Covariate Selection for Estimating Population Means and Procedures for Estimating Treatment Effects When Data Are Missing Not at Random

January 2025

Shintaro Yoneyama

A Thesis for the Degree of Ph.D. in Engineering

Covariate Selection for Estimating Population Means and Procedures for Estimating Treatment Effects When Data Are Missing Not at Random

January 2025

Graduate School of Science and Technology
Keio University

Shintaro Yoneyama

Summary

This thesis summarizes research on covariate selection for estimating the population means of the outcome, and on estimating the treatment effects when the outcome is missing and the mechanism is Missing Not at Random (MNAR). MNAR is a missing mechanism that depends not only on the observed values, but also on the missing values themselves, making proper adjustment more difficult than other missing mechanisms.

This thesis is organized as follows: Chapter 1 provides an introduction, reviewing studies on missing-data adjustment methods and discussing the limitations of recent MNAR data adjustment methods.

Chapter 2 describes Sun et al. (2018), which is used for the discussion in this thesis. We consider the use of instrumental variables (IV) in estimating population means. IV is a variable conditionally independent of the outcome and related to the missing mechanism when conditioning on the covariates in estimating the population mean when the outcome is MNAR. Sun et al. (2018) employ a two-step estimation procedure, first estimating the extended propensity scores, which are the probability of missingness conditional on the covariates, followed by the outcome and IV. Finally, the population mean is estimated using the aforementioned estimates.

In Chapter 3, we propose a strategy of covariate selection for the extended propensity score model in the population mean estimation using the two-step estimation procedure described in Chapter 2. First, we show the minimum set of covariates to be included in the aforementioned model for the consistency of the population mean estimation. Next, we show a good strategy is to include only the minimum set of covariates in the extended propensity score model to estimate a small bias and variance in the population mean estimate.

In Chapter 4, we propose a method to estimate the treatment effects using a two-step estimation procedure when the outcome is MNAR. As a confounder introduces bias when estimating treatment effects in observational studies, adjustments for missing and confounding factors need to be simultaneously considered. We propose a method for

estimating the average treatment effect, average treatment effect on the treatment and on the untreated without modeling the outcome in an observational study where the outcome is MNAR.

Chapter 5 summarizes the discussion on the covariate selection for estimating the population means and procedures for estimating the treatment effects when the outcome is MNAR. It also discusses the issues and future challenges of the two-stage estimation procedure for adjusting for MNAR.

Contents

Chapter 1 Introduction

Many empirical works encounter missing data. However, inferences that ignore missing data can yield biased results. For example, in cognitive function tests for older adults, the more cognitively impaired a person is, the more likely they will drop out of the test. Another example is the clinical trial for a new drug, where a prominent concern is the drug's safety. Here, subjects may withdraw from the trial due to worsening health conditions caused by its side effects. In such cases, inferences about cognitive function and the new drug's safety that ignore the missing data are not valid because they yield an overestimation of the average cognitive function and safety.

Missing mechanisms are generally classified into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR, MAR, and MNAR, the missing probability is independent of the data, depends only on the observed data, and depends on both the observed and missing data, respectively (Little & Rubin, 2002; Tsiatis, 2006). The most typical method to deal with MCAR data is listwise deletion. This specifically involves removing records with missing data, even in a single variable. Crucially, this method does not introduce bias with MCAR data due to missingness. However, it may introduce bias in the cases of MAR and MNAR data.

Numerous studies have investigated missing-data adjustment methods while assuming MAR, including likelihood models such as the selection and pattern mixture models, as outlined by Little (2008). However, properly identifying the distributions of both the outcome and missing data is required for likelihood models. An alternative approach is using the weighting method (Robins & Gill, 1997), wherein only the missing probability model must be properly identified. Specifically, the missing probabilities estimated in the first step are used to estimate the population mean of the outcome in the second step. Note that this method can be understood in parallel with the weighting method in causal inference (for example, Hirano, Imbens, & Ridder, 2003), which interprets potential outcomes as missing owing to assignment because of the similarities between MAR and assumptions in causal inference. The method of multiple imputation by chained

equations (MICE, Kennickell (1991); van Buuren & Groothuis-Oudshoorn (2011)) enables the adjustment of the MAR data even with missing data in several variables. MICE creates multiple imputed datasets, performs the estimation of interest on each imputed dataset, and merges the results.

Compared with the abundant literature on MAR data, few studies have investigated how to adjust for MNAR data. However, MNAR data can easily occur, such as in the aforementioned examples of cognitive function tests and clinical trials. Hereafter, we consider a case in which only the outcome may be missing, and the missing mechanism is MNAR. One approach considered in this case is the likelihood method. However, for this method to work effectively, the distributions of both the missing mechanism and outcome must be properly specified. In addition, further assumptions must be made to maintain identifiability (Wang, Shao, & Kim, 2014; Miao, Ding, & Geng, 2016). To solve this problem, several semiparametric MNAR adjustment and identification methods have been proposed recently. These methods are characterized by the use of independency among variables. d'Haultfoeuille (2010) introduced an estimation method using variables independent of the missing mechanism under conditioning on the outcome but related to the outcome. Miao & Tchetgen Tchetgen (2016) proposed a method for estimating the population mean of the outcome using a shadow variable that is independent of the missing mechanism under the covariates and outcome conditioning but is related to the outcome when conditioned with the covariates. Sun, Liu, Miao, Wirth, Robins, & Tchetgen Tchetgen (2018) proposed a method using an instrumental variable (IV) conditionally independent of the outcome given the covariates and is related to the missing pattern. The most standard IV method is used in linear regression models to prevent bias in coefficients when covariates and error terms are correlated (e.g., Angrist & Krueger, 1999). In this case, the IV is a variable related to the covariates but not to the outcome, except through the covariates. An example is a physician's medication instructions in a drug efficacy trial, wherein some subjects do not comply with the instructions (Angrist, Imbens, & Rubin, 1996). This is because medication instructions affect health status only through the actual medication. Let us now return to the method proposed by Sun et al. (2018). Similar to the MAR weighting method, these methods utilize a two-step procedure to estimate the population mean. The difference between the method proposed by Sun et al. (2018) and the MAR weighting method is that, in the first step, the missing probability conditional on both the covariates and

outcome is estimated. Because the outcome is missing with MNAR, estimating the missing probability conditional on the outcome is not possible using ordinary methods; therefore, an alternative method, such as that of Sun et al. (2018), is required. The necessary modeling for adjusting MNAR data is not easy and requires more expertise and background knowledge because models of missing mechanisms of outcomes include the outcomes themselves.

Despite the ease with which MNAR data can be generated, some difficulties remain, such as the limited targets for estimation in extant studies and the difficulty of modeling. To alleviate this problem and make MNAR data adjustment more accessible, we make two proposals to the MNAR data adjustment method, in particular, to the method of Sun et al. (2018). First, we propose a covariate selection strategy for the extended propensity score model, which is good in terms of bias and precision. Second, we propose a method to estimate treatment effects when the outcome is MNAR in an observational study. The treatment effect is the difference between outcomes with and without the treatment and is one of the major research interests.

The covariate selection strategy not only improves bias and precision but also alleviates the difficulty of modeling the extended propensity score, which requires the qualitative consideration of variables. To our knowledge, no study has focused on covariate selection in the case of MNAR. Still, studies have focused on covariate selection in the case of MAR using a two-step procedure that adjusts the missing outcome and estimates its population mean (Seaman & White, 2013; Wang & Kim, 2022). These studies showed that including variables in the model of missing probability that are related only to missing but not related to the outcome may worsen the precision of estimating population means, which are of primary interest. Similar arguments are made for causal inference. Specifically, the mean of the potential outcome is estimated using a two-step procedure, which is understood in parallel with the estimation of the population mean of the outcome, which is MAR. However, some argue that including variables unrelated to the outcome in the model of assignment reduces the precision of estimating the mean of the potential outcome (Brookhart, Schneeweiss, Rothman, Glynn, Avorn, & Stürmer, 2006; De Luna, Waernbaum, & Richardson, 2011). Yoneyama & Minami (2024) extended the above discussion to the case of MNAR and proposed a covariate selection strategy of the extended propensity score model that is better in terms of bias and precision of the population mean estimates.

Compared with experimental studies, estimating treatment effects is more difficult in observational studies as the investigator has no control over the assignment. This is because differences in outcomes with and without assignment may not depend on the assignment but on the covariates related to both the assignment and outcome. Therefore, with the effect of such covariates, the confounder must be removed to identify the effect of the assignment. Here, we consider the estimation of treatment effects when the outcome is MNAR. Here, we need to simultaneously consider the adjustment for both the confounder and missing data. One such method is that of Li & Zhou (2017), who extended the MNAR adjustment method proposed by Miao & Tchetgen Tchetgen (2016) to estimate the treatment effects. However, the method of Li & Zhou (2017) requires outcome modeling, which undermines the advantage of the method proposed by Miao & Tchetgen Tchetgen (2016), wherein outcome modeling can be avoided. Therefore, Yoneyama & Minami (2023) proposed a method to simultaneously adjust the MNAR outcome and confounder without outcome modeling.

The remainder of this thesis is organized as follows: Chapter 2 outlines a method for estimating the population mean when the outcome data is missing. This includes Sun et al. (2018)'s extended propensity score method for adjustment in the case of MNAR, forming the basis of this thesis. Chapter 3 discusses the variables that should be included or excluded from the extended propensity score model to obtain an unbiased estimate of the population mean with small standard errors. First, we show which covariates, at a minimum, should be included in this model so that the population mean estimator of the outcome is consistent. Next, we show that including some covariates in the missing probability model generates a large variance in the population mean estimates, even if they explain the missing probability well. We verify these arguments using simulation experiments and argue that to obtain unbiased, small-variance estimates of the population mean, a desirable strategy is to only include the covariates necessary for consistency. Chapter 4 proposes a method to estimate the average treatment effect (ATE), average treatment effect on the treated (ATT), and average treatment effect on the untreated (ATU) without modeling the outcome when it is MNAR. We demonstrate that the proposed estimators for ATE, ATT, and ATU are consistent and asymptotically normal. We also demonstrate that the proposed estimators perform better in terms of bias and root mean squared error than other commonly used missing-data adjustment

methods. Moreover, we verify the proposed estimator's variance formula through simulated experiments. Chapter 5 provides a conclusion of this thesis. We summarize the covariate selection in the estimation of the population means and procedure for estimating treatment effects when the outcome is MNAR, and finally, discuss future research directions.

Chapter 2 Preliminaries of missing data adjustment

In preparation for later discussion, this chapter outlines the methods of previous studies on how to estimate the population mean of a variable of interest when there are missing data for the variable of interest. The mechanism of missing data is generally categorized into three types:

1. MCAR (missing completely at random): a missing mechanism that does not depend on either the missing values or the observed values.

2. MAR (missing at random): a missing mechanism that depends only on observed values, but not on missing values.

3. MNAR (missing not at random): a missing mechanism that depends not only on observed values but also on missing values.

In the following, we outline methods for estimating the population mean of a variable of interest when that variable has missing data, in each missing mechanism.

2.1 Estimation of population means in the case of MCAR and MAR

First, we introduce the common notation for the case of MCAR and MAR. Then, we describe how to estimate the population mean of the variable of interest when the missing mechanism is MCAR and MAR, respectively.

2.1.1 Notation of the case of MCAR and MAR

The variable of interest for the population mean is referred to here as the outcome. Let Y be the outcome and ϕ be its population mean, that is, $\phi = \mathbb{E}[Y]$. We assume that the outcome may be missing , and the mechanism is MAR. Let R be the binary missing indicator for Y, such that $R = 1$ if Y is observed and $R = 0$ otherwise. In the case of MAR, whether the data is missing depends on the observed variables, so we consider adjusting for missing data by using other observed variables. Let X be a vector of covariates, variables observed other than the outcome and missing indicator. We

assume that missing data occur only in the outcome and we observe N independent and identically distributed observations (Y^*, R, X) where Y^* is the actual observed outcome, and Y^* is Y when $R = 1$ and missing when $R = 0$.

2.1.2 Estimation of population means in the case of MCAR

Assume that the notation is Section 2.1.1 and that the missing mechanism is MCAR. From each observation is i.i.d. and the missing mechanism does not depend on either observed and missing values, the following equation holds:

$$\Pr(R = 1|Y) = \Pr(R = 1),$$

i.e.,

$$Y \perp\!\!\!\perp R.$$

In the case of MCAR, the population mean can be estimated without bias by the sample mean ignoring missing data because the outcome and the missing mechanism are independent. However, in the case of MAR, because the missing mechanism depend on other observed variables and these variables might be related to the outcome, the outcome and the missing mechanism may not be independent as well as in the case of MNAR. Thus, in the case of MAR and NMAR, ignoring missing data might lead to a bias in the mean estimate.

2.1.3 Estimation of population means in the case of MAR

More studies have investigated methods for adjusting missing data while assuming MAR than MNAR. Here we describe one of the MAR adjustment methods, which uses the propensity score for missingness to estimate the population mean of a variable of interest when the variable with MAR. The notation remains the same as in Section 2.1.1, and the missing mechanism is assumed to be MAR. From each observation is i.i.d. and the missing mechanism depends only on observed value, so the following equation holds:

$$\Pr(R = 1|Y, X) = \Pr(R = 1|X),$$

i.e.,

$$Y \perp\!\!\!\perp R|X. \tag{2.1}$$

This implies that the covariate vector X includes all variables related to both the outcome Y and the missing indicator R. Also, the assumed relationship between variables in the case of MAR can be expressed in directed acyclic graph (DAG) as follows.

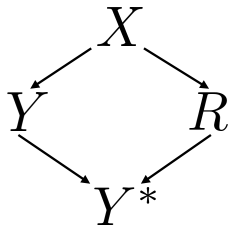

Figure 1: DAG illustrating the relationship of variables in the case of MAR.

The absence of an arrow between Y and R indicates that there is no direct relationship between Y and R.

One way to adjust for missing data in the case of MAR is to use a propensity score for missingness (c.f. Tsiatis (2006)). The propensity score for missingness $\pi_{R=1}(x)$ is the conditional probability that the outcome is observed given covariates x,

$$\pi_{R=1}(x) = \Pr(R = 1|X = x).$$

Here we assume that $\pi_{R=1}(x) > 0$ for all $x \in \mathrm{Supp}(X)$. Using the propensity score for missingness, the population mean of the outcome can be expressed as $\mathbb{E}[Y] = \mathbb{E}[RY/\pi_{R=1}(X)]$, because

$$\begin{aligned}
\mathbb{E}\left[\frac{RY}{\pi_{R=1}(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{RY}{\pi_{R=1}(X)}\middle|X\right]\right] && \text{(by the law of iterated expectaions)} \\
&= \mathbb{E}\left[\frac{\mathbb{E}[R|X]\mathbb{E}[Y|X]}{\pi_{R=1}(X)}\right] && \text{(by MAR (2.1))} \\
&= \mathbb{E}\left[\frac{\pi_{R=1}(X)\mathbb{E}[Y|X]}{\pi_{R=1}(X)}\right] \\
&= \mathbb{E}[Y].
\end{aligned}$$

Using this property, we can define the following inverse probability weighted (IPW) estimator for the population mean $\phi = \mathbb{E}[Y]$,

$$\hat{\phi}_{MAR} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i Y_i^*}{\hat{\pi}_{R=1}(X_i)},$$

where $\hat{\pi}_{R=1}(x)$ is the estimator of the propensity score for missingness. Note that $\hat{\phi}_{MAR}$ can be computed even if Y_i is missing since $R_i Y_i^*/\hat{\pi}_{R=1}(X_i) = 0$ when Y_i is missing. Suppose $\pi_{R=1}(x)$ can be represented as $\pi_{R=1}(x;\theta)$ with a parameter vector θ, and $\hat{\pi}_{R=1}(x)$ can be represented as $\pi_{R=1}(x;\hat{\theta})$ with an estimator $\hat{\theta}$ for θ. Assume $N^{-1}\sum_{i=1}^{N} R_i Y_i^*/\pi_{R=1}(X_i;\theta)$ converges in probability to $\mathbb{E}[RY^*/\pi_{R=1}(X;\theta)]$ uniformly in an open neighborhood of θ_0, the true value of θ, and $\mathbb{E}[RY^*/\pi_{R=1}(X;\theta)]$ is continuous at θ_0. Then, if $\hat{\pi}_{R=1}(x)$ is a consistent estimator for $\pi_{R=1}(x)$, i.e., $\hat{\theta} \xrightarrow{p} \theta_0$, $\hat{\phi}_{MAR}$ is a consistent estimator for ϕ by Lemma A.1 in Appendix A. As such, the method of estimating once what is not of direct interest and then using that estimate to estimate what is of original interest is referred to as a two-step procedure. Note that while other adjustment methods for missing data require modeling of the outcome, however, the method using the propensity scores does not require modeling of the outcome. Related to this, we state the two advantages of using the propensity score. The first is that modeling the propensity score is easier than directly modeling the outcome, including missing data. The second is that the propensity score is empirically robust to model misspecification (Drake, 1993; Cepeda, Boston, Farrar, & Strom, 2003).

2.2 Estimation of population mean in the case of MNAR

Next, we consider the case where the missing mechanism is MNAR. For this case, Sun et al. (2018) proposed the MNAR missing adjustment method to estimate the population mean. In this thesis, we base our discussion on their method. Here we outline the population estimation method of Sun et al. (2018).

2.2.1 Problem setting in the case of MNAR

We continue to consider the estimation of the population mean of the outcome $\phi = \mathbb{E}[Y]$. First, assume that the actual outcome Y^* and the missing indicator R are obtained and the missing mechanism is MNAR. In the case of MNAR, we additionally use an instrumental variable (IV) for adjustment. Let Z be an IV and here X be the covariates vector, observed variables other than the outcome, the missing indicator and the IV. Assume that the following two assumptions hold for Y, R, X, and Z.

Assumption 2.1 (exclusion restriction (Sun et al., 2018))**.** *Conditioning on covariates X, the outcome Y and the IV Z are independent, that is,*

$$Y \perp\!\!\!\perp Z|X.$$

Assumption 2.2 (IV relevance (Sun et al., 2018)). *For any x, z, and z' ($z \neq z'$), the following holds:*

$$\Pr(R = 1|X = x, Z = z) \neq \Pr(R = 1|X = x, Z = z').$$

These assumptions characterize the IV. Assumption 2.1 implies that the IV is not related to the outcome except through covariates. Assumption 2.2 implies that missing mechanism is related to the IV in the sense that the missing probability is different for different values of z with x fixed. Note that the IVs can be multidimensional. From the assumption that the missing mechanism is MNAR, the following conditional independence holds for Y, R, X, and Z:

$$Y \not\perp\!\!\!\perp R|X, Z.$$

Also, the assumed relationship between variables can be expressed in DAG as follows.

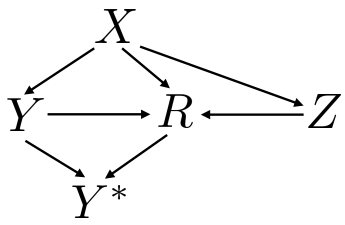

Figure 2: DAG illustrating the relationship of variables in the case of MNAR.

Again, we assume that missing data occur only in the outcome and we observe N independent and identically distributed observations (Y^*, R, X, Z).

2.2.2 Extended propensity score

Here, we outline the two-step population mean estimation method of Sun et al. (2018). It can be summarized as follows.

- At the first step, we estimate the extended propensity score, which is the probability that the outcome is observed conditional on the covariates, the outcome, and the IV.
- At the second step, we estimate the population mean with the IPW estimator using the extended propensity score estimates.

The extended propensity score is defined as follows.

Definition 2.1 (the extended propensity score (Sun et al., 2018))**.** *The extended propensity score $\pi(X, Y, Z)$ is defined as the probability that the outcome is observed conditional on X, Y, and Z:*

$$\pi(X, Y, Z) = \Pr(R = 1|X, Y, Z).$$

We assume that the extended propensity score follows a logistic regression model with parameter vectors $\left(\theta^T, \nu^T\right)^T$ in which the linear predictor does not include the product term of Z and Y so that the model can be expressed as

$$\pi(X, Y, Z; \theta, \nu) = \text{expit}\left(\theta^T h(X, Z) + \nu^T g(X, Y)\right) \tag{2.2}$$

where $\text{expit}(a) = (1 + \exp(-a))^{-1}$, h and g are vector functions, and θ and ν are their coefficients. To uniquely decompose the linear predictor into h and g, we assume that $g(X, Y)$ consists only of elements associated with Y. We also assume that h and g are differentiable in Z and Y, respectively. These constraints on the model are necessary to guarantee the identifiability of the extended propensity score model (Sun et al., 2018, Example 2).

Let us first consider the estimation of the extended propensity score. Note that the extended propensity score cannot be directly estimated using a logistic regression model because variable Y contains missing values. To solve this problem, we use an IV. First, let $f(X; \xi)$ denote the expectation of an IV, Z, given X with parameter vector ξ:

$$f(X; \xi) = \mathbb{E}[Z|X; \xi]. \tag{2.3}$$

We assume that we can obtain an m-estimator (Newey & McFadden, 1994; van der Vaart, 1998; Tsiatis, 2006) $\hat{\xi}$ for ξ, such as the maximum likelihood estimator.

For estimation of θ and ν in the extended propensity score model (2.2), we define an estimating function U for θ and ν given $\hat{\xi}$ as:

$$U(O; \theta, \nu, \xi = \hat{\xi}) = \begin{pmatrix} U_1(O; \theta, \nu) \\ U_2(O; \theta, \nu, \xi = \hat{\xi}) \end{pmatrix}, \tag{2.4}$$

where $O = (Y^*, R, X, Z)$,

$$\begin{aligned} U_1(O;\theta,\nu) = &\quad \left\{ \frac{R}{\pi(X, Y^*, Z;\theta,\nu)} - 1 \right\} h^*(X, Z); \\ U_2(O;\theta,\nu,\xi) = &\quad \frac{R}{\pi(X, Y^*, Z;\theta,\nu)} \left(Z - f(X;\xi)\right) \otimes g^*(X, Y^*); \end{aligned}$$

h^* and g^* are vector-valued functions which are of the same sizes as θ and ν, respectively, such that the estimating functions satisfy the conditions 1) and 2) of the Lemma A.2 and conditions 2) – 4) of Lemma A.3; and $\otimes$ denotes the Kronecker product. Note that h^* and g^* need not necessarily be the same as h and g in the model (2.2) because how h^* and g^* are chosen does not affect the consistency for θ and ν as discussed Sun et al. (2018). In this chapter, we set h^* and g^* to be h and g, respectively.

Let $\hat{\theta}, \hat{\nu}$ be the solution of the following estimating equation for θ, ν under given $\hat{\xi}$:

$$\frac{1}{N} \sum_{i=1}^{N} U(O_i;\theta,\nu,\xi = \hat{\xi}) = \mathbf{0}. \tag{2.5}$$

Given that $\mathbb{E}[U(O;\theta_0,\nu_0,\xi_0)] = \mathbf{0}$ holds, $\hat{\theta}$ and $\hat{\nu}$ are m-estimators under suitable regularity conditions as shown in the proof of the Proposition 1 of Sun et al. (2018).

When the IVs are multidimensional, θ and ν can be estimated using the generalized method of moment (GMM, Newey & McFadden (1994)). Even in this case, the following arguments still hold with the GMM estimates $\hat{\theta}, \hat{\nu}$ for θ, ν.

Plugging $\hat{\theta}, \hat{\nu}$ obtained above into θ, ν of the extended propensity score model (2.2) yields an estimator for the extended propensity score. Note that estimates of the extended propensity score are obtained only if Y is observed.

In the second step, we employ the following IPW estimator $\hat{\phi}_{MNAR}$ for the population mean $\phi = \mathbb{E}[Y]$:

$$\hat{\phi}_{MNAR} = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i^*}{\pi(X_i, Y_i^*, Z_i;\hat{\theta},\hat{\nu})}. \tag{2.6}$$

For each term in the summation in $\hat{\phi}_{MNAR}$, if Y_i is missing, R_i is zero and the term is zero, so this estimator can be computed. The following property holds for $\hat{\phi}_{MNAR}$.

Proposition 2.1 (consistency and asymptotic normality (Sun et al., 2018))**.** *Suppose that Assumption 2.1 and 2.2 hold. Assume conditions of Lemma A.2 and Lemma A.3 hold when the IV is one-dimensional, or conditions of Lemma A.4 and Lemma A.5 hold*

when the IVs are multidimensional. Then, the IPW estimator $\hat{\phi}_{MNAR}$ *is consistent and asymptotically normal for the population mean* ϕ*.*

Using the extended propensity score estimator from the first step, we are able to estimate the population mean in the second step.

Chapter 3 Covariate selection strategy for the extended propensity score

In the method for estimating the population mean of the outcome in the case of MAR, which had been described in Section 2.1.3, there has been some discussion recently on the covariate selection of the propensity score model for missingness (Seaman & White, 2013; Wang & Kim, 2022). A similar argument is also made in the estimation using a two-step procedure for the mean of the potential outcome in causal inference, which is understood as a parallel to the adjustment of the MAR (Brookhart et al., 2006; De Luna et al., 2011). These studies showed that including covariates in the model of missing or assignment probabilities that are related only to missing or assignment but not related to the outcome may worsen the precision of estimating the outcome or potential outcome means, which are of primary interest. In the context of causal inference, Shortreed & Ertefaie (2017) and Zhou & Jia (2021) proposed methods for variable selection in the model of assignment probabilities by extending the adaptive lasso (Zou, 2006). Nonetheless, the abovementioned discussion has not taken place regarding the MNAR case.

In response to this issue, Yoneyama & Minami (2024) investigated which covariates should be included in or excluded from the extended propensity score model to obtain estimates with lower bias and variance for the population mean when the outcome is MNAR and the missing data is adjusted using Sun et al. (2018)'s method. First, we show which covariates, at a minimum, should be included in the extended propensity score model in order for the population mean estimator of the outcome to be consistent. Next, we show that inclusion of some covariates in the extended propensity score model results in a large variance of the population mean estimates even if they explain the extended propensity score well. Then, we verify these arguments using simulation experiments and argue that, in terms of bias and variance of the estimates, it is better to include in the extended propensity score model only those covariates that should be minimally included for consistency.

This chapter is organized as follows. First, we discuss the minimal set of covariates

that should be included in the extended propensity score to obtain a consistent and asymptotically normal estimator of the population mean (Section 3.1.1). We also discuss the existence of covariates whose inclusion in the extended propensity score model would inadvertently worsen the precision of estimating the population mean of the outcome (Section 3.1.2). We then verify the claims of Section 3.1 through simulations and further argue that including only the minimal set of covariates in the extended propensity score model is a good strategy for variable selection in terms of bias and precision (Section 3.2). Finally, we conclude with the discussion (Section 3.3).

3.1 Covariates for Estimating the Extended Propensity Score

This chapter continues to discuss estimating population means using extended propensity scores under the problem setting for the case of MNAR discussed in Section 2.2. In this section, we first discuss the minimal set of covariates that should be included in the extended propensity score model for the consistent and asymptotically normal estimation of the population mean of the outcome. Then, we argue that inclusion of some covariates in the extended propensity score model would worsen the precision of the mean estimate of the outcome even if the covariates improve the estimation of the propensity score.

3.1.1 Minimal sets of covariates

In this subsection we discuss the sets of covariates that must be included in the extended propensity score model to obtain a consistent and asymptotically normal estimator for the population mean. We first define the sets of covariates X_{IV} and X_{PS} as follows.

Definition 3.1 (Minimal sets of covariates X_{IV} and X_{PS})**.** *We define the minimal sets of covariates X_{IV} and X_{PS} as follows.*

- *X_{IV} is a set of covariates that satisfies the following.*
  - *When conditioning X_{IV}, the outcome and the IV become conditionally independent. That is,*
    $$Y \perp\!\!\!\perp Z | X_{IV}. \tag{3.1}$$
  - *If any covariate is removed from X_{IV}, the above conditional independence (3.1) does not hold.*
- *X_{PS} is a set of covariates that satisfies the following.*

– *When conditioning X_{PS}, Y, and Z, X_{IV} and the missing indicator become conditionally independent. That is,*

$$X_{IV} \perp\!\!\!\perp R | X_{PS}, Y, Z. \tag{3.2}$$

– *If any covariate is removed from X_{PS}, the above conditional independence (3.2) does not hold.*

Here, "minimal" means that each conditional independence is no longer valid if any covariate is removed from X_{IV} or X_{PS}. The assumed relationship between variables can be expressed in DAG as Figure 3.

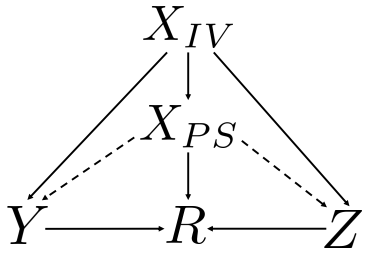

Figure 3: DAG illustrating the relationship among Y, Z, R, X_{IV} and X_{PS}. Dashed arrows indicate that there may or may not be a relationship between the variables connected by it.

If X_{IV} is not an empty set and there is no arrow from X_{IV} to R (Figure 4), then X_{PS} is an empty set. If X_{IV} is an empty set, then X_{PS} is also an empty set by its definition.

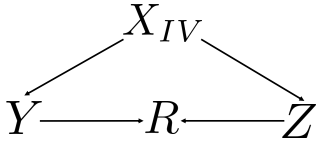

Figure 4: DAG illustrating the relationship among Y, Z, R and X_{IV} when X_{IV} is not an empty set but X_{PS} is an empty set.

In the following, we show that models for the IV and the extended propensity score should include X_{IV} and X_{PS}, respectively, to obtain the unbiased population mean estimate. Note that X_{IV} and X_{PS} may not be unique, but this does not affect the

consistency and asymptotic normality of estimation. We assume that IV relevance also holds for X_{PS}.

Assumption 3.1 (IV relevance with X_{PS})**.** *For any x_{ps}, z, and z' ($z \neq z'$), the following holds.*

$$\Pr(R = 1 | X_{PS} = x_{ps}, Z = z) \neq \Pr(R = 1 | X_{PS} = x_{ps}, Z = z').$$

Here, we show that the consistency and asymptotic normality of the IPW estimator can be maintained by using X_{IV} and X_{PS} in the model (2.3) and (2.2), respectively, instead of using the same X in both models. Allowing different covariate sets to be used in the two models differs from the approach of Sun et al. (2018).

We now consider the case where only the minimal sets of covariates X_{IV} and X_{PS} are used as covariates in the model (2.3) and (2.2), respectively. That is, let

$$f(X_{IV}; \xi) = \mathbb{E}[Z | X_{IV}; \xi] \tag{3.3}$$

be the model (2.3) for the conditional mean of the IV given X_{IV}, and let

$$\pi(X_{PS}, Y, Z; \theta, \nu) = \text{expit}\left(\theta^T h(X_{PS}, Z) + \nu^T g(X_{PS}, Y)\right), \tag{3.4}$$

be the model for the extended propensity score (2.2). We estimate $\hat{\xi}$ and $\hat{\theta}, \hat{\nu}$ by using models (3.3) and (3.4). Then, we employ the following IPW estimator $\hat{\phi}_{MS}$ for the mean ϕ of the outcome Y using the extended propensity score with a minimal set of covariates,

$$\hat{\phi}_{MS} \coloneqq \frac{1}{N} \sum_{i=1}^{N} \frac{R_i Y_i^*}{\pi(X_{PSi}, Y_i^*, Z_i; \hat{\theta}, \hat{\nu})}. \tag{3.5}$$

For this IPW estimator $\hat{\phi}_{MS}$, the following proposition holds.

Proposition 3.1. *Suppose that Assumption 3.1 holds. Assume conditions of Lemma A.2 and Lemma A.3 hold when the IV is one-dimensional, or conditions of Lemma A.4 and Lemma A.5 hold when the IVs are multidimensional. Then, the IPW estimator $\hat{\phi}_{MS}$ using the extended propensity score with the minimal set of covariates is consistent and asymptotically normal for the population mean ϕ.*

The proof and the asymptotic variance matrix is shown in Appendix B.1.

3.1.2 Covariates that should not be included in the extended propensity score model

For the case that the missing mechanism is MAR rather than MNAR in the same setting as we are considering, Robins, Rotnitzky, & Zhao (1994) proposed a two-step method to estimate the population mean of the outcome. Seaman & White (2013) and Wang & Kim (2022) reported that including covariates that are associated with the missing mechanism but not with the outcome in the model of the missing probability reduces the precision of the estimation of the population mean, even if the missing probability can be estimated with good precision.

We argue that also in the case with the MNAR, there are covariates whose inclusion in the extended propensity score estimation model reduces the precision of the population mean estimates. For example, covariates that are associated with missing but not with the outcome are likely to worsen the precision of the estimation of the population mean in the case of MNAR, as well as in the case of MAR. Furthermore, there may be other covariates that negatively affect the precision of the estimation, as they differ from MAR in some ways, such as estimating the model of the IV. To verify this, we conduct simulation experiments and check the following.

- The population mean can be estimated without bias when the minimal set of covariates X_{IV} and X_{PS} are used in the models of the IV and the extended propensity score, respectively.
- Removing even one covariate from the minimal set of covariates X_{PS} for the model (3.4) will cause a bias in the estimation of the population mean.
- The existence of covariates whose inclusion in the model of the extended propensity score reduces the precision of the estimation of the population mean.

3.2 Simulation experiments

In this section, we investigate the bias, standard deviation (SD), and root-mean-squared error (RMSE) of the mean estimates through simulation experiments using various settings for the extended propensity score model.

3.2.1 Simulation scenarios

We generated datasets for $3 \times 3 = 9$ scenarios with the following model:

$$\begin{aligned} X = (X_1, X_2, X_3)^T &\sim N(\mathbf{0},\, I)\,, \\ Z|X = 1 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon_z, &\quad \varepsilon_z \sim N(0,\, \sigma^2), \\ Y|X = 1 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_y, &\quad \varepsilon_y \sim N(0,\, 1)\,, \qquad \text{and} \\ R|X, Y, Z &\sim \text{Bernoulli}\left[\text{expit}\left(-1 + 0.35 X_1 + \gamma X_3 + 0.5 Z + 0.5 Y\right)\right], \end{aligned}$$

where $(\alpha_1, \alpha_2, \sigma)$ and $(\beta_1, \beta_2, \gamma)$ were set to be all combinations of the following patterns:

$(\alpha_1, \alpha_2, \sigma)$	A1: $(0,\ 0,\ 1)$	A2: $(0.7,\ 0,\ 0.7)$	A3: $(0,\ 0.7,\ 0.7)$
$(\beta_1, \beta_2, \gamma)$	B1: $(0.07,\ 0.07,\ 0.35)$	B2: $(0.21,\ 0.21,\ 0.35)$	B3: $(0.07,\ 0.07,\ 0.7)$.

We denote by Scenario i-j $(i, j = 1, 2, 3)$ the scenario conducted with values of the pattern Ai and the pattern Bj for the conditional distribution model of Z and Y given X, and R given X, Y, and Z.

In all scenarios, X_1 is related to both the missing mechanism and the outcome, that is, the confounder; X_2 is related to the outcome but not directly related to the missing mechanism; and X_3 is related to the missing mechanism but not related to the outcome.

In this model, the minimal sets of covariates X_{IV} and X_{PS} depend only on the values

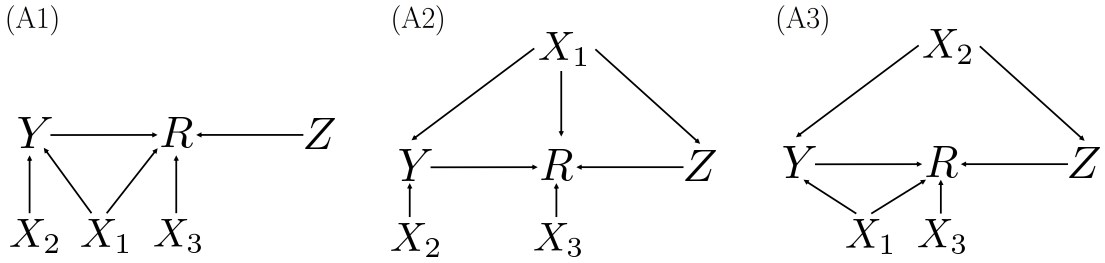

Figure 5: Left: DAG of pattern A1. Middle: DAG of pattern A2. Right: DAG of pattern A3.

For these DAGs, $X_{IV} = X_{PS} = \varnothing$ for A1, $X_{IV} = X_{PS} = \{X_1\}$ for A2, and $X_{IV} = \{X_2\}$ and $X_{PS} = \varnothing$ for A3. As for the setting of $(\beta_1, \beta_2, \gamma)$ in the conditional

distribution model of Y given X, and R given X, Y, and Z, we consider B1 as baseline. B2 has a stronger association between the covariates and the outcome than B1 has. With B3, X_3 is more relevant to missingness.

For each of the combinations of the 3×3 scenarios, we generated 10000 datasets with sample sizes $N = 5000$ and 10000 of observation vectors $O = (Y^*, R, X, Z)$.

3.2.2 Models for estimation

For each dataset of all scenarios, we estimated the population mean of the outcome ϕ using the following models:

1. Linear regression models with X_{IV} as covariates for model (2.3) of the IV, Z to estimate regression coefficients ξ (when X_{IV} is empty, $\hat{\xi}$ is the sample mean of Z).

2. The following eight models as the extended propensity score model (2.2). These models correspond to all possible combinations of $X = (X_1, X_2, X_3)$.

   Model 1: $\text{expit}(\theta_1 + \theta_2 Z + \nu Y)$ (no covariate)
   Model 2: $\text{expit}(\theta_1 + \theta_2 X_1 + \theta_3 Z + \nu Y)$ (covariate with confounder)
   Model 3: $\text{expit}(\theta_1 + \theta_2 X_2 + \theta_3 Z + \nu Y)$ (covariate not related missing)
   Model 4: $\text{expit}(\theta_1 + \theta_2 X_3 + \theta_3 Z + \nu Y)$ (covariate not related outcome)
   Model 5: $\text{expit}(\theta_1 + \theta_2 X_1 + \theta_3 X_2 + \theta_4 Z + \nu Y)$ (covariates related outcome)
   Model 6: $\text{expit}(\theta_1 + \theta_2 X_1 + \theta_3 X_3 + \theta_4 Z + \nu Y)$ (true model)
   Model 7: $\text{expit}(\theta_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 Z + \nu Y)$ (covariates without confounder)
   Model 8: $\text{expit}(\theta_1 + \theta_2 X_1 + \theta_3 X_2 + \theta_4 X_3 + \theta_5 Z + \nu Y)$ (full model)

   We included Y and Z in all eight models because they clearly play an essential role in adjusting for the MNAR outcome. We solved the estimating equation (2.5) for θ, ν with $\hat{\xi}$ obtained in 1 using R package GMM (Chaussé, 2010).

3. The IPW estimator (2.6) to estimate the population mean with the extended propensity score estimates obtained in 2.

Note that in Scenario 2-1 with sample size 5000, there was one dataset out of 10000 for which the program for solving estimating equation (2.5) of the R package GMM did not converge. This case was excluded from the evaluation of the results.

We also computed simple sample averages of the outcomes as a naive method.

3.2.3 Simulation results

Here, we review the simulation results from the following three aspects:

1) Relationship between covariates included in the extended propensity score model and bias in the population mean estimates of the outcome,

2) Existence of covariates that increase the SD of the population mean estimate when included in the extended propensity score model,

3) Which of the extended propensity score models 1-8 is best in terms of the RMSE of the population mean estimates

Figures 6, 7, and 8 show boxplots representing the results for scenarios A1, A2 and A3, respectively. First, let us point out that simple sample averages are biased in all cases.

Regarding the first aspect, let us confirm that when X_{PS} is included in the covariates of the extended propensity score models, the population mean estimates do not have bias. Figures 6 and 8 show that when X_{PS} is an empty set (scenarios A1 and A3), there is no bias in the population mean estimates with all extended propensity score models (i.e., 1-8). Also, the estimates of extended propensity score models 2, 5, 6, and 8 for scenario A2 are unbiased. However, the estimates of extended propensity score models 1, 3, 4 and 7 for scenario A2 are clearly biased. For scenarios A2, $X_{PS} = \{X_1\}$ and the models with biased estimates do not include X_1 as a covariate. This presence or absence of bias corresponds to whether all covariates in X_{PS} are included in the extended propensity score models, or not. This can be confirmed numerically from Table 1.

Next, for the second aspect, we examine how including a covariate that is not related to the outcome into the extended propensity score model can increase the SD of the population mean estimates. In this simulation setting, X_3 corresponds to this covariate, that is, it is related to the missing mechanism, but not related to the outcome. Table 1 shows that for all scenarios and sample sizes, both SD and RMSE increased when X_3 was additionally included. For all scenarios, the estimates of model 4 had larger SD and RMSE than those of model 1 where model 4 includes one more explanatory variable (i.e., X_3) than model 1 does. The same results are observed for the estimates of models $6, 7,$ and 8 compared to those of models $2, 3,$ and 5, respectively, where models $6, 7,$ and 8 include one more explanatory variable (i.e., X_3) than models $2, 3,$ and 5, respectively, do. In scenario B3, the increase in both SD and RMSE owing to the addition of X_3 stands

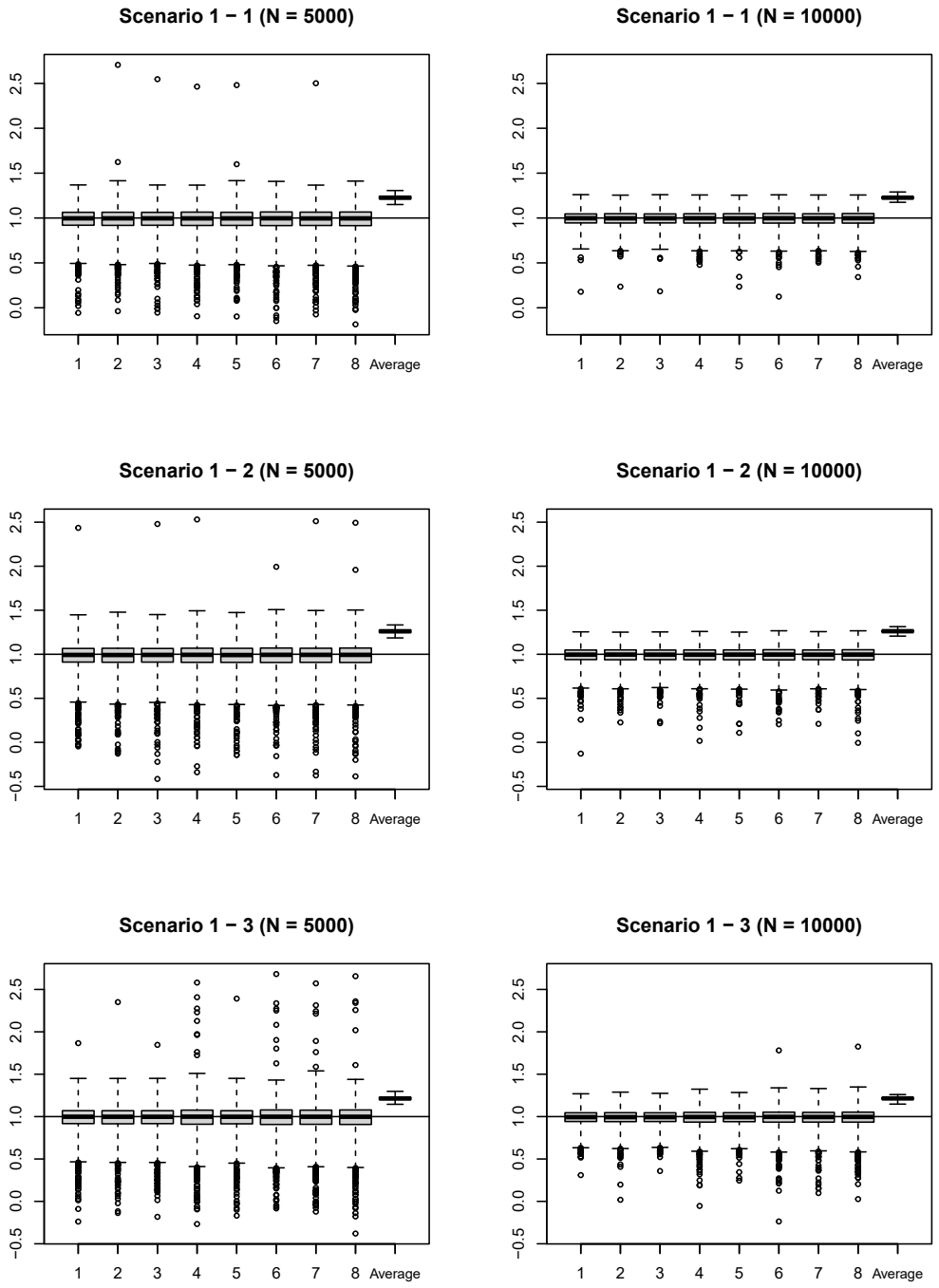

Figure 6: Boxplots of the population mean estimates for scenario A1 ($X_{IV} = X_{PS} = \varnothing$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side by side.

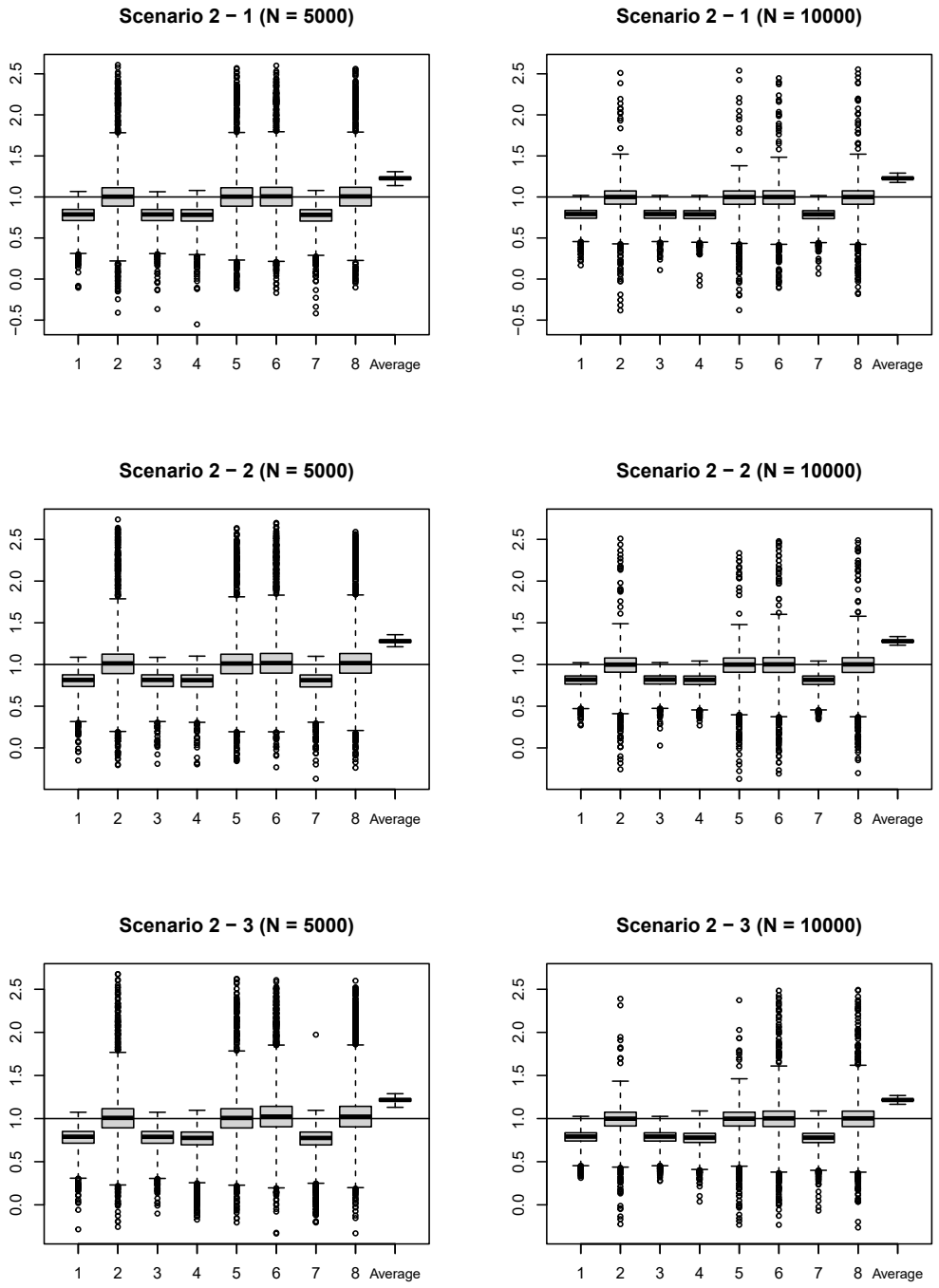

Figure 7: Boxplots of the population mean estimates for scenario A2 ($X_{IV} = X_{PS} = \{X_1\}$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side by side.

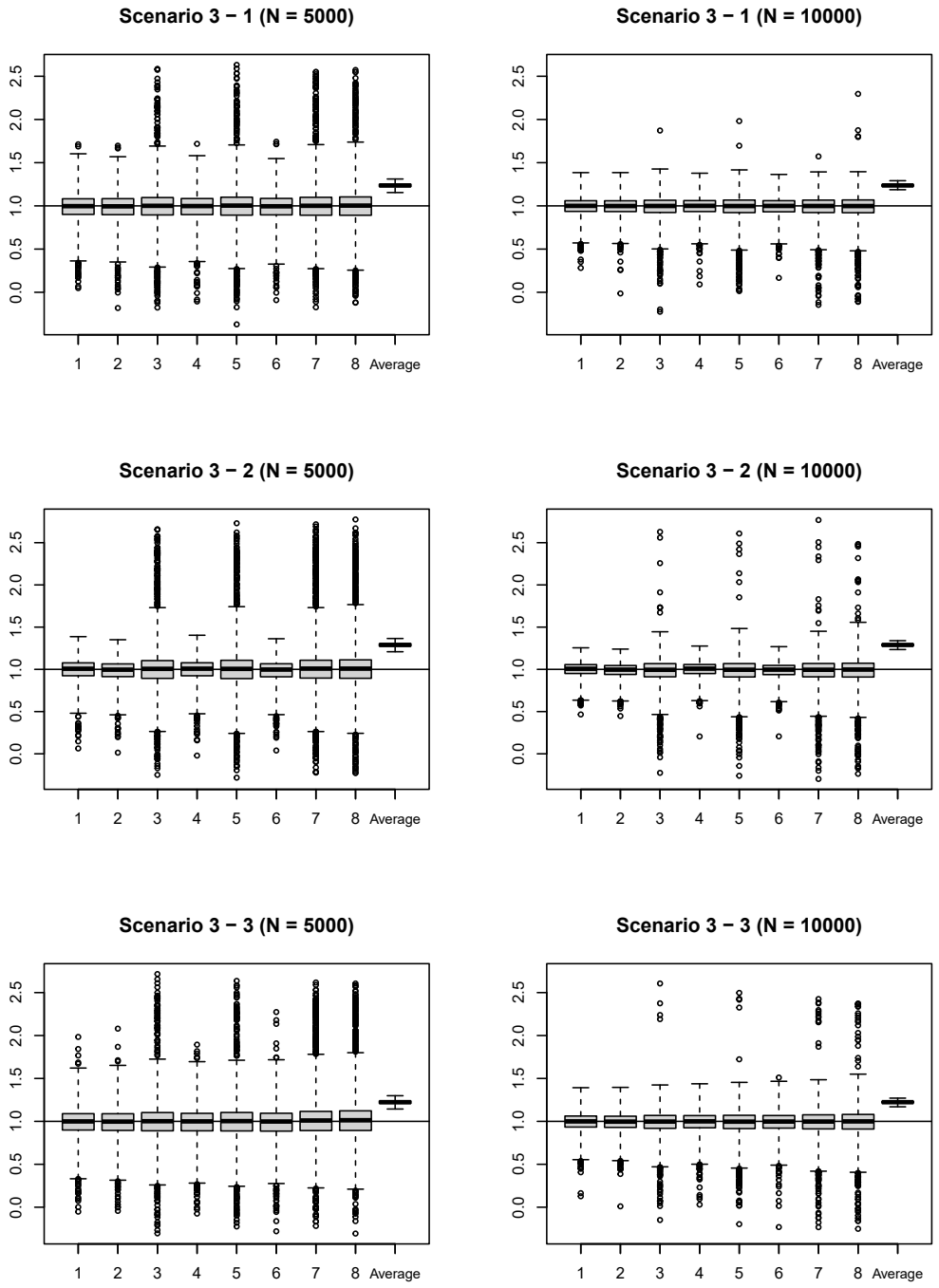

Figure 8: Boxplots of the population mean estimates for scenario A3 ($X_{IV} = \{X_2\}$, $X_{PS} = \varnothing$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side by side.

Table 1: Bias, standard deviation (SD), and root-mean-squared error (RMSE) of the population mean estimates for the simulation study using eight estimation models with a combination of covariates and simple sample averages.

Model		1	2	3	4	5	6	7	8	Average
Scenarios A1 ($X_{IV} = X_{PS} = \varnothing$)										
Scenario 1-1 (B1)										
$N = 5000$	Bias	−0.0160	−0.0168	−0.0163	−0.0174	−0.0173	−0.0185	−0.0180	−0.0191	0.2268
	SD	0.1211	0.1268	0.1228	0.1276	0.1273	0.1320	0.1286	0.1329	0.0195
	RMSE	0.1222	0.1279	0.1239	0.1288	0.1285	0.1333	0.1298	0.1342	0.2277
$N = 10000$	Bias	−0.0076	−0.0075	−0.0077	−0.0077	−0.0077	−0.0073	−0.0079	−0.0074	0.2267
	SD	0.0784	0.0803	0.0785	0.0804	0.0806	0.0830	0.0804	0.0829	0.0137
	RMSE	0.0788	0.0807	0.0789	0.0808	0.0810	0.0833	0.0808	0.0833	0.2271
Scenario 1-2 (B2)										
$N = 5000$	Bias	−0.0202	−0.0210	−0.0205	−0.0208	−0.0215	−0.0212	−0.0215	−0.0215	0.2614
	SD	0.1334	0.1354	0.1334	0.1364	0.1361	0.1396	0.1379	0.1411	0.0201
	RMSE	0.1349	0.1370	0.1350	0.1380	0.1378	0.1412	0.1396	0.1427	0.2622
$N = 10000$	Bias	−0.0108	−0.0109	−0.0109	−0.0114	−0.0111	−0.0111	−0.0115	−0.0114	0.2618
	SD	0.0873	0.0893	0.0869	0.0902	0.0895	0.0926	0.0896	0.0934	0.0141
	RMSE	0.0880	0.0900	0.0876	0.0909	0.0902	0.0932	0.0904	0.0940	0.2622
Scenario 1-3 (B3)										
$N = 5000$	Bias	−0.0138	−0.0149	−0.0140	−0.0157	−0.0156	−0.0158	−0.0168	−0.0170	0.2144
	SD	0.1300	0.1324	0.1298	0.1478	0.1339	0.1495	0.1483	0.1515	0.0196
	RMSE	0.1307	0.1332	0.1306	0.1486	0.1348	0.1503	0.1492	0.1524	0.2153
$N = 10000$	Bias	−0.0101	−0.0109	−0.0102	−0.0124	−0.0110	−0.0121	−0.0126	−0.0122	0.2143
	SD	0.0834	0.0863	0.0834	0.0958	0.0862	0.1002	0.0961	0.0996	0.0141
	RMSE	0.0840	0.0870	0.0840	0.0966	0.0869	0.1009	0.0969	0.1003	0.2148
Scenarios A2 ($X_{IV} = X_{PS} = \{X_1\}$)										
Scenario 2-1 (B1)										
$N = 5000$	Bias	−0.2293	−0.0001	−0.2306	−0.2338	0.0005	0.0050	−0.2351	0.0053	0.2275
	SD	0.1151	0.2218	0.1174	0.1206	0.2238	0.2260	0.1224	0.2272	0.0198
	RMSE	0.2566	0.2218	0.2588	0.2631	0.2238	0.2261	0.2651	0.2273	0.2284
$N = 10000$	Bias	−0.2176	−0.0155	−0.2180	−0.2209	−0.0163	−0.0143	−0.2213	−0.0142	0.2277
	SD	0.0802	0.1439	0.0803	0.0835	0.1438	0.1504	0.0833	0.1511	0.0140
	RMSE	0.2319	0.1447	0.2323	0.2361	0.1447	0.1511	0.2365	0.1517	0.2281
Scenario 2-2 (B2)										
$N = 5000$	Bias	−0.2041	0.0129	−0.2043	−0.2078	0.0101	0.0229	−0.2083	0.0219	0.2795
	SD	0.1193	0.2475	0.1204	0.1218	0.2418	0.2567	0.1242	0.2545	0.0199
	RMSE	0.2364	0.2479	0.2371	0.2408	0.2420	0.2577	0.2426	0.2554	0.2802
$N = 10000$	Bias	−0.1928	−0.0154	−0.1924	−0.1964	−0.0166	−0.0144	−0.1962	−0.0155	0.2792
	SD	0.0835	0.1538	0.0839	0.0867	0.1543	0.1652	0.0872	0.1626	0.0145
	RMSE	0.2102	0.1545	0.2099	0.2147	0.1551	0.1658	0.2147	0.1633	0.2796
Scenario 2-3 (B3)										
$N = 5000$	Bias	−0.2278	0.0033	−0.2287	−0.2427	0.0021	0.0329	−0.2436	0.0327	0.2163
	SD	0.1169	0.2199	0.1174	0.1320	0.2189	0.2544	0.1329	0.2565	0.0197
	RMSE	0.2561	0.2199	0.2570	0.2762	0.2189	0.2565	0.2775	0.2586	0.2172
$N = 10000$	Bias	−0.2181	−0.0141	−0.2184	−0.2316	−0.0145	−0.0083	−0.2323	−0.0086	0.2164
	SD	0.0808	0.1419	0.0809	0.0928	0.1434	0.1692	0.0938	0.1691	0.0136
	RMSE	0.2326	0.1426	0.2329	0.2495	0.1442	0.1694	0.2505	0.1693	0.2169
Scenarios A3 ($X_{IV} = \{X_2\}, X_{PS} = \varnothing$)										
Scenario 3-1 (B1)										
$N = 5000$	Bias	−0.0149	−0.0170	−0.0109	−0.0156	−0.0101	−0.0171	−0.0086	−0.0071	0.2366
	SD	0.1518	0.1559	0.1903	0.1556	0.1982	0.1589	0.2012	0.2074	0.0195
	RMSE	0.1525	0.1568	0.1906	0.1564	0.1984	0.1598	0.2013	0.2075	0.2374
$N = 10000$	Bias	−0.0074	−0.0095	−0.0129	−0.0077	−0.0133	−0.0094	−0.0141	−0.0135	0.2365
	SD	0.1023	0.1047	0.1218	0.1048	0.1257	0.1066	0.1259	0.1307	0.0139
	RMSE	0.1026	0.1051	0.1225	0.1051	0.1264	0.1070	0.1267	0.1314	0.2369
Scenario 3-2 (B2)										
$N = 5000$	Bias	−0.0089	−0.0208	−0.0024	−0.0087	−0.0014	−0.0207	0.0034	0.0050	0.2896
	SD	0.1257	0.1263	0.2144	0.1271	0.2173	0.1281	0.2268	0.2326	0.0199
	RMSE	0.1260	0.1280	0.2144	0.1274	0.2173	0.1298	0.2268	0.2326	0.2903
$N = 10000$	Bias	−0.0024	−0.0144	−0.0186	−0.0022	−0.0180	−0.0141	−0.0188	−0.0176	0.2893
	SD	0.0883	0.0896	0.1372	0.0904	0.1414	0.0916	0.1437	0.1487	0.0141
	RMSE	0.0883	0.0907	0.1384	0.0904	0.1425	0.0927	0.1449	0.1497	0.2897
Scenario 3-3 (B3)										
$N = 5000$	Bias	−0.0134	−0.0157	−0.0065	−0.0151	−0.0058	−0.0164	0.0075	0.0132	0.2236
	SD	0.1573	0.1600	0.2004	0.1709	0.2037	0.1781	0.2326	0.2391	0.0199
	RMSE	0.1579	0.1608	0.2005	0.1716	0.2038	0.1788	0.2328	0.2395	0.2245
$N = 10000$	Bias	−0.0081	−0.0108	−0.0130	−0.0107	−0.0141	−0.0123	−0.0152	−0.0140	0.2236
	SD	0.1078	0.1102	0.1300	0.1212	0.1340	0.1243	0.1498	0.1556	0.0138
	RMSE	0.1081	0.1107	0.1307	0.1217	0.1348	0.1249	0.1506	0.1562	0.2240

out at about 10%, whereas the increase in other cases is only a small percentage. This may be related to the fact that the association between X_3 and the missing mechanism in scenario B3 is stronger than in the other scenarios. In conclusion, it can be said that if a covariate is not related to the outcome, it should not be included in the missing model, even if it explains the missing mechanism well because it may increase the SD of the estimates for the mean of the outcome.

Continuing with the second aspect, we also considered that there might be other types of covariates whose inclusion in the extended propensity score model would increase the SD of the mean estimate for the outcome. We found them to be the covariates in X_{IV}. For scenario A2, $X_{IV} = X_{PS} = \{X_1\}$. The corresponding Figure 7 shows that models 2, 5, 6, and 8, which include X_1 in the extended propensity score model, are unbiased, but have large variations in their estimates. For scenario A3, $X_{IV} = \{X_2\}, X_{PS} = \varnothing$. The corresponding Figure 8 shows that the estimates by all models are unbiased, but models 3, 5, 7, and 8, which include X_2 in the extended propensity score model, clearly have larger variations in their estimates than other models have. These results lead us to think that the inclusion of a covariate from X_{IV} into the extended propensity score model may increase the SD of the estimates. If the variable is included in both X_{IV} and X_{PS}, then it becomes a problem of a trade-off between bias and SD. However, if a covariate is in X_{IV}, but not in X_{PS}, then we think that it should not be included in the extended propensity score model.

Finally, we discuss the third aspect: which of the extended propensity score models 1-8 is the best in terms of RMSE of the population mean estimate. Table 1 shows that among the 18 data generation patterns for the 9 scenarios, each with 2 sample sizes, 14 patterns had the smallest RMSE when only X_{PS} was included in the extended propensity score model. The four exceptions were scenarios 1-2 ($N = 10000$), 1-3 ($N = 5000$), 2-2 ($N = 5000$), and 2-3 ($N = 5000$). The difference between the model with the smallest RMSE and the one including only X_{PS} in the extended propensity score model was at most about 5 % in these four patterns. The RMSE minimum models for these four patterns included those with a bias in the estimates of the population mean. When the comparison was made only with the RMSE minimum model with small bias, that is, X_{PS} was included, the difference from the case in which only X_{PS} was included in the extended propensity score was quite small, at most about 2 %.

Based on the above, we argue that including only X_{PS} as covariates in the extended

propensity score model is a good variable selection strategy. There are mainly two reasons for this. The first is that including only X_{PS} in the extended propensity score model has theoretical validity in terms of consistency and asymptotic normality, as shown in the proposition3.1. Second, in the simulation experiments, including only X_{PS} in the extended propensity score model was not biased, and in many cases, the RMSE values were the best, and even when they were not, they were close to the best. Including only X_{PS} in the extended propensity score model is not necessarily the best strategy, but it is the next best strategy available. Therefore, we argue that the variable selection strategy of including only X_{PS} as a covariate in the extended propensity score model is a good available method to estimate the population mean of the outcome at or near the minimum RMSE without bias.

3.3 Conclusion of Chapter 3

3.3.1 Summary

In this paper, we discussed which variables should be included in the extended propensity score model to obtain an unbiased estimate with small RMSE values for the population mean using the method of Sun et al. (2018) that adjusts for the MNAR outcome. First, we defined the minimal sets of covariates, X_{IV} and X_{PS}, and showed that including the variables from X_{PS} only as covariates in the extended propensity score model yields the asymptotically normal estimation for the population mean. This holds even if X_{PS} are not equal to X_{IV}, that is, the covariates to be included in the IV mean model; this is the difference in methodology from Sun et al. (2018), who treat covariates in the IV mean model and the extended propensity score model as the same X.

Through simulation experiments, we confirmed the following three points: 1) including X_{PS} in the extended propensity score model provides unbiased estimates for the population mean; 2) if a covariate is not related to the outcome or is included in X_{IV}, but not in X_{PS}, then the covariate should not be included in the extended propensity score model because including it in the model increases the standard error without decreasing the bias; and 3) including only X_{PS} in the extended propensity score as covariates is a good covariate selection strategy because it allows estimating the population mean of the outcome with the best or almost the best RMSE. Further, when X_{IV} and X_{PS} are different, our covariate selection strategy can avoid the loss of estimation precision, whereas this is not possible with Sun et al. (2018)' s study, which uses the same X in

the IV mean model and the extended propensity score model. In addition, we were able to reduce the difficulty experienced when modeling the extended propensity score that includes variables with missing data by providing a strategy for covariate selection.

3.3.2 Discussion on the IVs

Let us now discuss the IVs. Using multiple IVs may reduce the variance in the estimates. For example, when a variable is associated only with the missing mechanism, it reduces estimation precision if it is included in the extended propensity score model, but it may improve estimation precision if it is used as an additional IV, as it satisfies the condition for the IV. However, IVs need to be carefully judged for their eligibility, as it cannot be determined from the data whether the exclusion restriction is satisfied or not (Heckman, 1997). In particular, bias might occur even if one ineligible variable is used as the IV. Thus, whether a variable is included in the model as an IV, or not, should be carefully considered. Note that identifying X_{PS} is as difficult as determining whether the exclusion constraint holds.

We argue that a useful solution to the problems in finding an appropriate IV and to identify X_{PS} is to obtain a variable Z satisfying $Y \perp\!\!\!\perp Z$ by means of a trial design and use it as an IV, as proposed by Sajons (2020) and Section 7 of Yoneyama & Minami (2023). This is because it not only gives us a highly accurate IV, but also eliminates the need for the mean model of the IV (2.3) and the identification of X_{PS}, as X_{IV} and X_{PS} are empty from $Y \perp\!\!\!\perp Z$. Therefore, as long as the necessary assumptions are met, only the IV and the outcome need to be included in the extended propensity score model to obtain a consistent and asymptotically normal estimator for the population mean of the outcome. It should also be noted that the fact that X_{IV} is empty avoids the phenomenon that the inclusion of X_{IV} in the extended propensity score model reduces the precision of the population mean estimate of the outcome. However, when conducting an analysis using existing data, it is not possible to obtain an IV that is guaranteed to satisfy the conditions. In such cases, we need to carefully examine the properties and relationships among variables to obtain IVs and identify X_{PS} to reduce bias and improve estimation precision.

3.3.3 Limitation of this study and future work

Finally, we discuss a limitation of the study and the resulting scope for future work. As seen in Figures 6-8 and Table 1, the method using the extended propensity score results

in no bias, but causes a larger SD than the method using simple sample averages does. Population mean estimates of outcomes using sample means have bias and can lead to erroneous judgments; however, adjustments using the extended propensity score, which includes X_{PS} as a covariate, has no bias and does not cause such erroneous judgments, even if the variance is large. However, the method using the extended propensity score has low power owing to its large variance. Therefore, in future, we would like to consider an estimation method of the population mean of the outcome that has smaller variance without bias; we would also like to include a theoretical discussion on why including covariates related only to the missing mechanism and X_{IV} covariates in the extended propensity score model reduces the precision of the estimation.

Chapter 4 Treatment effects estimation with MNAR data

Up to Chapter 3, we were interested in the estimation of the population mean when there are missing data in the outcome. In practice, on the other hand, we are often interested in estimating other than the population mean. One of the major interests, other than the population mean, is the estimation of the treatment effects, which is the difference of the outcomes with and without the treatment. In observational studies, where the investigator cannot experimentally assign, estimating the mean by assignment group and taking the difference does not necessarily indicate a difference due to assignment. This is because there are confounders that affect both the assignment and the outcome.

Let us consider an example, suppose we are interested in whether a regular exercise routine has a positive effect on cognitive ability in the elderly, and questionnaires and tests to measure cognitive ability have been administered to randomly selected elderly people. Suppose we do not intervene in the exercise routine, but ask about it in the questionnaires. In this case, even if we compare cognitive performance between people with and without regular exercise routine, we cannot say that this difference is an effect of regular exercise routine. This is because preference for regular exercise routine is also associated with other covariates that affect cognitive ability, such as obesity and hypertension (Barnes & Yaffe, 2011). Therefore, it is natural to assume that confounding occurs, and a simple comparison would also include the effects of covariates other than regular exercise routine. At the same time, cognitive testing in the older aged people may result in missing data about cognitive function because they may drop out during testing due to fatigue. In this chapter, we consider how to simultaneously adjust for the confounder and the missing data in response to make appropriate inferences in cases such as the above.

For confounder adjustment, an orthodox procedure is to specify a model that expresses outcomes in terms of treatments and covariates, find the estimates for all parameters, and express the treatment effects with the estimates. However, in many cases, identifying the exact relationship between the variables is not an easy task. To overcome

this difficulty, the propensity score for the assignment was introduced by Rosenbaum & Rubin (1983), and methods for estimating treatment effects without specifying an outcome model have been proposed, including matching (Rosenbaum & Rubin, 1985), subclassification (Rosenbaum & Rubin, 1984) and weighting (Horvitz & Thompson, 1952; Rosenbaum, 1987; Hirano et al., 2003). The propensity score for the assignment is the probability of receiving an assignment when conditioning on covariates. As for assignment, the methods using the propensity score still have the advantage of avoiding direct modeling of the outcome and of being resistant to misidentification of the propensity score model.

Here, we consider the estimation of the treatment effects when the outcome has MNAR data. However, the MNAR adjustment methods described in the previous chapters were to estimate the population mean, and cannot be used directly to estimate treatment effects in data with MNAR. Li & Zhou (2017) extended the method of the Miao & Tchetgen Tchetgen (2016)'s shadow variable to causal inference and proposed a method to estimate the mediation effect and the average treatment effect when the outcome is MNAR. However, their method requires the identification of the mean structure of the outcome.

Yoneyama & Minami (2023) proposed a method to estimate treatment effects without specifying the model of the outcome using IVs when the outcome is MNAR. The proposing method is an extension of Sun et al. (2018) for causal inferences. We introduce the concept of potential missingness in the situation of MNAR and an extended dual propensity score, which is the conditional probability that the assignment is a given status, and that the potential outcome is observed given the values of covariates, the potential outcome for the given status, and IVs. Then, we propose estimators defined with the extended dual propensity score for the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the untreated (ATU). We show that these estimators are consistent and asymptotically normal under regularity conditions.

The remainder of this chapter is presented as follows. We begin by outlining the basic ideas and methods of causal inference in Section 4.1. In Section 4.2, we describe the notation and assumptions about the relationships between variables. In Section 4.3, we introduce an extended dual propensity score. We propose an estimator of the extended dual propensity score using the IV in Section 4.4.1. In Section 4.4.2, we propose inverse

probability weighted (IPW) estimators for ATE, ATT, and ATU with estimators of extended dual propensity scores. We show that these IPW estimators are consistent and asymptotically normal when the outcome is MNAR. In Section 4.5, we report the simulation experiments conducted to investigate the performance of the proposed method and to compare it with other commonly used methods for the adjustment of missing data. In Section 4.6, we further compare the proposed method with other methods for adjusting for missing data using real data. In Section 4.7, we conclude with a discussion and present an example of a data collection strategy to obtain the IV.

4.1 Basic method of causal inference

Here we first outline the basic method of causal inference, especially that based on Rubin causal model (Holland, 1986).

4.1.1 Rubin causal model

Let W be the assignment indicator, where W is binary such that $W = 1$ if the individual is treated and $W = 0$ otherwise. Rubin causal model assumes that for each individual there is a potential outcome for either receiving or not receiving the treatment, but that only one case is actually observed and the other is missing due to assignment. Let Y^1 be the potential outcome if the treatment is received and Y^0 be the potential outcome if the treatment is not received. Then, the outcome Y can then be expressed as $Y = WY^1 + (1 - W)Y^0$. The following three types of treatment effects are:

Definition 4.1. *The average treatment effect (ATE) ϕ_{ATE}, the average treatment effect on the treated (ATT) ϕ_{ATT}, and the average treatment effect on the untreated (ATU) ϕ_{ATU} are defined respectively as:*

$$\phi_{ATE} = \mathbb{E}[Y^1 - Y^0], \phi_{ATT} = \mathbb{E}[Y^1 - Y^0|W = 1], \text{ and } \phi_{ATU} = \mathbb{E}[Y^1 - Y^0|W = 0].$$

The ATE represents an average over the population for the difference that occurred with and without the treatment. The ATT and ATU represent averages over the subpopulations of treated and untreated people for the difference, respectively. The ATT is an important measure of whether an intervention is effective for the recipients of the intervention. The ATU is an indicator of the effect of an intervention on people who would not normally receive such an intervention and plays an important role when considering whether to implement a more aggressive intervention.

4.1.2 Estimation of treatment effects

In estimating these treatment effects, the following formula is often assumed to hold (e.g. Rosenbaum & Rubin, 1983; Hirano et al., 2003):

$$\left(Y^1, Y^0\right) \perp\!\!\!\perp W|X, \tag{4.1}$$

where X is a vector of covariates. This assumption is called unconfoundedness and indicates that there are no unobserved confounder. The Rubin causality model is based on regarding assignment as a missing potential outcome. Since the assumption on the potential outcome in unconfoundedness (4.1) is parallel to the expression of MAR (2.1) when the outcome has MAR data, treatment effects can be estimated by using the conditional assignment probability as in the case of MAR data for the outcome. Let $\pi_{W=1}(x)$ be the propensity score for the assignment, i.e.,

$$\pi_{W=1}(x) = \Pr(W = 1|X = x).$$

Assume that $0 < \pi_{W=1}(x) < 1$ for all $x \in \mathrm{Supp}(X)$. Under appropriate regularity conditions, the following IPW estimator $\hat{\Phi}_{ATE}$ for the ATE ϕ_{ATE} is consistent (Hirano et al., 2003),

$$\hat{\Phi}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i}{\hat{\pi}_{W=1}(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i}{1-\hat{\pi}_{W=1}(X_i)},$$

where $\hat{\pi}_{W=1}(x)$ is the consistent estimator of the propensity score for assignment. The first and second terms on the right-hand side correspond to the estimators of $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, respectively. Similarly, Hirano et al. (2003) showed that weighted estimators can be constructed for ATT and ATU estimation, and these are consistent under the regularity conditions. Hirano et al. (2003) defined $\hat{\Phi}_{ATT}$, the estimator for the ATT ϕ_{ATT} and $\hat{\Phi}_{ATU}$, the estimator for the ATU ϕ_{ATU} as follows:

$$\hat{\Phi}_{ATT} = \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N}\hat{\pi}_{W=1}(X_i)} - \frac{\sum_{i=1}^{N}\dfrac{\hat{\pi}_{W=1}(X_i)(1-W_i)Y_i}{1-\hat{\pi}_{W=1}(X_i)}}{\sum_{i=1}^{N}\hat{\pi}_{W=1}(X_i)},$$

$$\hat{\Phi}_{ATU} = \frac{\sum_{i=1}^{N}\dfrac{\left(1-\hat{\pi}_{W=1}(X_i)\right)W_i Y_i}{\hat{\pi}_{W=1}(X_i)}}{\sum_{i=1}^{N}\left(1-\hat{\pi}_{W=1}(X_i)\right)} - \frac{\sum_{i=1}^{N}(1-W_i)Y_i}{\sum_{i=1}^{N}(1-\hat{\pi}_{W=1}(X_i))}.$$

4.2 Problem setting for estimating treatment effects with MNAR data

From here, we consider estimating treatment effects when an outcome is missing not at random (MNAR), but all other variables are observed. Let Y be the outcome, X be a vector of covariates, and R be the missing indicator for Y, where R is binary such that $R = 1$ if Y is observed and $R = 0$ otherwise. Let W be the assignment indicator, where W is binary such that $W = 1$ if the individual is treated and $W = 0$ otherwise.

To formulate the effect of the treatment, we introduce the potential outcomes Y^1 and Y^0 where Y^w represents the potential outcome under the assignment status $w \in \{1, 0\}$. The outcome Y can be expressed as

$$Y = WY^1 + (1 - W)Y^0.$$

Further, taking into account missing of the outcome, we define Y^* as the actual observed outcome,

$$Y^* = \begin{cases} Y^1 & \text{if} \quad W = 1 \text{ and } R = 1 \\ Y^0 & \text{if} \quad W = 0 \text{ and } R = 1 \\ \text{missing} & \text{if} \quad R = 0. \end{cases} \tag{4.2}$$

The table below summarizes the observed status of the potential outcomes according to the assignment and missing indicator.

Table 2: Observed status of potential outcome

	$W = 1$		$W = 0$	
	$R = 1$	$R = 0$	$R = 1$	$R = 0$
Y^1: Potential outcome if treated	observed	missing	missing	missing
Y^0: Potential outcome if not treated	missing	missing	observed	missing

We consider estimating the ATE ϕ_{ATE}, the ATT ϕ_{ATT}, and the ATU ϕ_{ATU} defined in Definition 4.1. In the following discussion, we further introduce the potential variables for the missing indicator R. Let R^1 be the potential missing indicator when treated and let R^0 be the potential missing indicator when not treated. We can represent the relationship of R, R^1 and R^0 as follows:

$$R = WR^1 + (1 - W)R^0. \tag{4.3}$$

Note that Zhao & Ding (2022) also introduced the concept of potential missing; however, they discussed missing covariates, whereas we examine them for the outcome, which is a different setting.

The MNAR can be interpreted as the occurrence of the potential missingness in a counterfactual world depending on the potential outcome, conditioning the observed variables, and the assignment determines which counterfactual world will be observed ($Y^w \not\perp\!\!\!\perp R^w | X, W$).

To find consistent estimators for treatment effects under MNAR outcomes, we adopt an IV instead of specifying the model for the outcome. We define Z as a p-dimensional ($p \geq 1$) random vector of IVs that satisfies the following Assumptions 4.1 and 4.2. Note that this IV is used to adjust for MNAR in the outcome, and its assumptions are different from the customary one aimed at adjusting for unobserved confounding.

Assumption 4.1 (exclusion restriction (Yoneyama & Minami, 2023))**.**

$$\left(Y^1, Y^0\right) \perp\!\!\!\perp Z | X$$

Assumption 4.2 (IV relevance (Yoneyama & Minami, 2023))**.** *Let us denote* $Z_{-k} = (Z_1 \ \ldots \ Z_{k-1} \, Z_{k+1} \ \ldots \ Z_p)^T$ *($k \in \{1, \ldots, p\}$ and z_{-k} likewise). Then, for all x, z_{-k}, z_k, and z'_k ($z_k \neq z'_k$),*

$$\Pr(R^w = 1 | X = x, Z_{-k} = z_{-k}, Z_k = z_k) \neq \Pr(R^w = 1 | X = x, Z_{-k} = z_{-k}, Z_k = z'_k)$$

holds for all $k \in \{1, \ldots, p\}$ and $w \in \{1, 0\}$.

Assumption 4.1 (exclusion restriction) means that the IVs and the potential outcomes are conditionally independent given X, i.e., the IVs do not have a direct effect on the potential outcomes, and indirect effects exist only through covariates X. Assumption 4.2 states that each element of the IVs remains relevant to the missing mechanism when other elements of IV and X are fixed. Including the IV, MNAR can be represented as follows:

$$Y^w \not\perp\!\!\!\perp R^w | X, Z, W \tag{4.4}$$

In addition, we make the following assumptions on assignments and missing data.

Assumption 4.3 (unconfoundedness (Yoneyama & Minami, 2023))**.**

$$(Y^1, Y^0) \perp\!\!\!\perp W | X, Z$$

Assumption 4.4. *The following conditional independence holds for any of the assignments* $w \in \{1, 0\}$.

$$R^w \perp\!\!\!\perp W | X, Y^w, Z$$

Assumption 4.3 is the common assumption in causal inference, which means that all variables that affect both the outcome and the assignment are included in the covariates or IVs. Assumption 4.4 states that in each counterfactual world, whether the potential outcome is observed or not, and the assignment are conditionally independent given covariates, a potential outcome, and IVs. From MNAR (4.4) and Assumption 4.1 – 4.4, the relationship between the variables can be expressed as a single world intervention graph (c.f. Richardson (2013)) shown in Figure 9.

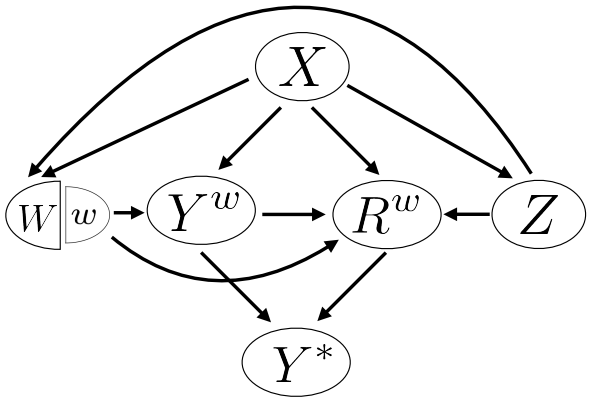

Figure 9: Single world intervention graph illustrating the relationship of variables.

We assume that N observation vectors $O_i = (Y_i^*, X_i, W_i, R_i, Z_i)$ $(i = 1, \ldots, N)$ are independent and identically distributed under the above assumptions.

4.3 Extended dual propensity score

In this section, we propose an extended dual propensity score to adjust for confounders and missing data simultaneously. We also show that the average treatment effect can be estimated consistently when we can obtain the consistent estimator of this extended dual propensity score.

We consider estimating the treatment effects with MNAR outcome. We first define an extended dual propensity score as follows:

Definition 4.2 (extended dual propensity score)**.** *For assignment $w \in \{1, 0\}$, let us define an extended dual propensity score $\pi_{W=w,R^w=1}(x, y^w, z)$ as the conditional probability that the assignment is w and the outcome is observed (i.e., $R^w = 1$) given the values of covariates, the potential outcome for assignment w and IVs, that is,*

$$\pi_{W=w,R^w=1}(x, y^w, z) = \Pr(W = w, R^w = 1 | X = x, Y^w = y^w, Z = z).$$

When a consistent estimator is obtained for this extended dual propensity score, the average treatment effect can be estimated consistently using the inverse probability weighting approach, as follows. The specific method of obtaining this consistent estimator will be discussed later.

Proposition 4.1. *Assume that we have consistent estimators $\hat{\pi}_{W=1,R^1=1}(x, y^1, z)$ and $\hat{\pi}_{W=0,R^0=1}(x, y^0, z)$ for the extended dual propensity scores for assignments 1 and 0, respectively. Then, the inverse probability weighted (IPW) estimator $\hat{\varphi}_{ATE}$ defined by*

$$\hat{\varphi}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} A_i^1 Y_i^* - \frac{1}{N} \sum_{i=1}^{N} A_i^0 Y_i^* \tag{4.5}$$

is a consistent estimator for $\phi_{ATE} = \mathbb{E}[Y^1 - Y^0]$ under conditions 1) and 3) of Lemma A.1 in Appendix A, where

$$A_i^w = \frac{\mathbf{1}_{\{W_i=w\}} R_i}{\hat{\pi}_{W=w,R^w=1}(X_i, Y_i^*, Z_i)},$$

and $\mathbf{1}_{\{W=w\}}$ is an indicator function,

$$\mathbf{1}_{\{W=w\}} = \begin{cases} 1 & \text{if} \quad W = w \\ 0 & \text{if} \quad W \neq w. \end{cases}$$

Proof: First, we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}_{\{W_i=w\}}R_iY_i^*}{\pi_{W=w,R^w=1}(X_i,Y_i^*,Z_i)}\\
&\quad\xrightarrow{p}\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}RY^*}{\pi_{W=w,R^w=1}(X,Y^*,Z)}\right]\quad\text{(by the law of large numbers)}\\
&\quad=\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R^wY^w}{\pi_{W=w,R^w=1}(X,Y^w,Z)}\right]\quad\text{(by (4.2) and (4.3))}\\
&\quad=\mathbb{E}\left[\frac{Y^w\mathbb{E}\left[\mathbf{1}_{\{W=w\}}R^w|X,Y^w,Z\right]}{\pi_{W=w,R^w=1}(X,Y^w,Z)}\right]\quad\text{(by the law of iterated expectations)}\\
&\quad=\mathbb{E}[Y^w],\quad\text{(since }\mathbb{E}[\mathbf{1}_{\{W=w\}}R^w|X,Y^w,Z]=\pi_{W=w,R^w=1}(X,Y^w,Z)).
\end{aligned}$$

Suppose $\pi_{W=w,R^w=1}(X,Y^w,Z)$ can be represented as $\pi_{W=w,R^w=1}(X,Y^w,Z;\theta^w)$ with a parameter vector $\theta^w\,(w\in\{1,0\})$, and $\hat{\pi}_{W=w,R^w=1}(X,Y^w,Z)$ can be represented as $\pi_{W=w,R^w=1}(X,Y^w,Z;\hat{\theta}^w)$ with an estimator $\hat{\theta}$ for θ^w. Assume conditions 1) and 3) of Lemma A.1 hold, that is,

$$\sup_{\theta\in\mathcal{N}}\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}_{\{W_i=w\}}R_i^wY_i^*}{\pi_{W=w,R^w=1}(X,Y^w,Z;\theta^w)}-\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R^wY^*}{\pi_{W=w,R^w=1}(X,Y^w,Z;\theta^w)}\right]\right\|\xrightarrow{p}0,$$

where $\mathcal{N}$ is an open neighborhood of the tule value θ_0 of θ, and $\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R^wY^*}{\pi_{W=w,R^w=1}(X,Y^w,Z;\theta^w)}\right]$ is continuous at θ_0^w. By Lemma A.1, if the estimator of the extended dual propensity score is consistent, i.e., $\hat{\theta}^w\xrightarrow{p}\theta_0^w$, then,

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{1}_{\{W_i=w\}}R_iY_i^*}{\hat{\pi}_{W=w,R^w=1}(X_i,Y_i^*,Z_i)}=\frac{1}{N}\sum_{i=1}^{N}A_i^wY_i^*\xrightarrow{p}\mathbb{E}[Y^w].\tag{4.6}$$

Therfore,

$$\hat{\varphi}_{ATE}=\frac{1}{N}\sum_{i=1}^{N}A_i^1Y_i^*-\frac{1}{N}\sum_{i=1}^{N}A_i^0Y_i^*\xrightarrow{p}\mathbb{E}[Y^1]-\mathbb{E}[Y^0]=\phi_{ATE}.$$

Hence, $\hat{\varphi}_{ATE}$ is a consistent estimator for the ATE. □

The first term on the right-hand side of (4.5) is an estimator of $\mathbb{E}[Y^1]$, and the second term is an estimator of $\mathbb{E}[Y^0]$. Note that this proposition does not require Assumptions

1 – 4.

One might wonder whether a consistent estimator for the extended dual propensity score can be obtained even when Y^* has MNAR missing; however, it is possible. We propose a method to obtain them by using IVs.

4.4 Estimation of the extended dual propensity score and treatment effects

The extended dual propensity score includes the outcome, which is MNAR; hence, it cannot be estimated properly by naive regression methods. In response to this, we propose a consistent estimator for the extended dual propensity score using the IVs in Section 4.4.1.

In Section 4.4.2, we show that the IPW estimator for the ATE using the extended dual propensity score estimates by the method in Section 4.4.1 is not only consistent but also asymptotically normal. We also propose estimators for the ATT and the ATU that are consistent and asymptotically normal.

4.4.1 Modeling and estimation of extended dual propensity score

To express the extended dual propensity score in a simple form, we define a propensity score for the assignment and an extended propensity score for missingness.

Definition 4.3 (propensity score for assignment)**.** *We define a propensity score $\pi_{W=w}(x, z)$ for assignment $w \in \{1, 0\}$ as a conditional probability that the assignment is w given covariates x and IVs z, that is,*

$$\pi_{W=w}(x, z) = \Pr(W = w | X = x, Z = z).$$

Note here that $\pi_{W=0}(x, z) = 1 - \pi_{W=1}(x, z)$, but we employ the above notation for simplicity.

Intuitively, we can assume that missing proportions of outcome and relationship between missingness and covariates are different between the treatment and control groups. Thus, we define an extended propensity score for missingness for each assignment as follows:

Definition 4.4 (extended propensity score for missingness)**.** *We define an extended propensity score $\pi_{R^w=1}(x, y^w, z)$ for missingness for assignment $w \in \{1, 0\}$ as the conditional probability that the outcome is observed in a counterfactual world with assignment*

w given covariate x, potential outcome y^w, and IVs z,

$$\pi_{R^w=1}(x, y^w, z) = \Pr(R^w = 1 | X = x, Y^w = y^w, Z = z).$$

Under Assumptions 4.3 and 4.4, the extended dual propensity score can be expressed as the product of the propensity score for assignment and the extended propensity score for missingness as follows.

$$\begin{aligned}
\pi_{W=w,R^w=1}(X, Y^w, Z) &= \Pr(W = w, R^w = 1 | X, Y^w, Z) \\
&= \Pr(W = w | X, Y^w, Z) \Pr(R^w = 1 | X, Y^w, Z) \\
&= \Pr(W = w | X, Z) \Pr(R^w = 1 | X, Y^w, Z) \\
&= \pi_{W=w}(X, Z) \pi_{R^w=1}(X, Y^w, Z)
\end{aligned}$$

This expression implies that we can consistently estimate the extended dual propensity score if we have consistent estimators for the propensity score for assignment and the extended propensity score for missingness.

Let γ be the parameter vector in a model for the propensity score for assignment $w = 1$ and denote the propensity score for assignment $w = 1$ as $\pi_{W=1}(X, Z; \gamma)$. The propensity score for assignment $w = 0$ is then expressed as $\pi_{W=0}(X, Z; \gamma) = 1 - \pi_{W=1}(X, Z; \gamma)$. Here, we assume that a consistent m-estimator (for example, Newey & McFadden (1994); van der Vaart (1998); Tsiatis (2006)) $\hat{\gamma}$ such as the maximum likelihood estimator is available for γ. This is a reasonable assumption because we assume MNAR for only the outcome and no missing data in variables related to the assignment W, covariates X, and IVs Z.

For the extended propensity score for missingness, we assume, for ease of discussion, the following logistic regression model with parameter vectors θ^w and ν^w:

$$\begin{aligned}
\pi_{R^w=1}(X, Y^w, Z; \theta^w, \nu^w) &= \Pr(R^w = 1 | X, Y^w, Z) \\
&= \frac{1}{1 + \exp\left\{-\theta^{wT} h^w(X, Z) - \nu^{wT} g^w(X, Y^w)\right\}} \qquad (4.7)
\end{aligned}$$

where $h^w(X, Z)$ is the vector of functions of X and Z, and $g^w(X, Y^w)$ is that of X, Y^w. To uniquely decompose the linear predictor into h^w and g^w, we assume that $g^w(X, Y^w)$ consists only of elements associated with Y^w. Note that in the linear predictor of the

logistic regression model (4.7), no term depends on both the IV and the outcome, and this model is identifiable. Proof of this identifiablity is given in the Appendix B.2.

Next, we consider the estimation method of parameters θ^w and ν^w in model (4.7) for the extended propensity score. For the estimation of θ^w, we focus on the fact that the following equation holds:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}-1\Big|X,Y^w,Z\right]\\
&=\frac{\mathbb{E}[\mathbf{1}_{\{W=w\}}\mid X,Y^w,Z]\mathbb{E}[R^w\mid X,Y^w,Z]}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}-1\\
&=\frac{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}-1\\
&=0,
\end{aligned}\tag{4.8}$$

where γ_0, θ_0^w and ν_0^w are true values of γ, θ^w and ν^w, respectively. The first equality holds by Assumption 4.4 and the second equality holds by Assumption 4.3. Using the equation (4.8), we construct the following estimating equation U_1^w:

$$U_1^w(O;\gamma,\theta^w,\nu^w)=\left\{\frac{\mathbf{1}_{\{W=w\}}R}{\pi_{W=w}(X,Z;\gamma)\pi_{R^w=1}(X,Y^*,Z;\theta^w,\nu^w)}-1\right\}\tilde{h}(X,Z)\tag{4.9}$$

where $\tilde{h}(X,Z)$ is a vector-valued function for X, Z with the same size as θ^w, for example $h^w(X,Z)$. Note that we assume $\tilde{h}(X,Z)$ is a function such that U_1^w satisfies conditions 1) and 2) of the Lemma A.2 and 2) – 4) of Lemma A.3 in the one-dimensional IV case, or conditions 1), 3) and 4) of the Lemma A.4 and 2) – 6) of Lemma A.5 in the case of multidimensional IVs. Because $\mathbf{1}_{\{W=w\}}R$ is 0 when Y^* is missing, $U_1^w(O;\gamma,\theta^w,\nu^w)$ can always be evaluated using the observed values. The proof of $\mathbb{E}[U_1^w(O;\gamma_0,\theta_0^w,\nu_0^w)]=\mathbf{0}$ is given in appendix B.3.

For the estimation of ν^w, the same procedure as that of θ^w cannot be adopted because even if $(\mathbf{1}_{\{W=w\}}R/\{\pi_{W=w}(X,Z)\pi_{R^w=1}(X,Y^w,Z)\}-1)$ is multiplied by the function of Y^w, calculating it from the observed values may not be possible due to the MNAR missing in Y^w. Here, we utilize the IV to estimate ν^w. The IVs are variables that are conditionally independent of the potential outcomes given covariates and related to the missing data, as explained in Assumptions 4.1 and 4.2. These assumptions are

modifications of Sun et al. (2018)'s IV assumptions for causal inference. Unlike general IVs (for example, Angrist & Krueger (1999)), these assumptions are not representative of the orthogonality for IVs and unobserved confounding. We discuss in the following how to estimate ν^w using the IV.

Let us consider the modeling of the expectation of Z given X with parameter ξ, that is, $\mathbb{E}_\xi[Z|X] = f(X;\xi)$, where $f(x;\xi)$ is a function of the mean structure of Z given $X = x$, parameterized by ξ. Because no data is assumed to be missing, in both X and Z, we assume that we have an m-estimator $\hat{\xi}$ for ξ. For example, if Z is a continuous variable, the maximum likelihood estimator for a linear regression model explaining Z by X is applicable. Because the IVs Z and the potential outcome Y^w are conditionally independent of the covariates X by the exclusion restriction (Assumption 4.1), it holds for the true value ξ_0 of ξ that

$$\begin{aligned}
&\mathbb{E}\Big[(Z - f(X;\xi_0)) \otimes \tilde{g}(X, Y^w)\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\left[(Z - f(X;\xi_0)) \otimes \tilde{g}(X, Y^w)\Big|X\right]\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\Big[Z - f(X;\xi_0)\Big|X\Big] \otimes \mathbb{E}\Big[\tilde{g}(X, Y^w)\Big|X\Big]\Big] \\
&\quad = \mathbb{E}\Big[\Big(\mathbb{E}[Z|X] - f(X;\xi_0)\Big) \otimes \mathbb{E}\Big[\tilde{g}(X, Y^w)\Big|X\Big]\Big] \\
&\quad = \mathbb{E}\Big[\Big(f(X;\xi_0) - f(X;\xi_0)\Big) \otimes \mathbb{E}\Big[\tilde{g}(X, Y^w)\Big|X\Big]\Big] \\
&\quad = \mathbf{0}, \qquad (4.10)
\end{aligned}$$

where $\otimes$ is the Kronecker product and $\tilde{g}(X, Y^w)$ is a vector-valued function for X, Y^w of the same dimension as ν^w. Here, using this orthogonality, we consider the following estimating function $U_2^w(O;\xi,\gamma,\theta^w,\nu^w)$,

$$\begin{aligned}
&U_2^w(O;\xi,\gamma,\theta^w,\nu^w) \\
&\quad = \frac{\mathbf{1}_{\{W=w\}}R}{\pi_{W=w}(X,Z;\gamma)\pi_{R^w=1}(X,Y^*,Z;\theta^w,\nu^w)}\,(Z - f(X;\xi)) \otimes \tilde{g}(X, Y^*). \qquad (4.11)
\end{aligned}$$

This U_2^w can be calculated regardless of the missing Y^w, since the formula is multiplied by $\mathbf{1}_{\{W=w\}}R$. Note that we also assume $\tilde{g}(X, Y^w)$ is a function such that U_2^w satisfies

conditions 1) and 2) of the Lemma A.2 and 2) – 4) of Lemma A.3 in the one-dimensional IV case, or conditions 1), 3) and 4) of the Lemma A.4 and 2) – 6) of Lemma A.5 in the multidimensional IVs case.

The proof of $\mathbb{E}[U_2^w(O;\xi_0,\gamma_0,\theta_0^w,\nu_0^w)] = \mathbf{0}$ is given in supplemental appendix B.3. Note that the size of U_2^w is the product of the number of IVs and the size of the parameter vector ν^w. In the case of multidimensional IVs, the dimension of equation (4.11) is higher than the parameter ν^w, but this is not a problem. We describe the specific estimation method later.

We let

$$U^w(O;\xi,\gamma,\theta^w,\nu^w) = \begin{pmatrix} U_1^w(O;\gamma,\theta^w,\nu^w) \\ U_2^w(O;\xi,\gamma,\theta^w,\nu^w) \end{pmatrix}. \tag{4.12}$$

For the one-dimensional IV case, the size of $U^w(O;\xi,\gamma,\theta^w,\nu^w)$ is equal to the number of elements in the parameter vectors θ^w and ν^w. Thus, we define the solutions $\hat{\theta}^w$ and $\hat{\nu}^w$ for θ^w and ν^w of the estimating equations under given m-estimators $\hat{\xi}$ and $\hat{\gamma}$ for ξ and γ:

$$\frac{1}{N}\sum_{i=1}^{N} U^w(O_i;\theta^w,\nu^w|\xi=\hat{\xi},\gamma=\hat{\gamma}) = \mathbf{0} \tag{4.13}$$

as our estimators for θ^w and ν^w, respectively. These $\hat{\theta}^w$ and $\hat{\nu}^w$ are m-estimators, they are consistent under the conditions of Lemma A.2 and A.3.

In the case of the multidimensional IVs, the size of the moment condition (4.12) is larger than that of parameters in $U^w(O;\theta^w,\nu^w|\xi=\hat{\xi},\gamma=\hat{\gamma})$. For such a case, we can employ the generalized method of moments (GMM, Newey & McFadden (1994)) to find consistent estimators for θ^w and ν^w under the conditions of Lemma A.4 and A.5.

We can obtain consistent estimators of the extended dual propensity score for assignment $w \in \{1,0\}$ by plugging the estimators obtained above into the parameters as follows.

$$\begin{aligned}\hat{\pi}_{W=w,R^w=1}(X,Y^w,Z) &= \pi_{W=w,R^w=1}(X,Y^w,Z;\hat{\gamma},\hat{\theta}^w,\hat{\nu}^w) \\ &= \pi_{W=w}(X,Z;\hat{\gamma})\pi_{R^w=1}(X,Y^w,Z;\hat{\theta}^w,\hat{\nu}^w).\end{aligned}$$

4.4.2 Estimation of treatment effects

We now consider the IPW estimator $\hat{\phi}_{ATE}$ for the ATE with specific extended dual propensity score estimates in Section 4.4.1, that is,

$$\hat{\phi}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} \bar{A}_i^1 Y_i^* - \frac{1}{N}\sum_{i=1}^{N} \bar{A}_i^0 Y_i^* \tag{4.14}$$

where $\bar{A}_i^w$ is

$$\bar{A}_i^w = \frac{\mathbf{1}_{\{W_i=w\}} R_i}{\pi_{W=w}(X_i, Z_i; \hat{\gamma})\pi_{R^w=1}(X_i, Y_i^*, Z_i; \hat{\theta}^w, \hat{\nu}^w)}, \tag{4.15}$$

$\hat{\gamma}$ is an m-estimator of γ, the parameter of the propensity score for assignment such as logistic regression coefficients, and $\hat{\theta}^1, \hat{\theta}^0, \hat{\nu}^1$ and $\hat{\nu}^0$ are consistent estimators of $\theta^1, \theta^0, \nu^1$ and ν^0 discussed in Section 4.4.1.

Note that $\sum_{i=1}^{N} \bar{A}_i^w = N$ when the linear predictor of the missing model contains a constant term since

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \bar{A}_i^w - 1 \right\} \times 1 = \mathbf{0}$$

holds as part of the estimating equation (4.13). Although it is often preferable to divide by the sum of the weights rather than N when using the IPW estimator (c.f. Hernán & Robins (2020)), dividing by N and by the sum of the weights are usually equal in our setting.

We also propose the following IPW estimators for the ATT $\phi_{ATT} = \mathbb{E}[Y^1 - Y^0 | W = 1]$ and the ATU $\phi_{ATU} = \mathbb{E}[Y^1 - Y^0 | W = 0]$:

$$\hat{\phi}_{ATT} = \frac{\sum_{i=1}^{N} \pi_{W=1}(X_i, Z_i; \hat{\gamma})\bar{A}_i^1 Y_i^*}{\sum_{i=1}^{N} \pi_{W=1}(X_i, Z_i; \hat{\gamma})} - \frac{\sum_{i=1}^{N} \pi_{W=1}(X_i, Z_i; \hat{\gamma})\bar{A}_i^0 Y_i^*}{\sum_{i=1}^{N} \pi_{W=1}(X_i, Z_i; \hat{\gamma})}, \tag{4.16}$$

$$\hat{\phi}_{ATU} = \frac{\sum_{i=1}^{N} \pi_{W=0}(X_i, Z_i; \hat{\gamma})\bar{A}_i^1 Y_i^*}{\sum_{i=1}^{N} \pi_{W=0}(X_i, Z_i; \hat{\gamma})} - \frac{\sum_{i=1}^{N} \pi_{W=0}(X_i, Z_i; \hat{\gamma})\bar{A}_i^0 Y_i^*}{\sum_{i=1}^{N} \pi_{W=0}(X_i, Z_i; \hat{\gamma})}. \tag{4.17}$$

For $\hat{\phi}_{ATT}$, each term on the right-hand side of (4.16) is an estimator of $\mathbb{E}[Y^1 | W = 1]$ and $\mathbb{E}[Y^0 | W = 1]$, respectively. These are extensions of Hirano et al. (2003)'s estimators for $\mathbb{E}[Y^1 | W = 1]$ and $\mathbb{E}[Y^0 | W = 1]$ discussed in Section 4.1.2, to the cases with MNAR outcome.

For these proposed IPW estimators for ATE $\hat{\phi}_{ATE}$, ATT $\hat{\phi}_{ATT}$ and ATU $\hat{\phi}_{ATU}$, the following theorem holds.

Theorem 4.1 (consistency and asymptotic normality (Yoneyama & Minami, 2023))**.** *Suppose that Assumptions 1 – 4 hold. Assume conditions of Lemma A.2 and Lemma A.3 hold when the IV is one-dimensional, or conditions of Lemma A.4 and Lemma A.5 hold when the IVs are multidimensional. Then, the IPW estimators for treatment effects* $\hat{\phi}_{ATE}$*,* $\hat{\phi}_{ATT}$*, and* $\hat{\phi}_{ATU}$ *given by (4.15), (4.16) and (4.17) are consistent and asymptotically normal.*

Proof and asymptotic variances are given in the supplemental appendix B.4.

We also provide the sandwich variance estimator for the asymptotic variance of $\hat{\phi}_{ATE}$, $\hat{\phi}_{ATT}$, and $\hat{\phi}_{ATU}$ in the appendix B.4. Another method to evaluate the variability of an estimator is to use the bootstrap method (Efron & Tibshirani, 1994). The bootstrap method has the advantage of not requiring the theoretical computation of the sandwich variance estimator. In the simulation study described in the next section, variances of estimates were calculated using both the asymptotic variance and bootstrap methods.

4.5 Simulation studies

In this section, we report the simulation studies conducted to evaluate the performance of the IPW estimator using the extended dual propensity score for the average treatment effect. In this experiment, we focused on the following points.

(i) Whether the bias of the estimates is sufficiently small, particularly when the sample size is large.

(ii) Whether the variability of the proposed estimators for the average treatment effect can be appropriately evaluated by the sandwich variance estimates $\hat{\phi}_{ATE}$ given by formula (B.5) in the supplemental appendix B.4.

(iii) Whether the bootstrap estimates for the variance can be used to assess the variability of the estimates.

(iv) How the strength of the association between the IV and the missing variable makes a difference in the estimates by the proposed method.

(v) Differences in performance between the proposed method and other commonly used methods to adjust for missing data, specifically the listwise-deletion method and the multiple imputation method.

4.5.1 Data generation

Simulation was conducted in three scenarios, changing the influence of the IV on missing and increasing the variance of the covariates. For each scenario, we generated 1000 independent datasets with sample sizes $N = 3000, 5000$, and 10000, with different parameter values for the missing model. Each dataset consists of observation vectors $O = (Y^*, X, W, R, Z)$, where the covariate vector is bivariate, $X = (X_1, X_2)$, and the IV Z is univariate.

The specific models were as follows,

$$\begin{aligned}
(X_1, X_2)^T &\sim N\left(\mathbf{0},\, \sigma_x^2 I\right) \\
Z|X &\sim N\left(0.7 + 0.3X_1,\, 0.3\right) \\
W|X, Z &\sim \text{Bernoulli}\big\{\text{expit}(\gamma^T(1, X_1, X_2, Z)^T)\big\} \\
Y^1|X &\sim N\left(2.1 + 1.5X_1 + 1.2X_2, 0.2\right) \\
Y^0|X &\sim N\left(1.6 + X_1 + 0.7X_2, 0.2\right) \\
R^1|X, Y^1, Z &\sim \text{Bernoulli}\big\{\text{expit}({\theta^1}^T(1,\, X_2,\, Z)^T + \nu^1 Y^1)\big\} \quad \text{and} \\
R^0|X, Y^0, Z &\sim \text{Bernoulli}\big\{\text{expit}({\theta^0}^T(1,\, X_2,\, Z)^T + \nu^0 Y^0)\big\},
\end{aligned}$$

where $\text{expit}(x) = (1 + \exp(-x))^{-1}$ and $\theta^1,\ \nu^1,\ \theta^0$ and ν^0 are parameters that took different values depending on the scenario. These parameters were set as follows.

Scenario 1

$$\sigma_x^2 = 0.1,\ \gamma = (-0.7, 0.7, 0.6, 0.9)^T$$
$$\theta^1 = (-1.6, 0, 1.4)^T,\ \nu^1 = 0.6,\ \theta^0 = (1.3, 0, 1.1)^T,\ \nu^0 = -1$$

Scenario 2

$$\sigma_x^2 = 0.1,\ \gamma = (-0.7, 0.7, 0.6, 0.9)^T$$
$$\theta^1 = (-2.4, 0.8, 1.4)^T,\ \nu^1 = 1,\ \theta^0 = (2, -0.6, 1.1)^T,\ \nu^0 = -1.4$$

Scenario 3

$$\sigma_x^2 = 0.5,\ \gamma = (-0.4, 0.35, 0.3, 0.45)^T$$
$$\theta^1 = (-0.9, 0.4, 0.7)^T,\ \nu^1 = 0.5,\ \theta^0 = (1.2, -0.3, 0.55)^T,\ \nu^0 = -0.7$$

The difference between scenarios 1 and 2 is the strength of the association between the IV and the missing with the association being stronger in scenario 1 than in scenario

2. In the linear predictors of the logistic regression models for missing, although the coefficients of the IV Z are the same at 1.4 in both scenarios, the coefficients of X_2 are zero in scenario 1, while they are 0.8 and -0.6 for R_1 and R_0, respectively, in scenario 2. The absolute values of the coefficients of Y^w are also larger in scenario 2. The correlations of Z and the linear predictors of the logistic regression models for R^1 and R^0 are 0.88 and 0.69 in scenario 1, and 0.69 and 0.51 in scenario 2. Thus, the relative association between the IV Z and the missing is stronger in scenario 1 than in scenario 2.

In scenario 3, the variance of the covariates was changed from 0.1 to 0.5 based on scenario 2. Simply setting the variance of the covariate to 0.5 makes it easier to obtain the extended dual propensity scores close to 0. Therefore, the coefficients other than the constant term of the (extended) propensity scores for assignment and missing were halved, and the constant term was set so that the treatment and missing probabilities were about the same as in scenario 2. The correlations of Z and linear predictors of the logistic regression models for R^1 and R^0 are 0.56 and 0.18, respectively, in scenario 3.

In this setting, the missing mechanism of the outcome is MNAR and Assumptions 1 – 4 are satisfied. The true value of ATE ϕ_{ATE} is 0.5. In scenarios 1 and 2, the probability of treatment $\Pr(W = 1)$ is 0.484 and $\Pr(W = 1)$ is 0.480 in scenario 3. The probabilities of the non-missing outcome $P(R = 1)$ are 0.632 in scenario 1, 0.628 in scenario 2 and 0.627 in scenario 3.

4.5.2 Estimating methods for the ATE

For the generated datasets, we computed the proposed IPW estimates for MNAR missing $\hat{\phi}_{ATE}$ for the average treatment effect ϕ_{ATE}. For comparison, we also computed the estimates for ϕ_{ATE} using the listwise deletion method and the multiple imputation method.

We briefly describe the estimation procedure for the IPW estimates $\hat{\phi}_{ATE}$. We started by finding the estimate $\hat{\xi}$ for ξ in the model of $\mathbb{E}[Z|X;\xi]$ by a simple linear regression model of Z on X_1, and the estimate $\hat{\gamma}$ for γ in the model of $\pi_{W=w}(X, Z; \gamma)$ using a logistic regression model of W on X and Z. Then, we estimated θ^w and ν^w in the model of $\pi_{R^w=1}(X, Y^w, Z; \theta^w, \nu^w)$ by solving the estimating equations (4.13) with $\hat{\xi}$ and $\hat{\gamma}$ as given. Finally, $\hat{\phi}_{ATE}$ was computed using these estimates by (4.14). To solve the estimating equations (4.13), we first used the Nelder–Mead method with zeros as initial values for all the parameters, then switched to the Newton's method because the

Nelder–Mead method tends to have a broad convergence region and Newton's method has the property of quadratic convergence. For the Nelder–Mead method, we have used the R package GMM (Chaussé, 2010). The Nelder–Mead iterations were repeated until the relative convergence tolerance condition with $r = 1.490 \times 10^{-8}$ was satisfied, and then, the Newton iterations were repeated until the squared norm of the estimating function values $\|U^w\|^2$ became less than 10^{-24} and the sum of the squares of the changes in the parameters θ^w and ν^w in one step was less than 10^{-8}.

In the listwise-deletion method, all the observations of items with missing outcome Y^* were excluded from analysis and the ATE was estimated by the ordinary IPW method using the dataset without the missing data by listwise-deletion. The listwise-deletion method is often used because of its ease of procedure. The estimator by this method is guaranteed to be unbiased only if the missing type is MCAR.

For the multiple imputation method, we used predictive mean matching (PMM, (Rubin, 1986; Little, 1988)). We generated 100 complete datasets by imputing the missing data using PMM and computed the ATE estimates using the ordinary IPW method with the obtained complete datasets. The average of 100 estimates was then used as the multiple imputation estimate. The PMM method requires modeling of the outcome. We used linear regression models with covariates X_1 and X_2, which are the correctly identified models. The size of the donor pool required for PMM was set to five. The R package mice (van Buuren & Groothuis-Oudshoorn, 2011) was used to perform these calculations for the multiple imputation estimation method. In the case of MAR and missing data of less than 50%, multiple imputation with PMM is often used because it is known to give good results in terms of bias and coverage rate through simulation experiments (Marshall, Altman, Royston, & Holder, 2010).

Note that the Nelder–Mead sequences satisfied the given condition before the maximum iteration number in all datasets, but Newton's sequences did not satisfy the termination condition in some datasets. For scenario 1, there were eight datasets where the sequence did not converge when $N = 3000$ and one dataset did not converge when $N = 5000$, but it converged in all the datasets with $N = 10000$. For scenario 2, there were fifteen datasets where the sequence did not converge when $N = 3000$, four datasets when $N = 5000$, and two datasets when $N = 10000$. For scenario 3, there were 83 datasets where the sequence did not converge when $N = 3000$, 60 datasets when $N = 5000$, and 15 datasets when $N = 10000$. The difference in convergence between the scenarios

may be due to the strength of the association between the instrumental and missing variables, and it can be inferred that the stronger the association, the more stable is the convergence of the algorithm. A large variance in the covariates may also have made it easier for the extended dual propensity scores to take a value close to zero, which may lead to differences in convergence. Hereinafter, unless otherwise noted, the results of the proposed IPW method using the extended dual propensity score do not include those of the datasets in which Newton's method did not converge.

4.5.3 Comparison of estimation results

Figure 10 shows boxplots of the ATE estimates for 1000 datasets with sample size 3000, 5000 and 10000 by the three methods; the extended dual propensity score (ext. dual p.s.), multiple imputation (mi) and listwise-deletion (listwise) methods. Bias, standard deviation (SD), and root mean squared error (RMSE) of the ATE estimates are reported in Table 3. As the probability of not missing outcomes was approximately 60% in all scenarios, the approximate numbers of observed outcomes are 1800 when $N = 3000$, 3000 when $N = 5000$, and 6000 when $N = 10000$, respectively. The probability of the assignment $\Pr(W = 1)$ was approximately 0.5; thus, the effective sample size for estimating the population average of each treatment group were approximately $900, 1500$, and 3000 for $N = 3000, 5000$, and 10000, respectively.

As shown in Figure 10 and Table 3, for each case, the proposed IPW estimator using the extended dual propensity score had the smallest absolute bias. The multiple imputation estimator and the estimator with the listwise-deletion were apparently biased. Conversely, the proposed IPW estimator had the largest standard deviation in all cases. We can also see that there were outliers of estimates in $N = 3000$ in scenario 1 and all sample size cases of scenario 2 and 3 in Figure 10. This is because the outcomes in the tail of the distribution were heavily weighted. However, overall, the proposed IPW estimator had the smallest RMSE in all cases.

When the results for the datasets for which Newton's method did not converge were included, the absolute biases were approximately from 5 to 15%, and the RMSE was approximately 3% larger than those without these datasets in scenario 1 and 2. In Scenario 3, conversely, there were many cases where the algorithm did not converge, and the bias in the estimates was large in those cases, so the bias was at most three times larger when the datasets of algorithm nonconvergence were considered. The RMSE increased by approximately 18% for $N = 3000$ and $N = 5000$, and by 7% for $N = 10000$,

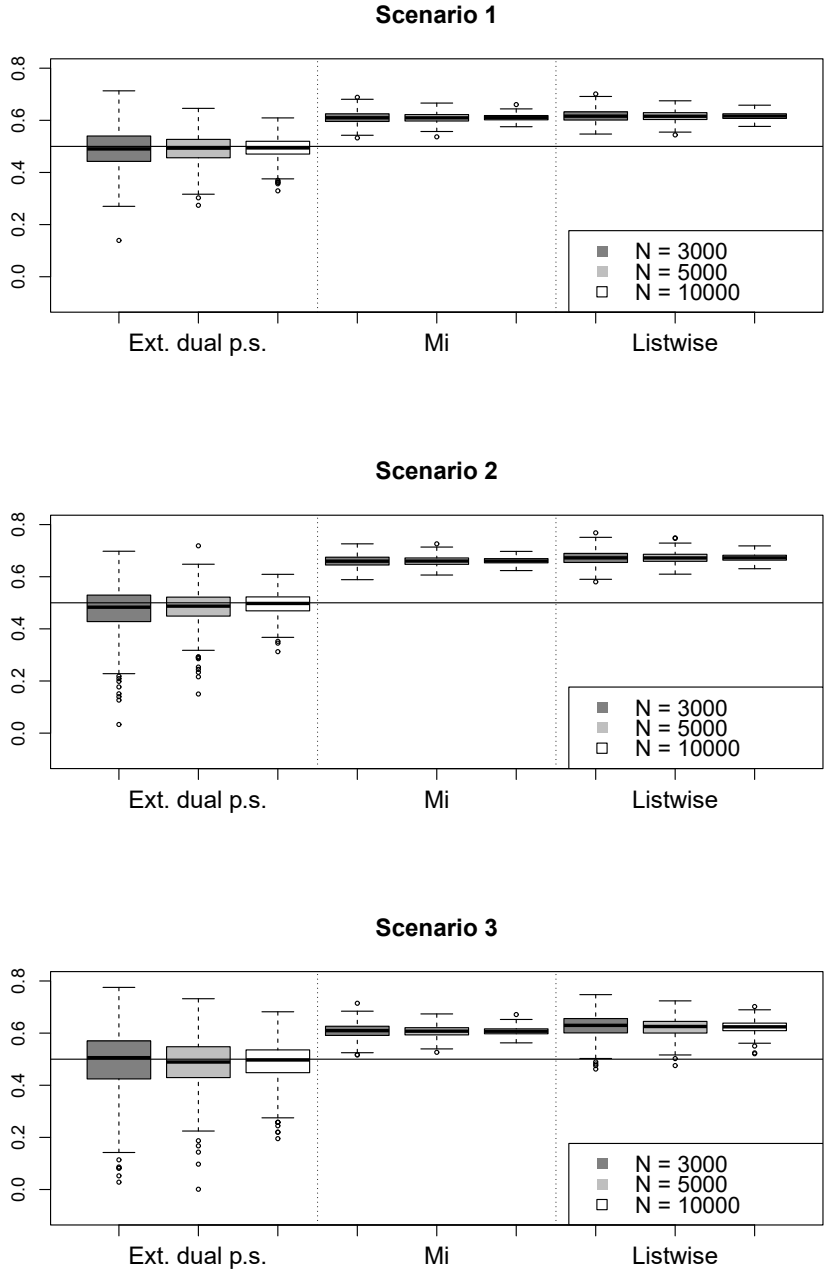

Figure 10: Boxplots of the ATE estimates for the simulation study computed by three methods: (i) proposed IPW estimator using extended dual propensity score (ext. dual p.s.), (ii) multiple imputation (mi) and (iii) listwise-deletion (listwise). The horizontal line represents the true value, $\mathbb{E}[Y^1 - Y^0] = 0.5$. Whiskers extend to the most extreme data points, less than twice the interquartile range from the box.

Table 3: Bias, standard deviation (SD) and root mean squared error (RMSE) of the ATE estimates for the simulation study by three methods: (i) the proposed IPW estimator using extended dual propensity score (ext. dual p.s.), (ii) multiple imputation (mi) and (iii) listwise-deletion (listwise).

		ext. dual p.s.			mi			listwise		
	N	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
Scenario 1	3000	−0.0096	0.0715	0.0721	0.1100	0.0226	0.1122	0.1168	0.0241	0.1192
($\Pr(R = 1) = 0.632$)	5000	−0.0088	0.0543	0.0550	0.1096	0.0178	0.1111	0.1160	0.0193	0.1176
	10000	−0.0056	0.0379	0.0383	0.1098	0.0119	0.1104	0.1162	0.0130	0.1170
Scenario 2	3000	−0.0248	0.0786	0.0824	0.1602	0.0221	0.1617	0.1719	0.0261	0.1738
($\Pr(R = 1) = 0.628$)	5000	−0.0158	0.0615	0.0634	0.1604	0.0170	0.1612	0.1724	0.0205	0.1737
	10000	−0.0056	0.0404	0.0408	0.1609	0.0120	0.1613	0.1729	0.0140	0.1735
Scenario 3	3000	−0.0080	0.1106	0.1108	0.1093	0.0271	0.1126	0.1268	0.0437	0.1341
($\Pr(R = 1) = 0.627$)	5000	−0.0158	0.0897	0.0910	0.1072	0.0214	0.1094	0.1227	0.0341	0.1273
	10000	−0.0105	0.0674	0.0681	0.1072	0.0142	0.1082	0.1235	0.0223	0.1255

including the case of algorithm nonconvergence. However, the extended dual propensity score method still had the smallest absolute bias and RMSE was also the smallest except for the case of $N = 3000$ in scenario 3.

4.5.4 Assessment of confidence interval estimation

Next, we examine the performance of two confidence interval estimation methods: the sandwich variance estimation and the bootstrap method for the proposed IPW estimator for MNAR outcome. For each dataset, we computed confidence intervals using these two methods. With the sandwich variance estimate $\hat{\sigma}^2_{ATE}$, we constructed the Wald 95% confidence interval for ϕ_{ATE} by formula (B.5) in appendix B.4.

$$\hat{\phi}_{ATE} \pm \Phi^{-1}(0.975)\sqrt{\hat{\sigma}^2_{ATE}/N} \tag{4.18}$$

where $\hat{\phi}_{ATE}$ is the estimate for the ATE and $\Phi^{-1}(x)$ is the inverse function of the distribution function of the standard normal distribution.

In the bootstrap procedure, the number of items in the treatment and control groups for resampling were fixed to be the same as those for each dataset. In addition, the number of missing outcomes for resampling was fixed to be the same as that for each dataset. This is an extension of the bootstrap in two-sample data (Tu & Zhou, 2002) to the case of two groups of data with missing values to avoid extreme ratios during resampling. The bootstrap size was $B = 1000$, and bootstrap replications of $\hat{\phi}_{ATE}$ were computed for each resampling. When Newton's method did not converge, we discarded this bootstrap sample and generated another one. Based on the obtained bootstrap replications, we computed the bootstrap estimate of the standard error $\widehat{\text{se}}_B$ (Efron & Tibshirani, 1994). Bootstrap confidence intervals for ϕ_{ATE} were computed by (4.18) using $\widehat{\text{se}}_B$ instead of $\sqrt{\hat{\sigma}^2_{ATE}/N}$. We report the empirical coverage rates of the confidence intervals for the ATE computed by the sandwich variance estimation method and the bootstrap method in Table 4.

On examining the results from Table 4, we observe that the empirical coverage rates of confidence intervals for the ATE computed by the sandwich variance estimation method are close to the nominal level (95%) except for the sandwich variance estimation method in scenario 3. The reasons for are as follows. The sandwich method uses the parameter estimates to estimate the asymptotic variance; hence, if the parameter estimate itself has a bias, the asymptotic variance estimate will also have a bias. As can be seen from

Figure 10, scenario 3 has greater variability in ATE estimates than the other methods. Similarly, the variability of the parameter estimates for the extended dual propensity score is also expected to be greater in scenario 3 than in the other scenarios. Therefore, as the parameter estimates are often far from the true values in scenario 3, confidence intervals by the sandwich estimator method have not been correctly estimated.

Conversely, the bootstrap method evaluates variation by resampling, making it less affected by bias in the parameter estimates. In fact, the empirical coverage rates by the bootstrap method are close to the nominal level even in scenario 3. Judging from these results, we recommend bootstrap method for the variance estimation of $\hat{\phi}_{ATE}$ other than the sandwich variance estimation method to avoid permissiveness.

Table 4: Empirical coverage rates of 95% confidence intervals for the ATE for the simulation study by sandwich variance estimation method and bootstrap method with the proposed IPW estimator using the extended dual propensity score.

	N	sandwich	bootstrap
Scenario 1	3000	94.2	95.1
	5000	94.7	95.6
	10000	95.2	94.8
Scenario 2	3000	94.0	94.2
	5000	92.4	94.7
	10000	94.2	96.2
Scenario 3	3000	89.7	94.8
	5000	88.7	95.0
	10000	86.1	95.1

4.6 Evaluation of the proposed method using the real data

In this section, we evaluate the proposed method by the analysis for the data in which an IV and missing are generated artificially for the real data. We describe below the real data used, the artificial data generation method, and the evaluation method of the proposed estimator.

4.6.1 Used Real Data

We used Framingham Heart Study data from sashelp library of SAS 9.4. This study is a follow-up epidemiological study of cardiovascular disease that has been ongoing in Framingham, Massachusetts, USA, since 1948. These data contain 5209 records (one record per person) with following variables: status (dead/alive), cause of death, age

coronary heart disease diagnosed, sex, age at start of this study, height, weight, diastolic blood pressure, systolic blood pressure, Metropolitan Relative Weight, number of cigarettes smoked per day, age at death, cholesterol value, cholesterol status (borderline/desirable/high), blood pressure status (normal/optimal/high), weight status (underweight/normal/overweight), smoking status (non-smoker/light/moderate/heavy/very heavy).

Using these data, we will consider estimating the impact of smoking on mortality. As we are interested in binary assignment here, we group all but non-smokers as smokers and consider the estimation of the difference in death rates between smokers and non-smokers. According to this definition of smoking status, the percentage of smokers is 51.7%. The naive percentage of deaths by group is 40.3% in the smoker group and 35.6% in the non-smoker group.

4.6.2 Missing data generation

There were no missing values for the status and there were 166 records with missing values for any one variable, except for cause of death, age coronary heart disease diagnosed, and age at death, which are variables whose values occur only when an event occurs. Here, cause of death and age at death were removed from the data, and age coronary heart disease diagnosed was replaced by a two-level variable, diagnosed if value was present and non-diagnosed if value was absent. We call the 5043 records obtained by listwise deleting these 166 records the "original data". Note that the original data is complete data.

To evaluate the proposed method, we generated an IV Z and MNAR missing for the outcome of the original data R as the following model:

$$\begin{aligned} Z &\sim N(0,1), \\ R \mid Y, Z, W &\sim \begin{cases} \text{Bernoulli}(\text{expit}(0.4Z - 1.3Y + 1.1)) & (W = \text{smoker}) \\ \text{Bernoulli}(\text{expit}(0.4Z - 1.4Y + 1.2)) & (W = \text{non-smoker}) \end{cases}, \end{aligned}$$

where W is the assignment and Y is the binary variable which takes 1 when the status is dead and 0 when the status is alive. This missing mechanism assumes that "when a person is dead, the response to the survey is lost and is likely to be missing". According to this model of the IV and missing for the outcome, 1000 datasets were generated. The missing outcome percentage was 36.1% on average for the 1000 datasets.

4.6.3 ATE estimation and comparison of results

As in Section 4.5.2, we estimated the ATE by the proposed method, multiple imputation, and listwise-deletion, for 1000 datasets generated in Section 4.6.2. For comparison, we also estimated the ATE using the IPW estimation method with the propensity score for assignment to the original data.

The propensity score for assignment was estimated by fitting a logistic regression model with smoking status (smoker/non-smoker) as the explained variable and sex, age at start, cholesterol status, blood pressure status, weight status and coronary heart disease diagnosis status (diagnosed/non-diagnosed), as the explanatory variables. The models for the extended propensity scores for missingness were assumed as the logistic regression model with missing indicator as the explained variable and IV Z and the outcome, status for dead or alive Y as explanatory variables. The estimates of the IV were their sample mean.

Figure 11 presents boxplots of the estimates for the ATE, and $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, which were obtained in the ATE estimation process by the proposed method, multiple imputation, and listwise-deletion. Table 5 shows bias, SD, RMSE for ATE, $\mathbb{E}[Y^1]$, $\mathbb{E}[Y^0]$ by three methods. For the calculation of bias and RMSE, the estimates of each target by the original data were used instead of the true values.

Figure 11 and Table 5 show that the proposed method can estimate ATE, $\mathbb{E}[Y^1]$, and $\mathbb{E}[Y^0]$ without bias, although there is a larger variation in the estimates than the other two methods. For the multiple imputation method, the biases are large for $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$ but shows a small bias for ATE. This can be thought to be because by taking the difference between $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, the respective biases cancel each other out and the ATE bias becomes smaller. The listwise-deletion method produced large biases for ATE and $\mathbb{E}[Y^1]$, and a slightly smaller bias for $\mathbb{E}[Y^0]$.

For RMSE, multiple imputation method is the smallest for the ATE estimates. However, for $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$, the proposed method has smaller RMSEs. In the listwise-deletion method, the RMSE for $\mathbb{E}[Y^0]$, is smaller than the proposed method, but the RMSEs for ATE, which is of primary estimation interest, and $\mathbb{E}[Y^1]$ are larger than that of the proposed method.

From the above, the proposed method is considered to have consistently small systematic biases for ATE, $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$. Indeed, there are many situations where the evaluation of $\mathbb{E}[Y^1]$ and $\mathbb{E}[Y^0]$ is also needed in the evaluation of treatment effects.

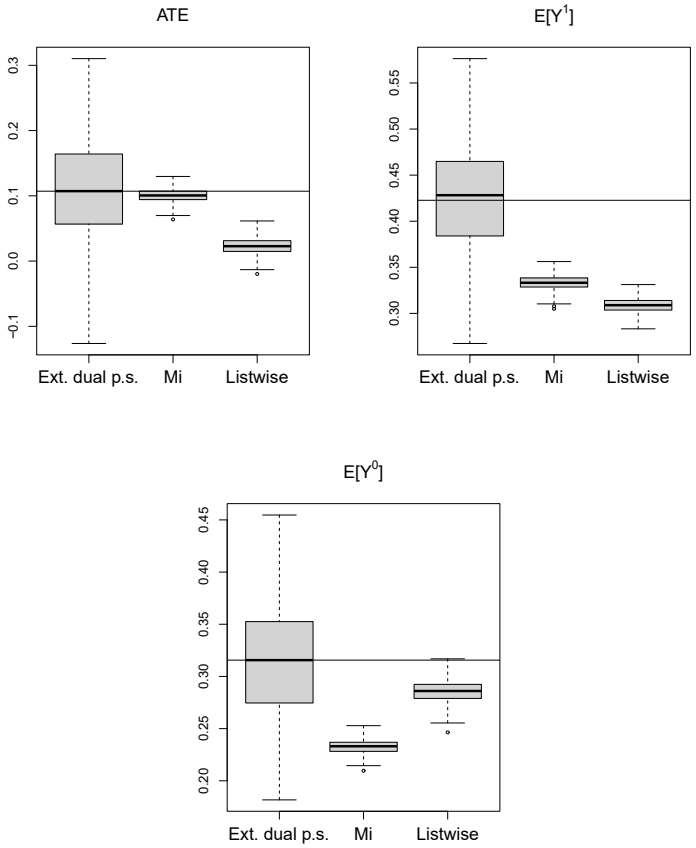

Figure 11: Boxplots of the ATE estimates, $\mathbb{E}[Y^1]$ estimates and $\mathbb{E}[Y^0]$ estimates for the datasets generated in Section 4.6.2 by three methods: (i) proposed IPW estimator using extended dual propensity score (ext. dual p.s.), (ii) multiple imputation (mi) and (iii) listwise-deletion (listwise). The horizontal line represents the estimates for original data. Whiskers extend to the most extreme data points, less than twice the interquartile range from the box.

Table 5: Bias, standard deviation (SD) and root mean squared error (RMSE) of the ATE estimates, $\mathbb{E}[Y^1]$ estimates and $\mathbb{E}[Y^0]$ estimates for the datasets generated in Section 4.6.2 by three methods: (i) proposed IPW estimator using extended dual propensity score (ext. dual p.s.), (ii) multiple imputation (mi) and (iii) listwise-deletion (listwise).

	ext. dual p.s.			mi			listwise		
Estimating target	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
ATE	0.00194	0.07726	0.07725	−0.00636	0.00955	0.01147	−0.08372	0.01243	0.08463
$\mathbb{E}[Y^1]$	0.00043	0.05546	0.05543	−0.08933	0.00743	0.08964	−0.11385	0.00777	0.11411
$\mathbb{E}[Y^0]$	−0.00151	0.05401	0.05400	−0.08297	0.00743	0.08321	−0.03013	0.00777	0.03177

Therefore, the proposed method is useful for evaluating treatment effects.

4.7 Discussion of Chapter 4

In this study, we proposed an estimating method for the ATE, ATT, and ATU when the outcome has MNAR missing. The estimators by this method are consistent and asymptotically normal even though the method does not require specifying the outcome model. With simulation experiments, we showed that the proposed method performs better than the conventional method and confirmed that the empirical coverage rates for the ATE computed by the sandwich variance estimator and the bootstrap method were close to the nominal level, particularly, the latter is better in our settings.

The existence of IVs is often discussed about methods that use these variables. Here, we describe an example of obtaining an IV. Let us assume that a researcher is interested in the effect of regular gym usage on physical strength. Here, as participation in the survey is voluntary and it is easy to gather people who are confident in their physical strength, MNAR is expected to occur. To properly estimate the treatment effects, we consider a set of pre-survey covariates and an invitation to a physical fitness test to obtain the IV, as elaborated below. First, the researcher administers a questionnaire for pre-survey covariates such as age, sex, financial situation, health awareness, treatment status, and whether the participant uses the gym regularly. After completion of the questionnaire, the researcher invites the participants to a central location test to measure their physical fitness. When inviting them, the researcher tells them that they will be given an honorarium for participating in the central location test, along with an honorarium amount that is randomly determined by the researcher for each survey participant. Then, the amount of honorarium can be considered as an IV because the amount is not directly related to the survey participant's health status, but it does influence whether the participant will attend the central location test.

Our future work will focus on introducing robust estimation methods to tackle the challenges of misspecification of the extended dual propensity score model such as doubly robust estimation (Scharfstein, Rotnitzky, & Robins, 1999), and instability of estimates due to enormous IPWs. Additionally, we hope to develop approaches to avoid any model identification, such as causal forests (Wager & Athey, 2018).

Chapter 5 Conclusion and future work

Missing data are often encountered in empirical studies. In particular, the MNAR case can easily arise, wherein both the observed and missing values affect the missing mechanism. However, MNAR has not been studied as much as other missing mechanisms. Additionally, certain issues remain, such as the difficulty of modeling outcomes and missing mechanisms, and limitation of the parameters that can be estimated. To address these gaps, this thesis focused on the case in which the outcome is MNAR and offered the following two proposals: First, we provided a strategy for covariate selection, wherein covariates should be included in the extended propensity score model when estimating the population mean of the outcome using the two-step estimation procedure with the extended propensity score. Second, we proposed the method used to estimate ATE, ATT, and ATU as treatment effects without modeling the outcome.

Chapter 3 proposed a strategy for selecting covariates to be included in the extended propensity score model to estimate the population mean of the outcome using a two-step estimation procedure with the extended propensity score. First, we showed the minimum set of covariates to be included in the model to establish consistency and asymptotic normality when estimating the population mean of the outcome. Next, we discussed the existence of covariates whose inclusion in the model would worsen the precision of the population mean estimation. The simulation results showed that a good strategy is to include only the minimum set of covariates in the model to establish consistency and asymptotic normality and reduce both the bias and variance of the estimated values.

Chapter 4 considered the case of an observational study with MNAR in the outcome and proposed a method to estimate the ATE, ATT, and ATU without modeling the outcome by using an IV when the outcome is MNAR. We showed that the resulting estimators are consistent and asymptotically normal. The simulation results also showed that the proposed method performs better in terms of RMSE than other commonly used methods for adjusting missing. Through the simulations, we also verified that we can properly estimate the confidence intervals for the proposed estimator.

These two proposals ensure the adjustment for missing data accessible, even when the outcome is MNAR.

Next, we describe the directions for future research. First, although we proposed a procedure for estimating the treatment effects and the covariate selection strategy in the extended propensity score model for estimating the population mean of the outcome, we have not discussed the covariate selection in the extended dual propensity score model estimating the treatment effects. Therefore, we would like to investigate the covariate selection using the extended dual propensity score model. In particular, we aimed to verify whether a combination of good covariate selection strategies for the propensity score for assignment and extended propensity score for missingness is a good covariate selection strategy for the extended dual propensity score as a whole. Second, as mentioned in previous chapters, both methods have a smaller bias but a larger SD in the estimates than the method that ignores missing data. Therefore, we would like to find a method to improve the precision of the estimation under MNAR other than the covariate selection.

Actively obtaining an IV through the data collection design solves the uncertainty of its existence and makes the adjustment of the MNAR more reliable, such as the population means estimation by Sun et al. (2018) and treatment effects estimation proposed in Chapter 4. Moreover, to use the IV thus obtained, the covariates in the extended propensity score model need not be included. Furthermore, the variance of the estimates decreases, as discussed in Section 3.3.2. Future work can focus on applying the proposed methods to real problems with MNAR by being involved from the data collection stage.

References

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Angrist, J. D. & Krueger, A. B. (1999). Empirical strategies in labor economics. In Ashenfelter, O. & Card, D., editors, *Handbook of Labor Economics*, volume 3, Part A, chapter 23, 1277–1366. Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. http://doi.org/10.1080/01621459.1996.10476902.

Barnes, D. E. & Yaffe, K. (2011). The projected effect of risk factor reduction on alzheimer's disease prevalence. *The Lancet Neurology*, *10*(9), 819–828. https://doi.org/10.1016/s1474-4422(11)70072-2.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, *163*(12), 1149–1156. https://doi.org/10.1093/aje/kwj149.

Cepeda, M., Boston, R., Farrar, J., & Strom, B. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology*, *158*, 280–7. http://doi.org/10.1093/aje/kwg115.

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with r. *Journal of Statistical Software*, *34*(11), 1–35. https://doi.org/10.18637/jss.v034.i11.

De Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*(4), 861–875. https://doi.org/10.1093/biomet/asr041.

d'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, *154*(1), 1–15. https://doi.org/10.1016/j.jeconom.2009.06.005.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*(4), 1231–1236. https://doi.org/10.2307/2532266.

Efron, B. & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *The Journal of Human Resources*, *32*(3), 441–462. https://doi.org/10.2307/146178.

Hernán, M. A. & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446.

Kennickell, A. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. In *Proceedings of the Section on Survey Research Methods*. 1990 Joint Statistical Meetings.

Li, W. & Zhou, X.-H. (2017). Identifiability and estimation of causal mediation effects with missing data. *Statistics in Medicine*, *36*(25), 3948–3965. https://doi.org/10.1002/sim.7413.

Little, R. (2008). Selection and pattern-mixture models. In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G., editors, *Longitudinal Data Analysis*, chapter 18, 409–432. Chapman and Hall/CRC.

Little, R. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Marshall, A., Altman, D. G., Royston, P., & Holder, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Medical Research Methodology*, *10*(7). https://doi.org/10.1186/1471-2288-10-7.

Miao, W. & Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, *103*(2), 475–482. https://doi.org/10.1093/biomet/asw016.

Miao, W., Ding, P., & Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, *111*(516), 1673–1683. https://doi.org/10.1080/01621459.2015.1105808.

Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. In Engle, R. & McFadden, D., editors, *Handbook of Econometrics*, volume 4, chapter 36, 2111–2245. Elsevier. https://doi.org/10.1016/s1573-4412(05)80005-4.

Richardson, T. S. (2013). Single world intervention graphs (swigs) : A unification of the counterfactual and graphical approaches to causality. In *Center for the Statistics and the Social Sciences, University of Washington Working Paper 128*, 1–146.

Robins, J. & Gill, R. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine*, *16*, 39–56. https://doi.org/10.1002/(sici)1097-0258(19970115)16:1⟨39::aid-sim535⟩3.0.co;2-d.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394. https://doi.org/10.1080/01621459.1987.10478441.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524. https://doi.org/10.1080/01621459.1984.10478078.

Rosenbaum, P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, *4*(1), 87–94. https://doi.org/10.2307/1391390.

Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *The Leadership Quarterly*, *31*(5), 101348. https://doi.org/10.1016/j.leaqua.2019.101348.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096–1120. https://doi.org/10.1080/01621459.1999.10473862.

Seaman, S. R. & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3), 278–295. https://doi.org/10.1177/0962280210395740.

Shortreed, S. & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*, 1111–1122. https://doi.org/10.1111/biom.12679.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., & Tchetgen Tchetgen, E. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, *28*(4), 1965–1983. https://doi.org/10.5705/ss.202016.0324.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics.

Tu, W. & Zhou, X.-H. (2002). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Services & Outcomes Research Methodology*, *3*, 135–147. https://doi.org/10.1023/A:1024212107921.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03.

van der Vaart, A. W. (1998). *Semiparametric Theory and Missing Data*. Springer Series in Statistics.

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113* (523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839.

Wang, H. & Kim, J. K. (2022). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv*. https://doi.org/10.48550/arXiv.2104.13469.

Wang, S., Shao, J., & Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, *24*(3), 1097–1116. https://doi.org/10.5705/ss.2012.074.

Yoneyama, S. & Minami, M. (2023). Treatment effects estimation with missing not at random data without outcome modeling. *Journal of Statistical Theory and Practice*, *3*(41), 41. https://doi.org/10.1007/s42519-023-00338-3.

Yoneyama, S. & Minami, M. (2024). Covariate selection strategy for the extended propensity score to adjust for missing not at random data. *International Journal of Statistics and Probability*, *13*(4), 26–41. https://doi.org/10.5539/ijsp.v13n4p26.

Zhao, A. & Ding, P. (2022). To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, *119*(545), 450–460. https://doi.org/10.1080/01621459.2022.2123814.

Zhou, K. & Jia, J. (2021). Propensity score adapted covariate selection for causal inference. *arXiv*. https://doi.org/10.48550/arXiv.2109.05155.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735.

Appendix A Definition and properties of estimators

A.1 Consistency of plug-in estimator

Lemma A.1 (Consistency of plug-in estimator (Theorem 4.5.1 of Amemiya (1985)))**.** *Let $g_N(\tau)$ be a nonnegative sequence of random variables depending on a parameter vector τ. Assume that following conditions hold:*

1) *$g_N(\tau)$ converges in probability to a nonstochastic function $g(\tau)$ uniformly in an open neighborhood of τ_0 as $N \to \infty$,*

2) *$\hat{\tau}_N \xrightarrow{p} \tau_0$ as $N \to \infty$,*

3) *$g(\tau)$ is continuous at τ_0.*

Then $g_N(\hat{\tau}_N) \xrightarrow{p} g(\tau_0)$ as $N \to \infty$.

A.2 Definition and properties of m-estimator

Consider a p-dimensional function $m(O, \tau)$ of observation vector O and parameter τ. The m-estimator $\hat{\tau}_N$ is defined as the solution of

$$\frac{1}{N}\sum_{i=1}^{N} m(O_i, \tau) = \mathbf{0}, \tag{A.1}$$

from a sample $O_1, \ldots, O_N \overset{\text{iid}}{\sim} p_O(\cdot, \tau)$, $\tau \in \mathcal{T} \subset \mathbb{R}^p$.

Lemma A.2 (consistency of m-estimator (Theorem 5.9 of van der Vaart (1998)))**.** *Assume that following conditions hold:*

1) $\sup_{\tau \in \mathcal{T}} \| \frac{1}{N}\sum_{i=1}^{N} m(O_i, \tau) - \mathbb{E}[m(O, \tau)] \| \xrightarrow{p} 0$,

2) $\inf_{\tau : \|\tau, \tau_0\| \geq \varepsilon} \|\mathbb{E}[m(O, \tau)]\| > 0 = \|\mathbb{E}[m(O, \tau_0)]\|$ *for every* $\varepsilon > 0$

Then, the any sequence of estimators $\hat{\tau}_N$ such that $\frac{1}{N}\sum_{i=1}^{N} m(O_i, \hat{\tau}_N) = o_p(1)$ converges in probability to τ_0.

Lemma A.3 (asymptotic normality of m-estimator (Theorem 5.21 of van der Vaart (1998))). *Assume that following conditions hold:*

1) *Each τ is in an open subset of Euclidean space,*

2) *$m(o,\tau)$ is a measurable vector-valued function such that, for every τ_1 and τ_2 in a neighborhood of τ_0 and a measurable function $\dot{m}(o)$ with $\mathbb{E}[\dot{m}^2(O)] < \infty$,*

$$\|m(o,\tau_1) - m(o,\tau_2)\| \leq \dot{m}(o)\|\tau_1 - \tau_2\|,$$

3) *$\mathbb{E}[\|m(O,\tau_0)\|^2] < \infty$,*

4) *$\mathbb{E}[m(O,\tau)]$ is differentiable at a zero τ_0, with nonsingular derivative matrix $\mathbb{E}\left[\frac{\partial m(O,\tau_0)}{\partial \tau^T}\right]$,*

5) *$\frac{1}{N}\sum_{i=1}^{N} m(O_i,\hat{\tau}_N) = o_p(N^{-1/2})$,*

6) *$\hat{\tau}_N \xrightarrow{p} \tau_0$ as $N \to \infty$.*

Then

$$\sqrt{N}(\hat{\tau}_N - \tau_0) = -\left\{\mathbb{E}\left[\frac{\partial m(O,\tau_0)}{\partial \tau^T}\right]\right\}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} m(O_i,\tau_0) + o_p(1).$$

In particular, the sequence $\sqrt{N}(\hat{\tau}_N - \tau_0)$ is asymptotically normal with mean zero and asymptotic variance $\left\{\mathbb{E}\left[\frac{\partial m(O,\tau_0)}{\partial \tau^T}\right]\right\}^{-1} \mathbb{E}[m(O,\tau_0)m^T(O,\tau_0)]\left\{\mathbb{E}\left[\frac{\partial m(O,\tau_0)}{\partial \tau^T}\right]\right\}^{-1^T}$.

A.3 Definition and properties of generalized method of moments

Suppose that $O_1,\ldots,O_N \overset{\text{iid}}{\sim} p_O(\cdot,\tau)$, $\tau \in \mathcal{T} \subset \mathbb{R}^p$. Let g be a q-dimensional function of observation O and parameter τ such that $q \geq p$ and $\mathbb{E}[g(O,\tau_0)] = \mathbf{0}$. The generalized method of moments (GMM) estimator $\hat{\tau}_N$ is defined as follows:

$$\hat{\tau}_N = \underset{\tau\in\mathcal{T}}{\operatorname{argmax}}\, Q_N(\tau)$$

where $Q_N(\tau) = -g_N^T(\tau)W_N g_N(\tau)$, $g_N(\tau) = \frac{1}{N}\sum_{i=1}^{N} g(O_i,\tau)$ and W_N is a $q \times q$ positive semi-definite matrix that converges in probability to a nonstochastic matrix W.

As long as W_N satisfies the conditions of the following theorems, the choice of W_N does not affect the consistency, but it does affect the asymptotic variance.

Lemma A.4 (consistency of the GMM estimator (Theorem 2.6 of Newey & McFadden (1994))). *Assume that following conditions hold:*

1) *W is positive semi-definite and $W\mathbb{E}[g(O,\tau)] = \mathbf{0}$ only if $\tau = \tau_0$,*

2) *$\tau_0 \in \mathcal{T}$ where $\mathcal{T}$ is compact,*

3) *$g(O,\tau)$ is continuous at each $\tau \in \mathcal{T}$ with probability one,*

4) *$\mathbb{E}[\sup_{\tau\in\mathcal{T}} \|g(O,\tau)\|] < \infty$.*

Then $\hat{\tau}_N \xrightarrow{p} \tau_0$ as $N \to \infty$.

Lemma A.5 (asymptotic normality of the GMM estimator (Theorem 3.4 of Newey & McFadden (1994)))**.** *Assume that in addition to the conditions in Lemma A.4, following conditions hold:*

1) *τ_0 is interior of $\mathcal{T}$,*

2) *$g(O,\tau)$ is continuously differentiable in a neighborhood $\mathcal{N}$ of τ_0, with probability approaching one,*

3) *$\mathbb{E}[\|g(O,\tau_0)\|^2] < \infty$,*

4) *$\mathbb{E}\left[\sup_{\tau\in\mathcal{N}} \left\|\frac{\partial g(O,\tau)}{\partial \tau^T}\right\|\right] < \infty$,*

5) *$G^T W G$ is nonsingular for $G = \mathbb{E}\left[\frac{\partial g(O,\tau_0)}{\partial \tau^T}\right]$,*

6) *$\Omega = \mathbb{E}[g(O,\tau_0)g^T(O,\tau_0)]$ exist.*

Then, $\sqrt{N}(\hat{\tau}_N - \tau_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, (G^T W G)^{-1} G^T W \Omega W G (G^T W G)^{-1})$.

With regard to W, it is optimal in terms of the asymptotic variance of the estimator to take $W = \Omega^{-1}$ if Ω is nonsingular. Then, the asymptotic variance of the GMM estimator is $(G^T \Omega^{-1} G)^{-1}$.

Appendix B Proofs of Propositions, Properties and Theorems

B.1 Proof of proposition 3.1

As in Sun et al. (2018), we prove Proposition 3.1 by showing that the expectation of the estimating function is equal to the zero vector. Let us first consider the case where the IV is one-dimensional.

Assume that we employ an estimating function $S(X_{IV}, Z; \xi)$ with the corresponding m-estimator $\hat{\xi}$ for ξ and it holds $\mathbb{E}\left[S(X_{IV}, Z; \xi_0)\right] = 0$ for the true value ξ_0. Then, the combined estimating function $\hat{\xi}, \hat{\theta}, \hat{\nu}$, and $\hat{\phi}$ defined in section 3.1 for ξ, θ, ν, and ϕ, respectively, is expressed as

$$m(O; \xi, \theta, \nu, \phi) = \begin{pmatrix} S(X_{IV}, Z; \xi) \\ U(O; \theta, \nu, \xi) \\ \dfrac{RY}{\pi(X_{PS}, Y, Z; \theta, \nu)} - \phi \end{pmatrix},$$

where $U(O; \theta, \nu, \xi)$ is the estimating function given by (2.4), but now it is expressed as

$$\begin{aligned} U(O; \theta, \nu, \xi) &= \left[U_1(O; \theta, \nu, \xi), U_2(O; \theta, \nu, \xi)\right]^T \\ U_1(O; \theta, \nu) &= \left\{\frac{R}{\pi(X_{PS}, Y, Z; \theta, \nu)} - 1\right\} h^*(X_{PS}, Z) \\ U_2(O; \theta, \nu, \xi) &= \frac{R}{\pi_R(X_{PS}, Y, Z; \theta, \nu)} \left(Z - f(X_{IV}; \xi)\right) \otimes g^*(X_{PS}, Y). \end{aligned}$$

Note that Y^* and Y in $U(O; \theta, \nu, \xi)$ are interchangeable because the functions containing Y^* are multiplied by R. If $R = 1$, then $Y^* = Y$ from the definition of Y^*, and if $R = 0$, then the parts multiplied by R are zero, regardless of Y. The estimating equation for ξ, θ, ν, and ϕ is given by

$$\frac{1}{N} \sum_{i=1}^{N} m(O_i; \xi, \theta, \nu, \phi) = \mathbf{0}.$$

Let ξ_0, θ_0, ν_0, and ϕ_0 denote the true values of the parameters ξ, θ, ν, and ϕ, respectively. We show that $\mathbb{E}[m(O; \xi_0, \theta_0, \nu_0, \phi_0)] = \mathbf{0}$ component by component. For the first component, we assumed the equation holds.

We first show that $\mathbb{E}[U(O; \xi_0, \theta_0, \nu_0)] = \mathbf{0}$. For $U_1(O; \theta_0, \nu_0)$, we have the following by the law of iterated expectations,

$$
\begin{aligned}
&\mathbb{E}\left[U_1(O; \theta_0, \nu_0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left\{\frac{R}{\pi(X_{PS}, Y, Z; \theta_0, \nu_0)} - 1\right\} h^*(X_{PS}, Z) \middle| X_{PS}, Y, Z\right]\right] \\
&= \mathbb{E}\left[\left\{\frac{\mathbb{E}[R|X_{PS}, Y, Z]}{\pi(X_{PS}, Y, Z; \theta_0, \nu_0)} - 1\right\} h^*(X_{PS}, Z)\right] \\
&= \mathbb{E}\left[\left\{\frac{\pi(X_{PS}, Y, Z; \theta_0, \nu_0)}{\pi(X_{PS}, Y, Z; \theta_0, \nu_0)} - 1\right\} h^*(X_{PS}, Z)\right] = \mathbf{0}.
\end{aligned}
$$

Similarly, the expectation of function U_2 is shown to be zero as

$$
\begin{aligned}
&\mathbb{E}\left[U_2(O; \theta_0, \nu_0, \xi_0)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{R}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} \left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y) \middle| X_{IV} \cup X_{PS}, Y, Z\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[R|X_{IV} \cup X_{PS}, Y, Z]}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} \left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[R|X_{PS}, Y, Z]}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} \left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y)\right] \quad (\because (3.2)) \\
&= \mathbb{E}\left[\frac{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} \left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y)\right] \\
&= \mathbb{E}\left[\left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Z - f(X_{IV}; \xi_0)\right) \otimes g^*(X_{PS}, Y) | X_{IV}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Z - f(X_{IV}; \xi_0) | X_{IV}\right] \otimes \mathbb{E}[g^*(X_{PS}, Y) | X_{IV}]\right] \quad (\because (3.1)) \\
&= \mathbb{E}\left[\left(\mathbb{E}[Z | X_{IV}; \xi_0] - f(X_{IV}; \xi_0)\right) \otimes \mathbb{E}[g^*(X_{PS}, Y) | X_{IV}]\right] \\
&= \mathbb{E}\left[\left(f(X_{IV}; \xi_0) - f(X_{IV}; \xi_0)\right) \otimes \mathbb{E}[g^*(X_{PS}, Y) | X_{IV}]\right] = \mathbf{0}.
\end{aligned}
$$

Similarly, the expectation of $RY/\pi(X_{PS}, Y, Z; \theta_0, \nu_0) - \phi$ is shown to be zero as

$$\begin{aligned}
\mathbb{E}\left[\frac{RY}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} - \phi_0\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{RY}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)}\middle| X_{PS}, Y, Z\right]\right] - \phi_0 \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left[R|X_{PS}, Y, Z\right]}{\pi_R(X_{PS}, Y, Z; \theta_0, \nu_0)} Y\right] - \phi_0 \\
&= \mathbb{E}[Y] - \phi_0 = 0.
\end{aligned}$$

Therefore, it holds that $\mathbb{E}[m(O; \xi_0, \theta_0, \nu_0, \phi_0)] = \mathbf{0}$. Here, assume that the parameter domain is a compact set, that $\mathbb{E}[m(O; \xi, \theta, \nu, \phi)]$ is continuous with respect to the parameters, and that $(\xi_0^T, \theta_0^T, \nu_0^T, \phi_0)^T$ is the only point satisfying $\mathbb{E}[m(O; \xi_0, \theta_0, \nu_0, \phi_0)] = \mathbf{0}$. Then the condition 2) of Lemma A.2 is satisfied. Furthermore, we assume condition 1) of Lemma A.2, i.e., $\frac{1}{N}\sum_{i=1}^{N} m(O_i; \xi, \theta, \nu, \phi)$ converges to $\mathbb{E}[m(O; \xi, \theta, \nu, \phi)]$ in probability uniformly. Then, by Lemma A.2, $(\hat{\xi}, \hat{\theta}, \hat{\nu}, \hat{\phi})$ converges in probability to $(\xi_0, \theta_0, \nu_0, \phi_0)$, that is, $\hat{\xi}, \hat{\theta}, \hat{\nu}$ and $\hat{\phi}$ are consistent estimators for ξ, θ, ν and ϕ.

As for the asymptotic normality, conditions 5), 6) in Lemma A.3 are already assumed or prooved. Assuming conditions 1), 2), 3), 4) additionally, Lemma A.3, which states asymptotic normality and asymptotic variance, holds. Condition 3) implies the assumption that the extended propensity score with the minimal set of covariates is bounded away from 0, that is, $\pi(x, y, z) > \sigma > 0$, for all $x \in \mathrm{Supp}(X_{PS})$, $y \in \mathrm{Supp}(Y)$, $z \in \mathrm{Supp}(Z)$, for some σ.

Asymptotic variances of $\hat{\xi}, \hat{\theta}, \hat{\nu}$, and $\hat{\phi}$ correspond to the diagonal components of the following asymptotic variance-covariance matrix:

$$\left\{\mathbb{E}\left[\frac{\partial m(O; \xi_0, \theta_0, \nu_0, \phi_0)}{\partial \tau^T}\right]\right\}^{-1} E\left[m(O; \xi_0, \theta_0, \nu_0, \phi_0) m^T(O; \xi_0, \theta_0, \nu_0, \phi_0)\right] \left\{\mathbb{E}\left[\frac{\partial m(O; \xi_0, \theta_0, \nu_0, \phi_0)}{\partial \tau^T}\right]\right\}^{-1^T},$$

where $\tau = (\xi^T, \theta^T, \nu^T, \phi)^T$.

When the IVs are multidimensional, we can consider GMM estimation with $\mathbb{E}[m(O; \xi_0, \theta_0, \nu_0, \phi_0)] = \mathbf{0}$ as the moment condition because this holds regardless of the dimension of the IV. As in the case where IV is one-dimensional, we can show about the asymptotic properties of the $\hat{\xi}, \hat{\theta}, \hat{\nu}$, and $\hat{\phi}$, using Lemmas A.4 and A.5. □

B.2 Proof of identifiability of the joint distribution $P(z, y^{w'}, r^{w'}, w)$

To discuss identifiability in the model of Section 4.4.1, we first parameterize the distribution. Consider the joint distribution $P(z, y^{w'}, r^{w'}, w)$ for fixed $w' \in \{1, 0\}$. Note that covariates X is omitted here for simplicity. Without loss of generality, by Assumptions

4.1, 4.3, and 4.4, we can decompose the conditional distribution as follows:

$$\begin{aligned} P(z, y^{w'}, r^{w'}, w) &= P(r^{w'}, w \mid z, y^{w'})P(z \mid y^{w'})P(y^{w'}) \\ &= P(r^{w'} \mid z, y^{w'})P(w \mid z)P(z)P(y^{w'}). \end{aligned}$$

Let us note that $\theta^{w'}, \gamma, \xi$ and $\phi^{w'}$ be the parameters for each distribution on the right-hand side, i.e.,

$$P_{\theta^{w'},\gamma,\xi,\phi^{w'}}(z, y^{w'}, r^{w'}, w) = P_{\theta^{w'}}(r^{w'} \mid z, y^{w'})P_{\gamma}(w \mid z)P_{\xi}(z)P_{\phi^{w'}}(y^{w'}). \tag{B.1}$$

Since the modeling of the joint distribution of y^1 and y^0 is not necessary for the estimation of causal effects, the causal effects can be specified if $\theta^{w'}, \gamma, \xi$ and $\phi^{w'}$ are identifiable for $w' = 1$ and 0, respectively. Assume that $P_{\gamma}(w|z)$ and $P_{\xi}(z)$ are identifiable because they have no missing data. Here we show that $P_{\theta^{w'},\gamma,\xi,\phi^{w'}}(z, y^{w'}, r^{w'}, w)$ is identifiable for fixed w', that is, for any pair of parameters $(\theta_1^{w'}, \phi_1^{w'})$ and $(\theta_2^{w'}, \phi_2^{w'})$ such that $\theta_1^{w'} \neq \theta_2^{w'}$ or $\phi_1^{w'} \neq \phi_2^{w'}$,

$$P_{\theta_1^{w'},\gamma,\xi,\phi_1^{w'}}(z, y^{w'}, r^{w'}, w) \neq P_{\theta_2^{w'},\gamma,\xi,\phi_2^{w'}}(z, y^{w'}, r^{w'}, w)$$

for some values of z and $y^{w'}$. More specifically, in the following proposition, the potential variable version of Result 1 of Sun et al. (2018) holds.

Proposition B.1. *Suppose that Assumptions 2.1, 4.3 and 4.4 hold, then the joint distribution* $P(z, y^{w'}, r^{w'}, w)$ *is identifiable for* $w' \in \{1, 0\}$ *if and only if* $P_{\theta^{w'}}(R^{w'}|Z, Y^{w'})$ *and* $P_{\phi^{w'}}(Y^{w'})$ *satisfy that for any pair of parameters* $(\theta_1^{w'}, \phi_1^{w'})$ *and* $(\theta_2^{w'}, \phi_2^{w'})$ *such that* $\theta_1^{w'} \neq \theta_2^{w'}$ *or* $\phi_1^{w'} \neq \phi_2^{w'}$*, it holds*

$$\frac{P_{\theta_1^{w'}}(R^{w'} = 1 \mid z, y^{w'})}{P_{\theta_2^{w'}}(R^{w'} = 1 \mid z, y^{w'})} \neq \frac{P_{\phi_2^{w'}}(y^{w'})}{P_{\phi_1^{w'}}(y^{w'})} \tag{B.2}$$

for at least one value of z *and* $y^{w'}$*.*

Proof: If the distribution is identifiable, it is obvious the equation (B.2) holds. Therefore, we show that the distribution is identifiable if the equation (B.2) holds. Assume that $P_{\gamma}(w|z)$ and $P_{\xi}(z)$ are identifiable because they have no missing data. Here we show the contraposition for the identifiability of $P_{\theta^{w'}}(r^{w'} \mid z, y^{w'})$ and $P_{\phi^{w'}}(y^{w'})$ when

the equation (B.2) holds. Suppose that, for a pair of parameters $(\theta_1^{w'},\phi_1^{w'})$ and $(\theta_2^{w'},\phi_2^{w'})$ such that $\theta_1^{w'}\neq\theta_2^{w'}$ or $\phi_1^{w'}\neq\phi_2^{w'}$. the following equations hold for any z and $y^{w'}$

$$P_{\theta_1^{w'},\gamma_1,\xi_1,\phi_1^{w'}}(z,y^{w'},R^{w'}=1,W=w')=P_{\theta_2^{w'},\gamma_1,\xi_1,\phi_2^{w'}}(z,y^{w'},R^{w'}=1,W=w')$$

Substituting these into (B.1), we obtain

$$\frac{P_{\theta_1^w}(R^w=1\mid y^w,z)}{P_{\theta_2^w}(R^w=1\mid y^w,z)}=\frac{P_{\phi_2^w}(y^w)}{P_{\phi_1^w}(y^w)}.$$

Thus, the contraposition has been shown. □

The same proposition as Result 1 in Sun et al. (2018) can be shown for potential variables. Therefore, their results can be used in the same way for the parameters of models that are identifiable or not, and a logistic regression model with no interaction between the potential outcome and the IV is one of the models, wherein their parameters can be identified from Example 2 of Sun et al. (2018).

B.3 Proof of $\mathbb{E}[U^w(O;\xi_0,\gamma_0,\theta_0^w,\nu_0^w)]=\mathbf{0}$

We show that the moment condition $\mathbb{E}[U^w(O;\xi_0,\gamma_0,\theta_0^w,\nu_0^w)]=\mathbf{0}$ discussed in Section 4.4.1 holds for the true parameter values $\xi_0,\gamma_0,\theta_0^w$, and ν_0^w, where U^w is given by (4.12). First, we consider the expectation of function U_1^w given by (4.9). According to the law of iterated expectations, we have the following:

$$\begin{aligned}
&\mathbb{E}\left[U_1^w(O;\gamma_0,\theta_0^w,\nu_0^w)\right]\\
&\quad=\mathbb{E}\left[\left\{\frac{\mathbf{1}_{\{W=w\}}R}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^*,Z;\theta_0^w,\nu_0^w)}-1\right\}\tilde{h}(X,Z)\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\left\{\frac{\mathbf{1}_{\{W=w\}}R^w}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}-1\right\}\tilde{h}(X,Z)\middle|X,Y^w,Z\right]\right]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\left\{\frac{\mathbf{1}_{\{W=w\}}R^w}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}-1\right\}\middle|X,Y^w,Z\right]\tilde{h}(X,Z)\right]\\
&\quad=\mathbf{0}.
\end{aligned}$$

The fourth equality holds by the equation (4.8). Similarly, the expectation of function U_2^w given by (4.11) is shown to be zero as follows:

$$\begin{aligned}
&\mathbb{E}\left[U_2^w(O;\xi_0,\gamma_0,\theta_0^w,\nu_0^w)\right] \\
&\quad= \mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^*,Z;\theta_0^w,\nu_0^w)}\left(Z-f(X,\xi_0)\right)\otimes\tilde{g}(X,Y^*)\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}_{\{W=w\}}R^w}{\pi_{W=w}(X,Z;\gamma_0)\pi_{R^w=1}(X,Y^w,Z;\theta_0^w,\nu_0^w)}\right.\right. \\
&\qquad\qquad\left.\left.\times\left(Z-f(X,\xi_0)\right)\otimes\tilde{g}(X,Y^w)\,\middle|\,X,Y^w,Z\right]\right] \\
&\quad= \mathbb{E}\left[\left(Z-f(X,\xi_0)\right)\otimes\tilde{g}(X,Y^w)\right] \\
&\quad= \mathbf{0}.
\end{aligned}$$

The third equality holds by the equation (4.8) and the last equality holds by the equation (4.10). □

B.4 Proof of Theorem 4.1

Let τ denote the vector of all the parameters $(\xi^T,\gamma^T,\theta^{1T},\nu^{1T},\theta^{0T},\nu^{0T},\phi_{ATE})^T$. We first show that for the case of the one-dimensional IV, the proposed estimator $\hat{\phi}_{ATE}$ for the ATE has consistency and asymptotic normality using the framework of an m-estimator.

The estimating function $m(O;\tau)$ for τ is defined as

$$\begin{aligned}
&m(O;\tau) \\
&\quad= \begin{pmatrix}
S_\xi(X,Z;\xi) \\
S_\gamma(X,Z,W;\gamma) \\
U^1(O;\xi,\gamma,\theta^1,\nu^1) \\
U^0(O;\xi,\gamma,\theta^0,\nu^0) \\
\dfrac{\mathbf{1}_{\{W=1\}}RY^*}{\pi_{W=1,R^1=1}(X,Y^*,Z;\gamma,\theta^1,\nu^1)}-\dfrac{\mathbf{1}_{\{W=0\}}RY^*}{\pi_{W=0,R^0=1}(X,Y^*,Z;\gamma,\theta^0,\nu^0)}-\phi_{ATE}
\end{pmatrix}
\end{aligned} \tag{B.3}$$

where $S_\xi(X,Z;\xi)$ and $S_\gamma(X,Z,W;\gamma)$ are estimating functions for m-estimator $\hat{\xi}$ and $\hat{\gamma}$,

respectively. Let $\hat{\tau} = (\hat{\xi}^T, \hat{\gamma}^T, \hat{\theta^1}^T, \hat{\nu^1}^T, \hat{\theta^0}^T, \hat{\nu^0}^T, \hat{\phi}_{ATE})^T$ be the solution of

$$\frac{1}{N}\sum_{i=1}^{N} m(O_i;\tau) = \mathbf{0}.$$

Then, each element of $\hat{\tau}$ is equal to the corresponding estimator obtained in Sections 4.3 and 4.4.2.

Now, we show that $\mathbb{E}[m(O;\tau_0)] = \mathbf{0}$ for the true parameter vector $\tau_0 = ({\xi_0}^T, {\gamma_0}^T, {\theta_0^1}^T, {\nu_0^1}^T, {\theta_0^0}^T, {\nu_0^0}^T, \phi_{ATE0})^T$. Because of the property of m-estimators, it holds that

$$\mathbb{E}\left[S_\xi(X, Z; \xi_0)\right] = \mathbf{0} \quad \text{and} \quad \mathbb{E}\left[S_\gamma(X, Z, W; \gamma_0)\right] = \mathbf{0}.$$

In appendix B.3, it is shown that $\mathbb{E}\left[U^w(O; \xi_0, \gamma_0, \theta_0^w, \nu_0^w)\right] = \mathbf{0}$ for $w = 1, 2$.

The expectation of the last element of $m(O;\tau)$ is shown to be zero from

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}_{\{W=1\}} R Y^*}{\pi_{W=1,R^1=1}(X, Y^*, Z; \gamma_0, \theta_0^1, \nu_0^1)}\right] - \mathbb{E}\left[\frac{\mathbf{1}_{\{W=0\}} R Y^*}{\pi_{W=0,R^0=1}(X, Y^*, Z; \gamma_0, \theta_0^0, \nu_0^0)}\right] \\
&\quad = \mathbb{E}\left[Y^1\right] - \mathbb{E}\left[Y^0\right] = \phi_{ATE0}.
\end{aligned}$$

From the above, we have $E[m(O;\tau_0)] = \mathbf{0}$ for the true parameter value τ_0. Here, we assume that the parameter domain is a compact set, that $\mathbb{E}[m(O;\tau)]$ is continuous with respect to the parameters, and that τ_0 is the only point satisfying $\mathbb{E}[m(O;\tau_0)] = \mathbf{0}$. Then the condition 2) of Lemma A.2 is satisfied. Furthermore, we assume the condition 1) of Lemma A.2 , i.e., $\frac{1}{N}\sum_{i=1}^{N} m(O_i;\tau)$ converges to $\mathbb{E}[m(O;\tau)]$ in probability uniformly. Then, by Lemma A.2, $\hat{\tau}$ converges in probability to τ_0, that is, $\hat{\tau}$ is a consistent estimator for τ.

As for the asymptotic normality, conditions 5), 6) in Lemma A.3 are already assumed or proved. Assuming conditions 1), 2), 3), 4) additionally, Lemma A.3, which states asymptotic normality and asymptotic variance, holds. Condition 3) implies the following necessary conditions for the IPW estimation (Robins et al., 1994; Sun et al., 2018).

$$\begin{aligned}
&0 < \rho_1 < \pi_{W=1}(x, z) < \rho_2 < 1 \\
&0 < \rho_3 < \pi_{R^1=1}(x, y^1, z), \quad 0 < \rho_4 < \pi_{R^0=1}(x, y^0, z)
\end{aligned}$$

with probability one for some constants $\rho_1, \ldots, \rho_4$.

Therefore, $\hat{\phi}_{ATE}$, an element of $\hat{\tau}$ is consistent and asymptotically normal. We can express the asymptotic normality of $\hat{\tau}$ with the asymptotic variance matrix V as

$$\sqrt{N}(\hat{\tau} - \tau_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, V),$$

where

$$V = \left\{ \mathbb{E}\left[\frac{\partial m(O;\tau_0)}{\partial \tau^T} \right] \right\}^{-1} \mathbb{E}\left[m(O;\tau_0) m(O;\tau_0)^T \right] \left\{ \mathbb{E}\left[\frac{\partial m(O;\tau_0)}{\partial \tau^T} \right] \right\}^{-1^T}.$$

Thus, the asymptotic variance σ^2_{ATE} of $\hat{\phi}_{ATE}$ is the bottom-right element of V.

The asymptotic variance V can be estimated using the "the sandwich estimator" $\hat{V}$ as follows:

$$\hat{V} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial m(O_i;\hat{\tau})}{\partial \hat{\tau}^T} \right\}^{-1} \left[\frac{1}{N} \sum_{i=1}^{N} m(O_i;\hat{\tau}) m(O_i;\hat{\tau})^T \right] \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial m(O_i;\hat{\tau})}{\partial \tau^T} \right\}^{-1^T} \tag{B.4}$$

The estimator of the asymptotic variance $\hat{\sigma}^2_{ATE}$ for σ^2_{ATE} can also be obtained from the bottom-right element of $\hat{V}$, i.e.,

$$\hat{\sigma}^2_{ATE} = (0 \ \cdots \ 0 \ 1)\hat{V}(0 \ \cdots \ 0 \ 1)^T. \tag{B.5}$$

For a case of multidimensional IVs, we employ the generalized method of moments (GMM) with the moment conditions $\mathbb{E}[m(O;\tau)] = \mathbf{0}$ for the estimation of τ, where $m(O;\tau)$ is the function given by (B.3). Because $\mathbb{E}[m(O;\tau_0)] = \mathbf{0}$ holds for the true parameter values τ_0, as shown above, the consistency and asymptotic normality of the corresponding estimator $\hat{\tau}$; thus, those of $\hat{\phi}_{ATE}$ can be established under conditions of Lemma A.4 and A.5.

Next, we consider the estimation of the average treatment effect on the treated (ATT). Here, we show that m-estimation or GMM can be applied for ATT in the same way as for the ATE. Let $\tilde{\tau}$ denote the vector of all the parameters for estimating ATT,

$\tilde{\tau}= (\xi^T\ \gamma^T\ \theta^{1T}\ \nu^{1T}\ \theta^{0T}\ \nu^{0T}\ \phi_{ATT})^T$. We define the function $\tilde{m}(O;\tilde{\tau})$ as follows:

$$\tilde{m}(O;\tilde{\tau}) = \begin{pmatrix} S_\xi(X,Z;\xi) \\ S_\gamma(X,Z,W;\gamma) \\ U^1(O;\xi,\gamma,\theta^1,\nu^1) \\ U^0(O;\xi,\gamma,\theta^0,\nu^0) \\ \dfrac{\mathbf{1}_{\{W=1\}}R\pi_{W=1}(X,Z;\gamma)Y^*}{\pi_{W=1,R^1=1}(X,Y^*,Z;\gamma,\theta^1,\nu^1)} - \dfrac{\mathbf{1}_{\{W=0\}}R\pi_{W=1}(X,Z;\gamma)Y^*}{\pi_{W=0,R^0=1}(X,Y^*,Z;\gamma,\theta^0,\nu^0)} - \pi_{W=1}(X,Z;\gamma)\phi_{ATT} \end{pmatrix}.$$

For $\mathbb{E}[\tilde{m}(O;\tilde{\tau}_0)] = \mathbf{0}$ to hold, we only need to show that the expectations of the last elements of $\tilde{m}(O;\tilde{\tau}_0)$ is zero, because we have already shown the others.

By the law of iterated expectations, and Assumption 4.4,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbf{1}_{\{W=1\}}R\pi_{W=1}(X,Z;\gamma_0)Y^*}{\pi_{W=1,R^1=1}(X,Y^*,Z;\gamma_0,\theta_0^1,\nu_0^1)}\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\frac{WR^1\pi_{W=1}(X,Z;\gamma_0)Y^*}{\pi_{W=1,R^1=1}(X,Y^1,Z;\gamma_0,\theta_0^1,\nu_0^1)}\middle| X,Y^1,Z,W\right]\right] \\
&\quad= \mathbb{E}\left[\frac{\mathbb{E}\left[R^1|X,Y^1,Z\right]WY^1\pi_{W=1}(X,Z;\gamma_0)}{\pi_{W=1}(X,Z;\gamma_0)\pi_{R^1=1}(X,Y^1,Z;\theta_0^1,\nu_0^1)}\right] \\
&\quad= \mathbb{E}\left[WY^1\right]
\end{aligned}$$

Further,

$$\begin{aligned}
\mathbb{E}\left[WY^1\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[WY^1|X,Z,W\right]|X,Z\right]\right] \\
&= \mathbb{E}\left[P(W=1|X,Z)\mathbb{E}\left[1\cdot Y^1|X,Z,W=1\right]\right. \\
&\quad \left.+P(W=0|X,Z)\mathbb{E}\left[0\cdot Y^1|X,Z,W=0\right]\right] \\
&= \mathbb{E}\left[P(W=1|X,Z)\mathbb{E}\left[Y^1|X,Z,W=1\right]\right] \\
&= \int_{\mathcal{X}\times\mathcal{Z}} P(W=1|X=x,Z=z)\mathbb{E}\left[Y^1|X=x,Z=z,W=1\right]f(x,z)d(x,z) \\
&= P(W=1)\int_{\mathcal{X}\times\mathcal{Z}} \mathbb{E}\left[Y^1|X=x,Z=z,W=1\right]f(x,z|W=1)d(x,z) \\
&= P(W=1)\mathbb{E}\left[\mathbb{E}\left[Y^1|X,Z,W=1\right]|W=1\right] \\
&= P(W=1)\mathbb{E}\left[Y^1|W=1\right].
\end{aligned}$$

The fifth equality holds by the Bayes' theorem. Similarly, we can show

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{W=0\}}R\pi_{W=1}(X,Z;\hat{\gamma})Y^*}{\pi_{W=0,R^0=1}(X,Y^*,Z;\gamma_0,\theta_0^0,\nu_0^0)}\right] = \mathbb{E}\left[WY^0\right] = P(W=1)\mathbb{E}[Y^0|W=1],$$

and $\mathbb{E}[\pi_{W=1}(X,Z;\gamma_0)] = P(W=1)$; thus,

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{W=1\}}R\pi_{W=1}(X,Z;\gamma_0)Y^*}{\pi_{W=1,R^1=1}(X,Y^*,Z;\gamma,\theta_0^1,\nu_0^1)} - \frac{\mathbf{1}_{\{W=0\}}R\pi_{W=1}(X,Z;\gamma)Y^*}{\pi_{W=0,R^0=1}(X,Y^*,Z;\gamma,\theta_0^0\nu_0^0)} - \pi_{W=1}(X,Z;\gamma_0)\phi_{ATT}\right] = \mathbf{0}.$$

As in the case of ATE, the consistency and asymptotic normality for the ATT estimator holds by applying Lemma A.2 and Lemma A.3, or Lemma A.4 and Lemma A.5, assuming that each of these lemma's conditions is satisfied. The ATU case can be shown in the same way as the ATT. □

Acknowledgment

First, I am grateful to my advisor, Professor Mihoko Minami. Not only did she provide me with generous research guidance, but she also supported me in times of trouble outside of research. Without her support, I and this thesis would not exist today. I am also deeply appreciative of the valuable feedback and suggestions provided by the members of my dissertation committee: Professor Hideo Suzuki, Professor Hiroshi Shiraishi, Professor Kei Kobayashi, and Professor Kenichi Hayashi. Their insightful comments greatly contributed to the development of my research.