1-2017

Efficient augmented inverse probability weighted estimation in missing data problems

Jing QIN
National Institutes of Health

Biao ZHANG
University of Toledo

Denis H. Y. Leung
Singapore Management University, denisleung@smu.edu.sg

Efficient Augmented Inverse Probability Weighted Estimation in Missing Data Problems

Jing QIN
National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD 20892 (*jingqin@niaid.nih.gov*)

Biao ZHANG
Department of Mathematics, University of Toledo, Toledo, OH 43606-3390 (*bzhang@utnet.utoledo.edu*)

Denis H.Y. LEUNG
School of Economics, Singapore Management University, Singapore 178903 (*denisleung@smu.edu.sg*)

When analyzing data with missing data, a commonly used method is the inverse probability weighting (IPW) method, which reweights estimating equations with propensity scores. The popularity of the IPW method is due to its simplicity. However, it is often being criticized for being inefficient because most of the information from the incomplete observations is not used. Alternatively, the regression method is known to be efficient but is nonrobust to the misspecification of the regression function. In this article, we propose a novel way of optimally combining the propensity score function and the regression model. The resulting estimating equation enjoys the properties of robustness against misspecification of either the propensity score or the regression function, as well as being locally semiparametric efficient. We demonstrate analytically situations where our method leads to a more efficient estimator than some of its competitors. In a simulation study, we show the new method compares favorably with its competitors in finite samples. Supplementary materials for this article are available online.

KEY WORDS: Inverse probability weighting; Missing data; Regression estimate; Semiparametric efficiency.

1. INTRODUCTION

In this article, we study regression problems in the presence of missing data. We assume that the data are missing at random (MAR; Little and Rubin 2002), which implies that, conditional on the observed, missingness and the unobserved data are independent. One way to approach this problem is to model the propensity score (Rosenbaum and Rubin 1983) or the probability of missingness given the covariates, and use the inverse of these estimates as weights to derive an estimator through inverse probability weighting (IPW) of observed outcomes (Horvitz and Thompson 1952). This approach was used by Imbens (1992) in choice-based sampling, Robins and Rotnitzky (1995) in nonlinear regression problems, Scharfstein, Robins, and Rotnitzky (1999) in panel data analysis, Wooldridge (2007) in M-estimation, and Hahn (1998) in treatment effects models. Another approach is to model the regression of the outcome given the covariates using the complete observations, and to derive an estimator based on the fitted values for observed and missing observations. Roberts, Rao, and Kumar (1987) used this approach to study unemployment rate data collected in a multi-level survey. Pepe, Reilly, and Fleming (1994) applied this method in regression analyses with incomplete covariate information. Cheng (1994) used this method to estimate the mean in a nonparametric regression. A third approach derives estimators using a combination of the propensity score and the regression model (Robins, Rotnitzky, and Zhao 1994). This approach has the attractive "doubly robust" property that estimators are consistent as long as either the propensity score or the outcome regression model is correctly specified. Furthermore, it attains the semiparametric efficiency bound (Bickel et al. 1993) if both the propensity score and outcome regression model are correct. Robins, Rotnitzky, and Zhao (1994, 1995) used this method for regression models with panel data. Bang and Robins (2005), Wooldridge (2007), Cattaneo (2010), Uysal (2015), among others, applied this method in treatment effects models. Słoczyński and Wooldridge (2014) provided a unified framework for various doubly robust estimators of the average treatment effect under unconfoundedness and Kang and Schafer (2007) showed that doubly robustness can also be achieved by other means (Särndal, Swensson, and Wretman 1989; Little and An 2004).

The class of estimators proposed by Robins, Rotnitzky, and Zhao (1994) achieves double robustness by augmenting the IPW by a function of the regression model, hence the estimators are often referred to as augmented inverse probability weighted (AIPW) estimators. Robins, Rotnitzky, and Zhao (1994) suggested using maximum likelihood to estimate the parameters in the propensity score and ordinary least squares to estimate the parameters in the outcome regression model. However, Kang and Schafer (2007) showed that it can be severely biased when (1) the regression model and the propensity score function are simultaneously misspecified or (2) the propensity score function is near zero for some observations. Kang and Schafer (2007) pointed out that a reason for the poor performance of the AIPW is due to inverse weighting of the propensity score. Recently,

Published in Journal of Business and Economic Statistics, 2017 January, Volume 35, Issue 1, Pages 86-97
http://doi.org/10.1080/07350015.2015.1058266

a number of works have suggested alternative doubly robust estimators to address the problems demonstrated by Kang and Schafer (2007). Most of these works differ from Robins, Rotnitzky, and Zhao (1994) in the way the outcome regression model is estimated. For example, Kang and Schafer (2007) and Rubin and van der Laan (2008) used weighted least squares to estimate the parameters in the regression model. Tan (2006) and Cao, Tsiatis, and Davidian (2009) proposed estimators by projection. On the other hand, Robins et al. (2007) and Cao, Tsiatis, and Davidian (2009) proposed using alternative estimates of the propensity score.

This article studies a new doubly robust method that addresses the problems illustrated by Kang and Schafer (2007). The additional robustness of the new method is achieved through a novel combination of the component estimation equations of the AIPW that is equivalent to the projection onto the largest linear subspace spanned by the component estimation equations (Tsiatis 2006, pp. 43–48). Previous works has focused on the problem of estimating the mean of the outcome where data are MAR, or in causal inference problems where the average treatment effect is of interest. In contrast, the method proposed here can be applied to more general missingness problems, where the parameter of interest need not be restricted to a mean or treatment effects parameter. Specifically, our method can be used to estimate regression coefficient parameters in the regression setup with MAR data. Even though doubly robust estimation is not new (see, e.g., Graham, Pinto, and Egel 2012, and references therein); however, we have proposed a new method which is more efficient than Robins et al.'s double robust estimator when the working regression model is misspecified.

2. MAIN RESULTS

Consider the situation where in the absence of missing data one would observe a random sample $V_1, \ldots, V_n$, where each V_i is a vector, typically $V_i = (Y_i, X_i, Z_i)$. The goal is to estimate a K-dimensional finite Euclidean-valued functional $\beta = \beta(F^*)$ of the law F^* of $V = (V^{*T}, W^T)^T$ under some model $\mathcal{F}$ for the law of V when a subvector V^* of V is missing in a subsample. Let D be the indicator function of missingness with value 1 if V^* is observed and 0 otherwise. This setup is completely general and applies to a large variety of missing data problems. If we let Y_i, X_i, Z_i denote, respectively, the potential outcome, covariates of interest, and additional variables of V_i, then in missing outcome problems, V^* consists of the subset of Y_i corresponding to $D_i = 0$; in missing covariate problems, V^* consists of the subset of X_i corresponding to $D_i = 0$. If we assume Z_i to be always observed and let Z_i contain information on auxiliary variables, then Z_i may be applied as proxies of missing outcome or covariates. Finally in treatment effects/causal inference models, $D_i = 0$ for all subjects, Z_i is a 0-1 treatment indicator, $Y_i = (1 - Z_i)Y_{i0} + Z_i Y_{i1}$ and V^* is the set of the unobserved outcomes $(Y_{10}, Y_{11}), \ldots, (Y_{i0}, Y_{i1}), \ldots, (Y_{n0}, Y_{n1})$.

In the absence of missing data, we assume there exists a $K \times 1$ unbiased estimating function $U(\beta) = U(V; \beta)$ so that we can obtain a full-data consistent and asymptotically normal estimator $\hat{\beta}_f$ of β under model $\{\mathcal{F}\}$ solving

$$E_n\{U(\hat{\beta}_f)\} = 0, \tag{2.1}$$

where E_n is the empirical mean operator based on V and 0 represents a zero vector. In the sequel, we use 0 to represent either a numeric zero, a zero vector or null matrix, where appropriate. Under MAR,

$$P(D = 1|V) = P(D = 1|W). \tag{2.2}$$

In practice, we may postulate a model

$$\omega(W; \eta) = P(D = 1|W, \eta), \tag{2.3}$$

where η is a p-dimensional vector parameter. The unknown parameter η can be estimated by the maximum likelihood estimator $\hat{\eta}$, obtained as the solution to

$$E_n\left[D\frac{\omega_\eta(W; \hat{\eta})}{\omega(W; \hat{\eta})} - (1 - D)\frac{\omega_\eta(W; \hat{\eta})}{1 - \omega(W; \hat{\eta})}\right] = 0,$$

where $\omega_\eta(W; \hat{\eta}) = \partial\omega(W; \eta)/\partial\eta|_{\eta=\hat{\eta}}$. Assuming $\omega(W; \eta)$ correctly specifies $P(D = 1|W)$, Robins, Rotnitzky, and Zhao (1994) defined a class of AIPW estimators $\hat{\beta}_{\text{AIPW}}$ by solving the following augmented estimating equation

$$E_n\left[\frac{D}{\omega(W, \hat{\eta})}U(\hat{\beta}_{\text{AIPW}}) - A\big(W, q(W; \hat{\beta}_{\text{AIPW}})\big)\right] = 0, \tag{2.4}$$

where

$$A(W; q) = \frac{D - \omega(W; \hat{\eta})}{\omega(W; \hat{\eta})}q(W; \beta)$$

and $q \equiv q(W; \beta)$ is any arbitrary $K \times 1$ function of W and β. Robins, Rotnitzky, and Zhao (1994) showed that, if model (2.3) is correct, then with no additional assumptions on the distribution of the data, all consistent and asymptotically normal estimators are derived from estimating equations of the form (2.4). The optimal $q(W; \beta)$ leading to the smallest asymptotic variance is $q_{\text{opt}}(W; \beta) = E[U(\beta)|W]$, where E denotes expectation taken under F^*. Robins, Rotnitzky, and Zhao (1994) demonstrated that estimators derived using estimation equations of the form (2.4) possess a "double robustness" property, that is, estimators are consistent if either the propensity score model (2.3) is correct or the outcome regression $q(W; \beta) = q_{\text{opt}}(W; \beta)$, given the observed data. When both (2.3) and $q(W; \beta)$ are correct, then using (2.4) leads to semiparametric locally efficient estimators. Robins, Rotnitzky, and Zhao (1994) proposed using maximum likelihood (ML) for estimating η. In the context of treatment effects and mean outcome estimation, Rubin and van der Laan (2008) and Tan (2008) proposed using weighted least squares to estimate η, which lead to an estimator with minimum asymptotic variance when $q(W; \beta) \neq q_{\text{opt}}(W; \beta)$. Other proposals that use the augmented estimating Equation (2.4) differ mainly in their methods of estimating the parameters η and β. For simplicity, in the following, we refer the estimator of Robins, Rotnitzky, and Zhao (1994) as AIPW.

In practice, without any knowledge on the distribution of W, $E[U(\beta)|W]$ is unknown. We may postulate a "working regression model" $h(W; \beta, \gamma)$ where γ is an extra parameter characterizing the relationship between W and other variables. If $h(W; \beta, \gamma) \neq E[U(\beta)|W]$, then using (2.4) no longer leads to an efficient estimator. This is because the simple difference of the estimating functions $\{D/\omega(W; \hat{\eta})\}U(\beta)$ and $A(W, q)$ may not produce the optimal combination of estimating equations.

Based on Godambe's (1960) optimal estimating function theory, it is not difficult to construct an optimal combination of these two estimating functions. However, the resulting optimal estimating equation may not have the double robustness property.

Intuitively, combining $\{D/\omega(W;\hat{\eta})\}U(\beta)$ and $A(W,q)$ is equivalent to combining the estimating functions

$$\frac{D}{\omega(W;\hat{\eta})}U(\beta) - A(W;q) \quad \text{and} \quad A(W,q). \tag{2.5}$$

Since AIPW has the double robustness property, it is possible to make the combined estimating equations inherit the same property.

The score estimating function for η can be written as

$$A(W;\tilde{q}) = \{D - \omega(W;\hat{\eta})\}\frac{\omega_\eta(W;\hat{\eta})}{\omega(W;\hat{\eta})\{1-\omega(W;\hat{\eta})\}},$$

where $\tilde{q} \equiv \tilde{q}(W) = \omega_\eta(W;\hat{\eta})/\{1-\omega(W;\hat{\eta})\}$. Define a $(K+p)$-dimensional vector $\mathbf{A}(W;q_1,q_2)^T = [A(W;q_1)^T, A(W;q_2)^T]^T$ for any q_1, q_2. We will show that the estimator $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ that solves

$$E_n\left[\frac{D}{\omega(W;\hat{\eta})}U(\hat{\beta}_{\text{AIPW}_{\text{new}}}) - A(W;\hat{q}) - \hat{\tau}^T\mathbf{A}(W;\hat{q},\tilde{q})\right] = 0 \tag{2.6}$$

has the double robustness property by using a careful choice of a $K \times (K+p)$ matrix $\hat{\tau} = \hat{B}_1\hat{B}_2^{-1}$, where $\hat{q} \equiv \hat{q}(W,\hat{\beta})$ is defined in Step 2 below and

$$\hat{B}_1 = n^{-1}\sum_{i=1}^{n}\frac{D_i\{1-\omega(W_i;\hat{\eta})\}\{U_i(\hat{\beta}) - \hat{q}(W_i;\hat{\beta})\}}{\omega^2(W_i;\hat{\eta})}\begin{bmatrix}\hat{q}(W_i;\hat{\beta})\\ \tilde{q}(W_i)\end{bmatrix}$$

$$\hat{B}_2 = n^{-1}\sum_{i=1}^{n}\frac{\{D_i - \omega(W_i,\hat{\eta})\}^2}{\omega^2(W_i,\hat{\eta})}\begin{bmatrix}\hat{q}(W_i;\hat{\beta})^T\\ \tilde{q}(W_i)^T\end{bmatrix}[\hat{q}(W_i;\hat{\beta}), \tilde{q}(W_i)].$$

We can observe two properties:

1. When model (2.3) is correct, $\hat{\tau}$ converges in probability to the least squares coefficient τ^* in the population regression of $\{D/\omega(W;\eta^*)\}U(\beta^*) - A(W;q^*)$ on $\mathbf{A}(W;q^*,q^{**})$ where β^* is the true value of β, $q^* = h(W;\beta^*,\eta^*)$, $q^{**} = \omega_\eta(W;\eta^*)/\{1-\omega(W;\eta^*)\}$.
2. When $q(W;\beta) = E[U(\beta)|W]$, $\hat{\tau}$ converges to 0 in probability.

Before we present the large sample properties of the proposed method, we summarize the overall approach. The method proposed here is a three-step procedure:

Step 1. A parametric model $\omega(W;\eta)$ for the propensity score function is fitted using the data. A commonly used model here is a logit model linear in W. This step gives $\omega(W;\hat{\eta})$.

Step 2. We require a working model $\hat{q}(W;\beta) \equiv h(W;\beta,\hat{\gamma})$ for $\text{E}[U(\beta)|W]$. A natural choice is to consider a regression model of $U(\beta)$ as a function of W. Then based on an initial estimate of β, say $\hat{\beta}$, we find $\hat{q}(W;\hat{\beta}) = h(W;\hat{\beta},\hat{\gamma})$ by regressing $U(\hat{\beta})$ on W. For example, we can use either the modified Cao, Tsiatis, and Davidian's (2009) method $\min_\gamma \sum_{i=1}^n d_i(1-w_i)/w_i^2[U_i(\hat{\beta}) - h(w_i,\hat{\beta},\gamma)]^2$ or the weighted least-square method $\min_\gamma \sum_{i=1}^n (d_i/w_i)[U_i(\hat{\beta}) - h(w_i,\hat{\beta},\gamma)]^2$.

Step 3. Using $\hat{q}(W;\hat{\beta})$ from Step 2, we can find $\hat{\tau}$. Finally, the estimate of β is estimated using (2.6). These three steps can be iterated until convergence.

We now give the large sample properties of our method:

Theorem 1. Under the same regularity conditions specified in Robins, Rotnitzky, and Zhao (1994).

(i) If either $\text{P}(D=1|W)$ or $E[U(\beta)|W]$ is correctly specified, then the estimator $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ is a consistent and asymptotically normal estimator for β^*, the true value of β.

(ii) If $\text{P}(D=1|W)$ is correctly specified, then for a "working regression function" $q(W;\beta)$(which need not be the same as $E[U(\beta) \mid W]$), $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ has asymptotic variance as least as small as those of the IPW estimator $\hat{\beta}_{\text{IPW}}$ and AIPW estimator $\hat{\beta}_{\text{AIPW}}$. There are no general results on the comparison of the asymptotic variances between $\hat{\beta}_{\text{IPW}}$ and $\hat{\beta}_{\text{AIPW}}$.

(iii) When both $\text{P}(D=1|W)$ and $E[U(\beta)|W]$ are correctly specified, then $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ and $\hat{\beta}_{\text{AIPW}}$ have the same asymptotic variance and both are asymptotically semiparametric efficient.

Proof.

(i) We first assume $\text{P}(D=1|W)$ is correctly specified. We note that

$$A(W;\hat{q}) - \hat{\tau}^T\mathbf{A}(W;\hat{q},\tilde{q}) = \frac{D-\omega(W;\hat{\eta})}{\omega(W;\hat{\eta})}\left[\hat{q} - \hat{\tau}_1^T\hat{q} - \hat{\tau}_2^T\tilde{q}\right],$$

which is in the form of $A(W;q)$ for $q = \hat{q} - \hat{\tau}_1^T\hat{q} - \hat{\tau}_2^T\tilde{q}$. Hence, (2.6) is in the form of (2.4) and $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ is consistent and asymptotically normal by the proof in Robins, Rotnitzky, and Zhao (1994).

On the other hand, if $q(W;\beta) = E[U(\beta)|W]$, then by the second property stated before the theorem, $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ is asymptotically the same as $\hat{\beta}_{\text{AIPW}}$ and the claim is satisfied.

(ii) When model (2.3) is correct, $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ has the same limiting distribution as the estimator that solves

$$E_n\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta) - A(W;q^*) - \tau^{*T}\mathbf{A}(W;q^*,q^{**})\right] = 0.$$

Since the left-hand second and third terms in the above equation are of the form $A(W;q)$ for $q = q^* - \tau_1^{*T}q^* - \tau_2^{*T}q^{**}$, the following expansion around the true value β^* of β follows from Robins, Rotnitzky, and Zhao (1994)

$$\sqrt{n}(\hat{\beta}_{\text{AIPW}_{\text{new}}} - \beta^*) = I^{-1}\left[E_n\left\{\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*) - \tau^{*T}\mathbf{A}(W;q^*,q^{**}) - \nu^{*T}A(W;q^{**})\right\}\right] + o_p(1),$$

where ν^* is the least-squares coefficient of the regression of $\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*) - \tau^{*T}\mathbf{A}(W;q^*,q^{**})$ on $A(W;q^{**})$ and $I = E[\frac{\partial}{\partial\beta}U(\beta)|_{\beta=\beta^*}]$.

For any S_1, S_2, let $\Lambda(S_1|S_2)$ denote the least-squares projection of S_1 on S_2 and write $M = \frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)$. Since $A(W;q^*)$ and $A(W;q^{**})$ are both in the space spanned by $\mathbf{A}(W;q^*,q^{**})$, we can write

$$\begin{aligned}
&M - \tau^{*T}\mathbf{A}(W;q^*,q^{**}) - \nu^{*T}A(W;q^{**}) \\
&\quad = M - \Lambda\left[M\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right] \\
&\qquad - \Lambda\left[M - \Lambda\left[M\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right]\middle|\,A(W;q^{**})\right] \\
&\quad = M - \Lambda\left[M\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right] \\
&\quad = \frac{D}{\omega(W;\hat{\eta})}U(\beta^*) \\
&\qquad - \Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right],
\end{aligned}$$

where the second identity is because $M - \Lambda[M|\mathbf{A}(W;q^*,q^{**})]$ is orthogonal to $A(W;q^{**})$ and the third identity is because the residual from the the projection of $\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)$ on the space spanned by $\mathbf{A}(W;q^*,q^{**})$ is the same as that from the projection of $\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)$ on the space spanned by $\mathbf{A}(W;q^*,q^{**})$.

The asymptotic variance of $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ is

$$\begin{aligned}
&I^{-1}\text{var}\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\right. \\
&\left.- \Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right]\right]I^{-1}
\end{aligned}$$

which cannot be larger than the asymptotic variance of $\hat{\beta}_{\text{IPW}}$, which is

$$\begin{aligned}
&I^{-1}\text{var}\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\right. \\
&\left.- \Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\,\middle|\,A(W;q^{**})\right]\right]I^{-1}
\end{aligned}$$

because the variance in the middle is for the residual of a projection into the smaller space spanned by $A(W;q^{**})$. It is also not larger than that of $\hat{\beta}_{\text{AIPW}}$, which has asymptotic variance

$$\begin{aligned}
&I^{-1}\text{var}\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)\right. \\
&\left.- \Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)\,\middle|\,A(W;q^{**})\right]\right]I^{-1}
\end{aligned}$$

because the variance in the middle is that of $\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)$ minus some linear combination of $A(W;q^*)$ and $A(W;q^{**})$.

This proves that the asymptotic variance of $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ is smaller than those of $\hat{\beta}_{\text{IPW}}$ and $\hat{\beta}_{\text{AIPW}}$.

(iii) When the working model is correct, then $A(W;q^*)$ is the projection of $\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)$ into the largest space spanned by all functions of the form $A(W;q)$ for any q as pointed by Robins, Rotnitzky, and Zhao (1994), hence

$$\begin{aligned}
&\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - \Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*)\,\middle|\,\mathbf{A}(W;q^*,q^{**})\right] \\
&\quad = \frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)
\end{aligned}$$

and likewise

$$\Lambda\left[\frac{D}{\omega(W;\hat{\eta})}U(\beta^*) - A(W;q^*)\,\middle|\,A(W;q^{**})\right] = 0.$$

Consequently, the asymptotic variances of $\hat{\beta}_{\text{AIPW}_{\text{new}}}$ and $\hat{\beta}_{\text{AIPW}}$ coincide when the working regression model is correct. Furthermore, they are semiparametric efficient due to the results in Robins, Rotnitzky, and Zhao (1994).

3. SIMULATION STUDY

In this section, we report the results of a Monte Carlo simulation study to evaluate the finite sample performance of the proposed estimator. Three separate sets of simulations were carried out. In the first set of simulations, the goal was to estimate the mean value of the outcome in a regression model. In the second set of simulations, the goal was to estimate the regression parameters. For the third set of simulations, we study the finite sample higher order bias. In each set of simulations, we compare the estimator to several alternative estimators.

3.1 Estimation of Mean

For the first set of simulations, the alternative estimators we consider are the inverse probability weighted method, henceforth IPW; the parametric imputation estimator as described in Rubin (1977), henceforth PI; the nonparametric regression estimator of Chen, Hong, and Tarozzi (2008), henceforth CHT; the augmented inverse probability weighted estimator of Robins, Rotnitzky, and Zhao (1994), henceforth AIPW; the projection estimator of Cao, Tsiatis, and Davidian (2009), henceforth CTD; the inverse probability tilting estimator of Graham, Pinto, and Egel (2012), henceforth IPT. We refer the estimator proposed in this article as AIPW_{new}.

The first set of simulations was set up in the following way. For the ith observation, let $Z_i = (Z_{i1}, Z_{i2}, Z_{i3}, Z_{i4})$ be generated from a standard multivariate normal distribution and let e_i be a standard normal deviate independent of Z_i. Furthermore, let $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})$ be the observed covariate vector. We let Y_i be the outcome and we allowed the outcome to be MAR for some observations and we denote the missingness indicator for the ith observation by D_i. We write $a = (a_0, a_1, a_2, a_3, a_4)^T$ and $b = (b_0, b_1, b_2, b_3, b_4)^T$, then we consider four different designs for the simulation study. For Designs 1 and 3, $Y_i = b^T(1, X_i) + e_i$ whereas for Designs 2 and 4, $Y_i = b^T(1, Z_i) + e_i$. For Designs 1 and 2, D_i is a Bernoulli variable such that $\text{logitP}(D_i = 1) = a^T(1, X_i)$, whereas for Designs 3 and 4, $\text{logitP}(D_i = 1) = a^T(1, Z_i)$. In Design 1, $X_i = Z_i$ and in Designs 2 to 4, we let $X_{i1} = \exp(Z_{i1}/2)$, $X_{i2} = Z_{i2}/\{1 + \exp(Z_{i1})\} + 10$, $X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$ and $X_{i4} = (Z_{i2} + Z_{i4})^2$. These models and the values of a and b

Table 1. Models and values of the parameters for the four designs used in the simulation study

	Design 1	Design 2	Design 3	Design 4
Y_i	$b^T(1, X_i) + e_i$	$b^T(1, Z_i) + e_i$	$b^T(1, X_i) + e_i$	$b^T(1, Z_i) + e_i$
$\text{logit}D_i$	$a^T(1, X_i)$	$a^T(1, X_i)$	$a^T(1, Z_i)$	$a^T(1, Z_i)$
a_0	0	−65	0	0
a_1	−1	−1.7	−1	−1
a_2	0.5	0.87	0.5	0.5
a_3	−0.25	0.23	−0.25	−0.25
a_4	−0.1	0	−0.1	−0.1
b_0	210	210	35.4	210
b_1	27.4	27.4	51.3	27.4
b_2	13.7	13.7	0	13.7
b_3	13.7	13.7	−130	13.7
b_4	13.7	13.7	0	13.7

used for each design are summarized in Table 1. These designs are similar to the simulation set-ups used in Kang and Schafer (2007), Cao, Tsiatis, and Davidian (2009) and others for evaluating the performance of doubly robust estimators. For each of the four designs, we use 5000 simulation runs with $N = 1000$ observations each.

Throughout the simulations, the estimators IPW, AIPW, IPT, CTD, and AIPW_{new} use a logistic model for the propensity model that is linear in X_i. For estimators that require an outcome regression model, a model linear in X_i was used for PI, AIPW, CTD, and AIPW_{new} and a polynomial model that includes a linear as well as a quadratic term in X_i is used in CHT. Therefore, for Design 1, all estimators are expected to be unbiased; for Design 2, estimators that used only a propensity score model (IPW) or the doubly robust estimators (IPT, AIPW, CTD, AIPW_{new}) are expected to be unbiased; for Design 3, estimators that only use an outcome regression model (PI, CHT) and the doubly robust estimators (IPT, AIPW, CTD, AIPW_{new}) are expected to be unbiased; for Design 4, none of the estimators are expected to be unbiased.

The results of this set of simulations are given in Tables 2 and 3. Recall that the goal of this study is to estimate the mean value of the outcome. For each estimator, we calculate the following statistics: (1) the median bias (Bias), which is defined as the Monte Carlo median of the difference between an estimate and the true value; (2) the median standard error (MSE), which is defined as the Monte Carlo standard error based on the asymptotic sandwich estimator; (3) the median absolute deviation (MAE), which is defined as the Monte Carlo median of the absolute value of the difference between an estimate and the true value; (4) the standard deviation (SD), which is defined as the Monte Carlo standard deviation of the estimates; (5) the root mean squared error (RMSE), which is defined as the square root of the Monte Carlo mean squared error and (6) the 95% coverage probability (COV) which is defined as the proportion of times the true value falls inside the 95% confidence interval based on estimate $\pm 1.96\tilde{}$SE.

Table 2, top panel, shows the results for Design 1. In this design, all estimators are expected to be unbiased in theory. However, the results show some bias for IPW, and so for this estimator, even though its standard error (as seen by the values of MSE) is usually considerably higher than those of the other estimators, there is still under coverage by its 95% confidence interval; furthermore, IPW has a much bigger SE than the other estimators. The results for the other estimators are similar; all are approximately unbiased and the coverages are all close to the 95% level.

Table 2. Simulation results based on 5000 Monte Carlo replications for Simulation Study 1 (Designs 1 and 2) with sample size $N = 1000$ each

Estimator	Bias	MSE	MAE	SD	RMSE	COV
			Design 1			
IPW	−0.175	1.496	1.133	1.749	1.750	0.930
IPT	0.011	1.138	0.795	1.165	1.165	0.943
CHT	0.010	1.145	0.791	1.164	1.164	0.944
PI	0.009	1.145	0.794	1.164	1.164	0.945
AIPW	0.010	1.145	0.794	1.165	1.165	0.944
CTD	0.014	1.145	0.796	1.165	1.165	0.944
AIPW_{new}	0.013	1.145	0.795	1.165	1.165	0.945
			Design 2			
IPW	−0.437	1.395	1.239	2.564	2.570	0.908
IPT	0.355	1.108	0.938	1.319	1.363	0.873
CHT	−1.023	1.340	1.266	1.556	1.870	0.846
PI	2.279	1.592	2.280	1.449	2.697	0.724
AIPW	0.546	1.493	1.297	3.302	3.307	0.917
CTD	0.186	1.265	0.969	1.425	1.438	0.921
AIPW_{new}	0.267	1.350	0.921	1.331	1.357	0.954

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

In Design 2 (Table 2, bottom panel), PI is expected to perform poorly and its poor performance is confirmed by the simulation results. All the other estimators are theoretically unbiased but all show some biases with the most serious bias seen in CHT. For

Table 3. Simulation results based on 5000 Monte Carlo replications for Simulation Study 1 (Designs 3 and 4) with sample size $N = 1000$ each

Estimator	Bias	MSE	MAE	SD	RMSE	COV
			Design 3			
IPW	6.263	3.771	6.263	23.08	26.28	0.757
IPT	0.063	1.081	0.679	1.010	1.012	0.963
CHT	0.080	0.990	0.677	1.012	1.015	0.948
PI	0.066	0.989	0.678	1.010	1.012	0.949
AIPW	0.072	0.993	0.692	1.780	1.780	0.948
CTD	0.066	0.989	0.680	1.011	1.013	0.949
AIPW_{new}	0.066	0.992	0.680	1.011	1.013	0.950
			Design 4			
IPW	2.171	2.474	2.584	11.07	12.20	0.886
IPT	−2.745	1.179	2.746	1.544	3.164	0.379
CHT	−2.192	1.356	2.199	1.542	2.717	0.631
PI	−0.802	1.658	1.190	1.500	1.692	0.949
AIPW	−5.224	2.415	5.224	293.7	294.5	0.606
CTD	−1.824	1.281	1.849	1.527	2.399	0.698
AIPW_{new}	−1.746	2.105	1.773	1.347	2.192	0.936

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

CHT, the logistic model for using a quadratic term in X_i failed to converge in a number of simulation runs, which indicates that including a quadratic term actually hurts in this case. The coverage for CHT suffers as a result. The standard error of IPT (as seen by MSE) is much smaller than those of the other estimators. This leads to a much shorter 95% confidence for IPT and subsequently for this estimator, there is significant under coverage. All the other doubly robust estimators (AIPW, CTD, AIPW_{new}) perform quite well in this case, with AIPW_{new} the best performance in terms of matching the coverage to the 95% level. Table 3, top panel, shows the results when it is not possible from the observed data to correctly identify model the propensity score function. In this situation, estimators that use a wrong propensity score model should perform poorly, as reflected in the results for IPW, which shows significant bias, as well as a very high SE and poor coverage. AIPW, though unbiased in this case, is much less efficient than the other estimators, as evidenced by its large SE and RMSE. The performance of the other estimators are similar but there is some over coverage in IPT due to its relatively large MSE.

None of the estimators is expected to be unbiased in Design 4 (Table 3, bottom panel); IPT, AIPW, CTD, and CHT performed particularly poorly. The best performances are seen in PI and AIPW_{new}.

In addition, we also calculated the semiparametric variance bounds based on Robins, Rotnitzky, and Zhao (1994), assuming the correct model is used, as benchmarks to evaluate the different methods. For the four designs in this section, the variance bounds are: 1.15, 1.15, 0.99, 1.15, respectively. These bounds are relevant only when the estimators correctly use the right model, as in the cases of Designs 1 and 3. Comparing the estimators to the variance bounds, we observe that for all estimators except IPW and AIPW, the SDs are close to variance bounds for both Designs 1 and 3. The well-known inefficiency of IPW is reflected in Design 1; for Design 3, IPW is biased by design. For AIPW, the results are close to the variance bound for Design 1 but not in Design 3.

We also carried out a set of simulations based on models originally considered in Kang and Schafer (2007). These models mimic situations when one or both of the propensity score function and the outcome regression model is slightly misspecified. Kang and Schafer (2007) showed that some estimators performed poorly under these models. Hence, this set of simulations is useful to study the robustness of the proposed method and how it compares to other estimators. The models used by Kang and Schafer (2007) are similar to those in the first set of simulations. Let Z_i, D_i, e_i, a, and b be defined as in the first set of simulations, and $X_{i1} = \exp(Z_{i1}/2)$, $X_{i2} = Z_{i2}/\{1 + \exp(Z_{i1})\} + 10$, $X_{i3} = (Z_{i1}Z_{i3}/25 + 0.6)^3$ and $X_{i4} = (Z_{i2} + Z_{i4} + 20)^2$. Furthermore, let $Y_i = b^T(1, Z_i) + e_i$ and $\text{logitP}(D_i = 1) = a^T(1, Z_i)$. The values of a and b are the same as those for Design 1 in Table 1. Four different situations were used in Kang and Schafer (2007), that correspond to (1) both the outcome regression model and the propensity score function are correctly fitted using Z_i, (2) only the propensity function is correctly fitted using Z_i, (3) only the outcome regression model is correctly fitted using Z_i, and (4) both models incorrectly fitted using X_i. We called these Designs 1', 2', 3', and 4'. Incidentally, Design 1' is the same as Design 1, but we include Design 1' here for completeness. This set of simulation also uses 5000 simulation runs of each situation, based on 1000 observations each.

Table 4. Simulation results based on 5000 Monte Carlo replications for Simulation Study 1' (Designs 1' and 2') with sample size $N = 1000$ each

Estimator	Bias	MSE	MAE	SD	RMSE	COV
			Design 1'			
IPW	0.013	1.145	0.794	1.163	1.163	0.945
IPT	0.010	1.138	0.795	1.165	1.165	0.943
CHT	0.009	1.145	0.791	1.164	1.164	0.944
PI	0.008	1.145	0.794	1.164	1.165	0.945
AIPW	0.009	1.145	0.794	1.165	1.165	0.944
CTD	0.012	1.149	0.796	1.165	1.165	0.946
AIPW_{new}	0.011	1.145	0.795	1.165	1.165	0.945
			Design 2'			
IPW	−0.066	1.496	1.105	1.721	1.721	0.943
IPT	−0.022	1.151	0.774	1.145	1.146	0.950
CHT	−1.931	1.197	2.313	1.494	2.759	0.510
PI	−0.492	1.658	1.167	1.496	1.698	0.950
AIPW	0.073	1.448	1.063	1.658	1.660	0.947
CTD	0.022	1.169	0.777	1.156	1.156	0.953
AIPW_{new}	0.073	1.602	0.865	1.261	1.268	0.986

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

The results of simulation study (Study 1') are given in Tables 4 and 5. The top panel of Table 4 shows similar results as those from Design 1, as expected. The bottom panel of Table 4 shows that both CHT and PI are biased, as to be expected because both methods fit a wrong outcome regression model and neither is doubly robust. The confidence interval coverage of CHT is severely affected by the bias, whereas, the moderate

Table 5. Simulation results based on 5000 Monte Carlo replications for Simulation Study 1' (Designs 3' and 4') with sample size $N = 1000$ each

Estimator	Bias	MSE	MAE	SD	RMSE	COV
			Design 3'			
IPW	0.862	2.455	2.536	9.850	10.93	0.880
IPT	−0.011	1.140	0.791	1.165	1.165	0.947
CHT	−0.012	1.146	0.791	1.165	1.165	0.946
PI	−0.013	1.146	0.788	1.165	1.165	0.947
AIPW	−0.011	1.150	0.796	2.392	2.392	0.947
CTD	−0.011	1.150	0.790	1.165	1.165	0.948
AIPW_{new}	−0.010	1.149	0.789	1.165	1.166	0.948
			Design 4'			
IPW	0.878	2.474	2.584	11.07	12.20	0.886
IPT	−2.328	1.179	2.746	1.544	3.164	0.379
CHT	−1.869	1.197	2.240	1.510	2.740	0.527
PI	−0.484	1.658	1.190	1.500	1.692	0.949
AIPW	−2.163	2.415	5.224	293.7	294.5	0.606
CTD	−0.931	1.348	1.328	1.268	1.772	0.873
AIPW_{new}	−0.829	2.105	1.773	1.347	2.192	0.936

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

bias of PI has no effect on the coverage, probably because of its large MSE. All other methods perform satisfactorily, in all dimensions considered. However, for AIPW_{new}, there is moderate over coverage. For Design 3', all methods except IPW perform satisfactorily. The performance of IPW is especially bad, with a large bias, large RMSE and its confidence interval severely conservative. AIPW, though being unbiased, has RMSE quite a bit larger than those of the remaining methods. In Design 4', no method is supposed to give unbiased estimate. However, the data affect the estimators quite differently. The bias of IPT is quite large here, and the confidence interval coverage is severely compromised; the results here is similar to IPT's performance in Design 4. IPW, CHT, and AIPW all perform poorly, in one dimension of another. There is moderate under coverage of CDT; the two methods that have the best overall performance are PI and AIPW_{new}. We note in passing that in these simulations, PI seems to perform quite satisfactorily. This is because these simulations were designed to mimic situations where the regression outcome model is only slightly misspecified, and so PI, which uses a regression model, is not severely affected.

The variance bounds for Designs 1'–4' are all 1.15. A similar pattern emerges for the different estimators, that is, for Designs 1' and 3', all estimators except IPW and AIPW perform well.

3.2 Estimation of Regression Parameters

In this set of simulations, for the ith observation, the outcome Y_i is generated using a linear model $Y_i = \beta^T(1, X_i, Z_i) + e_i$, where X_i and e_i are independent standard normal random variables, Z_i is a binary random variable with $\text{P}(Z_i = 1) = 0.5$ and $\beta \equiv (\beta_0, \beta_1, \beta_2)^T$ is the vector of regression parameters. We allow Z_i to be MAR for some observations. The missingness indicator for the ith observation, D_i is a Bernoulli variable such that $\text{logitP}(D_i = 1) = a^T(1, Y_i, X_i)$, where $a = (a_0, a_1, a_2)^T$. Throughout this set of simulations, we used $\beta = (-2, 1, 2)$ and we considered four different combinations of the values of $a = (-1, 0, 0), (-1, 0.2, 0.2), (-1, 0.4, 0.4), (-1, 0.6, 0.6)$. We call these Designs 5–8 and they correspond to different levels of dependency of the missingness on the observed data, ranging from missing completely at random to heavily MAR.

For this set of simulations, we compare AIPW_{new} to four competitors (IPW, AIPW, IPT, and CHT) that are designed to estimate regression parameters. For all estimators, we use a logistic model for the propensity model that is linear in Y_i and X_i. For AIPW and AIPW_{new}, we used an outcome regression model a logistic outcome regression model that is linear in Y_i and X_i and then we calculated $h(W; \beta, \hat{\gamma}) = \hat{q}(W; \beta)$ based on this logistic outcome regression model. Note that this logistic outcome regression model is not the correct model so we do not expect methods to be efficient. For all methods, we allow the form of the propensity score be known with the parameters estimated using the data. More details are given in a set of online supplementary materials, where additional simulation results are also given.

Once again, we use 5000 simulation runs with $N = 1000$ observations in each simulation run. The results of this set of simulations are recorded in Tables 6 and 7. For each regression parameter, we calculate the same statistics for each estimators, as we did in the first set of simulations, that is, Bias, MSE, MAE, SD, RMSE and COV. For Designs 5-6, IPW, IPT, and CHT are considerably less efficient than either AIPW or AIPW_{new}. Both AIPW and AIPW_{new} are similar for these two designs. All five estimators are approximately median unbiased for these designs. For Designs 7 and 8, however, when the estimated propensity scores for some of the observations can be close to zero, all three designs are affected in the sense that the median bias is non-negligible. The estimator AIPW stands out in having some very large Monte-Carlo SDs (Design 7), while for the remaining estimators, the order of efficiency are AIPW_{new}, IPT, IPW, and then CHT. Of note in Designs 7 and 8 is, except AIPW_{new}, the coverage probability of all estimators are quite seriously biased.

The semiparametric variance bounds for $\beta \equiv (\beta_0, \beta_1, \beta_2)^T$ are 0.051, 0.033, and 0.077, respectively. We notice that the relative efficiency of all methods decreases as the degree of MAR increases. Across all methods, AIPW_{new} again performs the best, but even so, for Design 8, its relative efficiency is still low. It is interesting to observe that whereas the propensity score is ancillary to the semiparametric variance bound for the

Table 6. Simulation results based on 5000 Monte Carlo replications for Simulation Study 2 (Designs 5 and 6) with sample size $N = 1000$ each

Parameter	Estimator	Bias	MSE	MAE	SD	RMSE	COV
				Design 5			
β_0	IPW	−0.003	0.107	0.055	0.078	0.078	0.993
	AIPW	−0.003	0.065	0.044	0.065	0.065	0.947
	AIPW_{new}	−0.006	0.075	0.043	0.064	0.064	0.976
	IPT	−0.003	0.078	0.054	0.078	0.078	0.949
	CHT	−0.003	0.086	0.055	0.079	0.079	0.970
β_1	IPW	0.000	0.061	0.042	0.063	0.063	0.949
	AIPW	−0.001	0.046	0.033	0.048	0.048	0.939
	AIPW_{new}	−0.001	0.052	0.033	0.048	0.048	0.961
	IPT	0.000	0.060	0.042	0.062	0.062	0.944
	CHT	0.000	0.061	0.043	0.063	0.063	0.942
β_2	IPW	0.002	0.123	0.083	0.123	0.123	0.946
	AIPW	0.006	0.089	0.062	0.091	0.091	0.947
	AIPW_{new}	0.009	0.107	0.058	0.086	0.087	0.980
	IPT	0.003	0.121	0.083	0.123	0.123	0.943
	CHT	0.004	0.122	0.084	0.124	0.124	0.946
				Design 6			
β_0	IPW	0.001	0.128	0.065	0.097	0.097	0.989
	AIPW	−0.003	0.068	0.047	0.069	0.069	0.945
	AIPW_{new}	−0.005	0.084	0.045	0.067	0.067	0.978
	IPT	0.000	0.086	0.061	0.088	0.088	0.945
	CHT	0.002	0.123	0.070	0.104	0.104	0.978
β_1	IPW	−0.002	0.072	0.049	0.074	0.074	0.942
	AIPW	−0.001	0.051	0.036	0.055	0.055	0.929
	AIPW_{new}	−0.002	0.058	0.036	0.055	0.055	0.961
	IPT	−0.003	0.068	0.048	0.073	0.073	0.932
	CHT	−0.006	0.081	0.061	0.093	0.093	0.918
β_2	IPW	0.001	0.140	0.095	0.140	0.140	0.952
	AIPW	0.006	0.095	0.067	0.098	0.098	0.941
	AIPW_{new}	0.012	0.117	0.064	0.094	0.095	0.973
	IPT	0.001	0.136	0.094	0.139	0.139	0.944
	CHT	−0.003	0.145	0.107	0.158	0.158	0.933

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

Table 7. Simulation results based on 5000 Monte Carlo replications for Simulation Study 2 (Designs 7 and 8) with sample size $N = 1000$ each

Parameter	Estimator	Bias	MSE	MAE	SD	RMSE	COV
				Design 7			
β_0	IPW	0.013	0.156	0.090	0.138	0.138	0.969
	AIPW	−0.008	0.073	0.051	0.859	0.860	0.938
	AIPW_{new}	−0.014	0.106	0.050	0.074	0.076	0.975
	IPT	0.007	0.098	0.071	0.108	0.108	0.920
	CHT	0.012	0.192	0.104	0.159	0.159	0.974
β_1	IPW	−0.013	0.086	0.068	0.104	0.104	0.911
	AIPW	0.000	0.057	0.046	0.777	0.777	0.912
	AIPW_{new}	−0.002	0.066	0.046	0.070	0.070	0.949
	IPT	−0.012	0.080	0.065	0.098	0.099	0.887
	CHT	−0.031	0.144	0.108	0.165	0.166	0.916
β_2	IPW	−0.001	0.168	0.118	0.180	0.180	0.947
	AIPW	0.023	0.103	0.077	1.495	1.495	0.928
	AIPW_{new}	0.039	0.145	0.079	0.112	0.116	0.964
	IPT	0.000	0.157	0.115	0.174	0.174	0.927
	CHT	−0.008	0.199	0.156	0.236	0.236	0.910
				Design 8			
β_0	IPW	0.042	0.183	0.132	0.202	0.204	0.914
	AIPW	−0.012	0.081	0.060	0.179	0.180	0.915
	AIPW_{new}	−0.021	0.159	0.060	0.094	0.097	0.978
	IPT	0.018	0.112	0.097	0.143	0.144	0.872
	CHT	0.036	0.292	0.155	0.240	0.242	0.956
β_1	IPW	−0.036	0.100	0.096	0.142	0.145	0.837
	AIPW	−0.006	0.062	0.059	0.151	0.151	0.883
	AIPW_{new}	−0.011	0.078	0.059	0.092	0.092	0.926
	IPT	−0.032	0.099	0.091	0.134	0.138	0.799
	CHT	−0.085	0.324	0.179	0.257	0.266	0.972
β_2	IPW	−0.016	0.200	0.158	0.243	0.243	0.914
	AIPW	0.042	0.113	0.094	0.248	0.251	0.888
	AIPW_{new}	0.071	0.202	0.105	0.149	0.161	0.947
	IPT	−0.007	0.182	0.151	0.231	0.231	0.890
	CHT	−0.033	0.284	0.220	0.337	0.338	0.900

NOTES: Bias: Median bias. MSE: Median standard error using asymptotic formula. MAE: Median absolute deviation. SD: Monte Carlo standard deviation. RMSE: Root mean squared error. COV: Monte Carlo coverage of 95% confidence intervals.

estimation of β (see, e.g., Hahn 1998), the level of precision to which the methods can estimate the unknown parameters is affected by the propensity score in finite sample problems.

3.3 Higher Order Bias

When both the outcome regression model and the propensity score functions are correctly specified, the four doubly robust estimators, i.e., AIPW, AIPW_{new}, IPT, and CTD, are all equivalent and they are all semiparametric equivalent, to order $O(N^{-1/2})$. Hence, one way to compare and contrast the estimators is to study their higher order bias behavior. Using results in Newey and Smith (2004), Graham, Pinto, and Egel (2012, Theorem 4.1) developed higher order bias expressions for their IPT estimator and the general doubly robust estimators. We will use these results to study the finite sample higher order of AIPW_{new} and compare the results to other estimators here. The results in Graham, Pinto, and Egel (2012) are derived under the assumptions that the propensity score function and outcome regression model can be correctly modeled (Assumption 1.5 and Assumption 2.1 in their article). However, we would like to compare the estimators under the scenarios when both the propensity score and outcome regression models are correctly specified, and also when only one of them is correctly specified. Hence, we derived the higher order bias expressions here for the four estimators under these more general situations (see the Appendix).

We focused on the data-generating mechanisms considered in simulation study 1' in this section. Following Newey and Smith (2004, Lemma A4), suppose θ is a L-dimensional vector of parameters defined by a set of estimating equations

$$E\{h(\theta)\} = 0, \tag{3.1}$$

and $\hat{\theta}$ is the solution to the equations

$$E_n\{h(\hat{\theta})\} = \sum_{i=1}^{N} h_i(\hat{\theta}) = 0. \tag{3.2}$$

Then under suitable regularity conditions as defined in Newey and Smith (2004), the higher order bias, to $O(n^{-1})$, is given by

$$\text{Bias}(\hat{\theta}) = \frac{-H^{-1}}{N}\left(\tilde{Q}\tilde{\phi} + \frac{1}{2}\sum_{l=1}^{L}\tilde{\phi}_l H_l \tilde{\phi}\right), \tag{3.3}$$

where

$$H = E\left(\frac{\partial h_i(\theta)}{\partial \theta^T}\right), \quad H_l = E\left(\frac{\partial^2 m_i(\theta)}{\partial \theta_l \partial \theta^T}\right),$$

$$\phi_i(\theta) = -H^{-1}h_i(\theta), \quad \tilde{\phi} = \frac{1}{N}\sum_{i=1}^{N}\phi_i(\theta),$$

$$Q_i = \frac{\partial h_i(\theta)}{\partial \theta^T} - H, \quad \tilde{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i,$$

and $\tilde{\phi}_l$ stands for the lth row of $\tilde{\phi}$. Note that (3.3) applies for doubly robust estimators as long as one or both of the outcome regression model or propensity score function is correctly specified, but not when both models are incorrect. Since all four doubly robust estimators can be written as the solution to estimating equations of the form (3.2), we can use (3.3) to study their higher order bias.

The ingredients for $\text{Bias}(\hat{\theta})$ for the four doubly robust estimators are given in the Appendix, under the designs in simulation study 1'–3'. We used simulations to verify the higher order biases. In each set of simulations, we generated $N = 500$ observations and repeated the simulation 5000 times. We evaluated the bias based on the asymptotic expression (3.3) for estimating μ and we compared them to the Monte Carlo biases, both scaled by the Monte Carlo SE. The results of these comparisons are given in Table 8. The results seem to reproduce the asymptotic bias expressions best for IPT. For this set of data, the higher order biases are similar in magnitudes between the different estimators, with no estimator dominating the others.

4. APPLICATION

In this section, we apply the method proposed in this article to a wage regression (Mincer 1974). In a wage regression, the effects of human capital on productivity are investigated using

Table 8. Higher order bias comparison for Simulation Study 1' (Designs 1', 2', and 3')

Design	Estimator	Bias_A	SE	Bias_{MC}
1'	IPT	−0.020331	1.605160	−0.017396
	AIPW	−0.017663	1.620410	−0.000100
	AIPW_{new}	−0.018226	1.620136	−0.000268
	CTD	−0.019966	1.620256	0.000012
2'	IPT	−0.019873	1.629051	−0.011349
	AIPW	0.090485	2.007442	−0.031858
	AIPW_{new}	0.007857	1.636378	−0.033795
	CTD	0.093400	1.778517	0.003977
3'	IPT	−0.019869	1.629053	−0.011349
	AIPW	−0.019344	1.623214	−0.002832
	AIPW_{new}	−0.020005	1.620506	−0.042134
	CTD	−0.020593	1.621180	−0.001617

NOTES: SE: Median SE based on Monte Carlo. Bias_A: Bias using asymptotic formula, scaled by SE. Bias_{MC}: Median bias based on Monte Carlo, scaled by SE.

a regression model, whereby the natural logarithm of a measure of wage is regressed upon education, experience, and ability. The use of experience allows economists to study the influence of education on wage, adjusting for individual differences in human capital acquired on the job. The data we use come from the 1980 wave of the National Longitudinal Surveys (NLS). The NLS sampled 5255 young men in 1966 to represent the civilian population of men aged 14–24 in the United States. The individuals selected into the NLS were followed longitudinally and were interviewed almost annually until 1981. The dataset consists of detailed information about each individual, in particular, measures of ability. However, the dataset suffers a high attrition rate such that by 1980, only 3438 (65.8%) of the men in the 1966 cohort were left in the study. To make matters worse, there is no evidence to suggest that attrition was completely at random.

We use the hourly wage as a measure of wage, and education is the highest grade completed. We use the following four variables as proxies for human capital: education, experience, IQ test score, KWW (Knowledge of the World of Work) test score. Experience is measured by age minus 6 minus years of education, IQ is coded as above or below median (104) and KWW has a range of 10–56. The wage regression equation is

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{ education} + \beta_2 \text{ experience} + \beta_3 \text{ experience}^2 + \beta_4 \text{ IQ} + \beta_5 \text{KWW} + \epsilon. \quad (4.4)$$

There are many evidences that show wage differences between race that cannot be accounted for by education, experience and ability, hence, following others, we focus our attention on the sample of 1784 white males in 1980. This sample accounts for about half of the 3438 men in the 1980 NLS because blacks were deliberately oversampled. Among the 1784 whites, a sub-sample of 1401 have complete records of IQ score. However, all individuals in that cohort also took another ability test KWW and the KWW test score is available for all 1784 men. Hence, $D = 1$ for the 1041 men with IQ information and $D = 0$ for the remaining $1784 - 1041 = 743$ men. We use five estimators, IPW, AIPW, AIPW_{new}, IPT, and CHT, to analyze wage regression model (4.1) based on the NLS data. Griliches, Hall, and Hausman (1978) showed that IQ score is not MAR and missingness may be related to age and education, among other things. Therefore, we use the inverse probability selection model $\text{logit}P(D = 1) = a^T(1, \log(\text{wage}), \text{education}, \text{experience}, \text{experience}^2, \text{KWW})$. For estimators that use a surrogate for the missing IQ score, we use a logistic model that is also linear in log(wage), experience, experience^2, and KWW.

The results of the analysis are given in Table 9. The results for all the estimators are similar. The coefficients for all the variables: education, experience, IQ, and KWW are positive, indicating a direct relationship of these on wage, which is to be expected. For experience^2, the coefficient is negative for all estimators, suggesting that influence of experience on wage does plateau off, which also seems to make sense.

5. CONCLUSION

Missing data are a common phenomenon in economics and social sciences research. In this article, we have proposed a robust and efficient method for handling missing data under a general missing data setup, which applies to a large variety of missing data problems such as missing outcome problems, missing covariate problems, and treatment effects/causal inference problems. The proposed method optimally combines working propensity score and regression functions by employing Godambe's (1960) optimal estimating function theory. The resulting estimator of the full-data model parameter is shown to be at least as efficient as the IPW estimator and augmented inverse probability weighted estimator in both large and small sample cases when the working propensity score model is correctly specified, as well as being locally semiparametric efficient when both the working propensity score model and the working regression model are correctly specified. The proposed estimator also enjoys the properties of double robustness against misspecification of either the propensity score model or the regression model. In addition, the proposed doubly robust method can be viewed as a projection approach for variance reduction in that

Table 9. Estimates (SE) of wage regression analysis based on 1784 white males from the 1980 NLS

	IPW	AIPW	AIPW_{new}	IPT	CHT
Intercept	0.7065 (0.2035)	0.6769 (0.1430)	0.6773 (0.1490)	0.7121 (0.1784)	0.6698 (0.2236)
Education	0.0488 (0.0085)	0.0480 (0.0067)	0.0478 (0.0067)	0.0515 (0.0082)	0.0514 (0.0093)
Experience	0.0646 (0.0205)	0.0611 (0.0147)	0.0613 (0.0133)	0.0598 (0.0203)	0.0729 (0.0207)
Experience^2	−0.1575 (0.0731)	−0.1553 (0.0509)	−0.1558 (0.0438)	−0.1342 (0.0749)	−0.1859 (0.0723)
IQ	0.0054 (0.0020)	0.0076 (0.0017)	0.0076 (0.0016)	0.0046 (0.0021)	0.0039 (0.0021)
KWW	0.0505 (0.0232)	0.0488 (0.0235)	0.0516 (0.0210)	0.0530 (0.0235)	0.0477 (0.0231)

the resulting estimating equation is obtained as the residual from the projection of the AIPW estimating function onto the linear space spanned by the working propensity score and regression functions. This consideration leads to the implementation of the proposed method as a three-step procedure: (1) fitting a working propensity score model using fully observed data W; (2) fitting a working regression model using complete data; (3) solving the proposed doubly robust estimating Equation (2.6) for estimating β. The proposed method compares favorably with its competitors in finite samples and is illustrated using an analysis of the data from the 1980 wave of the National Longitudinal Surveys (NLS).

Recently, Rothe and Firpo (2013) suggested using nonparametric estimators for the propensity score function ω and outcome regression function q in a double robustness set-up. They showed the assumptions on ω and q required in their method are weaker than those in classical semiparametric double robust estimators, leading to improved accuracy in drawing inference. nonparametric estimators of ω and q can also be used in the construction of our projection estimator.

For estimators that use inverse propensity scores as weights, there are always concerns about inverse weighting when the propensity scores get too close to zero. In such situations, observations with very low propensity scores that are occasionally observed in a sample create instability in the method. This problem is more severe when the propensity score function is known than when it is estimated in the sample and it does not disappear with large samples as in a large sample, there is a higher chance of such phenomenon to be observed. As pointed out by Robins et al. (2007) and Khan and Tamer (2010), when there is the possibility that some observations may assume propensity scores that are arbitrary close to zero, then in finite samples, the researcher must find a balance of choosing between a consistent estimator (at the risk of giving a totally unreasonable estimate) or an estimator that is bounded (but may not be consistent). Our method is not immune to this problem. The sensitivity analysis suggested by Robins et al. (2007) is one way of detecting this problem. When there is evidence that some of the propensity scores may be too close to zero, perhaps methods such as those in Khan and Tamer (2010) of bounding them could be employed. As Kang and Schafer (2007) pointed out, no method is foolproof and intuition and caution must always be exercised.

APPENDIX: DERIVATIONS OF THE HIGHER ORDER BIAS EXPRESSIONS FOR DOUBLY ROBUST ESTIMATORS

We will work out Bias($\hat{\theta}$) in (3.3) for the four doubly robust estimators, i.e., AIPW, AIPW$_{\text{new}}$, IPT, and CTD below, for the designs 1'–3' considered in simulation study 2 in Section 4. To economize notations, we omit references to the parameter θ or any part of it, where possible, for example, we write $\omega_i \equiv \omega(W_i;\eta) = P(D_i = 1|W_i,\eta)$. Under designs 1'–3', ω_i is fitted using a logistic model, in general, let $r_i = r(W_i)$ be a vector-valued function of W_i such that $\omega_i = \exp(\eta^T r_i)/(1+\exp(\eta^T r_i))$ so $\omega_{\eta i} = \partial\omega_i/\partial\eta = \omega_i(1-\omega_i)r_i$ and $\omega_{\eta\eta^T i} = \partial^2\omega_i/\partial\eta^2 = \omega_i(1-\omega_i)(1-\omega_i)r_i r_i^T$. Furthermore, let the regression function for the outcome model be $m_i = m(W_i) = \beta^T t(W_i) = \beta^T t_i$, $m_{\beta i} = \partial m_i/\partial\beta = t_i$ and $E(Y|W) = \mu$. We now work out the ingredients for evaluating (3.3) for the four estimators. For AIPW, $\theta = (\eta, \beta, \mu)$, and

$$h_i = \begin{pmatrix} \dfrac{D_i-\omega_i}{\omega_i(1-\omega_i)}\omega_{\eta i} \\ D_i(Y_i-m_i)m_{\beta i} \\ \dfrac{D_i}{\omega_i}(Y_i-m_i)+m_i-\mu \end{pmatrix} = \begin{pmatrix} (D_i-\omega_i)r_i \\ D_i(Y_i-m_i)t_i \\ \dfrac{D_i}{\omega_i}(Y_i-m_i)+m_i-\mu \end{pmatrix},$$

$$Q_i = \begin{pmatrix} -\omega_i(1-\omega_i)r_i r_i^T & 0 & 0 \\ 0 & -D_i t_i t_i^T & 0 \\ -\dfrac{D_i(1-\omega_i)}{\omega_i}(Y_i-m_i)r_i^T & -\dfrac{D_i-\omega_i}{\omega_i}t_i^T & -1 \end{pmatrix} - H.$$

Let p be the dimension of η, then for $H_l, l = 1,\ldots,p$, we have

$$H_l = E\begin{pmatrix} -\omega_i(1-\omega_i)(1-2\omega_i)r_{il}r_i r_i^T & 0 & 0 \\ 0 & 0 & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i}(Y_i-m_i)r_{il}r_i^T & \dfrac{D_i(1-\omega_i)}{\omega_i}r_{il}t_i^T & 0 \end{pmatrix}.$$

Let K be the dimension of β, then for $H_l, l = p+1,\ldots,p+K$, we have

$$H_l = E\begin{pmatrix} 0 & 0\,0 \\ 0 & 0\,0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i}t_{i(l-p)}r_i^T & 0\,0 \end{pmatrix}.$$

Finally,

$$H_{p+K+1} = 0.$$

For AIPW$_{\text{new}}$, $\theta = (\eta, \beta, \mu)$, and

$$h_i = \begin{pmatrix} \dfrac{D_i-\omega_i}{\omega_i(1-\omega_i)}\omega_{\eta i} \\ D_i(Y_i-m_i)m_{\beta i} \\ \dfrac{D_i}{\omega_i}(Y_i-m_i)+m_i-\mu-\hat{\tau}^T\mathbf{A}(W;\hat{q},\tilde{q}) \end{pmatrix} = \begin{pmatrix} (D_i-\omega_i)r_i \\ D_i(Y_i-m_i)t_i \\ \dfrac{D_i}{\omega_i}(Y_i-m_i)+m_i-\mu-\hat{\tau}^T\mathbf{A}(W;\hat{q},\tilde{q}) \end{pmatrix},$$

where $\hat{\tau} = \hat{B}_1\hat{B}_2^{-1}$.

$$Q_i = \begin{pmatrix} -\omega_i(1-\omega_i)r_i r_i^T & 0 & 0 \\ 0 & -D_i t_i t_i^T & 0 \\ -\dfrac{D_i(1-\omega_i)}{\omega_i}(Y_i-m_i)r_i^T - \hat{\tau}^T C_{1i} & -\dfrac{D_i-\omega_i}{\omega_i}t_i^T - \hat{\tau}_1^T C_{2i} & -1 \end{pmatrix} - H,$$

where

$$C_{1i} = -\left[\frac{D_i(1-\omega_i)}{\omega_i}\hat{q}_i r_i^T,\ \ \omega_i(1-\omega_i)r_i r_i^T\right];\quad C_{2i} = \frac{D_i-\omega_i}{\omega_i}t_i^T.$$

Let p be the dimension of η, then for $H_l, l = 1,\ldots,p$, we have

$$H_l = E\begin{pmatrix} -\omega_i(1-\omega_i)(1-2\omega_i)r_{il}r_i r_i^T & 0 & 0 \\ 0 & 0 & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i}(Y_i-m_i)r_{il}r_i^T - \hat{\tau}^T C_{1i}^* & \dfrac{D_i(1-\omega_i)}{\omega_i}r_{il}t_i^T - \hat{\tau}_1^T C_{2i}^* & 0 \end{pmatrix},$$

where

$$C_{1i}^* = \left[\frac{D_i(1-\omega_i)}{\omega_i}\hat{q}_i r_{il}r_i^T,\ \ -\omega_i(1-\omega_i)(1-2\omega_i)r_{il}r_i r_i^T\right];$$

$$C_{2i}^* = -\frac{D_i(1-\omega_i)}{\omega_i}r_{il}t_i^T.$$

Let K be the dimension of β, then for $H_l, l = p+1, \ldots, p+K$, we have

$$H_l = E\begin{pmatrix} 0 & 0\ 0 \\ 0 & 0\ 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i} t_{i(l-p)} r_i^T - \hat{\tau}_1^T C_{1i}^{**} & 0\ 0 \end{pmatrix},$$

where

$$C_{1i}^{**} = -\frac{D_i(1-\omega_i)}{\omega_i} t_{i(l-p)} r_i^T.$$

Finally,

$$H_{p+K+1} = 0.$$

Let t_i^* be the union of the nonoverlapping elements of (r_i, t_i), then IPT over-parameterizes ω_i by fitting a logistic function $\omega_i = \exp(\zeta^T t_i^*)/(1+\exp(\zeta^T t_i^*))$ so the parameters in this model are $\theta = (\zeta, \mu)$. Estimates of the parameters are solutions to (3.2) with

$$h_i = \begin{pmatrix} \left(\dfrac{D_i}{\omega_i} - 1\right)\omega_{\zeta i} \\ \dfrac{D_i}{\omega_i}(Y_i - \mu) \end{pmatrix} = \begin{pmatrix} \left(\dfrac{D_i}{\omega_i} - 1\right)\omega_i(1-\omega_i)t_i^* \\ \dfrac{D_i}{\omega_i}(Y_i - \mu) \end{pmatrix},$$

$$Q_i = \begin{pmatrix} -\dfrac{D_i(1-\omega_i)}{\omega_i} t_i^* t_i^{*T} & 0 \\ -\dfrac{D_i}{\omega_i^2}(Y-\mu)t_i^* & -\dfrac{D_i}{\omega_i} \end{pmatrix} - H.$$

Let p be the dimension of ζ, then for $H_l, l = 1, \ldots, p$, we have

$$H_l = E\begin{pmatrix} \dfrac{D_i(1-\omega_i)}{\omega_i} t_{il}^* t_i^* t_i^{*T} & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i}(Y-\mu)t_{il}^* t_i^* & \dfrac{D_i(1-\omega_i)}{\omega_i} t_{il}^* \end{pmatrix}.$$

$$H_{p+1} = E\begin{pmatrix} 0 & 0 \\ 0 & \dfrac{D_i(1-\omega_i)}{\omega_i} \end{pmatrix}.$$

For CTD, $\theta = (\eta, \beta, c, \mu)$, and

$$h_i = \begin{pmatrix} \dfrac{D_i - \omega_i}{\omega_i(1-\omega_i)}\omega_{\eta i} \\ \dfrac{D_i(1-\omega_i)}{\omega_i^2} m_{\beta i}\left(Y_i - m_i - c^T \dfrac{\omega_{\eta i}}{1-\omega_i}\right) \\ \dfrac{D_i(1-\omega_i)}{\omega_i^2}\dfrac{\omega_{\eta i}}{1-\omega_i}\left(Y_i - m_i - c^T\dfrac{\omega_{\eta i}}{1-\omega_i}\right) \\ \dfrac{D_i}{\omega_i}(Y_i - m_i) + m_i - c^T\dfrac{D_i-\omega_i}{\omega_i(1-\omega_i)}\omega_{\eta i} - \mu \end{pmatrix}$$

$$= \begin{pmatrix} (D_i - \omega_i) r_i \\ \dfrac{D_i(1-\omega_i)}{\omega_i^2}(Y_i - m_i - \omega_i c^T r_i) t_i \\ \dfrac{D_i(1-\omega_i)}{\omega_i}(Y_i - m_i - \omega_i c^T r_i) r_i \\ \dfrac{D_i}{\omega_i}(Y_i - m_i) + m_i - (D_i - \omega_i) c^T r_i - \mu \end{pmatrix}.$$

$$Q_i = \begin{pmatrix} -\omega_i(1-\omega_i) r_i r_i^T & 0 & 0 & 0 \\ t_i C_{2i} & -\dfrac{D_i(1-\omega_i)}{\omega_i^2} t_i t_i^T & -\dfrac{D_i(1-\omega_i)}{\omega_i} t_i r_i^T & 0 \\ C_{3i} & -\dfrac{D_i(1-\omega_i)}{\omega_i} r_i t_i^T & -D_i(1-\omega_i) r_i r_i^T & 0 \\ C_{4i} & -\dfrac{D_i - \omega_i}{\omega_i} t_i^T & -(D_i - \omega_i) r_i^T & -1 \end{pmatrix} - H,$$

where $C_{2i} = C_{21i} - c^T C_{22i}$ and

$$C_{21i} = -\frac{D_i(2-\omega_i)(1-\omega_i)}{\omega_i^2}(Y_i - m_i - \omega_i c^T r_i) r_i^T,$$

$$C_{22i} = \frac{D_i(1-\omega_i)^2}{\omega_i} r_i r_i^T,$$

$$C_{3i} = \omega_i r_i C_{2i} + (1-\omega_i) r_i h_{3i}^T,$$

$$C_{4i} = -\frac{D_i(1-\omega_i)}{\omega_i}(Y_i - m_i) r_i^T + \omega_i(1-\omega_i) c^T r_i r_i^T.$$

Let p be the dimension of η, then for $H_l, l = 1, \ldots, p$, we have

$$H_l = E\begin{pmatrix} -\omega_i(1-\omega_i) \times (1-2\omega_i) r_{il} r_i r_i^T & 0 & 0 & 0 \\ t_i C'_{2i} & \dfrac{D_i(1-\omega_i)(2-\omega_i)}{\omega_i^2} r_{il} t_i t_i^T & \dfrac{D_i(1-\omega_i)}{\omega_i} r_{il} t_i r_i^T & 0 \\ C'_{3i} & \dfrac{D_i(1-\omega_i)}{\omega_i} r_{il} r_i t_i^T & D_i\omega_i(1-\omega_i) r_{il} r_i r_i^T & 0 \\ C'_{4i} & \dfrac{D_i(1-\omega_i)}{\omega_i} r_{il} t_i^T & \omega_i(1-\omega_i) r_{il} r_i^T & 0 \end{pmatrix},$$

where $C'_{2i} = C'_{21i} - c^T C'_{22i}$ and

$$C'_{21i} = D_i \frac{1-\omega_i}{\omega_i^2} r_{il}\big[(4-3\omega_i)(Y_i - m_i - \omega_i c^T r_i) + (2-\omega_i)\omega_i(1-\omega_i) c^T r_i\big] r_i^T,$$

$$C'_{22i} = -\frac{D_i(1-\omega_i)^2(1+\omega_i)}{\omega_i} r_{il} r_i r_i^T,$$

$$C'_{3i} = 2\omega_i(1-\omega_i) r_{il} r_i C_{2i} + \omega_i r_i C'_{2i} + (1-\omega_i)(1-2\omega_i) r_{il} r_i h_{3i}^T,$$

$$C'_{4i} = \left[\frac{D_i}{\omega_i^2}(Y_i - m_i) + (1-2\omega_i) c^T r_i\right]\omega_i(1-\omega_i) r_{il} r_i^T.$$

Let p be the dimension of c, then for $H_l, l = p+1, \ldots, 2p$, we have

$$H_l = E\begin{pmatrix} 0 & 0 & 0 & 0 \\ \dfrac{D_i(2-\omega_i)(1-\omega_i)}{\omega_i^2} t_{i(l-p)} t_i r_i^T & 0 & 0 & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i} t_{i(l-p)} r_i r_i^T & 0 & 0 & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i} t_{i(l-p)} r_i^T & 0 & 0 & 0 \end{pmatrix}.$$

Let K be the dimension of β, then for $H_l, l = 2p+1, \ldots, 2p+K$, we have

$$H_l = E\begin{pmatrix} 0 & 0 & 0 & 0 \\ \dfrac{D_i(1-\omega_i)}{\omega_i} r_{i(l-2p)} t_i r_i^T & 0 & 0 & 0 \\ D_i\omega_i(1-\omega_i) r_{i(l-2p)} r_i r_i^T & 0 & 0 & 0 \\ \omega_i(1-\omega_i) r_{i(l-2p)} r_i^T & 0 & 0 & 0 \end{pmatrix}.$$

Finally,

$$H_{2p+K+1} = 0.$$

SUPPLEMENTARY MATERIALS

The supplementary materials give additional details and results of the simulation study described in Section 3.

ACKNOWLEDGMENTS

We thank the referees for their perceptive comments and suggestions, that have led to a greatly improved version of this article. Denis Leung's research is

partially supported by the Research Center at Singapore Management University.

[Received October 2012. Revised May 2015.]

REFERENCES

Bang, H., and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972. [86]

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press. [86]

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving Efficiency and Robustness of Doubly Robust Estimator for a Population Mean with Incomplete Data," *Biometrika*, 96, 723–734. [87,88,89]

Cattaneo, M. D. (2010), "Efficient Semiparametric Estimation of Multivalued Treatment Effects Under Ignorability," *Journal of Econometrics*, 155, 138–154. [86]

Chen, X., Hong, H., and Tarozzi, A. (2008), "Semiparametric Efficiency in GMM Models With Auxiliary Data," *Annals of Statistics*, 36, 808–843. [89]

Cheng, P. E. (1994), "Nonparametric Estimation of Mean Functionals With Data Missing at Random," *Journal of the American Statistical Association*, 89, 81–87. [86]

Godambe, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation," *Annals of Mathematical Statistics*, 31, 1208–11. [88,94]

Graham, B. S., Pinto, C., and Egel, D. (2012), "Inverse Probability Tilting for Moment Condition Models With Missing Data," *Review of Economic Studies*, 79, 1052–1079. [87,89,93]

Griliches, Z., Hall, B. H., and Hausman, J. A. (1978), "Missing Data and Self-Selection in Large Panels," *Annales De L'Insee*, 30–31, 127–176. [94]

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331. [86,93]

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. [86]

Imbens, G. W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling," *Econometrica*, 60, 1187–1214. [86]

Kang, J. D. Y., and Schafer, J. L. (2007), "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean From Incomplete Data" (with discussion), *Statistical Science*, 22, 523–39. [86,87,90,91,95]

Khan, S., and Tamer, E. (2010), "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042. [95]

Little, R. J. A., and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data With Missing Values," *Statistica Sinica*, 14, 949–968. [86]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, New York: Wiley. [86]

Mincer, J. (1974), *Schooling, Experience, and Earnings*, New York: National Bureau of Economic Research. [93]

Newey, W. K., and Smith, R. J. (2004), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [93]

Pepe, M. S., Reilly, M., and Fleming, T. R. (1994), "Auxiliary Outcome Data and the Mean-Score Method," *Journal of Statistical Planning and Inference*, 42, 137–160. [86]

Roberts, G., Rao, J., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, 74, 1–12. [86]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [86]

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. [86,87,88,89,91]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121. [86]

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Performance of Double Robust Estimators When Inverse Probability Weights are Highly Variable," *Statistical Science*, 22, 544–559. [87,95]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [86]

Rothe, C., and Firpo, S. (2013), Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions," IZA Discussion Paper 7564, Institute for the Study of Labor, Bonn, Germany. [95]

Rubin, D. B. (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Education Statistics*, 2, 1–26. [89]

Rubin, D. B., and van der Laan, M. J. (2008), "Empirical Efficiency Maximization: Improved Locally Efficient Covariate Adjustment in Randomized Experiments and Survival Analysis," *International Journal of Biostatistics*, 4. [87]

Särndal, C. E., Swensson, B., and Wretman, J. (1989), "The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of a Finite Population Total," *Biometrika*, 76, 527–537. [86]

Scharfstein, D. O., Robins, J. M., and Rotnitzky, A. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models With Missing Data" (with comments), *Journal of the American Statistical Association*, 94, 1096–1146. [86]

Słoczyński, T., and Wooldridge, J. M. (2014), A General Double Robustness Result for Estimating Average Treatment Effects," IZA Discussion Paper 8084, Institute for the Study of Labor, Bonn, Germany. [86]

Tan, Z. (2006), "A Distributional Approach for Causal Inference Using Propensity Scores," *Journal of American Statistical Association*, 101, 1619–1637. [87]

——— (2008), "Comment: Improved Local Efficiency and Double Robustness," *International Journal of Biostatistics*, 4. [87]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [87]

Uysal, S. D. (2015), "Doubly Robust Estimation of Causal Effects With Multivalued Treatments: An Application to the Returns to Schooling," *Journal of Applied Econometrics*, 30, 763–786. [86]

Wooldridge, J. (2007), "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301. [86]