Title	Semiparametric Inference With Nonignorable Nonresponse Data
Author(s)	森川, 耕輔
Citation	大阪大学, 2017, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/69256
rights	
Note	

Semiparametric Inference With Nonignorable Nonresponse Data

Kosuke Morikawa

DECEMBER 2017

Semiparametric Inference With Nonignorable Nonresponse Data

A dissertation submitted to
THE GRADUATE SCHOOL OF ENGINEERING SCIENCE
OSAKA UNIVERSITY
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY IN SCIENCE

BY

Kosuke Morikawa
DECEMBER 2017

Abstract

Statistical inference with missing data has become a major issue in many empirical research fields including medical science, epidemiology, econometrics and psychometrics. If data involve missing values, conventional statistical methods cannot be directly applied. In this thesis, we study nonresponse, a typical type of missingness, which means that some data are missing at dependent variables.

A key concept for valid analysis with missing data is response mechanism or missing-data mechanism. When missingness depends on the missing value, the mechanism is said to be nonignorable; most missing data are nonignorable nonresponses and this type of missingness is most difficult to handle. Appropriate analysis of nonignorable nonresponse data requires strong unverified assumptions such as existence of instrumental variables. It is hard to specify a response model in general; even though it can be specified, identifiability of the response model often fails and indeed it is even difficult to check the identifiability.

The first contribution of the thesis is to introduce a semiparametric approach to estimate a response model to overcome the difficulties described above. The first semiparametric estimator developed in this thesis is based on the method of maximum likelihood, which does not require other than the correct specification of the response model. Unfortunately, the estimator is not the most efficient. Hence, we propose two alternative semiparametric estimators which attain the semiparametric efficiency bound.

The second contribution is to provide useful conditions for checking the model identifiability in the analysis of longitudinal data with binary outcomes. The condition can be checked by observed data only, and do not rely on any instrumental variables. Based on the conditions, some identifiable models are proposed to analyze binary longitudinal data with dropouts.

Numerical experiments are conducted to show that our semiparametric estimators outperform other existing estimators in terms of bias and variance. Real data analyses of the Korean labor and income panel survey data and depot medroxyprogesterone acetate data, with our new methods, are seen to give quite reasonable results.

Contents

Chapter 1

Introduction and preliminaries

1.1 Introduction

Missing data problems are ubiquitous and require cumbersome handling techniques in many empirical research areas including econometrics, epidemiology, clinical study, and psychometrics. For example, due to the expansion of information on the Internet, personal information and privacy have become more important; people often decline to reply to personal items in a questionnaire such as income and career. In the clinical study, subjects can easily drop out from the study in view of the human participant protection. For such data, existing methods for complete data analysis cannot be directly applied.

If analysts do not properly deal with missing data, the result can be distorted and leads to incorrect conclusions. An easy way for dealing with missing data is listwise deletion. The method makes a complete dataset artificially by deleting any observations or units with missing values. However, this method has two drawbacks: (1) loss of efficiency and (2) making a biased dataset. The first problem occurs because the listwise deletion discards some data being possible to use, which yields to decrease of the sample size. The second problem is more critical from a practical perspective. Suppose that one is interested in income in a population. If lower-income workers are likely to decline to answer, then the mean income calculated from the observed data will be higher than the real income. As a result, appropriate methods for analyzing missing data need to be developed. Assumptions required for the methods should be as weak as possible.

The required assumptions are strongly related to *response mechanism* or *missing-data mechanism*. The response mechanism is defined as the conditional distribution of response probability given the complete data. This conditional probability is also called *propensity score* in the literature of causal inference (Rosenbaum and Rubin, 1983). Rubin (1976) distinguished the response mechanisms into two main parts: (i) missing at random (MAR); (ii) not missing at random (NMAR). If the mechanism does not depend on the missing data, it is called MAR, otherwise, called NMAR. When the mechanism is MAR (NMAR), it is also called *ignorable* (*nonignorable*). There have been many estimators developed under the MAR mechanism, which can be analyzed under almost the same assumptions as those with complete data. On the other hand, when the mechanism is nonignorable or NMAR, existing estimators require unverified assumptions. Hence, statistical analysis assuming nonignorable nonresponse has been criticized and has been likely to be avoided. It is a deplorable situation, and new appropriate statistical methods have to be developed.

1.2 Basic theory of the analysis of nonresponse data

Let Y be a response variable, which is subject to missingness, and let $\boldsymbol{X}$ be a fully observed covariate vector. Let R be a response indicator of Y, i.e., it takes 1(0) if Y is observed (missing). Letting $\boldsymbol{Z} = (\boldsymbol{X}^\top, Y)^\top$ and $G_r(\boldsymbol{Z})$ be observed variables in $\boldsymbol{Z}$ for $R = r$, the r-th response pattern. Here we have $G_1(\boldsymbol{Z}) = (\boldsymbol{X}, Y)$ for $R = 1$ and $G_0(\boldsymbol{Z}) = \boldsymbol{X}$ for $R = 0$. The response mechanism is defined as the conditional probability given $\boldsymbol{z}$: $\pi(\boldsymbol{z}) = \Pr(R = 1 \mid \boldsymbol{z})$. Denote our interesting parameter by $\boldsymbol{\theta}$ which is defined as a solution to $E\{U(\boldsymbol{\theta}; \boldsymbol{Z})\} = \boldsymbol{0}$, where $U(\cdot;\cdot)$ is a known function. For example, if population mean is of our interest, $U(\theta; \boldsymbol{z}) = \theta - y$. Assume that $(\boldsymbol{z}_i, r_i)$ $(i = 1, \ldots, n)$ are independent and identically distributed with a joint distribution $[\boldsymbol{z}, r]$. Here, $[\cdot]$ is generic notation for describing the distribution of the variables inside the square bracket. Notice that $[y \mid \boldsymbol{x}]$ denotes the conditional distribution of y given $\boldsymbol{x}$. Because of missingness, we observe only $(G_{r_i}(\boldsymbol{z}_i), r_i)$ $(i = 1, \ldots, n)$. By using only observed data, Horvitz and Thompson (1952) proposed an estimator for $E(Y)$ based on an unbiased estimating equation as follows:

$$\begin{aligned} n^{-1}\sum_{i=1}^{n}\frac{r_i}{\pi(\boldsymbol{z}_i)}y_i &= E\left\{\frac{R}{\pi(\boldsymbol{Z})}Y\right\} + o_p(1) \\ &= E(Y) + o_p(1). \end{aligned}$$

An estimator defined by a solution to an unbiased estimating equation will be consistent and asymptotic normal under some regularity conditions. However, the response mechanism $\pi(\boldsymbol{z})$ is generally unknown and has to be estimated. Assume that a parametric model $\pi(\boldsymbol{z};\boldsymbol{\phi})$ is specified which is known up to a value of the q-dimensional parameter $\boldsymbol{\phi}$. Estimation of $\boldsymbol{\phi}$ depends on its mechanism. When the mechanism is MAR or ignorable, the maximum likelihood estimation can be applied to estimate $\boldsymbol{\phi}$ because the missing data in y are unnecessary. Suppose that the mechanism is NMAR or nonignorable. In the case, correct specification for the outcome model $[y \mid \boldsymbol{x}]$ and the response model $[r \mid \boldsymbol{z}]$ is required. The likelihood is then complicated and sometimes the model identification is not achieved. More seriously the model identification is hard to check. Therefore estimation under NMAR mechanism has been not recommended due to these strong assumptions. In this thesis, we derive three semiparametric estimators which do not require any outcome model specification. Especially the two estimators are constructed to improve the other one and attain the semiparametric efficiency bound, where it is the lower bound of the asymptotic variance of an estimator when only a response model is specified. In addition, we propose an easy-to-check identification condition.

Assume that the response model is specified as

$$\pi(\boldsymbol{z};\boldsymbol{\phi}) = \frac{1}{1+\exp(\phi_{\mathrm{x}0}+\phi_{\mathrm{x}1}x+\phi_{\mathrm{y}}y)},$$

where $\boldsymbol{\phi} = (\phi_{\mathrm{x}0},\phi_{\mathrm{x}1},\phi_{\mathrm{y}})^\top$. Note that this response model is NMAR because it depends on y. In this setup, Greenlees et al. (1982) constructed the fully observed likelihood as follows:

$$\prod_{i=1}^{n}\{\pi(\boldsymbol{z}_i;\boldsymbol{\phi})f(y_i \mid x_i;\boldsymbol{\beta})\}^{r_i}\left\{1-\int\pi(x_i,y;\boldsymbol{\phi})f(y \mid x_i;\boldsymbol{\beta})dy\right\}^{1-r_i},$$

where $f(y \mid x;\boldsymbol{\beta})$ is a model for the conditional distribution of $[y \mid x]$ known up to a finite dimensional parameter $\boldsymbol{\beta}$. The term for $r=0$ is integrated out with regard to

y since y can not be observed. Thus the correct specification of the true conditional distribution is needed, and misspecification of the distribution leads to inconsistency of any estimator for $\boldsymbol{\phi}$. The question is whether we can remove the assumption of the correct specification of $f(y \mid x; \boldsymbol{\beta})$. We will answer to this question in Chapter 2.

Conditions for the model identification are also to be clear. If we can observe complete data, the identification condition is

$$\begin{aligned} &\pi(y \mid x; \boldsymbol{\phi}) = \pi(y \mid x; \boldsymbol{\phi}') \quad \text{w.p. 1} \\ \Rightarrow\ &\boldsymbol{\phi} = \boldsymbol{\phi}', \end{aligned}$$

where w.p. 1 means with probability one with respect to the true distribution. When some values in y are subject to missingness, the identification condition is changed to,

$$\begin{aligned} &\pi(y \mid x; \boldsymbol{\phi}) f(y \mid x; \boldsymbol{\beta}) = \pi(y \mid x; \boldsymbol{\phi}') f(y \mid x; \boldsymbol{\beta}') \quad \text{w.p. 1} \\ \Rightarrow\ &\boldsymbol{\phi} = \boldsymbol{\phi}' \text{ and } \boldsymbol{\beta} = \boldsymbol{\beta}'. \end{aligned}$$

However, this condition does not generally hold. For example, Miao et al. (2016) showed if the response models are $\pi(\boldsymbol{z}; \boldsymbol{\phi}) = \operatorname{expit}(-\log 2 + y)$ and $\pi(\boldsymbol{z}; \boldsymbol{\phi}') = \operatorname{expit}(\log 2 - y)$, and outcome models are $f(y; \boldsymbol{\beta}) = 2e^{-2y}$ and $f(y; \boldsymbol{\beta}') = e^{-y}$, the identifiability does not hold, where "expit" is the inverse function of "logit". We will give a new necessary and sufficient condition for the model identification.

1.3 Basic theory of semiparametric inference

In this section, we consider semiparametric estimation in the sense that response and outcome models are parametric and nonparametric, respectively. Because the outcome model is nonparametric, we somehow need to restrict a class of semiparametric estimator to rule out *super efficient estimators* such as Hodges's estimator and Stein's estimator to discuss the efficiency bound, which is the lower bound of the asymptotic variance (Bickel et al., 1998; Tsiatis, 2006). In this section, we consider a case in which there is no missing data to illustrate the semiparametric estimation for the time being. We introduce Hodges's super efficient estimator as an illustration of such estimators.

Example 1.3.1. (Hodges's super efficient estimator)

Let $X_1, \ldots, X_n$ be n i.i.d. copies of $N(\mu_0, 1)$, where $\mu_0 \in \mathbb{R}$ is a constant. Here $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the maximum likelihood estimator (MLE) of μ_0, which has the following asymptotic normality:

$$\sqrt{n}(\bar{X}_n - \mu_0) \xrightarrow{L(\mu_0)} N(0,1),$$

where the above convergence stands for the weak convergence under the distribution of $\mu = \mu_0$. Let us consider the estimator defined by Hodges in 1951 (see Tsiatis, 2006, for more details):

$$\hat{\mu}_n = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| > n^{-1/4} \\ 0 & \text{if } |\bar{X}_n| \leq n^{-1/4} \end{cases}.$$

We can easily show that when $\mu_0 \neq 0$, $\sqrt{n}(\hat{\mu}_n - \mu_0) \xrightarrow{L(\mu_0)} N(0,1)$, where $\xrightarrow{L(\mu)}$ denotes convergence in law under the distribution of $f(x;\mu)$. When $\mu_0 = 0$, it holds that $\sqrt{n}(\hat{\mu}_n - 0) \xrightarrow{L(0)} N(0,0)$. Therefore, it seems that this estimator is more efficient than MLE, and that is why this estimator is called *super-efficient estimator*. However, this nice property is gained at the expense of poor estimation in a neighborhood of $\mu = 0$. In fact, when the distribution is assessed at $\mu = \mu_n = n^{-1/3}$, MLE has asymptotic normality:

$$\sqrt{n}(\bar{X}_n - \mu_n) \xrightarrow{L(\mu_n)} N(0,1).$$

However, the super-efficient estimator does not, that is, $\sqrt{n}(\hat{\mu}_n - \mu_n) \to \infty$. We need to rule out such estimators to discuss the efficiency. □

We introduce *regularity* to exclude super efficient estimators. Let $\boldsymbol{\xi}^{p\times 1} = (\boldsymbol{\theta}^\top, \boldsymbol{\eta}^\top)^\top$, where $\boldsymbol{\theta}^{q\times 1}$ is our interest, $\boldsymbol{\eta}^{a\times 1}$ is a finite dimensional nuisance parameter (this "finite" restriction is relaxed to "infinite" later), and $p = q + a$.

Definition 1.3.1. (Regularity)

Assume that data $\boldsymbol{z}_i$ $(i = 1, \ldots, n)$ are generated from $\boldsymbol{\xi}_n = (\boldsymbol{\theta}_n^\top, \boldsymbol{\eta}_n^\top)^\top$ satisfying $\sqrt{n}(\boldsymbol{\xi}_n - \boldsymbol{\xi}^*)$ converges to a constant. Then, an estimator of $\hat{\boldsymbol{\theta}}_n$ is called regular if the estimator $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n)$ converges to a limit distribution which does not depend on $\boldsymbol{\xi}^*$. □

Strictly speaking, this is the definition of *local regularity.* However in the literature of semiparametric estimation, it is often called just *regular* (see Ibragimov and Has'minskii, 1981, for the original *regularity* definition). Note that this regularity clearly rules out the Hodges's super-efficient estimator. In addition to the regularity, we need another condition, that is, *asymptotic linearity.*

Definition 1.3.2. (Asymptotic linearity)

For an estimator $\hat{\boldsymbol{\theta}}_n$, if there exists a function $\boldsymbol{\varphi}$ such that $E\{\boldsymbol{\varphi}(\boldsymbol{Z})\} = \mathbf{0}$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{Z}_i) + o_p(1),$$

and $E\{\boldsymbol{\varphi}(\boldsymbol{Z})\boldsymbol{\varphi}(\boldsymbol{Z})^\top\}$ is finite and nonsingular, the estimator is called *asymptotic linear*. Here the function $\boldsymbol{\varphi}(\cdot)$ is called an *influence function.* □

It can be shown that an asymptotic linear estimator has the unique influence function. Furthermore, the asymptotic variance of the estimator can be computed by $E\{\boldsymbol{\varphi}(\boldsymbol{Z})\boldsymbol{\varphi}(\boldsymbol{Z})^\top\}$. Thus it is enough to focus on finding the best influence function, which attains the lower bound of the estimator, and we can find such an influence function in the class of regular and asymptotic linear (RAL) estimators. However, as a consequence of Hájek (1970)'s representation theorem, it can be shown that the most efficient regular estimator is asymptotically linear(Tsiatis, 2006). For the above reasons, we construct semiparametric estimators among the class of RAL estimators in this paper. Our first goal is to find the most efficient influence function $\boldsymbol{\varphi}^*(\cdot)$.

Consider a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, where $\mathcal{P} = \{P_\xi : \boldsymbol{\xi} \in \Xi\}$, $\Xi \subset \mathbb{R}^p$ is the parameter space of $\boldsymbol{\xi}$, and P_ξ is the probability measure characterized by the parameter $\boldsymbol{\xi} = (\boldsymbol{\theta}^\top, \boldsymbol{\eta}^\top)^\top$. Suppose that $p(\boldsymbol{z}; \boldsymbol{\xi})$ be the density function of the distribution of $\boldsymbol{Z}$ and define a Hilbert space $\mathcal{H}$ by

$$\mathcal{H} = \{\boldsymbol{h} : \Omega \to \mathbb{R}^q \mid E\{\boldsymbol{h}(\cdot)\} = \mathbf{0}, \|\boldsymbol{h}\| < \infty\}$$

with the inner product $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = E(\boldsymbol{h}_1^\top \boldsymbol{h}_2)$ for all $\boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathcal{H}$, where this expectation is taken with respect to the true distribution. Denote the score vector of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ by

$$\boldsymbol{S}_\theta(\boldsymbol{Z}) := \frac{\partial \log p(\boldsymbol{Z}; \boldsymbol{\xi}_0)}{\partial \boldsymbol{\theta}}, \quad \boldsymbol{S}_\eta(\boldsymbol{Z}) := \frac{\partial \log p(\boldsymbol{Z}; \boldsymbol{\xi}_0)}{\partial \boldsymbol{\eta}},$$

and the nuisance tangent space by

$$\Lambda := \{B^{q \times a} \boldsymbol{S}_\eta(Z) \mid \text{for all } q \times a \text{ matrices } B\}.$$

Then, the efficient influence function is given as follows.

Lemma 1.3.1. (Corollary 2 of Tsiatis (2006))

When the parameter $\boldsymbol{\xi}$ can be partitioned as $(\boldsymbol{\theta}^\top, \boldsymbol{\eta}^\top)^\top$, the efficient influence function for $\boldsymbol{\theta}$ can be written as

$$\varphi_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0) = \{E(\boldsymbol{S}_{\text{eff}}\boldsymbol{S}_{\text{eff}}^\top)^{-1}\}\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0)\},$$

where

$$\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0) = \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{\xi}_0) - \Pi(\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{\xi}_0) \mid \Lambda)$$

and $\Pi(\cdot \mid \Lambda)$ is the projection operator onto the space Λ.

Hence the lower bound of the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is given by

$$\{E(\varphi_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0)\varphi_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0)^\top)\}^{-1} = [E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0)\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z}, \boldsymbol{\xi}_0)^\top\}]^{-1}. \tag{1.1}$$

This result can be extended to the models with an infinite dimensional nuisance parameter $\boldsymbol{\eta}$. To discuss the efficiency, we at first define the semiparametric efficiency bound. Let $\mathcal{P}$ be a class of models of $p(\boldsymbol{z}; \boldsymbol{\xi}_0)$ and $\mathcal{P}_{\boldsymbol{\theta},\boldsymbol{\alpha}}$ be a parametric submodel of $\mathcal{P}$ with a finite dimensional nuisance parameter $\boldsymbol{\alpha}$. Then, the semiparametric efficiency bound is defined as

$$\sup_{\{\text{all parametric submodels}\}} [E\{S^{\text{eff}}_{\boldsymbol{\theta},\boldsymbol{\alpha}}(S^{\text{eff}}_{\boldsymbol{\theta},\boldsymbol{\alpha}})^\top]^{-1}, \tag{1.2}$$

where $S^{\text{eff}}_{\boldsymbol{\theta},\boldsymbol{\alpha}}$ is the efficient score for a parametric submodel with $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\alpha})$. The semiparametric lower bound (1.2) is larger than or equal to any parametric submodel. Next, we shall define the infinite dimensional version of the nuisance tangent space.

Definition 1.3.3. (Nuisance tangent space with infinite dimensional parameter $\boldsymbol{\eta}$)
Let $p(\boldsymbol{z}; \boldsymbol{\theta}^{q\times 1}, \boldsymbol{\alpha}^{a\times 1})$ and the nuisance tangent space

$$\Lambda_{\boldsymbol{a}} = \{B^{q\times a}S^{a\times 1}_{\boldsymbol{\alpha}}(\boldsymbol{Z}; \boldsymbol{\theta}_0, \boldsymbol{a}_0) \mid \text{for all } q \times a \text{ matrices } B\},$$

where $(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0)$ is the true value of the parameter vector, and $\boldsymbol{S}_{\boldsymbol{\alpha}}$ is the score vector for the nuisance parameter $\boldsymbol{a}$ for the parametric submodel. Then the nuisance tangent space Λ for a semiparametric model with the infinite dimensional nuisance parameter $\boldsymbol{\eta}$ is defined as the mean-square closure of the set of the nuisance tangent spaces for all parametric submodels. Specifically, let the space $\Lambda \subset \mathcal{H}$, where $\Lambda = [\boldsymbol{h}^{q\times 1}(\boldsymbol{Z}) \in \mathcal{H}$ such that $\|\boldsymbol{h}\| < \infty$ and there exists a sequence $\{B_j\boldsymbol{S}_{\alpha j}(\boldsymbol{Z})\}_{j\in\mathbb{N}}$ such that

$$\|\boldsymbol{h}(\boldsymbol{Z}) - B_j\boldsymbol{S}_{\alpha j}(\boldsymbol{Z})\|^2 \to 0 \quad \text{as} \quad j \to \infty$$

for a sequence of parametric submodels indexed by j].

Definition 1.3.4. (Semiparametric efficient score with infinite dimensional parameter $\boldsymbol{\eta}$)

The semiparametric efficient score for $\boldsymbol{\theta}$ is defined as

$$\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z};\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) = \boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z};\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) - \Pi\{\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{Z};\boldsymbol{\theta}_0,\boldsymbol{\eta}_0) \mid \Lambda\}.$$

Then, it can be shown that the semiparametric efficiency bound defined in (1.2) is given as

$$[E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)\boldsymbol{S}_{\text{eff}}(\boldsymbol{Z},\boldsymbol{\theta}_0,\boldsymbol{\eta}_0)^\top\}]^{-1}, \tag{1.3}$$

which is the same as the finite dimensional nuisance parameter case defined in (1.1).

Lemma 1.3.2. (Theorem 4.1 of Tsiatis (2006))

The lower bound (1.2) is given as (1.3).

As a result of the lemma, once we could find the nuisance tangent space Λ, an estimator which attains the semiparametric efficiency bound could be obtained. We will find the nuisance tangent space for the semiparametric models with nonignorable nonresponse data, and propose semiparametric adaptive estimators later.

Chapter 2

Semiparametric estimation

2.1 Introduction

Handling missing data often requires some assumptions about the response mechanism. If an outcome variable does not affect the probability of the response, the response mechanism is called missing at random (MAR) (Rubin, 1976). If, on the other hand, the response probability of the outcome variable depends on that variable directly, the response mechanism is called not missing at random (NMAR) (Little and Rubin, 2002). Under NMAR, the response probability cannot be verified using the observed study variables only, therefore, additional assumptions about the study variable are often required.

Greenlees et al. (1982) and Diggle and Kenward (1994) proposed a fully parametric approach to analyze nonignorable nonresponse data; their method requires two parametric models: (i) an outcome model, $[y \mid \boldsymbol{x}]$, and (ii) a response model $[r \mid \boldsymbol{x}, y]$. In practice, it is difficult to verify models (i) and (ii) because some y are not observed. For the fully parametric approach, model identification and model misspecification can be a problem, and sensitivity analysis becomes necessary (Scharfstein et al., 1999; Rotnitzky et al., 2001; Verbeke et al., 2001; Tsiatis, 2006). Sverchkov (2008) and Riddles et al. (2016) proposed a fully parametric approach that uses different model specifications based on (i) $[y \mid \boldsymbol{x}, r = 1]$, and (ii) $[r \mid \boldsymbol{x}, y]$. Their approach is attractive because one can verify a model for $[y \mid \boldsymbol{x}, r = 1]$ from the observed responses; however, because it is a fully parametric approach, it is still subject to model misspecification.

Recently, some semiparametric approaches have been proposed for nonignorable nonresponses. Ma et al. (2003) studied identification and parameter estimation for binary study variables. Tang et al. (2003) also considered model identification using an instrumental variable and proposed a maximum pseudo likelihood estimator that does not require specification of the response mechanism. D'Haultfoeuille (2010) also used the same instrumental variable assumption and considered a regression analysis using the nonparametric propensity score model. Zhao and Shao (2015) extended the method of Tang et al. (2003) and relaxed the condition on the instrumental variable, which is called nonresponse instrumental variable (Wang et al., 2014). Fitzmaurice et al. (2005) and Skrondal and Rabe-Hesketh (2014) proposed protective estimators that do not require a specification of the response mechanism, but the application of this approach is limited to situations in which y is binary. In the meantime, Kim and Yu (2011) proposed a semiparametric method for estimating $E(Y)$ using a semiparametric response model, but a validation sample is required in order to estimate the parameters in the response mechanism. Tang et al. (2014) used the method of empirical likelihood to extend the method of Kim and Yu (2011) to estimate more general parameters. In Zhao et al. (2017), the method of Qin et al. (2002) was used to construct a $n^{1/2}$-consistent estimator without a validation sample. Chang and Kott (2008) and Kott and Chang (2010) and Wang et al. (2014) considered a generalized method of moments (GMM) estimator that uses the response model assumption only, but their method is generally lacking in efficiency. Recently, Shao and Wang (2016) proposed a semiparametric inverse propensity weighting method using the nonresponse instrumental variable assumption of Wang et al. (2014). However, the above papers do not address efficiency of their semiparametric estimation methods. Developing an optimal semiparametric estimator is an important research gap in missing data analysis.

In this section, we use a parametric model for $[r \mid \boldsymbol{x}, y]$ and construct optimal estimators for parameters both related to the response mechanisms and for the parameter of interest such as population mean. Efficiency under this setup has already been discussed by Rotnitzky and Robins (1997) and Robins et al. (1999). However, their estimator requires many working models to achieve the semiparametric efficiency bound. Misspecification of the working models may lead to loss of efficiency. Therefore, we consider an alternative approach and propose two semiparametric estimators that attain the semiparametric lower bound (Bickel et al., 1998) (1) with a working model assumption or (2) without requiring working model assumptions. The first estimator is an adaptive estimator similar in spirit to the generalized linear estimator proposed by Liang and Zeger (1986) using a working model for $[y \mid \boldsymbol{x}, r = 1]$. If the working model is correct, the first estimator attains the lower bound. The second one is based on the nonparametric regression model which does not require any additional assumptions, but it still attains the lower bound. All technical details are given in Appendix B.

The results in this chapter are mainly owe to Morikawa and Kim (2017a) and Morikawa, Kim, and Kano (2018).

2.2 Basic setup

Let $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)^\top$ $(i = 1, \dots, n)$ be independently and identically distributed realizations from unknown distribution $F(\boldsymbol{z})$. Suppose that the response model is $\pi(\boldsymbol{z}; \boldsymbol{\phi})$ with a q-dimensional parameter $\boldsymbol{\phi} \in \Phi$. Let $\theta \in \Theta$ be an one-dimensional parameter satisfying $E\{U(\boldsymbol{Z}; \theta)\} = 0$, where U is a known function of $\boldsymbol{z}$, which does not prescribe the distribution of $[\boldsymbol{x}, y]$. For example, if we are interested in $E(Y)$, then $U(\boldsymbol{z}; \theta) = y - \theta$. In this paper, we consider semiparametric estimation of $(\boldsymbol{\phi}, \theta)$ from partial observations. In particular, we seek the most efficient estimator among the RAL estimators defined in Chapter 1 and propose two adaptive estimators.

Recently, Riddles et al. (2016) proposed an efficient estimator that uses a parametric model for $[y \mid \boldsymbol{x}, r = 1]$. Using the mean score theorem (Louis, 1982), the maximum likelihood estimator can be obtained by solving

$$\sum_{i=1}^{n} [r_i \boldsymbol{s}_1(\boldsymbol{z}_i; \boldsymbol{\phi}) + (1 - r_i) E_0\{\boldsymbol{s}_0(\boldsymbol{Z}; \boldsymbol{\phi}) \mid \boldsymbol{x}_i\}] = 0, \tag{2.1}$$

where $\boldsymbol{s}_r(\boldsymbol{z}; \boldsymbol{\phi})$ is the score function of $\boldsymbol{\phi}$, that is,

$$\boldsymbol{s}_r(\boldsymbol{z}; \boldsymbol{\phi}) = \frac{\{r - \pi(\boldsymbol{z}; \boldsymbol{\phi})\} \dot{\boldsymbol{\pi}}(\boldsymbol{z}; \boldsymbol{\phi})}{\pi(\boldsymbol{z}; \boldsymbol{\phi})\{1 - \pi(\boldsymbol{z}; \boldsymbol{\phi})\}}, \tag{2.2}$$

$\dot{\boldsymbol{\pi}}(\boldsymbol{z}; \boldsymbol{\phi}) = \partial \pi(\boldsymbol{z}; \boldsymbol{\phi}) / \partial \boldsymbol{\phi}$, and $E_0(\cdot \mid \boldsymbol{x})$ is the conditional expectation conditional on $\boldsymbol{x}$ and $r = 0$. To compute $E_0(\cdot \mid \boldsymbol{x})$, under Bayes' formula, Riddles et al. (2016) proposed using

$$\sum_{i=1}^{n} \left[r_i \boldsymbol{s}_1(\boldsymbol{z}_i; \boldsymbol{\phi}) + (1 - r_i) \frac{E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \boldsymbol{s}_0(\boldsymbol{Z}; \boldsymbol{\phi}) \mid \boldsymbol{x}_i\}}{E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid \boldsymbol{x}_i\}} \right] = 0, \tag{2.3}$$

where $O(\boldsymbol{z};\boldsymbol{\phi}) = \{1-\pi(\boldsymbol{z};\boldsymbol{\phi})\}/\pi(\boldsymbol{z};\boldsymbol{\phi})$, and $E_1(\cdot \mid \boldsymbol{x})$ is the conditional expectation on y given $\boldsymbol{x}$ and $r=1$. The conditional expectation is computed by assuming a parametric model $f_1(y \mid \boldsymbol{x};\boldsymbol{\gamma}) = f(y \mid \boldsymbol{x}, r=1;\boldsymbol{\gamma})$. This may increase the efficiency, however, because misspecification of the f_1 model would cause the solution $\hat{\boldsymbol{\phi}}$ to be inconsistent. Then, the interesting parameter can be estimated as a solution to

$$\sum_{i=1}^{n} \frac{r_i U(\boldsymbol{z}_i;\theta)}{\pi(\boldsymbol{z}_i;\boldsymbol{\phi})} = 0. \tag{2.4}$$

We consider an extension of this method to seimparametric estimation in the next section.

Next, two existing semiparametric estimators, which do not require any outcome model, are introduced: (i) empirical likelihood (EL) approach; (ii) moment-base approach. Without loss of generality, assume that the first m elements are observed, and the remaining $(n-m)$ elements are missing in y_i, i.e., $r_i = 1$ for $i = 1,\dots,m$ and $r_i = 0$ for $i = m+1,\dots n$. Because we cannot observe y when $r=0$, the likelihood can be written as

$$\prod_{i=1}^{m} \pi(\boldsymbol{\phi};\boldsymbol{z}_i)dF(\boldsymbol{z}_i) \prod_{i=m+1}^{n} \int \{1-\pi(\boldsymbol{\phi};\boldsymbol{z})\}dF(\boldsymbol{z}). \tag{2.5}$$

Qin et al. (2002) discretized the distribution F by w_i $(i=1,\dots,m)$. The discretized distribution w_i can be estimated by maximizing $\prod_{i=1}^{m} w_i$ under the following constraints:

$$w_i \geq 0, \quad \sum_{i=1}^{m} w_i = 1, \quad \sum_{i=1}^{m} w_i\{\pi(\boldsymbol{\phi};\boldsymbol{z}_i) - W\} = 0,$$

$W = \Pr(R=1) = \int \pi(\boldsymbol{z};\boldsymbol{\phi}_0)dF(\boldsymbol{z})$, and

$$\sum_{i=1}^{m} w_i\{\boldsymbol{h}(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_n\} = 0, \tag{2.6}$$

where $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^{p_1}$ $(p_1 \geq q-1)$ is an arbitrary function of $\boldsymbol{x}$, and $\bar{\boldsymbol{h}}_n = n^{-1}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{x}_i)$. The $\boldsymbol{h}(\boldsymbol{x})$ function helps to improve the efficiency. By introducing Lagrange multipliers, the solution to the above optimization problem is

$$\hat{w}_i = \frac{1}{m[1 + \boldsymbol{\lambda}_1^\top\{\boldsymbol{h}(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_n\} + \lambda_2\{\pi(\boldsymbol{\phi};\boldsymbol{z}_i) - W\}]}. \tag{2.7}$$

By profiling out the unknown F with estimated $\hat{w}_i$ $(i = 1, \ldots, m)$ in (2.5) and taking logarithm, we obtain the profile pseudo-loglikelihood:

$$\begin{aligned}&\ell(\boldsymbol{\phi}, W, \boldsymbol{\lambda}_1)\\&= \sum_{i=1}^{m} \log \pi(\boldsymbol{\phi}; \boldsymbol{z}_i) - \sum_{i=1}^{m} \log[1 + \boldsymbol{\lambda}_1^\top \{\boldsymbol{h}(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_n\} + \lambda_2\{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i) - W\}]\\&+ (n-m)\log(1-W),\end{aligned} \tag{2.8}$$

where $\lambda_2 = (n/m - 1)/(1 - W)$. Qin et al. (2002) proposed a semiparametric estimator for $\boldsymbol{\phi}$ by maximizing the profile pseudo-loglikelihood. In the optimization procedure, it requires some computational techniques (see Chen et al., 2002), because the maximizer of (2.8) must satisfy $\hat{w}_i \geq 0$, which are defined in (2.7).

On the other hand, under the same assumptions, Chang and Kott (2008) and Kott and Chang (2010) proposed another semiparametric estimator by solving the following estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)} - 1 \right\} \boldsymbol{g}(\boldsymbol{x}_i) = 0, \tag{2.9}$$

where $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^q$ is an arbitrary function of $\boldsymbol{x}$. This equation is called "calibration" in the literature of survey sampling. A typical choice for $\boldsymbol{g}$ when $d = 1$ is $\boldsymbol{g}(\boldsymbol{x}) = (1, x, \ldots, x^{q-1})^\top$. It is hard to decide the variables in the calibration condition when $d > 1$. Also, when the dimension of $\boldsymbol{g}(\boldsymbol{x})$ is larger than q, say p_2, the generalized method of moments (GMM) method (Hansen, 1982) can be used to estimate $\boldsymbol{\phi}$. Because

$$\mathrm{Var}\left[\left\{\frac{R}{\pi(\boldsymbol{Z})} - 1\right\} \boldsymbol{g}(\boldsymbol{X})\right] = \mathrm{E}\{O(\boldsymbol{Z})\boldsymbol{g}(\boldsymbol{X})^{\otimes 2}\},$$

the GMM estimator can be constructed by

$$\hat{\boldsymbol{\phi}} := \operatorname*{argmin}_{\boldsymbol{\phi}} \sum_{i=1}^{n} \left\{ \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)} - 1 \right\}^2 \boldsymbol{g}(\boldsymbol{x}_i)^\top \hat{\boldsymbol{V}}^{-1} \boldsymbol{g}(\boldsymbol{x}_i), \tag{2.10}$$

where $B^{\otimes 2} = BB^\top$ for any matrix B, $\hat{\boldsymbol{V}} = n^{-1}\sum_{i=1}^{n} O(\boldsymbol{z}_i)\boldsymbol{g}(\boldsymbol{x}_i)^{\otimes 2}$, and $O(\boldsymbol{z}) = \{1 - \pi(\boldsymbol{z})\}/\pi(\boldsymbol{z})$. The optimization of both (2.9) and (2.10) are much simpler than that of Qin et al. (2002) since there is no constraint in the optimization.

The two semiparametric estimation methods use the same assumptions but seem to provide different estimation results. A natural question is "which one is better?". Or, "are there any real difference between the two methods?". We will answer to these questions in §2.5.

2.3 Semiparametric maximum likelihood approach

The proposed method is different from Riddles et al. (2016) in two ways. First, instead of assuming a parametric model for $f_1(y \mid \boldsymbol{x})$, we use a nonparametric regression approach to compute the conditional expectation in (2.1). Second, parameter estimation of θ is not based on the propensity score method in (2.4). Instead, we consider an approach using expected estimating equations (Wang and Pepe, 2000), which will lead to more efficient parameter estimation.

To compute the conditional expectation in (2.1) nonparametrically, note that

$$\begin{aligned} E_0\{\boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{x}, Y) \mid \boldsymbol{x}\} &= \frac{\int \boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{x}, Y) O(\boldsymbol{\phi}; \boldsymbol{x}, y) f_1(y \mid \boldsymbol{x}) dy}{\int O(\boldsymbol{\phi}; \boldsymbol{x}, y) f_1(y \mid \boldsymbol{x}) dy} \\ &=: \frac{\boldsymbol{C}_s(\boldsymbol{\phi}; \boldsymbol{x})}{D(\boldsymbol{\phi}; \boldsymbol{x})}. \end{aligned} \tag{2.11}$$

Using kernel smoothing (e.g. Wasserman, 2005), we can estimate $\boldsymbol{C}_s(\cdot)/D(\cdot)$ by $\hat{\boldsymbol{C}}_s(\cdot)/\hat{D}(\cdot)$, where

$$\begin{aligned} \hat{\boldsymbol{C}}_s(\boldsymbol{\phi}; \boldsymbol{x}) &:= (nh^d)^{-1} \sum_{j=1}^{n} r_j K_h(\boldsymbol{x}_j - \boldsymbol{x}) O(\boldsymbol{\phi}; \boldsymbol{x}, y_j) \boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{x}, y_j), \\ \hat{D}(\boldsymbol{\phi}; \boldsymbol{x}) &:= (nh^d)^{-1} \sum_{j=1}^{n} r_j K_h(\boldsymbol{x}_j - \boldsymbol{x}) O(\boldsymbol{\phi}; \boldsymbol{x}, y_j), \end{aligned}$$

d is the dimension of $\boldsymbol{x}$, $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function, $K_h(\boldsymbol{x}) = K(\boldsymbol{x}/h)$, and h is an appropriate bandwidth which satisfies certain regularity conditions. The mean score equation (2.1) is now approximated by

$$\hat{\mathbf{S}}(\boldsymbol{\phi}) := n^{-1} \sum_{i=1}^{n} \left[r_i \boldsymbol{s}_1(\boldsymbol{\phi}; x_i, y_i) + (1 - r_i) \frac{\hat{\boldsymbol{C}}_s(\boldsymbol{\phi}; \boldsymbol{x}_i)}{\hat{D}(\boldsymbol{\phi}; \boldsymbol{x}_i)} \right] = \mathbf{0}. \tag{2.12}$$

Thus, $\boldsymbol{\phi}$ can be estimated without specifying any parametric distributional assumptions on the outcome variable Y. The solution $\hat{\boldsymbol{\phi}}$ from (2.12) can be called the semiparametric maximum likelihood estimator of $\boldsymbol{\phi}$ because we use a parametric model for $[r \mid \boldsymbol{x}, y]$, but use a nonparametric model for $f_1(y \mid \boldsymbol{x})$.

Given the solution $\hat{\boldsymbol{\phi}}$ from (2.12), parameter θ can be estimated as the solution to

$$n^{-1} \sum_{i=1}^{n} \left[r_i U(\theta; \boldsymbol{z}_i) + (1 - r_i) E_0\{U(\theta; \boldsymbol{x}_i, Y) \mid \boldsymbol{x}_i, \hat{\boldsymbol{\phi}}\} \right] = 0. \tag{2.13}$$

The above estimating equation, called the expected estimating equation, leads to an efficient estimator of θ. To compute the conditional expectation in (2.13), we use the kernel smoothing method again; the left side of (2.13) can be approximated by

$$\hat{\mathrm{U}}(\theta, \hat{\boldsymbol{\phi}}) := n^{-1} \sum_{i=1}^{n} \left[r_i U(\theta; \boldsymbol{z}_i) + (1 - r_i) \frac{\hat{C}_U(\theta, \hat{\boldsymbol{\phi}}; \boldsymbol{x}_i)}{\hat{D}(\hat{\boldsymbol{\phi}}; \boldsymbol{x}_i)} \right], \tag{2.14}$$

where

$$\hat{C}_U(\theta, \boldsymbol{\phi}; \boldsymbol{x}) := (nh^d)^{-1} \sum_{j=1}^{n} r_j K_h(\boldsymbol{x}_j - \boldsymbol{x}) O(\boldsymbol{\phi}; \boldsymbol{x}, y_j) U(\theta; \boldsymbol{x}, y_j).$$

Remark 2.3.1. Choosing an appropriate bandwidth is a challenging problem in nonparametric estimation. In this paper, we consider the following cross-validation method for bandwidth selection. This method is only suitable for choosing a bandwidth for (2.12), but the bandwidth selection for (2.14) can be conducted similarly.

Step 1. For each bandwidth $h \in H$, compute $\hat{\boldsymbol{\phi}}_h$, where H is a class of candidates for the bandwidth.

Step 2. For a bandwidth candidate $h \in H$, calculate

$$K(h) = \sum_{i=1}^{n} \left\| r_i \begin{bmatrix} \hat{\boldsymbol{C}}_s(\hat{\boldsymbol{\phi}}_h; \boldsymbol{x}_i)/\hat{f}_1(\boldsymbol{x}_i) - \boldsymbol{s}_0(\hat{\boldsymbol{\phi}}_h; \boldsymbol{x}_i) O(\hat{\boldsymbol{\phi}}_h; \boldsymbol{z}_i) \\ \hat{D}(\hat{\boldsymbol{\phi}}_h; \boldsymbol{x}_i)/\hat{f}_1(\boldsymbol{x}_i) - O(\hat{\boldsymbol{\phi}}_h; \boldsymbol{z}_i) \end{bmatrix} \right\|^2, \tag{2.15}$$

where $\| \cdot \|$ is the Euclidean norm and $\hat{f}_1(\boldsymbol{x}) = \{n(h^{'})^d\}^{-1} \sum_{i=1}^{n} r_i K_{h'}(\boldsymbol{x}_i - \boldsymbol{x})$, where h' is a bandwidth of nonparametric estimator for $f_1(\boldsymbol{x}) = f(\boldsymbol{x} \mid r = 1)$ chosen by such as the Normal reference rule and a cross-validation method (Wasserman, 2005).

Step 3. Choose an optimal bandwidth h^* which minimizes the above $K(h)$.

2.3.1 Asymptotic Properties

Here, asymptotic properties of the proposed estimator are presented. In Theorem 2.3.1 the consistency is established. In Theorem 2.3.2 the asymptotic normality is established. Regularity conditions are presented in Appendix A.

Theorem 2.3.1. Under conditions (C1.1) and (C1.6)–(C1.12) in Appendix A, our estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\theta}$ converge in probability to $\boldsymbol{\phi}_0$ and θ_0, respectively, as $n \to \infty$.

Theorem 2.3.2. Under conditions (C1.1)–(C1.12) in Appendix A, the following results hold:

[1] $$\sqrt{n} \left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 \right) \xrightarrow{\mathcal{L}} N(0, \Sigma_{\phi}) \quad \text{as} \quad n \to \infty,$$

where

$$\begin{aligned}
\Sigma_\phi &= \mathcal{I}_{11}^{-1} E\left\{O(\boldsymbol{\phi}_0)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)^{\otimes 2}\right\}\left(\mathcal{I}_{11}^{-1}\right)^\top, \\
B^{\otimes 2} &= BB^\top, \\
\mathcal{I}_{11} &= -E\{\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)\dot{\boldsymbol{\pi}}^\top(\boldsymbol{\phi}_0)/\pi(\boldsymbol{\phi}_0)\}, \\
\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}) &= E_0\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x},Y)\mid \boldsymbol{x}\}, \\
\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}) &= \partial\pi(\boldsymbol{\phi})/\partial\boldsymbol{\phi}
\end{aligned}$$

and

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{\mathcal{L}} N(0,\Sigma_\theta) \quad \text{as} \quad n\to\infty, \tag{2}$$

where

$$\begin{aligned}
\Sigma_\theta &= \mathcal{I}_{22}^{-1}\left(\mathrm{var}\{U(\theta_0)\} + E[O(\boldsymbol{\phi}_0)\{U(\theta_0)-\bar{U}(\theta_0)+\boldsymbol{\kappa}\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)\}^{\otimes 2}]\right)\left(\mathcal{I}_{22}^{-1}\right)^\top, \\
\bar{U}(\theta) &= E_0\{U(\theta;\boldsymbol{x},Y)\mid\boldsymbol{x}\}, \\
\boldsymbol{\kappa} &= \mathcal{I}_{21}\mathcal{I}_{11}^{-1}, \\
\mathcal{I}_{21} &= E[\{U(\theta_0)-\bar{U}(\theta_0)\}\dot{\boldsymbol{\pi}}^\top(\boldsymbol{\phi}_0)/\pi(\boldsymbol{\phi}_0)], \\
\mathcal{I}_{22} &= -E\{\partial U(\theta_0)/\partial\theta\}.
\end{aligned}$$

Remark 2.3.2. As shown in (B.12) in Appendix B, the asymptotic variance Σ_ϕ can be rewritten as

$$\Sigma_\phi = \mathcal{I}_{11}^{-1}\mathrm{var}\left\{r\left\{\boldsymbol{s}_1(\boldsymbol{\phi}_0;\boldsymbol{z})+\boldsymbol{G}(\boldsymbol{\phi}_0;\boldsymbol{z})\right\}+(1-r)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0;\boldsymbol{x})\right\}\left(\mathcal{I}_{11}^{-1}\right)^\top,$$

where $\boldsymbol{G}(\boldsymbol{z}) = O(\boldsymbol{\phi};\boldsymbol{z})\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{z})-\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x})\right\}$. If $\boldsymbol{G}\equiv\boldsymbol{0}$, the asymptotic variance of $\hat{\boldsymbol{\phi}}$ becomes

$$\Sigma_\phi = \mathcal{I}_{11}^{-1}\mathrm{var}\left\{r\boldsymbol{s}_1(\boldsymbol{\phi}_0;\boldsymbol{z})+(1-r)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0;\boldsymbol{x})\right\}\left(\mathcal{I}_{11}^{-1}\right)^\top$$

which is equivalent to the asymptotic variance when the true $f_1(y\mid\boldsymbol{x})$ is known in advance. Thus, $\boldsymbol{G}(\cdot)$ can be viewed as the additional price we pay due to estimating $f_1(y\mid\boldsymbol{x})$ using nonparametric regression.

For variance estimation, we can use the linearization method of Kim and Yu (2011). As shown in (B.12) in Appendix B, Σ_θ can be written as $\mathrm{var}[\mathcal{I}_{22}^{-1}\{\hat{\mathrm{U}}^l(\boldsymbol{\phi}_0,\theta_0)-\mathcal{I}_{21}\mathcal{I}_{11}^{-1}\hat{\mathbf{S}}^l(\boldsymbol{\phi}_0)\}]$. Therefore, if we could estimate $\mathcal{I}_{11}$, $\mathcal{I}_{21}$, and $\mathcal{I}_{22}$, a consistent estimator of the asymptotic variance Σ_θ would be obtained by

$$\hat{\Sigma}_\theta = n^{-1}\sum_{i=1}^n\hat{\eta}_i^2-\left(n^{-1}\sum_{i=1}^n\hat{\eta}_i\right)^2$$

where $\hat{\eta}_i = \hat{\mathcal{I}}_{22}^{-1}\{\hat{\mathrm{U}}_i^l(\hat{\boldsymbol{\phi}}, \hat{\theta}) - \hat{\mathcal{I}}_{21}\hat{\mathcal{I}}_{11}^{-1}\hat{\mathbf{S}}_i^l(\hat{\boldsymbol{\phi}})\}$ and $\hat{\mathcal{I}}_{11}, \hat{\mathcal{I}}_{21}$, and $\hat{\mathcal{I}}_{22}$ are consistent estimators of $\mathcal{I}_{11}$, $\mathcal{I}_{21}$, $\mathcal{I}_{22}$,

$$\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}) = \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x}_i),$$

and

$$\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}, \theta) = \left[\frac{r_i U(\theta; \boldsymbol{z}_i)}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)} + \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} \bar{U}(\theta; \boldsymbol{x}_i)\right].$$

For example, $\mathcal{I}_{11}$ can be estimated by

$$\hat{\mathcal{I}}_{11} = -n^{-1}\sum_{i=1}^{n} r_i \bar{\boldsymbol{s}}_0(\hat{\boldsymbol{\phi}}; \boldsymbol{x}_i)\frac{\dot{\boldsymbol{\pi}}(\hat{\boldsymbol{\phi}}; \boldsymbol{z}_i)^\top}{\pi^2(\hat{\boldsymbol{\phi}}; \boldsymbol{z}_i)}. \tag{2.16}$$

Instead of the above linearization method, a bootstrap method or the empirical likelihood jackknife method of Zhong and Chen (2014) can be also used.

2.4 Semiparametric adaptive estimators

2.4.1 Efficiency bound

In this section, we provide an optimal estimator for the true parameter $(\boldsymbol{\phi}_0^\top, \theta_0)^\top$ that is the most efficient among all RAL estimators. If the optimal influence function $\boldsymbol{\varphi}_{\text{eff}}$ is found, the semiparametric lower bound is given as $E(\boldsymbol{\varphi}_{\text{eff}}\boldsymbol{\varphi}_{\text{eff}}^\top)$. We begin by presenting the efficient influence function in Lemma 2.4.1. Although θ is a parameter not prescribing the distribution of $[\boldsymbol{x}, y]$ as defined in §2.2, this limitation is just for simplicity and can be removed. For example, Rotnitzky and Robins (1997) derived the semiparametric efficiency bound for regression parameters, which prescribe the first moment of the distribution of $[y \mid \boldsymbol{x}]$. However, ideas used for adaptive estimators expressed in §2.4.2 are still applicable for such parameters.

In the following discussion, we abbreviate the parameter value or random variable, for example, $\pi(\boldsymbol{z}; \boldsymbol{\phi}_0) = \pi(\boldsymbol{z}) = \pi(\boldsymbol{\phi}_0)$, unless this would lead to ambiguity.

Lemma 2.4.1. Let $\boldsymbol{S}_{\text{eff}} = (\boldsymbol{S}_1^\top, \boldsymbol{S}_2)^\top$, where $\boldsymbol{S}_1 = \boldsymbol{S}_1(R, \boldsymbol{G}_R(\boldsymbol{Z}))$ and $\boldsymbol{S}_2 = \boldsymbol{S}_2(R, \boldsymbol{G}_R(\boldsymbol{Z}))$ are defined as

$$\boldsymbol{S}_1(R, \boldsymbol{G}_R(\boldsymbol{Z}); \boldsymbol{\phi}) = \left\{1 - \frac{R}{\pi(\boldsymbol{Z}; \boldsymbol{\phi})}\right\} \boldsymbol{g}^\star(X; \boldsymbol{\phi}_0), \tag{2.17}$$

$$\boldsymbol{S}_2(R, \boldsymbol{G}_R(\boldsymbol{Z}); \boldsymbol{\phi}, \theta) = \frac{R}{\pi(\boldsymbol{Z}; \boldsymbol{\phi})} U(\boldsymbol{Z}; \theta) + \left\{1 - \frac{R}{\pi(\boldsymbol{Z}; \boldsymbol{\phi})}\right\} U^\star(X; \boldsymbol{\phi}_0, \theta)\}, \tag{2.18}$$

$\boldsymbol{g}^\star(\boldsymbol{x};\boldsymbol{\phi}_0) = E^\star\{\boldsymbol{s}_0(\boldsymbol{Z};\boldsymbol{\phi}_0) \mid \boldsymbol{x};\boldsymbol{\phi}_0\}$, $U^\star(\boldsymbol{x};\boldsymbol{\phi}_0,\theta) = E^\star\{U(\boldsymbol{Z};\theta) \mid \boldsymbol{x};\boldsymbol{\phi}_0\}$, and

$$E^\star\{\boldsymbol{g}(\boldsymbol{Z}) \mid \boldsymbol{x};\boldsymbol{\phi}_0\} = \frac{E\{O(\boldsymbol{Z};\boldsymbol{\phi}_0)\boldsymbol{g}(\boldsymbol{Z}) \mid \boldsymbol{x}\}}{E\{O(\boldsymbol{Z};\boldsymbol{\phi}_0) \mid \boldsymbol{x}\}} \tag{2.19}$$

with $O(\boldsymbol{z};\boldsymbol{\phi}_0) = \{1-\pi(\boldsymbol{z};\boldsymbol{\phi}_0)\}/\pi(\boldsymbol{z};\boldsymbol{\phi}_0)$. Then, the efficient influence function is $\boldsymbol{\varphi}_{\text{eff}} = M^{-1}\boldsymbol{S}_{\text{eff}}$, where $M = E(\boldsymbol{S}_{\text{eff}}^{\otimes 2}) = E\left\{\partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}_0,\theta_0)/\partial(\boldsymbol{\phi}^\top,\theta)^\top\right\}$. Therefore, the semiparametric efficiency bound is given by $\{E(\boldsymbol{S}_{\text{eff}}^{\otimes 2})\}^{-1}$. In particular, the asymptotic variance of $\hat{\theta}$ is $V = \text{var}\{\boldsymbol{S}_2(\boldsymbol{\phi}_0,\theta_0) - \boldsymbol{\kappa}\boldsymbol{S}_1(\boldsymbol{\phi}_0)\}$, where $\boldsymbol{\kappa} = E\{(E^\star(U \mid \boldsymbol{x};\boldsymbol{\phi}_0) - U)\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0)^\top/\pi(\boldsymbol{\phi}_0)\}E\{\boldsymbol{g}^\star(\boldsymbol{\phi}_0)\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0)^\top/\pi(\boldsymbol{\phi}_0)\}^{-1}$.

This Lemma 2.4.1 implies that if we can compute $E^\star(\cdot \mid \boldsymbol{x})$, then estimating functions (2.17) and (2.18) will yield an optimal estimator. The optimal estimator will be the solution to

$$\sum_{i=1}^{n} \boldsymbol{S}_{\text{eff},i}(\boldsymbol{\phi},\theta) = \sum_{i=1}^{n} \{\boldsymbol{S}_1^\top(r_i,\boldsymbol{G}_{r_i}(\boldsymbol{z}_i);\boldsymbol{\phi}), \boldsymbol{S}_2(r_i,\boldsymbol{G}_{r_i}(\boldsymbol{z}_i);\boldsymbol{\phi},\theta)\}^\top = \mathbf{0}. \tag{2.20}$$

Because by the regularity conditions defined in Appendix A, we assume M is nonsingular; also note that multiplying a nonsingular matrix does not affect its asymptotic distribution.

Remark 2.4.1. The estimating equation (2.17) can be viewed as a special case of the estimator of Chang and Kott (2008) and Kott and Chang (2010) defined in (2.9). One might think that more information can be included with larger dimension of g because the above two methods can handle models when $q > d+1$ by solving (2.10). However, according to Lemma (2.4.1), there is no need to use more g functions and it is enough to consider only $\boldsymbol{g}^\star(\boldsymbol{x},\boldsymbol{\phi}_0)$ (i.e., $q = d+1$) as the calibration functions.

The equation based on $\boldsymbol{S}_1(\boldsymbol{\phi})$ in (2.17) gives an optimal estimator for $\boldsymbol{\phi}$, say $\hat{\boldsymbol{\phi}}$. Then, by using $\hat{\boldsymbol{\phi}}$, $\boldsymbol{S}_2(\hat{\boldsymbol{\phi}},\theta)$ in (2.18) can provide an optimal estimator for θ. However, the expectation $E^\star(\cdot \mid \boldsymbol{x})$ and the parameter $\boldsymbol{\phi}_0$ are unknown and need to be estimated. Also, to compute the conditional expectation, we may need to correctly specify the distribution of $[y \mid \boldsymbol{x}]$, which is subjective and unverifiable, as is stated in §1.2. Two adaptive estimators are proposed to work around the problem and also attain the lower bound derived in Lemma 2.4.1.

2.4.2 Adaptive Estimators and their asymptotic properties

We now propose two adaptive estimators for $(\boldsymbol{\phi}_0, \theta_0)$: (i) with a parametric working model for $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})$; (ii) with a nonparametric estimator for $f_1(y \mid \boldsymbol{x})$, where $f_1(y \mid \boldsymbol{x}) = f(y \mid \boldsymbol{x}, r = 1)$.

To discuss the first method, let $f_1(y \mid \boldsymbol{x})$ be known up to the parameter $\boldsymbol{\gamma} \in \Gamma$, and let $\hat{\boldsymbol{\gamma}}$ be the maximizer of $\sum_{i=1}^n r_i \log f_1(y_i \mid \boldsymbol{x}_i; \boldsymbol{\gamma})$. This can be easily implemented, and its validity can be checked by using information criteria such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). By using the similar idea to Sverchkov (2008) and Riddles et al. (2016), we can show that, for any function $\boldsymbol{g}(\boldsymbol{z})$,

$$E^{\star}\{\boldsymbol{g}(\boldsymbol{Z}) \mid \boldsymbol{x}; \boldsymbol{\phi}_0, \gamma\} = \frac{E_1\{\pi^{-1}(\boldsymbol{Z}; \boldsymbol{\phi}_0) O(\boldsymbol{Z}; \boldsymbol{\phi}_0) \boldsymbol{g}(\boldsymbol{Z}) \mid \boldsymbol{x}; \gamma\}}{E_1\{\pi^{-1}(\boldsymbol{Z}; \boldsymbol{\phi}_0) O(\boldsymbol{Z}; \boldsymbol{\phi}_0) \mid \boldsymbol{x}; \gamma\}}, \tag{2.21}$$

where $E_1(\cdot \mid \boldsymbol{x}) = E(\cdot \mid \boldsymbol{x}, r = 1)$. Thus, the expectation can be estimated by using the assumed model $f_1(y \mid \boldsymbol{x}; \hat{\gamma})$ and $\pi(\boldsymbol{z}; \boldsymbol{\phi}_0)$. However, $\boldsymbol{\phi}_0$ is unknown, thus we propose an estimating equation $\sum_{i=1}^n \boldsymbol{S}_{\text{eff},i}(\boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}) = 0$, where $\boldsymbol{S}_{\text{eff},i}(\boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}) = \{\boldsymbol{S}_1^{\top}(r_i, \boldsymbol{G}_{r_i}(\boldsymbol{z}_i); \boldsymbol{\phi}, \hat{\boldsymbol{\gamma}}), \boldsymbol{S}_2(r_i, \boldsymbol{G}_{r_i}(\boldsymbol{z}_i); \boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}})\}^{\top}$ with

$$\begin{aligned} \boldsymbol{S}_1(r, \boldsymbol{G}_r(\boldsymbol{z}); \boldsymbol{\phi}; \hat{\boldsymbol{\gamma}}) &= \left\{1 - \frac{r}{\pi(\boldsymbol{z}; \boldsymbol{\phi})}\right\} \boldsymbol{g}^{\star}(\boldsymbol{x}; \boldsymbol{\phi}, \hat{\boldsymbol{\gamma}}), \\ \boldsymbol{S}_2(r, \boldsymbol{G}_r(\boldsymbol{z}); \boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}) &= \frac{r}{\pi(\boldsymbol{z}; \boldsymbol{\phi})} U(\boldsymbol{z}; \theta) + \left\{1 - \frac{r_i}{\pi(\boldsymbol{z}; \boldsymbol{\phi})}\right\} U^{\star}(\boldsymbol{x}; \boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}). \end{aligned} \tag{2.22}$$

How about $f_1(y \mid \boldsymbol{x})$ is misspecified? One would expect the solution to the estimating equation with (2.22) to be inconsistent as a result. However, the estimator that uses the function on the right-hand side of (2.21) is consistent even when the assumed model for $f_1(y \mid \boldsymbol{x})$ is misspecified. Also, if the model is correctly specified, the estimator attains the lower bound. This leads us to Theorem 2.4.1.

Theorem 2.4.1. Let $(\hat{\boldsymbol{\phi}}^{\top}, \hat{\theta})^{\top}$ be the solution to $\sum_{i=1}^n \boldsymbol{S}_{\text{eff},i}(\boldsymbol{\phi}, \theta, \hat{\boldsymbol{\gamma}}) = 0$. Under conditions (C2.1)–(C2.6) given in Appendix A, $(\hat{\boldsymbol{\phi}}^{\top}, \hat{\theta})^{\top}$ has consistency and asymptotic normality with variance

$$E\left\{\frac{\partial \boldsymbol{S}_{\text{eff}}^{*}}{\partial(\boldsymbol{\phi}^{\top}, \theta)}\right\}^{-1} E(\boldsymbol{S}_{\text{eff}}^{*\otimes 2}) E\left\{\frac{\partial \boldsymbol{S}_{\text{eff}}^{*}}{\partial(\boldsymbol{\phi}^{\top}, \theta)}\right\}^{-1,\top},$$

even if $f_1(y \mid \boldsymbol{x}; \hat{\gamma})$ is misspecified, where $\boldsymbol{\gamma}^{*}$ is the limit of $\hat{\gamma}$, and $\boldsymbol{S}_{\text{eff}}^{*} = \{\boldsymbol{S}_1(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^{*})^{\top}, \boldsymbol{S}_2(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^{*})\}^{\top}$ is defined in (2.22). Also, the asymptotic variance of $\hat{\theta}$ is given as

$$V^{*} = \text{var}\{\boldsymbol{S}_2(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^{*}) - \boldsymbol{\kappa}^{*} s_1(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^{*})\}/\sigma_{\text{U}}^2, \tag{2.23}$$

where

$$\begin{aligned}
\boldsymbol{\kappa}^* &= \boldsymbol{\kappa}_1^*(\boldsymbol{\kappa}_2^*)^{-1},\\
\boldsymbol{\kappa}_1^* &= E[\{U^\star(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*) - U(\theta_0)\}\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0)^\top/\pi(\boldsymbol{\phi}_0)\}],\\
\boldsymbol{\kappa}_2^* &= E\{\boldsymbol{g}^\star(\boldsymbol{\phi}_0, \boldsymbol{\gamma}^*)\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0)^\top/\pi(\boldsymbol{\phi}_0)\},\\
\sigma_{\mathrm{U}} &= E\{\partial U(\theta_0)/\partial\theta\}.
\end{aligned}$$

In addition, if the model is correctly specified, the estimator attains the semiparametric efficiency bound.

Remark 2.4.2. There are two properties better than the estimator of Riddles et al. (2016). First, the parametric model f_1 is irrelevant to the consistency and asymptotic normality of the estimator here. Therefore, we call f_1 a working model, as in Liang and Zeger (1986). Second, our proposed estimator can attain the semiparametric efficiency bound derived in Lemma 2.4.1. However, though equation (2.18) has a form similar to that of the doubly robust estimator under MAR (Robins et al., 1994), unfortunately, our estimator does not have the doubly robustness property. This is because the computation for $E^\star(\cdot \mid \boldsymbol{x})$ relies on the correct response mechanism.

Numerical computation is needed to calculate the conditional expectation in (2.21). The expectation-maximization (EM) algorithm considered in Riddles et al. (2016) can be used with a minor modification. we can directly apply their method, once the weights w_{ij}^* defined in (15) of Riddles et al. (2016) are changed to

$$w_{ij}^* = \frac{r_j\pi^{-1}(\boldsymbol{x}_i, y_j; \boldsymbol{\phi})O(\boldsymbol{x}_i, y_j; \boldsymbol{\phi})f_1(y_j \mid \boldsymbol{x}_i; \boldsymbol{\gamma})/C(y_j; \boldsymbol{\gamma})}{\sum_{k=1}^n r_k\pi^{-1}(\boldsymbol{x}_i, y_k; \boldsymbol{\phi})O(\boldsymbol{x}_i, y_k; \boldsymbol{\phi})f_1(y_k \mid \boldsymbol{x}_i; \boldsymbol{\gamma})/C(y_k; \boldsymbol{\gamma})},$$

where $C(y; \boldsymbol{\gamma}) = \sum_{l=1}^n r_l f_1(y \mid \boldsymbol{x}_l; \boldsymbol{\gamma})$. The weight w_{ij}^* can be called fractional weights in the context of fractional imputation (Kim, 2011). With these weights, $E^\star\{\boldsymbol{g}(\boldsymbol{x}_i, Y) \mid \boldsymbol{x}_i; \boldsymbol{\gamma}, \boldsymbol{\phi}\}$ can be computed by $\sum_{j=1}^n w_{ij}^*\boldsymbol{g}(\boldsymbol{x}_i, y_j)$.

We now discuss the second adaptive estimation method based on nonparametric estimation for $f_1(y \mid \boldsymbol{x})$. Generally speaking, directly computing the expectation defined in (2.19) with nonparametrically estimated $f_1(y \mid \boldsymbol{x})$ would make the computation difficult. To avoid this problem, we consider methods of calculating the expectation directly. When x is discrete, such as when $\boldsymbol{x}$ is a dichotomous variable, the expectation can be computed by averaging the data conditioned by $X = x$ and $R = 1$, e.g., for

$x = 0, 1$,

$$\hat{E}^{\star}\{\boldsymbol{g}(x, Y) \mid x; \boldsymbol{\phi}\} = \frac{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \boldsymbol{\phi}) O(x, y_j; \boldsymbol{\phi}) \boldsymbol{g}(x, y_j)}{\sum_{j \in I_x} r_j \pi^{-1}(x, y_j; \boldsymbol{\phi}) O(x, y_j; \boldsymbol{\phi})} \tag{2.24}$$

is a consistent estimator of (2.21), where $I_x = \{j \in \{1, \ldots, n\} \mid X_j = x\}$.

When $\boldsymbol{x}$ is continuous, the Nadaraya-Watson estimator can be employed. That is, for any function $\boldsymbol{g}(\boldsymbol{z})$,

$$\hat{E}^{\star}\{\boldsymbol{g}(\boldsymbol{x}, Y) \mid \boldsymbol{x}; \boldsymbol{\phi}\} = \frac{\sum_{j=1}^{n} K_h(\boldsymbol{x} - x_j) r_j \pi^{-1}(\boldsymbol{x}, y_j; \boldsymbol{\phi}) O(\boldsymbol{x}, y_j; \boldsymbol{\phi}) \boldsymbol{g}(\boldsymbol{x}, y_j)}{\sum_{j=1}^{n} K_h(\boldsymbol{x} - x_j) r_j \pi^{-1}(\boldsymbol{x}, y_j; \boldsymbol{\phi}) O(\boldsymbol{x}, y_j; \boldsymbol{\phi})} \tag{2.25}$$

is consistent under the regularity conditions given in §S1 in the Supplementary Material. Here,

$$K_h(\boldsymbol{x} - \boldsymbol{w}) = K\left(\frac{\boldsymbol{x} - \boldsymbol{w}}{h}\right) = \prod_{k=1}^{d} K\left(\frac{x_k - w_k}{h}\right),$$

where K is a kernel function, and h is the bandwidth. We have the following result for the adaptive estimators obtained with the Nadaraya-Watson estimation.

Theorem 2.4.2. Let $(\hat{\boldsymbol{\phi}}^{\top}, \hat{\theta})^{\top}$ be the solution to $\sum_{i=1}^{n} \hat{\boldsymbol{S}}_{\text{eff},i}(\boldsymbol{\phi}, \theta) = 0$, where $\hat{\boldsymbol{S}}_{\text{eff},i}(\boldsymbol{\phi}, \theta)$ is defined in (2.20) with the estimated conditional expectation (2.25). Under Conditions (C2.1), (C2.2), and (C2.7)–(C2.14) given in Appendix A, the estimator attains the semiparametric efficiency bound.

The proposed estimator is attractive because it does not need any model assumptions on f_1, but it would not work well when the dimension of $\boldsymbol{x}$ is high, as is common in any nonparametric estimation.

Estimating variance is also a difficult problem in semiparametric estimation. When we consider a parametric working model $f_1(y \mid \boldsymbol{x})$,

$$\hat{\boldsymbol{V}} = n^{-1} \sum_{i=1}^{n} \{\boldsymbol{S}_2(r_i, \boldsymbol{G}_{r_i}(\boldsymbol{z}_i); \hat{\boldsymbol{\phi}}, \hat{\theta}, \hat{\boldsymbol{\gamma}}) - \hat{\boldsymbol{\kappa}} \boldsymbol{S}_1(r_i, \boldsymbol{G}_{r_i}(\boldsymbol{z}_i); \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\gamma}})\}^{\otimes 2} / \hat{\sigma}_{\mathrm{U}}^2 \tag{2.26}$$

converges to V^* in probability as defined in (2.23), where $\hat{\sigma}_{\mathrm{U}}$ and $\hat{\boldsymbol{\kappa}}$ are consistent estimators for σ_{U} and $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_1^*(\boldsymbol{\kappa}_2^*)^{-1}$, respectively, for $\boldsymbol{\kappa}_1^*$ and $\boldsymbol{\kappa}_2^*$ as defined in Theorem 2.4.1. To estimate $\boldsymbol{\kappa}_1^*$, we propose using the same method that we used to compute θ_0, i.e., let $\mathcal{U}(\boldsymbol{\phi}_0, k_1, \boldsymbol{\gamma}^*) = k_1 - (U^{\star}(\boldsymbol{\gamma}^*) - U)\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0)^{\top} / \pi(\boldsymbol{\phi}_0)$ be our new U-function and let the solution to $E\{\mathcal{U}(\boldsymbol{\phi}_0, k_1, \boldsymbol{\gamma}^*)\} = 0$ with respect to k_1 be our target parameter; solve the following equation:

$$\sum_{i=1}^{n}\left[\frac{r_i}{\pi(\boldsymbol{z}_i;\hat{\boldsymbol{\phi}})}\mathcal{U}(\boldsymbol{z}_i;\hat{\boldsymbol{\phi}},k_1,\hat{\boldsymbol{\gamma}}) + \left\{1-\frac{r_i}{\pi(\boldsymbol{z}_i;\hat{\boldsymbol{\phi}})}\right\}E^{\star}\{\mathcal{U}(\boldsymbol{Z};\hat{\boldsymbol{\phi}},k_1,\hat{\boldsymbol{\gamma}}) \mid \boldsymbol{x}_i;\hat{\boldsymbol{\gamma}}\}\right] = 0.$$

This is the optimal estimator for $(\boldsymbol{\phi}_0, \boldsymbol{\kappa}_1^*)$ in terms of the asymptotic variance, because $\mathcal{U}$ is a known function and Theorem 2.4.1 is applicable. The best estimator for $\boldsymbol{\kappa}_2^*$ can be obtained in the same way. When we use the nonparametric method stated in Theorem 2.4.2 to estimate θ_0, the variance can be also estimated by using the nonparametric method (2.24), instead of assuming a parametric model $f_1(y \mid \boldsymbol{x};\boldsymbol{\gamma})$ in (2.26).

2.5 Theoretical comparison among the semiparametric estimators

In this section, we show when $q = p_1 + 1 = p_2$, if we choose specific $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$ functions, the two estimators are exactly the same. Recall that q, p_1, and p_2 are the dimension of $\boldsymbol{\phi}$, $\boldsymbol{h}(\boldsymbol{x})$, and $\boldsymbol{g}(\boldsymbol{x})$.

Theorem 2.5.1. When $q = p_1 + 1 = p_2$, both EL and GMM estimators are exactly same if and only if $\boldsymbol{g}(\boldsymbol{x}) = \{1, \boldsymbol{h}(\boldsymbol{x})\}^\top$.

Consider that we now try to estimate $\boldsymbol{\phi}$ by using a method of Qin et al. (2002) with an $\boldsymbol{h}(\boldsymbol{x})$ function. Theorem 2.5.1 implies that, when $q = p_1 + 1$, there is no reason to use the procedure for EL, and just use the GMM estimating equation (2.9) with $\boldsymbol{g}(\boldsymbol{x}) = \{1, \boldsymbol{h}(\boldsymbol{x})^\top\}^\top$, which is much simpler in terms of computation.

Furthermore, we can see that the GMM estimator including $\boldsymbol{g}^\star(\boldsymbol{x})$ as the constraint also attains the semiparametric efficiency bound.

Theorem 2.5.2. When $q < p_1 + 1 = p_2$, the GMM estimator attains the semiparametric efficiency bound if $\boldsymbol{g}(\boldsymbol{x})$ defined in (2.10) contains $\boldsymbol{g}^\star(\boldsymbol{x})$, i.e., $\boldsymbol{g}(\boldsymbol{x}) = \{\boldsymbol{k}(\boldsymbol{x})^\top, \boldsymbol{g}^\star(\boldsymbol{x})^\top\}^\top$, where $\boldsymbol{k} : \mathbb{R}^d \to \mathbb{R}^\kappa$ is an arbitrary function of $\boldsymbol{x}$, and $\kappa \in \mathbb{R}^1$ is a positive integer.

This estimator enjoys two properties, (i) robustness for misspecification of the response model; (ii) semiparametric efficiency. The robustness is achieved because of the constraints, which is proposed by Qin and Zhang (2007) and Imai and Ratkovic (2014) for observational studies in the case of MAR. Our study can be considered as an extension of their works to NMAR.

Chapter 3

Identification for models with repeated measurement data

3.1 Introduction

In clinical studies, researchers generally hope to obtain complete data but this does not always happen. In fact, subjects are typically told that they can drop out anytime they want, in accordance with human participant protection protocols. There are thus many dropouts in some experiments. From a statistical perspective, missing values complicate the data analysis, because ignoring the response mechanism can lead to inappropriate inference.

For example, Machin et al. (1988) report results of a comparative trial of two dosages of depot medroxyprogesterone acetate (DMPA, 100 mg and 150 mg) where over 40% of subjects are missing at the endpoint. In the DMPA trial test, 1151 female subjects were divided into two dosages randomly and took DMPA in every quarter, over one year. They reported the results of DMPA as binary data: if a subject experienced amenorrhea, i.e., absence of a menstrual period in a woman of reproductive age, it was coded as 1; otherwise, it was coded 0. Whether a subject experienced amenorrhea or not was based on her menstrual diary. Each subject thus generated a sequence according to whether or not she experienced amenorrhea in the successive reference periods. The number of women with each sequence is shown in Table 3.1, where "×" means missingness. For example, "0 1 × ×" means amenorrhea is absent in first period, but present the next period, and the data were obtained for the third and fourth periods.

These data have been analyzed by several authors using various approaches; see, e.g., Birminghama et al. (2003); Matsuyama (2004); Wilkins and Fitzmaurice (2006). We will analyze the same data with a new approach.

In the analysis of repeated measure data, serial correlations among response variables in $\boldsymbol{Y}_i = (Y_{i1}, , \ldots , Y_{iT})$ may not be ignored and any statistical model for $\boldsymbol{Y}_i$ has to take the correlations into account. Serial correlation is typically incorporated into the model either via a conditional or a marginal approach. Conditional models describe the serial correlation by modeling Y_t, which is the response at time t, given not only covariates X but also $Y_1, , \ldots , Y_{t-1}$, which are responses recorded earlier than time t. The approach is intuitive and simple, and the serial correlations are obtained easily from the conditional model (Molenberghs and Verbeke, 2005).

While the likelihoods of conditional models are often simple, only the direct effects of covariates on the responses are easily obtained since response variables are conditioned on the past data. Total effects require complicated calculations. Marginal models and hybrid models easily yield total effects of covariates to the response variables, which are often most interesting. However, the likelihoods of marginal models and hybrid models can be unwieldy. In addition, the parameter space may be restricted by equality constraints. For example, in Molenberghs et al. (1997), equality constraints are made on the coefficients of the response mechanism over time. This requires that the probability of missingness be invariant throughout the experiment, which is an unnatural assumption since, in many cases, subjects who stay to the end are likely to influence the response variables more than those who drop out early.

In contrast, there exist several models for analyzing categorical data with non-ignorable missingness. For example, Fay (1986), Baker and Laird (1988), and Park and Brown (1994) used log-linear models. Marginal models are developed to describe the serial correlations by modeling various moments of $\boldsymbol{Y}$ given covariates $\boldsymbol{X}$, e.g., $\mathrm{E}(Y_1 \mid X)$, $\mathrm{E}(Y_1 Y_2 \mid X)$ by adopting a fully parametric approach or by modeling a limited number of lower-order moments only, versus modeling conditionally on the responses at previous times (Fitzmaurice and Laird, 1993; Molenberghs and Lesaffre, 1994; Molenberghs et al., 1997; Molenberghs and Verbeke, 2005). Furthermore, more elaborate models have recently been proposed, e.g., hybrid models, which retain advantageous features of the selection and pattern-mixture model approaches simultaneously (Wilkins and Fitzmaurice, 2006; Yuan and Little, 2009).

Nevertheless, model unidentifiability remains an essential issue (Fitzmaurice et al., 1995; Matsuyama, 2004). If a model is not identifiable, then any statistical inference using the model is distorted and asymptotic properties such as consistency and asymptotic normality may fail. Sufficient conditions have been studied for nonignorable nonresponse when $t = 1$. For example, Miao et al. (2016) explored identifiable parameters when an outcome variable is continuous such as Gaussian, Student, or a normal mixture. Wang et al. (2014) also found that nonresponse instrumental variables can make a semiparametric model identifiable when the response model is parametric and the outcome model is unspecified.

In this chapter, we mainly focus on the identification problem for models with discrete response variables. Main results with a response and an outcome models, both are parametric, are presented in §3.2 and §3.4. In §3.4, the results in §3.2 are extended for semiparametric models where either of the mechanism or the outcome model is unspecified. Some technical developments are relegated to Appendix B.

3.2 Fully parametric models

Assume that $\boldsymbol{Y}$ is binary, which is perhaps the most challenging case for identifiability given the poor level of information in $\boldsymbol{Y}$. We further assume throughout that the missing pattern type is drop-out.

Let $\boldsymbol{R} = (R_1, \ldots, R_T)$ be a vector of indicator variables R_t taking value 1 (0) if Y_t is observed (missing). First, we consider the identification problem when both the response and outcome models are parametric. We then extend our study to the case where one of them is nonparametric.

Suppose that $(\boldsymbol{R}, \boldsymbol{Y})$ has distribution g and, for each $t \in \{1, \ldots, T\}$. Write $\boldsymbol{R} = \boldsymbol{r}^{(t)}$ when the value 1 occurs exactly t times in this vector of indicators. Denote the observed and missing parts of $\boldsymbol{Y}$ by $\boldsymbol{Y}^{(t)}$ and $\boldsymbol{Y}^{(-t)}$, respectively. Note that $\boldsymbol{Y} = (\boldsymbol{Y}^{(t)\top}, \boldsymbol{Y}^{(-t)\top})^\top$ always holds. We also assume $\Pr(R_1 = 1) = 1$. Note that $(Y_2, , \ldots, Y_T)$ is subject to missingness and that there are no covariates. We will consider the case with covariates later.

To model the relation between $\boldsymbol{Y}$ and the missing indicator $\boldsymbol{R}$, we can introduce a limited number of parameters because of the poor information in $\boldsymbol{Y}$. For example, in

Table 3.1: Results of DMPA Trial

Time	Amenorrhea sequence		DMPA (mg)		Time	Amenorrhea sequence		DMPA (mg)	
			100	150					
1	0× × ×		76	68	4	0 0 0 0		142	119
	1× × ×		23	31		0 0 0 1		49	36
		Total	99	99		0 0 1 0		14	26
2	0 0××		43	39		0 0 1 1		41	44
	0 1××		14	27		0 1 0 0		7	4
	1 0××		3	6		0 1 0 1		8	12
	1 1××		8	15		0 1 1 0		4	7
		Total	68	87		0 1 1 1		32	48
3	0 0 0×		20	11		1 0 0 0		6	3
	0 0 1×		13	10		1 0 0 1		7	6
	0 1 0×		1	0		1 0 1 0		0	2
	0 1 1×		5	6		1 0 1 1		10	12
	1 0 0×		2	1		1 1 0 0		4	1
	1 0 1×		2	1		1 1 0 1		4	3
	1 1 0×		0	1		1 1 1 0		3	2
	1 1 1×		5	6		1 1 1 1		30	28
		Total	48	36			Total	361	353

Table 3.1, the total number of cells is $2 + 4 + 8 + 16 = 30$; thus, we can use at most 29 parameters. If we let T be the endpoint of the experiment, then we can use at most

$$\sum_{t=1}^{T} 2^t - 1 = 2^{T+1} - 3 \tag{3.1}$$

parameters. Because this condition is necessary but not sufficient, there could exist many unidentifiable models having at most (3.1) parameters.

3.2.1 Logistic AR(1) model and its identification

Suppose that a model is such that $\Pr(y_t = 1 \mid \boldsymbol{v}_{t-1}) = \Pr(y_t = 1 \mid \boldsymbol{v}^p_{t-1})$, where, for each $t \in \{1, \ldots, T\}$,

$$\boldsymbol{v}^p_t = \begin{cases} (y_{t-p}, \ldots, y_t) & \text{if } t - p \geq 1, \\ (y_1, \ldots, y_t) & \text{otherwise,} \end{cases}$$

and $\boldsymbol{v}_t = (Y_1, \ldots, Y_t)^\top = \boldsymbol{v}_t^{t-1}$. In this so-called AR($p$) model, it is assumed that Y_t depends on its past own values from the last p periods or less. Also let the response model be

$$\operatorname{logit} \Pr(R_t = 1 \mid R_{t-1} = 1,\, y_{t-1}, y_t; \phi_{t,0}, \phi_{t,t-1}, \phi_{t,t}) = \phi_{t,0} + \phi_{t,t-1} y_{t-1} + \phi_{t,t} y_t \tag{3.2}$$

and the marginal distribution of $\boldsymbol{Y}$ be expressed in the form

$$\Pr(Y_1 = 1; \beta_1) = \beta_1, \quad \operatorname{logit} \Pr(y_t = 1 \mid y_{t-1}; \beta_{t,0}, \beta_{t,t-1}) = \beta_{t,0} + \beta_{t,t-1} y_{t-1}, \tag{3.3}$$

where $\beta_{t,0}$ and $\phi_{t,0}$ are each an intercept in the model, while $\beta_{t,t-1}$, $\phi_{t,t-1}$, and $\phi_{t,t}$ are coefficients in the model. We call this model the Logistic AR(1) model. Note that when $\phi_{t,t-1} = \phi_{t,t} = 0$ for $t \in \{2, \ldots, T\}$, the response mechanism is missing completely at random (MCAR); when $\phi_{t,t} = 0$ for $t \in \{2, , \ldots, T\}$ and there exists t such that $\phi_{t,t-1} \neq 0$, the mechanism is missing at random (MAR); when there exists $t \in \{2, \ldots, T\}$ such that $\phi_{t,t} \neq 0$, it is not missing at random (NMAR). In the Logistic AR(1) model, there are one parameter β_1 and five parameters ($\boldsymbol{\xi}_t = (\beta_{t,0}, \beta_{t,t-1}, \phi_{t,0}, \phi_{t,t-1}, \phi_{t,t}) = (\boldsymbol{\beta}_t, \boldsymbol{\phi}_t)$) at each time $t \in \{2, \ldots, T\}$. Thus, the total number of parameters is

$$1 + 5(T - 1) = 5T - 4. \tag{3.4}$$

The relation between (3.1) and (3.4) is $2^{T+1} - 3 < 5T - 4$ if $T = 2$ and $2^{T+1} - 3 > 5T - 4$ when $T \geq 3$. Therefore, if $T = 2$, then the model is not identified, and if $T \geq 3$, then the model meets the necessary condition. As we will see in a later section, however, identifiability does not hold for every T. To see this, we must first define identifiability precisely. In the following discussion, we use "$\otimes$" as the direct product, e.g., $\{0, 1\}^{\otimes 2} = \{0, 1\} \otimes \{0, 1\} = \{\{0, 0\}, \{0, 1\}, \{1, 0\}, \{1, 1\}\}$.

3.2.2 Theoretical results

Let Ξ be a parameter space, $\boldsymbol{\xi}^* \in \Xi$ be a true value of the model, and $P_{\boldsymbol{\xi}^*}$ be a probability measure of a probability function of complete data $(\boldsymbol{R}, \boldsymbol{Y})$ prescribed by a true parameter $\boldsymbol{\xi}^*$, and denote a probability function of observed data $(\boldsymbol{R}, \boldsymbol{Y}^{(t)})$ derived from g by g_t for each $t \in \{1, \ldots, T\}$, which is represented as

$$g_t(\boldsymbol{r}^{(t)}, \boldsymbol{y}_i^{(t)}\,; \boldsymbol{\xi}) = \begin{cases} \sum_{\boldsymbol{y}^{(-t)} \in \{0,1\}^{\otimes (T-t)}} g(\boldsymbol{r}^{(t)}, \boldsymbol{y}_i^{(t)}, \boldsymbol{y}^{(-t)}\,; \boldsymbol{\xi}) & \text{if } t \in \{1, \ldots, T-1\}, \\ g(\boldsymbol{r}^{(T)}, \boldsymbol{y}_i\,; \boldsymbol{\xi}) & \text{if } t = T. \end{cases}$$

Then, a parametric model g is said to be identifiable if

$$\forall_{t\in\{1,\ldots,T\}} \quad g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)}\ ;\ \boldsymbol{\xi}) = g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)}\ ;\ \boldsymbol{\xi}^*) \text{ a.s. } P_{\boldsymbol{\xi}^*} \quad \Rightarrow \quad \boldsymbol{\xi} = \boldsymbol{\xi}^*. \tag{3.5}$$

Since $\boldsymbol{r}^{(t)}$ and $\boldsymbol{y}^{(t)}$ are binary random vectors, the condition in (3.5) is equivalent to

$$\forall_{\boldsymbol{y}^{(t)}} \quad g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi}) = g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi}^*) \quad \text{or} \quad g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi}^*) = 0 \quad \Rightarrow \quad \boldsymbol{\xi} = \boldsymbol{\xi}^*.$$

Here the likelihood involving $\boldsymbol{R}$ will be called the full information maximum likelihood (FIML) and written as $L_n(\boldsymbol{\xi})$, where n is the sample size. Then, the FIML can be written as

$$L_n(\boldsymbol{\xi}) = \prod_{i=1}^{n}\sum_{t=1}^{T} \mathbf{1}_{\{\boldsymbol{R}_i=\boldsymbol{r}^{(t)}\}} g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}_i^{(t)}\ ;\ \boldsymbol{\xi}).$$

Let $L(\boldsymbol{\xi})$ be a function which is the limit of the log-likelihood $n^{-1}\ln\{L_n(\boldsymbol{\xi})\}$ as $n \to \infty$. Such a function exists by the Strong Law of Large Numbers, and we have

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\ln\{L_n(\boldsymbol{\xi})\} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\ln\left\{\sum_{t=1}^{T}\mathbf{1}_{\{\boldsymbol{R}_i=\boldsymbol{r}^{(t)}\}} g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}_i^{(t)};\ \boldsymbol{\xi})\right\} \\
&= \mathrm{E}_{\boldsymbol{\xi}^*}[\ln\{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi})\}] \ (\equiv L(\boldsymbol{\xi})) \text{ a.s. } P_{\boldsymbol{\xi}^*} \\
&= \sum_{t=1}^{T}\sum_{\boldsymbol{y}\in\{0,\,1\}^{\otimes T}} \ln\{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi})\} g(\boldsymbol{r}^{(t)},\, \boldsymbol{y}\ ;\ \boldsymbol{\xi}^*) \text{ a.s. } P_{\boldsymbol{\xi}^*},
\end{aligned}$$

where $\mathrm{E}_{\boldsymbol{\xi}^*}$ represents the expectation under the probability measure $P_{\boldsymbol{\xi}^*}$. The following important relationship is proved in the Appendix:

$$\forall_{\varepsilon>0} \quad \sup_{\boldsymbol{\xi}\in\Xi_\varepsilon} L(\boldsymbol{\xi}) < L(\boldsymbol{\xi}^*). \tag{3.6}$$

It ensures the consistency of the MLE under two additional assumptions, namely compactness of Ξ and continuity of $L(\xi)$, where $\boldsymbol{\Xi}_\varepsilon = \{\boldsymbol{\xi} \in \Xi : |\boldsymbol{\xi} - \boldsymbol{\xi}^*| \geq \varepsilon\}$. For these reasons, we focus on whether (3.5) holds or not in the following discussion.

We assume an AR(p) model for the distribution of $\boldsymbol{Y}$. By transforming $g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi})$ appropriately, we can obtain the next result.

Lemma 3.2.1. For each $t \in \{1, \ldots, T\}$, the joint probability function of $(\boldsymbol{R},\, \boldsymbol{Y}^{(t)})$ can be expressed in the form

$$g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\ \boldsymbol{\xi}) = \begin{cases} \pi_1(y_1;\ \boldsymbol{\xi}_1)\displaystyle\prod_{s=2}^{t}\pi_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_s) \times \left\{1 - \sum_{y_{t+1}}\pi_{t+1}(\boldsymbol{v}_{t+1};\ \boldsymbol{\xi}_{t+1})\right\} & \text{if } t \leq T-1, \\ \pi_1(y_1;\ \boldsymbol{\xi}_1)\prod_{s=2}^{T}\pi_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_s) & \text{if } t = T, \end{cases}$$

where $\boldsymbol{\xi}_t = (\boldsymbol{\beta}_t^\top, \boldsymbol{\phi}_t{}^\top)^\top$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1{}^\top, \ldots, \boldsymbol{\xi}_T^\top)^\top$, $\boldsymbol{v}_t = \boldsymbol{v}_t^{t-1}$,

$$\pi_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \begin{cases} \Pr(Y_1 = y_1;\ \boldsymbol{\beta}_1) & \text{if } t = 1, \\ \Pr\left(R_t = 1 \mid R_{t-1} = 1, \boldsymbol{v}_t;\ \boldsymbol{\phi}_t\right) \times f_t\left(Y_t = y_t \mid \boldsymbol{v}_{t-1};\ \boldsymbol{\beta}_t\right) & \text{if } t \in \{2, \ldots, T\}, \end{cases}$$

and f_t is the probability function of $(Y_t \mid \boldsymbol{V}_{t-1})$. In particular, for the AR(p) model, one gets

$$g_t(\boldsymbol{y}^{(t)},\ \boldsymbol{r}^{(t)};\ \boldsymbol{\xi}) = \begin{cases} \pi_1(y_1;\ \boldsymbol{\xi}_1) \displaystyle\prod_{s=2}^{t} \pi_s(\boldsymbol{v}_s^p;\ \boldsymbol{\xi}_s) \left\{ 1 - \sum_{y_{t+1}} \pi_{t+1}(\boldsymbol{v}_{t+1}^p;\ \boldsymbol{\xi}_{t+1}) \right\} & \text{if } t \leq T-1, \\ \pi_1(y_1;\ \boldsymbol{\xi}_1) \prod_{s=2}^{T} \pi_s(\boldsymbol{v}_s^p;\ \boldsymbol{\xi}_s) & \text{if } t = T. \end{cases}$$

Note that this lemma holds even if the response variable is continuous, in which case the probability function $\Pr(Y_t = y_t \mid \boldsymbol{v}_{t-1}; \boldsymbol{\beta}_t)$ in the definition of $\pi_t(\boldsymbol{v}_t; \boldsymbol{\xi}_t)$ is just replaced by the density function $f_t(y_t \mid \boldsymbol{v}_{t-1}; \boldsymbol{\beta}_t)$.

Using this result, the likelihood $L_n(\boldsymbol{\xi})$ is represented by functions of $\pi_1, \ldots, \pi_T$. The next theorem follows from Lemma 1

Theorem 3.2.1. The slightly improved identification condition in (3.5) is equivalent to

$$\forall_t \quad \pi_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \pi_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t^*) \quad \text{a.s. } P_{\boldsymbol{\xi}^*} \quad \Rightarrow \quad \boldsymbol{\xi}_t = \boldsymbol{\xi}_t^*. \tag{3.7}$$

In particular, for the AR(p) model, it becomes

$$\forall_t \quad \pi_t(\boldsymbol{v}_t^p;\ \boldsymbol{\xi}_t) = \pi_t(\boldsymbol{v}_t^p;\ \boldsymbol{\xi}_t^*) \quad \text{a.s. } P_{\boldsymbol{\xi}^*} \quad \Rightarrow \quad \boldsymbol{\xi}_t = \boldsymbol{\xi}_t^*. \tag{3.8}$$

The slightly improved condition from the original one is a simultaneous equation for $\boldsymbol{\xi}$, so this is still cumbersome. In contrast, for each t, the condition in (3.7) is a simple equation in $\boldsymbol{\xi}_t$ and the condition can be examined for each t individually, so it is relatively easy to check. Therefore, all we have to do is to check Condition (3.7) for each $t \in \{2, \ldots, T\}$. It can be also considered that the transition models are divided into $T-1$ nonresponse problems. Then π_t is just a joint probability for a nonresponse problem in which subjects are restricted to only those who are observed until at least the tth period.

When $p = 1$, $t = 2$, the joint probability can be written as

$$\pi_2(\boldsymbol{v}_2;\ \boldsymbol{\xi}_2) = \Pr(R_2 = 1 \mid R_1 = 1,\ y_1, y_2;\ \boldsymbol{\phi}_2) \Pr(Y_2 = y_2 \mid y_1;\ \boldsymbol{\beta}_2).$$

This is a nonresponse problem at $t = 2$. The condition is equivalent to stating that $(\boldsymbol{\beta}_2, \boldsymbol{\phi}_2) = (\boldsymbol{\beta}_2^*, \boldsymbol{\phi}_2^*)$ when

$$\begin{aligned}\forall_{(y_1,\,y_2)\in(0,1)^{\otimes 2}} \quad &\Pr(R_2 = 1 \mid R_1 = 1,\, y_1,\, y_2\,;\, \boldsymbol{\phi}_2) \Pr(Y_2 = y_2 \mid y_1\,;\, \boldsymbol{\beta}_2) \\ &= \Pr(R_2 = 1 \mid R_1 = 1,\, y_1,\, y_2\,;\, \boldsymbol{\phi}_2^*) \Pr(Y_2 = y_2 \mid y_1\,;\, \boldsymbol{\beta}_2^*). \quad (3.9)\end{aligned}$$

There are four constraints in (3.9) and if all of them were linear equations, the number of parameters would have to be at most 4. Since we have five parameters in the Logistic AR(1) model, this would imply that the model is not identified. In binary data analysis, the constraints are usually non-linear, but it is worth verifying this condition, i.e., stating in general terms whether $\dim(\boldsymbol{\xi}_t)$, representing the number of parameters used at time t, is less than or equal to $2^{\dim(\boldsymbol{v}_t^p)}$, representing the number of constraints. Note that generally we have to check all the above expressions rigorously.

3.3 Logistic AR(p) models

In the Logistic AR(1) model defined in (3.2)–(3.3), there are five parameters at each time $t \geq 2$. From the previous discussion, it seems that the Logistic AR(1) model is not identified, and in fact, it is not.

Proposition 3.3.1. The logistic AR(1) model defined in (3.2)–(3.3) is not identified.

Proof. For simplicity, we write

$$\begin{aligned}(a_{2,0}, a_{2,1}, b_{2,0}, b_{2,1}, b_{2,2}) &= (\exp(\beta_{2,0}), \exp(\beta_{2,1}), \exp(-\phi_{2,0}), \exp(-\phi_{2,1}), \exp(-\phi_{2,2})), \\ (a_{2,0}^*, a_{2,1}^*, b_{2,0}^*, b_{2,1}^*, b_{2,2}^*) &= (\exp(\beta_{2,0}^*), \exp(\beta_{2,1}^*), \exp(-\phi_{2,0}^*), \exp(-\phi_{2,1}^*), \exp(-\phi_{2,2}^*)),\end{aligned}$$

and prove it only for $t = 2$, i.e., show that for $(a_{2,0}^*, a_{2,1}^*, b_{2,0}^*, b_{2,1}^*, b_{2,2}^*)$, there exists $(a_{2,0}, a_{2,1}, b_{2,0}, b_{2,1}, b_{2,2})$ such that

$$(a_{2,0}, a_{2,1}, b_{2,0}, b_{2,1}, b_{2,2}) \neq (a_{2,0}^*, a_{2,1}^*, b_{2,0}^*, b_{2,1}^*, b_{2,2}^*)$$

and (3.9) holds. To show this, we fix $a_{2,0}$ at some value (say, $\tilde{a}_{2,0}$) which is not $a_{2,0}^*$ and prove that (3.9) holds if and only if the rest of the parameters $(a_{2,1}, b_{2,0}, b_{2,1}, b_{2,2})$ can be written as functions of $(\tilde{a}_{2,0}, a_{2,0}^*,\ a_{2,1}^*, b_{2,0}^*, b_{2,1}^*, b_{2,2}^*)$ only, which shows that the Logistic

AR(1) model is not identified. In the Logistic AR(1) model, (3.9) is represented as

$$\begin{cases} (1+a_{2,0})(1+b_{2,0}) = (1+a^*_{2,0})(1+b^*_{2,0}), & (3.10)\\ \left(1+\dfrac{1}{a_{2,0}}\right)(1+b_{2,0}b_{2,2}) = \left(1+\dfrac{1}{a^*_{2,0}}\right)(1+b^*_{2,0}b^*_{2,2}), & (3.11)\\ (1+a_{2,0}a_{2,1})(1+b_{2,0}b_{2,1}) = (1+a^*_{2,0}a^*_{2,1})(1+b^*_{2,0}b^*_{2,1}), & (3.12)\\ \left(1+\dfrac{1}{a_{2,0}a_{2,1}}\right)(1+b_{2,0}b_{2,1}b_{2,2}) = \left(1+\dfrac{1}{a^*_{2,0}a^*_{2,1}}\right)(1+b^*_{2,0}b^*_{2,1}b^*_{2,2}), & (3.13) \end{cases}$$

where $a_{2,0} = \tilde{a}_{2,0}$ and all the parameters $a_{2,0},\ a_{2,1},\ b_{2,0},\ b_{2,1}, b_{2,2},\ a^*_{2,0},\ a^*_{2,1},\ b^*_{2,0},\ b^*_{2,1},\ b^*_{2,2}$ are positive. From (3.10), we have

$$b_{2,0} = \frac{(1+a^*_{2,0})(1+b^*_{2,0})}{1+\tilde{a}_{2,0}} - 1 \tag{3.14}$$

and from (3.11), we have

$$b_{2,2} = \frac{\tilde{a}_{2,0}(1+a^*_{2,0})(1+b^*_{2,0}b^*_{2,2}) - a^*_{2,2}(1+\tilde{a}_{2,0})}{a^*_{2,0}\{(1+a^*_{2,0})(1+b^*_{2,0}) - (1+\tilde{a}_{2,0})\}}. \tag{3.15}$$

To guarantee that $b_{2,0}, b_{2,2} > 0$, $\tilde{a}_{2,0}$ must satisfy

$$\frac{a^*_{2,0}}{1+b^*_{2,0}b^*_{2,2}(1+a^*_{2,0})} < \tilde{a}_{2,0} < (1+a^*_{2,0})(1+b^*_{2,0}) - 1.$$

We assume this condition for $\tilde{a}_{2,0}$. By multiplying (3.13) by $\tilde{a}_{2,0}a_{2,1}a^*_{2,0}a^*_{2,1}$ and dividing by (3.12), we obtain

$$a_{2,1} = \frac{a^*_{2,0}a^*_{2,1}(1+b^*_{2,0}b^*_{2,1})(1+b_{2,0}b_{2,1}b_{2,2})}{\tilde{a}_{2,0}(1+b^*_{2,0}b^*_{2,1}b^*_{2,2})(1+b_{2,0}b_{2,1})}. \tag{3.16}$$

By replacing $b_{2,0}, b_{2,2}$ with (3.14), (3.15), respectively, we have

$$\begin{aligned} b_{2,1} = {} & \frac{b^*_{2,0}b^*_{2,1}\{a^*_{2,0}a^*_{2,1}b^*_{2,2}(b^*_{2,0}b^*_{2,1}+1) + b^*_{2,0}b^*_{2,1}b^*_{2,2}+1\}}{\Big\{(a^*_{2,0}+1)(b^*_{2,0}+1)/(\tilde{a}_{2,0}+1) - 1\Big\}} \\ & \times \frac{a^*_{2,0}(a^*_{2,0}+1)(b^*_{2,0}+1) - (\tilde{a}_{2,0}+1)}{a^*_{2,0}a^*_{2,1}(b^*_{2,0}b^*_{2,1}+1)[\tilde{a}_{2,0}\{(a^*_{2,0}+1)b^*_{2,0}b^*_{2,2}+1\} - a^*_{2,0}]}. \end{aligned} \tag{3.17}$$

Furthermore, by substituting (3.14), (3.15), and (3.17) into (3.16), respectively, we can obtain $a_{2,1}$ as a function of $(\tilde{a}_{2,0}, a^*_{2,0},\ a^*_{2,1}, b^*_{2,0}, b^*_{2,1}, b^*_{2,2})$. Therefore, all the parameters $a_{2,0}, a_{2,1}, b_{2,0}, b_{2,1}, b_{2,2}$ are represented by them. □

Usually, if smaller models, e.g., the AR(1) model, do not have identifiability, then neither do larger models, e.g., the AR(2) model. However, this is not always true, since the poor information at time t makes the Logistic AR(1) model unidentified, but the use of past information may make the Logistic AR(2) model identified. Recall that the AR(1) model defined in (3.2)–(3.3) is not identified, because the model imposes $2^2 = 4$ constraints in (3.9) with a larger number of the parameters (5), at each time $t \geq 2$ in the model; see the graphical representation in Figure 3.1. In the figure, the parameters enclosed by broken lines are intercepts in the model and each of the other parameters is the regression coefficient of the nearest arrow in the model. As we have already developed an expression for the joint distribution function of the AR(p) model in Lemma 3.2.1, if $t = 3, p = 2$, one parameter is added for the $\boldsymbol{Y}$s serial correlation; thus, we have six parameters versus $2^3 = 8$ (> 6) constraints. This is represented visually in Figure 3.2. Thus, for $t \geq 3$, we can claim that (3.8) holds as follows.

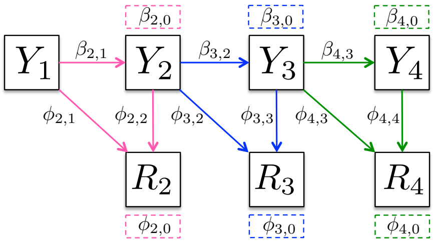

Figure 3.1: AR(1) model

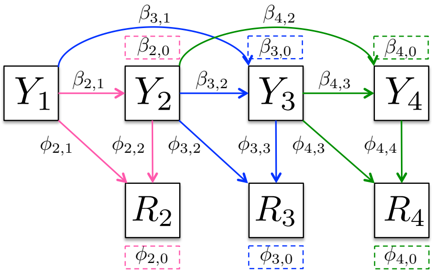

Figure 3.2: AR(2) model

Proposition 3.3.2. In the Logistic AR(2) model, for $t \geq 3$, the following implication holds if and only if $\beta_{t,t-2}$ is not zero:

$$\forall_{\boldsymbol{v}_t^2} \quad \pi_t(\boldsymbol{v}_t^2;\ \boldsymbol{\xi}_t) = \pi_t(\boldsymbol{v}_t^2;\ \boldsymbol{\xi}_t^*) \quad \Rightarrow \quad \boldsymbol{\xi}_t = \boldsymbol{\xi}_t^*.$$

Proof. The statement that we want to prove is that $(\boldsymbol{\beta}_t, \boldsymbol{\phi}_t) = (\boldsymbol{\beta}_t^*, \boldsymbol{\phi}_t^*)$ if, for every $(y_{t-2}, y_{t-1}, y_t)^\top \in \{0,1\}^{\otimes 3}$,

$$\frac{1}{[1+\exp\{-(\phi_{t,0}+\phi_{t,t-1}y_{t-1}+\phi_{t,t}y_t)\}][1+\exp\{(-1)^{y_t}(\beta_{t,0}+\beta_{t,t-2}y_{t-2}+\beta_{t,t-1}y_{t-1})\}]}$$
$$= \frac{1}{[1+\exp\{-(\phi_{t0}^*+\phi_{tt-1}^*y_{t-1}+\phi_{t,t}^*y_t)\}][1+\exp\{(-1)^{y_t}(\beta_{t0}^*+\beta_{tt-2}^*y_{t-2}+\beta_{tt-1}^*y_{t-1})\}]}. \tag{3.18}$$

We prove it here only for $t = 3$ since the above expression is identical for $t \geq 3$. The condition in (3.18) is equivalent to the following eight expressions:

$$\begin{cases}
(1 + a_{3,0})(1 + b_{3,0}) = (1 + a^*_{3,0})(1 + b^*_{3,0}) & (3.19)\\
(1 + a_{3,0}a_{3,1})(1 + b_{3,0}) = (1 + a^*_{3,0}a^*_{3,1})(1 + b^*_{3,0}) & (3.20)\\
(1 + a^{-1}_{3,0})(1 + b_{3,0}b_{3,3}) = (1 + a^{*-1}_{3,0})(1 + b^*_{3,0}b^*_{3,3}) & (3.21)\\
(1 + a^{-1}_{3,0}a^{-1}_{3,1})(1 + b_{3,0}b_{3,3}) = (1 + a^{*-1}_{3,0}a^{*-1}_{3,1})(1 + b^*_{3,0}b^*_{3,3}) & (3.22)\\
(1 + a_{3,0}a_{3,2})(1 + b_{3,0}b_{3,2}) = (1 + a^*_{3,0}a^*_{3,2})(1 + b^*_{3,0}b^*_{3,2}) & (3.23)\\
(1 + a_{3,0}a_{3,1}a_{3,2})(1 + b_{3,0}b_{3,2}) = (1 + a^*_{3,0}a^*_{3,1}a^*_{3,2})(1 + b^*_{3,0}b^*_{3,2}) & (3.24)\\
(1 + a^{-1}_{3,0}a^{-1}_{3,2})(1 + b_{3,0}b_{3,2}b_{3,3}) = (1 + a^{*-1}_{3,0}a^{*-1}_{3,2})(1 + b^*_{3,0}b^*_{3,2}b^*_{3,3}) & \\
(1 + a^{-1}_{3,0}a^{-1}_{3,1}a^{-1}_{3,2})(1 + b_{3,0}b_{3,2}b_{3,3}) = (1 + a^{*-1}_{3,0}a^{*-1}_{3,1}a^{*-1}_{3,2})(1 + b^*_{3,0}b^*_{3,2}b^*_{3,3}) &
\end{cases}$$

where

$$\begin{aligned}
&(a_{3,0}, a_{3,1}, a_{3,2}, b_{3,0}, b_{3,2}, b_{3,3})\\
&= (\exp(\beta_{3,0}), \exp(\beta_{3,1}), \exp(\beta_{3,2}), \exp(-\phi_{3,0}), \exp(-\phi_{3,2}), \exp(-\phi_{3,3})),\\
&(a^*_{3,0}, a^*_{3,1}, a^*_{3,2}, b^*_{3,0}, b^*_{3,2}, b^*_{3,3})\\
&= (\exp(\beta^*_{3,0}), \exp(\beta^*_{3,1}), \exp(\beta^*_{3,2}), \exp(-\phi^*_{3,0}), \exp(-\phi^*_{3,2}), \exp(-\phi^*_{3,3})).
\end{aligned}$$

By dividing (3.19) by (3.20) and (3.22) by (3.21), we have

$$\frac{1 + a_{3,0}}{1 + a_{3,0}a_{3,1}} = \frac{1 + a^*_{3,0}}{1 + a^*_{3,0}a^*_{3,1}}, \tag{3.25}$$

$$\frac{a_{3,1}(1 + a_{3,0})}{1 + a_{3,0}a_{3,1}} = \frac{a^*_{3,1}(1 + a^*_{3,0})}{1 + a^*_{3,0}a^*_{3,1}}. \tag{3.26}$$

By substituting (3.25) into (3.26), we obtain $a_{3,1} = a^*_{3,1}$. Plugging this equation back into (3.25), $(a^*_{3,1} - 1)(a_{3,0} - a^*_{3,0}) = 0$. Thus, if $a^*_{3,1} \neq 1$, $a_{3,0} = a^*_{3,0}$, otherwise $a_{3,0} = c$, where c is an arbitrary positive constant. Hence, when $a^*_{3,1} = 1$, the model is not identifiable. If $a^*_{3,1} \neq 1$, then dividing (3.23) by (3.24) and substituting $a_{3,0} = a^*_{3,0}$ and $a_{3,1} = a^*_{3,1}$, we find $(a^*_{3,1} - 1)(a_{3,2} - a^*_{3,2}) = 0$, which implies $a_{3,2} = a^*_{3,2}$. Therefore, we have

$$a^*_{3,1} \neq 1 \quad \Leftrightarrow \quad (a_{3,0},\, a_{3,1},\, a_{3,2}) = (a^*_{3,0},\, a^*_{3,1},\, a^*_{3,2}).$$

From (3.19), (3.21), and (3.23), it is obvious that

$$(a_{3,0},\, a_{3,1},\, a_{3,2}) = (a^*_{3,0},\, a^*_{3,1},\, a^*_{3,2}) \quad \Leftrightarrow \quad (b_{3,0},\, b_{3,2},\, b_{3,3}) = (b^*_{3,0},\, b^*_{3,2},\, b^*_{3,3})$$

holds. Hence,

$$a^*_{3,1} \neq 1 \quad \Leftrightarrow \quad (a_{3,0},\, a_{3,1},\, a_{3,2},\, b_{3,0},\, b_{3,2},\, b_{3,3}) = (a^*_{3,0},\, a^*_{3,1},\, a^*_{3,2},\, b^*_{3,0},\, b^*_{3,2},\, b^*_{3,3}).$$

Thus, we have reached the desired conclusion. □

Evidently, as shown in the graphical representations in Figures 3.1–3.2, the AR(2) model is not identified when $t = 2$ for the same reason as the AR(1) model. We need additional information or restrictions on the model to achieve identifiability.

Consider the case when there are no missing values at time $t = 2$. In this case, we can fix the two parameters $\phi_{2,1} = \phi_{2,2} = 0$. Hence, there are 3 $(= 5 - 2)$ parameters versus 4 (> 3) constraints. The model is identifiable but the proof is omitted because it is similar to that of Proposition 2. Another case is a model with one binary covariate which is invariant for all times and has no missing data such as the information of dose in Machin et al. (1988). More specifically, consider the AR(1) model with a binary covariate x as

$$\Pr(Y_1 = 1 \mid x; \beta_{1,0}, \beta_1) = \mathrm{expit}(\beta_{1,0} + \beta_1 x),$$
$$\forall_{t \geq 2} \quad \Pr(Y_t = 1 \mid y_{t-1}, x\,;\, \beta_{t,0}, \beta_{t,t-1}, \beta_t) = \mathrm{expit}(\beta_{t,0} + y_{t-1}\beta_{t,t-1} + \beta_t x).$$

A graphical representation is shown in Figure 3.3. There are six parameters versus eight constraints for each time $t \geq 3$, the same situation as in the AR(2) model for each $t \geq 3$. At time $t = 1$, we can fix the two parameters to be zero, which has an effect on the corresponding missing data indicator since there are no missing data on the covariate. Thus, the model is identified for the same reason as when there are no missing data at $t = 2$. Therefore, this model is identifiable but the proof is omitted because it is similar to that of Proposition 3.3.2. Moreover, we can show that the AR(2) with one covariate model graphically represented in Figure 3.4 is also identifiable. Note that the models shown in Figures 3.3–3.4 are still identifiable even if there are direct paths from the covariate to the response indicators.

3.4 Semiparametric models

In practice, checking model identification is very hard when both response and outcome models are specified parametrically. The specifications tend to be subjective. It would be preferable to specify either the outcome or the response model nonparametrically.

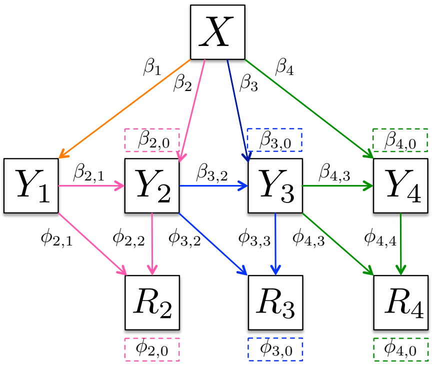

Figure 3.3: AR(1) model with one covariate

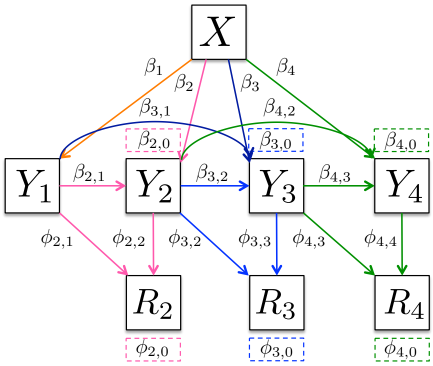

Figure 3.4: AR(2) model with one covariate

Such a model can be estimated by semiparametric methods (e.g., Tang et al., 2003; Zhao and Shao, 2015) following the literature on nonignorable nonresponse in which R has only two patterns, i.e., 0 or 1.

Recently, the identification problem of semiparametric models in the situation of nonignorable nonresponse has been studied by Wang et al. (2014) and Zhao and Shao (2015). Their key assumption is the existence of a "nonresponse instrumental variable". Let us consider the situation of a nonresponse problem first. Suppose that a covariate vector $\boldsymbol{X}$ is completely observed and response variables Y are subject to missingness. In this case, R is a binary variable, which takes 1 (0) if Y is observed (missing). Roughly speaking, their claim is that if the covariate $\boldsymbol{X}$ has two components, $\boldsymbol{X} = (\boldsymbol{Z}^\top, \boldsymbol{U}^\top)^\top$, such as $\boldsymbol{Z} \perp R \mid \boldsymbol{U}$ and $Y \not\perp \boldsymbol{Z} \mid \boldsymbol{U}$, then the semiparametric model is identifiable. A random vector $\boldsymbol{Z}$ satisfying such conditions is called a nonresponse instrumental variable.

In this section, we consider identification conditions for the semiparametric models without using any nonresponse instrumental variable. We also expect our estimator to be more efficient than other estimators because our estimator is based on maximum likelihood. We consider two cases: (i) the outcome model is parametric; (ii) the response model is parametric. The first case occurs when the effects of covariates on a response variable is our main interest. In this case, the response model is a nuisance. The second case is important when our interest is nonparametric statistics such as the mean and quantile, defined by a solution to an estimating equation $\mathrm{E}\{U(\theta; \boldsymbol{X}, Y)\} = 0$, where U

is a known function corresponding to θ, e.g., $U(\theta) = y - \theta$, when our interest is the mean of Y. In such a case, the outcome model is a nuisance, and we want to avoid imposing a parametric model on it.

Because, as we have already seen, the model identification in repeated measurement data can be considered separately in each time by Theorem 3.2.1, we consider a model identification of π_3 only.

3.4.1 Parametric outcome model

Let outcome follow the Logistic AR(2) model and the response model be nonparametric, that is,

$$\begin{aligned}\Pr(Y_3 = 1 \mid y_1, y_2\,;\, \boldsymbol{\beta}_3) &= \operatorname{expit}(\beta_{3,0} + y_1\beta_{3,1} + y_2\beta_{3,2}),\\ \Pr(R_3 = 1 \mid y_1, y_2, y_3\,;\boldsymbol{\phi}_3) &= \phi_{3,y_1,y_2,y_3},\end{aligned}$$

where $\boldsymbol{\beta}_3 = (\beta_{3,0}, \beta_{3,1}, \beta_{3,2})^\top$ and $\boldsymbol{\phi}_3 = (\phi_{3,0,0,0}, \dots, \phi_{3,1,1,1})^\top$. This model is not identifiable, because there are 11 unknown parameters but we have only eight constraints. This also implies that we need an assumption on the response mechanism because the nonparametric part already has eight parameters. If we could assume an independence $R_3 \perp Y_1 \mid (Y_2, Y_3)$, i.e., the AR(2) property for the response model, then that would make the model identifiable because this assumption reduces the number of parameters from 11 to 7 (< 8).

3.4.2 Parametric response mechanism

Let the outcome model be nonparametric and the response model be a logistic model, viz.

$$\begin{aligned}\Pr(Y_3 = 1 \mid y_1, y_2\,;\, \boldsymbol{\beta}_3) &= \beta_{3,y_1,y_2},\\ \Pr(R_3 = 1 \mid y_1, y_2, y_3\,;\boldsymbol{\phi}_3) &= \operatorname{expit}(\phi_{3,0} + y_1\phi_{3,1} + y_2\phi_{3,2} + y_3\phi_{3,3}),\end{aligned}$$

where $\boldsymbol{\beta}_3 = (\beta_{3,0,0}, \dots, \beta_{3,1,1})^\top$ and $\boldsymbol{\phi}_3 = (\phi_{3,0}, \phi_{3,1}, \phi_{3,2}, \phi_{3,3})^\top$. There are 8 unknown parameters versus eight constraints. Therefore, this model is identifiable without using any instrumental variable. Also, we can use the maximum likelihood to estimate $\boldsymbol{\beta}_3$ and $\boldsymbol{\xi}_3$. Identification at the other periods can be checked in the same way.

3.5 Other response variables

So far, we have considered a case where all variables are dichotomous. What happens when the response variable takes more than three values and continuous one? We consider the nonresponse case in this section because we now know that it is enough to consider only the situation from Theorem 3.2.1. Suppose that a covariate variable x is completely observed and a response variable y is subject to missingness. In this case, r is a binary variable, which takes 1 (0) if y is observed (missing). Then the nonresponse instrumental variable assumption on x corresponds to assume

$$X \perp R \mid Y, \quad \text{and} \quad Y \not\perp X. \tag{3.27}$$

The identification condition proposed in §3.5.2 does not require such conditions.

3.5.1 Ternary response variables

Let x and y be ternary variables taking $0, 1$, and 2, and the outcome model be

$$\Pr(Y = 1 \mid x) = \frac{\exp(\beta_1 I_{\{x=1\}})}{\exp(\beta_0 I_{\{x=0\}}) + \exp(\beta_1 I_{\{x=1\}}) + \exp(\beta_2 I_{\{x=2\}})},$$
$$\Pr(Y = 2 \mid x) = \frac{\exp(\beta_2 I_{\{x=2\}})}{\exp(\beta_0 I_{\{x=0\}}) + \exp(\beta_1 I_{\{x=1\}}) + \exp(\beta_2 I_{\{x=2\}})},$$

where I_A takes 1 (0) if an event A is true (false). Thus, there are 3 parameters for outcome and 3 parameters for the response mechanism. Therefore, the model is identifiable, unlike binary variables, because there are 6 unknown parameters versus $3^2 = 9$ constraints. Also, it is easily seen that a continuous covariate x in Figure 3.3, which is a mixture of discrete past response variables and a continuous covariate, makes the model identifiable.

3.5.2 Continuous response variables

However, a continuous response vector y makes the model identification problem much more difficult. For example, Miao et al. (2016) showed that when the outcome model $y \mid x$ is normal or a normal mixture, and the response mechanism is logistic, the model is not identifiable. However, if the response mechanism is probit or tobit, then the model is identifiable. This result is hard to understand intuitively, and we cannot check the model identification by just comparing the numbers of parameters and constraints.

A sufficient condition for the continuous response variable is guaranteed by using a nonresponse instrumental variable given in (3.27).

Although the existence of such a nonresponse instrumental variable is a sufficient condition, it is hard to verify it from the observed data. Therefore, we propose an alternative condition for the model identification by assuming a restriction on $[y \mid x, r = 1]$, not only on the response mechanism. Let a response model be $\pi(\boldsymbol{z}; \boldsymbol{\phi}) = \Pr(R = 1 \mid x, y; \boldsymbol{\phi})$, a parametric model on $f_1(y \mid x) = f(y \mid x, r = 1)$ be $f_1(y \mid x; \boldsymbol{\gamma})$, and that on $f(y \mid x)$ be $f(y \mid x; \boldsymbol{\beta})$. Also denote odds function of the response model by $O(\boldsymbol{z}; \boldsymbol{\phi}) = 1/\pi(\boldsymbol{z}; \boldsymbol{\phi}) - 1$ and the expectation on any function $g(\boldsymbol{z})$ given observed data by $E_1\{g(x, Y) \mid x; \boldsymbol{\gamma}\} = E\{g(x, Y) \mid r = 1, x; \boldsymbol{\gamma}\}$. Recall that the model identification condition is

$$\begin{aligned} &f(y \mid x; \boldsymbol{\beta})\pi(\boldsymbol{z}; \boldsymbol{\phi}) = f(y \mid x; \boldsymbol{\beta}')\pi(\boldsymbol{z}; \boldsymbol{\phi}') \quad \text{w.p. 1}, \\ &\Rightarrow \boldsymbol{\beta} = \boldsymbol{\beta}' \text{ and } \boldsymbol{\phi} = \boldsymbol{\phi}', \end{aligned} \tag{3.28}$$

where w.p. 1 implies with probability one with respect to the true distribution.

Theorem 3.5.1. Suppose that for each $(\boldsymbol{\phi}, \boldsymbol{\gamma}) \in \Phi \times \Gamma$, $E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid x; \boldsymbol{\gamma}\} < \infty$ with probability one with respect to the true distribution. Then, if

$$\begin{aligned} &f_1(y \mid x; \boldsymbol{\gamma}) = f_1(y \mid x; \boldsymbol{\gamma}') \text{ and } E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid x; \boldsymbol{\gamma}\} = E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}') \mid x; \boldsymbol{\gamma}'\} \quad \text{w.p.1}, \\ &\Rightarrow \boldsymbol{\gamma} = \boldsymbol{\gamma}' \text{ and } \boldsymbol{\phi} = \boldsymbol{\phi}', \end{aligned}$$

with probability one, implies $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$ and $\boldsymbol{\phi} = \boldsymbol{\phi}'$. This is a necessary and sufficient condition of the model identification (3.28).

For example, if the response mechanism is specified as $\pi(\boldsymbol{z}; \boldsymbol{\phi}) = 1/\{1 + \exp(\phi_{\text{x}0} + \phi_{\text{x}1} x + \phi_{\text{y}} y)\}$, where $\boldsymbol{\phi} = (\phi_{\text{x}0}, \phi_{\text{x}1}, \phi_{\text{y}})^{\text{T}}$. Then, $E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid x\}$ is written as

$$E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid x\} = \exp\{\phi_{\text{x}0} + \phi_{\text{x}1} x + K_{\phi_{\text{y}}}(x)\},$$

where $K_{\phi_2}(x)$ is the cumulant-generating function of $[y \mid x, r = 1]$. Therefore, we have only to check whether $K_{\phi_{\text{y}}}(x)$ is linear with respect to x or not. If f_1 is a parametric model, the model identification for $\boldsymbol{\phi}$ is easy to check. For example, if $[y \mid x, r = 1]$ belongs to an exponential family with the density function

$$f_1(y \mid x; \phi, \psi) = \exp\left[\frac{y\phi(x) - b\{\phi(x)\}}{\psi} + c(y, \psi)\right],$$

where ψ is the dispersion parameter and ϕ, b, c are known functions, then the cumulant-generating function reduces to $K_{\phi_{\mathrm{y}}}(x) = \{b(\phi_{\mathrm{y}}\psi + \phi(x)) - b(\phi(x))\}/\psi$, from which we can verify the model identification. For example, for model identification, b is allowed to be any polynomial function except for the 1st- and 2nd-order function of x such as log-function (e.g. Gamma distribution), exponential-function (e.g. Poisson distribution), etc. However, when b is a 2nd-order polynomial function, for example, $b(\phi) = \phi^2/2$, which means f_1 follows normal distribution, then $K_{\phi_{\mathrm{y}}}(x) = \phi(x)\phi_{\mathrm{y}} + \phi_{\mathrm{y}}^2\psi^2/2$. Also, we obtain

$$E_1\{O(\boldsymbol{Z};\boldsymbol{\phi}) \mid x\} = \exp\{\phi_{\mathrm{x}0} + \phi_{\mathrm{x}1}x + \phi(x)\phi_{\mathrm{y}} + \phi_{\mathrm{y}}^2\psi^2/2\}.$$

Theorem 3.5.1 implies ϕ is identifiable unless the mean structure $\phi(x)$ is linear. In practice, it will be worth checking the linearity if $f_1(y \mid x)$ is close to normal distribution. If $\phi(x)$ is linear we may use a nonresponse instrumental variable or consider other approaches (see §5.1 for more details). we can check identifiability for other distributions of $[y \mid x, r = 1]$ in a similar way.

Chapter 4

Simulation study

In this chapter, we conduct two simulation studies. One is to evaluate the performance of our proposed estimators stated in §2.4 and compare their efficiency with other methods in finite samples. The other simulation study is related to a topic discussed in §2.5. The main aim of the simulation is (i) to confirm asymptotic behavior of EL and GMM estimators when $q < p_1 + 1 = p_2$ and $\boldsymbol{g}(\boldsymbol{x}) = \{1, \boldsymbol{h}(\boldsymbol{x})\}^\top$; (ii) to check the performance of estimators proposed in Theorem 2.5.2 when both a response and $f_1(y \mid x)$ models are misspecified.

4.1 Comparison between proposed and existing semiparametric estimators

We conducted a Monte Carlo simulation study with four scenarios. In each scenario, we used a covariate $X \sim U(-1, 1)$, set the response mechanism to a Bernoulli distribution with $\mathrm{logit}\{\pi_{\mathrm{y}}(y)\} = \phi_{\mathrm{x}0} + \phi_{\mathrm{y}} y$, and generated the response outcome variable from $Y \mid (x, r = 1) \sim \mathcal{N}(\mu_s(x), \sigma_s^2)$. In Scenarios 1–3, $\mu_s(x)$ is defined as the s-th order polynomial: $\mu_1(x) = x - 0.121$, $\sigma_1^2 = 1/3$; $\mu_2(x) = 0.8x^2 - 0.3415$, $\sigma_2^2 = 1/4$; $\mu_3(x) = 2x(x - 3/4)(x + 3/4) - 0.0802$, $\sigma_3^2 = 1/3$. In Scenario 4, $\mu_4(x) = \{\cos(x\pi) + 2\sin(2x\pi)\}/2 - 0.06$, $\sigma_4^2 = 1/4$. We generated missing data by the response mechanism with $(\phi_{\mathrm{y}0}, \phi_{\mathrm{y}}) = (1.03, -1.2)$, $(0.91, -1)$, $(0.9, -0.8)$, $(0.91, -0.8)$ in Scenarios 1–4 respectively, so that the response rate is about 70 % and $E(Y) = 0$.

We note that x is a nonresponse instrumental variable (Wang et al., 2014); thus the parameters are identifiable in all scenarios. We also consider the case when the response mechanism is over-specified as $\mathrm{logit}\{\pi_{\mathrm{xy}}(x,y)\} = \phi_{\mathrm{x0}} + \phi_{\mathrm{x1}}x + \phi_{\mathrm{y}}y$. In this case, there is no instrumental variable, but all the parameters are identifiable except for Scenario 1 by Theorem 2.4.1. However, by using Theorem 2.4.1, it is possible to make the response model in Scenario 1 identifiable at the risk of misspecification of response mechanism. This problem is dealt with in the next section. We estimate the parameters for the two response mechanisms $\pi_{\mathrm{y}}(y;\boldsymbol{\phi})$ and $\pi_{\mathrm{xy}}(x,y;\boldsymbol{\phi})$, as well as $\theta = E(Y)$. For the response mechanisms, only ϕ_{y} is reported.

From each sample, we computed four estimators, as follows:

[1] MAR: A naive estimator based on the assumption that the missing data are missing-at-random:

$$\sum_{i=1}^{n} \delta_i(\theta - y_i)/\hat{\pi}_i = 0, \tag{4.1}$$

where $\hat{\pi}_i$ is an estimated response mechanism, that is, $\hat{\pi}_i = \{1 + \exp(\hat{\phi}_{\mathrm{x0}} + \hat{\phi}_{\mathrm{x}}x_i)\}^{-1}$, where $(\hat{\phi}_{\mathrm{x0}}, \hat{\phi}_{\mathrm{x}})$ is the maximum likelihood estimator.

[2] CK: The estimator of Chang and Kott (2008) and Kott and Chang (2010) . We use the estimating equation (2.9), setting g as $(1,x)^\top$ for $\pi_{\mathrm{y}}(y)$ and $(1,x,x^2)^\top$ for $\pi_{\mathrm{xy}}(x,y)$; θ is estimated by using (4.1) with the estimated response mechanism.

[3] RKI: The estimator of Riddles et al. (2016). In all scenarios, we specified a parametric model on f_1 based on normal distribution with mean structure $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. A best model among $2^4 - 1$ candidate models was chosen by using AIC; θ is estimated by using (4.1) with the estimated response mechanism. We take notice of the fact that a misspecified model was used in Scenario 4.

[4] New: Proposed estimator in §2.4. As for the working model for f_1, the same model specification as in the RKI method was used. We also assume the nonparametric estimator proposed in Theorem 2.4.2. We call the parametric method "P" and the nonparametric method "NP" in this section.

Monte Carlo samples of size $n = 500$ and 2,000 were independently generated 2,000 times. We used the correct models $\pi_{\mathrm{y}}(y)$ and $\pi_{\mathrm{xy}}(x,y)$ for the response mechanism, except for MAR.

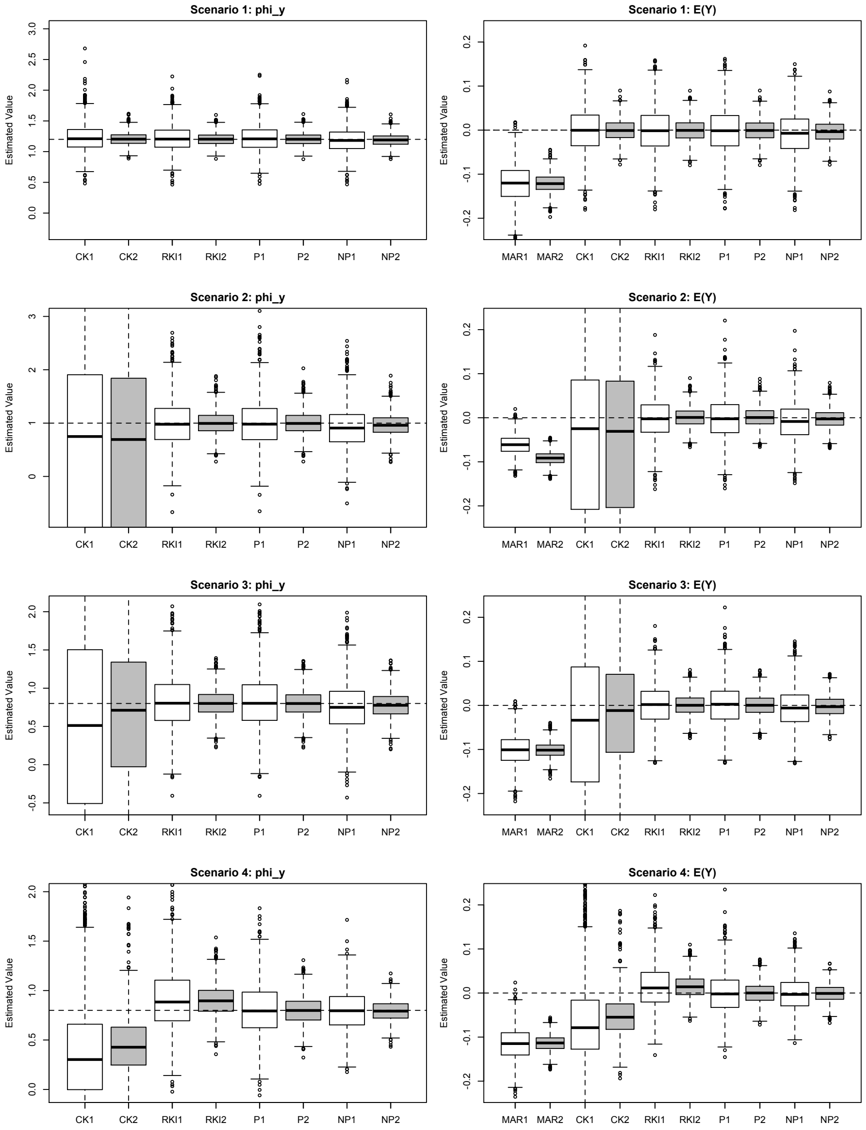

Figure 4.1: Boxplot of Monte Carlo results for ϕ_{y} and $\theta\{= E(Y)\}$ under four scenarios when $\phi_{\mathrm{x}1}$ is set to be 0. The four estimators are MAR (missing at random), CK (Chang & Kott's estimator), RKI (Riddles' estimator), P (our proposed estimator with parametric f_1 model) NP (our proposed estimator with nonparametric method). Numbers 1 and 2 stand for $n = 500$ and $n = 2,000$, respectively. The broken line shows the true value.

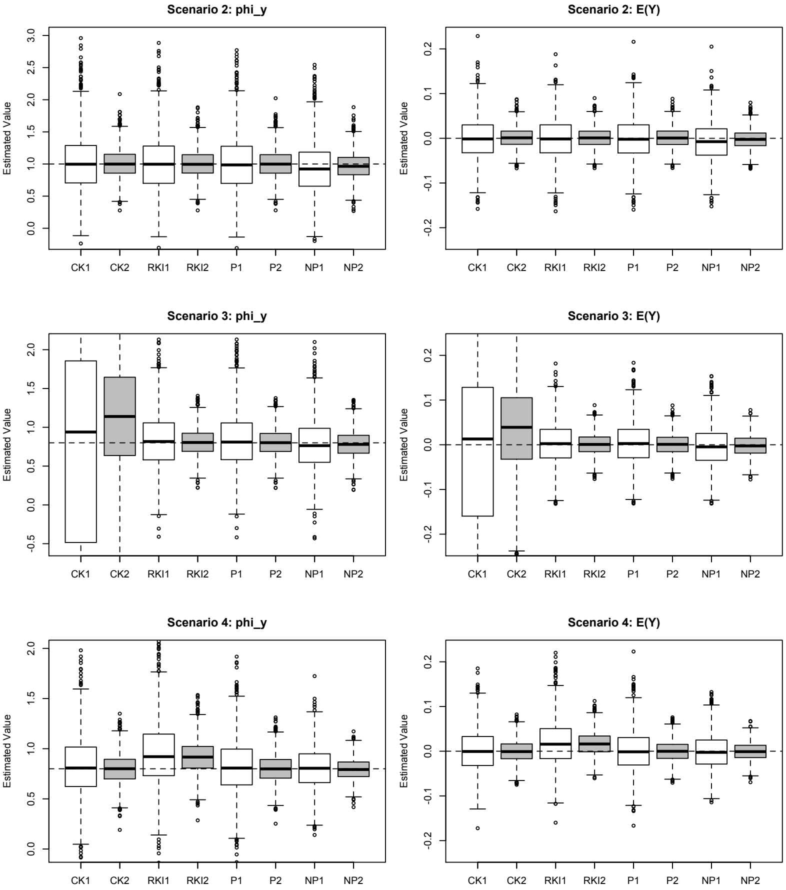

Figure 4.2: Boxplot of Monte Carlo results for ϕ_{y} and $\theta\{= E(Y)\}$ under four scenarios when ϕ_{x1} is estimated. The four estimators are MAR (missing at random), CK (Chang & Kott's estimator), RKI (Riddles' estimator), P (our proposed estimator with parametric f_1 model) NP (our proposed estimator with nonparametric method). Numbers 1 and 2 stand for $n = 500$ and $n = 2,000$, respectively. The broken line shows the true value.

4.1 Comparison between proposed and existing semiparametric estimators

Figure 4.1 shows the Monte Carlo simulation results with the response mechanism $\pi_{\mathrm{y}}(y)$ in all scenarios; Figure 4.2 shows the results with $\pi_{\mathrm{xy}}(x, y)$. In Figure 4.2, only the results for Scenarios 2-4 are shown because the parameters are not identifiable in Scenario 1; the result of MAR is not shown because it is already shown in Figure 4.1. In the CK method with $\pi_{\mathrm{y}}(y)$ mechanism, we encountered some numerical problems in Scenarios 2–4 and there was no solution because the estimate of ϕ did not converge. The following is the summary of the simulation results shown in Figure 4.1 and Figure 4.2:

[1] In all scenarios, the naive estimator using the MAR assumption is significantly biased, since this assumption does not hold.

[2] The CK method with $\pi_{\mathrm{y}}(y)$ model works well in Scenario 1, but the performance suffers from numerical problems in the other scenarios. However, the CK method with $\pi_{\mathrm{xy}}(x, y)$ model works well even in Scenarios 2 and 4 (though less efficient compared to RKI and our proposed estimators). This is because the calibration condition on $g(x) = (1, x)^{\top}$ falls short of estimating the parameters when the relationship between x and y becomes more complicated.

[3] The RKI method performs quite well in Scenarios 1–3 for both response mechanisms, but the estimators in Scenario 4 are somewhat positively biased in RKI due to the misspecification of the f_1 model.

[4] In all scenarios, our proposed estimators perform better than any other methods. We note that in Scenario 4, the estimator using parametric f_1 is still consistent despite misspecification of f_1. However it is less efficient compared to the nonparametric method because of the misspecification.

Table 1 shows the estimated coverage probability with 95% coverage confidence interval for our proposed estimators. We applied (2.26) to estimate the variance of our estimators both using the parametric f_1 model and the nonparametric model (see §2.4.2). My proposed variance estimator works well in all scenarios.

Table 4.1: The coverage probability of the confidence interval with 95% coverage rate for our proposed estimators with sample size $n = 500$ and $n = 2,000$ when ϕ_{x1} is set to be 0 (fix) and estimated (est).

Scenario	Method	n	ϕ_{x1}		Scenario	Method	n	ϕ_{x1}	
			fix	est				fix	est
1	P	500	0.939	–	3	P	500	0.958	0.950
		2000	0.944	–			2000	0.953	0.946
	NP	500	0.930	–		NP	500	0.964	0.940
		2000	0.937	–			2000	0.943	0.944
2	P	500	0.953	0.949	4	P	500	0.948	0.953
		2000	0.943	0.946			2000	0.941	0.943
	NP	500	0.942	0.959		NP	500	0.943	0.942
		2000	0.946	0.951			2000	0.946	0.949

P: method using a parametric working model for f_1, NP: nonparametric method.

4.2 Robustness for misspecification of response mechanism

In Theorem 2.5.1, we have shown that both EL and GMM estimators are exactly the same when $q = p_1 + 1 = p_2$. However, we have not shown any relationships of the two estimators when $q < p_1 + 1 = p_2$. Thus, we check the relationships here. As for (ii), in Theorem 2.5.2, we have shown that any function $\boldsymbol{k}(\boldsymbol{x})$ can be taken into our GMM constraints without loss of efficiency as long as $\boldsymbol{g}^*(\boldsymbol{x})$ is used. We check that the other function $\boldsymbol{k}(\boldsymbol{x})$ has a potential to reduce bias when both models are misspecified.

Let a covariate $X \sim \mathcal{N}(-1, 1)$, $Y \mid X = x \sim \mathcal{N}(2x(x-3/4)(x+3/4), 1/3)$. We prepare six different response mechanisms: M1(linear nonignorable): $\mathrm{logit}\{\pi(y)\} = 0.90 - 1.0y$; M2(linear ignorable): $\mathrm{logit}\{\pi(x)\} = 0.90 - 1.0x$; M3(quadratic nonignorable): $\mathrm{logit}\{\pi(y)\} = 1.3 - 0.5y - y^2$; M4(linear ignorable with a cross term): $\mathrm{logit}\{\pi(x, y)\} = 0.9 - 0.5y - 0.5xy$; M5(probit nonignorable): $\pi(y) = \Phi(0.6 + 0.5y - y^2)$; M6(log-log nonignorable): $\pi(x, y) = 1 - \exp\{-\exp(0.25 + 0.5x - y)\}$. In this simulation study, we specify a response model by

$$\pi(y; \boldsymbol{\phi}) = \frac{1}{1 + \exp(\phi_0 + \phi_1 y)},$$

where $\boldsymbol{\phi} = (\phi_0, \phi_1)^\top$. Therefore all models except for M1 are misspecified.

Furthermore, we assume that $f_1(y \mid x)$ is normally distributed with a mean function $\mu(x; \boldsymbol{\beta})$ and variance σ^2. Although, strictly speaking, this $f_1(y \mid x)$ misspecifies the true f_1, because it is given by

$$f_1(y \mid x) = \frac{\pi(x, y) f(y \mid x)}{\int \pi(x, y) f(y \mid x) dy}.$$

Thus, the true $f_1(y \mid x)$ depends on its response mechanism as well as the distribution of outcome, and it is not normal. We consider two settings for the mean function

1. (Heavily misspecified model) $\mu_1(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x$;

2. (Slighly misspecified model) $\mu_2(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$,

where, in the second model, an appropriate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ is chosen by AIC (Akaike Information Criterion). The first mean function is linear with regard to x, on the other hand, the second function is cubic. It is expected that estimators with the second mean function is more efficient than the first one.

For EL estimators, we consider three types of $\boldsymbol{h}(\boldsymbol{x})$ function: (1) $\boldsymbol{h}_1(x) = (x, x^2)^\top$; (2) $\boldsymbol{h}_2(x) = \boldsymbol{g}^\star(x)$; (3) $\boldsymbol{h}_3(x) = \{x, \boldsymbol{g}^\star(x)^\top\}^\top$. For GMM estimtaors, we consider four types of $\boldsymbol{g}(\boldsymbol{x})$ function: (1) $\boldsymbol{g}_1(x) = (1, x, x^2)$; (2) $\boldsymbol{g}_2(x) = \boldsymbol{g}^\star(x)$; (3) $\boldsymbol{g}_3(x) = \{1, \boldsymbol{g}^\star(x)^\top\}^\top$; (4) $\boldsymbol{g}_4(x) = \{1, x, \boldsymbol{g}^\star(x)^\top\}^\top$. We can check (i) asymptotic behavior of EL and GMM estimators when $q < p_1 + 1 = p_2$ and $\boldsymbol{g}(\boldsymbol{x}) = \{1, \boldsymbol{h}(x)\}^\top$, for example, by comparing estimators using $\boldsymbol{h}_1(x)$ and $\boldsymbol{g}_1(x)$, because $\boldsymbol{g}_1(x) = \{1, \boldsymbol{h}_1(x)^\top\}^\top$. The parameter $\boldsymbol{\phi}$ is estimated by (2.9) when $q = p_1 + 1 = p_2$, and by (2.10) when $q < p_1 + 1 = p_2$, for which two-step GMM estimation procedure is used (Hansen et al., 1996). The algorithm can be easily implemented by using a package "gmm" in R programing language.

It is meaningless to compare the estimated parameters of the response model because it is misspecified in models M2–M6, and the true parameter value is unknown. Therefore, we assess the robustness by estimating $\theta = E(Y)$ with the estimated response model. It can be estimated by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{r_i y_i}{\pi(y_i; \hat{\boldsymbol{\phi}})} + \left\{ 1 - \frac{r_i}{\pi(y_i; \hat{\boldsymbol{\phi}})} \right\} \mathrm{E}^\star(Y \mid x_i; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}}) \right],$$

where

$$E^{\star}(Y \mid x; \boldsymbol{\phi}, \boldsymbol{\beta}) = \frac{E_1[Y\{1-\pi(Y;\boldsymbol{\phi})\}/\pi(Y;\boldsymbol{\phi})^2 \mid x]}{E_1[\{1-\pi(Y;\boldsymbol{\phi})\}/\pi(Y;\boldsymbol{\phi})^2 \mid x]}.$$

This estimator attains the semiparametric efficiency bound when both $\pi(y;\boldsymbol{\phi})$ and $f_1(y \mid x;\boldsymbol{\beta})$ are correctly specified; still has consistency even if $f_1(y \mid x;\boldsymbol{\beta})$ is misspecified by Theorem 2.4.1.

Under this setup, samples are generated with 2,000 sample size and 2,000 replications. The results with response mechanisms M1– M3 are shown in Figure 4.3, and M4–M6 in Figure 4.4. In the boxplots, samples which have not converged are removed in advance. For estimators using $\boldsymbol{g}_2(x)$ and $\boldsymbol{h}_2(x)$ with $\mu_1(x)$ mean function, almost half of the samples did not converge. As for the other estimators, most of estimators converged except for that using $\boldsymbol{h}_1(x)$ in M5 and $\boldsymbol{g}_4(x)$ with $\mu_1(x)$ in M3 and M4. This indicates difficulty of optimization of EL method and importance of correct specification of $f_1(y \mid x;\beta)$ model. As for estimation, all estimators are correctly estimated in M1, i.e., when the response model is correct. Even when the response model is misspecified, estimators using $\mu_2(x)$ are totally well estimated. Surprisingly, the estimators in M2 still work well though the response mechanism is MAR.

At first, we would like to note on (i). It can be inferred from the results of estimators using $\boldsymbol{g}_1(x)$ and $\boldsymbol{h}_1(x)$ under mechanisms M1–M4 that Theorem 2.5.1 does not hold when $q < p_1+1 = p_2$. This implies that the EL estimator does not attain the semiparametric efficiency bound. As for (ii), we can see that more constraints lead to less biased estimators by comparing estimators using $\boldsymbol{g}_2(x)$, $\boldsymbol{g}_3(x)$, and $\boldsymbol{g}_4(x)$ with $\mu_1(x)$ mean, though some estimators using the most constrained function $\boldsymbol{g}_4(x)$ were not estimated well. However, on the flip side, it follows from the results of estimators using $\boldsymbol{g}_4(x)$ with $\mu_2(x)$ mean that more constraints lead to difficulty in the optimization computation or lack of identification. Therefore, choosing an appropriate constraint function $\boldsymbol{k}(x)$ in $\boldsymbol{g}(x) = \{\boldsymbol{g}^{\star}(x)^{\top}, \boldsymbol{k}(x)^{\top}\}^{\top}$ from observed data is important.

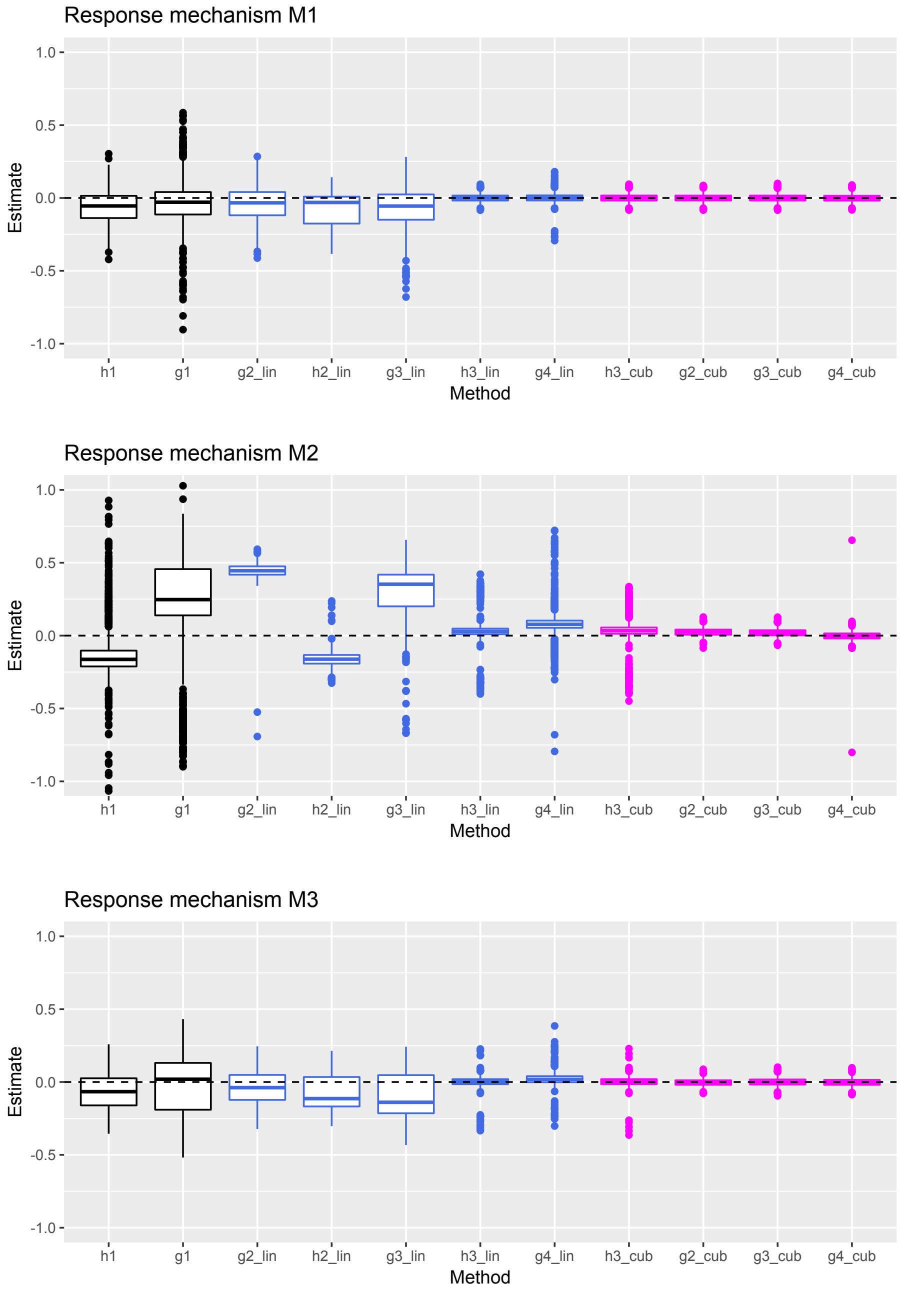

Figure 4.3: Boxplots of eleven estimators for $\mathrm{E}(Y)$ with M1–M3 response mechanisms. Colors stand for without $f_1(y \mid x)$ model for black, with mean function $\mu_1(x)$ (heavily misspecified model) for blue, and with $\mu_2(x)$ (slightly misspecified model) for pink boxplots. The first element of the method implies the constraint used for the estimators, and the second element “lin” and “cub”mean the linear and cubic models for the mean function of $f_1(y \mid x)$.

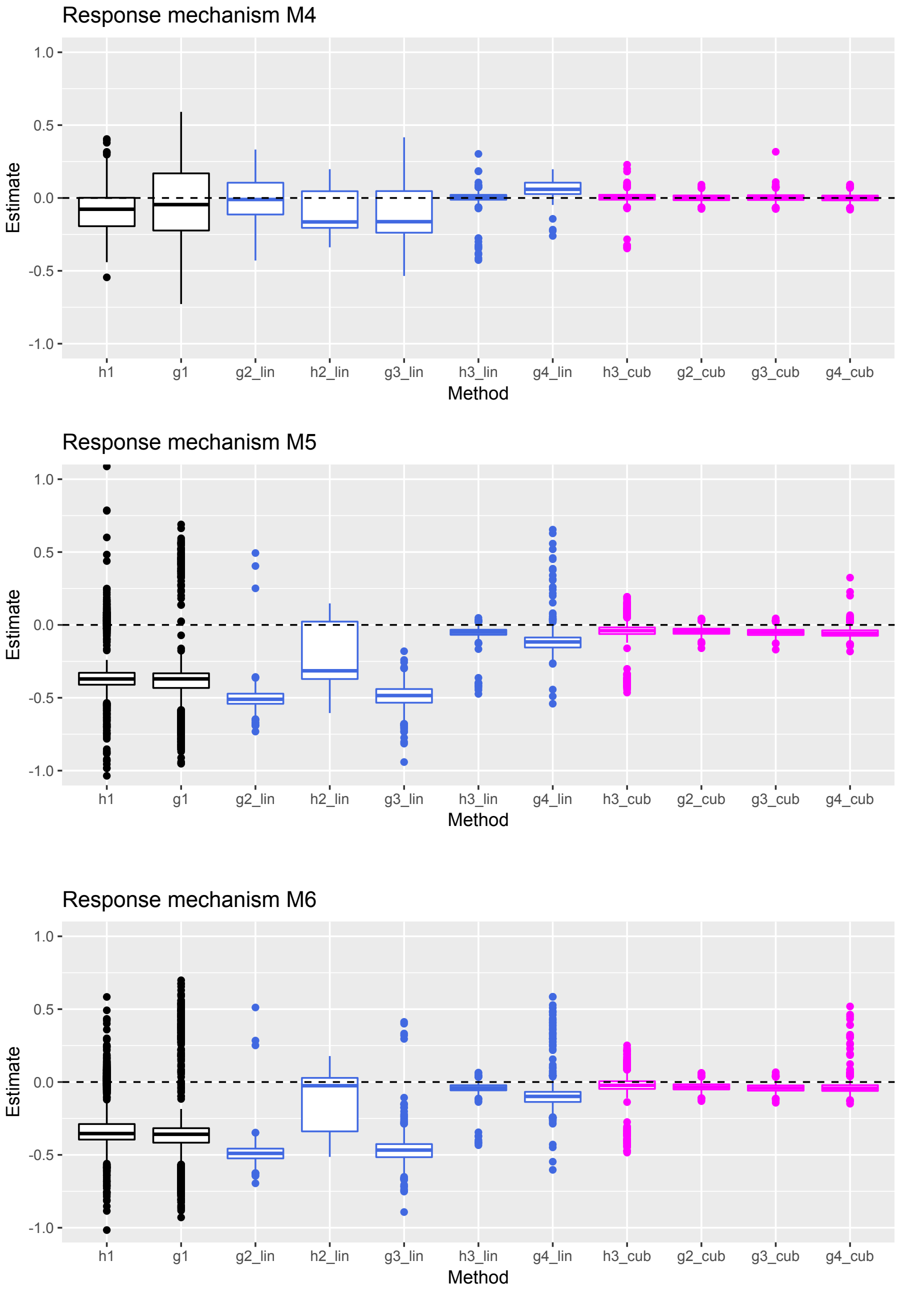

Figure 4.4: Boxplots of eleven estimators for $\mathrm{E}(Y)$ with M4–M6 response mechanisms. Colors stand for without $f_1(y \mid x)$ model for black, with mean function $\mu_1(x)$ (heavily misspecified model) for blue, and with $\mu_2(x)$ (slightly misspecified model) for pink boxplots. The first element of the method implies the constraint used for the estimators, and the second element "lin" and "cub"mean the linear and cubic models for the mean function of $f_1(y \mid x)$.

Chapter 5

Real data analysis

Two real data sets: (i) Korea labor and income panel survey(KLIPS) data; (ii) Depot medroxyprogesterone acetate(DMPA) data are analyzed by the methods proposed in Chapter 2 and 3.

5.1 Korea labor and income panel survey data

In this section, our proposed estimators are applied to the KLIPS data, which have been analyzed by many authors (Kim and Yu, 2011; Wang et al., 2014; Shao and Wang, 2016). The data contain $n = 2,506$ Korean wage earners; the response variable y is total wage income (10^6 Korean Won) in year 2008. There are three fully observed covariates: x_1: total wage income in the previous year (2007); x_2: gender; x_3: age. While x_1 is a continuous variable, x_2 has two categories 1 and 2 for male and female, and x_3 has three categories 1-3: $x_3 < 35$, $35 \leq x_3 < 51$, and $x_3 \geq 51$. We also identified three data points as outliers and excluded them from further analysis.

Although the data are completely observed, we took the approach of Kim and Yu (2011) and made eight artificial incomplete datasets by assuming the following eight response mechanisms: M1 (linear nonignorable without (x_2, x_3)): $\text{logit}(\pi) = 0.48 - 0.3x_1 - 0.5y$; M2 (linear nonignorable): $\text{logit}(\pi) = -0.85 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.4y$; M3 (nonlinear nonignorable, quadratic in x_1 without (x_2, x_3)): $\text{logit}(\pi) = 0.33 - 0.3x_1 - 0.1x_1^2 - 0.3y$; M4 (nonlinear nonignorable, quadratic in x_1): $\text{logit}(\pi) = -0.89 - 0.4x_1 - 0.1x_1^2 + 0.5x_2 + 0.2x_3 - 0.4y$; M5 (nonlinear nonignorable, quadratic in y without (x_2, x_3)): $\text{logit}(\pi) = 0.24 - 0.25x_1 - 0.25y - 0.1y^2$; M6 (nonlinear nonignorable, quadratic in y): $\text{logit}(\pi) = -0.93 - 0.2x_1 + 0.5x_2 + 0.2x_3 - 0.2y - 0.1y^2$; M7 (jump nonignorable without x) $\pi = 0.5I(y \leq 1.7) + 0.9(y > 1.7)$; M8 (jump nonignorable) $\pi = 0.5I(0.5x_2 + 0.2x_3 + y \leq 2.6) + 0.9(0.5x_2 + 0.2x_3 + y > 2.6)$, where $I(A)$ is the indicator function that takes 1(0) if an event A is true (false). Note that there are nonresponse instrumental variables for models M2, M4, M6, and M8. For all data sets, the response rate is about 70%. We estimated $\theta = E(Y)$ as considered in the simulation. The "true" average income in 2008 is $\hat{\theta}_n = 1.846$ as calculated using the complete data. In order to estimate the parameters, we assumed a response mechanism $\text{logit}\{\pi(x, y; \phi)\} = \phi_{\text{x}0} + \phi_{\text{x}1}x_1 + \phi_{\text{x}2}x_2 + \phi_{\text{x}3}x_3 + \phi_{\text{y}}y$. Therefore M1 and M2 are correctly specified while M3-M8 are misspecified.

We specified unknown f_1 models as normal distribution $Y \mid (x_1, x_2 = i, x_3 = j, r = 1) \sim \mathcal{N}(\mu_{i,j}(x_1), \sigma^2_{i,j})$ ($i = 1, 2$; $j = 1, 2, 3$), where $\mu_{i,j}(x_1) = \gamma_{0i,j} + \gamma_{1i,j}x_1 + \gamma_{2i,j}x_1^2 + \gamma_{3i,j}x_1^3 + \gamma_{4i,j}x_1^4$; $(\gamma_{1i,j}, \gamma_{2i,j}, \gamma_{3i,j}, \gamma_{4i,j})$ is the regression parameter when $(x_2, x_3) = (i, j)$. We chose the best model by AIC among 2^5-1 models for each (x_2, x_3)'s 2×3 pattern. By using Theorem 3.5.1, this model is identifiable as one of the 6 mean structures with $\mu_{i,j}$ being nonlinear with respect to x_1. However, in the real data, the correlation between x_1 and y is too high because wage income does not change considerably in one year; the mean structure is almost linear even when stratified by x_2 and x_3. Therefore, to obtain valid estimator of θ, we considered two different approaches: [1] find nonresponse instrumental variables used; [2] transform x_1 so that the relationship can be nonlinear. For the first approach, we specified x_2, x_3, and (x_2, x_3) as instrumental variables in applying our proposed method, which will lead to inconsistency for models M2 because there is actually no instrumental variable. For the second approach, we transformed x_1 to $\log(x_1)$ so that the relationship between y and $\log(x_1)$ is nonlinear. Although this transformation made the model identifiable, this also changed the assumed mechanism to $\text{logit}\{\pi(x, y; \phi)\} = \phi_{\text{x}0} + \phi_{\text{x}1}\log(x_1) + \phi_{\text{x}2}x_2 + \phi_{\text{x}3}x_3 + \phi_{\text{y}}y$. This may be a potential cause of biased estimation. On the flip side, this approach uses all information of covariates, which helps to reduce bias and gain efficiency. We show the result of this approach under both parametric and nonparametric f_1 models.

In Table 5.1, deviation of estimators from the full sample estimate $\hat{\theta}_n$ and estimated standard errors are shown. The methods using instrumental variable encountered some numerical problems even for correctly specified models: M3, M5, and M7. This is because the effect of the instrumental variables on the outcome variable is not so strong;

Table 5.1: $\hat{\theta} - \hat{\theta}_n$ (S.E. $(\hat{\theta})$): deviation of our proposed estimator $\hat{\theta}$ from the full sample estimate $\hat{\theta}_n = 1.846$ (and estimated their standard error) for datasets M1–M8 by two approaches: [1] using instrumental variable (IV); [2] using transformed x_1 with parametric (P) and nonparametric (NP) f_1 model. NA stands for not applicable due to numerical problems. All values are multiplied by 1,000.

	Approach				
	[1]			[2]	
IV	x_2	x_3	(x_2, x_3)	None	None
method	P	P	P	P	NP
M1	-8 (24)	16 (59)	14 (35)	-5 (23)	-6 (23)
M2	-73 (25)	-8 (23)	-25 (28)	-9 (23)	-4 (23)
M3	-25 (38)	NA (NA)	NA (NA)	-22 (23)	-20 (22)
M4	-19 (27)	13 (27)	-1 (24)	-4 (23)	5 (23)
M5	41 (206)	56 (373)	54 (563)	-10 (23)	-9 (23)
M6	23 (158)	-10 (31)	-57 (25)	-13 (24)	9 (23)
M7	26 (9500)	32 (7366)	36 (664)	10 (22)	-9 (23)
M8	50 (3985)	183 (NA)	128 (NA)	15 (25)	-18 (23)

the instrumental variable used is not useful enough. In terms of efficiency, the methods with transformed x_1 outperform by far those using instrumental variables. They are also more robust against misspecification of the response model.

5.2 Depot medroxyprogesterone acetate data

We analyzed the DMPA data by two approaches: (i) both $[\boldsymbol{Y} \mid X]$ and the missing-data mechanism $[\boldsymbol{R} \mid \boldsymbol{Y}, X]$ are parametric; (ii) only $[\boldsymbol{Y} \mid X]$ is parametric. For the missing-data mechanism, we assume an independence $R_t \perp X \mid (Y_1, \ldots, Y_t)$ $(t = 2, \ldots, T)$ because it is natural to consider the dose as having no effect on the response rate. In order to compare the results of the two approaches, we use the same parametric model for $[\boldsymbol{Y} \mid X]$, where $[\boldsymbol{Y} \mid X]$ is the AR(2) model with one covariate as shown in Figure 3.4.

5.2.1 Parametric approach

Parameters were estimated by maximizing the following likelihood:

$$
\begin{aligned}
& L_{1151}(\boldsymbol{\xi}) \\
&= \prod_{t=1}^{4} \prod_{i \in I_t} \operatorname{expit}\{(-1)^{y_{1i}+1}(\beta_{1,0} + \beta_1 x_i)\} \left\{ \prod_{s=2}^{t} \zeta_s(y_{s-1i}, y_{si}, x_i;\ \boldsymbol{\xi}_t) \right\}^{\mathbf{1}_{\{t \geq 2\}}} \\
&\quad \times \left\{ 1 - \mathbf{1}_{\{t \leq T-1\}} \sum_{y_{t+1}=0}^{1} \zeta_{t+1}\left(y_{ti}, y_{t+1}, x_i;\ \boldsymbol{\xi}_{t+1}\right) \right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\zeta_t(y_{t-1}, y_t, x;\ \boldsymbol{\xi}_t) &= \operatorname{expit}\{(-1)^{y_t+1}(\beta_{t,0} + \beta_{t,t-2} y_{t-2} + \beta_{t,t-1} y_{t-1} + \beta_t x)\} \\
&\quad \times \operatorname{expit}(\phi_{t,0} + \phi_{t,t-1} y_{t-1} + \phi_{t,t} y_t),
\end{aligned}
$$

$I_t = \{i : \boldsymbol{r}_i = \boldsymbol{r}_i^{(t)}\}$, and the function $\mathbf{1}_{\{A\}}$ is an indicator function which takes 1 (0) if an event A is true (false) and the sample size is $N =$ 1,151. The MLE of $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_4$ can be calculated independently since these parameters are separated in the above likelihood. We used the "optim" function in the programing language R to maximize $L_n(\boldsymbol{\xi})$.

The results of β_t show that the (direct) effect of the independent variable X on contraception is significantly different from zero at 6 and 9 months ($t = 2, 3$), which means that there are significant differences in effects on the contraception between 100 mg and 150 mg at these times based on the p-values. Matsuyama (2004) found the same results on the total effects of X. Next, the estimate of the serial correlation is strongly positive, which means once DMPA takes effect, it is also liable to have an effect at the next time point for both doses. Finally, $\boldsymbol{\phi}_2$, $\boldsymbol{\phi}_3$, $\boldsymbol{\phi}_4$, which are parameters on the missing-data mechanism, are difficult to interpret since the values of the standard deviation (SD) are larger than those of the other parameters though this model is theoretically identifiable. Due to this fact, all the parameters are not significantly different from zero. Next, we construct some sub-models of the full model to choose a best one by the likelihood test.

If data involve missing values, a naive information criterion such as AIC or BIC cannot be applied. Here, we choose a best model heuristically: the likelihood ratio test. First, we test the missing-data mechanism as "MCAR v.s. NMAR" and "MAR v.s. NMAR" following Diggle and Kenward (1994). Denote the MLE under the constraints

Table 5.2: Results of Parameter Estimation by Parametric AR(2) Model

Parameter	Estimate	SD	p-value
β_1	0.124	0.149	0.406
β_2	0.390	0.152	0.010
β_3	0.440	0.160	0.006
β_4	0.124	0.149	0.406
$\beta_{2,1}$	1.851	0.215	0.000
$\beta_{3,2}$	2.014	0.195	0.000
$\beta_{4,3}$	1.794	0.228	0.000
$\beta_{3,1}$	0.852	0.235	0.000
$\beta_{4,2}$	1.382	0.233	0.000
$\phi_{2,1}$	-0.506	0.680	0.457
$\phi_{3,2}$	-0.276	0.546	0.613
$\phi_{4,3}$	-1.067	0.506	0.035
$\phi_{2,2}$	-0.079	1.544	0.959
$\phi_{3,3}$	-0.719	1.231	0.559
$\phi_{4,4}$	0.939	0.931	0.313

$\phi_{2,1} = \phi_{2,2} = \phi_{3,2} = \phi_{3,3} = \phi_{4,3} = \phi_{4,4} = 0$ by $\hat{\boldsymbol{\xi}}_{MCAR}$, under the constraints $\phi_{2,2} = \phi_{3,3} = \phi_{4,4} = 0$ by $\hat{\boldsymbol{\xi}}_{MAR}$, and under no constraints by $\hat{\boldsymbol{\xi}}_{NMAR}$, i.e., the full-model MLE. We set the probability of type I error to 0.05 in the following two tests of the missing-data mechanisms. With these settings, the deviance between MCAR and NMAR is

$$-2\ln\frac{L_{1151}(\hat{\boldsymbol{\xi}}_{MCAR})}{L_{1151}(\hat{\boldsymbol{\xi}}_{NMAR})} = 27.157 > \chi^2_6(0.05) = 12.592,$$

and the deviance between MAR and NMAR is

$$-2\ln\frac{L_{1151}(\hat{\boldsymbol{\xi}}_{MAR})}{L_{1151}(\hat{\boldsymbol{\xi}}_{NMAR})} = 1.204 < \chi^2_3(0.05) = 7.814.$$

Following the results, the missing-data mechanism is not MCAR and probably not NMAR. Now, we study model selection in more detail. We consider the sub-models with at most three parameters selected from those in ($\phi_{2,1}$, $\phi_{2,2}$, $\phi_{3,2}$, $\phi_{3,3}$, $\phi_{4,3}$, $\phi_{4,4}$): ${}_6C_1 + {}_6C_2 + {}_6C_3 = 41$ as shown in Table 5.3 The reason why we consider only three parameters is the deviance being not so different from that when we add more than three. As indicated by bold sub-model numbers in Table 5.3, the sub-models with the smallest deviance among the set of models with the same number of parameters are Nos. 4, 9, and 29 for the sets with one, two, and three parameters, respectively. We see that all the sub-models with one or two parameters are rejected according to the deviance, while the deviance relatively clearly identifies accepted sub-models with three parameters. The deviances of sub-models No. 27, 29, 33, and 35 are small and not significantly different from each other. Therefore, we assert that these four models, whose parameters are chosen one at each period, are preferable. We could not say which sub-model is the best of all the four sub-models based on the analysis.

5.2.2 Semiparametric approach

We estimate the same outcome model without specifying any missing-data mechanism. The likelihood is given by

$$\begin{aligned} &L^{\mathrm{SP}}_{1151}(\boldsymbol{\xi}) \\ &= \prod_{t=1}^{4}\prod_{i\in I_t} \mathrm{expit}\{(-1)^{y_{1i}+1}(\beta_{1,0}+\beta_1 x_i)\}\left\{\prod_{s=2}^{t}\zeta^{\mathrm{SP}}_s(\boldsymbol{v}_{si}, x_i;\ \boldsymbol{\xi}_t)\right\}^{\mathbf{1}_{\{t\geq 2\}}} \\ &\quad\times\left\{1-\mathbf{1}_{\{t\leq T-1\}}\sum_{y_{t+1}=0}^{1}\zeta^{\mathrm{SP}}_{t+1}\left(\boldsymbol{v}_{ti}, y_{t+1}, x_i;\ \boldsymbol{\xi}_{t+1}\right)\right\}, \end{aligned}$$

Table 5.3: Sub-model deviance

No.	Parameter	Deviance	No.	parameter	deviance
1	$\phi_{2,1}$	18.670	22	$\phi_{2,1}, \phi_{2,2}, \phi_{3,2}$	7.744
2	$\phi_{2,2}$	19.268	23	$\phi_{2,1}, \phi_{2,2}, \phi_{3,3}$	7.670
3	$\phi_{3,2}$	16.242	24	$\phi_{2,1}, \phi_{2,2}, \phi_{4,3}$	12.124
4	$\phi_{3,3}$	16.160	25	$\phi_{2,1}, \phi_{2,2}, \phi_{4,4}$	14.290
5	$\phi_{4,3}$	20.616	26	$\phi_{2,1}, \phi_{3,2}, \phi_{3,3}$	7.390
6	$\phi_{4,4}$	22.802	27	$\phi_{2,1}, \phi_{3,2}, \phi_{4,3}$	1.204
7	$\phi_{2,1}, \phi_{2,2}$	18.666	28	$\phi_{2,1}, \phi_{3,2}, \phi_{4,4}$	3.392
8	$\phi_{2,1}, \phi_{3,2}$	7.748	**29**	$\phi_{2,1}, \phi_{3,3}, \phi_{4,3}$	1.130
9	$\phi_{2,1}, \phi_{3,3}$	7.672	30	$\phi_{2,1}, \phi_{3,3}, \phi_{4,4}$	3.316
10	$\phi_{2,1}, \phi_{4,3}$	12.128	31	$\phi_{2,1}, \phi_{4,3}, \phi_{4,4}$	11.282
11	$\phi_{2,1}, \phi_{4,4}$	14.314	32	$\phi_{2,2}, \phi_{3,2}, \phi_{3,3}$	7.988
12	$\phi_{2,2}, \phi_{3,2}$	8.346	33	$\phi_{2,2}, \phi_{3,2}, \phi_{4,3}$	1.804
13	$\phi_{2,2}, \phi_{3,3}$	8.272	34	$\phi_{2,2}, \phi_{3,2}, \phi_{4,4}$	3.990
14	$\phi_{2,2}, \phi_{4,3}$	12.726	35	$\phi_{2,2}, \phi_{3,3}, \phi_{4,3}$	1.728
15	$\phi_{2,2}, \phi_{4,4}$	14.912	36	$\phi_{2,2}, \phi_{3,3}, \phi_{4,4}$	3.916
16	$\phi_{3,2}, \phi_{3,3}$	15.878	37	$\phi_{2,2}, \phi_{3,3}, \phi_{4,4}$	11.880
17	$\phi_{3,2}, \phi_{4,3}$	9.692	38	$\phi_{2,2}, \phi_{4,3}, \phi_{4,4}$	9.336
18	$\phi_{3,2}, \phi_{4,4}$	11.878	39	$\phi_{3,2}, \phi_{3,3}, \phi_{4,3}$	11.522
19	$\phi_{3,3}, \phi_{4,3}$	9.618	40	$\phi_{3,2}, \phi_{3,3}, \phi_{4,4}$	8.846
20	$\phi_{3,3}, \phi_{4,4}$	11.804	41	$\phi_{3,2}, \phi_{4,3}, \phi_{4,4}$	8.772
21	$\phi_{4,3}, \phi_{4,4}$	19.770			

where $\boldsymbol{v}_t = (y_1, \ldots, y_t)^\top$,

$$\zeta_t^{\mathrm{SP}}(\boldsymbol{v}_t, x;\ \boldsymbol{\xi}_t) = \mathrm{expit}\{(-1)^{y_t+1}(\beta_{t,0} + \beta_{t,1}y_1 + \ldots + \beta_{t,t-1}y_{t-1} + \beta_t x)\}\phi_{t,y_1\ldots y_t},$$

and $\phi_{t,y_1\ldots y_t}$ is the nonparametric model for $\Pr(R_t = 1 \mid y_1, \ldots, y_t)$. This model is identifiable in view of §3.4.1. The parameters can be estimated by maximizing the likelihood $L_{1151}^{\mathrm{SP}}(\boldsymbol{\xi})$.

Results are shown in Table 5.4. We can see that the results are not very different from those with the parametric missing-data mechanism, but the SD is larger, especially at the third period. This is due to the additional parameters for the missing-data mechanism and poor information of dichotomous data. A parametric model for the missing-data mechanism might be desirable in such a case. Also the likelihood ratio test can be used in a similar way to as in §5.2.1, though we omitted the results of the likelihood ratio test since they are almost the same as in §5.2.1. If we used Zhao and Shao (2015)'s approach, the parameters could also be estimated without specifying any missing-data mechanism. However, the likelihood ratio test could not be used because the semiparametric approach is not based on the likelihood. Therefore, use of the semiparametric approach is more limited than that of our approach.

Table 5.4: Results of parameter estimation by the semiparametric AR(2) model

Parameter	Estimate	SD	p-value
β_1	0.124	0.149	0.406
β_2	0.402	0.148	0.007
β_3	0.437	0.163	0.007
β_4	-0.135	0.169	0.417
$\beta_{2,1}$	2.454	0.187	0.000
$\beta_{3,2}$	1.900	0.711	0.007
$\beta_{4,3}$	1.784	0.244	0.000
$\beta_{3,1}$	0.980	0.705	0.170
$\beta_{4,2}$	1.126	0.275	0.000
$\beta_{4,1}$	0.573	0.283	0.043

Chapter 6

Discussion

We have proposed three semiparametric estimators with nonignorable nonresponse data without assuming any parametric outcome model. The first one is a kernel-based semiparametric estimator, where semiparametric maximum likelihood is used to estimate the response mechanism parameters. While the method of maximum likelihood usually produces efficient estimators, it has been found that the semiparametic estimator does not attain the semiparametric efficiency bound. Thus we have improved the first one to create the other two semiparametric estimators which attain the lower bound. The relationship between the two semiparametric methods, EL and GMM approaches, for estimation of a response model have been also investigated. We have shown that estimators obtained with EL can be solved by using the moment based method when $q = p_1 + 1 = p_2$, where p_1 and p_2 are defined in §2.2. In addition, we have pointed out the two estimators may not be asymptotically equivalent when $q < p_1 + 1 = p_2$ through numerical study. Furthermore, we have shown a constraint function including the best function $g^{\star}(\boldsymbol{x})$ also attains the semiparametric efficiency bound (Rotnitzky and Robins, 1997). This property is useful in practice because the estimated response model is robust for model misspecification due to other constraints, and the true response mechanism is generally unknown.

However, as shown in the simulation study, more constraints may lead to lack of identification. Thus, it is necessary to develop a method to choose an appropriate constraint function. Also, in this paper, we restrict that a parametric response model is fully specified, though some semiparametric models

$$\pi(\boldsymbol{x}, y; \phi) = \frac{1}{1 + \exp\{w(\boldsymbol{x}) + \phi y\}},$$

have been already proposed, where $w(\boldsymbol{x})$ is an arbitrary function of $\boldsymbol{x}$. Sun et al. (in press) derive the semparametric efficiency bound for the semiparametric model. Proposal of an adaptive and robust estimator for the model will also be a topic of our future research. Alternatively, developing an appropriate information criterion for choosing the response mechanism will be a future research work. Instead of specifying a single response model, one can consider multiple response models, which possesses consistency property of model selection when one of the specified response models is correct. This multiple robustness property has been investigated in the ignorable nonresponse setup (Han, 2014; Chen and Haziza, 2017). Extension of multiple robustness to the nonignorable nonresponse case will also be a topic of our future research.

As for the model identification for nonignorable nonresponse, we have proposed some conditions under which a model for repeated measurement data with NMAR missingness is identifiable. Two different situations have been distinguished: (i) both the outcome model and the missing-data mechanism are correctly specified; (ii) either of the two models is correctly specified and the other model is unspecified. Even if the two models are parametric, the models may be unidentifiable (Miao et al., 2016). We have given an example of an identifiable model, that is, the AR(p) model which depends on the history only through the previous p responses. Then, an easy-to-check necessary and sufficient condition on the identification of the model has been proposed. For example, it can be proved from the derived condition that even the Logistic AR(1) model is not identified, the model becomes identifiable when there are covariates as in Figure 3.4 or there are no missing values at the first two periods. The estimation confines a situation where the outcome model is correctly specified. This naturally holds only when both response and covariates variables are discrete. Therefore, an alternative condition depending on $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}) := f(y \mid \boldsymbol{x}, r = 1; \boldsymbol{\gamma})$, but not depending on $f(y \mid \boldsymbol{x})$ have been developed. This condition is attractive because the nuisance parameter $\boldsymbol{\gamma}$ can be estimated separately from $\boldsymbol{\phi}$, and the condition can be easily checked from observed data.

Our identification condition is useful only for conditional models with repeated measurement data. Another extension giving concise conditions to check identifiability for more complicated models such as marginal models and hybrid models (Molenberghs and Verbeke, 2005; Wilkins and Fitzmaurice, 2006) need to be developed.

Appendix A

Regularity conditions

To discuss asymptotic properties of our estimator, we assume the following regularity conditions. Here and throughout, $\|\cdot\|$ denotes the Euclidean norm. For example, for a matrix $A = [a_{ij}]$, $\|A\| = (\sum_{i,j} a_{ij}^2)^{1/2}$.

C1.1. Φ and Θ are compact.

C1.2. Let

$$\hat{\mathbf{S}}^l(\boldsymbol{\phi}) := \left\{1 - \frac{r}{\pi(\boldsymbol{\phi}; \boldsymbol{z})}\right\} \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x})$$

and

$$\hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta) := \frac{r}{\pi(\boldsymbol{\phi}; \boldsymbol{z})} U(\theta; \boldsymbol{x}) + \left\{1 - \frac{r}{\pi(\boldsymbol{\phi}; \boldsymbol{z})}\right\} \bar{U}(\boldsymbol{\phi}, \theta; \boldsymbol{x}).$$

Both $\hat{\mathbf{S}}^l(\boldsymbol{\phi})$ and $\hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)$ are continuously differentiable at each $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ with probability one, and there exists a $d(\boldsymbol{z}, r)$ with $\|(\hat{\mathbf{S}}^l(\boldsymbol{\phi})^\top,\ \hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)^\top)^\top\| \leq d(\boldsymbol{z}, r)$ for all $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ and $E\{d(\boldsymbol{Z}, R)\} < \infty$.

C1.3. The infimum $\inf_{\boldsymbol{\phi}, \theta} E\{\hat{\mathbf{S}}^l(\boldsymbol{\phi}),\ \hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)\} = (\mathbf{0}, 0)$ attains only at $(\boldsymbol{\phi}_0, \theta_0) \in \Phi \times \Theta$.

C1.4. $\partial\{\hat{\mathbf{S}}^l(\boldsymbol{\phi})^\top, \hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)^\top\}^\top / \partial(\boldsymbol{\phi}^\top, \theta^\top)$ is continuous at $(\boldsymbol{\phi}_0, \theta_0)$ with probability one, and there is a neighborhood $\Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}}$ of $(\boldsymbol{\phi}_0, \theta_0)$ such that

$$E\left\{\sup_{(\boldsymbol{\phi}, \theta) \in \Phi_{\mathcal{N}} \times \Theta_{\mathcal{N}}} \left\| \frac{\partial\{\hat{\mathbf{S}}^l(\boldsymbol{\phi})^\top, \hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)^\top\}^\top}{\partial(\boldsymbol{\phi}^\top, \theta^\top)} \right\| \right\} < \infty.$$

C1.5. $E[\partial\{\hat{\mathbf{S}}^l(\boldsymbol{\phi})^\top, \hat{\mathrm{U}}^l(\boldsymbol{\phi}, \theta)^\top\}^\top / \partial(\boldsymbol{\phi}^\top, \theta^\top)]$ is nonsingular at $(\boldsymbol{\phi}_0, \theta_0)$.

C1.6. The kernel $K(u)$ is bounded and satisfies $\int K(u)\mathrm{d}u = 1$, and has zero moments of order $\leq m-1$ and a nonzero mth order moment.

C1.7. Let $b_1(\boldsymbol{z};\boldsymbol{\phi}) = O(\boldsymbol{z};\boldsymbol{\phi})$, $b_2(\boldsymbol{z};\boldsymbol{\phi}) = O(\boldsymbol{z};\boldsymbol{\phi})s_0(\boldsymbol{z};\boldsymbol{\phi})$, and $b_3(\boldsymbol{z};\boldsymbol{\phi},\theta) = O(\boldsymbol{z};\boldsymbol{\phi})U(\boldsymbol{z};\theta)$, and $\mathfrak{X}$ be a compact set that is contained in the support of $\boldsymbol{x}$. Supposed that $E\{Rb_i(\boldsymbol{x},Y;\boldsymbol{\phi},\theta) \mid \boldsymbol{x}\}$ for $i=1,2,3$ are differentiable with respect to $\boldsymbol{x}$ up to order m for each $\boldsymbol{\phi}\in\Phi$ and $\theta\in\Theta$. Also, assume that, for $i=1,2,3$,

$$\begin{aligned}
&\sup_{(\boldsymbol{\phi},\theta)\in\Phi\times\Theta} \|E\{Rb_i(\boldsymbol{x},Y;\boldsymbol{\phi};\theta) \mid \boldsymbol{x}\}\|_m \\
:= &\sup_{(\boldsymbol{\phi},\theta)\in\Phi\times\Theta} \max_{m_1+\ldots+m_d=m} \sup_{\boldsymbol{x}\in\mathfrak{X}} \left\| \frac{\partial^m E\{Rb_i(\boldsymbol{x},Y;\boldsymbol{\phi};\theta) \mid \boldsymbol{x}\}}{\partial x_1^{m_1}\ldots\partial x_d^{m_1}} \right\| < \infty.
\end{aligned}$$

C1.8. There exists a $v>2$ such that, for $i=1,2,3$, and for each $\boldsymbol{\phi}\in\Phi$ and $\theta\in\Theta$, $E\{R\|b_i(\boldsymbol{Z};\boldsymbol{\phi};\theta)\|^v\}<\infty$ and $E\{R\|b_i(\boldsymbol{x},Y;\boldsymbol{\phi};\theta)\|^v \mid \boldsymbol{x}\}$ is bounded.

C1.9. The bandwidth h satisfies $h\to 0$, $n^{1-(2/v)}h^d/\ln n\to\infty$, $\sqrt{n}h^d/\ln n\to\infty$, and $\sqrt{n}h^{2m}\to 0$, where d is the dimension of $\boldsymbol{x}$, and m are given in the condition (C1.8), and v is given in the condition (C1.10).

C1.10. There exist $|b_i(\boldsymbol{z};\boldsymbol{\phi},\theta)|\leq \tilde{b}_i(\boldsymbol{z})$ for $i=1,2,3$ such that $\sup_{\boldsymbol{x}\in\mathcal{X}} E\{R\tilde{b}_i(\boldsymbol{Z}) \mid \boldsymbol{x}\} < \infty$.

C1.11. Let

$$\boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{z}) := \frac{E_1\{O(\boldsymbol{\phi}_0;\boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi};\boldsymbol{z}) \mid \boldsymbol{x}\}} O(\boldsymbol{\phi};\boldsymbol{z})\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{z}) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x})\right\}$$

and

$$H_1(\boldsymbol{\phi},\theta;\boldsymbol{z}) := \frac{E_1\{O(\boldsymbol{\phi}_0;\boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi};\boldsymbol{z}) \mid \boldsymbol{x}\}} O(\boldsymbol{\phi};\boldsymbol{z})\left\{U(\theta;\boldsymbol{z}) - \bar{U}(\boldsymbol{\phi},\theta;\boldsymbol{x})\right\}.$$

The condition (C1.2) holds for $\boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{z})$, $H_1(\boldsymbol{\phi},\theta;\boldsymbol{z})$, $\partial\boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{z})/\partial\boldsymbol{\phi}$, $\partial H_1(\boldsymbol{\phi},\theta;\boldsymbol{z})/\partial\boldsymbol{\phi}$, and $\partial H_1(\boldsymbol{\phi},\theta;\boldsymbol{z})/\partial\theta$ instead of $\hat{\mathbf{S}}^l(\boldsymbol{\phi})$ and $\hat{\mathrm{U}}^l(\boldsymbol{\phi},\theta)$.

C1.12. For functions, $\partial b_1(\boldsymbol{z};\boldsymbol{\phi})/\partial\boldsymbol{\phi}$, $\partial b_2(\boldsymbol{z};\boldsymbol{\phi})/\partial\boldsymbol{\phi}^\top$, $\partial b_3(\boldsymbol{z};\boldsymbol{\phi},\theta)/\partial\boldsymbol{\phi}$, and $\partial b_3(\boldsymbol{z};\boldsymbol{\phi},\theta)/\partial\theta$, (C1.7), (C1.8), and (C1.10) hold instead of $b_i(\boldsymbol{z};\boldsymbol{\phi},\theta)$.

C2.1. Φ and Θ are compact.

C2.2. Γ is compact, $\boldsymbol{S}_\gamma(\boldsymbol{\gamma}) = \partial \log f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}$ is continuously differentiable at $\boldsymbol{\gamma} \in \Gamma$ with probability one, there exists $e(W)$ such that $\|\boldsymbol{S}_\gamma(\boldsymbol{\gamma})\| \leq e(W)$ for all $\boldsymbol{\gamma} \in \Gamma$ and $E\{e(W)\} < \infty$, $E\{\boldsymbol{S}_\gamma(\boldsymbol{\gamma})\} = 0$ has a unique solution $\boldsymbol{\gamma}^* \in \Gamma$, $\partial \boldsymbol{S}_\gamma(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^\top$ is continuous at $\boldsymbol{\gamma}^*$ with probability one, and there is a neighborhood $\Gamma_{\mathfrak{N}}$ of $\boldsymbol{\gamma}^*$ such that $\|E\{\sup_{\boldsymbol{\gamma} \in \Gamma_{\mathfrak{N}}} \partial \boldsymbol{S}_\gamma(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}^\top\}\| < \infty$.

C2.3. The infimum $\inf_{\boldsymbol{\phi},\theta} E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}^*)\} = (\mathbf{0}, 0)$ attains only at $(\boldsymbol{\phi}_0, \theta_0) \in \Phi \times \Theta$, where $\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}) = (\boldsymbol{S}_1(\boldsymbol{\phi}, \boldsymbol{\gamma})^\top, \boldsymbol{S}_2(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}))^\top$ defined in (2.22).

C2.4. $\partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma})/\partial(\boldsymbol{\phi}^\top, \theta, \boldsymbol{\gamma}^\top)$ is continuous at $(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*)$ with probability one, and there is a neighborhood $\Phi_{\mathfrak{N}} \times \Theta_{\mathfrak{N}} \times \Gamma_{\mathfrak{N}}$ of $(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*)$ such that

$$\|E\{\sup_{(\boldsymbol{\phi},\theta,\boldsymbol{\gamma}^*) \in \Phi_{\mathfrak{N}} \times \Theta_{\mathfrak{N}} \times \Gamma_{\mathfrak{N}}} \partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma})/\partial(\boldsymbol{\phi}^\top, \theta, \boldsymbol{\gamma}^\top)\}\| < \infty.$$

C2.5. $\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma})$ is continuously differentiable at each $(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}) \in \Phi \times \Theta \times \Gamma$ with probability one, and there exists $d_1(W)$ such that $\|\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma})\| \leq d_1(W)$ for all $(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}) \in \Phi \times \Theta \times \Gamma$ and $E\{d_1(W)\} < \infty$.

C2.6. $E\{\partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \boldsymbol{\gamma}^*)/\partial(\boldsymbol{\phi}^\top, \theta, \boldsymbol{\gamma}^\top)\}$ is nonsingular at $(\boldsymbol{\phi}_0, \theta_0, \boldsymbol{\gamma}^*)$.

C2.7. The conditions (C2.1)-(C2.6) hold for known distribution $f_1(y \mid \boldsymbol{x}; \gamma_0)$, i.e., $E\{\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta, \gamma_0)\} = 0$ has a unique solution $(\boldsymbol{\phi}_0, \theta_0) \in \Phi \times \Theta$, where $\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta) = (\boldsymbol{S}_1(\boldsymbol{\phi}, \gamma_0)^\top, \boldsymbol{S}_2(\boldsymbol{\phi}, \theta, \gamma_0))^\top$; $\partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)/\partial(\boldsymbol{\phi}^\top, \theta)$ is continuous at $(\boldsymbol{\phi}_0, \theta_0)$ with probability one, and there is a neighborhood $\Phi_{\mathfrak{N}} \times \Theta_{\mathfrak{N}}$ of $(\boldsymbol{\phi}_0, \theta_0)$ such that

$$\|E\{\sup_{(\boldsymbol{\phi},\theta) \in \Phi_{\mathfrak{N}} \times \Theta_{\mathfrak{N}}} \partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)/\partial(\boldsymbol{\phi}^\top, \theta)\}\| < \infty;$$

$\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)$ is continuously differentiable at each $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ with probability one, and there exists $d_2(W)$ such that $\|\boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)\| \leq d_2(W)$ for all $(\boldsymbol{\phi}, \theta) \in \Phi \times \Theta$ and $E\{d_2(W)\} < \infty$; $E\{\partial \boldsymbol{S}_{\text{eff}}(\boldsymbol{\phi}, \theta)/\partial(\boldsymbol{\phi}^\top, \theta)\}$ is nonsingular at $(\boldsymbol{\phi}_0, \theta_0)$.

C2.8. The kernel $K(u)$ is bounded and satisfies $\int K(u) \mathrm{d}u = 1$, and has zero moments of order $\leq m - 1$ and a nonzero mth order moment.

C2.9. Let $c_i(\boldsymbol{z}; \boldsymbol{\phi}) = \pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi}) b_i(\boldsymbol{z}; \boldsymbol{\phi})$ for $i = 1, 2, 3$, and $\mathfrak{X}$ be a compact set that is contained in the support of $\boldsymbol{x}$. Supposed that $E\{R c_i(\boldsymbol{x}, Y; \boldsymbol{\phi}, \theta) \mid \boldsymbol{x}\}$ for $i = 1, 2, 3$

are differentiable with respect to $\boldsymbol{x}$ up to order m for each $\boldsymbol{\phi} \in \Phi$ and $\theta \in \Theta$. Also, assume that, for $i = 1, 2, 3$ and for each θ and $\boldsymbol{\phi}$,

$$\begin{aligned} &\sup_{(\boldsymbol{\phi},\theta)\in\Phi\times\Theta} \|E\{Rb_i(\boldsymbol{x}, Y; \boldsymbol{\phi}; \theta) \mid \boldsymbol{x}\}\|_m \\ &:= \sup_{(\boldsymbol{\phi},\theta)\in\Phi\times\Theta} \max_{m_1+\ldots+m_d=m} \sup_{\boldsymbol{x}\in\mathcal{X}} \left\| \frac{\partial^m E\{Rb_i(\boldsymbol{x}, Y; \boldsymbol{\phi}; \theta) \mid \boldsymbol{x}\}}{\partial x_1^{m_1} \ldots \partial x_d^{m_1}} \right\| < \infty. \end{aligned}$$

C2.10. There exists a $v > 2$ such that, for $i = 1, 2, 3$, $E_1\{\|c_i(\boldsymbol{Z}; \boldsymbol{\phi}; \theta)\|^v\} < \infty$ and $E_1\{\|c_i(\boldsymbol{x}, Y; \boldsymbol{\phi}; \theta)\|^v \mid \boldsymbol{x}\}f_1(\boldsymbol{x})$ is bounded.

C2.11. The bandwidth h satisfies $h \to 0$, $n^{1-(2/v)}h^d/\ln n \to \infty$, $\sqrt{n}h^d/\ln n \to \infty$, and $\sqrt{n}h^{2m} \to 0$.

C2.12. There exist $|c_i(\boldsymbol{z}; \boldsymbol{\phi}, \theta)| \leq \tilde{c}_i(\boldsymbol{z})$ for $i = 1, 2, 3$ such that $\sup_{\boldsymbol{x}\in\mathcal{X}} E\{R\tilde{c}_i(\boldsymbol{Z}) \mid \boldsymbol{x}\} < \infty$.

C2.13. Let $\boldsymbol{I}(\boldsymbol{\phi}; \boldsymbol{z}_i) = I_1(\boldsymbol{\phi}; \boldsymbol{x}_i)\boldsymbol{I}_2(\boldsymbol{\phi}; \boldsymbol{z}_i)$, $J(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i) = I_1(\boldsymbol{\phi}; \boldsymbol{x}_i)J_2(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i)$, and

$$\begin{aligned} I_1(\boldsymbol{\phi}; \boldsymbol{x}_i) &= 1 - E\left\{ \frac{\pi(\boldsymbol{\phi}_0; \boldsymbol{Z})}{\pi(\boldsymbol{\phi}; \boldsymbol{Z})} \;\middle|\; \boldsymbol{x}_i \right\}, \\ \boldsymbol{I}_2(\boldsymbol{\phi}; \boldsymbol{z}_i) &= \frac{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{z}_i)O(\boldsymbol{\phi}; \boldsymbol{z}_i)\{\boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{z}_i) - g^\star(\boldsymbol{\phi}; \boldsymbol{x}_i)\}}{E_1\{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{Z})O(\boldsymbol{\phi}; \boldsymbol{Z}) \mid \boldsymbol{x}_i\} \Pr(R = 1 \mid \boldsymbol{x}_i)}, \\ J_2(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i) &= \frac{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{z}_i)O(\boldsymbol{\phi}; \boldsymbol{z}_i)\{U(\theta; \boldsymbol{z}_i) - U^\star(\theta, \boldsymbol{\phi}; \boldsymbol{x}_i)\}}{E_1\{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{Z})O(\boldsymbol{\phi}; \boldsymbol{Z}) \mid \boldsymbol{x}_i\} \Pr(R = 1 \mid \boldsymbol{x}_i)}. \end{aligned}$$

and

$$H_1(\boldsymbol{\phi}, \theta; \boldsymbol{z}) := \frac{E_1\{O(\boldsymbol{\phi}_0; \boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi}; \boldsymbol{z}) \mid \boldsymbol{x}\}} O(\boldsymbol{\phi}; \boldsymbol{z}) \left\{U(\theta; \boldsymbol{z}) - \bar{U}(\boldsymbol{\phi}, \theta; \boldsymbol{x})\right\}.$$

Then, the condition (C2.5) holds for $\boldsymbol{I}(\boldsymbol{\phi}; \boldsymbol{z})$, $J(\boldsymbol{\phi}, \theta; \boldsymbol{z})$, $\partial\boldsymbol{I}(\boldsymbol{\phi}; \boldsymbol{z})/\partial\boldsymbol{\phi}$, $\partial J(\boldsymbol{\phi}, \theta; \boldsymbol{z})/\partial\boldsymbol{\phi}$, and $\partial J(\boldsymbol{\phi}, \theta; \boldsymbol{z})/\partial\theta$ for known distribution $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}_0)$ instead of $\boldsymbol{S}_1(\boldsymbol{\phi})$, $\boldsymbol{S}_2(\boldsymbol{\phi}, \theta)$.

C2.14. For functions $\partial c_1(\boldsymbol{z}; \boldsymbol{\phi})/\partial\boldsymbol{\phi}$, $\partial c_2(\boldsymbol{z}; \boldsymbol{\phi})/\partial\boldsymbol{\phi}^\top$, $\partial c_3(\boldsymbol{z}; \boldsymbol{\phi}, \theta)/\partial\boldsymbol{\phi}$, and $\partial c_3(\boldsymbol{z}; \boldsymbol{\phi}, \theta)/\partial\theta$, (C2.9), (C2.10), and (C2.12) hold instead of $c_i(\boldsymbol{z}; \boldsymbol{\phi}, \theta)$.

Appendix B

Proofs of the technical results

Following uniform convergence of the kernel estimators used in Chapter 2 is obtained.

Theorem B.2.1. It follows from conditions (C1.1), (C1.2) and (C1.6)–(C1.10) that

$$\sup_{\boldsymbol{\phi}\in\Phi}\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{D}(\boldsymbol{\phi};\boldsymbol{x})-D(\boldsymbol{\phi};\boldsymbol{x})\|=o_p(n^{-1/4}), \tag{B.1}$$

$$\sup_{\boldsymbol{\phi}\in\Phi}\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{\boldsymbol{C}}_s(\boldsymbol{\phi};\boldsymbol{x})-\boldsymbol{C}_s(\boldsymbol{\phi};\boldsymbol{x})\|=o_p(n^{-1/4}),$$

$$\sup_{(\boldsymbol{\phi},\theta)\in\Phi\times\Theta}\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{C}_U(\boldsymbol{\phi},\theta;\boldsymbol{x})-C_U(\boldsymbol{\phi},\theta;\boldsymbol{x})\|=o_p(n^{-1/4}),$$

where

$$\begin{aligned}D(\boldsymbol{\phi};\boldsymbol{x})&=E_1\{O(\boldsymbol{\phi};\boldsymbol{x},Y)\mid\boldsymbol{x}\}f_1(\boldsymbol{x})\Pr(R=1),\\ \boldsymbol{C}_s(\boldsymbol{\phi};\boldsymbol{x})&=E_1\{O(\boldsymbol{\phi};\boldsymbol{x},Y)\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x},Y)\mid\boldsymbol{x}\}f_1(\boldsymbol{x})\Pr(R=1),\\ C(\boldsymbol{\phi};\boldsymbol{x})&=E_1\{U(\boldsymbol{\phi},\theta;\boldsymbol{x},Y)\mid\boldsymbol{x}\}f_1(\boldsymbol{x})\Pr(R=1),\end{aligned}$$

and $\hat{D}(\boldsymbol{\phi};\boldsymbol{x})$, $\hat{\boldsymbol{C}}_s(\boldsymbol{\phi};\boldsymbol{x})$, and $\hat{C}_U(\boldsymbol{\phi},\theta;\boldsymbol{x})$ are defined in §3.4.

In order to show the Theorem B.2.1, the following Lemma B.2.1 proved in Newey (1994) is to be referred. In what follows, only (B.1) is proved. Convergence of the other functions can be shown in a similar way.

Lemma B.2.1. For a fixed $\boldsymbol{\phi}\in\Phi$, under conditions (C1.1) and (C1.6)–(C1.8), it holds that

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|\hat{D}(\boldsymbol{\phi};\boldsymbol{x})-E\{\hat{D}(\boldsymbol{\phi};\boldsymbol{x})\}\|=O_p\{(\ln n)^{1/2}(nh^d)^{-1/2}\}, \tag{B.2}$$

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\|E\{\hat{D}(\boldsymbol{\phi};\boldsymbol{x})\}-D(\boldsymbol{\phi};\boldsymbol{x})\|=O_p(h^m). \tag{B.3}$$

Let $\delta_n = (\ln n)^{1/2}(nh^d)^{-1/2}$. Note that under the condition (C1.9), $O(\delta_n + h^m) = o(n^{-1/4})$. Thus, there is a gap between Theorem B.2.1 and Lemma B.2.1 that whether the convergence holds for each $\boldsymbol{\phi} \in \Phi$ or uniformly in Φ. The gap is filled as follows.

Proof of Theorem B.2.1. Denote a open ball with center $\boldsymbol{\phi}$ and radius e by $B(\boldsymbol{\phi}, e)$. It is enough to show both (B.2) and (B.3) hold uniformly in $\boldsymbol{\phi} \in B(\boldsymbol{\phi}, e)$ because of the continuity of $D(\boldsymbol{\phi}; \boldsymbol{x})$ with respect to $\boldsymbol{\phi}$ and compactness of the parameter space. Actually, It holds that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}} h^{-d}|\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-\boldsymbol{x}_i)O(\boldsymbol{x}, y_i; \boldsymbol{\phi}) - \inf_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-\boldsymbol{x}_i)O(\boldsymbol{x}, y_i; \boldsymbol{\phi})| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} 2h^{-d}|\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-\boldsymbol{x}_i)O(\boldsymbol{x}, y_i; \boldsymbol{\phi})| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} 2h^{-d}|RK_h(\boldsymbol{x}-\boldsymbol{x}_i)\tilde{b}_1(\boldsymbol{x}, y_i)|.
\end{aligned}$$

Here, the last equation is integrable by the condition (C1.6), (C1.7), and (C1.10) because

$$\sup_{\boldsymbol{x}\in\mathcal{X}} h^{-d}E\{RK_h(\boldsymbol{x}-\boldsymbol{x}_i)\tilde{b}_1(\boldsymbol{x}, y_i)\} = c\sup_{\boldsymbol{x}\in\mathcal{X}} E\{R\tilde{b}_1(\boldsymbol{Z}) \mid \boldsymbol{x}\} < \infty,$$

where $c = \sup_{\boldsymbol{x}} K(\boldsymbol{x})$. Therefore, for any $\varepsilon > 0$, we can choose e so that

$$h^{-d}\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|E\left\{\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-X)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\} - E\left\{\inf_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-X)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\}\right\| < \varepsilon$$

It follows from (B.2) that

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} \|\hat{D}(\boldsymbol{\phi}; \boldsymbol{x}) - E\{\hat{D}(\boldsymbol{\phi}; \boldsymbol{x})\}\| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}}\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)}\left\|\frac{1}{nh^d}\sum_{i=1}^n r_iK_h(\boldsymbol{x}-\boldsymbol{x}_i)O(\boldsymbol{x}, y_i; \boldsymbol{\phi}) - \frac{1}{h^d}E\left\{RK_h(\boldsymbol{x}-U)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\}\right\| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\frac{1}{nh^d}\sum_{i=1}^n \sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} r_iK_h(\boldsymbol{x}-\boldsymbol{x}_i)O(\boldsymbol{x}, y_i; \boldsymbol{\phi}) - \frac{1}{h^d}E\left\{\inf_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-U)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\}\right\| \\
&= O_p(\delta_n) \\
&\quad + h^{-d}\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|E\left\{\sup_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-X)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\} - E\left\{\inf_{\boldsymbol{\phi}\in B(\boldsymbol{\phi},e)} RK_h(\boldsymbol{x}-X)O(\boldsymbol{x}, Y; \boldsymbol{\phi})\right\}\right\| \\
&= O_p(\delta_n) + \varepsilon.
\end{aligned}$$

Therefore, (B.2) holds uniformly in the ball $B(\boldsymbol{\phi}, e)$. By using the compactness of Φ, we can easily show that this holds uniformly in Φ.

Let $w(\boldsymbol{x};\boldsymbol{\phi}) = E\{RO(\boldsymbol{x},Y;\boldsymbol{\phi}) \mid \boldsymbol{x}\}$. It follows from the Taylor's theorem and (C1.6) that, for some $0 < c < 1$,

$$\begin{aligned} E\{\hat{D}(\boldsymbol{\phi};\boldsymbol{x})\} - D(\boldsymbol{\phi};\boldsymbol{x}) &= \int K(\boldsymbol{u})\{w(\boldsymbol{x}+\boldsymbol{u}h;\boldsymbol{\phi}) - w(\boldsymbol{x};\boldsymbol{\phi})\}d\boldsymbol{u} \\ &\leq \frac{h^m}{m!}c^m \|w(\boldsymbol{x};\boldsymbol{\phi})\|_m \int \boldsymbol{u}^m K(\boldsymbol{u})d\boldsymbol{u}. \end{aligned}$$

By using the condition (C1.7),

$$\sup_{\boldsymbol{\phi}\in\Phi} \|w(\boldsymbol{\phi})\|_m < \infty$$

holds, and thus, (B.3) holds uniformly in Φ. Therefore (B.1) holds uniformly in Φ. Therefore, the desired conclusion is obtained.

Lemma B.2.2. Under conditions (C1.1)–(C1.5) and (C1.7)–(C1.10), we obtain the following asymptotic expansion of (2.12) and (2.14): for each $(\boldsymbol{\phi},\theta) \in \Phi \times \Theta$, it holds that

$$\hat{\mathbf{S}}(\boldsymbol{\phi}) = n^{-1}\sum_{i=1}^{n}\left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)}\right\}\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i) + n^{-1}\sum_{i=1}^{n} r_i \boldsymbol{G}_2(\boldsymbol{\phi};\boldsymbol{z}_i) + o_p(n^{-1/2}) \tag{B.4}$$

and

$$\begin{aligned} \hat{\mathrm{U}}(\boldsymbol{\phi},\theta) = n^{-1}\sum_{i=1}^{n}&\left[\frac{r_i U(\theta;\boldsymbol{z}_i)}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)} + \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)}\right\}\bar{U}(\theta;\boldsymbol{x}_i)\right] \\ &+ n^{-1}\sum_{i=1}^{n} r_i H_2(\boldsymbol{\phi};\boldsymbol{z}_i) + o_p(n^{-1/2}), \end{aligned} \tag{B.5}$$

where

$$\begin{aligned} \boldsymbol{G}_2(\boldsymbol{\phi};\boldsymbol{z}) &= \left\{\frac{E_1\{O(\boldsymbol{\phi}_0;\boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi};\boldsymbol{z}) \mid \boldsymbol{x}\}} - 1\right\} O(\boldsymbol{\phi};\boldsymbol{z})\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{z}) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x})\right\} \\ H_2(\boldsymbol{\phi};\boldsymbol{z}) &= \left\{\frac{E_1\{O(\boldsymbol{\phi}_0;\boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi};\boldsymbol{z}) \mid \boldsymbol{x}\}} - 1\right\} O(\boldsymbol{\phi};\boldsymbol{z})\left\{U(\theta;\boldsymbol{z}) - \bar{U}(\boldsymbol{\phi},\theta;\boldsymbol{x})\right\}. \end{aligned}$$

Proof of Lemma B.2.2. By using $\hat{a}/\hat{b} - a/b = b^{-1}\{1 - \hat{b}^{-1}(\hat{b}-b)\}\{(\hat{a}-a) - (a/b)(\hat{b}-b)\}$ and Lemma B.2.1, we obtain for each $\boldsymbol{\phi} \in \Phi$

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left\|\frac{\hat{\boldsymbol{C}}_s(\boldsymbol{\phi};\boldsymbol{x})}{\hat{D}(\boldsymbol{\phi};\boldsymbol{x})} - \frac{\boldsymbol{C}_s(\boldsymbol{\phi};\boldsymbol{x})}{D(\boldsymbol{\phi};\boldsymbol{x})} - E(\boldsymbol{\phi};\boldsymbol{x})\right\| = o_p(n^{-1/2}), \tag{B.6}$$

where

$$E(\boldsymbol{\phi};\boldsymbol{x}) := D(\boldsymbol{\phi};\boldsymbol{x})^{-1}\left\{(\hat{\boldsymbol{C}}_s(\boldsymbol{\phi};\boldsymbol{x}) - \boldsymbol{C}_s(\boldsymbol{\phi};\boldsymbol{x})) - \frac{\boldsymbol{C}_s(\boldsymbol{\phi};\boldsymbol{x})}{D(\boldsymbol{\phi};\boldsymbol{x})}(\hat{D}(\boldsymbol{\phi};\boldsymbol{x}) - D(\boldsymbol{\phi};\boldsymbol{x}))\right\}.$$

It follows that $\|\hat{s}_0(\boldsymbol{\phi};\boldsymbol{x}) - \bar{s}_0(\boldsymbol{\phi};\boldsymbol{x}) - E(\boldsymbol{\phi};\boldsymbol{x})\| = o_p(n^{-1/2})$.

Therefore, by using (B.6), we can show, for each $\boldsymbol{\phi} \in \Phi$,

$$\begin{aligned}
&\hat{\mathbf{S}}(\boldsymbol{\phi}) - n^{-1}\sum_{i=1}^{n}\{r_i \boldsymbol{s}_1(\boldsymbol{\phi};\boldsymbol{z}_i) + (1-r_i)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\} \\
&\quad - n^{-2}\sum_{i\neq j}\frac{(1-r_i)r_j h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)O(\boldsymbol{\phi};\boldsymbol{x}_i,y_j)\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x}_i,y_j) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\right\}}{D(\boldsymbol{\phi};\boldsymbol{x}_i)} \\
&\quad = o_p(n^{-1/2}).
\end{aligned} \tag{B.7}$$

Also, the double sum in (B.7) can be written as

$$\begin{aligned}
&n^{-2}\sum_{i\neq j}\frac{(1-r_i)r_j h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)O(\boldsymbol{\phi};\boldsymbol{x}_i,y_j)\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x}_i,y_j) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\right\}}{D(\boldsymbol{\phi};\boldsymbol{x}_i)} \\
&= \binom{n}{2}^{-1}\sum_{i<j} 2^{-1}\left[\boldsymbol{\zeta}_{ij} + \boldsymbol{\zeta}_{ji}\right] + o_p(n^{-1/2}),
\end{aligned} \tag{B.8}$$

where $\boldsymbol{\zeta}_{ij}$ can be expressed as

$$\boldsymbol{\zeta}_{ij} := \frac{(1-r_i)r_j h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)O(\boldsymbol{\phi};\boldsymbol{x}_i,y_j)\left\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x}_i,y_j) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\right\}}{D(\boldsymbol{\phi};\boldsymbol{x}_i)}.$$

Let $\boldsymbol{w} = (\boldsymbol{z}^\top, r)^\top$ and $\boldsymbol{h}(\boldsymbol{w}_i,\boldsymbol{w}_j) := (\boldsymbol{\zeta}_{ij} + \boldsymbol{\zeta}_{ji})/2$. According to U-statistic theory (e.g. van der Vaart, 1998, Ch.12), we have

$$\binom{n}{2}^{-1}\sum_{i<j}\boldsymbol{h}(\boldsymbol{w}_i,\boldsymbol{w}_j) = 2n^{-1}\sum_{i=1}^{n}E\{\boldsymbol{h}(\boldsymbol{w}_i,\boldsymbol{w}_j) \mid \boldsymbol{w}_i\} + o_p(n^{-1/2}). \tag{B.9}$$

Note that,

$$\begin{aligned}
&E(\boldsymbol{\zeta}_{ij} \mid \boldsymbol{w}_i) \\
&= \frac{1-r_i}{D(\boldsymbol{\phi};\boldsymbol{x}_i)}E[r_j h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)O(\boldsymbol{\phi};\boldsymbol{x}_i,y_j)\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x}_i,y_j) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\} \mid \boldsymbol{w}_i] \\
&= \frac{1-r_i}{D(\boldsymbol{\phi};\boldsymbol{x}_i)}E[h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)E[r_j O(\boldsymbol{\phi};\boldsymbol{x}_i,y_j)\{\boldsymbol{s}_0(\boldsymbol{\phi};\boldsymbol{x}_i,y_j) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)\} \mid \boldsymbol{x}_i] \mid \boldsymbol{w}_i] = 0
\end{aligned}$$

and

$$\begin{aligned}
E(\boldsymbol{\zeta}_{ji} \mid \boldsymbol{w}_i) &= r_i\int \boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{x}_j,y_i)h^{-d}K_h(\boldsymbol{x}_i - \boldsymbol{x}_j)\mathrm{d}x_j \\
&= r_i\boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{z}_i) + O(\boldsymbol{h}^m) \\
&= r_i\boldsymbol{G}_1(\boldsymbol{\phi};\boldsymbol{z}_i) + o(n^{-1/2}),
\end{aligned}$$

where

$$\boldsymbol{G}_1(\boldsymbol{\phi}; \boldsymbol{z}) = \frac{E_1\{O(\boldsymbol{\phi}_0; \boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi}; \boldsymbol{z}) \mid \boldsymbol{x}\}} O(\boldsymbol{\phi}; \boldsymbol{z}) \left\{\boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{z}) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x})\right\}.$$

Therefore, by using (B.7)-(B.9), the score function $\hat{\mathbf{S}}(\boldsymbol{\phi})$ can be written as

$$\begin{aligned}\hat{\mathbf{S}}(\boldsymbol{\phi}) &= n^{-1} \sum_{i=1}^{n} \left[r_i \left\{\boldsymbol{s}_1(\boldsymbol{\phi}; \boldsymbol{z}_i) + \boldsymbol{G}_1(\boldsymbol{\phi}; \boldsymbol{z}_i)\right\} + (1 - r_i)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x}_i)\right] + o_p(n^{-1/2}) \\ &=: n^{-1} \sum_{i=1}^{n} \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x}_i) + n^{-1} \sum_{i=1}^{n} r_i \boldsymbol{G}_2(\boldsymbol{\phi}; \boldsymbol{z}_i) + o_p(n^{-1/2}),\end{aligned}$$

where

$$\boldsymbol{G}_2(\boldsymbol{\phi}; \boldsymbol{z}) = \left\{\frac{E_1\{O(\boldsymbol{\phi}_0; \boldsymbol{z}) \mid \boldsymbol{x}\}}{E_1\{O(\boldsymbol{\phi}; \boldsymbol{z}) \mid \boldsymbol{x}\}} - 1\right\} O(\boldsymbol{\phi}; \boldsymbol{z}) \left\{\boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{z}) - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x})\right\}.$$

Therefore, (B.4) is obtained. The expansion of (B.5) is also obtained in a similar way. □

Consistency and asymptotic normality of our proposed estimator can be obtained by using Lemma B.2.2 in conjunction with theory of an *Asymptotically Unbiased Estimating Equation* proposed in Zhao and Lin (2012).

Proof of Theorem 2.3.1. We focus on proving (B.4) at first. Let

$$\hat{\mathbf{S}}_1(\boldsymbol{\phi}) := n^{-1} \sum_{i=1}^{n} \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x}_i) + n^{-1} \sum_{i=1}^{n} r_i \boldsymbol{G}_2(\boldsymbol{\phi}; \boldsymbol{z}_i).$$

To obtain consistency of our proposed estimator, it is only needed to show

$$\sup_{\boldsymbol{\phi} \in \Phi} \|\hat{\mathbf{S}}(\boldsymbol{\phi}) - \hat{\mathbf{S}}_1(\boldsymbol{\phi})\| = o_p(1)$$

and

$$\sup_{\boldsymbol{\phi} \in \Phi} \left\| \frac{\partial \hat{\mathbf{S}}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} - \frac{\partial \hat{\mathbf{S}}_1(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \right\| = o_p(1) \tag{B.10}$$

because the other conditions required by Theorem 5 and 6 in Zhao and Lin (2012) are met by the regularity conditions. These are easily shown by using Theorem B.2.1. In fact,

$$\begin{aligned}&\|\hat{\mathbf{S}}(\boldsymbol{\phi}) - \hat{\mathbf{S}}_1(\boldsymbol{\phi})\| \\ &= \left\| \frac{1}{n} \sum_{i=1}^{n} (1 - r_i) \left\{ \frac{\hat{\boldsymbol{C}}_s(\boldsymbol{\phi}; \boldsymbol{x}_i)}{\hat{D}(\boldsymbol{\phi}; \boldsymbol{x}_i)} - \bar{\boldsymbol{s}}_0(\boldsymbol{\phi}; \boldsymbol{x}_i) \right\} \right\| + \left\| \frac{1}{n} \sum_{i=1}^{n} r_i \boldsymbol{G}_1(\boldsymbol{\phi}; \boldsymbol{z}_i) \right\|.\end{aligned}$$

It follows from Theorem B.2.1 that the first term converges to zero uniformly by (C1.12). The second term also converges to zero uniformly by using the uniform law of large numbers with (C1.11). The other equation (B.10) can be shown in a similar way. □

Next, we simplify the influence function of $\hat{\boldsymbol{\phi}}$. So far, it has been shown that the influence function of $\hat{\mathbf{S}}(\boldsymbol{\phi})$ is given by

$$\left\{1-\frac{R}{\pi(\boldsymbol{\phi};\boldsymbol{Z})}\right\}\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{X})+\boldsymbol{G}_2(\boldsymbol{\phi};\boldsymbol{Z}). \tag{B.11}$$

Thus the asymptotic variance of $\hat{\boldsymbol{\phi}}$ can be derived by using the usual sandwich estimator. Here it follows from the facts $\boldsymbol{G}_2(\boldsymbol{\phi}_0)=\mathbf{0}$ and $E\{\partial\boldsymbol{G}_2(\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi}^\top\}=\mathbf{0}$ that the influence function (B.11) can be also written as

$$\left\{1-\frac{R}{\pi(\boldsymbol{\phi};\boldsymbol{Z})}\right\}\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{X}).$$

Using the same argument for $\hat{\mathrm{U}}(\boldsymbol{\phi},\theta)$, we can obtain the influence function for θ by

$$\frac{R}{\pi(\boldsymbol{\phi};\boldsymbol{Z})}U(\theta;\boldsymbol{X})+\left\{1-\frac{R}{\pi(\boldsymbol{\phi};\boldsymbol{Z})}\right\}\bar{U}(\theta;\boldsymbol{X}).$$

Denote the simplified influence functions by

$$\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}):=\left\{1-\frac{r_i}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)}\right\}\bar{\boldsymbol{s}}_0(\boldsymbol{\phi};\boldsymbol{x}_i)$$

and

$$\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}):=\frac{r_i}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)}U(\theta;\boldsymbol{x}_i)+\left\{1-\frac{r_i}{\pi(\boldsymbol{\phi};\boldsymbol{z}_i)}\right\}\bar{U}(\boldsymbol{\phi};\boldsymbol{x}_i).$$

The solution based on the estimating equations (B.4) and (B.5) is asymptotically equivalent to that to the following estimating equations:

$$\sum_{i=1}^n\hat{\mathbf{S}}_i^l(\boldsymbol{\phi})=\mathbf{0},$$
$$\sum_{i=1}^n\hat{\mathrm{U}}_i^l(\boldsymbol{\phi},\theta)=0.$$

Hereafter, we derive the asymptotic variance of the estimator based on the above estimating equations.

Lemma B.2.3. The following partial derivatives with respect to θ and $\boldsymbol{\phi}$ satisfy

$$\begin{aligned}E\left\{\partial\hat{\mathbf{S}}_n^l(\boldsymbol{\phi})/\partial\boldsymbol{\phi}^\top\right\}&=E\left\{\bar{\mathbf{s}}_0(\boldsymbol{\phi})\dot{\boldsymbol{\pi}}^\top(\boldsymbol{\phi})/\pi(\boldsymbol{\phi})\right\},\\ E\left\{\partial\hat{\mathrm{U}}_n^l(\boldsymbol{\phi},\theta)/\partial\boldsymbol{\phi}^\top\right\}&=E\left[\left\{\bar{U}(\theta)-U(\theta)\right\}\dot{\boldsymbol{\pi}}^\top(\boldsymbol{\phi})/\pi(\boldsymbol{\phi})\right],\\ E\left\{\partial\hat{\mathrm{U}}_n^l(\boldsymbol{\phi},\theta)/\partial\theta\right\}&=E\left\{\partial U(\theta)/\partial\theta\right\}.\end{aligned}$$

Finally, we give a proof of Theorem 2.3.2.

Proof of Theorem 2.3.2. By using standard arguments,

$$n^{1/2}\left[\begin{array}{c}\hat{\boldsymbol{\phi}}-\phi_0\\ \hat{\theta}-\theta_0\end{array}\right]=-\left(n^{-1}\sum_{i=1}^{n}\left[\begin{array}{cc}\partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\theta\\ \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\theta\end{array}\right]\right)^{-1}$$

$$\times\, n^{-1/2}\sum_{i=1}^{n}\left[\begin{array}{c}\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)\\ \hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)\end{array}\right]+o_p(1).$$

Let

$$\mathcal{J}:=-E\left[\begin{array}{cc}\partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\theta\\ \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\theta\end{array}\right]=\left[\begin{array}{cc}\mathcal{J}_{11} & O\\ \mathcal{J}_{21} & \mathcal{J}_{22}\end{array}\right].$$

By conditions (C1.1)–(C1.5), and Lemma 2.4 in Newey and McFadden (1994),

$$\left(-n^{-1}\sum_{i=1}^{n}\left[\begin{array}{cc}\partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathbf{S}}_i^l(\boldsymbol{\phi}_0)/\partial\theta\\ \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\boldsymbol{\phi}^\top & \partial\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)/\partial\theta\end{array}\right]\right)^{-1}$$

converges to $\mathcal{J}^{-1}$ in probability. On the other hand, by the Central Limit Theorem, $n^{-1/2}[\{\sum_{i=1}^n\hat{\mathbf{S}}_i^l(\phi_0)\}^\top,\ \{\sum_{i=1}^n\hat{\mathrm{U}}_i^l(\boldsymbol{\phi}_0,\theta_0)\}^\top]^\top$ has an asymptotic normal distribution with mean 0 and variance $\mathrm{var}[\{\hat{\mathbf{S}}^l(\phi_0)\}^\top,\ \{\hat{\mathrm{U}}^l(\phi_0,\theta_0)\}^\top]^\top$. Then, by Slutzky's Theorem, $n^{1/2}[(\hat{\boldsymbol{\phi}}-\phi_0)^\top,\ (\hat{\theta}-\theta_0)^\top]^\top$ has an asymptotic normal distribution with mean 0 and variance

$$\mathcal{J}^{-1}\mathrm{var}\left\{\begin{array}{c}\hat{\mathbf{S}}^l(\phi_0)\\ \hat{\mathrm{U}}^l(\phi_0,\theta_0)\end{array}\right\}\left(\mathcal{J}^{-1}\right)^\top=\mathrm{var}\left\{\left[\begin{array}{cc}\mathcal{J}_{11}^{-1} & O\\ -\mathcal{J}_{22}^{-1}\mathcal{J}_{21}\mathcal{J}_{11}^{-1} & \mathcal{J}_{22}^{-1}\end{array}\right]\left[\begin{array}{c}\hat{\mathbf{S}}^l(\phi_0)\\ \hat{\mathrm{U}}^l(\phi_0,\theta_0)\end{array}\right]\right\}$$

$$=\mathrm{var}\left\{\begin{array}{c}\mathcal{J}_{11}^{-1}\hat{\mathbf{S}}^l(\phi_0)\\ \mathcal{J}_{22}^{-1}\left\{\hat{\mathrm{U}}^l(\phi_0,\theta_0)-\mathcal{J}_{21}\mathcal{J}_{11}^{-1}\hat{\mathbf{S}}^l(\phi_0)\right\}\end{array}\right\}. \quad (\mathrm{B.12})$$

Finally, explicit expressions of the variance of $\hat{\mathbf{S}}^l(\phi_0)$ and $\hat{\mathrm{U}}^l(\phi_0,\theta_0)-\mathcal{J}_{21}\mathcal{J}_{11}^{-1}\hat{\mathbf{S}}^l(\phi_0)$ are given as follows:

$$\mathrm{var}\{\hat{\mathbf{S}}^l(\phi_0)\}=E\left[\mathrm{var}\left\{r\pi^{-1}(\boldsymbol{\phi}_0)\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)\mid\boldsymbol{z}\right\}\right]=E\{O(\boldsymbol{\phi}_0)\bar{\boldsymbol{s}}_0^{\otimes 2}\}$$

and

$$\begin{aligned}&\mathrm{var}\{\hat{\mathrm{U}}^l(\boldsymbol{\phi}_0,\theta_0)-\mathcal{J}_{21}\mathcal{J}_{11}^{-1}\hat{\mathbf{S}}^l(\phi_0)\}\\&=\mathrm{var}\{U(\theta_0)\}+E\left[\mathrm{var}\left\{r\pi^{-1}(\boldsymbol{\phi}_0)\{U(\theta_0)-\bar{U}(\theta_0)+\kappa\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)\}\mid\boldsymbol{z}\right\}\right]\\&=\mathrm{var}\{U(\theta_0)\}+E[O(\boldsymbol{\phi}_0)\{U(\theta_0)-\bar{U}(\theta_0)+\kappa\bar{\boldsymbol{s}}_0(\boldsymbol{\phi}_0)\}^{\otimes 2}],\end{aligned}$$

where $\kappa = \mathcal{I}_{21}\mathcal{I}_{11}^{-1}$. □

Next, we provide a proof of Lemma 2.4.1 and Theorem 2.4.1 and 2.4.2. In order to prove Lemma 2.4.1, we will assume $\theta = E(Y)$ for simplicity. Wespecify the joint distribution $\boldsymbol{z} = (\boldsymbol{x}^\top, y)^\top$ by $f(\boldsymbol{z}; \boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is an infinite-dimensional nuisance parameter, and $\boldsymbol{\eta}_0$ is the true value. By "full model" we refer to the class of models in which the data are completely observed, and by "obs model" those in which some Y are missing; that is, a full model consists of functions $\boldsymbol{h}(\boldsymbol{Z})$ and an obs model consists of $\boldsymbol{h}(R, \boldsymbol{G}_R(\boldsymbol{Z}))$. Furthermore, for each full and obs model, denote the nuisance tangent space by Λ^F and Λ, respectively, and its orthogonal complement by $\Lambda^{F\perp}$ and $\Lambda^\perp$, respectively. Let $\boldsymbol{S}_\phi$ be the score function with respect to $\boldsymbol{\phi}$. Consider a Hilbert space $\mathcal{H} = \{\boldsymbol{h}^{(q+1)\times 1} \mid E(\boldsymbol{h}) = 0; \|\boldsymbol{h}\| < \infty\}$ with inner product $\langle \boldsymbol{h}_1, \boldsymbol{h}_2 \rangle = E(\boldsymbol{h}_1^\top \boldsymbol{h}_2)$, where the expectation is taken under the true model. See Bickel et al. (1998) and Tsiatis (2006) for more details.

At first, we introduce a proposition of Rotnitzky and Robins (1997), which provides the efficient score for $(\boldsymbol{\phi}, \theta)$, as follows. Let B and D be functions of $(R, \boldsymbol{G}_R(\boldsymbol{Z}))$, and let B^* and D^* be functions of $\boldsymbol{Z}$. Also, let us define the following three linear operators: $\mathrm{g}(B^*) = E(B^* \mid R, \boldsymbol{G}_R(\boldsymbol{Z}))$, $\mathrm{m}(B^*) = E\{\mathrm{g}(B^*) \mid \boldsymbol{Z}\}$, and $\mathrm{u}(B^*) = RB^*/\pi(\boldsymbol{Z})$. Then, the efficient score for $(\boldsymbol{\phi}, \theta)$ can be derived by the following Lemma. See Proposition A1 in Rotnitzky and Robins (1997) for the proof.

Lemma B.2.4. The efficient score for $(\boldsymbol{\phi}, \theta)$ can be written as

$$\boldsymbol{S}_{\text{eff}} = \mathrm{u}(D^*_{\text{eff}}) - \Pi[\mathrm{u}(D^*_{\text{eff}}) \mid \Lambda_2] + A_{2,\text{eff}} = \mathrm{g}\{\mathrm{m}^{-1}(D^*_{\text{eff}})\} + A_{2,\text{eff}}, \tag{B.12}$$

where $\Pi[h \mid \Lambda_2]$ is the projection of h onto Λ_2, $\Lambda_2 = [\boldsymbol{h}(R, \boldsymbol{G}_R(\boldsymbol{Z})) : E(\boldsymbol{h}(R, \boldsymbol{G}_R(\boldsymbol{Z})) \mid \boldsymbol{Z}) = 0]$, and D^*_{eff} is a unique solution to

$$\Pi[\mathrm{m}^{-1}(D^*) \mid \Lambda^{F\perp}] = (Q, S^{F\perp}_{\text{eff},\theta}), \tag{B.13}$$

where $Q = \Pi[\mathrm{m}^{-1}[E\{\mathrm{g}(S^F_\phi) \mid L\}] \mid \Lambda^{F\perp}]$, $A_{2,\text{eff}} = (\Pi[\boldsymbol{S}_\phi \mid \Lambda_2]^\top, 0)^\top = (\mathrm{g}(S^F_\phi) - \mathrm{g}[\mathrm{m}^{-1}[E\{\mathrm{g}(S^F_\phi) \mid L\}]]^\top, 0)^\top$, and $S^{F\perp}_{\text{eff},\theta}$ is the efficient score function of θ in the full model.

This Lemma implies that the efficient score can be represented by (B.12) with D^*_{eff} satisfying condition (B.13). Thus, in the nonignorable nonresponse case, $\Lambda^{F\perp}$ needs to be calculated, and it can be done in a way similar to that shown in Section 4.5 of Tsiatis (2006).

Lemma B.2.5. The nuisance tangent space Λ^F and its orthogonal complement $\Lambda^{F\perp}$ in the full model are written as follows:

$$\begin{aligned}\Lambda^F &= [\boldsymbol{h}(\boldsymbol{Z}) \in \mathcal{H} \text{ such that } E\{Y\boldsymbol{h}(\boldsymbol{Z})\} = 0],\\ \Lambda^{F\perp} &= [\boldsymbol{k}(Y-\theta_0), \text{where } \boldsymbol{k} \text{ is any } q+1 \text{ dimensional vector}]\,.\end{aligned}$$

Finally, we give an explicit formula to calculate the projection onto Λ_2.

Lemma B.2.6. For $\boldsymbol{h}(R, \boldsymbol{G}_R(\boldsymbol{Z})) = R\boldsymbol{h}_1(\boldsymbol{Z}) + (1-R)\boldsymbol{h}_2(\boldsymbol{X})$, it holds that

$$\Pi(\boldsymbol{h} \mid \Lambda_2) = \left\{1 - \frac{R}{\pi(\boldsymbol{Z})}\right\} \frac{\boldsymbol{h}_2(\boldsymbol{X}) - E(\boldsymbol{h}_1(\boldsymbol{Z}) \mid \boldsymbol{X})}{E\{O(\boldsymbol{Z}) \mid \boldsymbol{X}\}}. \tag{B.14}$$

Proof of Lemma B.2.6. Obviously, the right-hand side of (B.14) belongs to Λ_2. Thus, it remains to check that for any g,

$$\left\langle \boldsymbol{h} - \left\{1 - \frac{R}{\pi(\boldsymbol{Z})}\right\} \frac{\boldsymbol{h}_2(\boldsymbol{X}) - E\{\boldsymbol{h}_1(\boldsymbol{Z}) \mid \boldsymbol{X}\}}{E\{O(\boldsymbol{Z}) \mid \boldsymbol{X}\}}, \left\{1 - \frac{R}{\pi(\boldsymbol{Z})}\right\} g(\boldsymbol{X}) \right\rangle = 0,$$

which can be proved easily. □

We now give a proof of Lemma 2.4.1.

Proof of Lemma 2.4.1. Note that $S^{F\perp}_{\text{eff},\theta} = Y - \theta_0$ by Lemma B.2.5, since there exists only one influence function, and it is the efficient one under the assumption that θ does not require any assumptions on the distribution of $\boldsymbol{Z}$ (see Tsiatis, 2006, Chap. 5). By the projection theorem, there exists a unique $\boldsymbol{k} = (\boldsymbol{k}_1, k_2^\top)^\top$ such that $D^*_{\text{eff}} = \boldsymbol{k}(Y - \theta_0)$.

Then, we calculate $A_{2,\text{eff}}$. The score function of ϕ is

$$\boldsymbol{S}_\phi = \mathrm{g}(S^F_\phi) = R\boldsymbol{s}_1(\boldsymbol{Z}; \boldsymbol{\phi}) + (1-R)\bar{\boldsymbol{s}}_0(\boldsymbol{X}; \boldsymbol{\phi}),$$

where $\boldsymbol{s}_r(\boldsymbol{\phi})$ is defined in (2.2). It follows from Lemma B.2.6 with $\boldsymbol{h}_1(z) = \boldsymbol{s}_1(\boldsymbol{\phi})$ and $\boldsymbol{h}_2(\boldsymbol{x}) = \bar{\boldsymbol{s}}_0(\boldsymbol{x}; \boldsymbol{\phi})$ in (B.14) that $\Pi(\boldsymbol{S}_\phi \mid \Lambda_2) = -\{1 - R/\pi(\boldsymbol{Z})\}g^\star(\boldsymbol{X})$. Thus, $A_{2,\text{eff}} = [0, -\{1-R/\pi(\boldsymbol{Z})\}g^\star(\boldsymbol{X})]$. Again, by using Lemma S3, it follows that $\Pi[\mathrm{u}(D^*_{\text{eff}}) \mid \Lambda_2] = -\{1 - R/\pi(\boldsymbol{Z})\}E^\star(Y - \theta_0 \mid \boldsymbol{X})$, by which (B.12) becomes

$$\boldsymbol{S}_1 = k_2 \left[\frac{R(Y-\theta_0)}{\pi(\boldsymbol{\phi}_0)} + \left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} E^\star(Y - \theta_0 \mid \boldsymbol{X})\right] - \left\{1 - \frac{R}{\pi(\boldsymbol{Z})}\right\} g^\star(\boldsymbol{X})$$

and

$$\boldsymbol{S}_2 = \boldsymbol{k}_1 \left[\frac{R(Y-\theta_0)}{\pi(\boldsymbol{\phi}_0)} + \left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} E^\star(Y - \theta_0 \mid \boldsymbol{X})\right].$$

This $\boldsymbol{S}_{\text{eff}} = (\boldsymbol{S}_1, S_2^\top)$ can be transformed into $\tilde{S}_{\text{eff}} = (\tilde{S}_1, \tilde{S}_2^\top) = A\boldsymbol{S}_{\text{eff}}$,

$$\tilde{S}_1 = \left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} g^\star(\boldsymbol{X}),$$
$$\tilde{S}_2 = \frac{R(Y-\theta_0)}{\pi(\boldsymbol{\phi}_0)} + \left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} E^\star(Y - \theta_0 \mid \boldsymbol{X})$$

with a nonsingular matrix A,

$$A = \begin{bmatrix} -I_q & -k_2^\top/\boldsymbol{k}_1 \\ 0^\top & \boldsymbol{k}_1^{-1} \end{bmatrix},$$

where I_q is a q-dimensional identity matrix. The score function multiplied by a nonsingular constant matrix does not have an influence on the asymptotic distribution. Thus, we have the desired efficient score. □

Proof of Theorem 2.4.1. Consistency and asymptotic normality are proved under the assumptions (C2.1)-(C2.7) in Appendix A by using the standard argument for GMM. Next, we give the explicit form of the asymptotic variance. Let $\xi = (\boldsymbol{\phi}^\top, \theta)^\top$. Recall that each $\hat{\gamma}$ and $\hat{\xi}$ is a solution to $\sum_{i=1}^n \partial \log f_1(y_i \mid \boldsymbol{x}_i; \gamma)/\partial\boldsymbol{\gamma} = \sum_{i=1}^n \boldsymbol{S}_{\boldsymbol{\gamma},i}(\gamma) = \boldsymbol{0}$ and $\sum_{i=1}^n \boldsymbol{S}_{\text{eff},i}(\hat{\gamma}, \xi) = \boldsymbol{0}$, respectively. By using standard asymptotic theory,

$$\begin{bmatrix} \hat{\gamma} - \boldsymbol{\gamma}^* \\ \hat{\xi} - \xi_0 \end{bmatrix} = -\mathcal{I}^{-1} n^{-1} \sum_{i=1}^n \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\gamma},i}(\boldsymbol{\gamma}^*) \\ \boldsymbol{S}_{\text{eff},i}(\boldsymbol{\gamma}^*, \xi_0) \end{bmatrix},$$

where

$$\mathcal{I} = E\begin{bmatrix} \partial\boldsymbol{S}_\gamma(\boldsymbol{\gamma}^*)/\gamma^\top & \partial\boldsymbol{S}_\gamma(\boldsymbol{\gamma}^*)/\xi^\top \\ \partial\boldsymbol{S}_{\text{eff}}(\boldsymbol{\gamma}^*, \xi_0)/\gamma^\top & \partial\boldsymbol{S}_{\text{eff}}(\boldsymbol{\gamma}^*, \xi_0)/\xi^\top \end{bmatrix} = E\begin{bmatrix} \partial\boldsymbol{S}_\gamma(\boldsymbol{\gamma}^*)/\gamma^\top & O \\ \partial\boldsymbol{S}_{\text{eff}}(\boldsymbol{\gamma}^*, \xi_0)/\gamma^\top & \partial\boldsymbol{S}_{\text{eff}}(\boldsymbol{\gamma}^*, \xi_0)/\xi^\top \end{bmatrix}.$$

Let the (i,j) block of $\mathcal{I}$ be $\mathcal{I}_{ij}$. Then,

$$\mathcal{I}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & O \\ -\mathcal{I}_{22}^{-1}\mathcal{I}_{21}\mathcal{I}_2^{-1} & \mathcal{I}_{22}^{-1} \end{bmatrix}.$$

Here, it follows that $\mathcal{I}_{21} = O$ because

$$E\left[\left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} \frac{\partial g^\star(\boldsymbol{\gamma}^*, \xi_0)}{\partial\boldsymbol{\gamma}^\top}\right] = O$$

and

$$E\left[\left\{1 - \frac{R}{\pi(\boldsymbol{\phi}_0)}\right\} \frac{\partial U^\star(\boldsymbol{\gamma}^*, \xi_0)}{\partial\boldsymbol{\gamma}^\top}\right] = 0^\top.$$

Therefore, we have,

$$\mathcal{I}^{-1} = \begin{bmatrix} \mathcal{I}_{11}^{-1} & O \\ O & \mathcal{I}_{22}^{-1} \end{bmatrix}.$$

By applying exactly the same arguments for $\mathcal{I}_{22}^{-1}$ used for $\mathcal{I}^{-1}$, we got the asymptotic variance of $\hat{\theta}$. □

Proof of Theorem 2.4.2. Consistency and asymptotic normality of my proposed estimator are similar to proving Lemma 2.3.1. We herein show our estimator attains the semiparametric lower bound derived in Lemma 2.3.1. Let $f_1(\boldsymbol{x})$ be the conditional distribution of $[\boldsymbol{x} \mid r = 1]$. From the same arguments that were used to prove Lemma 2.3.1, it can be shown that the estimating equation in Theorem 2.4.2, $\hat{S}_{\text{eff}}(\boldsymbol{\phi}, \theta) = \{\hat{S}_1(\boldsymbol{\phi})^\top, \hat{S}_2(\boldsymbol{\phi}, \theta)\}^\top$ is expanded as

$$
\begin{aligned}
\hat{S}_1(\boldsymbol{\phi}) &= n^{-1}\sum_{i=1}^{n}\left[\left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} g^\star(\boldsymbol{\phi}; \boldsymbol{x}_i) + r_i \boldsymbol{I}(\boldsymbol{z}_i; \boldsymbol{\phi})\right] + o_p(n^{-1/2}) \\
\hat{S}_2(\boldsymbol{\phi}, \theta) &= n^{-1}\sum_{i=1}^{n}\left[\frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)} U(\theta; \boldsymbol{z}_i) + \left\{1 - \frac{r_i}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}\right\} U^\star(\theta, \boldsymbol{\phi}; \boldsymbol{x}_i) + r_i J(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i)\right] \\
&\quad + o_p(n^{-1/2}),
\end{aligned}
$$

where $\boldsymbol{I}(\boldsymbol{\phi}; \boldsymbol{z}_i) = I_1(\boldsymbol{\phi}; \boldsymbol{x}_i)\boldsymbol{I}_2(\boldsymbol{\phi}; \boldsymbol{z}_i)$, $J(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i) = I_1(\boldsymbol{\phi}; \boldsymbol{x}_i)J_2(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i)$, and

$$
\begin{aligned}
I_1(\boldsymbol{\phi}; \boldsymbol{x}_i) &= 1 - E\left\{\frac{\pi(\boldsymbol{\phi}_0; \boldsymbol{Z})}{\pi(\boldsymbol{\phi}; \boldsymbol{Z})} \,\middle|\, \boldsymbol{x}_i\right\}, \\
\boldsymbol{I}_2(\boldsymbol{\phi}; \boldsymbol{z}_i) &= \frac{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{z}_i)O(\boldsymbol{\phi}; \boldsymbol{z}_i)\{\boldsymbol{s}_0(\boldsymbol{\phi}; \boldsymbol{z}_i) - g^\star(\boldsymbol{\phi}; \boldsymbol{x}_i)\}}{E_1\{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{Z})O(\boldsymbol{\phi}; \boldsymbol{Z}) \mid \boldsymbol{x}_i\}\Pr(R = 1 \mid \boldsymbol{x}_i)}, \\
J_2(\theta, \boldsymbol{\phi}; \boldsymbol{z}_i) &= \frac{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{z}_i)O(\boldsymbol{\phi}; \boldsymbol{z}_i)\{U(\theta; \boldsymbol{z}_i) - U^\star(\theta, \boldsymbol{\phi}; \boldsymbol{x}_i)\}}{E_1\{\pi^{-1}(\boldsymbol{\phi}; \boldsymbol{Z})O(\boldsymbol{\phi}; \boldsymbol{Z}) \mid \boldsymbol{x}_i\}\Pr(R = 1 \mid \boldsymbol{x}_i)}.
\end{aligned}
$$

Therefore, the asymptotic variance may increase due to the additional terms $r\boldsymbol{G}(\boldsymbol{\phi})$ and $rH(\boldsymbol{\phi})$, but this solution also attains the lower bound. At first, we focus on the estimator for ϕ. Once we get an unbiased estimating equation $\sum_{i=1}^{n}\varphi(\boldsymbol{z}_i; \boldsymbol{\phi}) = \boldsymbol{0}$, the asymptotic variance can be given as $\mathrm{Var}\{E(\dot{\varphi}(\boldsymbol{\phi}_0))^{-1}\varphi(\boldsymbol{\phi}_0)\}$, where $\dot{\varphi}(\boldsymbol{\phi}_0) = \partial\varphi(\boldsymbol{\phi}_0)/\partial\boldsymbol{\phi}^\top$. Thus, for the proving purpose, it suffices to show that $\boldsymbol{I}(\boldsymbol{\phi}_0) = \boldsymbol{0}$ and $E(R\dot{\boldsymbol{I}}(\boldsymbol{\phi}_0)) = O$. The former equation is trivial, so we only need to work on the latter equation, which can be written as $E(R\dot{\boldsymbol{I}}(\boldsymbol{\phi}_0)) = E(RI_1(\boldsymbol{\phi}_0)\dot{\boldsymbol{I}}_2(\boldsymbol{\phi}_0)) + E(R\boldsymbol{I}_2(\boldsymbol{\phi}_0)\dot{I}_1(\boldsymbol{\phi}_0))$. The first term is zero from $I_1(\boldsymbol{\phi}_0) = \boldsymbol{0}$. Also, the second term is $E(R\boldsymbol{I}_2(\boldsymbol{\phi}_0)\dot{I}_1(\boldsymbol{\phi}_0)) = E\{E(R\boldsymbol{I}_2(\boldsymbol{\phi}_0) \mid \boldsymbol{X})\dot{\boldsymbol{I}}_1(\boldsymbol{\phi}_0)\} = O$. Hence, the last equation holds by the definition of $g^\star(\boldsymbol{\phi}; \boldsymbol{x})$. Therefore, $r\boldsymbol{I}(\boldsymbol{\phi})$ has no effect on the asymptotic variance and my estimator also attains the semiparametric efficiency bound. The same conclusion can be made when estimating θ.

Proof of Theorem 2.5.1. According to equations (7)–(10) in Qin et al. (2002), the estimator is a solution to the following system of equations:

$$\sum_{i=1}^{n} r_i \frac{\boldsymbol{h}(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_n}{d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi})} = \mathbf{0}, \tag{B.15}$$

$$\sum_{i=1}^{n} r_i \frac{\pi(\boldsymbol{z}_i) - W}{d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi})} = 0, \tag{B.16}$$

$$\sum_{i=1}^{n} r_i \frac{\partial \log \pi(\boldsymbol{\phi}; \boldsymbol{z}_i)}{\partial \boldsymbol{\phi}} - \lambda_2 \sum_{i=1}^{n} r_i \frac{\partial \pi(\boldsymbol{\phi}; \boldsymbol{z}_i)/\partial \boldsymbol{\phi}}{d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi})} = \mathbf{0}, \tag{B.17}$$

$$\lambda_2 = \frac{n/m - 1}{1 - W}, \tag{B.18}$$

where $d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi}) = m[1 + \boldsymbol{\lambda}_1^{\top}\{\boldsymbol{h}(\boldsymbol{x}_i) - \bar{\boldsymbol{h}}_n\} + \lambda_2\{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i) - W\}]$ and $m = \sum_{i=1}^{n} r_i$. Note that the parameters to be estimated are $\boldsymbol{\lambda}_1, W$, and $\boldsymbol{\phi}$. It follows from (B.17) that

$$\sum_{i=1}^{n} r_i \frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}; \boldsymbol{z}_i)}{\pi(\boldsymbol{\phi}; \boldsymbol{z}_i)} = \lambda_2 \sum_{i=1}^{n} r_i \frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}; \boldsymbol{z}_i)}{d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi})}. \tag{B.19}$$

Thus, if we set $\boldsymbol{\lambda}_1 = \mathbf{0}$, $W = m/n$, $\lambda_2 = n/m$, and $d_i(\boldsymbol{\lambda}_1, W, \boldsymbol{\phi}) = \lambda_2 \pi(\boldsymbol{z}_i)$, (B.17) (or (B.19)) and (B.18) hold. By substituting the relevant terms in (B.15) and (B.16), we have (2.9). Because the probability that the estimator of Qin et al. (2002) has a unique solution goes to one almost sure as the sample size goes to infinity, this is the unique solution. □

Proof of Theorem 2.5.2. The proof can be made in a similar way to the proof of Corollary 3 of Qin and Lawless (1994). Define a function $\boldsymbol{K} : \mathbb{R}^1 \times \mathbb{R}^{d+1} \to \mathbb{R}^{p_2}$ through a function $\boldsymbol{k} : \mathbb{R}^d \to \mathbb{R}^{\kappa}$ by

$$\boldsymbol{K}(r, \boldsymbol{z}) = \left\{ \frac{r}{\pi(\boldsymbol{\phi}; \boldsymbol{z})} - 1 \right\} \boldsymbol{k}(\boldsymbol{x}).$$

Similarly, define $\boldsymbol{K}^{\star}(\cdot, \cdot)$ through $\boldsymbol{g}^{\star}(\boldsymbol{x})$. Recall that my estimating estimation is given as (2.10) with $\boldsymbol{g}(\boldsymbol{\phi}; \boldsymbol{x}) = \{\boldsymbol{g}^{\star}(\boldsymbol{\phi}; \boldsymbol{x})^{\top}, \boldsymbol{k}(\boldsymbol{x})^{\top}\}^{\top}$. Here we assume a working model $f_1(y \mid \boldsymbol{x}; \boldsymbol{\beta})$ is correctly specified such that $\boldsymbol{g}^{\star}(\boldsymbol{x})$ can be correctly computed. Wecompute the asymptotic variance of $\hat{\boldsymbol{\phi}}$ as if $\boldsymbol{\beta}_0$ were known, because the nuisance tangent space of $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$ are orthogonal, which implies the estimator $\hat{\boldsymbol{\beta}}$ does not affect the efficiency of $\hat{\boldsymbol{\phi}}$, as long as the working model is correctly specified. By using the standard argument of the GMM estimator, inverse of the asymptotic variance of the GMM estimator can be obtained by

$$
\begin{aligned}
&\left\{E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}^\top}\right),\ E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}^\top}\right)\right\}\left\{\begin{array}{cc} E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top}) & E(\boldsymbol{K}^\star\boldsymbol{K}^\top)\\ E(\boldsymbol{K}\boldsymbol{K}^{\star\top}) & E(\boldsymbol{K}\boldsymbol{K}^\top)\end{array}\right\}^{-1}\left\{\begin{array}{c} E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}}\right)\\ E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}}\right)\end{array}\right\}\\
&=\left\{E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}^\top}\right), -E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}^\top}\right)E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top})^{-1}E(\boldsymbol{K}^\star\boldsymbol{K}^\top)+E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}^\top}\right)\right\}\\
&\quad\times\left\{\begin{array}{cc} E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top})^{-1} & O_{q\times\kappa}\\ O_{\kappa\times q} & A_{22.1}^{-1}\end{array}\right\}^{-1}\left\{\begin{array}{c} E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}}\right)\\ -E(\boldsymbol{K}\boldsymbol{K}^{\star\top})E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top})^{-1}E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}}\right)+E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}}\right)\end{array}\right\},
\end{aligned}
$$

where $O_{p\times q}$ is a $p\times q$ zero matrix, and

$$A_{22.1}=E(\boldsymbol{K}\boldsymbol{K}^\top)-E(\boldsymbol{K}\boldsymbol{K}^{\star\top})\{E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top})\}^{-1}E(\boldsymbol{K}^\star\boldsymbol{K}^\top).$$

Therefore, it remains to show

$$E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}}\right)-E(\boldsymbol{K}\boldsymbol{K}^{\star\top})E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top})^{-1}E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}}\right)=\boldsymbol{0}.$$

This follows from the facts

$$E\left(\frac{\partial \boldsymbol{K}^\star}{\partial \boldsymbol{\phi}}\right)=-E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top}), \tag{B.20}$$

$$E\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}}\right)=-E(\boldsymbol{K}\boldsymbol{K}^{\star\top}). \tag{B.21}$$

We only show (B.20) because (B.21) can be shown in a similar way. The left-hand side of (B.20) can be calculated as

$$
\begin{aligned}
&E\left[\left\{\frac{\partial}{\partial\boldsymbol{\phi}^\top}\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})\right\}\left\{\frac{R}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}-1\right\}\right]+E\left[\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})\frac{\partial}{\partial\boldsymbol{\phi}^\top}\left\{\frac{R}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}-1\right\}\right]\\
&=E\left[\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})\frac{\partial}{\partial\boldsymbol{\phi}^\top}\left\{\frac{R}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}-1\right\}\right]\\
&=-E\left\{\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0;\boldsymbol{Z})^\top}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}\right\}\\
&=-E\left[\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})E\left\{\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0;\boldsymbol{Z})^\top}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}\;\middle|\;\boldsymbol{x}\right\}\right]\\
&=-E\left[\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})E\left\{\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\phi}_0;\boldsymbol{Z})^\top}{\pi(\boldsymbol{\phi}_0;\boldsymbol{Z})}\;\middle|\;\boldsymbol{x}\right\}\frac{O(\boldsymbol{\phi}_0;\boldsymbol{Z})}{E\{O(\boldsymbol{\phi}_0;\boldsymbol{Z})\mid\boldsymbol{x}\}}\right]\\
&=-E\left[O(\boldsymbol{\phi}_0;\boldsymbol{Z})\{\boldsymbol{g}^\star(\boldsymbol{\phi}_0;\boldsymbol{x})\}^{\otimes 2}\right]=-E(\boldsymbol{K}^\star\boldsymbol{K}^{\star\top}).
\end{aligned}
$$

Therefore the GMM estimator with $\boldsymbol{g}(\boldsymbol{\phi}; \boldsymbol{x}) = \{\boldsymbol{g}^{\star}(\boldsymbol{\phi}; \boldsymbol{x})^{\top}, \boldsymbol{k}(\boldsymbol{x})^{\top}\}^{\top}$ also attains the semiparametric efficiency bound. □

Proof of (3.6). Assume the condition in (3.5), compactness of the parameter space Ξ, and continuity of $L(\boldsymbol{\xi})$ to show $\sup_{\boldsymbol{\xi} \in \Xi_{\varepsilon}} L(\boldsymbol{\xi}) < L(\boldsymbol{\xi}_0)$ in (3.6), under three assumptions. First, we show that

$$L(\boldsymbol{\xi}) < L(\boldsymbol{\xi}_0) \quad \text{for } \boldsymbol{\xi} \neq \boldsymbol{\xi}_0. \tag{B.22}$$

In fact, we have

$$\begin{aligned}
&L(\boldsymbol{\xi}) - L(\boldsymbol{\xi}_0) \\
&= \sum_{t=1}^{T} \sum_{\boldsymbol{y} \in \{0,\,1\}^{\otimes T}} \ln \frac{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi})}{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi}_0)} g\left(\boldsymbol{r}^{(t)},\, \boldsymbol{y}\ ;\ \boldsymbol{\xi}_0\right) \\
&\leq \sum_{t=1}^{T} \sum_{\boldsymbol{y} \in \{0,\,1\}^{\otimes T}} \left\{ \frac{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi})}{g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi}_0)} - 1 \right\} g\left(\boldsymbol{r}^{(t)},\, \boldsymbol{y}\ ;\ \boldsymbol{\xi}_0\right) \\
&= 1 - 1 = 0,
\end{aligned}$$

where equality holds if and only if $g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi}) = g_t(\boldsymbol{r}^{(t)},\, \boldsymbol{y}^{(t)};\, \boldsymbol{\xi}_0)$ a.s. $P_{\boldsymbol{\xi}_0}$, which implies $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ from condition (5). Hence, we have (B.22).

Now we shall prove (6). There is a minor but important gap between the conditions in (B.22) and (6). To fill the gap, we have to show that there are no sequences $\{\boldsymbol{\xi}_n\}_{n \in \mathbb{N}}$ that tend to $\boldsymbol{\xi}^* \in \Xi \cap \Xi_{\varepsilon}$ such that they attain $L(\boldsymbol{\xi}^*) = L(\boldsymbol{\xi}_0)$. Suppose that there exists a sequence $\{\boldsymbol{\xi}_n\}_{n \in \mathbb{N}} \in \Xi \cap \Xi_{\varepsilon}$ such that $L(\boldsymbol{\xi}_n) \to L(\boldsymbol{\xi}_0)$. Due to the compactness of $\Xi \cap \Xi_{\varepsilon}$, there exists a subsequence $\{\boldsymbol{\xi}_{n_k}\}_{k \in \mathbb{N}}$ of $\{\boldsymbol{\xi}_n\}$ and $\boldsymbol{\xi}^* \in \Xi \cap \Xi_{\varepsilon}$ such that $\boldsymbol{\xi}_{n_k} \to \boldsymbol{\xi}^*$. By the continuity of L, $L(\boldsymbol{\xi}_{n_k}) \to L(\boldsymbol{\xi}^*) = L(\boldsymbol{\xi}_0)$, which contradicts (B.22). Thus, we have the conclusion. □

Proof of Lemma 3.2.1. We suppress the parameter $\boldsymbol{\xi}$ for simplicity in the following proof. For $t = 1, \ldots, T-1$, i.e., when data are missing,

$$\begin{aligned}
&g_t(\boldsymbol{y}^{(t)},\ \boldsymbol{r}^{(t)}) \\
&= \Pr(R_{t+1} = 0, R_t = 1, \boldsymbol{Y}^{(t)} = \boldsymbol{v}_t) \\
&= \sum_{y_{t+1}} \Pr(R_{t+1} = 0 \mid R_t = 1, \boldsymbol{v}_{t+1}) \\
&\quad \times \prod_{s=2}^{t} \Pr(R_s = 1 \mid R_{s-1} = 1, \boldsymbol{v}_{t+1}) \Pr(\boldsymbol{Y}^{(t+1)} = \boldsymbol{v}_{t+1}).
\end{aligned}$$

By the assumption of conditional independence, we have

$$\begin{aligned}
&g_t(\boldsymbol{y}^{(t)},\ \boldsymbol{r}^{(t)}) \\
&= \Pr(Y_1 = y_1) \prod_{s=1}^{t-1} \Pr(R_{s+1} = 1 \mid R_s = 1, \boldsymbol{v}_s) \Pr(Y_{s+1} = y_{s+1} \mid \boldsymbol{v}_s) \\
&\quad \times \sum_{y_{t+1}} \Pr(R_{t+1} = 0 \mid R_t = 1, \boldsymbol{v}_{t+1}) \Pr(Y_{t+1} = y_{t+1} \mid \boldsymbol{v}_T) \\
&= \pi_1(y_1) \prod_{s=2}^{t} \zeta_s(\boldsymbol{v}_s) \left(1 - \sum_{y_{t+1}} \zeta_{t+1}(\boldsymbol{v}_{t+1}) \right).
\end{aligned}$$

Also, for $t = T$, i.e., when data are completely observed, we have

$$g_T(\boldsymbol{y}^{(T)},\ \boldsymbol{r}^{(T)}) = \Pr(\boldsymbol{R}_T = 1, \boldsymbol{Y}^{(T)} = \boldsymbol{v}^{(T)}) = \pi_1(y_1) \prod_{s=2}^{T} \zeta_s(\boldsymbol{v}_s).$$

□

Proof of Theorem 3.2.1. We show that conditions (3.5) and (3.7) are equivalent when Y is binary, as is considered throughout §3.2, but its extension for the continuous case can be done analogously.

To show this, it suffices to prove that

$$g_t\left(\boldsymbol{r}^{(t)},\ \boldsymbol{y}^{(t)}\ ;\ \boldsymbol{\xi}\right) = g_t\left(\boldsymbol{r}^{(t)},\ \boldsymbol{y}^{(t)}\ ;\ \boldsymbol{\xi}_0\right) \quad \text{a.s. } P_{\boldsymbol{\xi}_0} \quad \forall t$$

and

$$\zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_{0,t}) \quad \text{a.s. } P_{\boldsymbol{\xi}_0} \quad \forall t$$

are equivalent, where

$$\begin{aligned}
&\pi_1(y_1;\ \boldsymbol{\xi}_1) := \Pr(Y_1 = y_1;\ \boldsymbol{\xi}_1), \\
&\zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) := \Pr\left(R_t = 1 \mid R_{t-1} = 1,\ y_{t-1},\ y_t;\ \boldsymbol{\xi}_t\right) \\
&\qquad\qquad \times \Pr\left(Y_t = y_t \mid \boldsymbol{v}_{t-1};\ \boldsymbol{\theta}_t\right) \qquad (t = 2, \ldots, T)
\end{aligned}$$

and $g_t\left(\boldsymbol{r}^{(t)},\ \boldsymbol{y}^{(t)}\right)$ is represented by the functions of $\pi_1, \ldots, \pi_T$ by Lemma 3.2.1. Note that

$$\zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_{0,t}) \quad \text{a.s. } P_{\boldsymbol{\xi}_0} \quad \forall t\ \Rightarrow \boldsymbol{\xi} = \boldsymbol{\xi}_0$$

and

$$\zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_{0,t}) \quad \text{a.s. } P_{\boldsymbol{\xi}_0} \ \Rightarrow \boldsymbol{\xi}_t = \boldsymbol{\xi}_{0,t} \quad \forall t$$

are equivalent since each expression of $\pi_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_t) = \pi_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_{0,t})$ a.s. $P_{\boldsymbol{\xi}_0}$ is in terms of each $\boldsymbol{\xi}_t$. It is obvious that (3.7) implies (3.5); thus, we only need to prove the reciprocal statement. From the condition when $t = T$, we have

$$\prod_{s=1}^{T} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}) = \prod_{s=1}^{T} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0),$$

where $\boldsymbol{v}_t = (\boldsymbol{v}_{t-1}, y_t)$. First, we assume that $\zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi})$ and $\zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0)$ are positive for all s. It follows that

$$\frac{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi})}{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0)} = \frac{\zeta_T(\boldsymbol{v}_T;\ \boldsymbol{\xi}_0)}{\zeta_T(\boldsymbol{v}_T;\ \boldsymbol{\xi})} \quad \forall y_T,$$

which implies

$$\frac{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi})}{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0)} = \frac{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi}_0)}{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi})} = \frac{\zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi}_0)}{\zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi})} \tag{B.23}$$

Similarly, from the condition when $t = T-1$, we can obtain

$$\frac{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi})}{\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0)} = \frac{(1 - \sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}_0)}{(1 - \sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}))}. \tag{B.24}$$

Thus, we have the following equalities:

$$\frac{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi}_0)}{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi})} = \frac{\zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi}_0)}{\zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi})} = \frac{(1 - \sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}_0)}{(1 - \sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}))}$$

in view of (B.23) and (B.24). This yields

$$\frac{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi}_0) + \zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi}_0)}{\zeta_T((\boldsymbol{v}_{T-1}, 0);\ \boldsymbol{\xi}) + \zeta_T((\boldsymbol{v}_{T-1}, 1);\ \boldsymbol{\xi})} = 1,$$

using the fact that if $a_1/a_2 = b_1/b_2 = c_1/c_2$, then $(a_1+b_1)/(a_2+b_2) = c_1/c_2$. Therefore, we obtain

$$\sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}) = \sum_{y_T} \zeta_T((\boldsymbol{v}_{T-1}, y_T);\ \boldsymbol{\xi}_0).$$

Substituting this into (B.24), we have

$$\prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}) = \prod_{s=1}^{T-1} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0).$$

Repeating this operation, it follows that

$$\prod_{s=1}^{t} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}) = \prod_{s=1}^{t} \zeta_s(\boldsymbol{v}_s;\ \boldsymbol{\xi}_0) \quad \forall t \in \{1, \dots T\},$$

which means

$$\zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}) = \zeta_t(\boldsymbol{v}_t;\ \boldsymbol{\xi}_0) \quad \forall t.$$

Then, we assume that there exists u such that $\zeta_u(\boldsymbol{v}_u;\ \boldsymbol{\xi}) = 0$ or $\zeta_u(\boldsymbol{v}_u;\ \boldsymbol{\xi}_0) = 0$, and define u_0 as

$$u_0 \equiv \min\{1 \leq u \leq T-1 \mid\ \zeta_u(\boldsymbol{v}_u;\ \boldsymbol{\xi}) = 0 \text{ or } \ \zeta_u(\boldsymbol{v}_u;\ \boldsymbol{\xi}_0) = 0\}.$$

Here, we assume $\zeta_{u_0}(\boldsymbol{v}_{u_0};\ \boldsymbol{\xi}) = 0$ and $\zeta_{u_0}(\boldsymbol{v}_{u_0};\ \boldsymbol{\xi}_0) > 0$ without loss of generality, and prove $\zeta_{u_0}(\boldsymbol{v}_{u_0};\ \boldsymbol{\xi}_0) = 0$. From the condition when $t = T$, the following proposition must be true:

$$\bigvee_{k=u_0+1}^{T} [\zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 0], \tag{B.25}$$

where $A \vee B$ is the logical sum of A and B, i.e., it is false if and only if both A and B are false and $\bigvee_{i=1}^{n} A_i = A_1 \vee A_2 \vee \cdots \vee A_n$. Similarly, from the condition when $t > u_0 + 1$, we have

$$\bigvee_{k=u_0+1}^{t} [\zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 0] \vee \left[\sum_{y_{t+1}} \zeta_{t+1}(\boldsymbol{v}_{t+1};\ \boldsymbol{\xi}_0) = 1\right] \tag{B.26}$$

is true, and when $t = u_0 + 1$,

$$\left[\sum_{y_{u_0+1}} \zeta_{u_0+1}(\boldsymbol{v}_{u_0+1};\ \boldsymbol{\xi}_0) = 1\right] \tag{B.27}$$

must hold. Considering the logical product of (B.25)–(B.27), it holds that

$$\bigvee_{k=u_0+1}^{T} \left[[\zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 0] \wedge \left[\sum_{y_k} \zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 1\right]\right] \tag{B.28}$$

is true. It follows from (B.28) that there exists $k \geq u_0 + 1$ such that $\zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 0$. Then, it follows that

$$\sum_{y_k} \zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 0,$$

which means that $\sum_{y_k} \zeta_k(\boldsymbol{v}_k;\ \boldsymbol{\xi}_0) = 1$ does not hold for all $k(\geq u_0 + 1)$. Thus, (B.28) cannot be true. Therefore, $\pi_{u_0}(\boldsymbol{v}_{u_0};\ \boldsymbol{\xi}_0) = 0$. □

Proof of Theorem 3.5.1. Let $\boldsymbol{\gamma}$ be a finite dimensional parameter of $f_1(y \mid \boldsymbol{x})$ and the true parameter be $\boldsymbol{\gamma}_0$. Denote $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$ if $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}) = f_1(y \mid \boldsymbol{x}; \gamma')$ holds for almost all z. Here, the distribution of $[y \mid \boldsymbol{x}]$ can be represented with $(\boldsymbol{\gamma}, \boldsymbol{\phi})$, because by using Bayes' formula, we have

$$f(y \mid \boldsymbol{x}; \gamma, \boldsymbol{\phi}) = \frac{f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})}{\int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})dy}. \tag{B.29}$$

We give a proof for Theorem 3.5.1 by taking two steps: (i) prove the identification condition for $f(y \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\phi})\pi(\boldsymbol{z}; \boldsymbol{\phi})$ is equivalent to that for $\int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})dy$ and the uniqueness of $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})$; (ii) derive the identification condition for $\int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})dy$.

(i). Wefirst show

$$f(y \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\phi})\pi(\boldsymbol{z}; \boldsymbol{\phi}) = f(y \mid \boldsymbol{x}; \boldsymbol{\gamma}', \boldsymbol{\phi}')\pi(\boldsymbol{z}; \boldsymbol{\phi}') \tag{B.30}$$

is equivalent to

$$f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}) = f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}') \tag{B.31}$$

and

$$\int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})dy = \int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}')\pi^{-1}(\boldsymbol{z}; \phi')dy \tag{B.32}$$

It follows from (B.29) that

$$f(y \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\phi})\pi(\boldsymbol{z}; \boldsymbol{\phi}) = \frac{f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})}{\int f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma})\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi})dy}. \tag{B.33}$$

Hence, (B.31) and (B.32) imply (B.30). On the contrary, by taking integration with respect to y in (B.30) by using (B.33), we have (B.32) and then, (B.31) follows from (B.32) and (B.33).

(ii). By the result of (i), f_1 is unique in the sense of (B.31). It only remains to show (B.32) with the true distribution $f_1(y \mid \boldsymbol{x}; \boldsymbol{\gamma}_0)$. By the definition of odds function $O(\boldsymbol{z}; \boldsymbol{\phi})$: $\pi^{-1}(\boldsymbol{z}; \boldsymbol{\phi}) = 1 + O(\boldsymbol{z}; \boldsymbol{\phi})$ and $\int f_1(y \mid \boldsymbol{x}; \gamma) dy{=}1$, the identification of $E_1\{O(\boldsymbol{Z}; \boldsymbol{\phi}) \mid \boldsymbol{x}; \boldsymbol{\gamma}_0\}$ is the necessary and sufficient condition for the identification of $f(y \mid \boldsymbol{x}; \boldsymbol{\gamma}, \boldsymbol{\phi})\pi(\boldsymbol{z}; \boldsymbol{\phi})$. □

References

Baker, S. G. and Laird, N. M. (1988) Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Amer. Statist. Assoc.*, **83**, 62–69. 24

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998) *Efficient and Adaptive Estimation for Semiparametric Models.* Springer. 4, 10, 74

Birminghama, J., Rotnitzky, A. and Fitzmaurice, G. M. (2003) Pattern–mixture and selection models for analysing longitudinal data with monotone missing patterns. *J. R. Statist. Soc. Ser. B*, **65**, 275–297. 24

Chang, T. and Kott, P. S. (2008) Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, **95**, 555–571. 10, 13, 18, 42

Chen, J., Sitter, R. R. and Wu, C. (2002) Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, **89**, 230–237. 13

Chen, S. and Haziza, D. (2017) Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, **104**, 439–453. 60

D'Haultfoeuille, X. (2010) A new instrumental method for dealing with endogeneous selection. *J. Econometrics*, **154**, 1–15. 10

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis. *J. R. Statist. Soc. Ser. C*, **43**, 49–93. 9, 54

Fay, R. E. (1986) Causal models for patterns of nonresponse. *J. Amer. Statist. Assoc.*, **81**, 354–365. 24

Fitzmaurice, G. M. and Laird, N. M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151. 24

Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim, J. G. (2005) A protective estimator for longitudinal binary data subject to non-ignorable non-monotone missingness. *J. R. Statist. Soc. Ser. B*, **168**, 723–735. 10

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995) Regression models for longitudinal binary responses with informative drop-outs. *J. R. Statist. Soc. Ser. B*, **57**, 691–704. 25

Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.*, **77**, 251–261. 3, 9

Hájek, J. (1970) A characterization of limiting distributions of regular estimates. *Z. Wahrscheinlichkeitstheorie verw. Gebiete.*, **4**, 323–330. 6

Han, P. (2014) Multiply robust estimation in regression analysis with missing data. *J. Amer. Statist. Assoc.*, **109**, 1159–1173. 60

Hansen, L. P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054. 13

Hansen, L. P., Heaton, H. and Yaron, A. (1996) Finite-sample properties of some alternative gmm estimators. *J. Bus. Econ. Statist.*, **14**, 262–280. 47

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685. 2

Ibragimov, I. and Has'minskii, R. Z. (1981) *Statistical Estimation: asymptotic theory.* Springer: New York. 5

Imai, K. and Ratkovic, M. (2014) Covariate balancing propensity score. *J. R. Statist. Soc. Ser. B*, **76**, 243–263. 22

Kim, J. K. (2011) Parametric fractional imputation for missing data analysis. *Biometrika*, **98**, 119–132. 20

Kim, J. K. and Yu, C. L. (2011) A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.*, **106**, 157–165. 10, 16, 51

Kott, P. S. and Chang, T. (2010) Using calibration weighting to adjust for nonignorable unit nonresponse. *J. Amer. Statist. Assoc.*, **105**, 1265–1275. 10, 13, 18, 42

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22. 11, 20

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Inference with Missing Data.* Wiley Series in Probability and Statistics, 2nd edition edn. 9

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. Ser. B*, **44**, 226–233. 11

Ma, W. Q., Geng, Z. and Hu, Y. H. (2003) Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *J. Multivar. Anal.*, **87**, 24–45. 10

Machin, D., Farley, T. M. M., Busca, B., Campbell, M. J. and d'Arcangues, C. (1988) Assessing changes in vaginal bleeding patterns in contracepting women. *Contraception*, **38**, 165–179. 23, 34

Matsuyama, Y. (2004) Analysis of missing data in longitudinal studies: A review. *Jpn. J. Biometrics*, **25**, 89–116. 24, 25, 54

Miao, W., Ding, P. and Geng, Z. (2016) Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.*, **111**, 1673–1683. 4, 25, 37, 60

Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997) The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, **84**, 33–44. 24

Molenberghs, G. and Lesaffre, E. (1994) Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *J. Amer. Statist. Assoc.*, **89**, 633–644. 24

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer-Verlag. 24, 61

Newey, W. K. (1994) Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, **10**, 233–253. 67

Newey, W. K. and McFadden, D. L. (1994) *Large Sample Estimation and Hypothesis Testing in Handbook of Econometrics*, vol. 4. Amsterdam: North Holland. 73

Park, T. and Brown, M. B. (1994) Models for categorical data with nonignorable nonresponse. *J. Amer. Statist. Assoc.*, **89**, 44–52. 24

Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating equations. *Ann. Math. Statist.*, **22**, 300–325. 78

Qin, J., Leung, D. and Shao, J. (2002) Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.*, **97**, 193–200. 10, 12, 13, 22, 78

Qin, J. and Zhang, B. (2007) Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc. Ser. B*, **69**, 101–122. 22

Riddles, M. K., Kim, J. K. and Im, J. (2016) Propensity-score-adjustment method for nonignorable nonresponse. *J. Surv. Statist. Methodol.*, **4**, 215–245. 9, 11, 14, 19, 20, 42

Robins, J. M., Rotnitzky, A. and Scharfstein, D. O. (1999) *Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models*, vol. 116 of *Statistical Models in Epidemiology: The Environment and Clinical Trials*. NY: Springer-Verlag. 10

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, **89**, 846–866. 20

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55. 2

Rotnitzky, A. and Robins, J. (1997) Analysis of semi-parametric regression models with non-ignorable non-response. *Statist. Med.*, **16**, 81–102. 10, 17, 59, 74

Rotnitzky, A., Scharfstein, D., Su, T.-L. and Robins, J. (2001) Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics*, **57**, 103–113. 9

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **61**, 581–592. 2, 9

Scharfstein, D. O., Rotnizky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.*, **94**, 1096–1146. 9

Shao, J. and Wang, L. (2016) Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, **103**, 175–187. 10, 51

Skrondal, A. and Rabe-Hesketh, S. (2014) Protective estimation of mixed-effects logistic regression when data are not missing at random. *Biometrika*, **101**, 175–188. 10

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen Tchetgen, E. J. (in press) Semiparametric estimation with data missing not at random using an instrumental variable. 60

Sverchkov, M. (2008) A new approach to estimation of response probabilities when missing data are not missing at random. *In Proc. Survey Res. Meth. Sect., Am. Statist. Ass.* Washington DC: American Statistical Association, 867-874. 9, 19

Tang, G., Little., R. J. A. and Raghunathan, T. E. (2003) Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, **90**, 747–764. 10, 35

Tang, N., Zhao, P. and Zhu, H. (2014) Empirical likelihood for estimating equations with nonignorably missing data empirical likelihood for estimating equations with nonignorably missing data. *Statist. Sinica*, **24**, 723–747. 10

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer. 4, 5, 6, 8, 9, 74, 75

van der Vaart, A. W. (1998) *Asymptotic Statistics.* Camblidge University Press. 70

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001) Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics*, **57**, 7–14. 9

Wang, C.-Y. and Pepe, M. S. (2000) Expected estimating equations to accommodate covariate measurement error. *J. R. Statist. Soc. Ser. B*, **62**, 509–524. 14

Wang, S., Shao, J. and Kim, J. K. (2014) An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statist. Sinica*, **24**, 1097–1116. 10, 25, 35, 41, 51

Wasserman, L. (2005) *All of Nonparametric Statistics.* New York: Springer. 14, 15

Wilkins, K. J. and Fitzmaurice, G. M. (2006) A hybrid model for nonignorable dropout in longitudinal binary responses. *Biometrics*, **62**, 168–76. 24, 61

Yuan, Y. and Little, R. J. A. (2009) Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, **65**, 478–86. 24

Zhao, H. and Lin, J. (2012) The large sample properties of the solutions of general estimating equations. *J. Syst. Sci. Complex.*, **25**, 315–328. 71

Zhao, J. and Shao, J. (2015) Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *J. Amer. Statist. Assoc.*, **110**, 1577–1590. 10, 35, 58

Zhao, P., Tnag, N., Qu, A. and Jiang, D. (2017) Semiparametric estimating equations inference with nonignorable nonresponse. *Statist. Sinica*, **27**, 89–113. 10

Zhong, P.-S. and Chen, S. (2014) Jackknife empirical likelihood inference with regression imputation and survey data. *J. Multivar. Anal.*, **129**, 193–205. 17

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Professor Yutaka Kano for his continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank Professor Jae Kwang Kim for allowing me to visit Iowa State University in U.S.A. and KAIST in Korea under his supervision. I could learn state of the art method in missing data analysis and survey sampling as well as the life of academic researchers in foreign countries.

I thank students of Kano laboratory for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. Also I thank alumni and alumnae of Kano laboratory for providing me advice and encouragement even after their graduation.

Last but not the least, I would like to thank my family: my parents and my sister for supporting me spiritually throughout writing this thesis and my life in general.

This thesis has been supported by the Grant-in Aid for Japan Society for the Promotion of Science (JSPS) Fellows.

List of Publications

[1]. Morikawa, K. and Kano, Y. (2014). "Statistical inference with different missing-data mechanisms", arXiv:1407.4971.

[2]. Morikawa, K. and Kano, Y. (2017). "Identification problem of transition models for repeated measurement data with nonignorable missing values", submitted to *Journal of Multivariate Analysis*, accepted for publication.

[3]. Morikawa, K. and Kim, J. K. (2017a). "Semiparametric optimal estimation with nonignorable nonresponse data", submitted to *Biometrika*.

[4]. Morikawa, K. and Kim, J. K. (2017b). "A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse", submitted to *Statistics & Probability Letters*.

[5]. Morikawa, K., Kim, J. K., and Kano, Y. (2017). "Semiparametric maximum likelihood estimation with data missing not at random", *Canadian Journal of Statistics*, **45**, 393–409.