Statistica Sinica Preprint No: SS-2018-0466	
Title	Semiparametric Inference of Causal Effect with Nonignorable Missing Confounders
Manuscript ID	SS-2018-0466
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0466
Complete List of Authors	Zhaohan Sun and Lan Liu
Corresponding Author	Lan Liu
E-mail	liux3771@umn.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Statistica Sinica

Semiparametric Inference of Causal Effect With Nonignorable Missing Confounders

Zhaohan Sun[1] and Lan Liu[2]

[1] *University of Waterloo, Waterloo, Ontario, Canada*

[2] *School of Statistics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, USA*

Abstract: We consider the estimation of a causal effect when the confounders are subject to missingness. We allow the missingness of the confounders to be nonignorable; that is, the missingness may depend on the missing confounders, conditional on the observed data. The identification has been discussed in the literature; however, few studies have focused on semiparametric causal inference with nonignorably missing confounders. To address this, we propose three semiparametric estimators: the inverse probability weighting (IPW), regression, and doubly robust (DR) estimators. The IPW and regression estimators require a correct specification of the propensity scores and the regression models for the confounders and outcome, respectively. Assuming the selection bias odds ratio function is always correctly specified, the DR estimator uses both sets of models and is consistent if either set of models, but not necessarily both, is correctly specified. We investigate the finite-sample performance of our proposed semiparametric estimators using simulation studies and apply our estimators to SO_2 emissions data.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Key words and phrases: Causal inference, doubly robustness, outcome-independent missingness, nonignorable missing, shadow variable

1. Introduction

The missing data problem is frequently encountered in biomedical research, the social sciences, and environmental studies. There are two types of missing-data mechanisms (Rubin, 1976): ignorable missingness and nonignorable missingness. The mechanism is ignorable if the missingness of each variable does not depend on missing values conditional on the observed data (Little, 1992; Little and Rubin, 2002). This is sometimes referred to as missing at random. When data are subject to ignorable missingness, the maximum likelihood approach (Dempster et al., 1977; Ibrahim, 1990), imputation methods (Rubin and Schenker, 1986; Rubin, 2004), fully Bayesian inference such as Gibbs sampling (Rubin, 1976), and semiparametric methods (Zhao et al., 1996; Robins et al., 1994) have all been proposed to estimate the parameters of interest.

The ignorable missingness assumption is unlikely to hold in some scenarios. For example, in environmental studies, records of operation times in a power plant may be missing as time exceeds the standard amount. The missing data are said to be nonignorable if the absence of the data depends on the missing values. A fundamental challenge of nonignorable

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

missingness is that the full data distribution is not fully identifiable without making assumptions.

Various identification and estimation methods have been proposed for regression problems in which the covariates or the responses are subject to nonignorable missingness. Heckman (1979) proposed the Heckman selection model, which requires models for the outcome regression and the selection process. Rubin and Schenker (1986) and Glynn et al. (1993) considered imputation-based methods for nonignorable missing data. Baker and Laird (1988) used the expectation-maximization (EM) algorithm to obtain maximum likelihood estimates from contingency tables. Lipsitz and Ibrahim (1996) discussed nonignorable missing responses for general binomial regression models. Ibrahim et al. (1999) developed a method for generalized linear models (GLMs) with nonignorable missing covariates. Chen and Little (1999) and Herring and Ibrahim (2001) estimated regression parameters in proportional hazards regression models with nonignorable missingness. Roy and Lin (2002) developed methods for nonignorable dropouts in a linear mixed model. Zhao and Shao (2016) developed a pseudo-likelihood method that uses an instrumental variable to facilitate the identification when both the response and the covariates are subject to nonignorable missingness. Ma et al. (2003), D'Haultfoeuille (2010), Wang et al. (2014), Kott (2014),

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Miao et al. (2016), and Miao and Tchetgen Tchetgen (2016) discuss the identification and estimation of a nonignorably missing outcome using a shadow variable.

In causal inference problems, there is growing interest in recovering the causal effect in the presence of nonignorable missing confounders or outcomes when the missingness is nonignorable. Yang et al. (2014) used the EM algorithm and an instrumental variable to estimate the effect of high-level (neonatal intensive care units)NICU on the mortality of premature babies. They regarded the compliance type as a proxy for the unmeasured risk of complications, and allowed the missingness of the confounders to depend on the fully observed outcome and the partially observed compliance class. Ding and Geng (2014) discussed the identifiability of the causal effect under the assumption that the missingness indicator of the confounders is independent of the outcome, conditional on the treatment and the possibly missing confounder values. Additional discussions on this assumption can be found in D'Haultfoeuille (2010), Kott (2014), Zhao and Shao (2016), Miao et al. (2016), and Miao and Tchetgen Tchetgen (2016). Shao and Wang (2016) constructed a kernel-type semiparametric inverse probability weighting estimator when the outcome is nonignorably missing. Lu and Ashmead (2017) proposed a sensitivity analysis method that uses matching

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

estimator to assess the effect of the nonignorable missing confounder on the estimation of the treatment effect. Under a cluster-specific nonignorable treatment assignment assumption, Kim et al. (2017) proposed a calibrated propensity score estimator for the average causal effect.

In this study, we focus on a semiparametric estimation of the causal effect when the confounders are nonignorably missing. Under the outcome-independent missingness assumption, we develop three semiparametric estimators, namely the inverse probability weighting (IPW), regression, and doubly robust (DR) estimators. The IPW estimator requires a correct specification of the propensity scores for the treatment selection and missing-data mechanism. The regression estimator requires a correct specification of the joint model of the confounders and the outcome, conditional on the treatment in the fully observed sample. Assuming the selection bias odds ratio function is always correctly specified, the DR estimator uses both sets of models and is consistent when either set of models, but not necessarily both, is correctly specified. We apply these three estimators to estimate the causal effect of a scrubber installation on the reduction of SO_2 emissions.

The remainder of the paper is organized as follows. In Section 2, we introduce the notation and assumptions. We propose three semiparametric estimators in Section 3. We further illustrate our methods using simulation

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

studies in Section 4, and analyze real data in Section 5. The paper concludes with a discussion in Section 6.

2. Notation and Assumptions

Let X denote the random vector of all confounders. The values of the confounders X are subject to missingness. Let $R = 1$ if all components of X are completely observed, and $R = 0$ if some components of X are missing. Note that we consider only binary missing patterns, for simplicity. We extend our results to the multiple missing patterns setting in the Supplementary Material. Let $A = 1$ if the treatment is received and $A = 0$ otherwise. Let Y denote the outcome of interest. We consider the case where the confounders X are subject to missingness, while the treatment A and the outcome Y are fully observed. Let $O = (A, R, RX, Y)$ denote the observed data. Let a, r, x, and y denote the possible values that A, R, X, and Y, respectively, can take. Let Y_1 and Y_0 denote the potential outcomes, that is, the outcome under treatment, and the control, respectively. The parameter of interest is the treatment effect on the outcome, that is, $\Delta = E(Y_1 - Y_0) = \mu_1 - \mu_0$, where $\mu_a = E(Y_a)$ is the average potential outcome if $A = a$. An extension of the proposed methods to estimate the treatment effect on the treated is given in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

By the causal consistency assumption, we have $Y = AY_1 + (1 - A)Y_0$. Additionally, we make an ignorability assumption that the treatment A is conditionally independent of the potential outcomes, given the confounders X; that is, $A \perp\!\!\!\perp (Y_0, Y_1)|X$. That is, we assume there are no unmeasured confounders, but confounders can have missing values for some units.

With the confounders X having missing values, Ding and Geng (2014) showed that without any assumptions, the joint distribution of (A, Y, X) is not identifiable. For the purpose of identification, we make the following outcome-independent missingness assumption.

Assumption 1 (Outcome-independent missingness)**.** *Assume that, given the treatment A and the confounders X, the missingness indicator R is independent of the outcome Y; that is, $R \perp\!\!\!\perp Y|A, X$.*

This assumption is also termed the instrumental missingness assumption (Zhao and Shao, 2016) or the shadow variable assumption in Ma et al. (2003), D'Haultfoeuille (2010), Miao et al. (2018), and Miao and Tchetgen Tchetgen (2016). This outcome-independent missingness assumption is reasonable for the missingness of confounders, given that the confounders X are collected at the baseline and the outcome Y is observed at a later time or when the missingness of the confounders is by design (Ding and Geng, 2014). Under this assumption, identifying the joint distribution of (A, Y, X)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

has been well studied in Ding and Geng (2014), Yang et al. (2019), Miao and Tchetgen Tchetgen (2017), and Miao et al. (2018). In this paper, we assume the underlying data distribution always satisfies the identification conditions, and focus on performing a semiparametric inference, assuming the confounders satisfy the outcome-independent missingness assumption.

3. Semiparametric Inference

3.1 IPW Estimator

The classic IPW estimator was first proposed by Horvitz and Thompson (1952) in the survey sampling literature. The idea is to create a pseudo population by weighting each subject by the inverse of the conditional probability of receiving the treatment, given confounders, that is, $1/\Pr(A|X)$. In this pseudo population, the treatment A can be viewed as completely randomized and its association with the confounders X is removed. The propensity score $\Pr(a|x)$ is, in general, unknown; thus, it is typically replaced with an estimate in the weights.

However, the propensity score $\Pr(a|x)$ is not directly estimable when the confounders are subject to missingness. To address this, we propose an IPW estimator, that extends the classic IPW estimator to the setting with nonignorable missing confounders. We assume that the joint propensity

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

score of the treatment and the missingness indicator is positive, that is, $\Pr(a, r|x) > 0$ for all a, r, and x. The construction of the IPW estimator relies on the following representation of the average potential outcome, the derivation of which is presented in the Supplementary Material.

Lemma 1. $\mu_a = E\left\{1(A = a)RY/\Pr(A, R|X)\right\}$.

In order to construct the IPW estimator using the above representation, we need to obtain an estimate for $\Pr(a, r|x)$. Because the confounders X are not fully observed, the estimation of the propensity score $\Pr(a, r|x)$ requires special parameterization. Specifically, we use a semiparametric odds ratio representation for this joint distribution (Chen, 2007),

$$\Pr(a, r|x) = \frac{\psi(a, a_0 = 1, r, r_0 = 1|x)\Pr(r|a_0 = 1, x)\Pr(a|r_0 = 1, x)}{\sum_{r=0}^{1}\sum_{a=0}^{1}\psi(a, a_0 = 1, r, r_0 = 1|x)\Pr(r|a_0 = 1, x)\Pr(a|r_0 = 1, x)}, \tag{3.1}$$

and

$$\psi(a, a_0 = 1, r, r_0 = 1|x) = \frac{\Pr(r|a, x)\Pr(r_0 = 1|a_0 = 1, x)}{\Pr(r|a = 1, x)\Pr(r_0 = 1|a, x)}.$$

According to (3.1), to model $\Pr(a, r|x)$, we need only model $\Pr(a|r_0 = 1, x)$ and $\Pr(r|a, x)$.

Let α and γ denote the parameters in the models $\Pr(a|r = 1, x; \alpha)$ and $\Pr(r|a, x; \gamma)$, respectively. The maximal likelihood estimates (MLE) for α,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

denoted by $\hat{\alpha}$, can be obtained by fitting a regression of the treatment A on the confounders X for those individuals with fully observed confounders ($R = 1$). However, owing to the missingness of the confounders X, one cannot calculate an estimate of γ by fitting a regression model directly. Instead, we obtain an estimator for γ by solving the following estimation equation:

$$\mathbb{P}_n\left[\left\{\frac{R}{\Pr(R|A,X;\gamma)}-1\right\}l(A,Y)\right]=0, \tag{3.2}$$

where $l(A,Y)$ is any differentiable vectorized function of A and Y that satisfies the regularity condition 1 given in the Supplementary Material. The function l can be chosen based on the model posited on the propensity score $\Pr(r|a,x)$. For example, assuming $\operatorname{logit}\Pr(r|a,x;\gamma)=\gamma_1+\gamma_2 a+\gamma_3 x$, where x is a scalar, the function l can be chosen as $l(A,Y)=(1,A,Y)^T$. Thus, the number of estimating equations is the same as the dimension of γ. We could also allow the dimension of l to exceed that of γ, and apply the generalized method of moments (GMM) to estimate γ (Hall, 2005).

An important feature of the estimating equation (3.2) is that the missing confounders X are only involved in the propensity score $\Pr(r|a,x;\gamma)$, the inverse of which is multiplied by the missing indicator R. Thus, the confounders are only needed for individuals with $R=1$, that is, those who

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

have fully observed confounders. Therefore, equation (3.2) is estimable. Let $\hat{\gamma}^{ipw}$ denote an estimator of γ by solving (3.2). As shown in the Supplementary Material, the consistency of $\hat{\gamma}^{ipw}$ relies on the outcome-independent missingness assumption 1.

Thus, we construct an IPW estimator with estimated parameters $\hat{\alpha}$ and $\hat{\gamma}^{ipw}$ as follows:

$$\hat{\mu}_a^{ipw} = \mathbb{P}_n \left\{ \frac{1(A=a)RY}{\Pr(A, R|X; \hat{\alpha}, \hat{\gamma}^{ipw})} \right\},$$

where $\mathbb{P}_n$ denotes the empirical average; that is, $\mathbb{P}_n \nu(O) = \sum_{i=1}^n \nu(O_i)/n$, for any functions $\nu(O)$.

As mentioned, when there are no missing values for the confounders X, we can remove the confounding of the treatment selection by assigning each individual a weight, the weight is an estimate of $1/\Pr(A|X)$, the inverse of the propensity score of the treatment selection. However, because the confounders are only fully observed for individuals with $R = 1$, we restrict our estimator to a summation over those individuals with fully observed confounders. To account for this selection, we further weight these individuals using an estimate of $1/\Pr(R = 1|A, X)$, the inverse probability of observing all confounders. Hence, the weight for our IPW estimator is an estimate of $1/\Pr(A, R|X)$, the inverse of the joint propensity score of the treatment and the missing indicator.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

The following proposition provides the consistency of the IPW estimator.

Proposition 1. Under Assumption 1, suppose $\Pr(r|a,x;\gamma)$ and $\Pr(a|r=1,x;\alpha)$ are correctly specified. Then the IPW estimator $\hat{\mu}_a^{ipw}$ is consistent for μ_a, where $\hat{\alpha}$ is the MLE of α, and $\hat{\gamma}^{ipw}$ is obtained from the estimation equation (3.2).

The choice of l will generally affect the efficiency, but does not affect the consistency, as long as the identification conditions hold and the required models are correctly specified. The choice of l that leads to the most efficient IPW estimator can be derived using the results in Newey and McFadden (1994). We relegate the asymptotic normality and the variance of the IPW estimator to the Supplementary Material.

3.2 Regression

If the confounders X are fully observed, then by the ignorability assumption, we can estimate the causal effect by regressing Y on X and A and then marginalizing over X, because $\mu_a = \int_x E(Y_a|x)f(x)dx = \int_x E(Y|a,x)f(x)dx$. Here $f(x)$ is either the probability density function of X if X is continuous or the probability mass function if X is discrete, and the integral is the Riemann–Stieltjes integral. When the confounders X are subject to

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

outcome-independent missingness, we have $E(Y|a,x) = E(Y|a,x,r=1)$. However, even though we are able to evaluate the regression coefficients in the model $E(Y|a,x)$, we are still unable to evaluate $E(Y|a,x)$ among those with missing confounders, with the confounders distributed in the population according to $f(x)$. Thus, we cannot directly estimate the average potential outcome μ_a in the study population.

Therefore, let $g_a(x) = E(Y|a,x,r=1)$. Then we have the following representation of the average potential outcome

Lemma 2. $\mu_a = E\{g_a(X)\} = E[Rg_a(X) + (1-R)E\{g_a(X)|A,R=0\}]$.

The proof of Lemma 2 is given in the Supplementary Material. Thus, to construct a regression estimator, we need to obtain estimates for $g_a(x)$ and $E\{g_a(X)|a,r=0\}$.

Let β denote the parameter in the model $f(x,y|a,r=1;\beta)$. When $R=1$, the confounders X are fully observed, in which case the MLE $\hat{\beta}$ for the parameter β can be estimated directly. Then, we can estimate $g_a(x)$ using $g_a(x;\hat{\beta}) = E(Y|a,x,r=1;\hat{\beta})$.

To obtain an estimate for $E\{g_a(X)|a,r=0\}$, we first estimate $f(x,y|a,r=0)$. Again, this is not straightforward, because the confounders X are subject to missingness when $R=0$. We use the following representation for $f(x,y|a,r=0)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Lemma 3.

$$f(x, y|a, r = 0) = \frac{\eta(r = 0, r_0 = 1, x, x_0 = 0|a) f(x, y|a, r = 1)}{E\{\eta(r = 0, r_0 = 1, X, x_0|a)|a, r = 1\}},$$

where

$$\eta(r, r_0, x, x_0|a) = \frac{\Pr(r|a, x) \Pr(r_0|a, x_0)}{\Pr(r_0|a, x) \Pr(r|a, x_0)}.$$

The derivation of Lemma 3 is given in the Supplementary Material. If the missingness of the confounders X is ignorable, then the missingness mechanism $\Pr(r|a, x)$ does not depend on x. In that case, $\eta(r, r_0, x, x_0|a) = 1$. Thus, the function $\eta(r, r_0, x, x_0|a)$ can be viewed as a selection bias function, because its deviation from one indicates that the missing mechanism of the confounders X is nonignorable.

We can parameterize the above representation of $f(x, y|a, r = 0)$ into two parts: $\eta(r, r_0, x, x_0|a)$ and $f(x, y|a, r = 1)$. Let ξ denote the parameter in the model $\eta(r, r_0, x, x_0|a; \xi)$. The estimation of ξ cannot be obtained directly, owing to the missingness of X. Therefore, for any function $l(A, Y)$, we have the following representation for $E\{l(a, Y)|a, r = 0\}$.

Lemma 4.

$$E\{l(a, Y)|a, r = 0\} = \frac{E\{\eta(r = 0, r_0, X, x_0|a) l(a, Y)|a, r = 1\}}{E\{\eta(r = 0, r_0, X, x_0|a)|a, r = 1\}}.$$

The derivation of Lemma 4 is given in the Supplementary Material. Note that on the right-hand side, the possibly missing confounders X are only

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

involved in the subpopulation where $R = 1$. Thus, $E\{l(a, Y)|a, r = 0\}$ can be evaluated using the observed data. Furthermore, the expectation can be evaluated numerically if it does not have a closed form. Under such a representation, we have the following estimation equation for ξ:

$$\mathbb{P}_n\left[(1-R)\left\{l(A,Y) - E\{l(A,Y)|A, R=0; \hat{\beta}, \xi\}\right\}\right] = 0, \tag{3.3}$$

where $\hat{\beta}$ is the MLE of β, and $l(A, Y)$ is a vectorized arbitrary differentiable function of A and Y that satisfies regularity condition 2 in the Supplementary Material. Once we obtain an estimate $\hat{\xi}^{reg}$ for ξ, we can estimate $E\{g_a(X)|a, r = 0\}$ using $E\{g_a(X; \hat{\beta})|a, r = 0; \hat{\beta}, \hat{\xi}^{reg}\} = \int g_a(x; \hat{\beta}) f(x|a, r = 0; \hat{\beta}, \hat{\xi}^{reg})dx$, where $g_a(x; \hat{\beta}) = E(Y|a, x, r = 1; \hat{\beta})$.

Thus, we construct the regression estimator as

$$\hat{\mu}_a^{reg} = \mathbb{P}_n\left[(1-R)E\{g_a(X; \hat{\beta})|a, r = 0; \hat{\beta}, \hat{\xi}^{reg}\} + Rg_a(X; \hat{\beta})\right]. \tag{3.4}$$

The regression estimator $\hat{\mu}_a^{reg}$ circumvents the problem of evaluating the distribution of the confounders by evaluating $g_a(x)$ and $E\{g_a(X)|a, r = 0\}$ instead. The expected values of $g_a(X)$ among individuals with and without missing confounders together form the population average of $g_a(X)$, which is the same as the mean potential outcome μ_a. The consistency of a regression estimator for μ_a is given in the following proposition.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Proposition 2. Under Assumption 1, assume $\eta(r = 0, r_0, x, x_0|a; \xi)$ and $f(x, y|a, r = 1; \beta)$ are correctly specified. Then the regression estimator $\hat{\mu}_a^{reg}$ given in (3.4) is consistent for μ_a, where $\hat{\beta}$ is the MLE of β and $\hat{\xi}^{reg}$ is obtained by (3.3).

The asymptotic normality and variance of the regression estimator can be derived similarly to those of the IPW estimator. We omit the derivation here, for simplicity.

3.3 DR

The DR estimator was first proposed by Robins et al. (1994) in the form of an augmented IPW estimator for regression coefficients with covariates missing at random. Further discussion on the existing DR estimator can be found in Lunceford and Davidian (2004), Carpenter et al. (2006), Leon et al. (2003), Davidian et al. (2005), Bang and Robins (2005), and Kang and Schafer (2007). Typically, a DR estimator involves a propensity score model and an outcome and a confounders regression model and is consistent when either model is correctly specified. It is also known to achieve the semiparametric efficiency bound when both models are correctly specified. This class of estimators also appears in the survey sampling literature, where they are referred to as model-assisted survey estimators (Särndal

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

et al., 2003).

If the confounders are missing at random, the specifications of the propensity scores and the regression models for the confounders and the outcome are independent. However, owing to the presence of nonignorable missing confounders X, the parameters in the propensity score of the missingness and the joint distribution model of X and Y, conditional on $A, R = 1$, are not variationally independent. To see this, we have the following representation of the propensity score for the missingness, the proof of which follows directly from the definition.

Lemma 5.

$$\Pr(r = 1|a, x) = \frac{\Pr(r = 1|a, x = 0)}{\Pr(r = 1|a, x = 0) + \eta(r = 0, r_0, x, x_0|a) \Pr(r = 0|a, x = 0)}.$$

Let δ denote the parameter in the model for $\Pr(r = 1|a, x = 0; \delta)$. Hence, we can parameterize $\Pr(r = 1|a, x; \gamma)$ into $\Pr(r = 1|a, x = 0; \delta)$ and $\eta(r = 0, r_0, x, x_0|a; \xi)$, where $\gamma = (\delta, \xi)$. Recall that the parameterization of $f(x, y|a, r = 0; \beta, \xi)$ in Section 3.2 also involves $\eta(r = 0, r_0, x, x_0|a; \xi)$. Thus, $\eta(r = 0, r_0, x, x_0|a; \xi)$ lies in the intersection of the propensity score model $\Pr(r = 1|a, x; \gamma)$ and the regression model $f(x, y|a, r = 0; \beta, \xi)$.

To construct a DR estimator, we assume that the functional form of $\eta(r = 0, r_0, x, x_0|a; \xi)$ is always correctly specified, with an unknown parameter ξ. We propose a DR estimator for μ_a in the sense that if either

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

set of baseline models (i) $\Pr(r=1|a,x=0;\delta)$ and $\Pr(a|r=1,x;\alpha)$ or (ii) $f(x,y|a,r=1;\beta)$ is correctly specified, the DR estimator is consistent for μ_a.

The key to constructing a DR estimator for μ_a is to first do so for the parameter ξ in the selection bias function $\eta(r=0,r_0,x,x_0|a;\xi)$ when either $\Pr(r=1|a,x=0;\delta)$ or $f(x,y|a,r=1;\beta)$ is correctly specified. We can obtain an estimator $\hat{\delta}$ for δ and a DR estimator $\hat{\xi}^{dr}$ for ξ by solving the following estimating equation:

$$\mathbb{P}_n\left[\left\{\frac{R}{\Pr(R|A,X;\delta,\xi)}-1\right\}\left\{l(A,Y)-E\{l(A,Y)|A,R=0;\hat{\beta},\xi\}\right\}\right]=0, \tag{3.5}$$

where $l(A,Y)$ is an arbitrary vectorized differentiable function that satisfies regularity condition 3 in the Supplementary Material. We can evaluate $E\{l(a,Y)|a,r=0\}$ using Lemma 4. The estimating equation (3.5) resembles (3.2) for the parameters in the propensity score of the IPW estimator, and resembles (3.3) for the parameters in the regression of the regression estimator. Equation (3.5) differs from (3.2) in that (3.5) replaces $l(A,Y)$ in (3.2) with a centered function $l(A,Y)-E\{l(A,Y)|A,R=0;\hat{\beta},\hat{\xi}^{dr}\}$. Equation (3.5) differs from (3.3) in that (3.5) involves the propensity score of the missingness in the weight. This centering and weighting achieve the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

DR estimation of the selection bias parameter ξ. Similarly to estimating equations (3.2) and (3.3), the confounders X are only involved in the individuals when $R = 1$. Thus, equation (3.5) can be evaluated using the observed data. A detailed proof of the DR property of the estimator $\hat{\xi}^{dr}$ is given in the Supplementary Material.

Hence, the DR estimator for μ can be constructed as

$$\hat{\mu}_a^{dr} = \mathbb{P}_n\Bigg[\frac{R}{\Pr(R|A,X;\hat{\delta},\hat{\xi}^{dr})}\Big\{\hat{h}_a(A,X,Y) - E\{\hat{h}_a(A,X,Y)|A,R=0;\hat{\beta},\hat{\xi}^{dr}\}\Big\} + E\{\hat{h}_a(A,X,Y)|A,R=0;\hat{\beta},\hat{\xi}^{dr}\}\Bigg], \tag{3.6}$$

where $\hat{h}_a(A,X,Y) = 1(A=a)\{Y - g_a(X;\hat{\beta})\}/\Pr(A|X;\hat{\alpha},\hat{\delta},\hat{\xi}^{dr}) + g_a(X;\hat{\beta})$, and

$$E\{\hat{h}_a(A,X,Y)|A,R=0;\hat{\beta},\hat{\xi}^{dr}\} = \frac{E\big\{R\eta(r=0,r_0,X,x_0|A)\hat{h}_a(A,X,Y)|A;\hat{\beta},\hat{\xi}^{dr}\big\}}{E\big\{R\eta(r=0,r_0,X,x_0|A)|A;\hat{\beta},\hat{\xi}^{dr}\big\}}.$$

Here, we give an intuitive illustration of the DR property of $\hat{\mu}_a^{dr}$. A more rigorous proof can be found in the Supplementary Material. Note that

$$\begin{aligned}\hat{\mu}_a^{dr} &= \hat{\mu}_a^{ipw} + \mathbb{P}_n\Bigg[\frac{R_i}{\Pr(R_i|A_i,X_i;\hat{\delta},\hat{\xi}^{dr})}\Big\{1 - \frac{1(A_i=a)}{\Pr(A_i|X_i;\hat{\alpha},\hat{\delta},\hat{\xi}^{dr})}\Big\}g_a(X_i;\hat{\beta})\Bigg] \\ &\quad + \mathbb{P}_n\Bigg[\Big\{1 - \frac{R_i}{\Pr(R_i|A_i,X_i;\hat{\delta},\hat{\xi}^{dr})}\Big\}E\{\hat{h}_a(X_i,Y_i)|A_i,R_i=0;\hat{\delta},\hat{\xi}^{dr}\}\Bigg].\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

If the baseline propensity scores are correctly specified, but $f(X, Y|A, R = 1)$ may be misspecified (i.e., (i) holds), then the estimator

converges in probability to

$$E\left[\hat{\mu}_a^{ipw} + \frac{R_i}{\Pr(R_i|A_i, X_i)}\left\{1 - \frac{1(A_i = a)}{\Pr(A_i|X_i)}\right\} g_a(X_i; \beta^\dagger)\right]$$
$$+E\left[\left\{1 - \frac{R_i}{\Pr(R_i|A_i, X_i)}\right\} E\{h_a^\dagger(A_i, X_i, Y_i)|A_i, R_i = 0\}\right],$$

where $h_a^\dagger(A, X, Y) = 1(A = a)\{Y - g_a(X; \beta^\dagger)\}/\Pr(A|X) + g_a(X; \beta^\dagger)$, and $\beta^\dagger$ is the limit of $\hat{\beta}$ under the possibly misspecified model. The expectation of the first term in the above equation is equal to μ_a, and that of the second and third terms both equal zero. Thus, the DR estimator is consistent. If the baseline outcome model $f(X, Y|A, R = 1)$ is correctly specified, but the baseline propensity scores may be misspecified(i.e., (ii) holds), then it can be shown that $\hat{\mu}_a^{dr}$ converges to

$$E[Rh_a^*(A, X, Y) + (1 - R)E\{h_a^*(A, X, Y)|A, R = 0\}],$$

where $h_a^*(A, X, Y) = 1(A = a)\{Y - g_a(X)\}/\Pr(A|X; \alpha^*, \delta^*) + g_a(X)$, α^* and δ^* denote the limits of $\hat{\alpha}$ and $\hat{\delta}$, respectively, in the misspecified propensity score models, and the correctly estimated parameter β and ξ are omitted. The right-hand side resembles the expectation of the regression estimation given in Lemma 2, to that with $g_a(X)$ replaced with $h_a^*(A, X, Y)$. Follow-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

ing a similar derivation as the one for the regression estimator, we have $E[Rh_a^*(A,X,Y) + (1-R)E\{h_a^*(A,X,Y)|A,R=0\}] = E\{h_a^*(A,X,Y)\}$. When (ii) holds, we have $E\{h_a^*(A,X,Y)\} = E\{g_a(X)\} = \mu_a$. Thus, the estimator $\hat{\mu}_a^{dr}$ is doubly robust.

The DR properties of $\hat{\xi}^{dr}$ and $\hat{\mu}_a^{dr}$ are given in the proposition below. The asymptotic normality and variance of the DR estimator $\hat{\mu}_a^{dr}$ can be derived similarly to those of the IPW estimator.

Proposition 3. Under Assumption 1 and assuming $\eta(r=0,r_0,x,x_0|a;\xi)$ is correctly specified, if (i) $\Pr(r=1|a,x=0;\delta)$ and $\Pr(a|r=1,x;\alpha)$ or (ii) $f(x,y|a,r=1;\beta)$ is correctly specified, then $\hat{\xi}^{dr}$ and $\hat{\mu}_a^{dr}$ are consistent for ξ and μ_a, respectively, where $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β, respectively, and $\hat{\xi}^{dr}$ and $\hat{\delta}$ are obtained by (3.5) and $\hat{\mu}_a^{dr}$ is given in (3.6).

As we show in the Supplementary Material, a correct specification of $\Pr(a|r=1,x;\alpha)$ is not necessarily needed for the DR property of $\hat{\xi}^{dr}$, but it is indispensable for the DR property of $\hat{\mu}_a^{dr}$.

Similarly to the IPW and the regression estimators, the user-specified function $l(A,Y)$ in (3.5) affects the efficiency, but not the consistency of the estimators $\hat{\xi}^{dr}$ and $\hat{\mu}_a^{dr}$. The choice of $l(A,Y)$ that leads to the most efficient estimators $\hat{\xi}^{dr}$ and $\hat{\mu}_a^{dr}$ can be derived following Newey and McFadden (1994). Under such a choice of $l(A,Y)$, the estimator $\hat{\mu}_a^{dr}$ achieves the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

semiparametric efficiency bound. Similarly to the argument in Miao et al. (2018), the semiparametric efficient estimator for μ_a only has a closed-form solution when all the variables are binary; thus, in practice, we do not recommend using it over the DR estimator $\hat{\mu}_a^{dr}$ proposed here.

4. Simulations

In this section, we study the finite-sample performance of the three proposed semiparametric methods using simulations. We consider four scenarios: (a) the baseline propensity models $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ and the baseline regression model $f(x, y|a, r = 1; \beta)$ are all correctly specified; (b) the baseline propensity scores $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ are correctly specified, but the regression model $f(x, y|a, r = 1; \beta)$ is misspecified; (c) the baseline regression model $f(x, y|a, r = 1; \beta)$ is correctly specified, but $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ are both misspecified; (d) neither the baseline propensity scores nor the baseline regression model are correctly specified. For scenario (a), we carried out the simulations in the following steps.

First, we generated a population of size $n = 4000$. For each individual, covariates X_1 and X_2 were generated independently from Bernoulli distributions with probability 0.3 and 0.8, respectively. We assumed logistic

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

models for the baseline propensity scores as $\text{logit}\Pr(a=1|r=1,x)=0.8+1.6x_1-0.4x_2$ and $\text{logit}\Pr(r=1|a,x)=0.5+1.2a+1.2x_1-0.4x_2$. The joint propensity score $\Pr(a,r|x)$ was then calculated using (3.1). The missingness indicator R and the treatment A were generated from $\Pr(a,r|x)$. Finally, we generated the outcome Y from the normal distribution $f(y|x_1,x_2,a)=\phi(0.5+2x_1-3x_2+0.5x_1x_2+3a,0.01)$, where $\phi(\mu,\sigma^2)$ denotes the normal density function with mean μ and variance σ^2.

Second, the propensity score models were correctly specified as $\text{logit}\Pr(a=1|r=1,x;\alpha)=\alpha_1+\alpha_2x_1+\alpha_3x_2$ and $\text{logit}\Pr(r=1|a,x;\gamma)=\delta_1+\delta_2a+\xi_1x_1+\xi_2x_2$. The parameter $\alpha=(\alpha_1,\alpha_2,\alpha_3)$ was estimated using the MLE, and $\gamma=(\delta_1,\delta_2,\xi_1,\xi_2)$ was estimated using equation (3.2) with $l(A,Y)=\{1,A,AY,A\sin(AY)\}$. Then, the IPW estimator $\hat{\mu}_a^{ipw}$ was calculated.

Third, the outcome regression model was correctly specified as $f(y|x,a,r=1;\beta)=\phi(\beta_1+\beta_2x_1+\beta_3x_2+\beta_4x_1x_2+\beta_5a,\beta_6^2)$. In addition, $\Pr(x|a,r=1;\beta)$ was correctly specified using a saturated model $\Pr(x|a,r=1;\beta)=\beta_7+\beta_8x_1+\beta_9x_2+\beta_{10}a+\beta_{11}x_1a+\beta_{12}x_2a$, and, we correctly specified $\eta(r=0,r_0,x,x_0;\xi)$ as $\eta(r=0,r_0,x,x_0;\xi)=\exp(\xi_1x_1+\xi_2x_2)$. The parameter β was estimated using the MLE, and ξ was estimated using equation (3.3) with $l(A;Y)=(AY,AY^2)$. Then, the outcome regression estimator

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

$\hat{\mu}_a^{reg}$ was calculated.

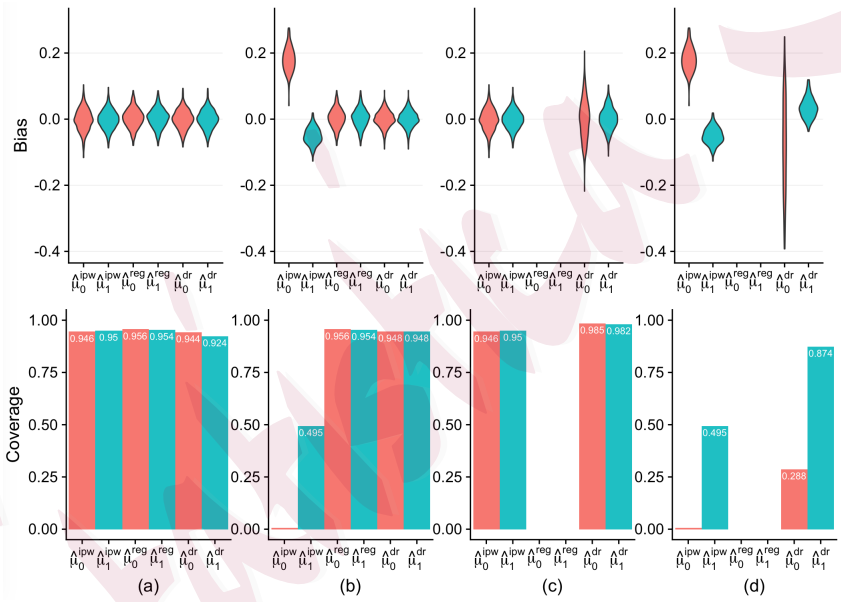
Fourth, the models $\Pr(a = 1|r = 1, x; \alpha)$, $f(y|x, a, r; \beta)$, $\Pr(x|a, r = 1; \beta)$, $\Pr(r = 1|a, x = 0; \delta)$, and $\eta(r = 0, r_0, x, x_0; \xi)$ were correctly specified, as before, and the estimate for the parameter $\gamma = (\delta, \xi)$ was obtained using equation (3.5) with $l(A, Y) = (Y, AY, Y^2, AY^2)$. Then, the DR estimator $\hat{\mu}_a^{dr}$ was calculated. We repeated the above procedures 1000 times.

The model for $f(y|x, a, r = 1; \beta)$ was correctly specified owing to the outcome-independent missingness assumption $R \perp\!\!\!\perp Y|A, X$. The simulations in scenario (b) are similar to those in scenario (a), except the outcome model $f(y|x, a, r = 1)$ was misspecified as an exponential distribution $f(y|x, a, r = 1; \beta) = |\beta_0 + \beta_1 x_1| \mathrm{e}^{-|\beta_0 + \beta_1 x_1| y}$. Note that we still correctly specified the model for $\Pr(x|a, r = 1)$ in this scenario. However, $f(x, y|a, r)$ was misspecified owing to the misspecification of $f(y|x, a, r = 1)$. The simulations in scenario (c) are similar to those in scenario (a), except that the propensity score models $\Pr(r = 1|a, x = 0)$ and $\Pr(a = 1|r = 1, x)$ were misspecified as $\operatorname{logit} \Pr(r = 1|a, x = 0) = \gamma_1$ and $\operatorname{logit} \Pr(a = 1|r = 1, x) = \alpha_1$, respectively. For scenario (d), both the propensity score and the outcome models in scenario (a) were replaced by the misspecified models above.

The biases and empirical coverages of the Wald-type 95% confidence intervals (CIs) for the IPW, regression and DR estimators for μ_a are pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Figure 1: Bias, coverage of the 95% Wald type confidence intervals of the IPW, regression, and DR estimators for μ_a when (a) both the baseline propensity and the outcome models are correctly specified, (b) only the baseline outcome model is correctly specified, (c) only the baseline propensity score models are correctly specified, and (d) none of the baseline models are correctly specified.



sented in Figure 1. The CIs were constructed using the estimated variance proposed in Section 3. We kept the ranges of the biases shown in Figure 1 the same for all scenarios; thus, seriously biased estimators are not shown in this figure. In scenario (a), because all the baseline models are correctly specified, the IPW, regression, and DR estimators all perform well in terms

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

of having small biases (the biases are <0.001, 0.002, and 0.001 for the three estimators, respectively, for μ_0), as well as having the CIs achieve close to the nominal levels (between 92% to 95%). In scenarios (b) and (c), neither the IPW nor the regression estimator performs well when the corresponding models are misspecified. For example, when the propensity score models are misspecified, the bias of the IPW estimator $\hat{\mu}_0^{ipw}$ is 0.18, and when the outcome model is misspecified, the bias for the regression estimator $\hat{\mu}_0^{reg}$ is 0.59. The DR estimator $\hat{\mu}_0^{dr}$ still showed relative small biases ($< 2e^{-3}$) in both cases. The coverage of the DR estimator remained at the nominal levels, while that of the IPW and the regression estimators dropped significantly when the corresponding models were misspecified (e.g., the coverage is 49.5% for the IPW estimator $\hat{\mu}_1^{ipw}$, zero for the regression estimator $\hat{\mu}_1^{reg}$, and 94.8% and 98.2% for the DR estimators $\hat{\mu}_1^{dr}$ in scenarios (b) and (c), respectively). In scenario (d), when the baseline models are all misspecified, the IPW and regression estimators showed relatively large biases, as mentioned before, as in scenarios (b) and (c). The DR estimator exhibited a bigger bias in scenario (d) than that in scenarios (a)–(c), but its bias was much smaller than, or at least comparable with, the other two estimators in scenario (d) (e.g., the bias for $\hat{\mu}_1^{dr}$ is 0.03).

The biases and empirical coverages of the Wald-type 95% CIs for the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

IPW, regression and DR estimators for the parameters ξ in the selection bias function $\eta(r = 0, r_0, x, x_0; \xi)$ are presented in Figure 2 in the Supplementary Material. Again, when an estimator is severely biased, the results are beyond the ranges shown in Figure 2. The results were similar to those for the estimators of μ_a, demonstrating the DR property of $\hat{\xi}^{dr}$.

5. Application

In this section, we further illustrate the proposed semiparametric estimators by means of an application to the sulfur dioxide (SO_2) emissions data. SO_2 is a toxic gas, and even short-term exposure harms the human respiratory system. In 1990, the Acid Rain Program was launched to reduce ambient PM2.5 (atmospheric particulate matter (PM) with a diameter of less than 2.5 μm) by limiting the emissions of multiple pollutants (SO_2, NO_x, and CO_2). The reduction was achieved mostly by cutting emissions from coal-fired electricity-generating units(EGUs), that is, by installing flue-gas desulfurization equipment ("scrubbers"). Thus, it is of interest to evaluate the causal effect of scrubber technologies on the reduction of SO_2 emissions.

Monthly emissions data for 2004 were collected from the emissions monitors on 258 coal-fired power plants (Zigler et al., 2016). A power-generating facility may consist of multiple EGUs and scrubbers are installed on the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

EGUs. The data set contains 223 EGUs with scrubbers installed, and 812 EGUs without scrubbers installed. Three important baseline characteristics of the EGUs are included in the data set: operation time, heat input rate of every unit, and coal sulfur content. Because these factors affect both the installation of the scrubbers and the emissions of SO_2, they serve as the confounders of the causal relationship between the treatment and the outcome.

Let X_1 denote the operation time, X_2 denote the sulfur content of each ton of coal, and X_3 denote the heat input rate (mmBtu/hr) at the baseline. A valid causal inference of the effect of the scrubbers on SO_2 emissions hinges on the missingness of the confounders. Here 3.35% of the data are missing for operation time, 0.78% are missing for the heat input rate of every unit, and 8.02% are missing for the coal sulfur content, with a total of about 9% missingness for at least one of the confounders. We delete observations with extreme weights because they make the estimators very unstable. These observations account for less than 3% of the data. A naive analysis, where we calculate the difference between the average SO_2 emissions among the EGUs with and without the scrubber installation, indicates that the scrubbers reduce SO_2 emissions by 472.35 tons, (standard error (SE) =14.66 and p-value < 0.001). Because the scrubbers were not

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

randomly assigned, this naive estimate cannot be interpreted as the causal effect of the scrubber installation on reducing the pollutant. To account for repeated measurements from the same EGUs, we assume a linear mixed effects model for the outcome Y as $f(y|x, a, r = 1; \beta) = \phi(\beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 a, \beta_6^2)$. We assume the confounders X_1, X_2, and X_3 are independent with each other, and conditional on A and $R = 1$, X_j follows a normal distribution $N(\beta_{7,j} + \beta_{8,j} a, \beta_{9,j}^2)$, for $j = 1, 2, 3$. We also assume that the missingness mechanism and the propensity scores of the scrubber installation $\Pr(r = 1|a, x)$ follow logistic regression models $\text{logit}\Pr(r = 1|a, x; \gamma) = \delta_1 + \delta_2 a + \xi_1 x_1 + \xi_2 x_2 + \xi_3 x_3$ and $\text{logit}\Pr(a = 1|r = 1, x; \alpha) = \alpha_1 + \alpha_2 x_1 + \alpha_3 x_2 + \alpha_4 x_3$, respectively. The parameters ares estimated as in Section 3.

From the models specified above, the selection bias function $\eta(r = 0, r_0, X, x_0; \xi)$ has the form $\eta(r = 0, r_0, X, x_0; \xi) = \exp(\xi_1 x_1 + \xi_2 x_2 + \xi_3 x_3)$. The IPW, regression and DR estimators for the parameters ξ in the selection bias function $\eta(r = 0, r_0, X, x_0; \xi)$ are $\hat{\xi}^{ipw} = (1.42, 1.00, 0.69)$ (SE=(0.26,3.17,1.46), p-value=(4.72×10^{-8},0.57,0.64)), $\hat{\xi}^{reg} = (1.85, -4.49, 8.45)$ (SE=(0.39,0.44,1.27), p-value= (2.09×10^{-6}, 1.9×10^{-24}, 2.86×10^{-11})), and $\hat{\xi}^{DR} = (4.40, -1.00, 1.76)$ (SE=(5.43,0.41,7.20), p-value = (0.42, 0.01, 0.81)). Hence, at least one component of ξ is estimated to be significantly away from zero for all three

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

estimators, indicating that it is less likely that the missingness of the confounders is ignorable. For the causal effect of interest, the IPW, regression and DR estimators suggest 594.60 (SE=26.34, p-value < 0.001), 648.99 (SE=4.85, p-value < 0.001), and 511.28 (SE=32.45, p-value < 0.001) tons of reduction, respectively, of SO_2 emissions after the installation of the scrubbers. Detailed parameter estimates and SEs are given in Table 2 in the Supplementary Material. All three estimators suggest that scrubbers are effective in reducing SO_2 emissions, and that this reduction is estimated to have a larger magnitude for all three estimates than that of the naive estimator.

6. Conclusion

We have proposed three semiparametric estimators for estimating the a causal effect in the presence of nonignorably missing confounders.

The proposed semiparameteric estimators are closely related to the estimators proposed in Sun et al. (2018) and Miao et al. (2018). The main difference is that our estimators are proposed under a causal inference setting with nonignorably missing confounders, which makes the construction and calculation of the estimators more complicated.

As pointed out by one reviewer, the stable unit treatment value assump-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

tion may be violated when a monitor is affected by the emission of several EGUs, that is, interference may be present. Under interference, the estimands and estimators need to be redefined to reflect the cohort effect and the individual effect. For example, Tchetgen Tchetgen and VanderWeele (2012) and Liu et al. (2019) evaluate the causal effects under a hypothetical randomization and proposed estimators for direct, indirect, total, and overall effects. Furthermore, additional detailed geographical information, such as population densities, GPS locations, and land use, and weather information, need to be collected to carry out further adjustments. Here, we use SO_2 emission data to illustrate our proposed method. We leave extensions of our method to interference to future research.

We have assumes that the treatment and outcomes are fully observed. Thus, constructing an efficient semiparametric estimator with missing treatment and outcome values and missing confounder values is also left to future studies.

Supplementary Material

Section S1 contains the proof of Lemma 1. Section S2 provides the proof of Proposition 1. Section S3 presents the asymptotic normality and variance for the IPW estimator. Section S4 presents the proof of Lemma 2. Section S5 shows the derivation of the equations for the regression estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Section S6 shows the proof of Proposition 2. Section S7 presents the proof of Proposition 3. Sections S8 and S9 extend the Lemmas, Propositions, and semiparametric estimators to the multiple missing patterns setting and to the average treatment effect on the treated as the parameter of interest setting. Section S10 includes additional tables and figures.

Acknowledgments

We would like to thank Prof. Chanmin Kim for providing us with the data set used in this study and valuable instructions on the data. We also thank the associate editor, and two reviewers for their helpful comments and suggestions. Liu's research is partially supported by NSF DMS 1916013 and NIH grant U24 DK 060990.

References

Baker, S. and Laird, N. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical association*, 83:62–69.

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

Carpenter, J., Kenward, M., and Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Statistical Society: Series A (Statistics in Society), 169:571–584.

Chen, H. (2007). A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421.

Chen, H. and Little, R. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94:896–908.

Davidian, M., Tsiatis, A., and Leon, S. (2005). Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20:261.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

D'Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154:1–15.

Ding, P. and Geng, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Statistics in Medicine*, 33:1121–1133.

Glynn, R., Laird, N., and Rubin, D. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association*, 88:984–993.

Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press Oxford.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47:53–161.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Herring, A. and Ibrahim, J. (2001). Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47:663–685.

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

Ibrahim, J., Lipsitz, S., and Chen, M. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61:173–190.

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

Kim, G.-S., Paik, M. C., and Kim, H. (2017). Causal inference with observational data under cluster-specific non-ignorable assignment mechanism. *Computational Statistics & Data Analysis*, 113:88–99.

Kott, P. S. (2014). Calibration weighting when model and calibration variables can differ. In *Contributions to Sampling Statistics*, pages 1–18. Springer.

Leon, S., Tsiatis, A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59:1046–1055.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Lipsitz, S. and Ibrahim, J. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83:916–922.

Little, R. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association*, 87:1227–1237.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd ed.

Liu, L., Hudgens, M. G., Saul, B., Clemens, J. D., Ali, M., and Emch, M. E. (2019). Doubly robust estimation in observational studies with partial interference. *Stat*, 8(1):e214.

Lu, B. and Ashmead, R. (2017). Propensity score matching analysis for causal effects with MNAR covariates. *Statistica Sinica, in press*.

Lunceford, J. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23:2937–2960.

Ma, W., Geng, Z., and Hu, Y. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *Journal of multivariate analysis*, 87:24–45.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. (2016). Identifying causal effects with shadow variables of an unmeasured confounder. *arXiv preprint arXiv:1609.08816*.

Miao, W. and Tchetgen Tchetgen, E. (2016). On varieties of doubly robust estimators under

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

missingness not at random with a shadow variable. *Biometrika*, 103:475–482.

Miao, W. and Tchetgen Tchetgen, E. (2017). Identification and inference with nonignorable missing covariate data. *Statistica Sinica, in press*.

Miao, W., Tchetgen Tchetgen, E., and Geng, Z. (2018+). Identification and doubly robust estimation of data missing not at random with a shadow variable. *technical report*.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89:846–866.

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association*, 97:40–52.

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

Rubin, D. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81:366–374.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

Särndal, C., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer.

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

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., and Tchetgen Tchetgen, E. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28:1965–1983.

Tchetgen Tchetgen, E. and VanderWeele, T. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research*, 21:55–75.

Wang, S., Shao, J., and Kim, J. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24:1097–1116.

Yang, F., Lorch, S., Small, D., et al. (2014). Estimation of causal effects using instrumental variables with nonignorable missing covariates: Application to effect of type of delivery NICU on premature infants. *The Annals of Applied Statistics*, 8:48–73.

Yang, S., Wang, L., and Ding, P. (2019). Causal inference with confounders missing not at random. *Biometrika*, 106:875–888.

Zhao, J. and Shao, J. (2016). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110:1577–1590.

Zhao, L., Lipsitz, S., and Lew, D. (1996). Regression analysis with missing covariate data using

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0466

estimating equations. *Biometrics*, 52:1165–1182.

Zigler, C.and Kim, C., Choirat, C., Hansen, J., Wang, Y., Hund, L., Samet, J., King, G., and Dominici, F. (2016). Causal inference methods for estimating long-term health effects of air quality regulations. *Research report (Health Effects Institute)*, 187:5–49.

Zhaohan Sun, Department of Statistics and Actuarial Science, University of Waterloo

E-mail: (z227sun@uwaterloo.ca)

Lan Liu, School of Statistics, University of Minnesota

E-mail: (liux3771@umn.edu)