

# CBPS-Based Inference in Nonlinear Regression Models with Missing Data

Donglin Guo<sup>1,2\*</sup>, Liugen Xue<sup>1</sup>, Haiqing Chen<sup>1</sup>

<sup>1</sup>College of Applied Sciences, Beijing University of Technology, Beijing, China

<sup>2</sup>School of Mathematics and Information Science, Shangqiu Normal University, Shangqiu, China

Email: \*gd11105@emails.bjut.edu.cn

Received 20 June 2016; accepted 22 August 2016; published 25 August 2016

Copyright © 2016 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

In this article, to improve the doubly robust estimator, the nonlinear regression models with missing responses are studied. Based on the covariate balancing propensity score (CBPS), estimators for the regression coefficients and the population mean are obtained. It is proved that the proposed estimators are asymptotically normal. In simulation studies, the proposed estimators show improved performance relative to usual augmented inverse probability weighted estimators.

## Keywords

Nonlinear Regression Model, Missing at Random, Covariate Balancing Propensity Score, GMM, Augmented Inverse Probability Weighted

---

## 1. Introduction

Consider the nonlinear regression model:

$$Y_i = f(X_i, \beta) + \varepsilon_i, \quad (1)$$

where  $Y_i$  is a scalar response variate,  $X_i$  is a  $d \times 1$  vector of covariate,  $\beta$  is a  $p \times 1$  vector of unknown regression parameter,  $f(\cdot)$  is a known function, and it is nonlinear with respect to  $\beta$ ,  $\varepsilon_i$  is a random statistical error with  $E[\varepsilon_i | X_i] = 0$ . In general,  $d$  is different from  $p$ . The model has been studied by many authors, such as Jennrich [1], Wu [2], Crainceanu and Ruppert [3] and so on.

Missing data is frequently encountered in statistical studies, and ignoring it could lead to biased estimation and misleading conclusions. Inverse probability weighting (Horvitz and Thompson [4]) and imputation are two main methods for dealing with missing data. Since Scharfstein *et al.* [5] noted that the augmented inverse

---

\*Corresponding author.

probability weighted (AIPW) estimator in Robins *et al.* [6] was double-robust, authors have proposed many estimators with the double-robust property, see Tan [7], Kang and Schafer [8], Cao *et al.* [9]. The estimator is doubly robust in the sense that consistent estimation can be obtained if either the outcome regression model or the propensity score model is correctly specified. The AIPW estimators have been advocated for routine use (Bang and Robins [10]). For model (1), in the absence of missing data, the weighted least squares estimator of  $\beta$  can be obtained by minimizing the objective function  $\sum_{i=1}^n (Y_i - f(X_i, \beta))^2$ . In the presence of missing data, the above-mentioned method can not be used directly, so we make use of AIPW method to consider the model (1).

Throughout this paper, we assume that  $X$ 's are observed completely,  $Y$  is missing at random (Rubin [11]). Thus, the data actually observed are independent and identically distributed  $(\delta_i Y_i, \delta_i, X_i) (i = 1, \dots, n)$ , where  $\delta_i = 1$  indicates that  $Y_i$  is observed and  $\delta_i = 0$  indicates that  $Y_i$  is missing. The missing at random (MAR) assumption implies that  $\delta$  and  $Y$  are conditionally independent given  $X$ , that is,  $P(\delta = 1 | X, Y) = P(\delta = 1 | X) \equiv \pi(X)$ . This probability is called the propensity score (Rosenbaum and Rubin [12]).

If  $f(X_i, \beta) = X_i^T \beta$ , model (1) is just the classical linear model. The linear models with missing data have been studied in existing papers, such as Wang and Rao ([13] [14]), Xue [15], Qin and Lei [16] and so on. The inverse probability weighted imputation methods of Xue [15] and other papers are based on the nonparametric estimators of the propensity score model. However, it is difficult to obtain the nonparametric estimators because of the ‘‘curse of dimensionality’’, and as mentioned in the Kang and Schafer [8], the AIPW estimators can be severely biased when both models are misspecified. In addition, there is little work done for model (1) with missing responses.

In this paper, we construct estimators for  $\beta$  and  $\mu$  of model (1), based on the covariate balancing propensity score (CBPS) method proposed by Imai and Ratkovic [17]. As mentioned in Imai and Ratkovic [18], the weights based on CBPS are robust in the sense that they improve covariate balance even when propensity score model is misspecified. Our estimator has the following merits: 1) it avoids the ‘‘curse of dimensionality’’; 2) it avoids selecting the optimal bandwidth; 3) it improves performance of the AIPW estimators in terms of bias, standard deviation (SD) and mean-squared error (MSE), even when both outcome regression model and propensity score model are misspecified.

The rest of this paper is organized as follows. In Section 2, based on the CBPS and the AIPW methods, the estimators for the regression parameter  $\beta$  and the population mean  $\mu$  are proposed, and the asymptotic properties of the estimators are investigated. In Section 3, simulation studies are carried out to assess the performance of the proposed method. In Section 4, concluding remarks are made. In Appendix, the proofs of the main results are given.

## 2. Construction of Estimators

The most popular choice of  $\pi(X)$  is a logistic regression function (Qin and Zhang [19]). We make the same choice and posit a logistic regression model for  $\pi(X)$

$$\pi(X) = \frac{\exp(\alpha^T X)}{1 + \exp(\alpha^T X)}, \tag{2}$$

where  $\alpha \in \Theta$  is  $d$ -dimensional unknown column vector parameter.

### 2.1. CBPS-Based Estimator for the Propensity Score

Based on  $(\delta_i, X_i) (i = 1, \dots, n)$ , people can obtain the estimator  $\hat{\alpha}$  by maximizing the log-likelihood function:

$$L = \sum_{i=1}^n [\delta_i \log \{\pi(X_i, \alpha)\} + (1 - \delta_i) \log \{1 - \pi(X_i, \alpha)\}]. \tag{3}$$

Assuming that  $\pi(X, \alpha)$  is twice continuously differentiable with respect to  $\alpha$ , so maximizing the (3) implies the first-order condition

$$\frac{1}{n} \sum_{i=1}^n s(\delta_i, X_i, \alpha) = 0, \quad s(\delta_i, X_i, \alpha) = \frac{(\delta_i - \pi(X_i, \alpha)) \pi'(X_i, \alpha)}{\pi(X_i, \alpha)(1 - \pi(X_i, \alpha))}, \tag{4}$$

where  $\pi'(X_i, \alpha) = \partial \pi(X_i, \alpha) / \partial \alpha^T$ . However, the main drawback of this standard method is that the propensity score model  $\pi(X)$  may be misspecified, yielding biased estimators for the interesting parameters, such as  $\beta$  and  $\mu$ . To overcome the drawback, we borrow the following ideas of Imai and Ratkovic [17]. Similar to arguments present by Imai and Ratkovic [17], we operationalize the covariate balancing property by using inverse propensity score weighting

$$E \left\{ \frac{\delta_i X_i}{\pi(X_i, \alpha)} - \frac{(1 - \delta_i) X_i}{1 - \pi(X_i, \alpha)} \right\} = 0. \quad (5)$$

Equation (5) ensures that the first moment of each covariate is balanced and the weights based on CBPS are robust even when propensity score model is misspecified. The key idea behind the CBPS is that propensity score model determines the missing mechanism and covariate balancing weights, see Imai and Ratkovic [17]. The sample analogue of the covariate balancing moment condition given in Equation (5) is

$$\frac{1}{n} \sum_{i=1}^n w(\delta_i, X_i, \alpha) X_i = 0, \quad w(\delta_i, X_i, \alpha) = \frac{\delta_i - \pi(X_i, \alpha)}{\pi(X_i, \alpha)(1 - \pi(X_i, \alpha))}. \quad (6)$$

According to Imai and Ratkovic [17], the CBPS is said to be *just identified* when the number of moment conditions equals that of parameters. If we use the covariate balancing conditions given in Equation (6) alone, the CBPS is just-identified. If we combine Equation (6) with the score condition given in Equation (4), then the CBPS is *overidentified* because the number of moment conditions exceeds that of parameters.

Combining Equation (6) with the score condition given in Equation (4), we obtain the following equation:

$$\bar{U}(\alpha) = \frac{1}{n} \sum_{i=1}^n U(\delta_i, X_i, \alpha) = 0, \quad U(\delta_i, X_i, \alpha) = \begin{pmatrix} s(\delta_i, X_i, \alpha) \\ w(\delta_i, X_i, \alpha) X_i \end{pmatrix}, \quad (7)$$

Let  $\hat{\alpha}$  be the solution to the Equation (7). For the overidentified CBPS, the GMM (Hansen [20]) estimator  $\hat{\alpha}$  can be obtained by minimizing the following equation with respect to  $\alpha$  for some positive-semidefinite symmetric weight matrix  $W$ :

$$Q(\delta_i, X_i, \alpha) = \bar{U}^T(\alpha) W \bar{U}(\alpha). \quad (8)$$

It is easy to show that, under some regularity conditions,  $\hat{\alpha}$  is a consistent estimator of  $\alpha_0$ , the true value of  $\alpha$ . For the just-identified CBPS, we borrow the ideas of Imai and Ratkovic [17] and still minimize Equation (8) without the score condition to find  $\hat{\alpha}$ .

**Theorem 1.** Suppose that  $X_i, i=1, \dots, n$  be a set of independent and identically distributed random vectors, under the Assumptions (A1)-(A3) in the Appendix, then  $\hat{\alpha} \xrightarrow{P} \alpha_0$  and  $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$ , where  $\hat{\alpha}$  minimizes Equation (8).

## 2.2. Estimator for the Regression Parameter

To make use of AIPW method, we borrow the idea of Seber and Wild [21] and define the least squares estimator of  $\beta$  based on complete-case data by solving the following estimating equation:

$$\sum_{i=1}^n \delta_i F_i(\beta) (Y_i - f(X_i, \beta)) = 0, \quad (9)$$

where  $F_i(\beta) = \partial f(X_i, \beta) / \partial \beta$ . There is no closed form of  $\hat{\beta}$ , but it can be obtained by the following iterative equation:

$$\beta^{(t)} = \beta^{(t-1)} + (F^T W F)^{-1} F^T W \varepsilon, \quad t = 0, 1, \dots \quad (10)$$

where  $F^T = (F_1, \dots, F_n)$ ,  $W = \text{diag}(\delta_1, \dots, \delta_n)$ ,  $\varepsilon = (Y_1 - f(X_1, \beta), \dots, Y_n - f(X_n, \beta))^T$ , and  $F, \varepsilon$  are evaluated at  $\beta^{(t-1)}$ . If  $\|\beta^{(t)} - \beta^{(t-1)}\| \leq c$ , where  $c$  is a prespecified tolerance and  $\|\cdot\|$  denotes the  $L_2$  norm, then we stop the above iterative algorithm and obtain the least squares estimator of  $\beta$ , denoted by  $\hat{\beta}_c$ .

Although the implementation of the complete case method is simple, it may result in misleading conclusion by simply excluding the missing data. In this section, we introduce an AIPW method based on CBPS to deal

with the problems of complete case method.

Denote  $\check{Y}_i = \frac{\delta_i}{\pi(X_i, \alpha)} Y_i + \left(1 - \frac{\delta_i}{\pi(X_i, \alpha)}\right) f(X_i, \hat{\beta}_c)$ , From Equation (1), we have  $E(\check{Y}_i | X_i) = f(X_i, \beta)$

under the MAR condition. Hence

$$\check{Y}_i = f(X_i, \beta) + \eta_i, \tag{11}$$

where  $\eta_i$ 's satisfy  $E(\eta_i | X_i) = 0$ . Formula (11) is a full data model without missing data. So similar to Equation (10), we can obtain an estimator  $\hat{\beta}_t$  of  $\beta$  by iterative equation

$$\beta^{(t)} = \beta^{(t-1)} + (F^T F)^{-1} F^T e, t = 0, 1, \dots, \text{ where } e = (\hat{Y}_1 - f(X_1, \beta), \dots, \hat{Y}_n - f(X_n, \beta))^T \text{ with}$$

$$\hat{Y}_i = \frac{\delta_i}{\pi(X_i, \hat{\alpha})} Y_i + \left(1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})}\right) f(X_i, \hat{\beta}_c) \text{ and } \hat{\alpha} \text{ is obtained by CBPS method.}$$

The following Theorem 2 gives the asymptotic normality of  $\hat{\beta}_t$ .

**Theorem 2.** Suppose that Assumptions (A1)-(A4) in the Appendix hold. Then we have

$$\sqrt{n}(\hat{\beta}_t - \beta) \xrightarrow{D} N(0, B),$$

where  $B = \Sigma^{-1} \Omega \Sigma^{-1}$ ,  $\Sigma = E[F_i F_i^T]$  and  $\Omega = E[\pi(X)^{-1} F_i F_i^T \varepsilon^2]$  with  $\varepsilon = Y - f(X, \beta)$ .

To apply Theorem 2 to construct the confidence region of  $\beta$ , we use  $\hat{B} = \hat{\Sigma}^{-1} \hat{\Omega} \hat{\Sigma}^{-1}$  to consistently estimate  $B$ . where  $\hat{\Sigma}$  and  $\hat{\Omega}$  are defined by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n F_i(\hat{\beta}_t) F_i^T(\hat{\beta}_t) \text{ and } \hat{\Omega} = \frac{1}{n} \sum_{i=1}^n \pi(X_i, \hat{\alpha})^{-1} F_i(\hat{\beta}_t) F_i^T(\hat{\beta}_t) (\check{Y}_i - X_i^T \hat{\beta}_t)^2.$$

Therefore, we have

$$\hat{B}^{-1/2} \sqrt{n}(\hat{\beta}_t - \beta) \xrightarrow{D} N(0, I_p) \tag{12}$$

and

$$(\hat{\beta}_t - \beta)^T n \hat{B}^{-1} (\hat{\beta}_t - \beta) \xrightarrow{D} \chi_p^2. \tag{13}$$

We can construct the confidence interval of  $\beta$  using (12) and (13).

### 2.3. Estimator for the Response Mean

It is of interest to estimate the mean of  $Y$ , say  $\mu$ , when there are missing data in the responses. We here make use of the method of Xue [15] to construct the estimators of  $\mu$ . Let

$$Y_i^* = \frac{\delta_i Y_i}{\pi(X_i, \alpha)} + \left(1 - \frac{\delta_i}{\pi(X_i, \alpha)}\right) f(X_i, \beta).$$

Under the MAR condition, we have  $E(Y_i^*) = \mu$  if  $\mu$  is the true parameter. Then the proposed estimator is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \left[ \frac{\delta_i Y_i}{\pi(X_i, \hat{\alpha})} + \left(1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})}\right) f(X_i, \hat{\beta}_c) \right]. \tag{14}$$

In the following theorem, we state the asymptotic properties of  $\hat{\mu}$ .

**Theorem 3.** Under the assumptions (A1)-(A4) in the Appendix, we have

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} N(0, V),$$

where  $V = E\left\{ \frac{\sigma^2(X)}{\pi(X)} \right\} + \text{var}\{f(X, \beta)\}$  with  $\sigma^2(X) = \text{var}[\varepsilon | X]$ .

Borrowing the method of Xue [15], we can obtain the following consistent estimator of  $V$ :

$$\hat{V} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - \hat{\mu})^2,$$

where

$$\hat{Y}_i = \frac{\delta_i Y_i}{\pi(X_i, \hat{\alpha})} + \left(1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})}\right) f(X_i, \hat{\beta}_c).$$

By Theorem 3, the normal approximation based confidence interval of  $\mu$  with confidence level  $1 - \alpha$  is  $\hat{\mu} \pm z_{1-\alpha/2} \sqrt{\hat{V}/n}$ .

### 3. Simulation Examples

We conducted simulation studies to examine the performance of the proposed estimation methods. The simulated data are generated from the model  $Y = \exp(X^T \beta) + \varepsilon$  with  $\beta = (1, 2, 1)^T$ . The components of  $X = (X_1, X_2, X_3)^T$  are generated from the uniform distribution  $U(0, 1)$  respectively and  $\varepsilon$  is generated from the standard normal distribution,  $\delta$  is generated from Bernoulli with true propensity score model  $\pi(X) = 1 / \{1 + \exp(-0.1X_1 - 0.3X_2 + 0.25X_3)\}$ .

When both models are misspecified or either of them is misspecified, we adopt the same way as Kang and Schafer [8] to examine whether our method can improve the empirical performance of doubly robust estimators or not. Similar to Kang and Schafer [8], only the  $X^* = \{\exp(X_1/2), X_2 / \{1 + \exp(X_1)\} + 10, (X_1 X_3 / 25 + 0.6)^3\}$  are observed. If  $Y$  is expressed as  $\exp(X^{*T} \beta) + \varepsilon$  or propensity score model is expressed as  $\pi(X^*)$ , the model is misspecified. As in the original study, we conduct simulations for population mean  $\mu$  under four scenarios:

- 1) both outcome and propensity score models are correctly specified;
- 2) only the propensity score model is correct;
- 3) only the outcome model is correct;
- 4) both outcome and propensity score models are correctly misspecified.

Due to the regression parameter  $\beta$  is in the outcome regression model, we only conduct simulations for  $\beta$  under (1) and (3) scenarios. For each scenario, we conduct 1000 simulations and calculate the bias, standard deviation (SD) and mean-squared error (MSE) for  $\beta$  and  $\mu$ . The results of our simulations are presented in Tables 1-3. For a given scenario, we examine the performance of estimators on the basis of four different propensity score methods:

**Table 1.** Relative performance of the estimators for regression parameter based on different propensity score estimation methods when both models are correct.

$\beta$	$n$	GLM			CBPS1			CBPS2			TRUE		
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
$\beta_1$	80	0.0010	0.0500	0.0025	-0.0002	0.0469	0.0022	0.0000	0.0481	0.0023	0.0001	0.0480	0.0023
	100	0.0011	0.0450	0.0020	0.0004	0.0439	0.0019	0.0003	0.0434	0.0019	0.0013	0.0428	0.0018
	120	0.0009	0.0405	0.0016	0.0013	0.0382	0.0015	-0.0017	0.0392	0.0015	-0.0005	0.0374	0.0014
$\beta_2$	80	-0.0015	0.0509	0.0026	-0.0008	0.0476	0.0023	0.0001	0.0498	0.0025	0.0001	0.0495	0.0024
	100	0.0003	0.0449	0.0020	-0.0007	0.0420	0.0018	-0.0007	0.0430	0.0019	-0.0004	0.0422	0.0018
	120	0.0001	0.0381	0.0014	-0.0017	0.0376	0.0014	0.0008	0.0391	0.0015	0.0022	0.0376	0.0014
$\beta_3$	80	-0.0007	0.0500	0.0025	0.0010	0.0485	0.0024	0.0001	0.0505	0.0025	0.0000	0.0502	0.0025
	100	-0.0007	0.0425	0.0018	0.0005	0.0426	0.0018	-0.0003	0.0430	0.0018	-0.0010	0.0418	0.0017
	120	-0.0022	0.0368	0.0014	0.0009	0.0364	0.0013	0.0011	0.0387	0.0015	-0.0016	0.0392	0.0015

**Table 2.** Relative performance of the estimators for regression parameter based on different propensity score estimation methods when only outcome model is correct.

$\beta$	n	GLM			CBPS1			CBPS2			TRUE		
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
$\beta_1$	80	0.0007	0.0502	0.0025	-0.0012	0.0490	0.0024	-0.0006	0.0492	0.0024	-0.0011	0.0485	0.0023
	100	0.0029	0.0422	0.0018	0.0003	0.0435	0.0018	-0.0016	0.0429	0.0018	-0.0013	0.0430	0.0019
	120	-0.0019	0.0398	0.0016	0.0005	0.0390	0.0015	0.0000	0.0383	0.0015	0.0004	0.0389	0.0015
$\beta_2$	80	-0.0002	0.0504	0.0025	0.0004	0.0504	0.0025	0.0002	0.0494	0.0024	0.0004	0.0498	0.0025
	100	-0.0025	0.0442	0.0020	-0.0007	0.0430	0.0018	-0.0018	0.0428	0.0018	-0.0019	0.0433	0.0019
	120	0.0002	0.0395	0.0016	0.0008	0.0389	0.0015	-0.0007	0.0382	0.0015	-0.0012	0.0405	0.0016
$\beta_3$	80	-0.0017	0.0477	0.0023	0.0006	0.0482	0.0023	-0.0001	0.0502	0.0025	0.0004	0.0476	0.0023
	100	-0.0014	0.0437	0.0019	-0.0004	0.0429	0.0018	0.0023	0.0421	0.0018	0.0023	0.0427	0.0018
	120	0.0008	0.0392	0.0015	-0.0012	0.0410	0.0016	0.0004	0.0374	0.0014	0.0005	0.0407	0.0017

**Table 3.** Relative performance of the doubly robust estimators based on different propensity score estimation methods for mean under the four different scenarios.

Scen	n	GLM			CBPS1			CBPS2			TRUE		
		Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
(1)	200	0.0040	0.5336	0.2845	0.0104	0.4931	0.2430	0.0134	0.5014	0.2513	0.0132	0.5012	0.2511
	500	0.0081	0.3197	0.1022	-0.0059	0.3162	0.0999	0.0060	0.3119	0.0972	-0.0133	0.3181	0.1013
(2)	200	0.0016	0.5623	0.3158	0.0423	0.5595	0.3145	0.0500	0.5361	0.2896	-0.0169	0.6446	0.4154
	500	0.0100	0.3497	0.1227	0.0390	0.3433	0.1193	0.0258	0.3438	0.1188	0.0066	0.4190	0.1754
(3)	200	0.0012	0.5167	0.2667	0.0094	0.4886	0.2386	0.0103	0.4931	0.2430	-0.0166	0.5110	0.2612
	500	-0.0118	0.3118	0.0973	0.0106	0.3106	0.0965	-0.0149	0.3074	0.0946	0.0122	0.3158	0.0998
(4)	200	-0.0084	0.5694	0.3239	0.0584	0.5597	0.3163	0.0385	0.5568	0.3112	0.0182	0.6418	0.4119
	500	-0.0104	0.3553	0.1262	0.0352	0.3472	0.1217	0.0553	0.3483	0.1243	-0.0130	0.4003	0.1602

Remark: 1) Both models are correct; 2) Only propensity score model is correct; 3) Only outcome model is correct; 4) Both models are incorrect.

- a) usual GLM method;
- b) the just-identified CBPS estimation with the covariate balancing moment conditions and without the score condition (CBPS1);
- c) the overidentified CBPS estimation with both the covariate balancing and score conditions (CBPS2);
- d) The true propensity score model which we do not need to estimate (TRUE).

From **Table 1** and **Table 2**, we can see that SD and MSE of our estimators for  $\beta$  decrease as  $n$  increases. Whether the propensity score model is specified correctly or not, the proposed estimators based on CBPS have smaller SD and MSE than the usual GLM estimators mostly. The CBPS with or without the score condition can substantially improve the performance of usual estimator. Compared with estimators based on true propensity score model, our proposed estimators perform as well as them in the terms of SD and MSE. **Table 3** shows that, under the four scenarios, the SD and MSE of our proposed estimators remain lower than the usual GLM estimators. Similar to Imai and Ratkovic [17], the final scenario illustrates the most important point made by Kang and Schafer [8] that doubly robust estimator can deteriorate when both the outcome and the propensity models are misspecified. Under this scenario, the doubly robust estimators based on usual GLM have a significant amount of bias and variance. However, the CBPS can improve the performance of doubly robust estimators. In a word, we obtain the same conclusion as Imai and Ratkovic [17] that the CBPS can yield robust estimators of population mean, even when both the outcome and propensity score models are misspecified.

## 4. Concluding Remarks

We have proposed an improved estimation method for the parameters of interest in the nonlinear regression model with missing responses. The estimators based on CBPS and AIPW method have the following merits: 1) They avoid the “curse of dimensionality” and avoid selecting the optimal bandwidth; 2) When either the outcome regression model or the propensity score model is correctly specified, the proposed estimators perform as well as estimators based on true propensity model in the terms of SD and MSE; 3) When both outcome regression and propensity score models are misspecified, as mentioned in Section 1, the usual AIPW estimator can be severely biased, but our method improves the performance of them and obtains an improved estimator for population mean. The simulation shows that the proposed method is feasible. Furthermore, with appropriately modification, the proposed method can be extended to other models with missing responses. The exhaustive procedure will be presented in our future work.

## Acknowledgements

We thank the Editor and the referee for their helpful comments that largely improve the presentation of the paper.

## References

- [1] Jennrich, R.I. (1969) Asymptotic Properties of Nonlinear Least Squares Estimators. *The Annals of Mathematical Statistics*, **40**, 633-643. <http://dx.doi.org/10.1214/aoms/1177697731>
- [2] Wu, C.F. (1981) Asymptotic Theory of Nonlinear Least Squares Estimation. *Ann. Statist. The Annals of Statistics*, **9**, 501-513. <http://dx.doi.org/10.1214/aos/1176345455>
- [3] Crainiceanu, C.M. and Ruppert, D. (2004) Likelihood Ratio Tests for Goodness-of-Fit of a Nonlinear Regression Model. *Journal of Multivariate Analysis*, **100**, 35-52. <http://dx.doi.org/10.1016/j.jmva.2004.04.008>
- [4] 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. <http://dx.doi.org/10.1080/01621459.1952.10483446>
- [5] Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999) Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, **94**, 1096-1120. <http://dx.doi.org/10.1080/01621459.1999.10473862>
- [6] Robins, J.M. and Rotnitzky, A. (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**, 846-866. <http://dx.doi.org/10.1080/01621459.1994.10476818>
- [7] Tan, Z. (2006) A Distributional Approach for Causal Inference Using Propensity Scores. *Journal of the American Statistical Association*, **101**, 1619-1637. <http://dx.doi.org/10.1198/016214506000000023>
- [8] 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. *Statistical Science*, **22**, 523-539. <http://dx.doi.org/10.1214/07-sts227>
- [9] Cao, W., Tsiatis, A. and Davidian, M. (2009) Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, **96**, 723-734. <http://dx.doi.org/10.1093/biomet/asp033>
- [10] Bang, H. and Robins, J.M. (2005) Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, **61**, 692-972. <http://dx.doi.org/10.1111/j.1541-0420.2005.00377.x>
- [11] Rubin, D.B. (1976) Inference and Missing Data. *Biometrika*, **63**, 581-592. <http://dx.doi.org/10.1093/biomet/63.3.581>
- [12] 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. <http://dx.doi.org/10.1093/biomet/70.1.41>
- [13] Wang, Q.H. and Rao, J.N.K. (2001) Empirical Likelihood for Linear Regression Models under Imputation for Missing Responses. *The Canadian Journal of Statistics*, **29**, 597-608. <http://dx.doi.org/10.2307/3316009>
- [14] Wang, Q.H. and Rao, J.N.K. (2002) Empirical Likelihood-Based Inference in Linear Models with Missing Data. *The Scandinavian Journal of Statistics*, **29**, 563-576. <http://dx.doi.org/10.1111/1467-9469.00306>
- [15] Xue, L.G. (2009) Empirical Likelihood for Linear Models with Missing Responses. *Journal of Multivariate Analysis*, **100**, 1353-1366. <http://dx.doi.org/10.1016/j.jmva.2008.12.009>
- [16] Qin, Y. and Lei, Q. (2010) On Empirical Likelihood for Linear Models with Missing Responses. *Journal of Statistical Planning and Inference*, **140**, 3399-3408. <http://dx.doi.org/10.1016/j.jspi.2010.05.001>

- [17] Imai, K. and Ratkovic, M. (2014) Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society, Series B*, **76**, 243-263. <http://dx.doi.org/10.1111/rssb.12027>
- [18] Imai, K. and Ratkovic, M. (2015) Robust Estimation of Inverse Probability Weights for Marginal Structural Models. *Journal of the American Statistical Association*, **110**, 1013-1022. <http://dx.doi.org/10.1080/01621459.2014.956872>
- [19] Qin, J. and Zhang, B. (2007) Empirical-Likelihood-Based Inference in Missing Response Problems and Its Application in Observational Studies. *Journal of the Royal Statistical Society, Series B*, **69**, 101-122. <http://dx.doi.org/10.1111/j.1467-9868.2007.00579.x>
- [20] Hansen, L.P. (1982) Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**, 1029-1054. <http://dx.doi.org/10.2307/1912775>
- [21] Seber, G.F. and Wild, C.J. (2003) *Nonlinear Regression*. Wiley, New York.
- [22] Newey, W. and McFadden, D. (1994) Large Sample Estimation and Hypothesis Testing. In: Engle, R.F. and McFadden, D.L., Eds., *Handbook of Econometrics*, **IV**, North-Holland, Amsterdam, 2111-2245.



## Appendix: Proofs of the Main Results

Throughout, let  $\alpha_0$  be the true value of  $\alpha$ , and  $\|A\| = \left(\sum_{j,k} a_{jk}^2\right)^{1/2}$  be the Euclidean norm for a matrix

$A = [a_{jk}]$ . Firstly we make the following assumptions.

(A1) For all  $X$ 's,  $\pi(X, \alpha) = \Phi(X^T \alpha)$  is a known, differentiable function from  $\mathfrak{R}$  to  $(0,1)$  for all  $\alpha$ 's in a neighborhood of  $\alpha_0$ .

(A2)  $E(X_i)$  and  $E[w_\beta(X_i)]$  exist and the matrix  $E(X_i X_i^T)$  is of full rank.

(A3) 1)  $W$  is positive semi-definite and  $WE[U(X, \alpha)] = 0$  only if  $\alpha = \alpha_0$ . 2)  $\alpha_0 \in \Theta$ , which is compact.

3)  $E\left[\|\pi_1(X, \alpha_0)\|^2\right] < \infty$  and  $E\left[\|X_i\|^2\right] < \infty$ , where  $\pi_1(X, \alpha) = \partial\pi(X, \alpha)/\partial\alpha$ .

(A4)  $\sup_x E(\varepsilon^2 | X = x) < \infty$  and  $E(\|X\|^2) < \infty$ .

To complete the proofs of Theorems 1-3, the following lemma is needed. If there is a function  $Q_0(\theta)$  such that 1)  $Q_0(\theta)$  is uniquely minimized at  $\theta_0$ ; 2)  $\Theta$  is compact; 3)  $Q_0(\theta)$  is continuous; 4)  $\hat{Q}_n(\theta)$  converges uniformly in probability to  $Q_0(\theta)$ , then  $\hat{\theta} \xrightarrow{p} \theta_0$ , where  $\hat{\theta}$  minimizes  $\hat{Q}_n(\theta)$  subject to  $\theta \in \Theta$ .

**Lemma 1.** is the fundamental consistency result for extremum estimators. Its proof can be found in Newey and McFadden [22], and we omit it here.

**Proof of Theorem 1.** Similar to Theorem 2.6 in Newey and McFadden [22], the proof of  $\hat{\alpha} \xrightarrow{p} \alpha_0$  is proceed by verifying the conditions of Lemma 1. Under assumption (A2), (A3) and Lemma 2.3 in Newey and McFadden (1994), we know that conditions 1) and 2) hold in Lemma 1. Let

$a(X, T; \alpha) = U(X, T; \alpha)$ ,  $U_0(\alpha) = E[U(X, T; \alpha)]$ , Under assumption (A1) and by Lemma 2.4 in Newey and McFadden (1994), we have  $\sup_{\alpha \in \Theta} \|\bar{U}(\alpha) - U_0(\alpha)\| \xrightarrow{p} 0$  and  $U_0(\alpha)$  is continuous. Thus, condition 3 in

Lemma 1 holds by  $Q_0(\alpha) = U_0^T(\alpha) W U_0(\alpha)$  continuous. By  $\Theta$  is compact,  $U_0(\alpha)$  is bounded on  $\Theta$ , and by the Cauchy-Schwartz inequalities,  $\sup_{\alpha \in \Theta} |Q_0(\alpha) - Q_0(\alpha)| \xrightarrow{p} 0$ , and condition 4) in Lemma 1 holds. According to Theorem 3.2 in Newey and McFadden [22], we can obtain that  $\hat{\alpha} - \alpha_0 = O_p(n^{-1/2})$ .

**Proof of Theorem 2.** Denote  $\hat{\eta}_{i,l}(\beta) = F_i(\beta)(Y_i - f(X_i, \beta))$ , where  $F_i(\beta) = \partial f(X_i, \beta)/\partial\beta$ . By the definition of  $\hat{\beta}_l$ ,

$$\sqrt{n}(\hat{\beta}_l - \beta) = \left[ \frac{1}{n} \sum_{i=1}^n F_i(\beta) F_i^T(\beta) \right]^{-1} \left[ \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,l}(\beta) \right] + o_p(1) \quad (15)$$

To prove Theorem 2, we will verify the asymptotically normality of  $\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,l}(\beta)$ . By direct calculation, we have

$$\hat{\eta}_{i,l}(\beta) = \hat{\eta}_{i,w}(\beta) + \left( 1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})} \right) F_i(\beta) F_i^T(\beta) (\hat{\beta}_c - \beta), \quad (16)$$

where

$$\hat{\eta}_{i,w}(\beta) = \frac{\delta_i}{\pi(X_i, \hat{\alpha})} F_i(\beta) (Y_i - f(X_i, \beta)). \quad (17)$$

Under MAR assumption, we have  $E\left[\left(1 - \frac{\delta_i}{\pi(X_i, \alpha)}\right) F_i(\beta) F_i^T(\beta)\right] = 0$ . This combines with Theorem 1 yields

$$\frac{1}{n} \sum_{i=1}^n \left( 1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})} \right) F_i(\beta) F_i^T(\beta) = o_p(1). \quad (18)$$

From the Theorem 5 in Wu (1981), we know that  $\hat{\beta}_c - \beta = O_p(n^{-1/2})$ . This together with (16) and (18) proves that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,l}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,w}(\beta) + o_p(1). \quad (19)$$

By Theorem 1,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,W}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{\pi(X_i, \alpha)} F_i(\beta) \varepsilon_i + o_p(1).$$

According to the assumptions given in model (1), we have

$$\text{cov} \left\{ \frac{\delta_i}{\pi(X_i, \alpha)} F_i(\beta) \varepsilon_i \right\} = E \left[ \pi(X)^{-1} F_i(\beta) F_i^T(\beta) \varepsilon^2 \right] = \Omega.$$

Then, it follows from the central limit theorem that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\eta}_{i,W}(\beta) \xrightarrow{D} N(0, \Omega).$$

Therefore, by using (19) and Slutsky theorem, the proof of Theorem 2 is completed.

**Proof of Theorem 3.** By direct calculation, we have we have

$$\sqrt{n}(\hat{\mu} - \mu) = S_1 + S_2 + S_3, \tag{20}$$

where

$$S_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\delta_i \varepsilon_i}{\pi(X_i, \alpha)} + (f(X_i, \beta) - \mu) \right\},$$

$$S_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{1}{\pi(X_i, \hat{\alpha})} - \frac{1}{\pi(X_i, \alpha)} \right) \delta_i \varepsilon_i,$$

$$S_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( 1 - \frac{\delta_i}{\pi(X_i, \hat{\alpha})} \right) F_i^T(\beta) (\hat{\beta}_c - \beta),$$

By the central theorem, we have  $S_1 \xrightarrow{D} N(0, V)$ . To prove the asymptotically normality of estimator, we need to prove that  $S_k = o_p(1), k = 2, 3, 4$ . For  $S_2$ ,

$$S_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{[\pi(X_i, \alpha) - \pi(X_i, \hat{\alpha})] \delta_i \varepsilon_i}{\pi^2(X_i, \alpha)} + o(1).$$

Similar to arguments of Qin and Lei [16], we have  $S_2 = o_p(1)$ .

For  $S_3$ , Under MAR assumption, we have  $E[(1 - \delta_i/\pi(X_i, \alpha)) F_i(\beta)] = 0$ . Therefore,

$\frac{1}{n} \sum_{i=1}^n [1 - \delta_i/\pi(X_i, \alpha)] F_i(\beta) = o_p(1)$ . This combines with Theorem 1 and  $\hat{\beta}_c - \beta = O_p(n^{-1/2})$  yields  $S_3 = o_p(1)$ . Then the Theorem 3 is proved.



**Submit or recommend next manuscript to SCIRP and we will provide best service for you:**

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.

A wide selection of journals (inclusive of 9 subjects, more than 200 journals)

Providing 24-hour high-quality service

User-friendly online submission system

Fair and swift peer-review system

Efficient typesetting and proofreading procedure

Display of the result of downloads and visits, as well as the number of cited articles

Maximum dissemination of your research work

Submit your manuscript at: <http://papersubmission.scirp.org/>