in a closed form for most missing data problems, which needs factorization and reparameterization of the likelihood. These can be achieved with multivariate normal model (Hartley & Hocking (1971)) or specific monotone patterns of missing (Little & Rubin (2014)). When likelihood factorization is not available, the EM algorithm is a popular technique for obtaining the maximum likelihood estimation (MLE) with ignorable missing categorical or continuous covariates (Fuchs (1982), Schluchter & Jackson (1989) and Ibrahim (1990)).

Another likelihood-based approach is Bayesian methods, which are straightforward in terms of concepts and inferences. On the other hand, it can be challenging to correctly specify the conditional covariate distribution and the joint priors over parameters. Ibrahim et al. (2002) considered Bayesian methods for MAR covariates in GLM with informative prior based on historical data.

The most popular approach in industry and software packages might be multiple imputation (MI). The basic idea of MI is to impute missing data to create a new "complete" sample, and then analyze it as if it were a complete data set. MI methods for MAR covariates in linear regression models are discussed in Rubin (2004) and Little & Rubin (2014). In terms of specifying the conditional covariate distribution, MI works similarly to the EM algorithm and the motivation is Bayesian, but the idea of MI itself is quite general and can be applied to other methods (Ibrahim et al. (2005)). Hsu et al. (2014) proposed a nearest neighbor-based nonparametric multiple imputation approach for missing covariate data by using the distance calculated from a two-dimensional summary score of information about the missing covariate and the missingness indicator.

In most situations, all the above three methods depend on the specification of the likelihood. When the distributional assumptions are correct, they are optimal. However, they can give biased estimation when the assumptions are violated. To have more robust estimation with less likelihood assumptions, some semiparametric methods such as weighted estimating equations (WEE) are

proposed. Robins et al. (1994) proposed a class of semiparametric estimators based on inverse-probability WEE

$$\Delta(\beta,\psi)=n^{-1/2}\sum_{i=1}^{n}\left\{\frac{R_i}{\pi_i}D_i(W_i)(y_i-W_i^\top\beta)+\left(1-\frac{R_i}{\pi_i}\right)\psi_i(Q_i)\right\}=0,$$

where n is the total sample size (including incomplete cases), $Q_i=(y_i,Z_i^\top)^\top$, $D_i(W_i)$ is a function satisfying a local identification condition as nonsingular $E\{D_i(W_i)W_i^\top\}$, $\psi_i(Q_i)$ is an arbitrary function of Q_i, R_i is the binary indicator such that $R_i=1$ if the covariate X_i is observed and $R_i=0$ otherwise, $\pi_i=E(R_i|Q_i)$ is the probability of observing X_i. Robins et al. (1994) discussed the choices of D_i and ψ_i, and showed that there exist unique $D_i^{(eff)}$ and corresponding $\psi_i^{(eff)}(Q_i)=E\{D_i^{(eff)}(W_i)(y_i-W_i^\top\beta)|Q_i\}$ that can achieve the semiparametric efficiency bound of $\hat{\beta}$. However, $D_i^{(eff)}$ does not always have a closed form, so we choose a convenient one as $D_i=\partial\mu_i/\partial\beta=W_i$, where $\mu_i=E(y_i|X_i,Z_i)$. Then $\psi_i(Q_i)=E(T_i|Q_i)$ with $T_i=W_i(y_i-W_i^\top\beta)$ the regular score function. If we let $\psi_i\equiv 0$, the resulting estimator is the inverse-probability weighted estimator (IPW). We usually fit a logistic regression model of R_i on Q_i to estimate π_i's in the equations. When the logistic regression model is incorrect for π_i, the estimation can be biased. To overcome this difficulty, Wang et al. (1997) proposed a nonparametric kernel smoother for the selection probabilities and developed asymptotic theory for the estimator, including the optimal bandwidth rate. When $\psi_i\neq 0$, the estimator is the augmented inverse-probability weighted estimator (AIPW). This estimator is generally more efficient than IPW because it also incorporates the incomplete cases. The most important advantage of AIPW is that it is doubly robust (DR) in the sense that the estimator will be consistent when either the selection probability model (π_i) or the augmentation model ($\psi_i(Q_i)$) is correctly specified. There are many works discussing the DR estimator and extensions, such as Bang & Robins (2005), Kang & Schafer (2007) and Robins & Ritov (1997). AIPW can still fail when both models are misspecified, and it always needs distributional assumptions on $p(x_i|y_i,z_i)$ to estimate $\psi_i(Q_i)$.

Although Robins et al. (1994) restricted π_i to be bounded away from 0, in practice you may

still get some positive but near-zero values for the estimates $\hat{\pi}_i$, which can make the inverse-probabilities highly variable and skewed (Kang & Schafer (2007), Robins et al. (2007)). Extending Wang et al. (1997), Wang & Wang (2001) considered kernel estimation for both π_i and $\psi_i(Q_i)$, developed several kernel assisted estimators, and showed their asymptotic equivalence. However, when the dimension of the continuous part in Q_i increases, we need multivariate kernel functions and the estimation procedure suffers from the "curse of dimensionality". This motivates us to propose using a single-index model on $E(T_i|Q_i)$. Then we can apply a univariate kernel function on the single-index $Q_i^\top \gamma$. Han & Wang (2013), Han (2014) and Han (2016) recently developed some methods to improve the robustness of AIPW estimators by allowing multiple working models for both the selection probability and the augmentation. Thus multiple robustness can be gained, which means the estimators are consistent if any of the working models is correctly specified. In addition, these estimators are not sensitive to near-zero selection probabilities. Our method, from another perspective, is numerically stable with near-zero selection probabilities simply by not including them in the point estimation procedure. Based on Wang & Wang (2001), we develop asymptotic distribution theory for the resulting estimator and compare it with IPW and AIPW. We also conduct simulation studies to investigate the finite-sample performance of the proposed estimator in comparison with existing methods.

The rest of the chapter is organized as follows. In Section 3.2 we briefly review IPW and AIPW. We then describe our new estimator in Section 3.3. In Section 3.4 we present asymptotic theory of the above estimators of β and compare their asymptotic efficiency. In particular, we show that under certain conditions and assumptions, our proposed estimator is as efficient as the existing methods based on standard kernel smoothing, which are often practically infeasible in the case of multiple covariates. In Section 3.6 we provide the results of our simulation studies. In Section 3.7 we apply our methods to Canada 2010/2011 Youth Smoking Survey (YSS) data. We make some concluding remarks in Section 3.8.

3.2 Brief Review of Existing Methods

3.2.1 Inverse-probability Weighted Estimator

Robins et al. (1994) proposed a class of estimators based on weighted estimating equations. One of them is IPW through the estimating equation

$$\Delta_1(\beta, \pi) = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} W_i (y_i - W_i^\top \beta) = 0. \tag{3.1}$$

The selection probabilities π_i's are usually unknown in observational studies. One can assume a parametric model for π_i, for example, a logistic regression model under MAR

$$\pi_i(\alpha) = \mathrm{P}(R_i = 1 | y_i, Z_i, \alpha) = \{1 + \exp(-\alpha_0 - \alpha_1 y_i - \alpha_2^\top Z_i)\}^{-1} = \{1 + \exp(-\alpha^\top Q_i)\}^{-1},$$

where α is unknown. In our estimation problem, α is a nuisance parameter and can be estimated by maximum likelihood estimator $\hat{\alpha}$. We denote the solution of $\Delta_1(\beta, \pi(\hat{\alpha})) = 0$ as $\hat{\beta}_{PIP}$ in the rest of the chapter.

Another approach is to estimate π_i nonparametrically. Wang et al. (1997) considered non-parametric kernel smoothers for the selection probabilities. Let d be the number of continuous components of Q, K be an rth-order kernel function, h_1 be the bandwidth parameter, and define $K_{h_1}(\cdot) = K(\cdot/h_1)$. Then the kernel estimator of $\pi(q)$ is given by

$$\hat{\pi}(q) = \frac{\sum_{i=1}^{n} R_i K_{h_1}(q - Q_i)}{\sum_{i=1}^{n} K_{h_1}(q - Q_i)}. \tag{3.2}$$

The resulting estimator is consistent, but is difficult to implement when Q is multi-dimensional.

A complete-case (CC) analysis is to use the observed data only treating the partial data set as a completely observed data set. This approach generally not only leads to inconsistent estimates when the missing mechanism is not MCAR, but also loses efficiency due to discarding information from incomplete cases. We will illustrate these points through simulations in Section 3.6.

3.2.2 Augmented Inverse-probability Weighted Estimator

The IPW does not incorporate the incomplete cases, which generally leads to inefficient estimates. Robins et al. (1994) proposed the AIPW by solving the following equations:

$$\Delta_2(\beta, \pi, \psi) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi_i} T_i + \left(1 - \frac{R_i}{\pi_i}\right) \psi_i \right\} = 0, \tag{3.3}$$

where $\psi_i = E(T_i|Q_i)$. Further, it also obtains the DR property. We can still estimate the selection probability π_i using a parametric model or a kernel smoother such as (3.2). To estimate ψ_i, Wang & Wang (2001) proposed to use a kernel estimator similar to that for π_i given by

$$\hat{\psi}(q) = \frac{\sum_{i=1}^{n} R_i T_i K_{h_2}(q - Q_i)}{\sum_{i=1}^{n} R_i K_{h_2}(q - Q_i)} \tag{3.4}$$

for another kernel bandwidth h_2. It can share the same order of kernel function and same rate of bandwidth with $\hat{\pi}(q)$ in (3.2). It is possible to use a parametric model on ψ or specify the conditional distribution $p(x_i|Q_i)$, but then it will fall into the same area of techniques as EM algorithm, Bayesian or MI.

3.2.3 Mean-score Estimator

The mean-score estimator (MS) solves

$$\Delta_3(\beta, \psi) = n^{-1/2} \sum_{i=1}^{n} \{R_i T_i + (1 - R_i)\psi_i\} = 0. \tag{3.5}$$

Reilly & Pepe (1995) proposed this estimator with all components of Q discrete, where ψ_i is estimated by

$$\hat{\psi}_i = \frac{1}{n^{(o)}_{y_i, Z_i}} \sum_{j \in V^{(o)}_{y_i, Z_i}} T_j(X_j; \beta | y_i, Z_i)$$

with $V^{(o)}_{y_i,Z_i}$ denoting the subset of complete cases for $y = y_i$, $Z = Z_i$, $n^{(o)}_{y_i,Z_i}$ the size of $V^{(o)}_{y_i,Z_i}$, $T_j(X_j;\beta|y_i,Z_i)$ the score function for the samples in $V^{(o)}_{y_i,Z_i}$. It simply uses the averaged score of the complete cases with the same Q_i as the estimate of ψ_i. Wang & Wang (2001) extended it to the setting where some components are continuous by (3.4). Unlike IPW or AIPW, MS does not need to estimate the selection probabilities π_i.

3.3 Proposed Methodology

3.3.1 The Issues of Small Selection Probabilities and Curse of Dimensionality

Although theoretically the IPW and AIPW estimators are unbiased estimators when either the model for selection probabilities (π) or for augmentation (ψ) is correctly specified, they may encounter numerical problems if some π_i's are small so that the inverse-probability weights are highly variable. In this case, some subjects may have very large weights to significantly influence the weighted averages, and the sampling distribution of a locally semiparametric efficient estimator (IPW, AIPW) can be markedly skewed and highly variable, leading to biased estimation. We will illustrate this point through simulations in Section 3.6. This phenomenon is observed and discussed by Kang & Schafer (2007) and Robins et al. (2007), existing at least when parametrically modeling π_i. In this sense, the mean-score estimator has its advantage as it does not need to model and use inverse-probability weights.

However, all the kernel-estimation-based estimators mentioned above, including the MS estimator, have the same problem: curse of dimensionality. If the dimension d of the continuous part in Q is more than one, the performance of kernel functions can be unsatisfying.

3.3.2 Single-index Model and the Proposed Estimator

To overcome the problem discussed above, we consider a single-index model on ψ. Notice that

$$\begin{aligned}\psi_i = E(T_i|Q_i) = E\left\{W_i(y_i - W_i^\top\beta)\big|Q_i\right\} = E(W_i|Q_i)y_i - E(W_iW_i^\top|Q_i)\beta \\ = \begin{pmatrix} 1 \\ E(X_i|Q_i) \\ Z_i \end{pmatrix} y_i - \begin{pmatrix} 1 & E(X_i|Q_i) & Z_i^\top \\ E(X_i|Q_i) & E(X_i^2|Q_i) & E(X_i|Q_i)Z_i^\top \\ Z_i & Z_iE(X_i|Q_i) & Z_iZ_i^\top \end{pmatrix}\beta. \end{aligned} \tag{3.6}$$

Thus we only need to model $E(X_i|Q_i)$ and $E(X_i^2|Q_i)$. Assume a single-index model (SIM)

$$X_i = g(Q_i^\top\gamma) + e_i, \tag{3.7}$$

where g is an unknown smooth univariate function, γ is the parameter of the model with the same dimension of Q_i, e_i's are random errors with zero mean. To guarantee identifiability, we assume the first non-zero element of γ is positive 1. If the number of complete cases is n_1, one estimator of $g(\cdot)$ based only on the complete cases is

$$\hat{g}(u|\gamma) = \frac{\sum_{j=1}^{n_1} X_j^{(o)} K_h\left(u - {Q_j^{(o)}}^\top\gamma\right)}{\sum_{j=1}^{n_1} K_h\left(u - {Q_j^{(o)}}^\top\gamma\right)} = \frac{\sum_{k=1}^{n} R_kX_kK_h(u - Q_k^\top\gamma)}{\sum_{k=1}^{n} R_kK_h(u - Q_k^\top\gamma)},$$

where $(X_j^{(o)}, Q_j^{(o)})$ are pairs of the complete cases. Then under the SIM condition, we have

$$\hat{E}(X_i|Q_i) = \hat{E}(X_i|Q_i^\top\gamma) = \hat{g}(Q_i^\top\gamma) = \frac{\sum_{k=1}^{n} R_kX_kK_h((Q_i - Q_k)^\top\gamma)}{\sum_{k=1}^{n} R_kK_h((Q_i - Q_k)^\top\gamma)}. \tag{3.8}$$

We can also apply this model to get an estimate of $E(X_i^2|Q_i)$ as

$$\hat{E}(X_i^2|Q_i) = \frac{\sum_{k=1}^{n} R_k X_k^2 K_h((Q_i - Q_k)^\top \gamma)}{\sum_{k=1}^{n} R_k K_h((Q_i - Q_k)^\top \gamma)}. \tag{3.9}$$

We can construct

$$\hat{\pi}_i^*(\gamma) = \hat{E}(R_i|Q_i^\top \gamma) = \frac{\sum_{k=1}^{n} R_k K_h((Q_i - Q_k)^\top \gamma)}{\sum_{k=1}^{n} K_h((Q_i - Q_k)^\top \gamma)} \tag{3.10}$$

as the estimated selection probabilities modeled by the SIM using the same (γ, h). Notice that (3.8) and (3.10) have the same forms as (3.4) and (3.2) (NW-estimators). But the former two are conditional on the single-index and thus can just use univariate kernel functions with the additional parameter γ. Due to this similarity, we can extend the asymptotic results by Wang & Wang (2001) to the single-index models. The details will be shown in Section 3.4. On the other hand, compared to Wang & Wang (2001), here we only estimate the first two moments of X_i given Q_i by using the local average when estimating ψ_i but keep the original y_i, Z_i since they are always observed, instead of using the local average of the whole score function like (3.4).

Let

$$\hat{\psi}_i(\gamma) = \frac{\sum_{k=1}^{n} R_k T_{i,k} K_h((Q_i - Q_k)^\top \gamma)}{\sum_{k=1}^{n} R_k K_h((Q_i - Q_k)^\top \gamma)},$$

where $T_{i,k} = W_{i,k}(y_i - W_{i,k}^\top \beta)$ with $W_{i,k} = (1, X_k, Z_i^\top)^\top$. This $\hat{\psi}_i(\gamma)$ is a kernel estimate of ψ_i by estimating only $E(X_i|Q_i)$ and $E(X_i^2|Q_i)$ with a kernel smoother via (3.6). Then the AIPW (3.3) and MS (3.5) estimators in the previous section can be extended using the SIM as follows:

(a) AIPW with a parametric model on selection probabilities:

$$\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma)) = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{R_i}{\pi_i(\hat{\alpha})} T_i + \left(1 - \frac{R_i}{\pi_i(\hat{\alpha})}\right) \hat{\psi}_i(\gamma) \right\} = 0; \tag{3.11}$$

(b) MS without inverse-probability weights:

$$\Delta_3(\beta, \hat{\psi}(\gamma)) = n^{-1/2} \sum_{i=1}^{n} \Big\{ R_i T_i + (1 - R_i)\hat{\psi}_i(\gamma) \Big\} = 0. \tag{3.12}$$

We use $\hat{\beta}_{PIPA}$ and $\hat{\beta}_A$ to denote the solutions of equations (3.11) and (3.12) respectively.

Generally, besides the main parameter β, γ is an unknown nuisance parameter which also needs to be estimated. However, in our case with the linear relationship between Y and $(X, Z^\top)^\top$, we have a special form of γ as $\gamma = (1, -\beta_Z^\top)^\top$, where β_Z is the regression coefficient of Z as $E(Y|X, Z) = \beta_0 + \beta_1 X + \beta_Z^\top Z$. In that sense, the single index is $u_i = Q_i^\top \gamma = y_i - \beta_Z^\top Z_i$ and γ is a part of β so that we do not need to estimate γ separately. Note also that the choice of bandwidth h is crucial. Technical details about bandwidth selection will be discussed in Section 3.4.

Under certain conditions, we can show that these two estimators are asymptotically equivalent (see Corollary 3.1 below), and they are both as efficient as the existing estimators using standard kernel smoothing, which are often practically infeasible in the case of multi-covariates. In practice, we prefer $\hat{\beta}_A$ because the estimation procedure of $\hat{\beta}_A$ has the clear advantage of not involving inverse of selection probabilities to avoid modeling π_i's thus it is simpler. Moreover, it is important to note that unlike all inverse probability weighted estimators, $\hat{\beta}_A$ is not sensitive to the positive near-zero π_i's since we do not use them in the point estimation procedure.

Of course, $\hat{\beta}_A$ no longer has the property of double robustness, and thus needs a consistent estimator of ψ. In this setting, the performance of $\hat{\beta}_A$ depends on whether the single-index model (3.7) is reasonable. Since the relationship between the response and covariates is assumed to be linear, it is not unreasonable to assume this model. Actually it is valid when (y_i, X_i, Z_i) jointly follows a multivariate normal distribution. More generally, it can still give reasonably robust results under other distributions, as is to be shown in our numerical studies in Section 3.6.

3.4 Asymptotic Properties

In this section, we will show the asymptotic behavior of the proposed estimator $\hat{\beta}_A$, and its asymptotic equivalence to some other estimators described above under certain conditions. For

simplicity, we define $\pi_i^*(\gamma) = E(R_i|Q_i^\top\gamma)$ as the selection probabilities conditional on the single-index $Q_i^\top\gamma$ with parameter γ, $\pi_i(\alpha)$ as the selection probabilities based on a parametric model with parameter α. We need the following regularity conditions to establish the asymptotic theory:

(i) The smoothing parameter h satisfies $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$.

(ii) All the selection probabilities π_i's are bounded away from zero.

(iii) The selection probability function on the single-index $\pi^*(\gamma)$ has r continuous and bounded partial derivatives a.e.

(iv) The density function $f(u)$ of U and the conditional density function $f_{U|R}(u)$ of $U|R$ have r continuous and bounded partial derivatives a.e.

(v) The conditional distributions $f_{U|R=0}(u)$ and $f_{U|R=1}(u)$ have the same support, and $b(u) = f_{U|R=0}(u)/f_{U|R=1}(u)$ is bounded over the support.

(vi) The conditional expectations $\psi(u|\gamma) = E(T|Q^\top\gamma = u)$ and $E(TT^\top|Q^\top\gamma)$ exist and have r continuous and bounded partial derivatives a.e.

(vii) For score T, $E(TT^\top)$ and $E\{(\partial/\partial\beta)T\}$ exist and are positive definite, and $(\partial^2/\partial\beta\partial\beta^\top)T$ exists and is continuous with respect to β a.e.

Recall that r is the order of the kernel function used in the estimation. From regularity condition (i), r is related to the rate of the bandwidth h. Since we are considering a SIM for estimation, a standard 2nd-order ($r = 2$) univariate kernel function seems reasonable in practice.

Let $\eta_n = \{nh^{2r} + (nh^2)^{-1}\}^{1/2}$. The following lemmas are important to prove our main theorems.

Lemma 3.1. *Under regularity conditions* (i)-(vii) *and assuming that the single-index model (3.7) is true, we have*

$$n^{-1/2}\sum_{i=1}^{n}(1-R_i)\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^{n} R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top\gamma) + O_p(\eta_n),$$

where $a(Q_i^\top\gamma) = \{1-\pi_i^*(\gamma)\}/\pi_i^*(\gamma)$, $T_i^0 = E_{Z_i|u_i,R_i=0}(T_i) = \int T_i f(Z_i|u_i, R_i=0)dZ_i$, $\psi_i^0(\gamma) = E_{Z_i|u_i,R_i=0}\{\psi_i(\gamma)\} = \int \psi_i(\gamma) f(Z_i|u_i, R_i=0)dZ_i$ *with* $u_i = Q_i^\top\gamma$ *as the single index.*

This is an extension of Lemma 1 in Wang & Wang (2001). The proof of this lemma is given in Section 3.5.1.

Note that with the SIM and the single index $u_i = y_i - \beta_Z^\top Z_i$, we can write T_i and $\psi_i(\gamma)$ as

$$T_i = \begin{pmatrix} u_i - \beta_0 - \beta_1 X_i \\ u_i X_i - \beta_0 X_i - \beta_1 X_i^2 \\ Z_i(u_i - \beta_0 - \beta_1 X_i) \end{pmatrix}, \ \psi_i(\gamma) = \begin{pmatrix} u_i - \beta_0 - \beta_1 E(X_i|u_i) \\ u_i E(X_i|u_i) - \beta_0 E(X_i|u_i) - \beta_1 E(X_i^2|u_i) \\ Z_i\{u_i - \beta_0 - \beta_1 E(X_i|u_i)\} \end{pmatrix}.$$

Since MAR implies $(X_i \perp R_i)|Q_i$, we also have

$$T_i^0 = \begin{pmatrix} u_i - \beta_0 - \beta_1 X_i \\ u_i X_i - \beta_0 X_i - \beta_1 X_i^2 \\ Z_i^{u|0}(u_i - \beta_0 - \beta_1 X_i) \end{pmatrix}, \ \psi_i^0(\gamma) = \begin{pmatrix} u_i - \beta_0 - \beta_1 E(X_i|u_i) \\ u_i E(X_i|u_i) - \beta_0 E(X_i|u_i) - \beta_1 E(X_i^2|u_i) \\ Z_i^{u|0}\{u_i - \beta_0 - \beta_1 E(X_i|u_i)\} \end{pmatrix}$$

with $Z_i^{u|0} = E(Z_i|u_i, R_i=0)$.

Lemma 3.1 is useful because it converts asymptotically a sum of dependent random variables to a sum of independent and identically distributed (i.i.d.) random variables. Then it is easier to be dealt with by applying standard asymptotic theory.

Lemma 3.2. *Under the same conditions as those in Lemma 3.1, we have*

a)

$$n^{-1/2}\sum_{i=1}^n R_i\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^n R_i\{T_i^1 - \psi_i^1(\gamma)\} + O_p(\eta_n);$$

b)

$$n^{-1/2}\sum_{i=1}^n \frac{R_i}{\hat{\pi}_i^*(\gamma)}\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^n \frac{R_i}{\pi_i^*(\gamma)}\{T_i^1 - \psi_i^1(\gamma)\} + O_p(\eta_n);$$

c) In addition, if the parametric model for selection probabilities is correctly specified and has a single-index model form with the same single index $u_i = Q_i^\top\gamma$ *as the augmentation, which*

means $\pi_i(\alpha) = \pi_i = \pi_i^*(\gamma)$*, then*

$$n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\alpha})}\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\{T_i^1 - \psi_i^1(\gamma)\} + O_p(\eta_n),$$

where $T_i^1 = E_{Z_i|u_i,R_i=1}(T_i)$, $\psi_i^1(\gamma) = E_{Z_i|u_i,R_i=1}\{\psi_i(\gamma)\}$.

The proof of Lemma 3.2 is given in Section 3.5.2. The idea of the proof is analogous to that of Lemma 3.1.

Define

$$U_i = R_iT_i + (1 - R_i)\psi_i(\gamma) + R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top\gamma).$$

Based on Lemmas 3.1 and 3.2, we have the following main theorems.

Theorem 3.1. *Under the regularity conditions* (i)-(vii) *and assuming that the single-index model (3.7) is true,* $\hat{\beta}_A$ *is asymptotically equivalent to the solution of the following estimating equation:*

$$n^{-1/2}\sum_{i=1}^{n}U_i = 0.$$

Furthermore, we have

$$n^{1/2}(\hat{\beta}_A - \beta) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Sigma}_A),$$

where $\boldsymbol{\Sigma}_A = \boldsymbol{D}^{-1}\boldsymbol{\mathcal{M}}\boldsymbol{D}^{-1}$ *with* $\boldsymbol{D} = -n^{-1}E(\partial T_1/\partial\beta^\top) = E(W_1W_1^\top)$ *and* $\boldsymbol{\mathcal{M}} = \mathrm{cov}(U_1) = \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} + 2\boldsymbol{\mathcal{C}}$ *for*

$$\begin{aligned}
\boldsymbol{\mathcal{A}} &= E\left(\pi_1T_1T_1^\top\right) + E\left\{(1-\pi_1)\psi_1\psi_1^\top\right\},\\
\boldsymbol{\mathcal{B}} &= E\left\{\pi_1(T_1^0 - \psi_1^0)(T_1^0 - \psi_1^0)^\top a^2(Q_1^\top\gamma)\right\},\\
\boldsymbol{\mathcal{C}} &= E\left\{\pi_1T_1(T_1^0 - \psi_1^0)^\top a(Q_1^\top\gamma)\right\}.
\end{aligned}$$

The main step of the proof is to obtain the asymptotic equivalence between our estimating equation $\Delta_3(\beta, \hat{\psi}(\gamma)) = 0$, and $n^{-1/2}\sum_{i=1}^{n}U_i = 0$ with true $\pi^*(\gamma)$ and $\psi(\gamma)$. The details can be

found in Section 3.5.3.

The asymptotic covariance matrix Σ_A of $\hat{\beta}_A$ can be estimated by first estimating $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\mathcal{C}}$ separately. However, this approach cannot guarantee the necessary property of non-negative definiteness of the resulting covariance estimate and it might lead to numerically unstable results. For this reason, we propose to estimate Σ_A directly as follows:

$$\hat{\boldsymbol{\Sigma}}_A = \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \left(\frac{1}{n} \sum_{i=1}^{n} \hat{U}_i \hat{U}_i^\top \right) \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}), \tag{3.13}$$

where

$$\hat{U}_i = R_i T_i(\hat{\beta}_A) + (1 - R_i)\hat{\psi}_i + R_i\{\hat{T}_i^0(\hat{\beta}_A) - \hat{\psi}_i^0\} a(Q_i^\top \hat{\gamma})$$

with $\hat{\psi}_i = \hat{\psi}_i(\hat{\beta}_A, \hat{\gamma})$, $\hat{\psi}_i^0 = \hat{\psi}_i^0(\hat{\beta}_A, \hat{\gamma})$ being the estimates of ψ_i, ψ_i^0 based on $\hat{\beta}_A$ and

$$\hat{\boldsymbol{D}}_n(\hat{\gamma}) = n^{-1} \sum_{i=1}^{n} \left\{ R_i W_i W_i^\top + (1 - R_i)\hat{E}(W_i W_i^\top | Q_i^\top \hat{\gamma}) \right\}.$$

Here $\hat{T}_i^0$ is T_i^0 with $Z_i^{u|0}$ estimated by

$$\hat{Z}_i^{u|0} = \hat{E}(Z_i | \hat{u}_i, R_i = 0) = \frac{\sum_{k=1}^{n} (1 - R_k) Z_k K_h((Q_i - Q_k)^\top \hat{\gamma})}{\sum_{k=1}^{n} (1 - R_k) K_h((Q_i - Q_k)^\top \hat{\gamma})}$$

with $\hat{u}_i = Q_i^\top \hat{\gamma}$.

Note that to get $\hat{E}(W_i W_i^\top | Q_i^\top \hat{\gamma})$, we only need to calculate $\hat{E}(X_i | Q_i)$ and $\hat{E}(X_i^2 | Q_i)$ through (3.8) and (3.9) because of the structure in (3.6).

Theorem 3.2. *Under the same conditions as in Theorem 3.1 and the additional conditions for Lemma 3.2(c), we have*

$$n^{1/2}(\hat{\beta}_{PIPA} - \beta) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Sigma}_{PA}),$$

where $\boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S}^* + \boldsymbol{V})\boldsymbol{D}^{-1}$, *with* $\boldsymbol{D} = E(W_1 W_1^\top)$, $\boldsymbol{S} = E\{T_1 T_1^\top / \pi_1^*(\gamma)\}$, $\boldsymbol{S}^* = E\{\psi_1 \psi_1^\top / \pi_1^*(\gamma)\}$ *and* $\boldsymbol{V} = E(\psi_1 \psi_1^\top)$.

It is readily seen that a consistent covariance matrix estimate of $\hat{\beta}_{PIPA}$ is given by

$$\hat{\boldsymbol{\Sigma}}_{PA} = \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{R_i}{\hat{\pi}_i^{*2}(\hat{\gamma})} \left(\hat{T}_i - \hat{\psi}_i \right) \left(\hat{T}_i - \hat{\psi}_i \right)^{\top} + \frac{R_i}{\hat{\pi}_i^{*}(\hat{\gamma})} \hat{\psi}_i \hat{\psi}_i^{\top} \right\} \right] \hat{\boldsymbol{D}}_n^{-1}(\hat{\gamma}) \qquad (3.14)$$

with $\hat{T}_i = \hat{T}_i(\hat{\beta}_{PIPA}, \hat{\gamma})$, $\hat{\psi}_i = \hat{\psi}_i(\hat{\beta}_{PIPA}, \hat{\gamma})$ being estimates of T_i, ψ_i based on $\hat{\beta}_{PIPA}$. Since $Pr(R_i = 1|u_i, Z_i) = Pr(R_i = 1|Q_i) = \pi_i = \pi_i^*(\gamma) = Pr(R_i = 1|u_i)$, the additional condition for Lemma 3.2(c) implies that $(Z_i \perp R_i)|u_i$. Then $T_i^0 = T_i^1 = E_{Z_i|u_i}(T_i)$, $\psi_i^0(\gamma) = \psi_i^1(\gamma) = E_{Z_i|u_i}\{\psi_i(\gamma)\}$.

Although the relationship between $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_{PA}$ is generally not clear even under the conditions of Theorem 3.2, numerically the SE of $\hat{\beta}_A$ is competitive (see Section 3.6) and $\hat{\beta}_A$ does not have the potential danger of having exceedingly high inverse-probability weights. The theorems also demonstrate the asymptotic normality of the above estimators, which helps us to make inferences with the estimators.

Corollary 3.1. *Under the same conditions as in Theorem 3.2 and further assuming $E(Z_i|u_i) = Z_i$, we have*

(a) $\hat{\beta}_A$ and $\hat{\beta}_{PIPA}$ are asymptotically equivalent and are both more efficient than $\hat{\beta}_{PIP}$;

(b) The estimators $\hat{\beta}_A$ and $\hat{\beta}_{PIPA}$ based on a single-index model are as efficient as those based on a standard multivariate kernel smoother.

It is intuitive to see that these estimators are asymptotically more efficient than $\hat{\beta}_{PIP}$ because they incorporate the incomplete cases. However, when the conditions are satisfied, it is surprising to see that the estimators based on the SIM can keep the efficiency of the standard kernel smoothers (such as (3.2), (3.4) proposed by Wang & Wang (2001)) with a lower dimension of information. Since IPW is asymptotically equivalent to AIPW and MS estimators with both selection probabilities and augmentation estimated by a standard kernel smoother (Wang & Wang (2001)), the proof of the corollary also shows that IPW using a standard kernel smoother is more efficient than $\hat{\beta}_{PIP}$, which was not discussed by Wang et al. (1997).

We define $\boldsymbol{\Sigma}_P$ as the asymptotic covariance matrix of $\hat{\beta}_{PIP}$ and $\tilde{\boldsymbol{\Sigma}}$ for $\hat{\beta}_A$, $\hat{\beta}_{PIPA}$ based on a standard kernel smoother. For two positive semi-definite covariance matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we define $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite. From the proof in Section 3.5.5, we see that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}$ under the conditions in Corollary 3.1.

The performance of the estimator $\hat{\beta}_A$ depends on the choice of the bandwidth h used in the kernel function $K_h(\cdot)$. In the regularity conditions, we require $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$. Therefore, the classical optimal rate of the bandwidth $O(n^{-1/5})$ does not work in our situation, as indicated in Sepanski et al. (1994). A reasonable choice is $h = Cn^{-1/3}$, where C is a constant. A plug-in method can be applied to estimate C. For simplicity, we can use $C = \hat{\sigma}_u$ as suggested by Wang et al. (1997) and Zhou et al. (2008), where $\hat{\sigma}_u$ is the sample standard deviation of the single index u_i. We use this formula to choose the bandwidth in our following numerical studies.

3.5 Proofs of the Main Theorems

3.5.1 Proof of Lemma 3.1

Proof. The idea in the proof is similar to that in the proof of Lemma 1 in Wang & Wang (2001). Recall that $u_i = Q_i^\top \gamma = y_i - \beta_Z^\top Z_i$ is the single index and that n_1 is the number of complete cases. Let

$$\hat{f}_{U|R=1}(u) = \frac{1}{n_1 h} \sum_{k=1}^{n} R_k K_h(u - u_k), \quad E_n(u) = \hat{f}_{U|R=1}(u) - f_{U|R=1}(u),$$

$$V_{ni} = \hat{f}_{U|R=1}(u_i), \quad W_{ni} = \frac{1}{n_1 h} \sum_{k=1}^{n} R_k T_{i,k} K_h(u_i - u_k).$$

Under the regularity conditions, we have $E\{E_n(u)\} = O(h^r)$ and $\mathrm{var}\{E_n(u)\} = O\{(nh)^{-1}\}$ by the Taylor expansions. Then by the Chebyshev inequality, $E_n(u) - E\{E_n(u)\} = O_p\{(nh)^{-1/2}\}$, which implies $E_n(u) = O_p\{h^r + (nh)^{-1/2}\}$, and thus $E_n(u_i) = O_p\{h^r + (nh)^{-1/2}\}$. Similarly, we have $W_{ni} - \psi_i V_{ni} = O_p\{h^r + (nh)^{-1/2}\}$.

Define $\delta_n = h^{2r} + (nh)^{-1}$. Under the SIM condition,

$$\hat{\psi}_i - \psi_i = \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} - \frac{(W_{ni} - \psi_i V_{ni}) E_n(u_i)}{V_{ni} f_{U|R=1}(u_i)} = \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} + O_p(\delta_n). \tag{3.15}$$

Let $Q_i^* = R_i Q_i$, $X_i^* = R_i X_i$ for $i = 1, ..., n$ as the values of the complete cases. Then

$$\begin{aligned}
&E\left\{ \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} \middle| R_i = 0, \text{all } (R, Q^*, X^*) \right\} \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Q|R=0}(Q_i) dQ_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \iint \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U,Z|R=0}(u_i, Z_i) dZ_i du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \left\{ \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Z|U,R=0}(Z_i) dZ_i \right\} f_{U|R=0}(u_i) du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \frac{(T_{i,k}^0 - \psi_i^0) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U|R=0}(u_i) du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k (T_k^0 - \psi_k^0) b(u_k) + O_p(h^r),
\end{aligned}$$

where $T_{i,k}^0 = E_{Z_i|u_i, R_i=0}(T_{i,k}) = \int T_{i,k} f(Z_i|u_i, R_i = 0) dZ_i$, $b(u)$ is defined in regularity condition (v). The last step is because of the concentration of u_i on u_k. Using the same idea and $\{\cdots\}$ to denote a repeat of the preceding term, we also have

$$\begin{aligned}
&\text{var}\left\{ \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} \middle| R_i = 0, \text{all } (R, Q^*, X^*) \right\} \\
&= \frac{1}{n_1^2} \sum_{k=1}^{n} R_k \left[\int \left\{ \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} \right\} \{\cdots\}^\top f_{Q|R=0}(Q_i) dQ_i \right. \\
&\left. - \left\{ \sum_{k=1}^{n} R_k \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Q|R=0}(Q_i) dQ_i \right\} \{\cdots\}^\top \right] + O_p\left(\frac{1}{nh}\right) \\
&= O_p\left(\frac{1}{nh}\right).
\end{aligned}$$

Let

$$S_n = n^{-1/2} \sum_{i=1}^{n} (1 - R_i) \left\{ \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} - \frac{1}{n_1} \sum_{k=1}^{n} R_k (T_k^0 - \psi_k^0) b(u_k) \right\}.$$

Then the summations with $R_i = 0$ in S_n are i.i.d. random variables conditioning on all (R, Q^*, X^*). Thus we have

$$\text{var}\{S_n | \text{all } (R, Q^*, X^*)\} = \frac{n - n_1}{n} \text{var} \left\{ \frac{W_{n1} - \psi_1 V_{n1}}{f_{U|R=1}(u_1)} \middle| \text{ all } (R, Q^*, X^*) \right\} = O_p \left(h^{2r} + \frac{1}{nh} \right).$$

Then $E(S_n) = O(h^r)$ and $\text{var}(S_n) = O(h^{2r} + (nh)^{-1})$ imply $S_n = O_p(\eta_n)$. Back to (3.15), we have

$$\begin{aligned} n^{-1/2} \sum_{i=1}^{n} (1 - R_i)(\hat{\psi}_i - \psi_i) &= n^{-1/2} \sum_{i=1}^{n} \left\{ (1 - R_i) \frac{1}{n_1} \sum_{k=1}^{n} R_k (T_k^0 - \psi_k^0) b(u_k) \right\} + O_p(\eta_n) \\ &= n^{-1/2} \sum_{k=1}^{n} R_k (T_k^0 - \psi_k^0) a(u_k) + O_p(\eta_n). \end{aligned}$$

□

3.5.2 Proof of Lemma 3.2

Proof. (a). The proof is analogous to that of Lemma 3.1. The main difference is that this is the summation of the complete cases. Thus we need to condition on $R_i = 1$. Then

$$
\begin{aligned}
& E\left\{ \frac{W_{ni} - \psi_i V_{ni}}{f_{U|R=1}(u_i)} \middle| R_i = 1, \text{all } (R, Q^*, X^*) \right\} \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Q|R=1}(Q_i) dQ_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \iint \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U,Z|R=1}(u_i, Z_i) dZ_i du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \left\{ \int \frac{(T_{i,k} - \psi_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{Z|U,R=1}(Z_i) dZ_i \right\} f_{U|R=1}(u_i) du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k \int \frac{(T^1_{i,k} - \psi^1_i) K_h(u_i - u_k)}{h f_{U|R=1}(u_i)} f_{U|R=1}(u_i) du_i \\
&= \frac{1}{n_1} \sum_{k=1}^{n} R_k (T^1_k - \psi^1_k) + O_p(h^r),
\end{aligned}
$$

where $T^1_{i,k} = E_{Z_i|u_i,R_i=1}(T_{i,k}) = \int T_{i,k} f(Z_i|u_i, R_i = 1) dZ_i$. The rest of the proof follows in the same manner as in the proof of Lemma 3.1.

(b). Similarly to the proof of (a), we have

$$
n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi^*_i(\gamma)} \{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} = n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi^*_i(\gamma)} \{T^1_i - \psi^1_i(\gamma)\} + O_p(\eta_n).
$$

According to the Hölder inequality for the sum of the product terms in the second term below, we have

$$
\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\hat{\pi}^*_i(\gamma)} \{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi^*_i(\gamma)} \{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} + n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\hat{\pi}^*_i(\gamma) \pi^*_i(\gamma)} \{\pi^*_i(\gamma) - \hat{\pi}^*_i(\gamma)\} \{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi^*_i(\gamma)} \{T^1_i - \psi^1_i(\gamma)\} + O_p(\eta_n).
\end{aligned}
$$

(c). The proof can be obtained analogously as in (b). □

3.5.3 Proof of Theorem 3.1

Proof. Based on the conclusion of Lemma 3.1,

$$\begin{aligned}\Delta_3(\beta, \hat{\psi}(\gamma)) &= n^{-1/2} \sum_{i=1}^{n} R_i T_i + (1 - R_i)\psi_i(\gamma) + (1 - R_i)\{\hat{\psi}_i(\gamma) - \psi_i(\gamma)\} \\ &= n^{-1/2} \sum_{i=1}^{n} R_i T_i + (1 - R_i)\psi_i(\gamma) + R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top \gamma) + O_p(\eta_n) \\ &= n^{-1/2} \sum_{i=1}^{n} U_i + O_p(\eta_n).\end{aligned}$$

Since $\Delta_3(\beta, \hat{\psi}(\gamma))$ is asymptotically equivalent to a sum of i.i.d. random variables, $\hat{\beta}_A$ is asymptotically normally distributed and has the asymptotic covariance $\boldsymbol{\Sigma}_A = \boldsymbol{D}^{-1}\boldsymbol{\mathcal{M}}\boldsymbol{D}^{-1}$ with

$$\begin{aligned}\boldsymbol{\mathcal{M}} &= \text{cov}\left(n^{-1/2} \sum_{i=1}^{n} U_i\right) = \text{cov}(U_1) \\ &= \text{cov}\left\{R_1 T_1 + (1 - R_1)\psi_i\right\} + \text{cov}\left[R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top \gamma)\right] \\ &\quad + 2\text{cov}\left(R_1 T_1 + (1 - R_1)\psi_i, R_i\{T_i^0 - \psi_i^0(\gamma)\}a(Q_i^\top \gamma)\right) \\ &= \boldsymbol{\mathcal{A}} + \boldsymbol{\mathcal{B}} + 2\boldsymbol{\mathcal{C}}.\end{aligned}$$

□

3.5.4 Proof of Theorem 3.2

Proof. We first consider the first part, $\Delta_1(\beta, \pi(\hat{\alpha}))$, of its estimating equation (3.11). By assumption, a correctly specified parametric model for the selection probabilities with parameter α is given by

$$\pi_i = \pi_i(\alpha) = E(R_i|Q_i) = \pi(\alpha|Q_i).$$

The log-likelihood is

$$l(\alpha) = \sum_{i=1}^{n} R_i \log\{\pi_i(\alpha)\} + (1 - R_i)\log\{1 - \pi_i(\alpha)\}.$$

The corresponding estimating equation for MLE $\hat{\alpha}$ is given by

$$n^{-1/2} \sum_{i=1}^{n} \frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1 - \pi_i(\alpha)\}} \{R_i - \pi_i(\alpha)\} = 0.$$

Then we have

$$\begin{aligned} n^{1/2}(\hat{\alpha} - \alpha) = & \left[E\left\{ \frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1 - \pi_1(\alpha)\}} \right\} \right]^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1 - \pi_i(\alpha)\}} \{R_i - \pi_i(\alpha)\} \right\} \\ & + O_p(n^{-1/2}). \end{aligned}$$

Moreover,

$$\begin{aligned} \Delta_1(\beta, \pi(\hat{\alpha})) &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\alpha})} T_i \\ &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\alpha)} T_i - n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i^2(\alpha)} T_i \pi_i'(\alpha)^\top (\hat{\alpha} - \alpha) + O_p(n^{-1/2}) \\ &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\alpha)} T_i - E\left\{ \frac{1}{\pi_1(\alpha)} \psi_1 \pi_1'(\alpha)^\top \right\} n^{1/2}(\hat{\alpha} - \alpha) + O_p(n^{-1/2}) \\ &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} T_i - E\left\{ \frac{1}{\pi_1(\alpha)} \psi_1 \pi_1'(\alpha)^\top \right\} \left[E\left\{ \frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1 - \pi_1(\alpha)\}} \right\} \right]^{-1} \\ &\quad \left\{ n^{-1/2} \sum_{i=1}^{n} \frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1 - \pi_i(\alpha)\}} \{R_i - \pi_i(\alpha)\} \right\} + O_p(n^{-1/2}) \\ &= \Delta_1(\beta, \pi) - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(n^{-1/2}), \end{aligned}$$

where $\boldsymbol{F}(\alpha) = E\left\{ \frac{1}{\pi_1(\alpha)} \psi_1 \pi_1'(\alpha)^\top \right\}$, $P_n(\alpha) = n^{-1/2} \sum_{i=1}^{n} \frac{\pi_i'(\alpha)}{\pi_i(\alpha)\{1 - \pi_i(\alpha)\}} \{R_i - \pi_i(\alpha)\}$, $\boldsymbol{C}(\alpha) = E\left\{ \frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{\pi_1(\alpha)\{1 - \pi_1(\alpha)\}} \right\}$.

We now consider the second part of the estimating equation. By Lemmas 3.1 and 3.2(a), we

obtain that

$$n^{-1/2}\sum_{i=1}^{n}\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^{n} R_i\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\} + n^{-1/2}\sum_{i=1}^{n}(1-R_i)\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\}$$
$$= n^{-1/2}\sum_{i=1}^{n} R_i\{T_i^1-\psi_i^1(\gamma)\} + n^{-1/2}\sum_{i=1}^{n} R_i\{T_i^0-\psi_i^0(\gamma)\}a(Q_i^\top\gamma)$$
$$+ O_p(\eta_n).$$

Recall that the additional condition for Lemma 3.2(c) requires $\pi_i = \pi_i^*(\gamma)$. This implies that $T_i^0 = T_i^1 = E_{Z_i|u_i}(T_i)$, $\psi_i^0(\gamma) = \psi_i^1(\gamma) = E_{Z_i|u_i}\{\psi_i(\gamma)\}$. Let $T_i^* = E_{Z_i|u_i}(T_i)$, $\psi_i^*(\gamma) = E_{Z_i|u_i}\{\psi_i(\gamma)\}$. Then

$$n^{-1/2}\sum_{i=1}^{n}\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\} = n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\{T_i^*-\psi_i^*(\gamma)\} + O_p(\eta_n) \tag{3.16}$$

Equation (3.16) and Lemma 3.2(c) imply that

$$n^{-1/2}\sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi_i(\hat{\alpha})}\right\}\{\hat{\psi}_i(\gamma)-\psi_i(\gamma)\} = O_p(\eta_n).$$

Then

$$n^{-1/2}\sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi_i(\hat{\alpha})}\right\}\hat{\psi}_i(\gamma) = n^{-1/2}\sum_{i=1}^{n}\left\{1-\frac{R_i}{\pi_i(\hat{\alpha})}\right\}\psi_i(\gamma) + O_p(\eta_n).$$

As in the proof for the first part $\Delta_1(\beta, \pi(\hat{\alpha}))$, we can show that

$$n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\alpha})}\psi_i(\gamma) = n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i}\psi_i(\gamma) - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)P_n(\alpha) + O_p(n^{-1/2}).$$

Finally we have

$$\begin{aligned}
\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma)) &= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\alpha})} T_i + \left\{ 1 - \frac{R_i}{\pi_i(\hat{\alpha})} \right\} \hat{\psi}_i(\gamma) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} T_i - \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + n^{-1/2} \sum_{i=1}^{n} \psi_i(\gamma) \\
&\quad - n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} \psi_i(\gamma) + \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(\eta_n) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i} T_i + n^{-1/2} \sum_{i=1}^{n} \left(1 - \frac{R_i}{\pi_i} \right) \psi_i(\gamma) + O_p(\eta_n) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i^*(\gamma)} T_i + n^{-1/2} \sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi_i^*(\gamma)} \right\} \psi_i(\gamma) + O_p(\eta_n) \\
&= \Delta_2(\beta, \pi^*(\gamma), \psi) + O_p(\eta_n).
\end{aligned}$$

In summary, we have shown that under certain conditions $\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma))$ is asymptotically equivalent to $\Delta_2(\beta, \pi^*(\gamma), \psi)$, which is a sum of i.i.d. terms. Hence, $\hat{\beta}_{PIPA}$ is asymptotically equivalent to the solution of $\Delta_2(\beta, \pi^*(\gamma), \psi) = 0$, having asymptotic normality with asymptotic covariance

$$\boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S}^* + \boldsymbol{V})\boldsymbol{D}^{-1}.$$

□

3.5.5 Proof of Corollary 3.1

Proof. By the fact that

$$\Delta_1(\beta, \pi(\hat{\alpha})) = \Delta_1(\beta, \pi) - \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) P_n(\alpha) + O_p(n^{-1/2}),$$

where $\boldsymbol{F}(\alpha)$ and $\boldsymbol{C}(\alpha)$ are given in the proof of Theorem 3.1, and by (A.1) in Wang et al. (1997), with an extension to a general parametric model, we have the asymptotic covariance for $\hat{\beta}_{PIP}$ as

$$\boldsymbol{\Sigma}_P = \boldsymbol{D}^{-1}\{\tilde{\boldsymbol{S}} - \boldsymbol{F}(\alpha) \boldsymbol{C}^{-1}(\alpha) \boldsymbol{F}(\alpha)^\top\} \boldsymbol{D}^{-1},$$

where $\tilde{\boldsymbol{S}} = E(T_1 T_1^\top/\pi_1)$. By Wang & Wang (2001),

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{D}^{-1}(\tilde{\boldsymbol{S}} - \tilde{\boldsymbol{S}}^* + \boldsymbol{V})\boldsymbol{D}^{-1}$$

is the asymptotic covariance matrix for $\hat{\beta}$ when $\hat{\psi}$ is based on a standard kernel smoother, where $\tilde{\boldsymbol{S}}^* = E(\psi_1 \psi_1^\top/\pi_1)$.

First we show that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}}$. By the construction of the covariances, we only need to show that $\tilde{\boldsymbol{S}}^* - \boldsymbol{V} \succeq \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top$. Define $\xi = (\sqrt{\frac{1-\pi_1}{\pi_1}}\psi_1, \frac{\pi_1'(\alpha)}{\sqrt{(1-\pi_1)\pi_1}})^\top$. Then we have

$$E(\xi\xi^\top) = \begin{pmatrix} E\left(\frac{1-\pi_1}{\pi_1}\psi_1\psi_1^\top\right) & E\left\{\frac{1}{\pi_1}\psi_1\pi_1'(\alpha)^\top\right\} \\ E\left\{\frac{1}{\pi_1}\pi_1'(\alpha)\psi_1^\top\right\} & E\left\{\frac{\pi_1'(\alpha)\pi_1'(\alpha)^\top}{(1-\pi_1)\pi_1}\right\} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{S}}^* - \boldsymbol{V} & \boldsymbol{F}(\alpha) \\ \boldsymbol{F}(\alpha)^\top & \boldsymbol{C}(\alpha) \end{pmatrix} \succeq 0.$$

By the Schur complement condition of the matrix above, we have

$$(\tilde{\boldsymbol{S}}^* - \boldsymbol{V}) - \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top \succeq 0.$$

Therefore, $\tilde{\boldsymbol{S}}^* - \boldsymbol{V} \succeq \boldsymbol{F}(\alpha)\boldsymbol{C}^{-1}(\alpha)\boldsymbol{F}(\alpha)^\top$, which implies that $\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}}$.

Next, we show that $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}$ and thus the asymptotic equivalence between $\hat{\beta}_A$ and $\hat{\beta}_{PIPA}$. Based on the results of Theorem 3.1, we can rewrite $\Delta_3(\beta, \hat{\psi}(\gamma))$ as

$$\begin{aligned}\Delta_3(\beta, \hat{\psi}(\gamma)) &= n^{-1/2}\sum_{i=1}^n U_i + O_p(\eta_n) \\ &= n^{-1/2}\sum_{i=1}^n \left[\frac{R_i}{\pi_i^*(\gamma)}T_i + \left\{1 - \frac{R_i}{\pi_i^*(\gamma)}\right\}\psi_i + R_i a(Q_i^\top\gamma)\left\{(T_i^0 - \psi_i^0) - (T_i - \psi_i)\right\}\right] + O_p(\eta_n).\end{aligned}$$

The condition $E(Z_i|u_i) = Z_i$ implies that $T_i^0 = T_i^1 = T_i$ and $\psi_i^0(\gamma) = \psi_i^1(\gamma) = \psi_i(\gamma)$. by Theorem 3.2, both $\Delta_2(\beta, \pi(\hat{\alpha}), \hat{\psi}(\gamma))$ and $\Delta_3(\beta, \hat{\psi}(\gamma))$ are asymptotically equivalent to $\Delta_2(\beta, \pi^*(\gamma), \psi)$ and thus have the same asymptotic covariance matrix as

$$\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA} = \boldsymbol{D}^{-1}(\boldsymbol{S} - \boldsymbol{S}^* + \boldsymbol{V})\boldsymbol{D}^{-1}.$$

Recall the condition of Lemma 3.2(c) that $\pi_i = \pi_i^*(\gamma)$. Then $\boldsymbol{S} = \tilde{\boldsymbol{S}}$, $\boldsymbol{S}^* = \tilde{\boldsymbol{S}}^*$. Thus we finally have

$$\boldsymbol{\Sigma}_P \succeq \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_{PA}.$$

□

3.6 Simulations

In this section, we investigate the performance of the proposed estimator $\hat{\beta}_A$ compared to other estimators, in terms of bias and standard error. We also examine the covariance estimation using the sandwich-formula (3.13), by comparing the asymptotic standard error with the empirical standard error. The empirical standard error is obtained from 1000 estimates through independent Monte Carlo simulations under the same data-generating conditions. The asymptotic normality of the estimators is examined by calculating the 95% coverage probabilities. We also use this numerical study as an example to illustrate the phenomenon of highly variable inverse-probabilities, as well as the robustness of our estimator under non-normal distributions.

There are two main scenarios in our simulations. For both of them, we consider $n = 250$ and 500. In the first scenario, we have (y_i, X_i, Z_i) generated from a multivariate normal distribution with $X_i \sim N(0,1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0,1)$, $i = 1, 2, ..., n$. Thus we have $p = 4$. The true regression coefficients $\beta = (0, 0.5, 1, -1, -0.5)^\top$, and $y_i = W_i\beta + \varepsilon_i$ with $W_i = (1, X_i, Z_i^\top)^\top$. The selection probabilities for observing X_i are $\pi_i = \{1+\exp(-\alpha_0 - \alpha_1 y_i - \alpha_2 Z_{i1} - \alpha_3 Z_{i2} - \alpha_4 Z_{i3})\}^{-1}$, which satisfy MAR on X. In this setting, the single-index model on the augmentation is easily seen to be valid. We have three different choices for the values of α. On average, there are about 20%, 40% and 60% of the cases that have X missing. We choose to use a second-order Gaussian kernel function ($r = 2$). The bandwidth selection has been discussed in the previous section. In practice, since X_i^2 is more variable than X_i, we use $h = 0.4\hat{\sigma}_u n^{-1/3}$ when estimating $E(X_i^2|Q_i)$. The coefficient parameters are estimated through the estimating equations (3.12) by iterations in R.

We use a logistic regression model to model the missing process parametrically for $\hat{\beta}_{PIP}$ and $\hat{\beta}_{PIPA}$. In this setting, this model is correctly specified so that theoretically they are unbiased esti-

mators. However, the (estimated) selection probabilities can be positive but near zero, which may lead to numerically biased estimates as we indicated earlier. Our empirical experience suggests that, since we only use the information in incomplete cases when estimating $\hat{E}(Z_i|u_i, R_i = 0)$, it would be helpful to include a correction factor matrix in the sandwich-formula (3.13) for small to moderate sample sizes, such as those in our simulation studies, especially when the percentage of missingness is high and the data is believed to be skewed. For example, we may replace the estimated asymptotic covariance by $\hat{\boldsymbol{\Sigma}}_A^* = \boldsymbol{F}_c \cdot \hat{\boldsymbol{\Sigma}}_A$, where $\boldsymbol{F}_c = \text{diag}\{a, ..., a, b, a, ..., a\}^{-1}$, $a = 1 - 0.3 \times \text{miss}\%$, $b = (1 - 0.7 \times \text{miss}\%) \cdot \min(\exp\{(n - 500)/5000\}, 1)$, and miss% means the percentage of missingness of X in the data set. The position of b matches the position of the coefficient of the missing covariate. This is what we used for $\hat{\beta}_A$ in our numerical results. For simulation purposes, we also show the results of the full data $\hat{\beta}_F$ as a benchmark for comparison.

The results for the first scenario are displayed in Tables 3.1–3.3. For each estimator, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error by formula over 1000 replications, and the third line is the 95% coverage probability. Since the conditions for Theorem 3.2 are not satisfied in the simulation studies, we do not have a closed form for $\hat{\Sigma}_{PA}$. Thus we put an "$*$" in the places of averaged asymptotic SE of $\hat{\beta}_{PIPA}$ and use the 1% trimmed empirical SE to calculate the 95% coverage probabilities. The reason to use the trimmed SE is that we have some extremely "bad" results caused by the near-zero selection probabilities and these few extreme values make the empirical SE too large compared to other estimators. As expected, the CC analysis produces biased estimates in this scenario. We also observe that $\hat{\beta}_{PIP}$ always has significant bias for each parameter, and $\hat{\beta}_{PIPA}$ has bias at least for β_1, the coefficient of X, even with $n = 500$. Moreover, the above two estimators have much larger standard errors than $\hat{\beta}_A$. Given the multivariate normal data and correctly specified logistic model for the selection probabilities in Tables 3.1–3.3, $\hat{\beta}_{PIP}$ and $\hat{\beta}_{PIPA}$ should be consistent. The deviation from the expectation arises from the estimated positive but near-zero selection probabilities. These inverse-probability weights make $\hat{\beta}_{PIP}$ and $\hat{\beta}_{PIPA}$ unstable and skewed distributed, resulting in large standard errors and biases. The near-zero selection probabilities also have influence on the

sandwich-formula of the asymptotic covariance, making the averaged asymptotic standard error very different from the empirical standard error and resulting in low coverages. On the other hand, our proposed estimator $\hat{\beta}_A$ performs well on bias and standard error. Its asymptotic standard error is also close to the empirical standard error. The 95% coverage probabilities of $\hat{\beta}_A$ are reasonable.

In the second scenario, we use non-normal distributions to generate data. Specifically, we generate X_i from a standardized gamma distribution $(Gamma(5,1)-5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$ and ε_i from a standardized t distribution with $df = 5$ as $t_5/\sqrt{5/3}$. We keep the same settings for the parameters. The results for the second scenario are displayed in Tables 3.4–3.6. In this setting, the parametric model for selection probabilities is still valid, but the single-index model on the augmentation is not. However, we get conclusions similar to those from the first scenario. Estimators $\hat{\beta}_{PIP}$ and $\hat{\beta}_{PIPA}$ still have large biases and standard errors. Our proposed estimator has a slightly low coverage for β_1, but is much better compared to other estimators in terms of bias, standard errors and coverage probabilities.

These simulation results illustrate our point on the numerical issues in $\hat{\beta}_{PIP}$ and $\hat{\beta}_{PIPA}$ that are mainly caused by estimated positive but near-zero selection probabilities. Our proposed estimator $\hat{\beta}_A$ has its advantage of not only the simplicity but also that it does not need to make parametric model assumption on the selection probabilities or the conditional covariate distribution $p(X|y, Z)$. It is not sensitive to the near-zero selection probabilities and gives pretty robust estimates even when the single-index model is misspecified.

Although both scenarios have a continuous missing covariate X, our method can also be applied to the situations with a categorical missing covariate. The parallel theory should still be valid as long as the single-index model $E(X_i|Q_i) = g(Q_i^\top \gamma)$ is still true (e.g., GLMs). When X is a binary variable, the estimation procedure can even be simpler because $E(X_i^2|Q_i) = E(X_i|Q_i)$.

3.7 Illustrative Example of Data Analysis

In this section we apply our proposed method to the data collected from the Canada 2010/2011 Youth Smoking Survey (YSS). The 2010/2011 Youth Smoking Survey (YSS) is a Health Canada sponsored pan-Canadian, classroom-based survey of a representative sample of students in grades

Table 3.1: Simulation results of 1000 replications for the normal data, $X_i \sim N(0,1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0,1)$, with $\alpha = (2.2, -0.9, -0.7, 0, 0)$, about 20% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0024 0.0020/0.0020 0.943	-0.0014 0.0021/0.0020 0.936	-0.0001 0.0020/0.0020 0.953	0.0006 0.0020/0.0020 0.944	0.0011 0.0020/0.0020 0.943
	$\hat{\beta}_{CC}$	-0.1527 0.0023/0.0022 0.436	-0.0307 0.0023/0.0022 0.917	-0.1051 0.0024/0.0024 0.712	0.0575 0.0023/0.0023 0.870	0.0309 0.0022/0.0022 0.924
	$\hat{\beta}_{PIP}$	-0.0120 0.0026/0.0022 0.909	-0.0026 0.0029/0.0024 0.904	-0.0150 0.0033/0.0025 0.854	0.0072 0.0029/0.0024 0.905	0.0068 0.0028/0.0024 0.910
	$\hat{\beta}_{PIPA}$	-0.0048 0.0021/$*$ 0.926	0.0100 0.0025/$*$ 0.928	0.0002 0.0022/$*$ 0.927	0.0074 0.0021/$*$ 0.931	0.0019 0.0021/$*$ 0.934
	$\hat{\beta}_A$	-0.0009 0.0021/0.0021 0.949	-0.0059 0.0023/0.0024 0.956	0.0010 0.0021/0.0020 0.934	-0.0002 0.0021/0.0020 0.944	0.0004 0.0021/0.0020 0.940
$n = 500$	$\hat{\beta}_F$	0.0004 0.0015/0.0014 0.941	-0.0014 0.0015/0.0014 0.938	0.0004 0.0013/0.0014 0.960	-0.0016 0.0014/0.0014 0.948	0.0024 0.0015/0.0014 0.941
	$\hat{\beta}_{CC}$	-0.1512 0.0017/0.0016 0.148	-0.0299 0.0016/0.0015 0.893	-0.1045 0.0016/0.0017 0.501	0.0573 0.0016/0.0016 0.779	0.0317 0.0016/0.0016 0.891
	$\hat{\beta}_{PIP}$	-0.0053 0.0019/0.0016 0.903	-0.0031 0.0021/0.0018 0.911	-0.0084 0.0024/0.0020 0.891	0.0036 0.0022/0.0019 0.904	0.0044 0.0021/0.0018 0.919
	$\hat{\beta}_{PIPA}$	0.0028 0.0015/$*$ 0.927	0.0294 0.0018/$*$ 0.934	0.0090 0.0016/$*$ 0.927	-0.0154 0.0016/$*$ 0.927	0.0007 0.0015/$*$ 0.928
	$\hat{\beta}_A$	0.0009 0.0015/0.0015 0.945	-0.0037 0.0017/0.0017 0.953	0.0005 0.0014/0.0014 0.946	-0.0018 0.0015/0.0014 0.941	0.0018 0.0015/0.0014 0.940

Table 3.2: Simulation results of 1000 replications for the normal data, $X_i \sim N(0,1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0,1)$, with $\alpha = (0.5, -1, -0.5, 0, 0)$, about 40% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0024 0.0020/0.0020 0.943	-0.0014 0.0021/0.0020 0.936	-0.0001 0.0020/0.0020 0.953	0.0006 0.0020/0.0020 0.944	0.0011 0.0020/0.0020 0.943
	$\hat{\beta}_{CC}$	-0.3499 0.0028/0.0028 0.027	-0.0578 0.0025/0.0025 0.884	-0.1642 0.0029/0.0028 0.543	0.1086 0.0027/0.0026 0.727	0.0536 0.0026/0.0025 0.880
	$\hat{\beta}_{PIP}$	-0.0462 0.0035/0.0028 0.814	-0.0166 0.0038/0.0030 0.869	-0.0432 0.0044/0.0032 0.793	0.0258 0.0043/0.0031 0.822	0.0151 0.0038/0.0030 0.861
	$\hat{\beta}_{PIPA}$	-0.0037 0.0026/$*$ 0.922	0.0197 0.0040/$*$ 0.918	-0.0075 0.0028/$*$ 0.926	0.0127 0.0026/$*$ 0.928	0.0018 0.0025/$*$ 0.934
	$\hat{\beta}_A$	-0.0002 0.0022/0.0023 0.956	-0.0127 0.0028/0.0029 0.959	0.0005 0.0021/0.0020 0.935	-0.0006 0.0021/0.0021 0.938	-0.0003 0.0021/0.0021 0.941
$n = 500$	$\hat{\beta}_F$	0.0004 0.0015/0.0014 0.941	-0.0014 0.0015/0.0014 0.938	0.0004 0.0013/0.0014 0.960	-0.0016 0.0014/0.0014 0.948	0.0024 0.0015/0.0014 0.941
	$\hat{\beta}_{CC}$	-0.3458 0.0020/0.0020 0.000	-0.0548 0.0018/0.0018 0.824	-0.1631 0.0020/0.0020 0.257	0.1080 0.0019/0.0019 0.554	0.0558 0.0018/0.0018 0.831
	$\hat{\beta}_{PIP}$	-0.0278 0.0027/0.0022 0.852	-0.0111 0.0029/0.0024 0.907	-0.0292 0.0035/0.0026 0.819	0.0203 0.0033/0.0025 0.844	0.0148 0.0029/0.0024 0.880
	$\hat{\beta}_{PIPA}$	-0.0088 0.0018/$*$ 0.927	-0.0210 0.0027/$*$ 0.925	-0.0023 0.0019/$*$ 0.928	0.0216 0.0018/$*$ 0.931	0.0121 0.0018/$*$ 0.934
	$\hat{\beta}_A$	0.0020 0.0016/0.0016 0.954	-0.0065 0.0020/0.0021 0.950	0.0005 0.0015/0.0014 0.947	-0.0024 0.0015/0.0015 0.936	0.0012 0.0015/0.0015 0.935

Table 3.3: Simulation results of 1000 replications for the normal data, $X_i \sim N(0,1)$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim N(0,1)$, with $\alpha = (-0.5, -0.5, -0.5, 0, 0)$, about 60% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0024 0.0020/0.0020 0.943	-0.0014 0.0021/0.0020 0.936	-0.0001 0.0020/0.0020 0.953	0.0006 0.0020/0.0020 0.944	0.0011 0.0020/0.0020 0.943
	$\hat{\beta}_{CC}$	-0.2830 0.0036/0.0036 0.288	-0.0227 0.0032/0.0030 0.929	-0.0902 0.0035/0.0033 0.832	0.0447 0.0032/0.0031 0.917	0.0243 0.0033/0.0031 0.915
	$\hat{\beta}_{PIP}$	-0.0269 0.0037/0.0030 0.861	-0.0037 0.0042/0.0035 0.872	-0.0174 0.0049/0.0036 0.827	0.0098 0.0046/0.0036 0.862	0.0085 0.0042/0.0035 0.897
	$\hat{\beta}_{PIPA}$	-0.0063 0.0026/$*$ 0.933	0.0290 0.0040/$*$ 0.922	-0.0041 0.0028/$*$ 0.930	0.0021 0.0028/$*$ 0.929	< 0.0001 0.0026/$*$ 0.922
	$\hat{\beta}_A$	-0.0023 0.0023/0.0025 0.964	-0.0083 0.0032/0.0035 0.968	0.0003 0.0021/0.0021 0.936	-0.0006 0.0022/0.0021 0.937	-0.0003 0.0022/0.0021 0.934
$n = 500$	$\hat{\beta}_F$	0.0004 0.0015/0.0014 0.941	-0.0014 0.0015/0.0014 0.938	0.0004 0.0013/0.0014 0.960	-0.0016 0.0014/0.0014 0.948	0.0024 0.0015/0.0014 0.941
	$\hat{\beta}_{CC}$	-0.2805 0.0026/0.0025 0.062	-0.0226 0.0022/0.0022 0.923	-0.0907 0.0024/0.0024 0.774	0.0428 0.0023/0.0022 0.888	0.0261 0.0023/0.0022 0.915
	$\hat{\beta}_{PIP}$	-0.0107 0.0026/0.0023 0.896	-0.0022 0.0030/0.0027 0.908	-0.0093 0.0036/0.0029 0.876	0.0033 0.0033/0.0028 0.890	0.0079 0.0031/0.0027 0.908
	$\hat{\beta}_{PIPA}$	0.0005 0.0019/$*$ 0.919	0.0107 0.0026/$*$ 0.933	0.0001 0.0020/$*$ 0.926	-0.0040 0.0019/$*$ 0.932	0.0034 0.0018/$*$ 0.933
	$\hat{\beta}_A$	0.0015 0.0017/0.0018 0.944	-0.0055 0.0022/0.0025 0.975	0.0004 0.0015/0.0015 0.954	-0.0022 0.0015/0.0015 0.933	0.0015 0.0016/0.0015 0.932

Table 3.4: Simulation results of 1000 replications for the normal data, $X_i \sim (Gamma(5,1) - 5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (2.2, -0.9, -0.7, 0, 0)$, about 20% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0012 0.0020/0.0020 0.950	0.0009 0.0021/0.0020 0.940	-0.0018 0.0020/0.0020 0.956	0.0005 0.0020/0.0020 0.939	0.0001 0.0020/0.0020 0.949
	$\hat{\beta}_{CC}$	-0.1481 0.0022/0.0022 0.447	-0.0284 0.0023/0.0022 0.918	-0.0988 0.0024/0.0024 0.725	0.0570 0.0023/0.0022 0.859	0.0294 0.0022/0.0022 0.921
	$\hat{\beta}_{PIP}$	-0.0160 0.0029/0.0022 0.883	-0.0065 0.0033/0.0024 0.884	-0.0209 0.0032/0.0025 0.877	0.0091 0.0034/0.0024 0.879	0.0075 0.0029/0.0024 0.906
	$\hat{\beta}_{PIPA}$	-0.0019 0.0021/$*$ 0.925	0.0029 0.0029/$*$ 0.930	0.0002 0.0023/$*$ 0.929	-0.0045 0.0021/$*$ 0.923	-0.0003 0.0021/$*$ 0.929
	$\hat{\beta}_A$	-0.0005 0.0021/0.0021 0.961	-0.0012 0.0025/0.0024 0.945	0.0001 0.0021/0.0020 0.941	-0.0008 0.0021/0.0020 0.937	0.0005 0.0020/0.0020 0.945
$n = 500$	$\hat{\beta}_F$	-0.0009 0.0014/0.0014 0.957	-0.0015 0.0014/0.0014 0.948	-0.0006 0.0013/0.0014 0.956	-0.0012 0.0015/0.0014 0.943	-0.0006 0.0014/0.0014 0.954
	$\hat{\beta}_{CC}$	-0.1479 0.0016/0.0016 0.146	-0.0317 0.0016/0.0016 0.886	-0.0980 0.0016/0.0017 0.558	0.0544 0.0016/0.0016 0.807	0.0279 0.0015/0.0015 0.908
	$\hat{\beta}_{PIP}$	-0.0103 0.0022/0.0017 0.895	-0.0089 0.0023/0.0019 0.902	-0.0138 0.0025/0.0020 0.885	0.0063 0.0025/0.0019 0.914	0.0041 0.0023/0.0018 0.918
	$\hat{\beta}_{PIPA}$	-0.0033 0.0015/$*$ 0.930	0.0083 0.0021/$*$ 0.934	-0.0040 0.0016/$*$ 0.932	-0.0012 0.0015/$*$ 0.923	0.0007 0.0015/$*$ 0.926
	$\hat{\beta}_A$	-0.0009 0.0015/0.0015 0.959	-0.0017 0.0018/0.0017 0.930	-0.0002 0.0014/0.0014 0.943	-0.0017 0.0015/0.0014 0.933	-0.0003 0.0014/0.0014 0.943

Table 3.5: Simulation results of 1000 replications for the normal data, $X_i \sim (Gamma(5,1) - 5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (0.5, -1, -0.5, 0, 0)$, about 40% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0012 0.0020/0.0020 0.950	0.0009 0.0021/0.0020 0.940	-0.0018 0.0020/0.0020 0.956	0.0005 0.0020/0.0020 0.939	0.0001 0.0020/0.0020 0.949
	$\hat{\beta}_{CC}$	-0.3292 0.0031/0.0029 0.048	-0.0527 0.0028/0.0027 0.899	-0.1561 0.0029/0.0029 0.594	0.1030 0.0027/0.0027 0.755	0.0542 0.0026/0.0025 0.877
	$\hat{\beta}_{PIP}$	-0.0546 0.0037/0.0027 0.758	-0.0183 0.0040/0.0031 0.860	-0.0467 0.0041/0.0030 0.794	0.0282 0.0040/0.0030 0.842	0.0166 0.0038/0.0029 0.869
	$\hat{\beta}_{PIPA}$	0.0002 0.0026/$*$ 0.931	0.0300 0.0044/$*$ 0.918	0.0035 0.0027/$*$ 0.937	-0.0077 0.0026/$*$ 0.930	-0.0020 0.0024/$*$ 0.933
	$\hat{\beta}_A$	< 0.0001 0.0022/0.0023 0.947	-0.0012 0.0032/0.0030 0.925	0.0003 0.0021/0.0020 0.942	-0.0001 0.0022/0.0021 0.932	0.0002 0.0021/0.0021 0.936
$n = 500$	$\hat{\beta}_F$	-0.0009 0.0014/0.0014 0.957	-0.0015 0.0014/0.0014 0.948	-0.0006 0.0013/0.0014 0.956	-0.0012 0.0015/0.0014 0.943	-0.0006 0.0014/0.0014 0.954
	$\hat{\beta}_{CC}$	-0.3306 0.0021/0.0021 0.000	-0.0570 0.0020/0.0019 0.838	-0.1562 0.0020/0.0020 0.288	0.1014 0.0019/0.0019 0.592	0.0526 0.0018/0.0018 0.837
	$\hat{\beta}_{PIP}$	-0.0317 0.0037/0.0023 0.792	-0.0188 0.0034/0.0026 0.863	-0.0291 0.0040/0.0026 0.824	0.0189 0.0033/0.0025 0.863	0.0154 0.0034/0.0024 0.882
	$\hat{\beta}_{PIPA}$	0.0004 0.0019/$*$ 0.924	0.0306 0.0032/$*$ 0.922	-0.0015 0.0019/$*$ 0.930	-0.0140 0.0018/$*$ 0.924	-0.0108 0.0018/$*$ 0.925
	$\hat{\beta}_A$	-0.0008 0.0016/0.0016 0.949	-0.0033 0.0024/0.0021 0.921	-0.0001 0.0015/0.0014 0.942	-0.0017 0.0016/0.0015 0.921	-0.0005 0.0015/0.0015 0.942

Table 3.6: Simulation results of 1000 replications for the normal data, $X_i \sim (Gamma(5,1) - 5)/\sqrt{5}$, $Z_i \sim N_3(0, \boldsymbol{I}_3)$, $\varepsilon_i \sim t_5/\sqrt{5/3}$, with $\alpha = (-0.5, -0.5, -0.5, 0, 0)$, about 60% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability. An "$*$" indicates the asymptotic standard error formula unavailable.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$	$\hat{\beta}_4 - \beta_4$
$n = 250$	$\hat{\beta}_F$	-0.0012 0.0020/0.0020 0.950	0.0009 0.0021/0.0020 0.940	-0.0018 0.0020/0.0020 0.956	0.0005 0.0020/0.0020 0.939	0.0001 0.0020/0.0020 0.949
	$\hat{\beta}_{CC}$	-0.2777 0.0040/0.0037 0.329	-0.0279 0.0034/0.0033 0.924	-0.0964 0.0035/0.0034 0.842	0.0449 0.0032/0.0032 0.918	0.0266 0.0033/0.0031 0.916
	$\hat{\beta}_{PIP}$	-0.0314 0.0038/0.0029 0.839	-0.0137 0.0043/0.0036 0.879	-0.0327 0.0047/0.0036 0.849	0.0068 0.0046/0.0034 0.877	0.0100 0.0043/0.0034 0.886
	$\hat{\beta}_{PIPA}$	-0.0477 0.0026/$*$ 0.931	0.1346 0.0042/$*$ 0.924	-0.1189 0.0027/$*$ 0.924	0.0309 0.0025/$*$ 0.925	0.0571 0.0025/$*$ 0.923
	$\hat{\beta}_A$	0.0013 0.0023/0.0025 0.960	-0.0029 0.0035/0.0035 0.947	-0.0005 0.0022/0.0021 0.937	-0.0005 0.0022/0.0021 0.927	0.0003 0.0022/0.0021 0.929
$n = 500$	$\hat{\beta}_F$	-0.0009 0.0014/0.0014 0.957	-0.0015 0.0014/0.0014 0.948	-0.0006 0.0013/0.0014 0.956	-0.0012 0.0015/0.0014 0.943	-0.0006 0.0014/0.0014 0.954
	$\hat{\beta}_{CC}$	-0.2790 0.0027/0.0026 0.067	-0.0256 0.0023/0.0023 0.924	-0.0958 0.0024/0.0024 0.768	0.0447 0.0022/0.0023 0.897	0.0284 0.0024/0.0022 0.906
	$\hat{\beta}_{PIP}$	-0.0168 0.0030/0.0023 0.877	-0.0104 0.0034/0.0028 0.902	-0.0196 0.0037/0.0029 0.868	0.0041 0.0036/0.0028 0.908	0.0101 0.0033/0.0027 0.893
	$\hat{\beta}_{PIPA}$	0.0016 0.0018/$*$ 0.928	0.0118 0.0029/$*$ 0.933	0.0073 0.0019/$*$ 0.922	-0.0031 0.0018/$*$ 0.929	-0.0012 0.0019/$*$ 0.926
	$\hat{\beta}_A$	-0.0002 0.0016/0.0017 0.962	< 0.0001 0.0025/0.0025 0.963	-0.0002 0.0015/0.0015 0.938	-0.0019 0.0016/0.0015 0.921	-0.0003 0.0016/0.0015 0.927

6 through 12. The 2010/2011 YSS was implemented in schools between October 2010 and June 2011 by provincial level teams located in the 9 participating provinces in Canada. More details can be found in *2010/2011 YOUTH SMOKING SURVEY MICRODATA USER GUIDE*, or from `https://uwaterloo.ca/canadian-student-tobacco-alcohol-drugs-survey`.

We focus on data collected from Asian students (Grade 6 through 8). The main interest is to explore the correlation between the students' self-esteem scores and smoking status, controlling other covariates as sex, marks and BMI. The variables used are displayed below:

1. $esteem$: a 0 to 12 score measuring the student's overall self-esteem;

2. sex: a binary variable indicating the student's gender (0 for female and 1 for male);

3. $marks$: a categorical variable with five levels describing the student's marks during the past year: mostly A's (1), mostly A's and B's (2), mostly B's and C's (3), mostly C's (4) and mostly below C's (5);

4. $smoke$: originally a categorical variable with 3 levels: currently smokes, formerly smoked and never smoked. In this data set of Asian students from Grade 6 to 8, we do not have students in status of "formerly smoked". Thus we can regard this variable as binary for smoking ($smoke = 1$) or not ($smoke = 0$);

5. BMI: a continuous variable that measures the respondent's body mass index.

We take a subset with size $n = 493$, which has complete observations on $esteem$, sex, $marks$ and $smoke$. In this data set, there are 121, 160 and 212 students in Grades 6 through 8, respectively. There are 252 female students and 241 male students, and only 9 smokers and 484 non-smokers. But 29.2% (144 out of 493) students have BMI missing. We consider a linear model on the self-esteem score as

$$esteem = \beta_0 + \beta_1 BMI + \beta_2 sex + \beta_3 marks + \beta_4 smoke + \varepsilon.$$

Table 3.7: 2010/2011 YSS data analysis focusing on Asian students ($n = 493$)

	$\hat{\beta}_0(intercept)$	$\hat{\beta}_1(BMI)$	$\hat{\beta}_2(sex)$	$\hat{\beta}_3(marks)$	$\hat{\beta}_4(smoke)$
$\hat{\beta}_{CC}$	12.1476(0.6649)	-0.0975(0.0291)	-0.0177(0.2322)	-0.6646(0.1747)	-1.1455(0.9920)
p-value	< 0.0001	0.0009	0.9392	0.0002	0.2488
$\hat{\beta}_{PIP}$	12.0189(0.7695)	-0.0966(0.0314)	0.0388(0.2251)	-0.5395(0.1858)	-1.4591(1.1036)
p-value	< 0.0001	0.0022	0.8633	0.0039	0.1867
$\hat{\beta}_{PIPA}$	12.4119(0.6205)	-0.1092(0.0338)	0.0038(0.1976)	-0.6034(0.1572)	-3.0881(1.2228)
p-value	< 0.0001	0.0133	0.9846	0.0001	0.0119
$\hat{\beta}_{A}$	12.2657(0.6332)	-0.0991(0.0314)	-0.0178(0.2046)	-0.6241(0.1512)	-3.2420(1.2059)
p-value	< 0.0001	0.0017	0.9305	< 0.0001	0.0074

Assume that the missing mechanism is MAR and that the parametric model for selection probabilities is

$$logit(\pi) = \alpha_0 + \alpha_1 esteem + \alpha_2 sex + \alpha_3 marks + \alpha_4 smoke.$$

After fitting a logistic model, we find the p-values for α_1, α_2 and α_3 as 0.0321, 0.0431 and 0.0159 respectively, so that $esteem$, sex and $marks$ are significant at the significance level of 0.05.

Before estimating the regression coefficients β, we first look at the self-esteem scores of the 8 smokers: $\{0, 0, 0, 2, 7, 8, 9, 9, 12\}$. We find most of them are lower than the average self-esteem score of the non-smokers of 9.24, and 4 of them have extremely low scores. The difference can also be found in Figure 3.1 for the exploratory data analysis. This in some sense implies that the smokers among the students have a lower self-esteem score compared to the non-smokers. The results of the analysis can be found in Table 3.7. In the estimating procedure of $\hat{\beta}_{PIPA}$, we find $\pi_i(\hat{\alpha})$ and $\hat{\pi}_i(\hat{\gamma})$ are very close. This indicates that the assumptions in Theorem 3.2 might be reasonable in this situation. The values in brackets are the standard errors of the corresponding estimators using the sandwich-formulas (3.13) and (3.14).

In Table 3.7 we observe that all methods conclude that BMI and $marks$ are significant in the linear model, while sex is insignificant. The significant effects show that higher body mass index and worse marks lead to lower self-esteem scores. The main difference lies in the effect of $smoke$. The complete-case analysis and inverse-probability method give insignificant results,

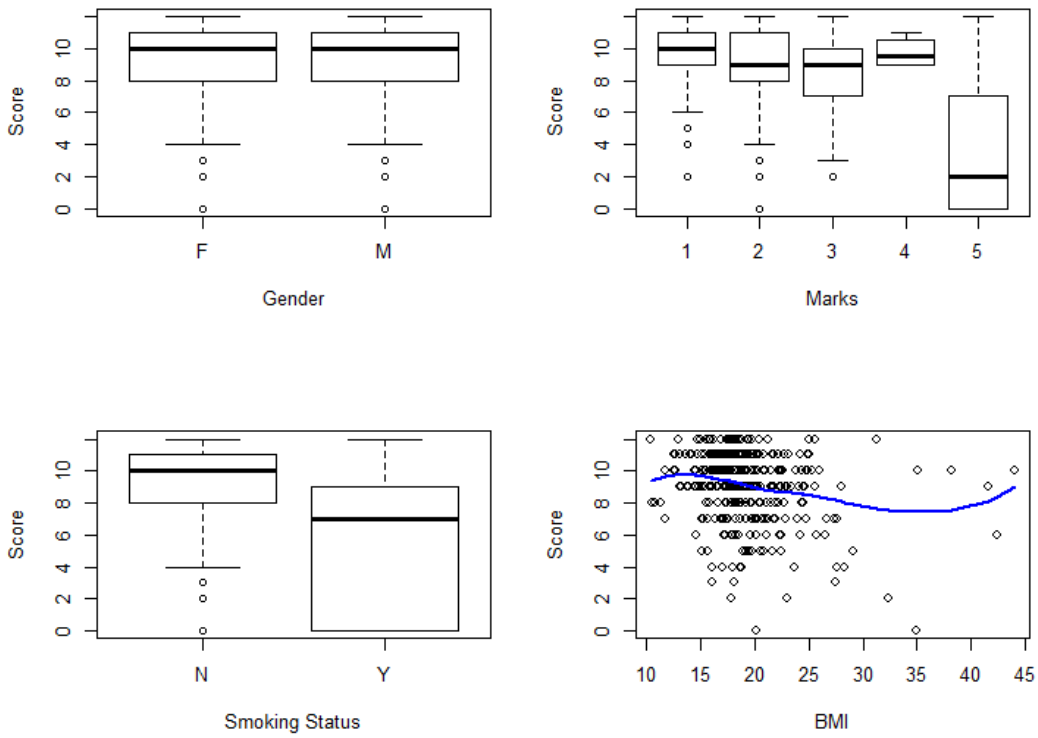

Figure 3.1: Plots for showing the relationship between self-esteem score and other variables. Top left: Side-by-side boxplot of Self-esteem score vs. Gender; Top right: Side-by-side boxplot of Self-esteem score vs. Marks; Bottom left: Side-by-side boxplot of Self-esteem score vs. Smoking Status; Bottom right: Scatterplot of Self-esteem score vs. BMI.

but the rest of the methods conclude significance. The difference is caused mainly by the smaller absolute value of the point estimates of the first two methods, compared to $\hat{\beta}_{PIPA}$ and $\hat{\beta}_{A}$. Since the missing mechanism is not completely missing at random, the results of $\hat{\beta}_{CC}$ are likely to be biased. Combined with the comparison of the self-esteem scores between smokers and non-smokers mentioned before, we believe that the results of significance are more reliable. The performance of $\hat{\beta}_{PIP}$ might be explained by the misspecification of the model on selection probabilities. Based on the analysis above, we conclude that Asian students in Grade 6 through 8 who smoke have a significantly lower self-esteem score compared to the non-smokers, controlling other covariates, BMI, sex and $marks$ during the past year.

3.8 Concluding Remarks

In this chapter we have proposed an unweighted mean-score-form estimator of regression coefficients through GEE with a single-index model when some covariates are missing at random. This is a semiparametric estimation approach since we only assume a single-index model on augmentation without making any distribution assumptions. We do not even specify a parametric model such as a logistic model for the missingness mechanism. We have also introduced the standard doubly robust estimator $\hat{\beta}_{PIPA}$ with the same single-index model on augmentation and parametrically modeled selection probabilities. We have presented the asymptotic distribution for $\hat{\beta}_{PIPA}$ and $\hat{\beta}_A$ in Theorems 3.1 and 3.2, along with the sandwich-formulas of the asymptotic covariances and the choice of the bandwidth. We also have shown the asymptotic equivalence between the two augmented estimators under certain conditions. However, one important advantage of our proposed estimator over the (augmented) inverse-probability weighted estimators is that it does not include selection probabilities in the point estimation procedure so that it does not need to model π_i's and avoids the situation of having highly variable inverse-probability weights, as described in Robins et al. (2007). In this sense, numerically our proposed estimator is not sensitive to positive but near-zero selection probabilities, while the performance of the inverse-probability weighted estimators are highly influenced by those near-zero π_i's. Furthermore, compared to using a standard multivariate kernel function, the SIM we use on augmentation not only avoids the curse of dimensionality, but also keeps the efficiency of standard kernel smoothing in some particular situations. The R code used in our simulations and the example can be found on the following website: `https://github.com/zhuoersun/Missing-Data`.

In this work, we only considered a single univariate covariate X in simulation studies and the real data example. The results can be easily extended to the particular case of a multivariate X when $R_i = 0$ means that all the covariates in X_i are missing at the same time. It would be interesting but more challenging to consider more complex missingness patterns such as monotone or non-monotone missingness in covariates. One can refer to Chen (2004) and Sinha et al. (2014) for more information. It would be natural to extend the proposed methodology to generalized

linear models. However, generalized linear models have a more complicated score function and do not have the simple form of augmentation like (3.6). Further investigation will be required in this important problem. Yet as another future research problem, it would also be interesting to apply this idea to longitudinal data with some covariates partially missing.

4. SEMIPARAMETRIC ESTIMATION IN REGRESSION WITH MISSING COVARIATES FOR LONGITUDINAL DATA

4.1 Introduction

Longitudinal data analysis is very common in many fields of research studies, especially in sociology, biomedical science, clinical trials and public health research. It is often of interest to estimate the parameter β of a longitudinal regression model with time-varying responses and covariates. However, it is quite possible that some of the covariates in the data are partially missing at some observation times. This can be caused by unavailability of covariate measurements, study subjects' refusal to answer the questions or to continue the participation, patients' deaths and many other reasons. An extensive literature discussed the situation of informative dropouts (Diggle (2002)) under the missing at random (MAR) mechanism (Little & Rubin (2014)). When the joint likelihood of the response and covariates is available through normal random effect model or generalized linear mixed model (GLMM) with the data MAR, regular maximum likelihood methods such as EM algorithm give consistent estimates (Horton & Laird (1999), Fuchs (1982), Schluchter & Jackson (1989) and Ibrahim (1990)). Specifically, by modeling the dropout process in addition, one can use selection models or pattern-mixture models based on two different factorizations of the joint density of the responses, covariates and missingness indicators (Little (1993), Little (1995)). These two models also work under the missing not at random (MNAR) mechanism (Ibrahim & Molenberghs (2009)). Bayesian methods (Daniels & Hogan (2008)) and multiple imputation (Schafer (1997), Rubin (2004)) can also be considered to handle the missing data problem. However, all the above methods usually require likelihood assumptions and can be sensitive to model misspecification.

For complete longitudinal data, Liang & Zeger (1986) proposed to perform the analysis based on generalized estimating equations (GEE). With a working covariance structure, this semiparametric method gives consistent results once the marginal mean of the outcomes at each time is cor-

rectly specified. But when we have incomplete data, GEE generally produces unbiased estimates only under missing completely at random (MCAR) mechanism. Robins et al. (1994) first introduced a class of inverse-probability weighted estimators (IPW) and augmented inverse-probability weighted estimators (AIPW) for i.i.d. data based on GEE when data are MAR. The weights are obtained from the parametric models for the selection probabilities. These models need to be correctly specified to guarantee the consistency of the estimation for IPW. By choosing the augmentation to be the conditional expectation of the score function in the first part of the estimating equations, AIPW is doubly robust (DR). That is to say, AIPW is consistent when either the selection probability model or the missing covariate model conditional on the observed data is correctly specified. Robins et al. (1995) extended the idea of IPW to longitudinal data with monotone missing response.

Recently, more research works have been done on DR estimators for incomplete longitudinal data. Lipsitz et al. (1999) first introduced the DR estimator for cross-sectional studies with a missing covariate and properties similar to maximum likelihood. Other literature includes Van der Laan & Robins (2003), Bang & Robins (2005) and Seaman & Copas (2009). But this literature mainly solves the problem for monotone missing incomplete response. Dealing with non-monotone missing values is generally more difficult because the variety of patterns makes the factorization of the likelihood very challenging. Chen et al. (2010) and Chen & Zhou (2011) discussed the DR estimator for both response and covariates non-monotone missing at random. These works are impressive for a more complicated situation, but many assumptions are needed for the identifiability of the models. It is not easy to calculate the marginal selection probabilities even under the correct parametric models and the augmentation may involve nontrivial integration.

The DR estimator can still be inconsistent if both of the selection probability model and the missing covariate model are misspecified. Moreover, although Robins et al. (1994) restricted the selection probabilities to be bounded away from 0, in practice you may still get some positive but near-zero values for their estimates, which can make the inverse-probabilities weights highly variable and the resulting estimators highly skewed distributed. This phenomenon is observed and

discussed at least in i.i.d. data (Kang & Schafer (2007), Robins et al. (2007)) and also exists in incomplete longitudinal data. Instead of modeling the missing process parametrically, Wang et al. (1997) and Wang & Wang (2001) proposed nonparametric kernel smoother for the selection probabilities and the augmentation. However, few works have been done for estimating the selection probabilities and augmentation nonparametrically in incomplete longitudinal data. In practice, using multivariate kernel functions also suffers from "curse of dimensionality". In this chapter, we propose a new semiparametric estimator for incomplete longitudinal data based on augmented GEE without inverse-probability weights when a covariate is non-monotone MAR. The augmentation is estimated through kernel smoothing based on a single-index model (SIM). Heteroscedasticity is allowed and we use a working independence (WI) correlation structure to simplify the estimation procedure. This approach is appealing because it does not require the specification of the joint distribution of the data or the parametric models for selection probabilities and the missing covariate. It is also simple because the point estimation procedure does not include any estimation for the selection probabilities, and thus avoids the situation of unstable inverse-probability weights.

The rest of the chapter is organized as follows. In Section 4.2 we introduce the necessary notations and briefly review IPW and AIPW. We then describe our new estimators in Section 4.3. In Section 4.4 we present an asymptotic theory of the proposed estimators and show their asymptotic consistency and normality along with sandwich formulas for asymptotic covariances. In Section 4.5 we provide the results of our simulation studies. In Section 4.6 we apply our methods to a real data example. Concluding remarks are made in Section 4.7.

4.2 Notations and Models

Consider the following longitudinal linear model:

$$Y_{ij} = \boldsymbol{W}_{ij}^{\top}\boldsymbol{\beta} + \varepsilon_{ij} = \beta_0 + X_{ij}\beta_1 + \boldsymbol{Z}_{ij}^{\top}\beta_2 + \varepsilon_{ij} \tag{4.1}$$

for $i = 1, ..., n$ and $j = 1, ..., m_i$, where $\boldsymbol{W}_{ij} = (1, X_{ij}, \boldsymbol{Z}_{ij}^{\top})^{\top}$, $\boldsymbol{\beta}$ is the vector of the regression coefficients, Y_{ij} is a continuous response and $(X_{ij}, \boldsymbol{Z}_{ij}^{\top})^{\top}$ are covariates of subject i observed at

time t_{ij} with corresponding random error ε_{ij}. Here we consider the sparse longitudinal case, which means that m_i is bounded when $n \to \infty$. In vector and matrix forms, let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{im_i})^\top$ denote the $m_i \times 1$ completely observed response vector for subject i, $\mathbf{X}_i = (X_{i1}, ..., X_{im_i})^\top$ be the covariate vector that may be partially missing at some time point t_{ij} and $\mathbf{Z}_i = (\boldsymbol{Z}_{i1}, ..., \boldsymbol{Z}_{im_i})^\top$ be the covariate matrix that is always observed. Then the whole covariates matrix can be expressed as $\mathbf{W}_i = (\boldsymbol{W}_{i1}, ..., \boldsymbol{W}_{im_i})^\top$. We assume that different subjects are mutually independent, but generally there is within-subject correlation for observations measured at different time points. Let $\mathbf{V}_i = \mathrm{Cov}(\mathbf{Y}_i|\mathbf{X}_i, \mathbf{Z}_i) = \mathrm{Cov}(\boldsymbol{\varepsilon}_i)$ with $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, ..., \varepsilon_{im_i})^\top$. Using the variance-correlation decomposition, $\mathbf{V}_i$ can be expressed as $\mathbf{V}_i = \mathbf{F}_i^{1/2}\mathbf{C}_i(\boldsymbol{\rho})\mathbf{F}_i^{1/2}$, where $\mathbf{F}_i = \mathrm{diag}(\sigma_{ij}^2)$ with $\sigma_{ij}^2 = \mathrm{var}(\varepsilon_{ij})$, $\mathbf{C}_i(\boldsymbol{\rho})$ is the correlation matrix of $\boldsymbol{\varepsilon}_i$ with parameters $\boldsymbol{\rho}$. When heteroscedasticity is allowed, we assume $\sigma_{ij}^2 = \sigma^2(t_{ij})$, where $\sigma^2(\cdot)$ is a smooth function.

To model the missing process of X_{ij}, let R_{ij} denote the indicator of the availability of X_{ij}. That is, let $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ if X_{ij} is missing. Let $\mathbf{R}_i = (R_{i1}, ..., R_{im_i})^\top$ be the indicator vector of subject i. As suggested by Chen et al. (2010) and Chen & Zhou (2011), instead of modeling $P(\mathbf{R}_i = \mathbf{r}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ directly, we can focus on the conditional models as $P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$, which reflects the nature of the observation process over time. Here $\bar{\mathbf{R}}_{ij} = (R_{i1}, ..., R_{i,j-1})^\top$ is the history of the indicators until time $t_{i,j-1}$. This form is very important for monotone missing pattern (dropout), which means that $R_{it} = 0$ implies $R_{i(t+1)} = 0$ (Robins et al. (1995)). It is also useful for non-monotone missing pattern (intermittent) because the joint distribution of $\mathbf{R}_i$ then can be expressed as

$$P(\mathbf{R}_i = \mathbf{r}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \prod_{j=2}^{m_i} P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) \cdot P(R_{i1} = r_{i1}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i). \quad (4.2)$$

When the data are missing at random (MAR) in the sense of Rubin (1976), we have $P(\mathbf{R}_i = \mathbf{r}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = P(\mathbf{R}_i = \mathbf{r}_i|\mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i)$, where $\mathbf{X}_i^o$ denotes the observed part of $\mathbf{X}_i$. Then a somewhat stronger condition is assumed as

$$P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \mathbf{Y}_i, \bar{\mathbf{X}}_{ij}^o, \mathbf{Z}_i), \quad (4.3)$$

where $\bar{\mathbf{X}}_{ij}^{o}$ represents the history of observed X_{ij} until time $t_{i,j-1}$. Since dropout means that once a subject leaves the study, return is not possible, it is not reasonable to assume that the covariate X_{ij} has the monotone missing pattern while the responses Y_{ij}'s are fully observed. Thus in this chapter, we only consider non-monotone missing patterns and assume the first measure is always observed ($R_{i1} = 1$).

Now our main interest is to consistently estimate the true regression parameters $\boldsymbol{\beta}$ when some X_{ij} are missing intermittently. For completely observed data, Liang & Zeger (1986) proposed to estimate $\boldsymbol{\beta}$ through GEE as

$$n^{-1/2}\sum_{i=1}^{n}\frac{\partial \boldsymbol{\mu}_i^{\top}}{\partial \boldsymbol{\beta}}\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})\} = n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_i^{\top}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta}) = 0,$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im_i})^{\top}$ is the vector of conditional mean with $\mu_{ij} = E(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$. The main benefits of GEE are that it does not require fully likelihood assumption on the data and the estimator is still unbiased when the correlation structure $\mathbf{C}_i$ or even the whole covariance $\mathbf{V}_i$ is misspecified. Therefore, Liang & Zeger (1986) refers $\mathbf{C}_i(\boldsymbol{\rho})$ as a 'working' correlation matrix which can be selected by the user with certain structure (compound symmetry, AR(1), etc.). A convenient choice is working independence (WI) structure, which means $\mathbf{C}_i(\boldsymbol{\rho}) = \mathbf{I}_{m_i}$. For the rest of the chapter, we will focus on the WI structure. Based on GEE, there are some existing methods when X_{ij} is partially missing.

4.2.1 Complete Case Analysis

Complete case analysis (CC) uses the following estimating equation

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{I}(\mathbf{R}_i = \mathbf{1}_{m_i})\mathbf{W}_i^{\top}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta}) = 0,$$

where $\mathbf{I}(\cdot)$ is the indicator function, $\mathbf{1}_{m_i}$ is a vector of 1 with length m_i. CC only makes use of the data of those subjects who have complete observations at each time point.

4.2.2 Available Case Analysis

Available case analysis (AC) uses the following estimating equation

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_i^\top\mathbf{M}_i(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\beta})=0,$$

where $\mathbf{M}_i = \mathbf{F}_i^{-1/2}(\mathbf{C}_i^{-1}\odot\boldsymbol{\Delta}_i)\mathbf{F}_i^{-1/2}$ with $\boldsymbol{\Delta}_i = [\delta_{ijk}]_{m_i\times m_i}$, $\delta_{ijk} = I(R_{ij}=1, R_{ik}=1)$ and $\odot$ denotes the component-wise product. Considering WI structure, $\mathbf{M}_i$ can be simplified to $\mathbf{M}_i^{\mathrm{WI}} = \mathbf{F}_i^{-1/2}\boldsymbol{\Delta}_i^{\mathrm{WI}}\mathbf{F}_i^{-1/2}$ with $\boldsymbol{\Delta}_i^{\mathrm{WI}} = \mathrm{diag}(R_{ij})$. Then the estimating equations can also be expressed as

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{\sigma_{ij}^2}R_{ij}\boldsymbol{W}_{ij}(Y_{ij}-\boldsymbol{W}_{ij}^\top\boldsymbol{\beta})=0.$$

It is obvious that AC is different from CC as it uses all observations at t_{ij} when X_{ij} is available instead of dropping the entire data of that subject. CC and AC have unbiased estimators when data is missing completely at random (MCAR), but they can lead to inconsistent estimators under MAR.

4.2.3 Inverse-probability Weighted Estimator (IPW)

To solve the problem of inconsistent estimation using GEE, Robins et al. (1994) first proposed a class of semiparametric estimators based on inverse-probability weighted GEE for i.i.d. data and Robins et al. (1995) extended it to longitudinal data with monotone missing responses. Chen et al. (2010) generalized the idea to longitudinal data with both response and covariate intermittently MAR. In our setting with missing covariate only, the estimating equation is

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_i^\top\mathbf{M}_i^*(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\beta})=0,$$

where $\mathbf{M}_i^* = \mathbf{F}_i^{-1/2}(\mathbf{C}_i^{-1} \odot \boldsymbol{\Delta}_i^*)\mathbf{F}_i^{-1/2}$ with $\boldsymbol{\Delta}_i^* = [\delta_{ijk}^*]_{m_i \times m_i}$, $\delta_{ijk}^* = I(R_{ij} = 1, R_{ik} = 1)/\pi_{ijk}$, $\pi_{ijk} = P(R_{ij} = 1, R_{ik} = 1|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$. Again with the WI structure, it is equivalent to have

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{W}_i^\top\mathbf{M}_i^{*\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{\sigma_{ij}^2}\frac{R_{ij}}{\pi_{ij}}\boldsymbol{W}_{ij}(Y_{ij} - \boldsymbol{W}_{ij}^\top\boldsymbol{\beta}) = 0, \tag{4.4}$$

where $\mathbf{M}_i^{*\mathrm{WI}} = \mathbf{F}_i^{-1/2}\boldsymbol{\Delta}_i^{*\mathrm{WI}}\mathbf{F}_i^{-1/2}$ with $\boldsymbol{\Delta}_i^{*\mathrm{WI}} = \mathrm{diag}(R_{ij}/\pi_{ij})$ and $\pi_{ij} = P(R_{ij} = 1|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$.

Although we have avoided calculating the joint conditional probabilities like π_{ijk} with the WI structure, it is still not straightforward to gain the marginal probabilities π_{ij}. Even under (4.2) and (4.3), one needs to do the integration by summing up all the possible outcomes of $\bar{\mathbf{R}}_{ij}$ (Chen et al. (2010)). For the purpose of illustrating the method more easily, we further assume $P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \bar{\mathbf{X}}_{ij}^o, \mathbf{Y}_i, \mathbf{Z}_i) = P(R_{ij} = r_{ij}|\bar{\mathbf{X}}_{ij}^o, \mathbf{Y}_i, \mathbf{Z}_i)$. Then (4.2) can be rewritten as

$$P(\mathbf{R}_i = \mathbf{r}_i|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \prod_{j=1}^{m_i} P(R_{ij} = r_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = \prod_{j=1}^{m_i}\pi_{ij}. \tag{4.5}$$

Based on this condition, we can model π_{ij} directly. Let $\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\alpha}) = 0$ be the estimating equation from the maximum likelihood of $\mathbf{R}_i$ based on a specified model (e.g., logistic regression model on π_{ij}) with the nuisance parameter $\boldsymbol{\alpha}$, $\mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{W}_i^\top\mathbf{M}_i^{*\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta})$. Then actually we are solving

$$n^{-1/2}\begin{pmatrix}\sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) \\ \sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\alpha})\end{pmatrix} = \mathbf{0} \tag{4.6}$$

simultaneously. Note that generally (4.5) is not required. IPW will lead to unbiased parameter estimators if the model on π_{ij} is correctly specified.

4.2.4 Augmented Inverse-probability Weighted Estimator (AIPW)

Although IPW gives consistent estimators under the correctly specified model, it still does not make use of the information (y_{ij} and $\boldsymbol{Z}_{ij}$) from the incomplete cases ($R_{ij} = 0$). Chen et al. (2010) added an augmentation term $\mathbf{A}_i$ into the equation (4.4) and $\mathbf{A}_i$ is an arbitrary function of

the available measurements that may not be included in (4.4) with mean 0.

Another problem of IPW is that it can still be biased when the model on π_{ij} is misspecified. Robins et al. (1994) and Chen & Zhou (2011) suggested to use

$$n^{-1/2}\sum_{i=1}^{n}\left[\mathbf{W}_i^\top\mathbf{M}_i^*(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\beta})+E_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top\mathbf{N}_i^*(\mathbf{Y}_i-\mathbf{W}_i\boldsymbol{\beta})\}\right]=0, \tag{4.7}$$

where $\mathbf{N}_i^*=\mathbf{F}_i^{-1/2}\{\mathbf{C}_i^{-1}\odot(\mathbf{1}\mathbf{1}^\top-\boldsymbol{\Delta}_i^*)\}\mathbf{F}_i^{-1/2}$, $\mathbf{X}_i^m$ denotes the missing part of $\mathbf{X}_i$. This AIPW has the doubly robust (DR) property. That is, the resulting estimators will be consistent when either the selection probability model on π_{ij} or the missing covariate model on the augmentation (the conditional expectation) is correctly specified. With the WI structure we can simplify it to

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1}{\sigma_{ij}^2}\left[\frac{R_{ij}}{\pi_{ij}}\boldsymbol{W}_{ij}(Y_{ij}-\boldsymbol{W}_{ij}^\top\boldsymbol{\beta})+\left(1-\frac{R_{ij}}{\pi_{ij}}\right)E_{\mathbf{X}_{ij}|\mathbf{Y}_i,\bar{\mathbf{X}}_{ij}^o,\mathbf{Z}_i}\{\boldsymbol{W}_{ij}(Y_{ij}-\boldsymbol{W}_{ij}^\top\boldsymbol{\beta})\}\right]=0.$$

To estimate the conditional expectation, likelihood assumptions or parametric models are specified on the covariate model. Chen & Zhou (2011) expressed the joint density as

$$f(\mathbf{X}_i|\mathbf{Y}_i,\mathbf{Z}_i;\boldsymbol{\gamma})=\prod_{j=2}^{m_i}f(X_{ij}|\bar{\mathbf{X}}_{ij},\mathbf{Y}_i,\mathbf{Z}_i;\boldsymbol{\gamma})\cdot f(X_{i1}|\mathbf{Y}_i,\mathbf{Z}_i;\boldsymbol{\gamma}), \tag{4.8}$$

where $\bar{\mathbf{X}}_{ij}=(X_{i1},...,X_{i,j-1})$ is the history of X_{ij} until time $t_{i,j-1}$ and $\boldsymbol{\gamma}$ is the nuisance parameter for the density. Then models can be specified on $f(X_{ij}|\bar{\mathbf{X}}_{ij},\mathbf{Y}_i,\mathbf{Z}_i;\boldsymbol{\gamma})$. The estimate of $\boldsymbol{\gamma}$ can be obtained through maximizing the observed likelihood function

$$L(\boldsymbol{\gamma})=\prod_{i=1}^{n}\int f(\mathbf{X}_i|\mathbf{Y}_i,\mathbf{Z}_i;\boldsymbol{\gamma})d\mathbf{X}_i^m.$$

Let $\partial\log L(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^\top=\sum_{i=1}^{n}\mathbf{O}_i(\boldsymbol{\gamma})=0$ be the corresponding estimating equation, $\mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma})$ be the component inside the summation in (4.7). Then the whole estimating process can be regarded

as solving

$$n^{-1/2}\begin{pmatrix} \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\gamma}) \\ \sum_{i=1}^{n} \mathbf{S}_i(\boldsymbol{\alpha}) \\ \sum_{i=1}^{n} \mathbf{O}_i(\boldsymbol{\gamma}) \end{pmatrix} = \mathbf{0}. \tag{4.9}$$

4.3 Proposed Method

Let $T_{ij} = \boldsymbol{W}_{ij}(Y_{ij} - \boldsymbol{W}_{ij}^{\top}\boldsymbol{\beta})$ and $\psi_{ij} = E_{\mathbf{X}_{ij}|\mathbf{Y}_i,\bar{\mathbf{X}}_{ij}^{o},\mathbf{Z}_i}(T_{ij})$. It is obvious that T_{ij} is the regular score function for linear regression, ψ_{ij} is the conditional expectation of T_{ij}. Although the AIPW in (4.7) has the DR property, it still can be biased when both the selection probability model and the covariate model are misspecified. Actually it is not easy to model $P(R_{ij} = r_{ij}|\bar{\mathbf{R}}_{ij}, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i)$ and $f(X_{ij}|\bar{\mathbf{X}}_{ij}, \mathbf{Y}_i, \mathbf{Z}_i; \boldsymbol{\gamma})$ correctly. Since the histories $\bar{\mathbf{R}}_{ij}$ and $\bar{\mathbf{X}}_{ij}$ have different length for different time t_{ij}, generally the nuisance parameters can be different as $\boldsymbol{\alpha}^{(j)}$ and $\boldsymbol{\gamma}^{(j)}$. There might be too many parameters at large j for correct model specification. It would need further assumptions to make things simpler. For example, Chen & Zhou (2011) assumed that R_{ij} depends only on the previously ($t_{i,j-1}$) observed outcomes and covariates, and (4.5) is another approach. There are other ways to model the joint densities instead of using the factorization through the hierarchical structure like (4.2) and (4.8), but correctly modeling the correlation structure within $\mathbf{R}_i$ or $\mathbf{X}_i$ is very challenging.

Even when both the selection probability model and the covariate model are correctly specified, IPW and AIPW may encounter numerical problems if some π_{ij}'s are near zero to cause the inverse-probability weights highly variable, resulting in biased estimators. This phenomenon is observed and discussed by Kang & Schafer (2007) and Robins et al. (2007). We will illustrate the phenomenon by numerical examples in Section 4.5.

Wang et al. (1997) and Wang & Wang (2001) proposed to estimate the selection probability and the conditional expectation in the augmentation using nonparametric kernel smoothing for i.i.d. data. Details can be found in Section 3.2 of Chapter 2. To extend the application of this idea to longitudinal data with the WI correlation structure, we first need the following assumption.

Assumption 4.1.

$$E(X_{ij}|\bar{\mathbf{X}}_{ij}^{o}, \mathbf{Y}_i, \mathbf{Z}_i) = E(X_{ij}|\mathbf{Q}_{ij}),$$

where $\mathbf{Q}_{ij} = (Y_{ij}, \boldsymbol{Z}_{ij}^{\top})^{\top}$.

Under this assumption, ψ_{ij} can be written as $\psi_{ij} = E(T_{ij}|\mathbf{Q}_{ij})$. Notice that

$$\begin{aligned}\psi_{ij} = E(T_{ij}|\mathbf{Q}_{ij}) &= E\left\{\boldsymbol{W}_{ij}(Y_{ij} - \boldsymbol{W}_{ij}^{\top}\boldsymbol{\beta})\middle|\mathbf{Q}_{ij}\right\} = E(\boldsymbol{W}_{ij}|\mathbf{Q}_{ij})Y_{ij} - E(\boldsymbol{W}_{ij}\boldsymbol{W}_{ij}^{\top}|\mathbf{Q}_{ij})\boldsymbol{\beta} \\ &= \begin{pmatrix} 1 \\ E(X_{ij}|\mathbf{Q}_{ij}) \\ \boldsymbol{Z}_{ij} \end{pmatrix} Y_{ij} - \begin{pmatrix} 1 & E(X_{ij}|\mathbf{Q}_{ij}) & \boldsymbol{Z}_{ij}^{\top} \\ E(X_{ij}|\mathbf{Q}_{ij}) & E(X_{ij}^2|\mathbf{Q}_{ij}) & E(X_{ij}|\mathbf{Q}_{ij})\boldsymbol{Z}_{ij}^{\top} \\ \boldsymbol{Z}_{ij} & \boldsymbol{Z}_{ij}E(X_{ij}|\mathbf{Q}_{ij}) & \boldsymbol{Z}_{ij}\boldsymbol{Z}_{ij}^{\top} \end{pmatrix} \boldsymbol{\beta}. \quad (4.10)\end{aligned}$$

Thus we only need to model $E(X_{ij}|\mathbf{Q}_{ij})$ and $E(X_{ij}^2|\mathbf{Q}_{ij})$. Let d denote the length of the continuous part in $\mathbf{Q}_{ij}$. Although under Assumption 4.1 the calculation of ψ_{ij} has already been simplified, the regular multivariate kernel smoother for $E(X_{ij}|\mathbf{Q}_{ij})$ and $E(X_{ij}^2|\mathbf{Q}_{ij})$ will suffer from "curse of dimensionality" when $d > 1$. To overcome this difficulty, we make another assumption.

Assumption 4.2. *Assume a single-index model (SIM)*

$$X_{ij} = g(\mathbf{Q}_{ij}^{\top}\boldsymbol{\gamma}) + e_{ij},$$

where g is an unknown smooth univariate function, $\boldsymbol{\gamma}$ is the parameter of the model with the same dimension of $\mathbf{Q}_{ij}$, and e_{ij}'s are random errors with zero mean.

To guarantee identifiability, we assume the first non-zero element of $\boldsymbol{\gamma}$ to be positive 1. If the number of available cases for subject i is $m_i^{(o)} = \sum_{j=1}^{m_i} R_{ij}$, one estimator of $g(\cdot)$ based only on the available cases is

$$\hat{g}(u|\boldsymbol{\gamma}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i^{(o)}} X_{ij}^{(o)} K_h\left(u - \mathbf{Q}_{ij}^{(o)\top}\boldsymbol{\gamma}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i^{(o)}} K_h\left(u - \mathbf{Q}_{ij}^{(o)\top}\boldsymbol{\gamma}\right)} = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}X_{kl}K_h(u - \mathbf{Q}_{kl}^{\top}\boldsymbol{\gamma})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}K_h(u - \mathbf{Q}_{kl}^{\top}\boldsymbol{\gamma})},$$

where $(X_{ij}^{(o)}, \mathbf{Q}_{ij}^{(o)})$ are pairs of the available cases, $K_h(\cdot) = K(\cdot/h)$ is a univariate kernel function with bandwidth h. Then under the SIM condition, we have

$$\hat{E}(X_{ij}|\mathbf{Q}_{ij}) = \hat{E}(X_{ij}|\mathbf{Q}_{ij}^\top\boldsymbol{\gamma}) = \hat{g}(\mathbf{Q}_{ij}^\top\boldsymbol{\gamma}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}X_{kl}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}. \tag{4.11}$$

We can also apply this model to get an estimate of $E(X_{ij}^2|\mathbf{Q}_{ij})$ as

$$\hat{E}(X_{ij}^2|\mathbf{Q}_{ij}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}X_{kl}^2K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}. \tag{4.12}$$

We can construct

$$\hat{\pi}_{ij}^*(\boldsymbol{\gamma}) = \hat{E}(R_{ij}|\mathbf{Q}_{ij}^\top\boldsymbol{\gamma}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k} K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})} \tag{4.13}$$

as the estimated selection probabilities modeled by the SIM using the same $(\boldsymbol{\gamma}, h)$. Notice that (4.11) and (4.13) are essentially NW-estimators with univariate kernel functions and the additional parameter $\boldsymbol{\gamma}$. Compared to Wang & Wang (2001), here we only estimate the first two moments of X_{ij} given $\mathbf{Q}_{ij}$ by using the local average when estimating ψ_{ij} but keep the original $\mathbf{Y}_i$, $\mathbf{Z}_i$ since they are always observed, instead of using the local average of the whole score function T_{ij}.

Let

$$\hat{\psi}_{ij}(\boldsymbol{\gamma}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}T_{ij}^{(kl)}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k} R_{kl}K_h((\mathbf{Q}_{ij}-\mathbf{Q}_{kl})^\top\boldsymbol{\gamma})},$$

where $T_{ij}^{(kl)} = \boldsymbol{W}_{ij}^{(kl)}(Y_{ij} - \boldsymbol{W}_{ij}^{(kl)^\top}\boldsymbol{\beta})$ with $\boldsymbol{W}_{ij}^{(kl)} = (1, X_{kl}, \boldsymbol{Z}_{ij}^\top)^\top$. This $\hat{\psi}_{ij}(\boldsymbol{\gamma})$ is a kernel estimate of ψ_{ij} by estimating only $E(X_{ij}|\mathbf{Q}_{ij})$ and $E(X_{ij}^2|\mathbf{Q}_{ij})$ with a kernel smoother via (4.10). Then we can first propose an Ordinary Least Square (OLS) semiparametric estimator without inverse-

probability weights as

$$\mathbf{U}(\boldsymbol{\beta}, \hat{\psi}(\boldsymbol{\gamma})) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left\{ R_{ij} T_{ij} + (1 - R_{ij}) \hat{\psi}_{ij}(\boldsymbol{\gamma}) \right\} \tag{4.14a}$$

$$= n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{W}_i^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) + \hat{E}_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}}) (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} \right] = 0, \tag{4.14b}$$

where $\hat{E}_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}}) (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \}$ is the estimate of $E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}}) (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \}$ based on $\hat{\psi}_{ij}$. Let $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ denote the solution of the above estimating equation. We will show that $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ is an unbiased estimator in Section 4.4. Clearly, (4.14b) ignores the variance structure σ_{ij}^2. Hence $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ works well for homoscedasticity ($\sigma_{ij}^2 \equiv \sigma^2$), but it is not efficient when heteroscedasticity is present. Recall the assumption that $\sigma_{ij}^2 = \sigma^2(t_{ij})$ when heteroscedasticity is allowed. Let $\hat{\varepsilon}_{ij} = Y_{ij} - \boldsymbol{W}_{ij}^\top \hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ be the residuals after fitting (4.14b). Obviously, some of the residuals cannot be obtained because some X_{ij}'s are missing. Following the idea of Fan et al. (2007), a kernel estimator with bandwidth h_σ for $\sigma^2(t)$ is

$$\hat{\sigma}^2(t) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} R_{ij} \hat{\varepsilon}_{ij}^2 K_{h_\sigma}(t - t_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} R_{ij} K_{h_\sigma}(t - t_{ij})},$$

which only uses the available residuals in the calculation. Then σ_{ij}^2 can be estimated through

$$\hat{\sigma}_{ij}^2 = \hat{\sigma}^2(t_{ij}) = \frac{\sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} \hat{\varepsilon}_{kl}^2 K_{h_\sigma}(t_{ij} - t_{kl})}{\sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} K_{h_\sigma}(t_{ij} - t_{kl})}. \tag{4.15}$$

Based on these estimated variances, a Weighted Least Square (WLS) estimator adjusted for heteroscedasticity is proposed as

$$\mathbf{U}(\boldsymbol{\beta}, \hat{\psi}(\boldsymbol{\gamma}), \hat{\sigma}^2(t)) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{1}{\hat{\sigma}_{ij}^2} \left\{ R_{ij} T_{ij} + (1 - R_{ij}) \hat{\psi}_{ij}(\boldsymbol{\gamma}) \right\} \tag{4.16a}$$

$$= n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{W}_i^\top \hat{\mathbf{M}}_i^{\mathrm{WI}} (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) + \hat{E}_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top \hat{\mathbf{N}}_i^{\mathrm{WI}} (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} \right] = 0, \tag{4.16b}$$

where $\hat{\mathbf{M}}_i^{\mathrm{WI}} = \hat{\mathbf{F}}_i^{-1/2}\boldsymbol{\Delta}_i^{\mathrm{WI}}\hat{\mathbf{F}}_i^{-1/2}$, $\hat{\mathbf{N}}_i^{\mathrm{WI}} = \hat{\mathbf{F}}_i^{-1/2}(\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}})\hat{\mathbf{F}}_i^{-1/2}$ with $\hat{\mathbf{F}}_i = \mathrm{diag}(\hat{\sigma}_{ij}^2)$. Let $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ denote the solution of the above estimating equation.

Generally, like the estimation procedure of AIPW, $\boldsymbol{\gamma}$ is an nuisance parameter which needs to be estimated. However, in our case with the linear relationship between Y_{ij} and $(X_{ij}, \boldsymbol{Z}_{ij}^\top)^\top$, we have a special form of $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma} = (1, -\beta_2^\top)^\top$. In this sense, the single index is $u_{ij} = \mathbf{Q}_{ij}^\top\boldsymbol{\gamma} = Y_{ij} - \boldsymbol{Z}_{ij}^\top\beta_2$ and $\boldsymbol{\gamma}$ is a part of $\boldsymbol{\beta}$ so that we do not need to estimate $\boldsymbol{\gamma}$ separately. Note that while the choice of the kernel functions does not have much influence on the performance of the estimators with a fixed order r, the choice of the bandwidths h, h_σ is crucial. More details about bandwidth selection will be discussed in Section 4.4.

Notice that the estimation procedures of $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ do not involve the marginal selection probabilities π_{ij}'s, which are usually difficult to model and calculate. The SIM structure allows us to use univariate kernel functions during the procedure and avoids suffering from "curse of dimensionality". A more important benefit is that unlike all inverse-probability weighted estimators (such as IPW, AIPW), $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ are not sensitive to those positive but near-zero π_{ij}'s since we do not use them in the point estimation procedure.

On the other hand, due to the construction of the estimating equations (4.14b) and (4.16b) without inverse-probability weights, $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ no longer have the property of double robustness, thus need a consistent estimator of ψ_{ij}. This consistency depends on whether Assumption 4.2 of a single-index model on X_{ij} is reasonable. Because of the linear relationship between Y_{ij} and $(X_{ij}, \boldsymbol{Z}_{ij}^\top)^\top$ in this setting, it seems not unreasonable to assume this model. Actually Assumption 4.2 is valid when $(\mathbf{Y}_i^\top, \mathbf{X}_i^\top, \boldsymbol{Z}_{i1}^\top, ..., \boldsymbol{Z}_{im_i}^\top)^\top$ jointly follows a multivariate normal distribution under the WI correlation structure. The two proposed estimators can still give robust results under other distributions or the assumptions are not exactly satisfied, which is to be illustrated through numerical studies in Section 4.5.

4.4 Asymptotic Properties

In this section, we will show the asymptotic behavior of the proposed estimators $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$. For simplicity, we define $\pi_{ij}^*(\boldsymbol{\gamma}) = E(R_{ij}|\mathbf{Q}_{ij}^\top\boldsymbol{\gamma})$ as the selection probabilities conditional

on the single-index $\mathbf{Q}_{ij}^{\top}\boldsymbol{\gamma}$ with parameter $\boldsymbol{\gamma}$. Let $U = \mathbf{Q}^{\top}\boldsymbol{\gamma}$ denote the single-index random variable. Let $N = \sum_{i=1}^{n} m_i$ be the total number of observations, and $N^{(o)}$ be the total number of available observations. We need the following regularity conditions to establish the asymptotic theory:

(i) The smoothing parameter h satisfies $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$.

(ii) The smoothing parameter h_σ satisfies $h_\sigma \to 0$ and $nh_\sigma \to \infty$, as $n \to \infty$.

(iii) All the selection probabilities π_{ij}'s are bounded away from zero.

(iv) The selection probability function on the single-index $\pi^*(\boldsymbol{\gamma})$ has r continuous and bounded partial derivatives a.e.

(v) The density function $f(u)$ of U and the conditional density function $f_{U|R}(u)$ of $U|R$ have r^{th} continuous and bounded partial derivative a.e.

(vi) The conditional distributions $f_{U|R=0}(u)$ and $f_{U|R=1}(u)$ have the same support, and $b(u) = f_{U|R=0}(u)/f_{U|R=1}(u)$ is bounded over the support.

(vii) The conditional expectations $\psi(u|\boldsymbol{\gamma}) = E(T|\mathbf{Q}^{\top}\boldsymbol{\gamma} = u)$ and $E(TT^{\top}|\mathbf{Q}^{\top}\boldsymbol{\gamma})$ exist and have r continuous and bounded partial derivative a.e.

(viii) For score T, $E(TT^{\top})$ and $E\{(\partial/\partial\boldsymbol{\beta})T\}$ exist and are positive definite, and $(\partial^2/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top})T$ exists and is continuous with respect to $\boldsymbol{\beta}$ a.e.

Recall that r is the order of the kernel function used in the estimation. From regularity condition (i), r is related to the rate of the bandwidth h. Since we are considering a SIM for estimation, a standard 2nd-order ($r = 2$) univariate kernel function seems reasonable in practice.

Let $\eta_n = \{nh^{2r} + (nh^2)^{-1}\}^{1/2}$. The following lemma is important to prove our main theorems.

Lemma 4.1. *Under regularity conditions* (i)-(viii) *and when Assumption 4.1 and 4.2 are true, we have*

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-R_{ij})\{\hat{\psi}_{ij}(\boldsymbol{\gamma})-\psi_{ij}(\boldsymbol{\gamma})\}=n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}R_{ij}\{T_{ij}^0-\psi_{ij}^0(\boldsymbol{\gamma})\}a(\mathbf{Q}_{ij}^\top\boldsymbol{\gamma})+O_p(\eta_n),$$

and

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{1-R_{ij}}{\hat{\sigma}_{ij}^2}\{\hat{\psi}_{ij}(\boldsymbol{\gamma})-\psi_{ij}(\boldsymbol{\gamma})\}=n^{-1/2}\kappa\sum_{i=1}^{n}\sum_{j=1}^{m_i}R_{ij}\{T_{ij}^0-\psi_{ij}^0(\boldsymbol{\gamma})\}a(\mathbf{Q}_{ij}^\top\boldsymbol{\gamma})+O_p(\eta_n),$$

where $a(\mathbf{Q}_{ij}^\top\boldsymbol{\gamma})=\{1-\pi_{ij}^*(\boldsymbol{\gamma})\}/\pi_{ij}^*(\boldsymbol{\gamma})$, $T_{ij}^0=E_{\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0}(T_{ij})=\int T_{ij}f(\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0)d\boldsymbol{Z}_{ij}$, $\psi_{ij}^0(\boldsymbol{\gamma})=E_{\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0}\{\psi_{ij}(\boldsymbol{\gamma})\}=\int\psi_{ij}(\boldsymbol{\gamma})f(\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0)d\boldsymbol{Z}_{ij}$ *with* $u_{ij}=\mathbf{Q}_{ij}^\top\boldsymbol{\gamma}$ *as the single index and* $\kappa=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{1-R_{ij}}{\sigma_{ij}^2}\right)/\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-R_{ij})=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-R_{ij})/\sigma_{ij}^2}{N-N^{(o)}}$.

Proof. The idea in the proof is similar to that in the proof of Lemma 1 in Wang & Wang (2001). Recall that $u_{ij}=\mathbf{Q}_{ij}^\top\boldsymbol{\gamma}=Y_{ij}-\beta_2^\top\boldsymbol{Z}_{ij}$ is the single index. Let

$$\hat{f}_{U|R=1}(u)=\frac{1}{N^{(o)}h}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}K_h(u-u_{kl}),\quad E_n(u)=\hat{f}_{U|R=1}(u)-f_{U|R=1}(u),$$

$$A_{nij}=\hat{f}_{U|R=1}(u_{ij}),\quad B_{nij}=\frac{1}{N^{(o)}h}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}T_{ij}^{(kl)}K_h(u_{ij}-u_{kl}).$$

Under the regularity conditions with the sparse longitudinal data setting, we have $E\{E_n(u)\}=O(h^r)$ and $\mathrm{var}\{E_n(u)\}=O\{(nh)^{-1}\}$ by the Taylor expansions. Then by the Chebyshev inequality, $E_n(u)-E\{E_n(u)\}=O_p\{(nh)^{-1/2}\}$, which implies $E_n(u)=O_p\{h^r+(nh)^{-1/2}\}$, and thus $E_n(u_{ij})=O_p\{h^r+(nh)^{-1/2}\}$. Similarly, we have $B_{nij}-\psi_{ij}A_{nij}=O_p\{h^r+(nh)^{-1/2}\}$.

Define $\delta_n=h^{2r}+(nh)^{-1}$. When Assumption 4.1 and 4.2 are true,

$$\hat{\psi}_{ij}-\psi_{ij}=\frac{B_{nij}-\psi_{ij}A_{nij}}{f_{U|R=1}(u_{ij})}-\frac{(B_{nij}-\psi_{ij}A_{nij})E_n(u_{ij})}{A_{nij}f_{U|R=1}(u_{ij})}=\frac{B_{nij}-\psi_{ij}A_{nij}}{f_{U|R=1}(u_{ij})}+O_p(\delta_n).$$

Let $\mathbf{Q}_{ij}^*=R_{ij}\mathbf{Q}_{ij}$, $X_{ij}^*=R_{ij}X_{ij}$ for $i=1,...,n$, $j=1,...,m_i$ as the values of the available

cases. Then

$$
\begin{aligned}
&E\left\{\frac{B_{nij}-\psi_{ij}A_{nij}}{f_{U|R=1}(u_{ij})}\middle| R_{ij}=0, \text{all } (\mathbf{R},\mathbf{Q}^*,\mathbf{X}^*)\right\}\\
&=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\int\frac{(T_{ij}^{(kl)}-\psi_{ij})K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_{ij})}f_{\mathbf{Q}|R=0}(\mathbf{Q}_{ij})d\mathbf{Q}_{ij}\\
&=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\iint\frac{(T_{ij}^{(kl)}-\psi_{ij})K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_{ij})}f_{U,\boldsymbol{Z}|R=0}(u_{ij},\boldsymbol{Z}_{ij})d\boldsymbol{Z}_{ij}du_{ij}\\
&=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\int\left\{\int\frac{(T_{ij}^{(kl)}-\psi_{ij})K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_{ij})}f_{\boldsymbol{Z}|U,R=0}(\boldsymbol{Z}_{ij})d\boldsymbol{Z}_{ij}\right\}f_{U|R=0}(u_{ij})du_{ij}\\
&=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\int\frac{(T_{ij}^{(kl)0}-\psi_{ij}^0)K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_{ij})}f_{U|R=0}(u_{ij})du_{ij}\\
&=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}(T_{kl}^0-\psi_{kl}^0)b(u_{kl})+O_p(h^r),
\end{aligned}
$$

where $T_{ij}^{(kl)0}=E_{\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0}\left\{T_{ij}^{(kl)}\right\}=\int T_{ij}^{(kl)}f(\boldsymbol{Z}_{ij}|u_{ij},R_{ij}=0)d\boldsymbol{Z}_{ij}$, $b(u)$ is defined in regularity condition (vi). The last step is because of the concentration of u_{ij} on u_{kl}. Using the same idea and $\{\cdots\}$ to denote a repeat of the preceding term, we also have

$$
\begin{aligned}
&\text{var}\left\{\frac{B_{nij}-\psi_{ij}A_{nij}}{f_{U|R=1}(u_{ij})}\middle| R_{ij}=0, \text{all } (\mathbf{R},\mathbf{Q}^*,\mathbf{X}^*)\right\}\\
&=\frac{1}{\{N^{(o)}\}^2}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\left[\int\left\{\frac{(T_{ij}^{(kl)}-\psi_{ij})K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_i)}\right\}\{\cdots\}^\top f_{\mathbf{Q}|R=0}(\mathbf{Q}_{ij})d\mathbf{Q}_{ij}\right.\\
&\left.-\left\{\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}\int\frac{(T_{ij}^{(kl)}-\psi_{ij})K_h(u_{ij}-u_{kl})}{hf_{U|R=1}(u_{ij})}f_{\mathbf{Q}|R=0}(\mathbf{Q}_{ij})d\mathbf{Q}_{ij}\right\}\{\cdots\}^\top\right]+O_p\left(\frac{1}{nh}\right)\\
&=O_p\left(\frac{1}{nh}\right).
\end{aligned}
$$

Thus,

$$
\hat{\psi}_{ij}-\psi_{ij}=\frac{1}{N^{(o)}}\sum_{k=1}^{n}\sum_{l=1}^{m_k}R_{kl}(T_{kl}^0-\psi_{kl}^0)b(u_{kl})+O_p(\delta_n).
$$

Let

$$S_n = n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} (1 - R_{ij}) \left\{ \frac{B_{nij} - \psi_{ij} A_{nij}}{f_{U|R=1}(u_{ij})} - \frac{1}{N^{(o)}} \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) b(u_{kl}) \right\}.$$

Then the summands with $R_{ij} = 0$ in S_n are i.i.d. random variables conditioning on all $(\mathbf{R}, \mathbf{Q}^*, \mathbf{X}^*)$. Thus we have

$$\begin{aligned} \text{var}\{S_n | \text{all } (\mathbf{R}, \mathbf{Q}^*, \mathbf{X}^*)\} &= \frac{N - N^{(o)}}{n} \text{var} \left\{ \frac{B_{n11} - \psi_{11} A_{n11}}{f_{U|R=1}(u_{11})} \middle| \text{ all } (\mathbf{R}, \mathbf{Q}^*, \mathbf{X}^*) \right\} \\ &= O_p \left(h^{2r} + \frac{1}{nh} \right). \end{aligned}$$

Then $E(S_n) = O(h^r)$ and $\text{var}(S_n) = O(h^{2r} + (nh)^{-1})$ imply $S_n = O_p(\eta_n)$. Hence we have

$$\begin{aligned} & n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} (1 - R_{ij})(\hat{\psi}_{ij} - \psi_{ij}) \\ &= n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \left\{ (1 - R_{ij}) \frac{1}{N^{(o)}} \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) b(u_{kl}) \right\} + O_p(\eta_n) \\ &= n^{-1/2} \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) a(u_{kl}) + O_p(\eta_n). \end{aligned}$$

The proof of the second part is similar. We still have the same conclusions on the conditional expectation and variance of $\hat{\psi}_{ij} - \psi_{ij}$. Then let

$$S_n^w = n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \frac{1 - R_{ij}}{\sigma_{ij}^2} \left\{ \frac{B_{nij} - \psi_{ij} A_{nij}}{f_{U|R=1}(u_{ij})} - \frac{1}{N^{(o)}} \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) b(u_{kl}) \right\}.$$

We have

$$\begin{aligned} \text{var}\{S_n^w | \text{all } (\mathbf{R}, \mathbf{Q}^*, \mathbf{X}^*)\} &= \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (1 - R_{ij}) / \sigma_{ij}^4}{n} \text{var} \left\{ \frac{B_{n11} - \psi_{11} A_{n11}}{f_{U|R=1}(u_{11})} \middle| \text{ all } (\mathbf{R}, \mathbf{Q}^*, \mathbf{X}^*) \right\} \\ &= O_p \left(h^{2r} + \frac{1}{nh} \right). \end{aligned}$$

Again we have $S_n^w = O_p(\eta_n)$. Let $\delta_n^\sigma = h_\sigma^{2r} + (nh_\sigma)^{-1}$. Since $\hat{\sigma}_{ij}^2 - \sigma_{ij}^2 = O_p(\delta_n^\sigma)$, finally

$$
\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \frac{1 - R_{ij}}{\hat{\sigma}_{ij}^2} (\hat{\psi}_{ij} - \psi_{ij}) \\
&= n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \frac{1 - R_{ij}}{\hat{\sigma}_{ij}^2} (\hat{\psi}_{ij} - \psi_{ij}) - n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \frac{\hat{\sigma}_{ij}^2 - \sigma_{ij}^2}{\hat{\sigma}_{ij}^2 \sigma_{ij}^2} (1 - R_{ij}) (\hat{\psi}_{ij} - \psi_{ij}) \\
&= n^{-1/2} \sum_{i=1}^{n} \sum_{l=1}^{m_k} \left\{ \frac{1 - R_{ij}}{\sigma_{ij}^2} \frac{1}{N^{(o)}} \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) b(u_{kl}) \right\} + O_p(\eta_n) \\
&= n^{-1/2} \kappa \sum_{k=1}^{n} \sum_{l=1}^{m_k} R_{kl} (T_{kl}^0 - \psi_{kl}^0) a(u_{kl}) + O_p(\eta_n).
\end{aligned}
$$

□

Note that with the SIM and the single index $u_{ij} = Y_{ij} - \beta_2^\top \boldsymbol{Z}_{ij}$, we can write T_{ij} and $\psi_{ij}(\boldsymbol{\gamma})$ as

$$
T_{ij} = \begin{pmatrix} u_{ij} - \beta_0 - \beta_1 X_{ij} \\ u_{ij} X_{ij} - \beta_0 X_{ij} - \beta_1 X_{ij}^2 \\ \boldsymbol{Z}_{ij} (u_{ij} - \beta_0 - \beta_1 X_{ij}) \end{pmatrix},
$$

$$
\psi_{ij}(\boldsymbol{\gamma}) = \begin{pmatrix} u_{ij} - \beta_0 - \beta_1 E(X_{ij} | u_{ij}) \\ u_{ij} E(X_{ij} | u_{ij}) - \beta_0 E(X_{ij} | u_{ij}) - \beta_1 E(X_{ij}^2 | u_{ij}) \\ \boldsymbol{Z}_{ij} \{ u_{ij} - \beta_0 - \beta_1 E(X_{ij} | u_{ij}) \} \end{pmatrix}.
$$

Since MAR and Assumption 4.1 imply $(X_{ij} \perp R_{ij}) | \mathbf{Q}_{ij}$, we also have

$$
T_{ij}^0 = \begin{pmatrix} u_{ij} - \beta_0 - \beta_1 X_{ij} \\ u_{ij} X_{ij} - \beta_0 X_{ij} - \beta_1 X_{ij}^2 \\ \boldsymbol{Z}_{ij}^{u|0} (u_{ij} - \beta_0 - \beta_1 X_{ij}) \end{pmatrix},
$$

$$
\psi_{ij}^0(\boldsymbol{\gamma}) = \begin{pmatrix} u_{ij} - \beta_0 - \beta_1 E(X_{ij} | u_{ij}) \\ u_{ij} E(X_{ij} | u_{ij}) - \beta_0 E(X_{ij} | u_{ij}) - \beta_1 E(X_{ij}^2 | u_{ij}) \\ \boldsymbol{Z}_{ij}^{u|0} \{ u_{ij} - \beta_0 - \beta_1 E(X_{ij} | u_{ij}) \} \end{pmatrix}
$$

with $\boldsymbol{Z}_{ij}^{u|0} = E(\boldsymbol{Z}_{ij}|u_{ij}, R_{ij} = 0)$.

Lemma 4.1 is useful because it converts asymptotically a sum of dependent random variables to a sum of independent and i.i.d. random variables. Then it is easier to derive the following theorems by applying standard asymptotic theory.

Define

$$\begin{aligned}\mathbf{L}_i^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma})) = \mathbf{W}_i^\top \boldsymbol{\Delta}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta}) + E_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top(\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}})(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta})\} \\ + \mathbf{W}_i^{0\top}\boldsymbol{\Delta}_i^{\mathrm{WI}}\mathbf{a}(\boldsymbol{u}_i)(\mathbf{Y}_i - \mathbf{W}_i^0\boldsymbol{\beta}) - E_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^{0\top}\boldsymbol{\Delta}_i^{\mathrm{WI}}\mathbf{a}(\boldsymbol{u}_i)(\mathbf{Y}_i - \mathbf{W}_i^0\boldsymbol{\beta})\},\end{aligned}$$

where $\mathbf{W}_i^0 = (\boldsymbol{W}_{i1}^0, ..., \boldsymbol{W}_{im_i}^0)^\top$ with $\boldsymbol{W}_{ij}^0 = \left(1, X_{ij}, \boldsymbol{Z}_{ij}^{u|0\top}\right)^\top$, $\mathbf{a}(\boldsymbol{u}_i) = \mathrm{diag}\{a(u_{ij})\}$.

Theorem 4.1. *Under the regularity conditions* (i)-(viii) *and assuming that Assumptions 4.1 and 4.2 are true,* $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ *is asymptotically equivalent to the solution of the following estimating equation:*

$$n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_i^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma})) = 0.$$

Furthermore, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Sigma}_{\mathbf{o}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{o}} = {\mathbf{G}_{\mathbf{o}}}^{-1}\boldsymbol{\Omega}_{\mathbf{o}}{\mathbf{G}_{\mathbf{o}}}^{-1}$ *with* $\mathbf{G}_{\mathbf{o}} = -n^{-1}E\{\partial\mathbf{U}(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))/\partial\boldsymbol{\beta}^\top\} = E(\mathbf{W}_1\mathbf{W}_1^\top)$ *and* $\boldsymbol{\Omega}_{\mathbf{o}} = \mathrm{cov}(\mathbf{L}_1^o) = E\{\mathbf{L}_1^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))\mathbf{L}_1^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))^\top\}$.

Proof. Based on the conclusion of Lemma 4.1,

$$
\begin{aligned}
&\mathbf{U}(\boldsymbol{\beta}, \hat{\psi}(\boldsymbol{\gamma})) \\
&= n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Big[R_{ij} T_{ij} + (1 - R_{ij}) \psi_{ij}(\boldsymbol{\gamma}) + (1 - R_{ij}) \{ \hat{\psi}_{ij}(\boldsymbol{\gamma}) - \psi_{ij}(\boldsymbol{\gamma}) \} \Big] \\
&= n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[R_{ij} T_{ij} + (1 - R_{ij}) \psi_{ij}(\boldsymbol{\gamma}) + R_{ij} \{ T_{ij}^0 - \psi_{ij}^0(\boldsymbol{\gamma}) \} a(\mathbf{Q}_{ij}^\top \boldsymbol{\gamma}) \right] + O_p(\eta_n) \\
&= n^{-1/2} \sum_{i=1}^{n} \Bigg[\mathbf{W}_i^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) + E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}}) (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} \\
&\quad + \mathbf{W}_i^{0\top} \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i) (\mathbf{Y}_i - \mathbf{W}_i^0 \boldsymbol{\beta}) - E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^{0\top} \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i) (\mathbf{Y}_i - \mathbf{W}_i^0 \boldsymbol{\beta}) \} \Bigg] + O_p(\eta_n) \\
&= n^{-1/2} \sum_{i=1}^{n} \mathbf{L}_i^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma})) + O_p(\eta_n).
\end{aligned}
$$

It is obvious that

$$
E \left[\mathbf{W}_i^{0\top} \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i) (\mathbf{Y}_i - \mathbf{W}_i^0 \boldsymbol{\beta}) - E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^{0\top} \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i) (\mathbf{Y}_i - \mathbf{W}_i^0 \boldsymbol{\beta}) \} \right] = 0.
$$

Thus we have

$$
\begin{aligned}
&E\{\mathbf{U}_i(\boldsymbol{\beta}, \hat{\psi}(\boldsymbol{\gamma}))\} \\
&= E\{\mathbf{U}_i(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))\} + O_p(\eta_n) \\
&= E \left[E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} + E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{I}_{m_i} - \boldsymbol{\Delta}_i^{\mathrm{WI}}) (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} \right] \\
&\quad + O_p(\eta_n) \\
&= E \left[E_{\mathbf{X}_i^m | \mathbf{Y}_i, \mathbf{X}_i^o, \mathbf{Z}_i} \{ \mathbf{W}_i^\top (\mathbf{Y}_i - \mathbf{W}_i \boldsymbol{\beta}) \} \right] + O_p(\eta_n) \\
&= O_p(\eta_n).
\end{aligned}
$$

Hence $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ is asymptotically unbiased. Since $\mathbf{U}(\boldsymbol{\beta}, \hat{\psi}(\boldsymbol{\gamma}))$ is asymptotically equivalent to a sum of i.i.d. random variables $\mathbf{L}_i^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))$, $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ is asymptotically normally distributed and has the

asymptotic covariance $\mathbf{\Sigma_o} = \mathbf{G_o}^{-1}\mathbf{\Omega_o}\mathbf{G_o}^{-1}$ with

$$\mathbf{G_o} = -n^{-1}E\{\partial \mathbf{U}(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))/\partial \boldsymbol{\beta}^\top\} = E(\mathbf{W}_1\mathbf{W}_1^\top),$$

$$\mathbf{\Sigma_o} = \mathrm{cov}\left(n^{-1/2}\sum_{i=1}^{n}\mathbf{L}_i^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))\right) = \mathrm{cov}(\mathbf{L}_1^o) = E\{\mathbf{L}_1^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))\mathbf{L}_1^o(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}))^\top\}.$$

□

The $\mathbf{\Sigma_o}$ can be estimated by

$$\hat{\mathbf{\Sigma}}_o = \hat{\mathbf{G}}_o^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{L}}_i^o(\hat{\mathbf{L}}_i^o)^\top\right\}\hat{\mathbf{G}}_o^{-1}, \tag{4.17}$$

where

$$\begin{aligned}
\hat{\mathbf{L}}_i^o &= \mathbf{W}_i^\top\mathbf{\Delta}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\hat{\boldsymbol{\beta}}) + \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top(\mathbf{I}_{m_i} - \mathbf{\Delta}_i^{\mathrm{WI}})(\mathbf{Y}_i - \mathbf{W}_i\hat{\boldsymbol{\beta}})\} \\
&\quad + (\hat{\mathbf{W}}_i^0)^\top\mathbf{\Delta}_i^{\mathrm{WI}}\mathbf{a}(\hat{\boldsymbol{u}}_i)(\mathbf{Y}_i - \hat{\mathbf{W}}_i^0\hat{\boldsymbol{\beta}}) - \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{(\hat{\mathbf{W}}_i^0)^\top\mathbf{\Delta}_i^{\mathrm{WI}}\mathbf{a}(\hat{\boldsymbol{u}}_i)(\mathbf{Y}_i - \hat{\mathbf{W}}_i^0\hat{\boldsymbol{\beta}})\} \\
&= \sum_{j=1}^{m_i}\left[R_{ij}T_{ij}(\hat{\boldsymbol{\beta}}) + (1 - R_{ij})\hat{\psi}_{ij}(\hat{\boldsymbol{\gamma}}) + R_{ij}\left\{\hat{T}_{ij}^0(\hat{\boldsymbol{\beta}}) - \hat{\psi}_{ij}^0(\hat{\boldsymbol{\gamma}})\right\}a(\hat{u}_{ij})\right]
\end{aligned}$$

with $\hat{\mathbf{W}}_i^0 = (\hat{\boldsymbol{W}}_{i1}^0, ..., \hat{\boldsymbol{W}}_{im_i}^0)^\top$, $\hat{\boldsymbol{W}}_{ij}^0 = \left(1, X_{ij}, (\hat{\boldsymbol{Z}}_{ij}^{u|0})^\top\right)^\top$, $\mathbf{a}(\hat{\boldsymbol{u}}_i) = \mathrm{diag}\{\mathbf{Q}_{ij}^\top\hat{\boldsymbol{\gamma}}\}$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and

$$\hat{\mathbf{G}}_o = n^{-1}\sum_{i=1}^{n}\left[\mathbf{W}_i^\top\mathbf{\Delta}_i^{\mathrm{WI}}\mathbf{W}_i + \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top(\mathbf{I}_{m_i} - \mathbf{\Delta}_i^{\mathrm{WI}})\mathbf{W}_i\}\right].$$

Here $\boldsymbol{Z}_{ij}^{u|0}$ can be estimated by

$$\hat{\boldsymbol{Z}}_{ij}^{u|0} = \hat{E}(\boldsymbol{Z}_{ij}|\hat{u}_{ij}, R_i = 0) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{m_k}(1 - R_{kl})\boldsymbol{Z}_{kl}K_h((\mathbf{Q}_{ij} - \mathbf{Q}_{kl})^\top\hat{\boldsymbol{\gamma}})}{\sum_{k=1}^{n}\sum_{l=1}^{m_k}(1 - R_{kl})K_h((\mathbf{Q}_{ij} - \mathbf{Q}_{kl})^\top\hat{\boldsymbol{\gamma}})}.$$

Note that to get $\hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top(\mathbf{I}_{m_i} - \mathbf{\Delta}_i^{\mathrm{WI}})\mathbf{W}_i\}$, we only need to calculate $\hat{E}(X_{ij}|\mathbf{Q}_{ij})$ and $\hat{E}(X_{ij}^2|\mathbf{Q}_{ij})$ through (4.11) and (4.12) because of the structure in (4.10).

Similarly, define

$$\begin{aligned}
&\mathbf{L}_i^w(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}), \sigma^2(t)) \\
&= \mathbf{W}_i^\top \mathbf{M}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta}) + E_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top \mathbf{N}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\boldsymbol{\beta})\} \\
&+ \kappa {\mathbf{W}_i^0}^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i)(\mathbf{Y}_i - \mathbf{W}_i^0\boldsymbol{\beta}) - E_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\kappa {\mathbf{W}_i^0}^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\boldsymbol{u}_i)(\mathbf{Y}_i - \mathbf{W}_i^0\boldsymbol{\beta})\}.
\end{aligned}$$

Theorem 4.2. *Under the regularity conditions* (i)-(viii) *and assuming that Assumptions 4.1 and 4.2 are true, $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ is asymptotically equivalent to the solution of the following estimating equation:*

$$n^{-1/2} \sum_{i=1}^{n} \mathbf{L}_i^w(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}), \sigma^2(t)) = 0.$$

Furthermore, we have

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(0, \boldsymbol{\Sigma}_{\mathbf{w}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{w}} = {\mathbf{G}_{\mathbf{w}}}^{-1} \boldsymbol{\Omega}_{\mathbf{w}} {\mathbf{G}_{\mathbf{w}}}^{-1}$ with $\mathbf{G}_{\mathbf{w}} = -n^{-1}E\{\partial \mathbf{U}(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}), \sigma^2(t))/\partial \boldsymbol{\beta}^\top\} = E(\mathbf{W}_1 \mathbf{F}_1^{-1} \mathbf{W}_1^\top)$ and $\boldsymbol{\Omega}_{\mathbf{w}} = \mathrm{cov}(\mathbf{L}_1^w) = E\{\mathbf{L}_1^w(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}), \sigma^2(t))\mathbf{L}_1^w(\boldsymbol{\beta}, \psi(\boldsymbol{\gamma}), \sigma^2(t))^\top\}$.

The proof is analogous to the proof of Theorem 4.1. $\boldsymbol{\Sigma}_{\mathbf{w}}$ can be estimated by

$$\hat{\boldsymbol{\Sigma}}_w = \hat{\mathbf{G}}_w^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{L}}_i^w (\hat{\mathbf{L}}_i^w)^\top \right\} \hat{\mathbf{G}}_w^{-1}, \tag{4.18}$$

where

$$\begin{aligned}
\hat{\mathbf{L}}_i^w &= \mathbf{W}_i^\top \hat{\mathbf{M}}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\hat{\boldsymbol{\beta}}) + \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top \hat{\mathbf{N}}_i^{\mathrm{WI}}(\mathbf{Y}_i - \mathbf{W}_i\hat{\boldsymbol{\beta}})\} \\
&+ \hat{\kappa}(\hat{\mathbf{W}}_i^0)^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\hat{\boldsymbol{u}}_i)(\mathbf{Y}_i - \hat{\mathbf{W}}_i^0\hat{\boldsymbol{\beta}}) - \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\hat{\kappa}(\hat{\mathbf{W}}_i^0)^\top \boldsymbol{\Delta}_i^{\mathrm{WI}} \mathbf{a}(\hat{\boldsymbol{u}}_i)(\mathbf{Y}_i - \hat{\mathbf{W}}_i^0\hat{\boldsymbol{\beta}})\} \\
&= \sum_{j=1}^{m_i} \left[\frac{R_{ij}}{\hat{\sigma}_{ij}^2} T_{ij}(\hat{\boldsymbol{\beta}}) + \frac{(1 - R_{ij})}{\hat{\sigma}_{ij}^2} \hat{\psi}_{ij}(\hat{\boldsymbol{\gamma}}) + \hat{\kappa} R_{ij} \left\{ \hat{T}_{ij}^0(\hat{\boldsymbol{\beta}}) - \hat{\psi}_{ij}^0(\hat{\boldsymbol{\gamma}}) \right\} a(\hat{u}_{ij}) \right]
\end{aligned}$$

with $\hat{\kappa} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(\frac{1-R_{ij}}{\hat{\sigma}_{ij}^2}\right) / \sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-R_{ij})$, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ and

$$\hat{\mathbf{G}}_w = n^{-1}\sum_{i=1}^{n}\left[\mathbf{W}_i^\top \hat{\mathbf{M}}_i^{\mathrm{WI}}\mathbf{W}_i + \hat{E}_{\mathbf{X}_i^m|\mathbf{Y}_i,\mathbf{X}_i^o,\mathbf{Z}_i}\{\mathbf{W}_i^\top \hat{\mathbf{N}}_i^{\mathrm{WI}}\mathbf{W}_i\}\right].$$

Note that Theorem 4.2 reduces to Theorem 4.1 when the random errors are homoscedastic ($\sigma^2(t) \equiv \sigma^2$).

The performance of the estimators AOLS and AWLS depend on the choice of the bandwidth h used in the kernel function $K_h(\cdot)$ for the estimation of ψ_{ij}. In the regularity conditions, we require $nh^2 \to \infty$ and $nh^{2r} \to 0$, as $n \to \infty$. Therefore, the classical optimal rate of the bandwidth $O(n^{-1/5})$ does not work in our situation, as indicated by Sepanski et al. (1994). A reasonable choice is $h = CN^{-1/3}$ for some constant C. A plug-in method can be applied to estimate C. For simplicity, we can use $C = \hat{\sigma}_u$ as suggested by Wang et al. (1997) and Zhou et al. (2008), where $\hat{\sigma}_u$ is the sample standard deviation of the single index u_{ij}. On the other hand, the classical optimal rate of the bandwidth $O(n^{-1/5})$ still works for the kernel smoother (4.15) of $\sigma^2(t)$. We can also use a plug-in bandwidth as $h_\sigma = \hat{\sigma}_t N^{-1/5}$, where $\hat{\sigma}_t$ is the sample standard deviation of the observation times t_{ij}. We use these formulas to choose the bandwidths h and h_σ in our following empirical studies.

4.5 Empirical Studies

In this section, we investigate the performance of our proposed estimators $\hat{\boldsymbol{\beta}}_{\mathrm{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{AWLS}}$ for finite samples, compared to other commonly used GEE-based methods. In the simulation studies setting, the true regression parameter is of dimension 4 as $\boldsymbol{\beta} = (0, 0.5, 1, -1)^\top$. $\boldsymbol{Z}_{ij}$ is of length 2 as $\boldsymbol{Z}_{ij} = (Z_{1ij}, Z_{2ij})^\top$ and are independently generated from $N_2(\mathbf{0}, \mathbf{I}_2)$. $\mathbf{X}_i$ is jointly generated from a multivariate normal or gamma distribution with mean 0, variances 1 and an exchangeable correlation structure ($\rho_x = 0.6$). Marginally the gamma distribution is $(Gamma(5,1) - 5)/\sqrt{5}$. While X_{i1} is always available, X_{ij} ($j > 2$) might be missing according to the following missing process:

$$\mathrm{logit}(\pi_{ij}) = \alpha_0 + \alpha_1 R_{i,j-1}X_{i,j-1} + \alpha_2 Y_{ij} + \alpha_3 Z_{1ij} + \alpha_4 Z_{2ij},$$

which satisfies the MAR mechanism. There are 3 sets of values for the true parameter $\boldsymbol{\alpha}$, making the covariate 18%, 20% and 40% missing on average respectively. The time of measuring t_{ij} is generated in a similar way as Fan et al. (2007): each subject has a set of "scheduled" time points, $\{0, 1, 2, ..., 5\}$, and each scheduled time except time 0 has a 20% chance of being skipped. Then the selected scheduled times plus a uniform $[0, 1]$ random variable makes t_{ij}. Thus we produce unbalanced longitudinal data and on average the number of observations for each subject $\bar{m} = \sum_{i=1}^{n} m_i/n$ is about 5.

The random errors are generated as $\boldsymbol{\varepsilon}_i = \mathrm{chol}(\mathbf{V}_i)\boldsymbol{\varepsilon}_i^0$, where $\mathrm{chol}(\mathbf{V}_i)$ is the lower triangle matrix from Cholesky decomposition of the covariance matrix $\mathbf{V}_i$, and $\boldsymbol{\varepsilon}_i^0$ are independently generated from the standard normal or standardized t distribution ($t_5/\sqrt{5/3}$). Then $\mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \mathbf{V}_i$. We know $\mathbf{V}_i = \mathbf{F}_i^{1/2}\mathbf{C}_i(\boldsymbol{\rho})\mathbf{F}_i^{1/2}$ with $\mathbf{F}_i = \mathrm{diag}(\sigma_{ij}^2)$. There are mainly two scenarios for $\mathbf{V}_i$:

1. The within-subject correlation structure is compound symmetry with $\rho = 0.6$. That is, for $\boldsymbol{\varepsilon}(t_{ij}) = \varepsilon_{ij}$, $\mathrm{Cor}(\boldsymbol{\varepsilon}(t_1), \boldsymbol{\varepsilon}(t_2)) = \rho$ for $t_1 \neq t_2$ and $\sigma_{ij}^2 \equiv 1$.

2. The within-subject correlation structure is AR(1) with $\rho = 0.6$. That is, for $\boldsymbol{\varepsilon}(t_{ij}) = \varepsilon_{ij}$, $\mathrm{Cor}(\boldsymbol{\varepsilon}(t_1), \boldsymbol{\varepsilon}(t_2)) = \rho^{|t_1 - t_2|}$ for $t_1 \neq t_2$. The variances are heteroskedastic through $\mathrm{var}(\varepsilon_{ij}) = \sigma^2(t_{ij}) = 0.25\exp(t/6)$.

We have two types of data: normal data with both $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}_i$ from multivariate normal distributions; and non-normal data with $\mathbf{X}_i$ from a gamma distribution and $\boldsymbol{\varepsilon}_i$ from a t distribution as described above. For IPW and AIPW, we fit a logistic regression model for R_{ij} when $j > 2$ on $\boldsymbol{H}_{ij} = (1, R_{i,j-1}X_{i,j-1}, Y_{ij}, \boldsymbol{Z}_{ij}^\top)^\top$. Thus the selection probability model is correctly specified, which should guarantee unbiased estimators for IPW/AIPW theoretically. Then the estimating equations mentioned in (4.6) have $\mathbf{S}_i(\boldsymbol{\alpha}) = \mathbf{H}_i^\top(\mathbf{R}_i - \boldsymbol{\pi}_i(\boldsymbol{\alpha}))$, where $\mathbf{H}_i = (\mathbf{0}, \boldsymbol{H}_{i2}, ..., \boldsymbol{H}_{im_i})^\top$, $\boldsymbol{\pi}_i(\boldsymbol{\alpha}) = (1, \pi_{i2}, ..., \pi_{im_i})$. For the augmentation in AIPW, we fit a linear regression model for X_{ij} on $\boldsymbol{Q}_{ij} = (1, Y_{ij}, \boldsymbol{Z}_{ij}^\top)^\top$. This model specification is correct at least for the normal data. We will have $\mathbf{O}_i(\boldsymbol{\gamma}) = \mathbf{Q}_i^\top(\mathbf{X}_i - \mathbf{Q}_i\boldsymbol{\gamma})$, where $\mathbf{Q}_i = (\boldsymbol{Q}_{i1}, ..., \boldsymbol{Q}_{im_i})^\top$ in (4.9). Chen et al. (2010), Chen & Zhou (2011) and Robins et al. (1995) all give asymptotic theories for IPW/AIPW based on the

estimating equations (4.6) and (4.9). We will use the sandwich formulas from these papers to gain the asymptotic covariance in the simulation studies without detailed explanation.

For $\hat{\boldsymbol{\beta}}_{\text{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{AWLS}}$, we choose to use a second-order Gaussian kernel function ($r = 2$). The bandwidth selection has been discussed in the previous section. In this simulation study, we use $h = 0.7\hat{\sigma}_u N^{-1/3}$ and $h_\sigma = \hat{\sigma}_t N^{-1/5}$. Assumptions 4.1 and 4.2 are valid at least for the normal data scenario. The asymptotic standard errors are obtained through (4.17) and (4.18). Our empirical experience suggests that, since we only use the information in incomplete cases when estimating $E(\boldsymbol{Z}_{ij}|u_{ij}, R_{ij} = 0)$, it would be helpful to include a correction factor matrix in the sandwich-formulas (4.17) and (4.18) for small to moderate sample sizes, such as those in our simulation studies, especially when the percentage of missingness is high and the data are believed to be skewed. For example, we may replace the estimated asymptotic covariance by $\hat{\boldsymbol{\Sigma}}^*_{\text{AOLS}}(\hat{\boldsymbol{\Sigma}}^*_{\text{AWLS}}) = \boldsymbol{F}_c \cdot \hat{\boldsymbol{\Sigma}}_{\text{AOLS}}(\hat{\boldsymbol{\Sigma}}_{\text{AWLS}})$, where $\boldsymbol{F}_c = \text{diag}\{a, ..., a, b, a, ..., a\}^{-1}$, $a = 1 - 0.8 \times$ miss%, $b = (1 - \text{miss}\%) \cdot \min(\exp\{(n - 1500)/5000\}, 1)$, and miss% means the percentage of missingness of X in the data set. The position of b matches the position of the coefficient of the missing covariate. This is what we used for $\hat{\boldsymbol{\beta}}_{\text{AOLS}}$ and $\hat{\boldsymbol{\beta}}_{\text{AWLS}}$ in our numerical results.

Table 4.1 displays the results of normal data with stable selection probabilities. For each estimator, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error over 1000 replications, and the third line is the 95% coverage probability. We run simulations for both of the covariance structures mentioned above. The example boxplot of π_{ij} for one simulation run can be found in Figure 4.1. As we can see, there are no near-zero values that may cause the highly variable weights discussed in Section 4.3. As expected, CC and AC produce biased estimates because of the MAR mechanism. IPW and AIPW are unbiased, which is reasonable because the selection probability model and covariate model are both correctly specified in this situation. AOLS and AWLS also give consistent results, and are as efficient as AIPW based on similar standard errors. When heteroskedasticity is present, AWLS has slightly smaller standard errors than AOLS.

The other two normal data cases with 20% and 40% missing rate can be found in Tables 4.2

and 4.4. As the boxplots show in Figure 4.1, unlike the simulation setting for Table 4.1, these two cases have very unstable π_{ij}'s. A few of the selection probabilities are near zero, which makes the corresponding inverse-probability weights very large and several terms dominate the summations in (4.4) and (4.7). This numerical issue leads to highly skewed distributions for IPW and AIPW, resulting in large bias, standard errors and deviation from the asymptotic normal distributions though the parametric models are still correctly specified as in Table 4.1. The near-zero selection probabilities also have influence on the sandwich formula of the asymptotic covariances of IPW and AIPW, making the averaged asymptotic standard error very different from the empirical standard error and resulting in low coverages. To make the results comparable, we used 1% trimmed empirical SE for AIPW and put a "*" when the asymptotic SE is ridiculously large. On the other hand, the proposed estimators AOLS and AWLS perform well on bias and standard error. They may have small bias when the missing rate is high at 40%, but are still much better than IPW and AIPW. The asymptotic standard errors are generally close to the empirical standard errors. The 95% coverage probabilities of AOLS and AWLS are also reasonable.

The results of non-normal data are shown in Table 4.3 and 4.5. We keep the same parameters of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ as in Tables 4.2 and 4.4 respectively except the generating distributions for $\mathbf{X}_i$ and $\boldsymbol{\varepsilon}_i$. In this setting, the parametric model for selection probability is still valid, but the parametric model for covariate and the SIM for X_{ij} (Assumption 4.2) are not. However, we get similar conclusions as for the normal data. Actually we can use the histograms of the 1000 replications for different estimators in Figure 4.2 as an example to explain the results. Although IPW and AIPW should give unbiased estimation due to the correctly specified selection probability model and DR property, the numerical issues caused by near-zero π_{ij}'s makes the estimators highly skewed distributed, reflected by the first two rows of histograms in Figure 4.2. But AOLS and AWLS still have normal-shape sampling distributions with the peaks located near the true values of $\boldsymbol{\beta}$ although the coverages are little bit low for β_1.

These results illustrate the numerical problems in IPW and AIPW that are mainly caused by the positive but near-zero selection probabilities. The proposed AOLS and AWLS are not sensitive

to the near-zero π_{ij}'s and robust to the misspecification on the single-index model. They do not require modeling the selection probabilities and likelihood assumptions on $f(\mathbf{X}|\mathbf{Y}_i, \mathbf{Z}_i)$, and hence have a simpler point estimation procedure.

To be mentioned, although the simulations only have a continuous missing covariate X_{ij}, our methodology can also be applied to categorical random variables. The theory is still true as long as the SIM $E(X_{ij}|\mathbf{Q}_{ij}) = g(\mathbf{Q}_{ij}^\top \boldsymbol{\gamma})$ and Assumption 4.1 are valid. When X_{ij} is binary, the estimation procedure can even be simpler because $E(X_{ij}^2|\mathbf{Q}_{ij}) = E(X_{ij}|\mathbf{Q}_{ij})$.

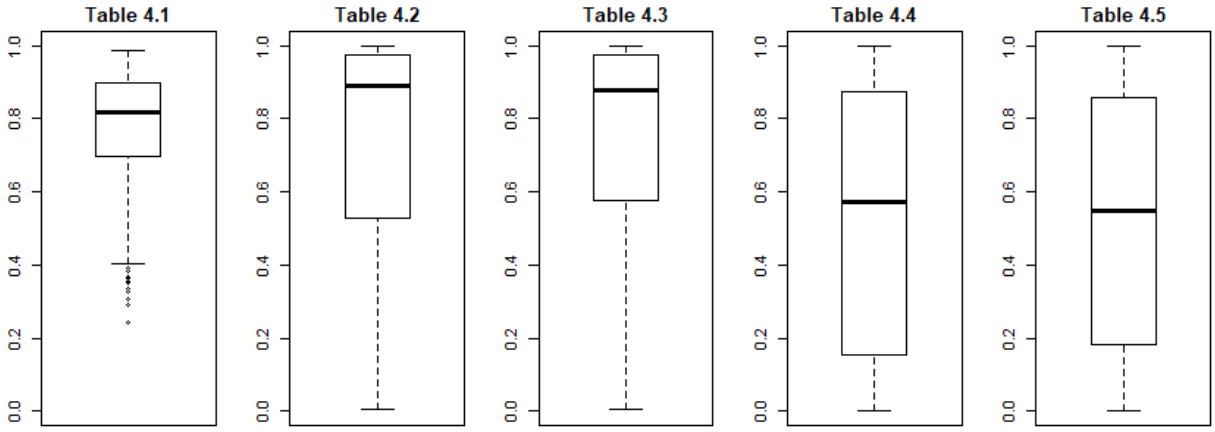

Figure 4.1: Boxplots of the true selection probabilities π_{ij}'s for each table with compound symmetry correlation structure at one simulation run.

4.6 Real Data Examples

In this section, we apply our proposed method to the data from a double-blinded randomized trial in primary biliary cirrhosis of the liver (PBC) for comparing the drug D-penicillamine (DPCA) with a placebo, conducted by the Mayo Clinic between January, 1974 and May, 1984. PBC is a rare but fatal liver disease with a prevalence of about 50-cases-per-million population. There were 312 patients involved in this trial with information gathered routinely during the follow-up. More details about the trial can be found in Fleming & Harrington (2011) and the data are available in

Table 4.1: Simulation results of 1000 replications for the normal data ($n = 100$, $\bar{m} = 5$), under two different correlation structures, with homoskedastic/heteroskedastic errors, and $\alpha = (1.5, -0.5, -0.5, 0, 0)$, about 18% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$
Compound Symmtry ($\rho = 0.6$) $\sigma_{ij}^2 \equiv \sigma^2$	CC	-0.2437 0.0042/0.0040 0.512	-0.0288 0.0025/0.0025 0.919	-0.0201 0.0024/0.0022 0.914	0.0214 0.0023/0.0022 0.926
	AC	-0.0762 0.0027/0.0026 0.843	-0.0128 0.0017/0.0017 0.934	-0.0196 0.0015/0.0015 0.923	0.0205 0.0015/0.0015 0.922
	IPW	-0.0025 0.0027/0.0026 0.922	0.0000 0.0018/0.0017 0.926	0.0015 0.0017/0.0016 0.932	-0.0007 0.0017/0.0016 0.930
	AIPW	-0.0027 0.0026†/0.0026 0.930	0.0001 0.0017†/0.0017 0.938	0.0014 0.0014†/0.0014 0.924	-0.0004 0.0014†/0.0014 0.935
	AOLS	-0.0003 0.0027/0.0027 0.949	-0.0031 0.0017/0.0019 0.958	0.0017 0.0015/0.0015 0.934	-0.0009 0.0014/0.0015 0.954
	AWLS	0.0005 0.0027/0.0027 0.946	-0.0029 0.0018/0.0018 0.957	0.0020 0.0015/0.0015 0.937	-0.0013 0.0014/0.0015 0.955
AR(1) ($\rho = 0.6$) $\sigma_{ij}^2 = \sigma^2(t_{ij})$	CC	-0.0729 0.0022/0.0021 0.791	-0.0105 0.0017/0.0016 0.928	-0.0103 0.0016/0.0016 0.929	0.0132 0.0016/0.0016 0.921
	AC	-0.0404 0.0014/0.0014 0.835	-0.0067 0.0011/0.0011 0.927	-0.0100 0.0011/0.0010 0.937	0.0119 0.0011/0.0011 0.923
	IPW	-0.0016 0.0014/0.0013 0.933	0.0004 0.0012/0.0011 0.917	0.0014 0.0011/0.0011 0.934	0.0003 0.0012/0.0011 0.925
	AIPW	-0.0012 0.0013†/0.0013 0.938	0.0013 0.0011†/0.0011 0.925	0.0015 0.0010†/0.0010 0.922	0.0000 0.0010†/0.0010 0.929
	AOLS	-0.0006 0.0014/0.0014 0.951	0.0038 0.0011/0.0012 0.963	0.0013 0.0010/0.0010 0.942	-0.0001 0.0010/0.0010 0.942
	AWLS	0.0001 0.0013/0.0013 0.943	0.0009 0.0011/0.0011 0.962	0.0014 0.0010/0.0009 0.928	-0.0001 0.0009/0.0009 0.940

†: 1% trimmed empirical standard errors.

Table 4.2: Simulation results of 1000 replications for the normal data ($n = 100$, $\bar{m} = 5$), under two different correlation structures, with homoskedastic/heteroskedastic errors, and $\alpha = (2, -0.5, -1.5, 0, 0)$, about 20% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$
Compound Symmtry ($\rho = 0.6$) $\sigma_{ij}^2 \equiv \sigma^2$	CC	-0.4855 0.0041/0.0039 0.041	-0.0599 0.0026/0.0024 0.852	-0.0752 0.0025/0.0023 0.794	0.0759 0.0025/0.0023 0.796
	AC	-0.1830 0.0026/0.0025 0.375	-0.0432 0.0017/0.0017 0.856	-0.0784 0.0016/0.0016 0.652	0.0806 0.0016/0.0016 0.627
	IPW	-0.0215 0.0033/0.0026 0.861	-0.0125 0.0027/0.0020 0.856	-0.0203 0.0025/0.0019 0.826	0.0171 0.0027/0.0019 0.799
	AIPW	-0.0072 0.0027†/0.0077 0.937	0.0052 0.0022†/0.0171 0.930	-0.0024 0.0017†/0.0054 0.930	0.0051 0.0016†/0.0193 0.928
	AOLS	-0.0012 0.0027/0.0028 0.950	-0.0048 0.0019/0.0019 0.955	0.0017 0.0015/0.0014 0.942	-0.0008 0.0014/0.0015 0.955
	AWLS	0.0006 0.0027/0.0027 0.949	-0.0046 0.0019/0.0019 0.952	0.0023 0.0015/0.0015 0.941	-0.0016 0.0015/0.0015 0.953
AR(1) ($\rho = 0.6$) $\sigma_{ij}^2 = \sigma^2(t_{ij})$	CC	-0.2048 0.0026/0.0025 0.259	-0.0360 0.0019/0.0017 0.864	-0.0456 0.0017/0.0017 0.833	0.0447 0.0017/0.0017 0.855
	AC	-0.0970 0.0015/0.0015 0.457	-0.0246 0.0011/0.0011 0.887	-0.0468 0.0011/0.0011 0.718	0.0471 0.0011/0.0011 0.716
	IPW	-0.0065 0.0020/0.0016 0.874	-0.0037 0.0018/0.0013 0.883	-0.0103 0.0018/0.0013 0.842	0.0086 0.0019/0.0013 0.834
	AIPW	0.0024 0.0016†/0.0646 0.928	0.0102 0.0014†/0.1314 0.914	0.0014 0.0013†/0.0752 0.927	-0.0019 0.0013†/0.0444 0.929
	AOLS	-0.0001 0.0016/0.0016 0.947	0.0058 0.0012/0.0012 0.958	0.0006 0.0010/0.0010 0.943	0.0003 0.0010/0.0010 0.939
	AWLS	0.0025 0.0015/0.0015 0.944	0.0027 0.0011/0.0012 0.950	0.0017 0.0010/0.0010 0.951	-0.0004 0.0010/0.0010 0.940

†: 1% trimmed empirical standard errors.

Table 4.3: Simulation results of 1000 replications for the non-normal data ($n = 100$, $\bar{m} = 5$), under two different correlation structures, with homoskedastic/heteroskedastic errors, and $\alpha = (2, -0.5, -1.5, 0, 0)$, about 20% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$
Compound Symmtry ($\rho = 0.6$) $\sigma_{ij}^2 \equiv \sigma^2$	CC	-0.4674 0.0044/0.0041 0.046	-0.0676 0.0029/0.0027 0.828	-0.0714 0.0024/0.0024 0.830	0.0684 0.0024/0.0024 0.821
	AC	-0.1743 0.0026/0.0026 0.401	-0.0405 0.0017/0.0017 0.873	-0.0753 0.0016/0.0016 0.677	0.0726 0.0016/0.0016 0.691
	IPW	-0.0224 0.0033/0.0027 0.858	-0.0123 0.0028/0.0020 0.860	-0.0177 0.0028/0.0019 0.848	0.0160 0.0025/0.0019 0.840
	AIPW	-0.0027 $0.0028^{\dagger}$/0.0068 0.949	-0.0157 $0.0025^{\dagger}$/0.0164 0.916	0.0025 $0.0017^{\dagger}$/0.0106 0.937	-0.0006 $0.0017^{\dagger}$/0.0116 0.929
	AOLS	0.0007 0.0027/0.0028 0.958	-0.0041 0.0019/0.0019 0.947	0.0009 0.0014/0.0015 0.958	-0.0034 0.0015/0.0015 0.950
	AWLS	0.0036 0.0027/0.0027 0.953	-0.0037 0.0019/0.0019 0.942	0.0017 0.0014/0.0015 0.957	-0.0039 0.0015/0.0015 0.944
AR(1) ($\rho = 0.6$) $\sigma_{ij}^2 = \sigma^2(t_{ij})$	CC	-0.2000 0.0027/0.0025 0.276	-0.0345 0.0020/0.0019 0.884	-0.0410 0.0017/0.0017 0.871	0.0409 0.0017/0.0017 0.861
	AC	-0.0948 0.0015/0.0015 0.467	-0.0238 0.0011/0.0011 0.898	-0.0442 0.0011/0.0011 0.762	0.0433 0.0011/0.0011 0.745
	IPW	-0.0104 0.0019/0.0016 0.878	-0.0062 0.0018/0.0013 0.852	-0.0084 0.0018/0.0013 0.865	0.0075 0.0019/0.0013 0.839
	AIPW	0.0221 $0.0017^{\dagger}/*$ 0.941	0.0681 $0.0016^{\dagger}/*$ 0.906	-0.0176 $0.0013^{\dagger}/*$ 0.930	-0.0146 $0.0013^{\dagger}/*$ 0.920
	AOLS	-0.0031 0.0016/0.0015 0.944	0.0042 0.0012/0.0012 0.942	-0.0004 0.0010/0.0010 0.943	-0.0008 0.0010/0.0010 0.947
	AWLS	0.0010 0.0015/0.0014 0.942	0.0016 0.0012/0.0011 0.940	0.0004 0.0009/0.0009 0.940	-0.0018 0.0010/0.0009 0.941

†: 1% trimmed empirical standard errors.

Table 4.4: Simulation results of 1000 replications for the normal data ($n = 100$, $\bar{m} = 5$), under two different correlation structures, with homoskedastic/heteroskedastic errors, and $\alpha = (0.2, -0.5, -1.5, 0, 0)$, about 40% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$
Compound Symmtry ($\rho = 0.6$) $\sigma_{ij}^2 \equiv \sigma^2$	CC	-0.7716 0.0069/0.0061 0.045	-0.0698 0.0040/0.0037 0.841	-0.0914 0.0043/0.0037 0.795	0.0966 0.0043/0.0038 0.812
	AC	-0.2745 0.0029/0.0028 0.114	-0.0382 0.0018/0.0018 0.892	-0.0715 0.0018/0.0018 0.739	0.0745 0.0018/0.0018 0.731
	IPW	-0.0497 0.0042/0.0030 0.778	-0.0196 0.0036/0.0025 0.818	-0.0350 0.0038/0.0024 0.770	0.0377 0.0037/0.0024 0.743
	AIPW	0.0000 $0.0031^{\dagger}$/* 0.922	0.0393 $0.0038^{\dagger}$/* 0.905	-0.0019 $0.0023^{\dagger}$/* 0.912	-0.0305 $0.0022^{\dagger}$/* 0.921
	AOLS	-0.0009 0.0027/0.0029 0.962	-0.0063 0.0020/0.0020 0.949	0.0017 0.0015/0.0016 0.948	-0.0006 0.0015/0.0016 0.969
	AWLS	0.0018 0.0027/0.0029 0.963	-0.0062 0.0020/0.0020 0.942	0.0019 0.0015/0.0016 0.952	-0.0007 0.0015/0.0016 0.964
AR(1) ($\rho = 0.6$) $\sigma_{ij}^2 = \sigma^2(t_{ij})$	CC	-0.3357 0.0049/0.0042 0.272	-0.0405 0.0035/0.0029 0.840	-0.0516 0.0034/0.0030 0.843	0.0527 0.0034/0.0031 0.866
	AC	-0.1546 0.0016/0.0016 0.128	-0.0233 0.0012/0.0012 0.891	-0.0429 0.0012/0.0012 0.777	0.0444 0.0012/0.0012 0.773
	IPW	-0.0229 0.0025/0.0019 0.806	-0.0111 0.0023/0.0017 0.844	-0.0185 0.0025/0.0017 0.805	0.0196 0.0025/0.0017 0.785
	AIPW	-0.0207 $0.0020^{\dagger}$/* 0.917	0.0419 $0.0023^{\dagger}$/* 0.912	-0.0019 $0.0018^{\dagger}$/* 0.917	0.0615 $0.0018^{\dagger}$/* 0.902
	AOLS	-0.0076 0.0016/0.0016 0.938	0.0123 0.0013/0.0013 0.935	0.0001 0.0011/0.0011 0.945	0.0009 0.0010/0.0011 0.940
	AWLS	-0.0035 0.0015/0.0015 0.942	0.0057 0.0012/0.0013 0.950	0.0004 0.0010/0.0010 0.941	0.0006 0.0010/0.0010 0.951

†: 1% trimmed empirical standard errors.

Table 4.5: Simulation results of 1000 replications for the non-normal data ($n = 100$, $\bar{m} = 5$), under two different correlation structures, with homoskedastic/heteroskedastic errors, and $\alpha = (0.2, -0.5, -1.5, 0, 0)$, about 40% missing at random on average. For each entry, the first line displays the bias, the second line is the empirical standard error/averaged asymptotic standard error, and the third line is the 95% coverage probability.

		$\hat{\beta}_0 - \beta_0$	$\hat{\beta}_1 - \beta_1$	$\hat{\beta}_2 - \beta_2$	$\hat{\beta}_3 - \beta_3$
Compound Symmtry ($\rho = 0.6$) $\sigma_{ij}^2 \equiv \sigma^2$	CC	-0.7585 0.0082/0.0070 0.053	-0.0896 0.0052/0.0046 0.843	-0.0955 0.0048/0.0040 0.806	0.0968 0.0046/0.0040 0.819
	AC	-0.2637 0.0029/0.0028 0.144	-0.0350 0.0020/0.0020 0.905	-0.0702 0.0018/0.0018 0.760	0.0686 0.0018/0.0018 0.771
	IPW	-0.0550 0.0041/0.0029 0.760	-0.0223 0.0037/0.0025 0.802	-0.0379 0.0035/0.0023 0.760	0.0346 0.0035/0.0023 0.757
	AIPW	2.0517 $0.0033^{\dagger}/*$ 0.941	6.3115 $0.0040^{\dagger}/*$ 0.893	3.5297 $0.0022^{\dagger}/*$ 0.924	-0.9407 $0.0026^{\dagger}/*$ 0.921
	AOLS	0.0001 0.0027/0.0029 0.962	-0.0047 0.0022/0.0020 0.938	0.0013 0.0015/0.0016 0.970	-0.0028 0.0015/0.0016 0.961
	AWLS	0.0042 0.0027/0.0029 0.957	-0.0046 0.0022/0.0020 0.935	0.0016 0.0015/0.0016 0.970	-0.0026 0.0015/0.0016 0.958
AR(1) ($\rho = 0.6$) $\sigma_{ij}^2 = \sigma^2(t_{ij})$	CC	-0.3322 0.0051/0.0045 0.301	-0.0416 0.0041/0.0034 0.837	-0.0509 0.0036/0.0030 0.838	0.0526 0.0036/0.0031 0.843
	AC	-0.1492 0.0017/0.0016 0.169	-0.0198 0.0013/0.0013 0.921	-0.0428 0.0012/0.0012 0.787	0.0410 0.0012/0.0012 0.805
	IPW	-0.0283 0.0025/0.0018 0.784	-0.0120 0.0024/0.0016 0.817	-0.0177 0.0025/0.0017 0.801	0.0207 0.0024/0.0017 0.789
	AIPW	-0.0052 $0.0021^{\dagger}$/0.0125 0.926	0.0254 $0.0022^{\dagger}$/0.0433 0.870	0.0012 $0.0018^{\dagger}$/0.0258 0.927	0.0030 $0.0017^{\dagger}$/0.0140 0.920
	AOLS	-0.0091 0.0016/0.0016 0.940	0.0122 0.0013/0.0013 0.923	-0.0013 0.0010/0.0011 0.950	-0.0001 0.0011/0.0011 0.945
	AWLS	-0.0026 0.0015/0.0015 0.941	0.0067 0.0013/0.0012 0.927	-0.0009 0.0010/0.0010 0.948	-0.0006 0.0010/0.0010 0.948

†: 1% trimmed empirical standard errors.

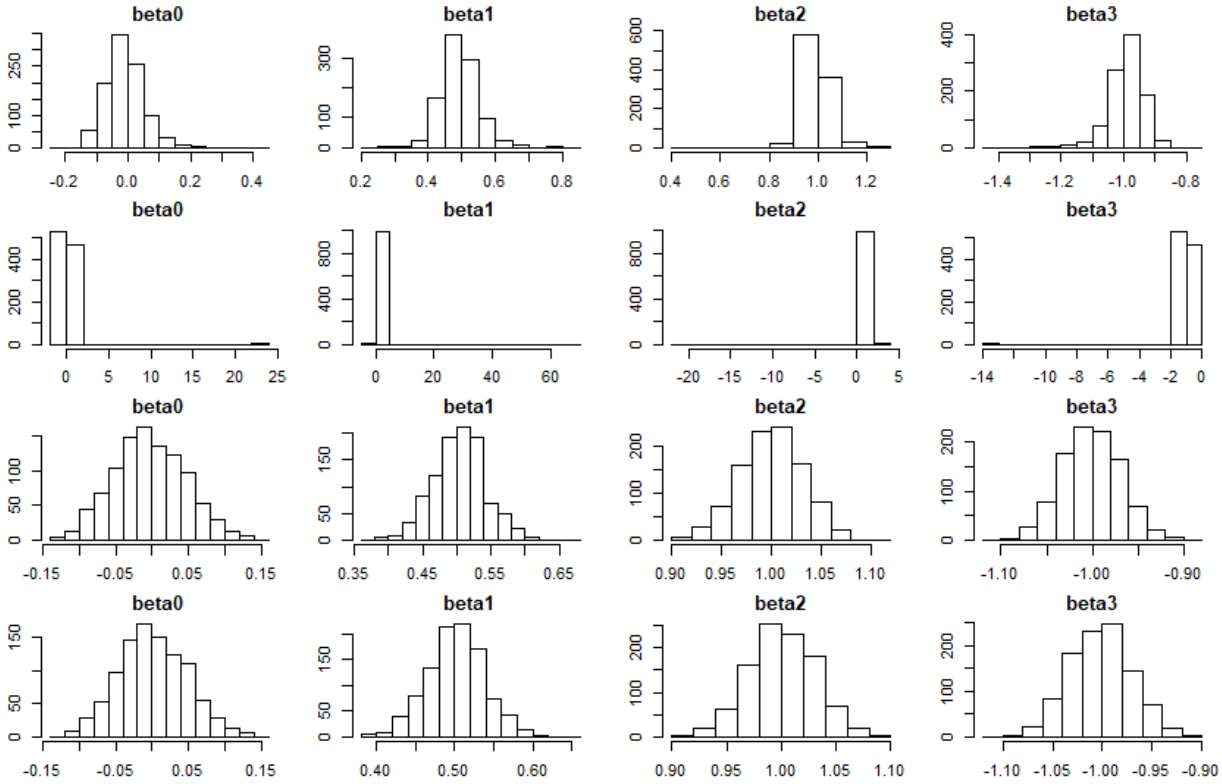

Figure 4.2: Histograms of the estimators for 1000 times simulations under the setting of Table 4.3 with AR(1) correlation structure. 1st row: IPW; 2nd row: AIPW; 3rd row: AOLS; 4th row: AWLS.

Appendix D of the book. You can also find the data online at `http://lib.stat.cmu.edu/datasets/pbcseq`.

The original interest of this study is to compare the survival distributions for the two groups and establish a Cox proportional hazards regression model for making estimation (Murtaugh et al. (1994)). Komarek & Komárková (2013) proposed a clustering method for multivariate longitudinal data and used the PBC data as an illustrative example. The original dataset has 19 attributes including some demographic variables like age and sex, clinical measurements like the presence/absence of ascites, and biochemical measurements such as the levels of bilirubin, albumin. At the end of the study, there are three types of patients' status: alive, dead or liver transplanted. For simplicity and case control, we focus on the subset of alive patients with the following variables:

1. Y-*LogBi*: logarithm of serum bilirubin (mg/dl), which is a liver bile pigment;

2. X-*LogCh*: logarithm of serum cholesterol (mg/dl), which is a blood lipoprotein;

3. Z_1-*Drug*: a binary variable indicating the treatment group (0 for placebo and 1 for DPCA);

4. Z_2-*Age*: a continuous variable as the patient's age at the measurement time;

5. t-*Year*: a continuous variable recording the years between the enrollment and this visit, obtained by dividing the original data in days by 365. Start with 0 as the measurement time.

Since bilirubin is a very important prognostic factor in PBC (Shapiro et al. (1979)), here we would like to explore the difference between the two groups on bilirubin levels, controlling other variables as cholesterol level and age. Note that the data for cholesterol is non-monotone missing. For the purpose of illustration, we delete several patients' records with missing cholesterol at the enrollment ($t = 0$). The final data we use has $n = 127$ patients with totally $N = 919$ measurement. There are 64 of them in the DPCA group and 63 in the placebo group. Z_1 is a baseline value for each patient, while other variables are time-varying. The number of measurements for each patient m_i ranges from 1 to 16, with an average as 7.24. X (log cholesterol) is always available at the first measurement, but may be missing intermittently during the follow-up with a 32.3% overall missing proportion.

We first perform simple exploratory data analysis on the data. Results are visualized in Figure 4.3. The blue curves in the right two scatterplots are obtained through locally weighted smoothing. From the plots, we find the relationships between Y and X, Z_2 are almost linear, and the distributions of Y for the two groups do not differ too much. Based on this, we consider a linear model as

$$LogBi = \beta_0 + \beta_1 LogCh + \beta_2 Drug + \beta_3 Age + \epsilon.$$

Assume the missing mechanism for X is MAR and the parametric model for selection probabilities is

$$\text{logit}(\pi_{ij}) = \alpha_0 + \alpha_1 R_{i,j-1} X_{i,j-1} + \alpha_2 Y_{ij} + \alpha_3 Z_{1ij} + \alpha_4 Z_{2ij}.$$

After fitting the logistic model, the p-values for α_1, α_2 and α_4 are all smaller than 0.001. This

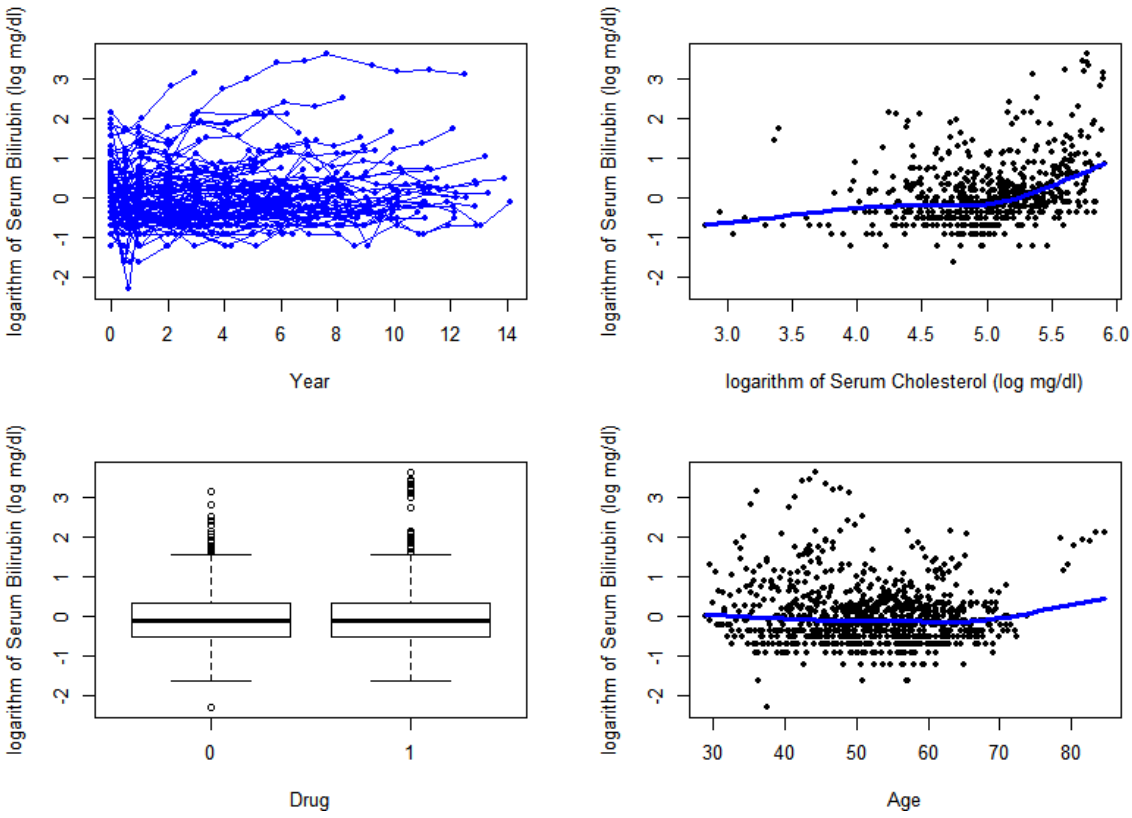

Figure 4.3: Plots for showing the relationship between log bilirubin and other variables. Top left: Scatterplot of Log Bilirubin vs. Measurement Time. Each curve presents the change in Y over time for each patient; Top right: Scatterplot of Log Bilirubin vs. Log Cholesterol; Bottom left: Side-by-side boxplot of Log Bilirubin vs. Drug; Bottom right: Scatterplot of Log Bilirubin vs. Age.

at least indicates that the missingness of X is related to Y and observed part of X, thus MCAR mechanism should not be considered and the MAR assumption seems reasonable.

We applied our proposed method to this incomplete data and compare it to other estimators just like what we did in Section 4.5. For the augmentation in AIPW, we still fit a linear regression model for X_{ij} on $\boldsymbol{Q}_{ij} = (1, Y_{ij}, \boldsymbol{Z}_{ij}^\top)^\top$. The results of the analysis can be found in Table 4.6. The values in the brackets are the standard errors of the estimators. From the table we observe that all methods conclude that *LogCh* is significant in this linear model while Age is insignificant. And *LogCh* has a positive relationship with *LogBi*. As expected, CC has larger standard errors than other estimators. The rest of the listed estimators have similar efficiency, though AWLS always

Table 4.6: PBC Data Analysis ($n = 127$, $\bar{m} = 7.24$)

	$\hat{\beta}_0(Intercept)$	$\hat{\beta}_1(LogCh)$	$\hat{\beta}_2(Drug)$	$\hat{\beta}_3(Age)$
CC	-3.4542(1.2083)	0.8600(0.1999)	-0.7904(0.3615)	-0.0.0016(0.0095)
p-value	0.0050	< 0.0001	0.0307	0.8669
AC	-2.1783(0.6921)	0.5211(0.1480)	0.0100(0.1399)	-0.0.0061(0.0091)
p-value	0.0021	0.0006	0.9432	0.5043
IPW	-2.0945(0.6302)	0.4949(0.1310)	-0.0156(0.1239)	-0.0061(0.0074)
p-value	0.0012	0.0002	0.9002	0.4094
AIPW	-2.1266(0.6034)	0.4874(0.1307)	0.0021(0.1197)	-0.0050(0.0078)
p-value	0.0006	0.0003	0.9863	0.5275
AOLS	-2.1552(0.5740)	0.4798(0.1549)	0.0199(0.1411)	-0.0040(0.0089)
p-value	0.0003	0.0024	0.8881	0.6533
AWLS	-2.1369(0.5323)	0.4703(0.1399)	0.0068(0.1318)	-0.0037(0.0085)
p-value	0.0001	0.0010	0.9591	0.6661

has smaller SE than AOLS for considering heteroscedasticity. The main difference lies in the effect of *Drug*. CC gives significant result while the rest conclude insignificance based on the p-values. Since MCAR is not reasonable here, we believe that CC can draw a wrong conclusion about the relationship between $Drug$ and $LogBi$. In this case we do not have unstable inverse-probability weights, so the results of IPW, AIPW, AOLS and AWLS are very close and more reliable than CC/AC under MAR. Based on these results, we can conclude that there is no significant difference of serum bilirubin levels between the DPCA and the placebo group. This is consistent with the boxplot in Figure 4.3 and the analysis results mentioned in Fleming & Harrington (2011) as "there are no detectable differences between the distributions of survival times for the DPCA and placebo treatment groups".

4.7 Conclusion Remarks

In this chapter we have proposed a new semiparametric estimator for longitudinal regression parameters based on augmented GEE without inverse-probability weights using a single-index model for augmentation when the covariate is non-monotone missing at random. Except the SIM and some necessary regularity conditions, we do not need to specify any likelihood or parametric models for the missing process. Since we do not include the selection probabilities in the point

estimation procedure, our method not only has a simpler algorithm, but also avoids the situation of highly variable inverse-probability weights. In this sense, numerically our proposed estimator is not sensitive to positive but near-zero selection probabilities, while IPW and AIPW can be highly influenced by those near-zero π_{ij}'s. Equally importantly, compared to using a standard multivariate kernel function, the SIM we use on augmentation avoids the problem of curse of dimensionality. Variance functions are considered for heteroscedasticity. Asymptotic theory has been developed for the proposed estimators showing the asymptotic consistency and normality, which is important for making statistical inference. All these conclusions are supported by our numerical studies.

In this work, we only considered a single univariate covariate X in simulation studies and the real data example. The results can be easily extended to the particular case of a multivariate $\boldsymbol{X}$ when $R_{ij} = 0$ means that all the covariates in $\boldsymbol{X}_{ij}$ are missing at the same time. It is interesting to consider several covariates missing separately, but this is out of the scope of this research. Further research problems include the estimation of the parameter $\boldsymbol{\rho}$ of the true correlation structure $\mathbf{C}(\boldsymbol{\rho})$, the estimation and related theory using similar methods but with another working correlation structure instead of WI, and extension to generalized linear models.

5. SUMMARY AND CONCLUSIONS

5.1 Summary

In this dissertation, we have proposed a simple semiparametric estimator for linear regression parameters based on augmented GEE without inverse-probability weights using a single-index model for augmentation when the covariate is MAR for both i.i.d. data and longitudinal data. In particular, the missing pattern in the incomplete longitudinal data is as general as being non-monotone missing.

The proposed method does not need to specify any likelihood or parametric models for the selection probability or the missing covariate except the SIM and some necessary regularity conditions. One important advantage of our proposed estimator over the (augmented) inverse-probability weighted estimators is that it does not include selection probabilities in the point estimation procedure so that it does not need to model them and avoids the situation of having highly variable inverse-probability weights. In this sense, numerically our proposed estimator is simple and not sensitive to positive but near-zero selection probabilities, while the performance of IPW and AIPW are highly influenced by those near-zero selection probabilities. Equally importantly, compared to using a standard multivariate kernel function, the SIM we use on augmentation avoids the problem of curse of dimensionality. Asymptotic theory has been developed for the proposed estimator showing the asymptotic consistency and normality, along with the sandwich formulas for asymptotic covariances. Additionally, for i.i.d. data, we have shown that our proposed estimator is asymptotically equivalent to AIPW with the same estimation for the augmentation under certain conditions, and the SIM we use also keeps the efficiency of standard kernel smoothing in some particular situations. For longitudinal data, a WI correlation structure is applied to simplify the estimation procedure, and heteroscedasticity is allowed with variance functions being estimated nonparametrically based on the partially observed residuals derived from an initial estimate.

Simulation studies for data generated from different distributions (multivariate normal or non-

normal) with different sample sizes, missing proportions and correlation structures were conducted. The results support our theoretical conclusions and illustrate the phenomenon of unstable inverse-probability weights for IPW and AIPW. The proposed method is applied to one real data example for i.i.d. data and longitudinal data, respectively.

5.2 Further Study

There are still some interesting open problems for future research. One problem is to extend the methodology to GLM or even more flexible regression models. This extension seems straightforward but actually is complicated because they no longer have the simple form of the score functions for linear models. Another possible topic is to consider several covariates missing separately, which means they are not always observed or missing at the same time. Furthermore, for longitudinal data, it is still not clear how to construct the estimation procedure and establish the theory with another structured correlation matrix instead of the WI. Finally, it will be interesting to consider generalizing the method to MNAR mechanism.

REFERENCES

Bang, H. & Robins, J. M. (2005), 'Doubly robust estimation in missing data and causal inference models', *Biometrics* **61**, 962–973.

Bishop, Y. M., Fienberg, S. E., Holland, P. W., Light, R. J. & Mosteller, F. (1977), 'Book review: Discrete multivariate analysis: Theory and practice', *Applied Psychological Measurement* **1**, 297–306.

Chen, B., Yi, G. Y. & Cook, R. J. (2010), 'Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random', *Journal of the American Statistical Association* **105**, 336–353.

Chen, B. & Zhou, X.-H. (2011), 'Doubly robust estimates for binary longitudinal data analysis with missing response and missing covariates', *Biometrics* **67**, 830–842.

Chen, H. Y. (2004), 'Nonparametric and semiparametric models for missing covariates in parametric regression', *Journal of the American Statistical Association* **99**, 1176–1189.

Daniels, M. J. & Hogan, J. W. (2008), *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*, CRC Press.

Diggle, P. (2002), *Analysis of longitudinal data*, Oxford University Press.

Fan, J., Huang, T. & Li, R. (2007), 'Analysis of longitudinal data with semiparametric estimation of covariance function', *Journal of the American Statistical Association* **102**, 632–641.

Fitzmaurice, G. M. & Laird, N. M. (1993), 'A likelihood-based method for analysing longitudinal binary responses', *Biometrika* **80**, 141–151.

Fleming, T. R. & Harrington, D. P. (2011), *Counting processes and survival analysis*, Vol. 169, John Wiley & Sons.

Friedman, J., Hastie, T. & Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics New York.

Fuchs, C. (1982), 'Maximum likelihood estimation and model selection in contingency tables with missing data', *Journal of the American Statistical Association* **77**, 270–278.

Gasser, T. & Müller, H.-G. (1984), 'Estimating regression functions and their derivatives by the kernel method', *Scandinavian Journal of Statistics* pp. 171–185.

Han, P. (2014), 'Multiply robust estimation in regression analysis with missing data', *Journal of the American Statistical Association* **109**, 1159–1173.

Han, P. (2016), 'Combining inverse probability weighting and multiple imputation to improve robustness of estimation', *Scandinavian Journal of Statistics* **43**, 246–260.

Han, P. & Wang, L. (2013), 'Estimation with missing data: beyond double robustness', *Biometrika* **100**, 417–430.

Härdle, W. (1990), *Applied nonparametric regression*, number 19, Cambridge university press.

Härdle, W., Hall, P. & Ichimura, H. (1993), 'Optimal smoothing in single-index models', *The Annals of Statistics* **21**, 157–178.

Hartley, H. & Hocking, R. (1971), 'The analysis of incomplete data', *Biometrics* **27**, 783–823.

Horton, N. J. & Laird, N. M. (1999), 'Maximum likelihood analysis of generalized linear models with missing covariates', *Statistical Methods in Medical Research* **8**, 37–50.

Hsu, C.-H., Long, Q., Li, Y. & Jacobs, E. (2014), 'A nonparametric multiple imputation approach for data with missing covariate values with application to colorectal adenoma data', *Journal of Biopharmaceutical Statistics* **24**, 634–648.

Ibrahim, J. G. (1990), 'Incomplete data in generalized linear models', *Journal of the American Statistical Association* **85**, 765–769.

Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. (2002), 'Bayesian methods for generalized linear models with covariates missing at random', *Canadian Journal of Statistics* **30**, 55–78.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R. & Herring, A. H. (2005), 'Missing-data methods for generalized linear models: A comparative review', *Journal of the American Statistical Association* **100**, 332–346.

Ibrahim, J. G. & Molenberghs, G. (2009), 'Missing data methods in longitudinal studies: a review', *Test* **18**, 1–43.

Ichimura, H. (1993), 'Semiparametric least squares (sls) and weighted sls estimation of single-

index models', *Journal of Econometrics* **58**, 71–120.

Kang, J. D. & Schafer, J. L. (2007), 'Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data', *Statistical Science* **22**, 523–539.

Komarek, A. & Komárková, L. (2013), 'Clustering for multivariate continuous and discrete longitudinal data', *The Annals of Applied Statistics* pp. 177–200.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Lipsitz, S. R., Ibrahim, J. G. & Zhao, L. P. (1999), 'A weighted estimating equation for missing covariate data with properties similar to maximum likelihood', *Journal of the American Statistical Association* **94**, 1147–1160.

Little, R. J. (1993), 'Pattern-mixture models for multivariate incomplete data', *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. (1995), 'Modeling the drop-out mechanism in repeated-measures studies', *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. & Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons. New Jersey.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L. & Gips, C. H. (1994), 'Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits', *Hepatology* **20**, 126–134.

Nadaraya, E. A. (1964), 'On estimating regression', *Theory of Probability & Its Applications* **9**, 141–142.

Reilly, M. & Pepe, M. S. (1995), 'A mean score method for missing and auxiliary covariate data in regression models', *Biometrika* **82**, 299–314.

Robins, J. M. & Ritov, Y. (1997), 'Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models', *Statistics in Medicine* **16**, 285–319.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some

regressors are not always observed', *Journal of the American statistical Association* **89**, 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121.

Robins, J., Sued, M., Lei-Gomez, Q. & Rotnitzky, A. (2007), 'Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable', *Statistical Science* **22**, 544–559.

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, Vol. 81, John Wiley & Sons. New Jersey.

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, CRC press.

Schluchter, M. D. & Jackson, K. L. (1989), 'Log-linear analysis of censored survival data with partially observed covariates', *Journal of the American Statistical Association* **84**, 42–52.

Seaman, S. & Copas, A. (2009), 'Doubly robust generalized estimating equations for longitudinal data', *Statistics in Medicine* **28**, 937–955.

Sepanski, J., Knickerbocker, R. & Carroll, R. (1994), 'A semiparametric correction for attenuation', *Journal of the American Statistical Association* **89**, 1366–1373.

Shapiro, J., Smith, H. & Schaffner, F. (1979), 'Serum bilirubin: a prognostic factor in primary biliary cirrhosis.', *Gut* **20**, 137–140.

Sinha, S., Saha, K. K. & Wang, S. (2014), 'Semiparametric approach for non-monotone missing covariates in a parametric regression model', *Biometrics* **70**, 299–311.

Van der Laan, M. J. & Robins, J. M. (2003), *Unified methods for censored longitudinal data and causality*, Springer Science & Business Media.

Wang, C., Wang, S., Zhao, L.-P. & Ou, S.-T. (1997), 'Weighted semiparametric estimation in regression analysis with missing covariate data', *Journal of the American Statistical Association* **92**, 512–525.

Wang, S. & Wang, C. (2001), 'A note on kernel assisted estimators in missing covariate regression', *Statistics & Probability Letters* **55**, 439–449.

Watson, G. S. (1964), 'Smooth regression analysis', *Sankhyā: The Indian Journal of Statistics, Series A* pp. 359–372.

Zhou, Y., Wan, A. T. K. & Wang, X. (2008), 'Estimating equations inference with missing data', *Journal of the American Statistical Association* **103**, 1187–1199.