International Journal of Statistics and Probability; Vol. 13, No. 4; December 2024
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Covariate Selection Strategy for the Extended Propensity Score to Adjust for Missing Not at Random Data

Shintaro Yoneyama[1] & Mihoko Minami[2]

[1] Graduate School of Science and Technology, Keio University, Kanagawa, Japan
[2] Department of Mathematics, Keio University, Kanagawa, Japan
Correspondence: Shintaro Yoneyama, Graduate School of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa, Japan

Received: October 9, 2024 Accepted: November 20, 2024 Online Published: November 29, 2024

doi:10.5539/ijsp.v13n4p26 URL: https://doi.org/10.5539/ijsp.v13n4p26

Abstract

Missing data can introduce biases in the estimation of the indicator of interest if appropriate adjustments are not made. The case of Missing Not at Random (MNAR), a missing mechanism in which the missingness also depends on the missing values themselves, has been under-explored. When an outcome has MNAR data, one method to estimate the population mean of the outcome is using the extended propensity score. This method first estimates the extended propensity score, which is the missing probability conditional on the outcome and covariates. Then, the population mean of the outcome is estimated using these estimates. In this paper, we discuss which variables should be included in 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 the model of missing probability so that the population mean estimator of the outcome is consistent. Next, we show that the inclusion of some covariates in the missing probability model results in a large variance of the population mean estimates even if they explain the missing probability well. Then, we verify these arguments using simulation experiments and argue that to obtain unbiased, small-variance estimates of the population mean, it is desirable to include only those covariates necessary for consistency. This study allows us to obtain such estimates when the outcome is MNAR and adjusted by the extended propensity score.

Keywords: extended propensity score, instrumental variable, minimal set of covariates, missing not at random, variable selection

1. Introduction

Many empirical studies encounter missing data. Inferences that ignore missing data can lead to biased results. For example, in a test of cognitive function in older people, the more cognitively impaired they are, the more likely they are to drop out of the test. In such a case, inferences about cognitive function that ignore the missing data will not be valid because they will lead to an overestimation of average cognitive function.

Missing mechanisms are generally classified into three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR is the case where the missing probability is independent of the data, MAR is the case where the missing probability depends only on the observed data, and MNAR is the case where the missing probability depends not only on the observed data but also on the missing data itself (Little & Rubin, 2020; Tsiatis, 2006). The most typical method to deal with MCAR data is listwise deletion, which is an analysis method that removes records that have missing data even in one variable. In the case of MCAR data, this method does not introduce bias owing to missingness. However, in the case of MAR and MNAR data, listwise deletion may introduce bias.

Numerous studies have investigated methods for adjusting missing measurements while assuming MAR. One such method is to use likelihood models, such as the selection model and pattern mixture model, as outlined by Little (2008). However, the proper identification of the distribution of both the outcome and the missing data is required for this method. An alternative approach is to use the weighting method (J. Robins & Gill, 1997), where only the missing probability model needs to be properly identified. In this method 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 (e.g. Hirano, Imbens, & Ridder, 2003), which interprets potential outcomes as missing owing to assignment because of the similarities between MAR and the assumptions placed in causal inference. The method of multiple imputation by chained equations (MICE, Kennickell (1991); van Buuren and Groothuis-Oudshoorn (2011)) enables the adjustment for MAR data even when the missing data occur in several variables. MICE creates multiple imputed datasets, performs the estimation of interest on each imputed dataset, and merges the results.

In contrast to the plentiful literature on MAR data, few studies have investigated how to adjust for MNAR data. Hereafter, we consider the 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, both distributions of the missing mechanism and the outcome must be properly specified. In addition, further assumptions must be made to maintain identifiability (Miao, Ding, & Geng, 2016; S. Wang, Shao, & Kim, 2014). As a solution to this problem, several semiparametric MNAR adjustment and identification methods have recently been proposed. These methods are characterized by the use of independency among variables. d'Haultfoeuille (2010) introduced an estimation method using variables that are independent of the missing mechanism under conditioning on the outcome, but are related to the outcome. Miao and 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 the outcome conditioning, but is related to the outcome when conditioned with the covariates. Sun et al. (2018) proposed a method using an instrumental variable (IV) that is conditionally independent of the outcome given the covariates and is related to the missing pattern. Similar to the MAR weighting method, these methods utilize a two-step procedure to estimate the population mean. The difference of Sun et al. (2018)'s method from the MAR weighting method is that in the first step, the missing probability conditional on the covariates, as well as on the outcome, is estimated. As 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. We note that the necessary modeling is not easy and requires more expertise and background knowledge because models of missing mechanisms of outcomes include outcomes themselves.

In recent years, considering the two-step estimation of adjusting for missing outcome data with MAR and causal inference, the issue of variable selection in the first step to improve the precision of the second step estimation has been discussed regarding causal inference (Brookhart et al., 2006; De Luna, Waernbaum, & Richardson, 2011) and MAR (Seaman & White, 2013; H. Wang & Kim, 2022). These studies showed that including variables 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 population means and causal effects, which are of primary interest. In the context of causal inference, Shortreed and Ertefaie (2017) and Zhou and 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.

Therefore, in this paper, we discuss which covariates should be included in or excluded from a missing probability model to obtain a lower bias and variance estimator for the population mean when the outcome has MNAR data and is adjusted using Sun et al. (2018)'s method. First, we show which covariates, at a minimum, should be included in the model of missing probability in order for the population mean estimator of the outcome to be consistent. Next, we show that inclusion of some covariates in the missing probability model results in a large variance of the population mean estimates even if they explain the missing probability 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 missing probability model only those covariates that should be minimally included for consistency.

This paper is organized as follows. First, we set up the problem and outline the method of Sun et al. (2018) (Section 2). Next, we discuss the minimal set of covariates that should be included in the model of missing probability to obtain a consistent and asymptotically normal estimator of the population mean (Section 3.1). We also discuss the existence of covariates whose inclusion in the model of missing probability would inadvertently worsen the precision of estimating the population mean of the outcome (Section 3.2). We then verify the claims of Section 3 through simulations and further argue that including only the minimal set of covariates in the model of missing probability is a good strategy for variable selection in terms of bias and precision (Section 4). Finally, we conclude with the discussion (Section 5).

2. Preliminaries

We consider the case where our interest is the population mean of a variable, but the variable is MNAR. For this situation, Sun et al. (2018) proposed the MNAR missing adjustment method to estimate the population mean. In this paper, we base our discussion on their method. In this section, we first set up the problem and then outline the population estimation method of Sun et al. (2018).

2.1 Notation and Assumptions

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, $\phi = \mathrm{E}[Y]$. We assume that the outcome may be missing, and the mechanism is MNAR. Let R be the binary missing indicator for Y, such that $R = 1$ if Y is observed and $R = 0$ otherwise. We consider using covariates and the IV to adjust the MNAR outcome and estimate the population mean of the outcome. Let X be a vector of observed covariates and Z be an IV that satisfies the following two assumptions.

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:

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

Note that the IV can be multidimensional.

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

2.2 Extended Propensity Score and Inverse Probability Weighted (IPW) Estimator

Here, we outline the two-step population mean estimation method of Sun et al. (2018). The two-step procedure 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

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

is defined as the probability that the outcome is observed conditional on X, Y, and Z.

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

$$\pi(X, Y, Z; \theta) = \text{expit}\left(\theta^{h^T} h(X, Z) + \theta^{g^T} g(X, Y)\right) \tag{1}$$

where $\text{expit}(a) = (1+\exp(-a))^{-1}$, h and g are vector functions, and θ^h and θ^g 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:

$$f(X; \xi) = E[Z|X; \xi]. \tag{2}$$

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

For estimation of θ in the extended propensity score model (1), we define an estimating function U for θ given $\hat{\xi}$ as:

$$U(O; \theta, \xi = \hat{\xi}) = \begin{pmatrix} U_1(O; \theta) \\ U_2(O; \theta, \xi = \hat{\xi}) \end{pmatrix}, \tag{3}$$

where

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

h^* and g^* are arbitrary vector-valued functions of the same sizes as θ^h and θ^g, respectively; and $\otimes$ denotes the Kronecker product. Note that h^* and g^* need not necessarily be the same as h and g in the model (1) because how h^* and g^* are chosen does not affect the consistency for θ as long as they satisfy the regularity conditions for the m-estimator, as discussed Sun et al. (2018). In this paper, we set h^* and g^* to be h and g, respectively.

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

$$\frac{1}{N}\sum_{i=1}^{N} U(O_i;\theta,\xi=\hat{\xi}) = \mathbf{0}. \tag{4}$$

Given that $E_{\theta,\xi}[U(O;\theta,\xi)] = \mathbf{0}$ holds, $\hat{\theta}$ is an m-estimator under suitable regularity conditions as shown in the proof of the Proposition 1 of Sun et al. (2018).

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

Plugging $\hat{\theta}$ obtained above into θ of the extended propensity score model (1) 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 $\hat{\phi}$ for the population mean $\phi := E[Y]$:

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i Y_i^*}{\pi(X_i, Y_i^*, Z_i;\hat{\theta})}. \tag{5}$$

For each term in the summation in $\hat{\phi}$, 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 $\hat{\phi}$.

Proposition 2.1 [consistency and asymptotic normality (Sun et al., 2018)] Suppose that Assumption 2.1 and 2.1 hold. Then, the IPW estimator $\hat{\phi}$ is consistent and asymptotically normal for the population mean ϕ under suitable regularity conditions. Using the extended propensity score estimator from the first step, we are able to estimate the population mean in the second step.

3. Covariates for Estimating the Extended Propensity Score

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 Minimal Sets of Covariates

In this subsection we discuss the sets of covariates that must be included in the model for the extended propensity score 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,
  $$Y \perp\!\!\!\perp Z | X_{IV}. \tag{6}$$
  - If any covariate is removed from X_{IV}, the above conditional independence (6) 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,
  $$X_{IV} \perp\!\!\!\perp R | X_{PS}, Y, Z. \tag{7}$$
  - If any covariate is removed from X_{PS}, the above conditional independence (7) 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}. 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.

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

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) and (1), 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) and (1), respectively. That is, let

$$f(X_{IV}; \xi) = E[Z|X_{IV}; \xi] \tag{8}$$

be the model (2) for the conditional mean of the IV given X_{IV}, and let

$$\pi(X_{PS}, Y, Z; \theta) = \text{expit}\Big({\theta^h}^T h(X_{PS}, Z) + {\theta^g}^T g(X_{PS}, Y)\Big), \tag{9}$$

be the model for the extended propensity score (1). We estimate $\hat{\xi}$ and $\hat{\theta}$ by using models (8) and (9). 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,

$$\hat{\phi}_{MS} \coloneqq \frac{1}{N}\sum_{i=1}^{N} \frac{R_i Y_i^*}{\pi(X_{PS\,i}, Y_i^*, Z_i; \hat{\theta})}. \tag{10}$$

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

Proposition 3.1 Suppose that Assumption 3.1 holds, 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 under suitable regularity conditions. The proof and the asymptotic variance matrix is shown in Appendix A.

3.2 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, J. M. Robins, Rotnitzky, and Zhao (1994) proposed a two-step method to estimate the population mean of the outcome. Seaman and White (2013) and H. Wang and 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 (9) 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.

4. 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.

4.1 Simulation Scenarios

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

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

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.

The minimal sets of covariates X_{IV} and X_{PS} depend only on the values of $(\alpha_1, \alpha_2, \sigma)$ in the conditional distribution model of Z given X. $X_{IV} = X_{PS} = \varnothing$ 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 $O = (Y^*, R, X, Z)$.

4.2 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) 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 (1). These models correspond to all possible combinations of $X = (X_1\ X_2\ X_3)$.

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

   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 (4) for θ with $\hat{\xi}$ obtained in 1 using R package GMM (Chaussé, 2010).

3. The IPW estimator (5) 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 (4) 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.

4.3 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 1, 2, and 3 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 1 and 3 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 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 2 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 3 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

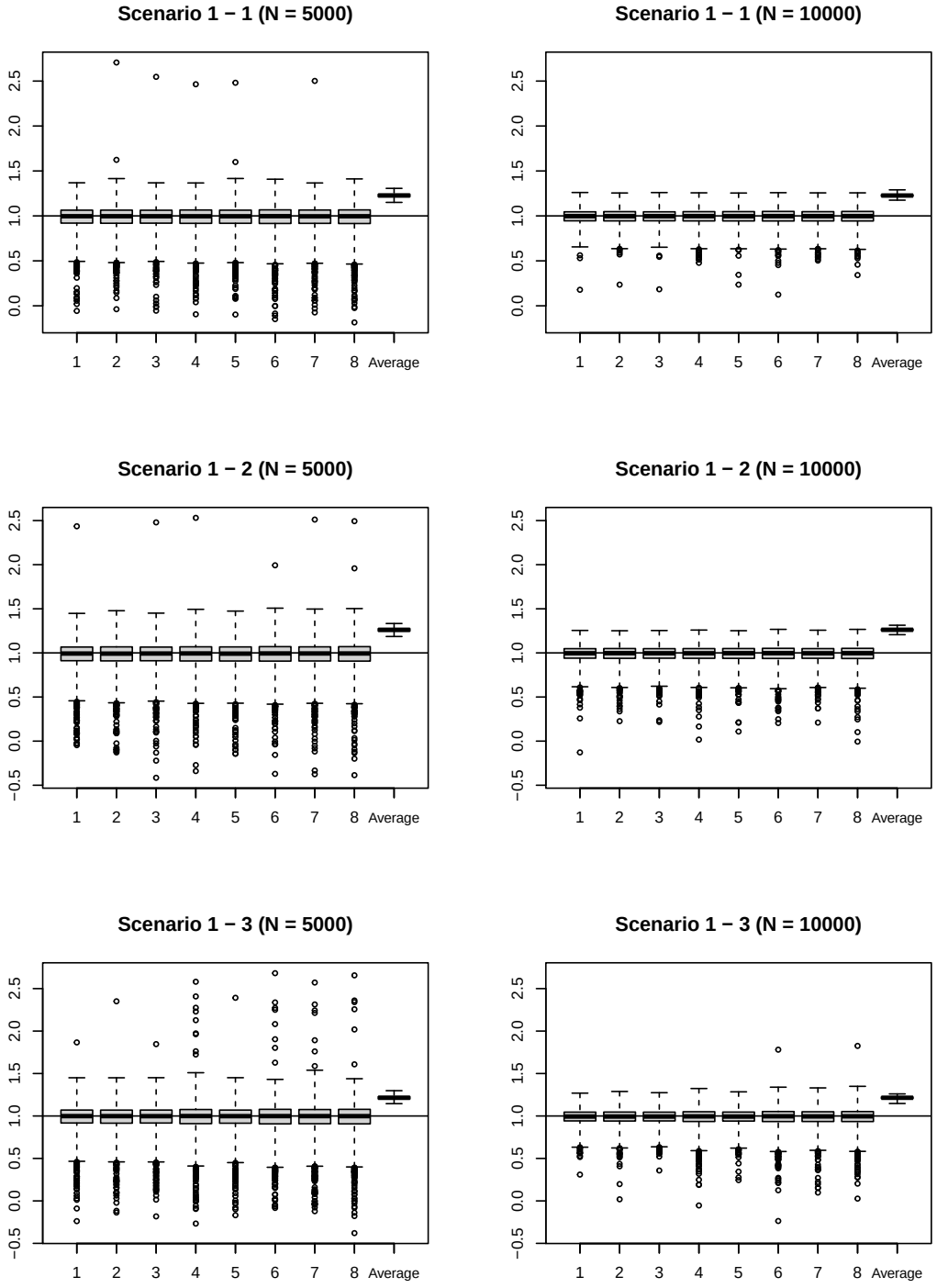

Figure 1. Boxplots of the population mean estimates for scenario A1 ($X_{IV} = X_{PS} = \varnothing$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side by side

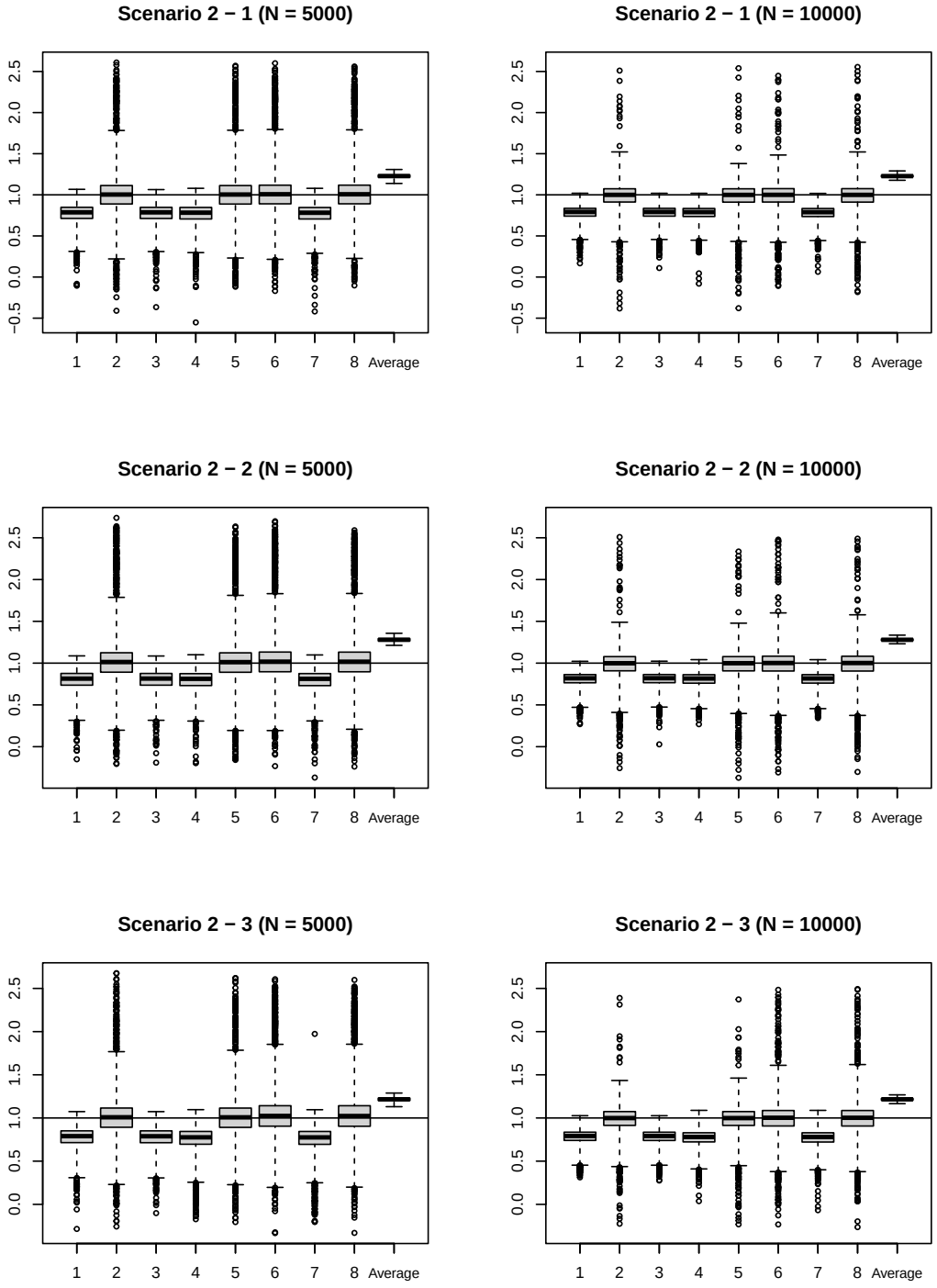

Figure 2. Boxplots of the population mean estimates for scenario A2 ($X_{IV} = X_{PS} = \{X_1\}$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side by side

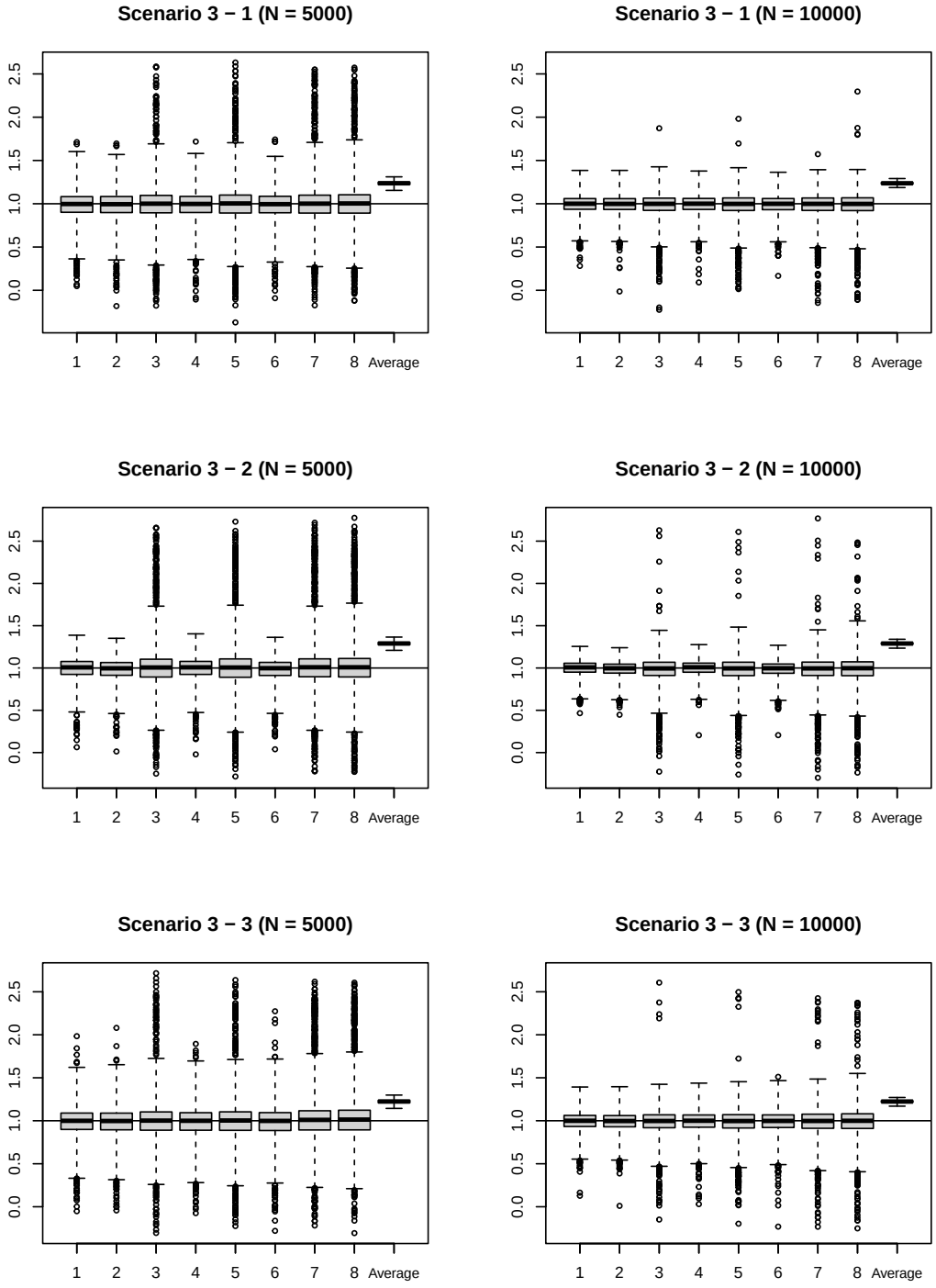

Figure 3. Boxplots of the population mean estimates for scenario A3 ($X_{IV} = \{X_2\},\ X_{PS} = \varnothing$) with B1, B2, and B3 (vertically), and with sample sizes 5000 and 10000 (horizontally). In each plot, the results by eight extended propensity models and simple sample average are placed side 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

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.

5. Conclusion

5.1 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.

5.2 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 and 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) 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.

5.3 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 1-3 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.

References

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

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

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. doi: https://doi.org/10.2307/146178

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

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.

Little, R. (2008). Selection and pattern-mixture models. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 409–432). Chapman and Hall/CRC.

Little, R., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

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

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

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

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

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

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.

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

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*, *17*(41). https://doi.org/10.1007/s42519-023-00338-3

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

A. Proof of Proposition 3.1

As in Sun et al. (2018), it is proven using the fact that $\hat{\xi}, \hat{\theta}$, and $\hat{\phi}$ are m-estimators. First, consider the case where the IV is one-dimensional.

The estimating function for ξ, θ, and ϕ is defined as follows

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

where $S(X_{IV}, Z;\xi)$ is an estimating function for the m-estimator $\hat{\xi}$. The solution for

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

$\hat{\theta}$ equals the solution to (4), and $\hat{\phi}$ equals (5).

Let ξ_0, θ_0, and ϕ_0 denote the true values of the parameter ξ, θ, and ϕ, respectively. To show that $m(O;\xi,\theta,\phi)$ is appropriate as an estimating function, we show that $E[m(O;\xi_0,\theta_0,\phi_0)] = \mathbf{0}$. For S, from the assumption that it is an estimating function of the m-estimator, $E[S(X_{IV}, Z;\xi_0)] = \mathbf{0}$ is satisfied. Therefore, we show that for U and $RY/\pi(X_{PS}, Y, Z;\theta) - \phi$, the expectation is zero when the true values of the parameters are given.

We show $E[U(O;\xi_0,\theta_0)] = \mathbf{0}$. From the definition of U, (3), we show that for each of U_1 and U_2, the expectation is zero given the true values of the parameters. For U_1, according to the law of iterated expectations, we have the following:

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

Similarly, note $X_{PS} \subset X_{IV}$ from the definition of X_{PS}, the expectation of function U_2 is shown to be zero as follows:

$$\begin{aligned}
&E\left[U_2(O;\theta_0,\xi_0)\right] \\
&= E\left[E\left[\frac{R}{\pi_R(X_{PS}, Y, Z;\theta_0)}(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)\middle| X_{IV}, Y, Z\right]\right] \\
&= E\left[\frac{E[R|X_{IV}, Y, Z]}{\pi_R(X_{PS}, Y, Z;\theta_0)}(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)\right] \\
&= E\left[\frac{E[R|X_{PS}, Y, Z]}{\pi_R(X_{PS}, Y, Z;\theta_0)}(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)\right] \quad (\because (7)) \\
&= E\left[\frac{\pi_R(X_{PS}, Y, Z;\theta_0)}{\pi_R(X_{PS}, Y, Z;\theta_0)}(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)\right] \\
&= E\left[(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)\right] \\
&= E\left[\left[(Z - f(X_{IV};\xi_0)) \otimes g^*(X_{PS}, Y)|X_{IV}\right]\right] \\
&= E\left[E\left[Z - f(X_{IV};\xi_0)|X_{IV}\right] \otimes E[g^*(X_{PS}, Y)|X_{IV}]\right] \quad (\because (6)) \\
&= E\left[(E[Z|X_{IV};\xi_0] - f(X_{IV};\xi_0)) \otimes E[g^*(X_{PS}, Y)|X_{IV}]\right] \\
&= E\left[(f(X_{IV};\xi_0) - f(X_{IV};\xi_0)) \otimes E[g^*(X_{PS}, Y)|X_{IV}]\right] = \mathbf{0}.
\end{aligned}$$

Note that $E[m(U;\xi_0,\theta_0)] = \mathbf{0}$ holds regardless of the form of h or g.

Similarly, for $RY/\pi(X_{PS}, Y, Z;\theta) - \phi$, it can be shown that the expectation is zero when the true value of the parameters are

given,

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

From the above, we show that $E[m(O; \xi_0, \theta_0, \phi_0)] = \mathbf{0}$. Thus, the framework of the m-estimator can be applied for $\hat{\xi}, \hat{\theta}$, and $\hat{\phi}$. Under the regularity conditions for m-estimator (Newey & McFadden, 1994) and the extended propensity score with minimal set of covariates is bounded away from 0, that is,

$$^{\exists}\sigma > 0, \text{ such that } \pi(X_{PS}, Y, Z) > \sigma > 0 \text{ with probability 1,}$$

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

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

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

As $E[m(O; \xi_0, \theta_0, \phi_0)] = \mathbf{0}$ holds regardless of the dimension of the IV, we can consider estimation by GMM with this as the moment condition when the IVs are multidimensional. □

Acknowledgments

Not applicable.

Authors contributions

Shintaro Yoneyama was responsible for the contents of this paper. Prof. Mihoko Minami gave advice on theoretical aspects and on writing the manuscript. All authors read and approved the final manuscript.

Funding

This work was partly supported by JSPS KAKENHI Grant Numbers JP21K11794.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education. The journals policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.