Statistica Sinica **28** (2018), 1965-1983
doi:https://doi.org/10.5705/ss.202016.0324

SEMIPARAMETRIC ESTIMATION WITH DATA MISSING NOT AT RANDOM USING AN INSTRUMENTAL VARIABLE

BaoLuo Sun[1], Lan Liu[2], Wang Miao[3], Kathleen Wirth[1], James Robins[1] and Eric J. Tchetgen Tchetgen[1]

[1]*Harvard,* [2]*University of Minnesota at Twin Cities and* [3]*Peking University*

Abstract: Missing data occur frequently in empirical studies in the health and social sciences, and can compromise our ability to obtain valid inference. An outcome is said to be missing not at random (MNAR) if, conditional on the observed variables, the missing data mechanism still depends on the unobserved outcome. In such settings, identification is generally not possible without imposing additional assumptions. Identification is sometimes possible, however, if an instrumental variable (IV) is observed for all subjects that satisfies the exclusion restriction that the IV affects the missingness process without directly influencing the outcome. In this paper, we provide necessary and sufficient conditions for nonparametric identification of the full data distribution under MNAR with the aid of an IV. In addition, we give sufficient identification conditions that are more straightforward to verify in practice. For inference, we focus on estimation of a population outcome mean, for which we develop a suite of semiparametric estimators that extend methods previously developed for data missing at random. Specifically, we propose a novel doubly robust estimator of the mean of an outcome subject to MNAR. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer characteristics such as gender, age and years of experience as IVs.

Key words and phrases: Doubly robust, instrumental variable, inverse probability weighting, missing not at random.

1. Introduction

Selection bias is a major problem in the health and social sciences, and is said to be present in an empirical study if features of the underlying population of primary interest are entangled with features of the selection process not of scientific interest. Selection bias can occur in practice due to incomplete data if the observed sample is not representative of the underlying population. While various ad hoc methods exist to adjust for missing data, such methods may be subject to

bias unless under fairly strong assumptions. For example, complete-case analysis is easy to implement and is routinely used in practice. However, complete-case analysis can be biased when the outcome is not missing completely at random (MCAR) (Little and Rubin (2002)). Progress can still be made if data are missing at random (MAR), such that the missing data mechanism is independent of unobserved variables conditional on observed variables. Principled methods for handling MAR data abound, including likelihood-based procedures (Little and Rubin (2002); Horton and Laird (1998)), multiple imputation (Rubin (1987); Kenward and Carpenter (2007a); Horton and Lipsitz (2001); Schafer (1999)), inverse probability weighting (Robins, Rotnitzky and Zhao (1994); Tsiatis (2007); Li, Shen and Robins (2013)) and doubly robust estimation (Scharfstein, Rotnitzky and Robins (1999); Lipsitz, Ibrahim and Zhao (1999); Robins, Rotnitzky and Scharfstein (2000); Robins and Rotnitzky (2001); Neugebauer and van der Laan (2005); Tsiatis (2007); Tchetgen Tchetgen (2009)).

The MAR assumption is strictly not testable in a nonparametric model without an additional assumption (Gill, van der Laan and Robins (1997); Potthoff et al. (2006)) and is often untenable. An outcome is said to be missing not at random (MNAR) if it is neither MCAR nor MAR, such that conditional on the observed variables, the missingness process depends on the unobserved variables (Little and Rubin (2002)). Identification is generally not available under MNAR without an additional assumption (Robins and Ritov (1997)). A possible approach is to make sufficient parametric assumptions (Little and Rubin (2002); Roy (2003); Wu and Carroll (1988)) about the full data distribution for identification. However, this approach can fail even with commonly used fully parametric models (Miao, Ding and Geng (2016); Wang, Shao and Kim (2014)). Alternative strategies for MNAR include positing instead sufficiently stringent modeling restrictions on a model for the missing data process (Rotnitzky and Robins (1997)) or conducting sensitivity analysis and constructing bounds (Moreno-Betancur and Chavance (2013); Kenward and Carpenter (2007b); Robins, Rotnitzky and Scharfstein (2000); Vansteelandt, Rotnitzky and Robins (2007)). A framework for identification and semiparametric inference was recently proposed by Miao, Tchetgen Tchetgen and Geng (2015) and Miao and Tchetgen Tchetgen (2016), building on earlier work by D'Haultfoeuille (2010), Wang, Shao and Kim (2014) and Zhao and Shao (2015), under the assumption that a shadow variable is fully observed which is associated with the outcome prone to missingness, but independent of the missingness process conditional on covariates and the possibly unobserved outcome. Another common identification approach involves leverag-

ing an instrumental variable (IV) (Manski (1985); Winship and Mare (1992)). Heckman's framework (Heckman (1979, 1997)) is perhaps the most common IV approach used primarily in economics and other social sciences to account for outcome MNAR. A valid IV is known to satisfy the following conditions:

(i) The IV is not directly related to the outcome in the underlying population, conditional on a set of fully observed covariates, and

(ii) The IV is associated with the missingness mechanism conditional on the fully observed covariates.

Therefore a valid IV must predict a person's propensity to have an observed outcome, without directly influencing the outcome.

In principle, one can use a valid IV to obtain a nonparametric test of the MAR assumption. However access to an IV does not generally point identify the joint distribution of the full data nor its functionals. Heckman's selection model consists of an outcome model that is associated with the selection process through correlated latent variables included in both models (Heckman (1979)). It is generally not identifiable without an assumption of bivariate normal latent error in defining the model (Wooldridge (2010)). Estimation using Heckman-type selection models may be sensitive to these parametric assumptions (Winship and Mare (1992); Puhani (2000)), although there has been significant work towards relaxing some of the assumptions (Manski (1985); Newey, Powell and Walker (1990); Das, Newey and Vella (2003); Newey (2009)). An alternative sufficient identification condition was considered by Tchetgen Tchetgen and Wirth (2017) which involves restricting the functional form of the selection bias function due to non-response on the additive, multiplicative or odds ratio scale. However, their approach for estimation is fully parametric and may be sensitive to bias due to model misspecification. Therefore a more robust approach is warranted.

In this paper, we develop a general framework for nonparametric identification of selection models based on an IV. We describe necessary and sufficient conditions for identifiability of the full data distribution with a valid IV. For inference we focus on estimation of an outcome mean, although the proposed methods are easy to adapt to other functionals. We develop semiparametric approaches including inverse probability weighting (IPW) and outcome regression (OR) that extend analogous methods previously developed for missing at random (MAR) settings, and introduce a novel doubly robust (DR) estimation approach. The consistency of each estimator relies on correctly specified models for different parts of the data generating model. We note that IPW in MNAR

via calibration weighting (Deville (2000); Kott (2006); Chang and Kott (2008)) has previously been proposed to account for sample nonresponse in survey design settings, and typically requires matching of weighted estimates to population totals for benchmark variables. Besides assuming a correctly specified model for nonresponse, identification in such settings is made possible by availability of known or estimated population totals, an assumption we do not require. Extensive simulation studies are used to investigate the finite sample properties of proposed estimators. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV seroprevalence in Mochudi, Botswana, using interviewer characteristics including gender, age and years of experience as IVs. All proofs are relegated to a Supplemental Appendix.

2. Notation and Assumptions

Suppose that one has observed n independent and identically distributed observations (X, Y, R, Z) with fully observed covariates X. Let R be the missingness indicator for Y, with $R = 1$ if Y is observed to take a value in its sample space Ω and $R = 0$ if $Y = Y^*$, indicating any value in Ω. The variable Z is a fully observed IV that satisfies conditions (i) and (ii) formalized below. In the evaluation of HIV prevalence in Mochudi, X includes all demographic and behavioral variables collected for all persons in the sample, while HIV status Y may be missing for individuals who failed to be tested. Let $\tilde{\pi}(X, Z) = \Pr(R = 1|X, Z)$ denote the propensity score for the missingness mechanism given (X, Z). As a valid IV, we assume that Z satisfies the following assumptions.

(IV.1) Exclusion restriction:

$$P_{Y|X,Z}(y|x, z) = P_{Y|X}(y|x) \quad \forall\, x, z.$$

(IV.2) IV relevance:

$$\tilde{\pi}(x, z) \neq \tilde{\pi}(x, z') \quad \forall\, x.$$

Exclusion restriction (IV.1) specifies that the IV does not have a direct effect on the outcome, which places restrictions on the full data law. IV relevance requires that the IV remains associated with the missingness mechanism even after conditioning on X. In spite of (IV.2), (IV.1) implies that Z cannot reduce the dependence between R and Y, therefore under MNAR $\pi(x, y, z) = P(R = 1|x, y, z)$ remains a function of y even after conditioning on (x, z). In addition, (IV.1) and (IV.2) imply that under MNAR the IV remains relevant in $\pi(x, y, z)$

conditional on (x, y). These facts will be used repeatedly throughout. $\tilde{\pi}(x, z)$ is typically referred to as the propensity score for the missingness process, and we shall refer to $\pi(x, y, z)$ as the extended propensity score.

3. Identification

Although (IV.1) reduces the number of unknown parameters in the full data law, identification is still only available for a subset of all possible full data laws. As an illustration, consider the case of binary outcome and IV. For simplicity and without loss of generality, we omit covariates X. Assumption (IV.1) implies $P(z, y) = P(y)P(z)$. We are only able to identify the quantities $P(z, y|R = 1)$, $P(z|R = 0)$, $P(R = 1)$ from the observed data. These quantities are functions of the unknown parameters: $P(Z = 1)$, $P(Y = 1)$, and $P(R = 1|z, y)$. So we have six unknown parameters, but only five available independent equations, one for each identified parameter given above. As a result, the full data law is not identifiable, and $P(Y = 1)$ is not identifiable.

The IV model becomes identifiable once one sufficiently restricts the class of models for the joint distribution of (Z, Y, R). Let $\mathcal{P}_\theta(R, Z, Y)$, $\mathcal{P}_\eta(Z)$ and $\mathcal{P}_\xi(Y)$ denote the collection of such candidates for $P(R = 1|z, y)$, $P(z)$ and $P(y)$, respectively. Members of the sets are indexed by parameters θ, η, and ξ, which may be infinite dimensional. The identifiability of the model is determined by the relationship between its members.

Result 1. Suppose that Assumption (IV.1) holds, then the joint distribution $P(z, y, r)$ is identifiable if and only if $\mathcal{P}_\theta(R, Z, Y)$ and $\mathcal{P}_\xi(Y)$ satisfy that for any pair of candidates

$$\{P_{\theta_1}(R = 1|z, y), P_{\xi_1}(y)\} \text{ and } \{P_{\theta_2}(R = 1|z, y), P_{\xi_2}(y)\}$$

in the model,

$$\frac{P_{\theta_1}(R = 1|z, y)}{P_{\theta_2}(R = 1|z, y)} \neq \frac{P_{\xi_2}(y)}{P_{\xi_1}(y)} \tag{3.1}$$

holds for at least one value of z and y.

Result 1 presents a necessary and sufficient condition for identifiability of the joint distribution of the full data, and thus a sufficient condition for identifiability of its functionals. We provide a more convenient condition to verify.

Corollary 1. *If Assumption (IV.1) holds, then the joint distribution $P(z, y, r)$ is identifiable if $\forall\ \theta_1, \theta_2$ such that $\theta_1 \neq \theta_2$, the ratio $P_{\theta_1}(R = 1|z, y)/P_{\theta_2}(R = 1|z, y)$ is either a constant or varies with z.*

Although this provides a sufficient condition for identification of the joint distribution of the full data, it may be used to establish identifiability in parametric or semi-parametric models, as we illustrate in a number of examples. Let $\mathcal{M}_{\text{IV}}$ denote the collection of models with valid IV.

Example 1. Suppose both Y and Z are binary and consider the model $\mathcal{M}_1 \cap \mathcal{M}_{\text{IV}}$, where

$$\mathcal{M}_1 = \big\{P(R=1|Z,Y) = \text{expit}\,(\theta_0 + \theta_1 Z + \theta_2 Y + \theta_3 ZY) : (\theta_0,\theta_1,\theta_2,\theta_3) \in \mathbb{R}^4\big\},$$

which includes the nonparametric model. It is shown in the Supplemental Appendix that this model does not satisfy (3.1) and therefore the joint distribution of (Z,Y,R) cannot be identified without reducing the dimension of θ. In contrast, Corollary 1 confirms that the smaller model $\mathcal{M}_2 \cap \mathcal{M}_{\text{IV}}$ is identified, where

$$\mathcal{M}_2 = \big\{P(R=1|Z,Y) = \text{expit}\,(\theta_0 + \theta_1 Z + \theta_2 Y) : (\theta_0,\theta_1,\theta_2) \in \mathbb{R}^3\big\}.$$

Thus the IV model becomes identified upon imposing a no-interaction assumption between Y and Z in the logistic model for the extended propensity score. An analogous result holds for possibly continuous Y and Z, assuming a logistic generalized additive model for the extended propensity score.

Example 2. The model $\mathcal{M}_{\text{SL}} \cap \mathcal{M}_{\text{IV}}$ is identified for the separable logistic missing data mechanism

$$\mathcal{M}_{\text{SL}} = \{P(R=1|Z,Y) = \text{expit}(q(Z) + h(Y))\}, \tag{3.2}$$

where $q(\cdot)$ and $h(\cdot)$ are unknown functions differentiable with respect to Z and Y, respectively.

4. Estimation and Inference

In this section, we consider estimation and inference under a variety of semi-parametric IV models shown to satisfy Result (1). We denote the collection of such identifiable models as $\mathcal{M}^*_{\text{IV}}$. Although in principle the identification results given in the previous section allow for nonparametric inference, in practice estimation often involves specifying parametric models, at least for parts of the full data law. This is generally the case when a large number of covariates X or Z are present and therefore the curse of dimensionality precludes the use of nonparametric regression to model conditional densities or their mean functions required for IV inferences (Robins and Ritov (1997)). As a measure of departure from MAR, we introduce the selection bias function

$$\eta(x,y,z)=\log\left[\frac{\{P(R=1|x,y,z)/P(R=0|x,y,z)\}}{\{P(R=1|x,Y=0,z)/P(R=0|x,Y=0,z)\}}\right]. \tag{4.1}$$

η quantifies the degree of association between Y and R given (X,Z) on the log odds ratio scale. Under MAR, $P(R=1|x,y,z)=P(R=1|x,z)$ and $\eta=0$. The conditional density $P(r,y,z|x,z)$ can be represented in terms of the selection bias function η and baseline densities as

$$\begin{aligned}P(r,y,z|x)=C(x,z)^{-1}\exp\{(r-1)\eta(x,y,z)\}\\ \times f(y|R=1,x,z)P(r|Y=0,x,z)q(z|x),\end{aligned} \tag{4.2}$$

where $C(x,z)<+\infty$ for all (x,z) is a normalizing constant, and $q(z|x)$ models the density of the IV conditional on the covariates (Chen (2007); Tchetgen Tchetgen, Robins and Rotnitzky (2010)). As we show below, the selection bias function η in (4.2) needs to be correctly specified for any of the three proposed estimators to be consistent. To fix ideas, throughout we suppose that one aims to make inferences about the population mean $\phi=E(Y)$, although the proposed methods are easy to extend to other full data functionals.

IPW estimation requires a correctly specified model for the extended propensity score $\pi(x,y,z)$, which under logit link function is

$$\pi(x,y,z)=\frac{1}{[1+\exp\{-\eta(x,y,z)-\lambda(x,z)\}]}, \tag{4.3}$$

where $\eta(x,y,z)$ is the selection bias function given in (4.1), and $\lambda(x,z)=\log\{P(R=1|Y=0,x,z)/P(R=0|Y=0,x,z)\}$ is a person's baseline conditional odds of observing complete data. In principle, one could use any well-defined link function for the propensity score, but we simplify the presentation by only considering the logit case. We consider IPW estimation in the model $\mathcal{M}_{\text{IPW}}\subset\mathcal{M}^{*}_{\text{IV}}$, where

$$\begin{aligned}\mathcal{M}_{\text{IPW}}=\Big\{&P(r,y,z|x):\eta(x,y,z;\zeta),P(r|Y=0,x,z;\omega),q(z|x;\xi);\\ &\text{unrestricted } P(y|R=1,x,z)\Big\},\end{aligned}$$

and the parametric models indexed by parameters ζ, ω, and ξ, respectively, are assumed to be correctly specified, while the baseline outcome model $f(y|R=1,x,z)$ in (4.2) is unrestricted.

Outcome regression-based estimation under MAR requires a model for $f(y|R=1,x,z)=f(y|x,z)$, which can be estimated based on complete-cases. However, under MNAR $f(y|R=1,X,Z)\neq f(y|R=0,X,Z)$ and estimation of $f(y|R=0,x,z)$ is not readily available since outcome is not observed for this

subpopulation. However, by (4.2),

$$f(y|r,x,z)=\frac{P(y,r|x,z)}{\int P(y,r|x,z)\mathrm{d}\mu(y)}=\frac{\exp\{-(1-r)\eta(x,y,z)\}f(y|R=1,x,z)}{E[\exp\{-(1-r)\eta(x,Y,z)\}|R=1,x,z]}, \quad (4.4)$$

and therefore the density $f(y|R=0,x,z)$ can be expressed in terms of the selection bias function η and baseline outcome model $f(y|R=1,x,z)$ for complete-cases. We consider OR estimation in the model $\mathcal{M}_{\text{OR}} \subset \mathcal{M}^*_{\text{IV}}$ where

$$\mathcal{M}_{\text{OR}}=\Big\{P(r,y,z|x):\eta(x,y,z;\zeta),P(y|R=1,x,z;\theta),q(z|x;\xi);\\ \text{unrestricted } \lambda(x,z)\Big\},$$

which allows the baseline missing data model $P(r|Y=0,x,z)$ to remain unrestricted while the models indexed by parameters ζ, θ and ξ are assumed to be correctly specified.

We also propose a doubly robust estimator which is consistent in the union model $\mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$, provided the models $\eta(x,y,z;\zeta)$ and $q(z|x;\xi)$ are correctly specified, and either $P(r|Y=0,x,z;\omega)$ or $P(y|R=1,x,z;\theta)$, but not necessarily both, are correctly specified, thus giving the analyst two chances, instead of one, to obtain valid inferences.

Throughout the next section, we let $\hat{\theta}_{\text{MLE}}$ denote the complete-case maximum likelihood estimator which maximizes the conditional log-likelihood $\sum_{i:R_i=1}\log P(y_i|x_i,z_i;\theta)$, and let $\hat{\xi}_{\text{MLE}}$ denote the maximum likelihood estimator which maximizes the log-likelihood $\sum_{i=1}^n \log q(z_i|x_i;\xi)$. Let $\mathbb{P}_n$ denote the empirical measure $\mathbb{P}_n f(O)=n^{-1}\sum_{i=1}^n f(O_i)$.

4.1. Inverse probability weighted estimation under $\mathcal{M}_{\text{IPW}}$

IPW is a well-known approach to acount for missing data under MAR. In this section we describe an analogous approach under MNAR. Standard approaches for estimating the propensity score under MAR, such as maximum likelihood of a logistic regression model of the propensity score, cannot be used here since the extended propensity score $\pi(x,y,z)$ depends on Y which is only observed when $R=1$. Therefore, we adopt an alternative method of moments approach which resolves this difficulty. Under the model $\mathcal{M}_{\text{IPW}}$, $(\hat{\zeta},\hat{\omega})$ solves

$$\mathbb{P}_n\left\{\boldsymbol{U}^{\text{IPW}}\left(\hat{\xi}_{\text{MLE}},\hat{\zeta},\hat{\omega}\right)\right\}=\mathbf{0} \quad (4.5)$$

where $\boldsymbol{U}^{\text{IPW}}(\cdot)$ consists of the estimating functions

$$\left\{\frac{R}{\pi\left(\hat{\zeta},\hat{\omega}\right)}-1\right\}\boldsymbol{h}_1(X,Z) \tag{4.6}$$

$$\frac{R}{\pi\left(\hat{\zeta},\hat{\omega}\right)}\boldsymbol{g}(X,Y)\left[\boldsymbol{h}_2(Z,X)-E\left\{\boldsymbol{h}_2(Z,X)\middle|X;\hat{\xi}_{\text{MLE}}\right\}\right]. \tag{4.7}$$

Functions (4.6) and (4.7) estimate unknown parameters in $P(r|Y=0,x,z;\omega)$ and $\eta(x,y,z;\zeta)$ respectively, where $\boldsymbol{h}_1$ is an user-specified function of (x,z) with the same dimension as ω, while $\boldsymbol{g}$ and $\boldsymbol{h}_2$ are user-specified functions of (x,y) and (x,z), respectively, with the same dimension as ζ. Specific choices of $(\boldsymbol{h}_1,\boldsymbol{h}_2,\boldsymbol{g})$ can generally affect efficiency, but not consistency.

Proposition 1. *Consider a model $\mathcal{M}_{IPW}\subset\mathcal{M}^*_{IV}$ that satisfies Result* (1)*. Then the IPW estimator*

$$\hat{\phi}^{IPW}=\mathbb{P}_n\left\{\frac{RY}{\pi(\hat{\eta})}\right\} \tag{4.8}$$

is consistent and asymptotically normal as $n\to\infty$,

$$\sqrt{n}\left(\hat{\phi}^{IPW}-\phi_0\right)\xrightarrow{d}N\left(0,V_{IPW}\right)$$

in model $\mathcal{M}_{IPW}$ under suitable regularity conditions, where V_{IPW} is given in the Supplemental Appendix.

4.2. Outcome regression estimation under $\mathcal{M}_{\text{OR}}$

Next, consider inferences under a parametric model for the outcome, $\mathcal{M}_{\text{OR}}$. Using the parametrization given in (4.4), consider the parametric model

$$P(y|R=0,x,z;\zeta,\hat{\theta}_{\text{MLE}})=\frac{\exp\{-\eta(x,y,z;\zeta)\}\,f\left(y|R=1,x,z;\hat{\theta}_{\text{MLE}}\right)}{E\left[\exp\{-\eta(x,Y,z;\zeta)\}|R=1,x,z;\hat{\theta}_{\text{MLE}}\right]},$$

and the estimator $\tilde{\zeta}$ solving

$$\begin{aligned}
&\mathbb{P}_n\left\{\boldsymbol{U}^{\text{OR}}\left(\tilde{\zeta},\hat{\xi}_{\text{MLE}},\hat{\theta}_{\text{MLE}},\boldsymbol{q}_1,\boldsymbol{q}_2\right)\right\}\\
&=\mathbb{P}_n\left[\boldsymbol{q}_1(X,Z)-E\left\{\boldsymbol{q}_1(X,Z)\middle|X;\hat{\xi}_{\text{MLE}}\right\}\right]\times\\
&\quad\left\{(1-R)E\left(\boldsymbol{q}_2(X,Y)\middle|R=0,X,Z;\tilde{\zeta},\hat{\theta}_{\text{MLE}}\right)+R\boldsymbol{q}_2(X,Y)\right\}\\
&=\mathbf{0},
\end{aligned} \tag{4.9}$$

where $\boldsymbol{q}_1,\boldsymbol{q}_2$ are vectors of the same dimensions as ζ.

Proposition 2. *Consider a model $\mathcal{M}_{OR}\subset\mathcal{M}^*_{IV}$ that satisfies Result* (1)*. Then*

the outcome regression estimator

$$\hat{\phi}^{OR} = \mathbb{P}_n \left\{ RY + (1-R)E\left(Y \middle| R=0, X, Z; \tilde{\zeta}, \hat{\theta}_{MLE}\right)\right\} \tag{4.10}$$

is consistent and asymptotically normal as $n \to \infty$,

$$\sqrt{n}\left(\hat{\phi}^{OR} - \phi_0\right) \xrightarrow{d} N\left(0, V_{OR}\right)$$

in model $\mathcal{M}_{OR}$ *under suitable regularity conditions, where* V_{OR} *is given in the Supplemental Appendix.*

4.3. Doubly robust estimation under $\mathcal{M}_{\text{DR}}$

Estimation approaches described thus far depend on correct specification of extended propensity score for IPW and outcome model for OR. Here we describe a doubly robust estimator that remains consistent if the conditional density $q(z|x;\xi)$ is correctly specified, and either $P(y|R,X,Z;\theta)$ or $P(r|Y,X,Z;\omega)$ is correctly specified, but not necessarily both. We write $\mathcal{M}_{\text{DR}} = \mathcal{M}_{\text{IPW}} \cup \mathcal{M}_{\text{OR}}$. Our construction requires first obtaining the DR estimator $\hat{\zeta}_{\text{DR}}$ of the parameter indexing selection bias function $\eta(\zeta)$ that remains consistent in $\mathcal{M}_{\text{DR}}$. In this vein, let

$$\begin{aligned}
&\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z; \zeta, \omega, \hat{\theta}_{\text{MLE}}, \boldsymbol{u}\right) \\
&= \frac{R}{\pi(\zeta,\omega)}\boldsymbol{u}(X,Y) - \frac{R-\pi(\zeta,\omega)}{\pi(\zeta,\omega)} E\left\{\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right\} \\
&= \frac{R}{\pi(\zeta,\omega)}\left[\boldsymbol{u}(X,Y) - E\left\{\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right\}\right] \\
&\quad + E\left\{\boldsymbol{u}(X,Y)\middle| R=0, X, Z; \zeta, \hat{\theta}_{\text{MLE}}\right\},
\end{aligned} \tag{4.11}$$

where $\boldsymbol{u}(X,Y)$ is of the same dimensions as ζ. We obtain $(\hat{\zeta}_{\text{DR}}, \hat{\omega})$ as the solution to the estimating equation (4.6), combined with

$$\begin{aligned}
&\mathbb{P}_n\left\{\boldsymbol{U}^{\text{DR}}\left(\hat{\zeta}_{\text{DR}}, \hat{\omega}, \hat{\theta}_{\text{MLE}}, \hat{\xi}_{\text{MLE}}, \boldsymbol{u}, \boldsymbol{v}\right)\right\} \\
&= \mathbb{P}_n\Bigg(\left[\boldsymbol{v}(X,Z) - E\left\{\boldsymbol{v}(X,Z)\middle| X; \hat{\xi}_{\text{MLE}}\right\}\right] \\
&\qquad \times \left\{\boldsymbol{G}^{\text{DR}}\left(R, X, Y, Z; \hat{\zeta}_{\text{DR}}, \hat{\omega}, \hat{\theta}_{\text{MLE}}, \boldsymbol{u}\right)\right\}\Bigg) \\
&= \mathbf{0}.
\end{aligned} \tag{4.12}$$

Proposition 3. *Consider a model* $\mathcal{M}_{DR} \subset \mathcal{M}^*_{IV}$ *that satisfies Result* (1)*. Then the doubly robust estimator*

$$\hat{\phi}^{DR} = \mathbb{P}_n\left\{\boldsymbol{G}^{DR}\left(R, X, Y, Z, \hat{\zeta}_{DR}, \hat{\omega}, \hat{\theta}_{MLE}, \boldsymbol{u}^{\dagger}\right)\right\}, \tag{4.13}$$

where $\boldsymbol{u}^{\dagger}(X,Y) = Y$*, is consistent and asymptotically normal as* $n \to \infty$*,*

$$\sqrt{n}\left(\hat{\phi}^{DR} - \phi_0\right) \xrightarrow{d} N\left(0, V_{DR}\right)$$

in the model $\mathcal{M}_{DR}$ *under suitable regularity conditions, where* V_{DR} *is given in the Supplemental Appendix.*

The notion of doubly robust estimation was first introduced in the context of semi-parametric non-response models under MAR (Scharfstein, Rotnitzky and Robins (1999)), and the approach was further studied by others (Lipsitz, Ibrahim and Zhao (1999); Robins, Rotnitzky and Scharfstein (2000); Lunceford and Davidian (2004); Neugebauer and van der Laan (2005)) with theoretical underpinnings given by Robins and Rotnitzky (2001) and van der Laan and Robins (2003). A doubly robust version of estimating equation (4.13) of mean outcome under MNAR was previously described by Vansteelandt, Rotnitzky and Robins (2007), who assume that the selection bias function η is known a priori within the context of a sensitivity analysis. An important contribution here is to derive a large class of DR estimators of the selection bias using an IV. To the best of our knowledge, this is the first time that a DR estimator for the mean outcome has been constructed in the context of an IV for data subject to MNAR.

5. Simulation Study

In order to investigate the finite-sample performance of proposed estimators, we carried out a simulation study involving i.i.d. data (Y, Z, X), where $X = (X_1, X_2)$. For each sample size $n = 2{,}000$, $5{,}000$, we simulated 1,000 data sets as follows:

$$X_1 \sim \text{Bernoulli}(p = 0.4), \quad X_2 \sim \text{Bernoulli}(p = 0.6),$$
$$Z \sim \text{Bernoulli}\left[p = \left\{1 + \exp\left(-0.4 - 0.9X_1 + 0.7X_2 + 0.8X_1X_2\right)\right\}^{-1}\right],$$
$$Y \sim \text{Bernoulli}\left[p = \left\{1 + \exp\left(-1.0 + 1.2X_1 - 1.5X_2\right)\right\}^{-1}\right],$$
$$R \sim \text{Bernoulli}\left[p = \left\{1 + \exp\left(1.5 - 2.5Z - 0.8X_1 + 1.2X_2 - 1.8Y\right)\right\}^{-1}\right],$$

such that Y is only observed if $R = 1$. Under this data generating mechanism, Z satisfies **(IV.1)** and **(IV.2)**, with the true value of $\phi_0 = E(Y) = 0.769$. The selection bias model is $\alpha(x, y, z) = \zeta y$ with true value $\zeta_0 = 1.8$. The model is identified since the missing data mechanism follows the separable logistic regression model described in Example 2 of Section 3. For IPW estimation, we specified the correct extended propensity score and model for $P(Z = 1|X_1, X_2; \xi)$, with

$h_1 = (Z, X_1, ZX_1)^T$, $g = Y$ and $h_2 = Z$ in (4.6) and (4.7). For OR estimation, we let $(q_1, q_2) = (Z, Y)$ in (4.9) and specified a saturated logistic regression for Y with all 2-way and 3-way interactions included. DR estimation was carried out as described in the previous section. While Chang and Kott (2008) only considered a survey design setting, here the IPW approach is analogous to a form of calibration weighted estimation that matches the weighted sample estimates of benchmark variables $L_{CW} = \{1, Z, X_1, X_2, Y\{Z - P(Z = 1|X_1, X_2)\}\}$ to their estimated population totals, where the last variable in L_{CW} has known population total value of zero by **(IV.1)**.

To study the performance of the proposed estimators in situations where some models may be mis-specified, we also evaluated the estimators where either the extended propensity score model or the complete-case outcome model was mis-specified by replacing them with models

$$P(R = 1|X, Y, Z) = \text{expit}(\omega_0 + \omega_1 X_1 + \omega_2 Z + \omega_3 X_1 Z + \zeta Y)$$

and

$$P(Y = 1|R = 1, X, Z) = \text{expit}(\theta_0 + \theta X_1),$$

respectively.

In each simulated sample, we evaluated the standard error of the estimator using the sandwich estimator. Wald 95% confidence interval coverage rates were evaluated across 1,000 simulations. Estimating equations were solved using the R package BB (Varadhan and Gilbert (2009)). Figures 1 and 2 present results for estimation of the selection bias parameter ζ_0 and the outcome mean ϕ_0, respectively, while Table 1 shows the empirical coverage rates. Under correct model specification, all estimators have negligible bias for ϕ_0 and ζ_0 that diminishes with increasing sample size, with empirical coverage near the nominal 95% level. In agreement with our theoretical results, the IPW and OR estimators are biased with poor empirical coverages when the extended propensity score or the complete-case outcome model is misspecified, respectively. The DR estimator performs well in terms of bias and coverage when either model is misspecified but the other is correct.

6. Applications

To illustrate the proposed IV approach, we obtained data from a household survey in Mochudi, Botswana to estimate HIV seroprevalence among adults adjusting for selective missingness of HIV test results. The data consist of 4,997

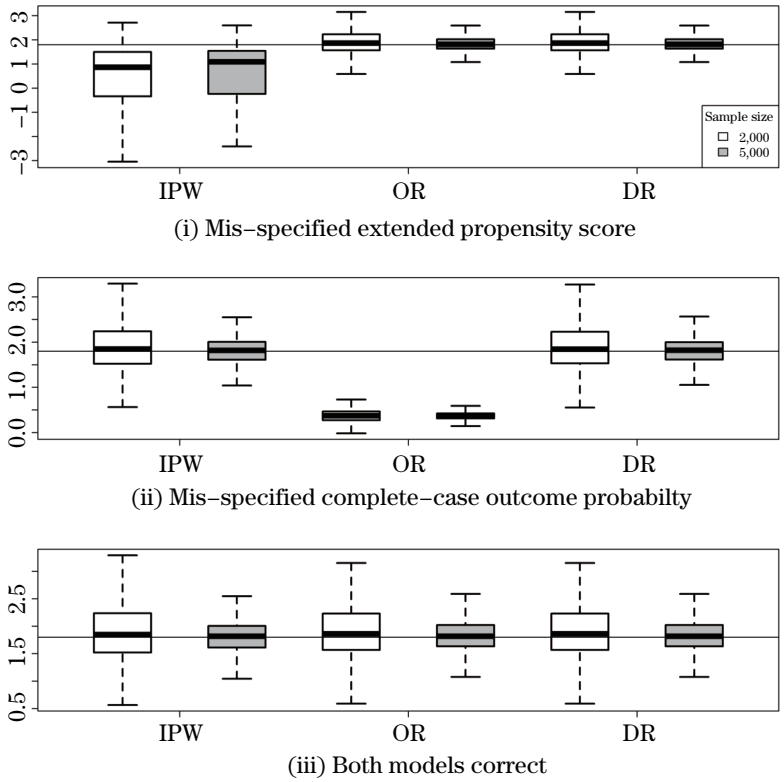

Figure 1. Boxplots of inverse probability weighted (IPW), outcome regression (OR) and doubly-robust (DR) estimators of the selection bias parameter, for which the true value $\zeta_0 = 1.8$ is marked by the horizontal lines.

Table 1. Empirical coverage rates based on 95% Wald confidence intervals under three scenarios: (i) mis-specified extended propensity score, (ii) mis-specified complete-case outcome probability and (iii) both models are correct. In each scenario, the first row presents results for $n = 2{,}000$ and the second row for $n = 5{,}000$.

	ζ			ϕ		
	IPW	OR	DR	IPW	OR	DR
(i)	86.4	95.4	95.4	81.3	95.2	95.2
	57.8	95.1	95.1	50.1	94.9	94.9
(ii)	95.0	0.0	94.4	95.1	65.6	95.2
	94.7	0.0	94.5	95.0	29.9	94.5
(iii)	95.0	95.4	95.4	95.1	95.2	95.2
	94.7	95.1	95.1	95.0	94.9	94.9

adults between the ages of 16 and 64 who were contacted for the survey, out of whom 4,045 (81%) had complete information on HIV testing. Of those who

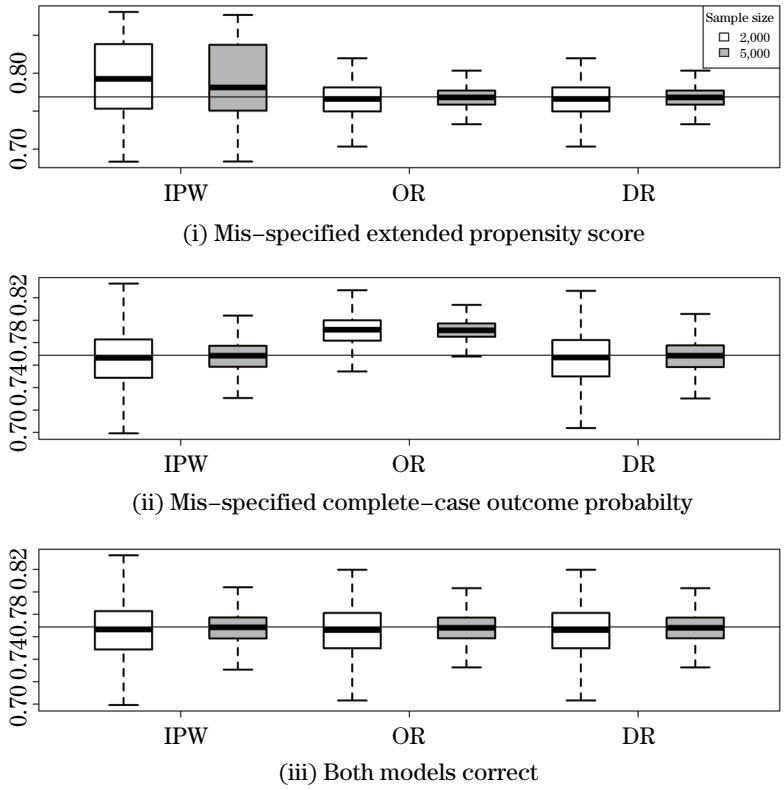

Figure 2. Boxplots of inverse probability weighted (IPW), outcome regression (OR) and doubly-robust (DR) estimators of the outcome mean, for which the true value $\phi_0 = 0.769$ is marked by the horizontal lines.

did not have HIV test results ($R = 0$), 111 (2%) agreed to participate in the HIV test but their final HIV outcomes are unknown, and 841 (17%) refused to participate in the HIV testing component. It is likely that refusal to participate in the survey when contact is established presents a possible source of selection bias.

Fully available individual characteristics from the survey include participant gender (X). Candidate IVs include interviewer gender (Z_1), age (Z_2), and years of experience (Z_3). These interviewer characteristics are likely to influence the response rates of individuals who were contacted for the survey, but are unlikely to directly influence an individual's HIV status given that interviewer deployment was determined at random prior to the survey. We implemented the proposed IPW, OR, and DR estimators by making use of interviewer gender, age, and years of experience as IVs. For IPW estimation, the missingness propensity

Table 2. Estimation for HIV seroprevalence (ϕ) and magnitude of selection bias (ζ) in Mochudi, Botswana with 95% Wald confidence intervals.

Estimator	$\hat{\phi}$	$\hat{\zeta}$	$\hat{\zeta}$ p-val
CC	0.214 (0.202, 0.227)	–	–
MAR IPW	0.213 (0.201, 0.226)	–	–
IV IPW	0.260 (0.175, 0.341)	−1.601 (−2.992, −0.210)	0.02
IV OR	0.241 (0.175, 0.307)	−0.757 (−1.889, 0.376)	0.19
IV DR	0.258 (0.174, 0.342)	−1.121 (−2.433, 0.191)	0.09

score is specified as a main effects only logistic regression, with the selection bias function specified as $\alpha(x, y, z) = \zeta y$, where Y is HIV status. The posited missing data mechanism belongs to the separable logistic class, therefore the average HIV prevalence can be identified, by Example 2. For OR estimation, we specified the regression model

$$\text{logit } P(Y = 1|R = 1, X, \boldsymbol{Z}) = \theta_0 + \theta_1 X + \theta_2 Z_1 + \theta_3 Z_2 + \theta_4 Z_3. \tag{6.1}$$

Finally, the doubly robust estimator is implemented by incorporating both models. Because more than one IV was available, estimating equations $\boldsymbol{U}^{\text{IPW}}$, $\boldsymbol{U}^{\text{OR}}$, and $\boldsymbol{U}^{\text{DR}}$ were solved using the generalized method of moments (GMM) package in R (Chaussé (2010)). Standard errors were obtained using the proposed sandwich estimator. For comparison, we also carried out standard complete-case analysis and standard IPW estimation assuming MAR conditional on (x, z) using a main effects only logistic regression to model the propensity score. Results are presented in Table 2.

IV estimates of HIV seroprevalence are $12.6 - 21.5\%$ higher than the crude estimate of 0.214 (95% CI: 0.202–0.227) based on complete-cases only. Standard IPW (i.e. assuming MAR) produced similar estimates as complete-case analysis. Negative point estimates of the selection bias parameter ζ suggest that HIV-infected persons are less likely to participate in the HIV testing component of the survey, although this difference is statistically significant at 0.05 α-level only for IPW. The larger confidence intervals of the three IV estimators of ϕ_0 compared to those of the CC and MAR estimators are a more accurate reflection of the amount of uncertainty involving inferences about ϕ_0, since the CC and MAR estimators do not take into account the uncertainty about the underlying MNAR mechanism by assuming MCAR and MAR, respectively, i.e. setting selection bias parameter $\zeta = 0$. $\hat{\phi}^{\text{IV IPW}}$ and $\hat{\phi}^{\text{IV DR}}$ are close to each other. This comparison is useful as an informal goodness of fit test in that their similarity suggests that the missingness propensity score may be specified nearly correctly (Robins and Rotnitzky (2001)). In addition, by incorporating all possible pairwise interaction

terms in the outcome logistic regression model, and therefore allowing it to be more flexible, the OR point estimate $\hat{\phi}^{\text{IV OR}}$ increases to 0.246 (95% CI: 0.179-0.314), thus even closer to $\hat{\phi}^{\text{IV IPW}}$ and $\hat{\phi}^{\text{IV DR}}$.

7. Conclusion

In this paper, we have considered a pernicious form of selection bias which can arise from outcome missing not at random. We have argued that under fairly reasonable assumptions this problem can be made more tractable with the aid of an IV, and proposed a general framework for establishing identifiability of parametric, semiparametric, and nonparametric models. In addition, we have characterized the set of all influence functions of regular and asymptotically linear estimators as well as the semiparametric efficient score of (ζ, ϕ) in model $\mathcal{M}_{np}$ which assumes that Z is a valid IV, the selection bias function $\eta(X, Y, Z; \zeta)$ is correctly specified, and the joint likelihood of (Y, X, Z, R) is otherwise unrestricted. The efficient score is not generally available in closed-form, except in special cases, such as when Z and Y are both polytomous. Due to space constraints, local efficiency results are available in Sun et al. (2016).

Supplementary Materials

The proofs for results, propositions and examples are included in an online Supplemental Appendix.

References

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

Chaussé, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *Journal of Statistical Software* **34**, 1–35.

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

Das, M., Newey, W. K. and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies* **70**, 33–58.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. 65–76. In *COMPSTAT*, Springer.

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

Gill, R. D., van der Laan, M. J. and Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Lecture Notes in Statistics* (Edited by D. Y. Lin and T. R. Fleming). Springer-Verlag.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47**, 153–161.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources* **32**, 441–462.

Horton, N. J. and Laird, N. M. (1998). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* **8**, 37–50.

Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician* **55**, 244–254.

Kenward, M. and Carpenter, J. (2007a). Multiple imputation: Current perspectives. *Statistical Methods in Medical Research* **16**, 199–218.

Kenward, M. and Carpenter, J. (2007b). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research* **16**, 259–275.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133–142.

Li, L., Shen, C., Li, X. and Robins, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research* **22**, 14–30.

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

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

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.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *The Econometrics Journal* **27**, 313–333.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

Miao, W., Tchetgen Tchetgen, E. and Geng, Z. (2015). Identification and doubly robust estimation of data missing not at random with an ancillary variable. *arXiv preprint arXiv:1509.02556*.

Miao, W. and Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.

Moreno-Betancur, M. and Chavance, M. (2013). Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: Methodology and application in a clinical trial with drop-outs. *Statistical Methods in Medical Research* **25**, 1471–1489.

Neugebauer, R. and van der Laan, M. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference* **129**, 405–426.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal* **12**, S217–S229.

Newey, W. K., Powell, J. and Walker, J. (1990). Semiparametric estimation of selection models: some empirical results. *The American Economic Review* **80**, 324–328.

Potthoff, R. F., Tudor, G. E., Pieper, K. S. and Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical*

Research **15**, 213–234.

Puhani, P. (2000). The heckman correction for sample selection and its critique. *Journal of Economic Surveys* **14**, 53–68.

Robins, J., Rotnitzky, A. and Scharfstein, D. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Edited by M. E. Halloran and D. Berry). Springer-Verlag.

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

Robins, J. M. and Rotnitzky, A. (2001). Comment on "inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11**, 920–936.

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

Rotnitzky, A. and Robins, J. (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine* **16**, 81–102.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59**, 829–836.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schafer, J. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94**, 1096–1146.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen Tchetgen, E. (2016). Semiparametric Estimation with Data Missing Not at Random Using an Instrumental Variable. *ArXiv e-prints*.

Tchetgen Tchetgen, E. (2009). A simple implementation of doubly robust estimation in logistic regression with covariates missing at random. *Epidemiology* **20**, 391–394.

Tchetgen Tchetgen, E. J., Robins, J. M. and Rotnitzky, A. (2010). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97**, 171–180.

Tchetgen Tchetgen, E. J. and Wirth, K. (2017). A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics* published online ahead of print, DOI: 10.1111/biom.12670.

Tsiatis(2007). *Semiparametric Theory and Missing Data*. Springer.

van der Laan, M. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag.

Vansteelandt, S., Rotnitzky, A. and Robins, J. M. (2007). Estimation of regression models for the mean of repeated outcomes under non-ignorable non-monotone non-response. *Biometrika* **94**, 841–860.

Varadhan, R. and Gilbert, P. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* **32**, 1–26.

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

Winship, C. and Mare, R. (1992). Models for sample selection bias. *Annual Review of Sociology* **18**, 327–350.

Wooldridge, J. (2010). *Economic Analysis of Cross Section and Panel Data*. MIT press.

Wu, M. and Carroll, R. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

Zhao, J. and Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association* **110**, 1577–1590.

Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Ave. Building 2, 4th Floor, Boston, Massachusetts 02115 USA.

E-mail: bsun@hsph.harvard.edu

School of Statistics, University of Minnesota at Twin Cities, 224 Church St., Minneapolis, Minnesota 55455 USA.

E-mail: liux3771@umn.edu

Department of Business Statistics & Econometrics, Guanghua School of Management, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China.

E-mail: mwfy@pku.edu.cn

Departments of Epidemiology and Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Kresge Building, Boston, Massachusetts 02115 USA.

E-mail: kwirth@hsph.harvard.edu

Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Kresge Building Room 823, Boston, Massachusetts 02115 USA.

E-mail: robins@hsph.harvard.edu

Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Kresge Building Room 822, Boston, Massachusetts 02115 USA.

E-mail: etchetge@hsph.harvard.edu

(Received June 2016; accepted April 2017)