Published in final edited form as:

Stat Med. 2016 November 30; 35(27): 5009–5028. doi:10.1002/sim.7038.

An alternative empirical-likelihood method in missing response problems and causal inference

Kaili Ren^{a,*}, Christopher A. Drummond^b, Pamela S. Brewster^b, Steven T. Haller^b, Jiang Tian^b, Christopher J. Cooper^b, and Biao Zhang^a

^aDepartment of Mathematics and Statistics, The University of Toledo, Toledo, OH 43606, U.S.A

^bDepartment of Medicine, The University of Toledo, Toledo, OH 43614, U.S.A

Abstract

Missing responses are common problems in medical, social, and economic studies. When responses are missing at random, a complete-case data analysis may result in biases. A popular debias method is inverse probability weighting (IPW) proposed by Horvitz and Thompson [1]. To improve efficiency, Robins *et al.* [2, 3] proposed an augmented inverse probability weighting (AIPW) method. The AIPW estimator has a double-robustness property and achieves the semiparametric efficiency lower bound when the regression model and propensity score model are both correctly specified. In this paper, we introduce an empirical-likelihood-based estimator as an alternative to Qin and Zhang [4]. Our proposed estimator is also doubly robust and locally efficient. Simulation results show that the proposed estimator has better performance when the propensity score is correctly modeled. Moreover, the proposed method can be applied in the estimation of average treatment effect in observational causal inferences. Finally, we apply our method to an observational study of smoking, using data from the Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL) clinical trial [5].

Keywords

Average treatment effect; Causal inference; Empirical likelihood; Missing at random; Observational study; Propensity score

1. Introduction

Missing data occur frequently in medical, social, and economic studies. Missing responses, missing covariates, or both are possible missing data patterns. In this paper, we focus on missing response problems. We assume that responses are missing at random [6], which means that the missing indicator variable is conditionally independent of responses given the covariates. The simplest way to deal with missing data is complete-case (CC) analysis, i.e., deleting subjects with missing values and analyzing the subjects with complete observations. However, this approach may lose efficiency and lead to biased results unless the missing data mechanism is missing completely at random [6].

*Correspondence to: Kaili Ren, Department of Mathematics and Statistics, The University of Toledo, Toledo, OH 43606, U.S.A.; Kaili.Ren@rockets.utoledo.edu.

The regression method is an efficient approach to estimate the mean response. After fitting a regression model from the complete-case data, the regression estimator can be derived by the mean of the fitted values from all subjects; this is motivated from survey sampling methods [7]. Another common approach is the inverse probability weighting (IPW) method motivated by Horvitz and Thompson [1]. This method weights the complete-case response by the inverse of estimated selection probability or propensity score [8]. Both of these methods are simple to use; however, they are not consistent if the regression model or the propensity score model is misspecified. Robins *et al.* [2, 3] proposed an augmented inverse probability weighting (AIPW) method, which extends the IPW method by adding a regression term as an augmentation. The AIPW estimator has a double-robustness property, i.e., the estimator is consistent if either the outcome regression model or the propensity score model is correctly specified. It also achieves the semiparametric efficiency lower bound when the regression model and the propensity score model are both correctly specified. Various doubly robust and multiply robust estimators have been studied in recent years; see for example, [4, 9–15].

In this paper, we propose an empirical-likelihood-based estimator as an alternative to Qin and Zhang [4], which made use of the empirical likelihood approach [16–19]. Our work differs from that of Qin and Zhang [4] in three aspects as described below. Firstly, instead of using a set of known functions $a(x)$ as in Qin and Zhang [4], we postulate a working regression model which possibly involves unknown parameters but reduces the number of constraints and lowers the implementation difficulty. Secondly, instead of constructing the constraint equations by calibrating both the propensity score and the functions $a(x)$ as in Qin and Zhang [4], we only employ the calibration constraint equation to match the first-order moment conditions of the estimated working regression function between the complete-case subsample and the full sample, with no calibration weighting on the constraint of the working propensity score; this gives rise to a different empirical likelihood method than that of Qin and Zhang [4] by treating the expected working propensity score as an unknown parameter in the constraint equations and estimating it through a profile empirical likelihood function. Although an additional parameter is introduced in the constraint equations because no calibration is imposed on the working propensity score, it turns out that the reciprocal of the expected propensity score parameter is constrained to be equal to one of the two Lagrange multipliers. This implies that in terms of calculating the probability masses supported on the observed data, the computational effort of our proposed empirical likelihood method is comparable to that in Qin and Zhang [4] because the number of the unknown parameters in our estimating equations is the same as that in the estimating equations of Qin and Zhang [4] when the dimension of $a(x)$ equals one. Finally, our proposed partial calibration weighting method using empirical likelihood yields a different asymptotic variance of the estimated mean response from that produced by the full calibration weighting empirical likelihood method of Qin and Zhang [4]; neither dominates each other asymptotically, though simulation results show that our proposed estimator has better performance when the working propensity score is correctly specified. Apart from these three different aspects, our proposed estimator shares the double-robustness property and achieves the semiparametric efficiency lower bound when the working regression model and the working propensity score model are both correctly specified.

Estimation of the average treatment effect (ATE) is often the basis of epidemiologic and econometric studies. In a randomized experiment, the ATE can be estimated by simply employing the sample mean difference. However, in an observational study, the treatment assignment may depend on covariates; thus, the sample mean difference may no longer be consistent. In this case, the estimation of ATE requires adjustment for confounding factors. The propensity score proposed by Rosenbaum and Rubin [8] plays a critical role in observational causal inference. It can be used to adjust for confounding factors through matching, stratification, and weighting; see for example, [4, 20–25]. If we consider the control responses as missing data for subjects in the treated group, the estimation of the mean potential control response can be viewed as a one-sample missing response problem. In the same manner, the mean potential treated response can be estimated by viewing the treated responses as missing data for subjects in the control group. As a result, methods developed in the missing data problems can be applied in the estimation of the ATE in observational causal inference.

Cigarette smoking may worsen renal function in people with diabetes mellitus and primary kidney diseases [26–29]; however, the effect of cigarette smoking on renal function in patients with atherosclerotic renal artery stenosis (ARAS) is uncertain. In this context, we apply our proposed method and several other methods on a dataset from the Cardiovascular Outcomes in Renal Atherosclerotic Lesions (CORAL) clinical trial [5] to study the causal effect of cigarette smoking on renal function in patients with ARAS; see Drummond *et al.* [30]. The CORAL study was a prospective, international, randomized clinical trial which compared medical therapy only with medical therapy plus stenting in patients with ARAS, followed from May 2005 to January 2010. Randomization was carried out by an interactive voice randomization system (IVRS) with the use of a permuted block design [5]. Active cigarette smoking, defined as regular tobacco use within one year prior to enrollment in the study, was observed after randomization; hence, the study of the causal effect of smoking on renal function of ARAS patients is an observational study. In the CORAL study, 277 (30%) of the 931 enrolled patients were self-reported smokers.

The paper is organized as follows. Section 2 presents empirical-likelihood-based estimators in one-sample missing response problem and two-sample missing response problem (causal inference). Section 3 provides theoretical properties and asymptotic distributions of the proposed estimators. Simulations studied for both one-sample and two-sample missing data problems are given in Section 4. Section 5 presents an application of the proposed method based on a dataset from the CORAL clinical trial [5]. Section 6 contains concluding remarks. Proofs of theoretical results are given in the Appendix.

2. Methodology

2.1. One-sample missing response problem

As in Qin and Zhang [4], we consider the standard missing data set-up. Let Y , X , D be the response variable, covariate vector, and missing indicator respectively, where $D = 1$ or 0 as Y is observed or missing, and X is always observed. Our goal is to estimate the population mean

$$\mu = E(Y) = \int \int y f(y, x) dx dy,$$

where $f(y, x)$ is the joint density function of (Y, X) , and μ_0 is the true value of μ .

Let (D_i, Y_i, X_i) , $i = 1, \dots, n$ denote the observed data. Without loss of generality, we index the subjects with observed response by $i = 1, \dots, n_1$, where $n_1 = \sum_{i=1}^n D_i$. Our proposed method requires making assumptions about the propensity score $P(D = 1/X = x)$ and the conditional expectation $E(Y|X = x)$, which are denoted as $\pi(x)$ and $m(x)$, respectively. A parametric working propensity score model $\pi(x, \gamma)$ for $\pi(x)$ is often postulated by researchers, where γ is a $\pi \times 1$ unknown vector parameter, and γ is often estimated by maximizing the binomial likelihood

$$\prod_{i=1}^n \pi(X_i, \gamma)^{D_i} \{1 - \pi(X_i, \gamma)\}^{1-D_i}. \quad (1)$$

The most common choice of the propensity score model is the logistic regression model $\pi(x, \gamma) = \exp(\gamma^T x) / \{1 + \exp(\gamma^T x)\}$. Similarly, we can posit a parametric working regression model $m(x, \beta)$ for $m(x)$, where β is a $q \times 1$ unknown vector parameter that can be estimated from complete-case data.

The conditional likelihood founded on (Y_i, X_i) , given $D_i = 1$, $i = 1, \dots, n_1$ can be written as

$$L = \prod_{i=1}^{n_1} \frac{\pi(X_i, \gamma) p_i}{\theta}, \quad (2)$$

where

$$\theta = \int \pi(x) f(y, x) dx dy = E\{\pi(X)\}$$

and $p_i = f(Y_i, X_i) = dF(Y_i, X_i)$, $i = 1, \dots, n_1$, denote positive jumps with sum 1, where $F(y, x)$ is the joint cumulative distribution function of (Y, X) . Accordingly, we can now treat the inference on the conditional likelihood (2) as a biased sampling problem similar to Vardi [31, 32]. To obtain a more efficient empirical-likelihood-based estimator, we maximize the conditional likelihood (2) under the following constraints

$$\sum_{i=1}^{n_1} p_i = 1, \quad \sum_{i=1}^{n_1} p_i \{\pi(X_i, \hat{\gamma}) - \theta\} = 0, \quad \sum_{i=1}^{n_1} p_i \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\} = 0, \quad (3)$$

where $\hat{\gamma}$ is the maximizer of binomial likelihood (1), $\hat{\beta}$ is the coefficient of the regression model $m(x, \beta)$, and $\hat{m}(\hat{\beta}) = n^{-1} \sum_{i=1}^n m(X_i, \hat{\beta})$. The first constraint corresponds to the truth that the total jumps equals 1. The second constraint reflects the selection bias. The third and final constraint is to improve efficiency by using the regression function. For fixed (θ, γ) , applying the method of Lagrange multipliers shows that the maximum value of L is attained at

$$p_i(\theta) = \frac{1}{n_1} \frac{1}{1 + \lambda_1 \{\pi(X_i, \hat{\gamma}) - \theta\} + \lambda \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}, \quad i=1, \dots, n_1, \quad (4)$$

where λ_1 and λ are Lagrange multipliers. Substituting $\pi_i(\theta)$'s into the conditional likelihood (2), the profiling likelihood of θ is

$$L(\theta) = \prod_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma})}{\theta} \frac{1}{n_1} \frac{1}{1 + \lambda_1 \{\pi(X_i, \hat{\gamma}) - \theta\} + \lambda \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}.$$

Maximizing the profiling likelihood function by differentiating the log-likelihood $\ell(\theta)$, where $\ell(\theta) = \log \{L(\theta)\}$, with respect to θ and setting the derivative to 0, we obtain

$$\lambda_1 = \frac{1}{\theta}.$$

Under constraints (3), θ and λ satisfy

$$\begin{aligned} \sum_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma}) - \theta}{\theta^{-1} \pi(X_i, \hat{\gamma}) + \lambda \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} &= 0, \\ \sum_{i=1}^{n_1} \frac{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})}{\theta^{-1} \pi(X_i, \hat{\gamma}) + \lambda \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} &= 0. \end{aligned} \quad (5)$$

Suppose $(\hat{\theta}, \hat{\lambda})^T$ is a solution of equations (5). Then from (4) we obtain

$$\hat{p}_i = \frac{1}{n_1} \frac{1}{\hat{\theta}^{-1} \pi(X_i, \hat{\gamma}) + \hat{\lambda} \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}}, \quad i=1, \dots, n_1. \quad (6)$$

It turns out that our proposed estimator is given by

$$\hat{\mu} = \sum_{i=1}^{n_1} \hat{p}_i Y_i = \frac{1}{n_1} \sum_{i=1}^n \frac{1}{\hat{\theta}^{-1} \pi(X_i, \hat{\gamma}) + \hat{\lambda} \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} D_i Y_i. \quad (7)$$

Remark 1—A major difference between our proposed method and Qin and Zhang's [4] empirical likelihood method is in the second constraint of (3). Their approach is to calibrate the estimated working propensity score $\pi(x, \hat{\gamma})$ by matching its first-order moments from the complete-case covariate vector $\{(X_i, D_i = 1), i = 1, \dots, n\}$ to the full-data covariate vector $(X_1, \dots, X_n)$. By contrast, our approach matches the first-order moment of $\pi(x, \hat{\gamma})$ from $\{(X_i, D_i = 1), i = 1, \dots, n\}$ to the expected working propensity score $\theta = E\{\pi(X, \gamma)\}$, an additional unknown parameter introduced into the second constraint equation in (3) due to no calibration on $\pi(x, \hat{\gamma})$. In other words, Qin and Zhang [4] performed calibration on both the propensity score $\pi(x, \hat{\gamma})$ and the known functions $a(x)$ through matching between the complete-case subsample and the full sample on the covariate vector X , whereas we only perform partial calibration on $m(x, \hat{\beta})$ and impose no calibration constraint on $\pi(x, \hat{\gamma})$. As a result, the expected working propensity score θ is estimated differently depending on whether or not calibration is performed on $\pi(x, \hat{\gamma})$. In Qin and Zhang [4], θ is estimated by the sample mean propensity score $\tilde{\theta} = n^{-1} \sum_{i=1}^n \pi(X_i, \hat{\gamma})$, which only uses the observed data on (D, X) and the estimated working propensity score $\pi(x, \hat{\gamma})$. It is worth pointing out that $\tilde{\theta}$ reduces to the non-missing proportion n_1/n for any logistic propensity score with an intercept. In contrast to their estimator $\tilde{\theta}$, our proposed estimator $\hat{\theta}$ is obtained by maximizing the profile likelihood function $L(\theta)$ and has the potential advantages of utilizing all the observed data on (DY, X, D) and extracting useful information from both working models $p(x, \hat{\gamma})$ and $m(x, \hat{\beta})$.

2.2. Causal inference or two-sample missing response problem

Let D denote an indicator of treatment exposure such that $D = 1$ if treated and $D = 0$ if control. Let X denote a covariate vector which is not affected by either treatment. Denote $Y(0)$ and $Y(1)$ as potential outcomes, respectively, when $D = 0$ and $D = 1$. $Y_i(1) - Y_i(0)$ represents the treatment effect for the i th subject. However, it cannot be observed, since each subject can only be in the treated group or in the control group, not both. Accordingly, the actual observed outcome Y is written as

$$Y = DY(1) + (1-D)Y(0),$$

and (Y_i, X_i, D_i) , $i = 1, \dots, n$ are observed values. Nevertheless, causal effects are still comparisons between potential outcomes, which can be measured by the average treatment effect (ATE), defined as

$$\Delta = E\{Y(1) - Y(0)\} = \mu^1 - \mu^0. \quad (8)$$

In an observational study, the propensity score is the conditional probability of receiving treatment given the covariate vector X [8], which is

$$\pi(x) = P(D=1|X=x), \quad 0 < \pi(x) < 1.$$

In addition, if the strongly ignorable assumption [8] holds, which is

$$\{Y(0), Y(1)\} \perp\!\!\!\perp D|X,$$

the estimation of μ^1 in causal inference can be considered as a two-sample missing response problem under the missing at random assumption. The two samples are $(Y_i(1), D_i, X_i)$ and $(Y_i(0), D_i, X_i)$, $i = 1, \dots, n$, where $Y_i(1)$ and $Y_i(0)$ are missing if $D_i = 0$ and $D_i = 1$, respectively. Then we can estimate μ^1 and μ^0 in (8) separately by our proposed method in Section 2.1. Denote $m_j(x) = E\{Y(j)|X=x\}$, $j = 0, 1$. Then we can postulate parametric models $\pi(x, \gamma)$, $m_0(x, \beta^0)$, and $m_1(x, \beta^1)$, for $\pi(x)$, $m_0(x)$, and $m_1(x)$, respectively, where γ can be estimated from the binomial likelihood function which has the same format as (1), β^j can be estimated from the complete-case data of $(Y_i(j), D_i, X_i)$, $i = 1, \dots, n$, and $j = 0, 1$.

Write $\hat{m}_j(\hat{\beta}^j) = n^{-1} \sum_{i=1}^n m_j(X_i, \hat{\beta}^j)$, $n_1 = \sum_{i=1}^n D_i$, and $n_0 = n - n_1$. On the basis of the methodology in Section 2.1, μ^1 is estimated by

$$\hat{\mu}^1 = \frac{1}{n_1} \sum_{i=1}^n \frac{1}{(\hat{\theta}^1)^{-1} \pi(X_i, \hat{\gamma}) + \hat{\lambda}^1 \{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} D_i Y_i, \quad (9)$$

where $\hat{\theta}^1$ and $\hat{\lambda}^1$ satisfy the equations

$$\begin{aligned} \sum_{i=1}^{n_1} \frac{\pi(X_i, \hat{\gamma}) - \theta^1}{(\theta^1)^{-1} \pi(X_i, \hat{\gamma}) + \lambda^1 \{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} &= 0, \\ \sum_{i=1}^{n_1} \frac{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)}{(\theta^1)^{-1} \pi(X_i, \hat{\gamma}) + \lambda^1 \{m_1(X_i, \hat{\beta}^1) - \hat{m}_1(\hat{\beta}^1)\}} &= 0. \end{aligned}$$

Similarly, μ^0 is estimated by

$$\hat{\mu}^0 = \frac{1}{n_0} \sum_{i=1}^n \frac{1}{(\hat{\theta}^0)^{-1} \{1 - \pi(X_i, \hat{\gamma})\} + \hat{\lambda}^0 \{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} (1 - D_i) Y_i, \quad (10)$$

where $\hat{\theta}^0$ and $\hat{\lambda}^0$ satisfy the equations

$$\sum_{i=1}^{n_0} \frac{1 - \pi(X_i, \hat{\gamma}) - \theta^0}{(\theta^0)^{-1} \{1 - \pi(X_i, \hat{\gamma})\} + \lambda^0 \{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} = 0,$$

$$\sum_{i=1}^{n_0} \frac{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)}{(\theta^0)^{-1} \{1 - \pi(X_i, \hat{\gamma})\} + \lambda^0 \{m_0(X_i, \hat{\beta}^0) - \hat{m}_0(\hat{\beta}^0)\}} = 0.$$

According to (9) and (10), we propose the estimate by $\hat{\mu} = \hat{\mu}^1 - \hat{\mu}^0$.

3. Theoretical Properties

3.1. One-sample missing response: consistency when the working regression model is correctly specified

Suppose that $m(x)$ is correctly modeled by $m(x, \beta)$. Denote the true value of β as β_0 such that $m(x, \beta_0) = m(x)$, then $\hat{\beta} \rightarrow \beta_0$ in probability, and, moreover, $\hat{m}(\hat{\beta}) \rightarrow \mu_0$ in probability. Applying the results of White [33], $\hat{\gamma} \rightarrow \gamma_0^*$ in probability under suitable regularity conditions, where γ_0^* is the value that minimizes the Kullback-Leibler discrepancy

$$\int \log\{\pi(x)/\pi(x, \gamma)\} \pi(x) dx$$

with respect to γ . In addition, suppose $(\hat{\theta}, \hat{\lambda})^T$ is a solution of equations (5). We can prove that $(\hat{\theta}, \hat{\lambda})^T \rightarrow (\theta_0^*, \lambda_0^*)^T$ under suitable regularity conditions. Then the consistency of $\hat{\mu}$ can be obtained as follows after some algebra, which is

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{n_1} \hat{p}_i \{Y_i - m(X_i, \hat{\beta})\} + \sum_{i=1}^{n_1} \hat{p}_i m(X_i, \hat{\beta}) \\ &= \frac{n}{n_1} \frac{1}{n} \sum_{i=1}^n \frac{D_i \{Y_i - m(X_i, \hat{\beta})\}}{\hat{\theta}^{-1} \pi(X_i, \hat{\gamma}) + \hat{\lambda} \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} + \hat{m}(\hat{\beta}) \\ &\rightarrow \frac{1}{P(D=1)} E \left[\frac{D\{Y - m(X)\}}{\theta_0^{*-1} \pi(X, \gamma^*) + \lambda_0^* \{m(X) - \mu_0\}} \right] + \mu_0 = \mu_0 \end{aligned}$$

in probability. It follows that

Theorem 1—If the regression model $m(x, \beta)$ is correctly specified, $\hat{\mu}$ is a consistent estimator of μ_0 .

3.2. One-sample missing response: asymptotic distribution, consistency, and efficiency when the working propensity score is correctly specified

Now suppose that $\pi(x)$ is correctly modeled by $\pi(x, \gamma)$. Denote the true value of γ as γ_0 such that $\pi(x, \gamma_0) = \pi(x)$. Applying the results of White [33], $\hat{\beta} \rightarrow \beta_0^*$ in probability under suitable regularity conditions. Furthermore, $\hat{m}(\hat{\beta}) \rightarrow m_0^*$ in probability, where $m_0^* = E\{m(X, \beta_0^*)\}$. Write

$$\begin{aligned}
\pi_1(x, \gamma) &= \partial \pi(x, \gamma) / \partial \gamma, \\
A &= \frac{\{D - \pi(X, \gamma_0)\} \pi_1(X, \gamma_0)}{\pi(X, \gamma_0) \{1 - \pi(X, \gamma_0)\}}, \\
B &= E \left[\frac{1}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\}^2 \right], \\
G &= E \left[\frac{Y - \mu_0}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\} \right], \\
H &= \frac{D(Y - \mu_0)}{\pi(X, \gamma_0)} - G B^{-1} \frac{D - \pi(X, \gamma_0)}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\}. \quad (11)
\end{aligned}$$

Theorem 2—If the propensity score model $\pi(x, \gamma)$ is correctly specified. Under suitable regularity conditions, $\hat{\mu}$ is a consistent estimator of μ_0 . Moreover, as $n \rightarrow \infty$,

$$n^{1/2}(\hat{\mu} - \mu_0) \rightarrow N(0, \text{Var}(K))$$

in distribution, where

$$K = H - E(HA^T) \{E(AA^T)\}^{-1} A. \quad (12)$$

The proof of theorem 2 is given in the Appendix.

From a geometric viewpoint, the term $E(HA^T) \{E(AA^T)\}^{-1} A$ in (12) can be viewed as the orthogonal projection of H onto the linear space spanned by A , and the influence function K of the proposed empirical likelihood estimator $\hat{\mu}$ can be regarded as the residual from this projection. As a result, $\text{Var}(K) = \text{Var}(H)$. In contrast, the empirical likelihood estimator $\hat{\mu}_{\text{EL}}$ of μ proposed by Qin and Zhang [4] possesses a different influence function in that it is the residual $K_{\text{EL}} = H_{\text{EL}} - E(H_{\text{EL}}A^T) \{E(AA^T)\}^{-1} A$ from the projection of H_{EL} onto the linear space spanned by A , where

$$H_{\text{EL}} = \frac{D(Y - \mu_0)}{\pi(X, \gamma_0)} - G_{\text{EL}}^T B_{\text{EL}}^{-1} \frac{D - \pi(X, \gamma_0)}{\pi(X, \gamma_0)} \begin{pmatrix} \pi(X, \gamma_0) - \theta_0 \\ a(X) - a_0 \end{pmatrix}$$

with $a_0 = E\{a(X)\}$ and

$$\begin{aligned}
G_{\text{EL}} &= E \left\{ \frac{Y - \mu_0}{\pi(X, \gamma_0)} \begin{pmatrix} \pi(X, \gamma_0) - \theta_0 \\ a(X) - a_0 \end{pmatrix} \right\}, \\
B_{\text{EL}} &= E \left\{ \frac{1}{\pi(X, \gamma_0)} \begin{pmatrix} \pi(X, \gamma_0) - \theta_0 \\ a(X) - a_0 \end{pmatrix} \begin{pmatrix} \pi(X, \gamma_0) - \theta_0 \\ a(X) - a_0 \end{pmatrix}^T \right\}.
\end{aligned}$$

This implies that the proposed estimator $\hat{\mu}$ and the estimator $\hat{\mu}_{\text{EL}}$ of Qin and Zhang [4] have different asymptotic variances $\text{Var}(K)$ and $\text{Var}(K_{\text{EL}})$; neither estimator appears to dominate

the other one in terms of having a smaller asymptotic variance. The simulation study presented in the next section shows favorable results and improvement of $\hat{\mu}$ under the correct working propensity score.

To optimize the cost and accuracy of a study, planned missing data designs are usually applied by researchers. In this case, $\pi(x)$ is known. It follows that,

Corollary 1—*If the true propensity score $\pi(x)$ is known, we substitute $\pi(x)$ for $\pi(x, \hat{\gamma})$ in (7). Under suitable regularity conditions, as $n \rightarrow \infty$,*

$$n^{1/2}(\hat{\mu} - \mu_0) \rightarrow N(0, \text{Var}(H))$$

in distribution, where H is defined in (11).

The proof of Corollary 1 is similar to that of Theorem 2 and is omitted. Although $\pi(x)$ is known in planned missing data designs, since $\text{Var}(K) = \text{Var}(H)$, we can improve the efficiency of $\hat{\mu}$ by postulating a model for $\mu(x)$. This is a well-known counterintuitive result [3].

Corollary 2—*If the propensity score model $\pi(x, \gamma)$ and the working regression model $m(x, \beta)$ are both correctly specified, the asymptotic variance $\text{Var}(H_B)$ reaches the semiparametric efficiency lower bound, where*

$$H_B = \frac{D(Y - \mu_0)}{\pi(X)} - \frac{D - \pi(X)}{\pi(X)} \{m(X) - \mu_0\} = \frac{D\{Y - m(X)\}}{\pi(X)} + m(X) - \mu_0.$$

Proof: If $\pi(x, \gamma)$ and $m(x, \beta)$ are both correctly specified, in (11) and (12),

$$B = G = E \left[\frac{1}{\pi(X)} \{m(X) - \mu_0\}^2 \right],$$

which leads to

$$H = H_B = \frac{D(Y - \mu_0)}{\pi(X)} - \frac{D - \pi(X)}{\pi(X)} \{m(X) - \mu_0\} \text{ and } E(HA^T) = 0.$$

Therefore, we have $K = H_B$, and the asymptotic variance $\text{Var}(H_B)$ equals the semiparametric efficiency lower bound [22, 34]. The proof is complete.

3.3. Theoretical properties in causal inference

Let (μ_0^0, μ_0^1) be the true value of (μ^0, μ^1) . When $\pi(x)$ is correctly modeled by $\pi(x, \gamma)$, denote the true value of γ as γ_0 such that $\pi(x, \gamma_0) = \pi(x)$. Applying the results of White [33], $\hat{\beta}^j \rightarrow \beta_0^{j*}, j=0, 1$, in probability under suitable regularity conditions. Furthermore, for $j=0, 1$, $\hat{m}_j(\hat{\beta}^j) \rightarrow m_{j0}^*$ in probability, where $m_{j0}^* = E\{m_j(X, \beta_0^{j*})\}$. For $j=0, 1$, write

$$\begin{aligned} \pi_1(x, \gamma) &= \partial \pi(x, \gamma) / \partial \gamma, \\ A &= \frac{\{D - \pi(X, \gamma_0)\} \pi_1(X, \gamma_0)}{\pi(X, \gamma_0) \{1 - \pi(X, \gamma_0)\}}, \\ B_j &= E \left[\frac{1}{\{\pi(X)\}^j \{1 - \pi(X)\}^{1-j}} \{m_j(X, \beta_0^{j*}) - m_{j0}^*\}^2 \right], \\ G_j &= E \left[\frac{Y - \mu_0^j}{\{\pi(X)\}^j \{1 - \pi(X)\}^{1-j}} \{m_j(X, \beta_0^{j*}) - m_{j0}^*\} \right], \\ H_j &= \frac{D^j (1-D)^{1-j} (Y - \mu_0^j)}{\{\pi(X)\}^j \{1 - \pi(X)\}^{1-j}} - (-1)^{j+1} G_j B_j^{-1} \frac{D - \pi(X)}{\{\pi(X)\}^j \{1 - \pi(X)\}^{1-j}} \{m_j(X, \beta_0^{j*}) - m_{j0}^*\}, \\ H_\Delta &= H_1 - H_0. \end{aligned} \quad (13)$$

Theorem 3—Under suitable regularity conditions, our proposed estimator $\hat{\Delta}$ has the following properties:

- a. $\hat{\Delta}$ is a doubly robust estimator, i.e. it is a consistent estimator of Δ_0 if either the propensity score model $\pi(x, \gamma)$ or the regression models $m_j(x, \beta^j), j=0, 1$, are correctly specified.
- b. If the propensity score model $\pi(x, \gamma)$ is correctly specified. As $n \rightarrow \infty$,

$$n^{1/2}(\hat{\Delta} - \Delta_0) \rightarrow N(0, \text{Var}(K_\Delta))$$

in distribution, where

$$K_\Delta = H_\Delta - E(H_\Delta A^T) \{E(AA^T)\}^{-1} A.$$

- c. If the true propensity score $\pi(x)$ is known, we substitute $\pi(x)$ for $\pi(x, \gamma)$ in (9) and (10). As $n \rightarrow \infty$,

$$n^{1/2}(\hat{\Delta} - \Delta_0) \rightarrow N(0, \text{Var}(H_\Delta))$$

in distribution, where H is defined in (13).

- d. If the propensity score model $\pi(x, \gamma)$ and the working regression models $m_j(x, \beta^j)$ are both correctly specified. The asymptotic variance $\text{Var}(H_B)$ reaches the semiparametric efficiency lower bound, where

$$H_{\Delta B} = \frac{D(Y-\mu_0^1)}{\pi(X)} - \frac{(1-D)(Y-\mu_0^0)}{1-\pi(X)} - \left[\frac{D-\pi(X)}{\pi(X)} \{m_1(X) - \mu_0^1\} - \frac{\pi(X)-D}{1-\pi(X)} \{m_0(X) - \mu_0^0\} \right] \\ = \frac{D\{Y-m_1(X)\}}{\pi(X)} - \frac{(1-D)\{Y-m_0(X)\}}{1-\pi(X)} + m_1(X) - m_0(X) - \Delta_0.$$

The proof of Theorem 3 is similar to proofs of Theorem 1, Theorem 2, Corollary 1, and Corollary 2, and is omitted.

4. Simulation studies

4.1. One-sample missing response problem

The first simulation study is presented to compare the performance of relative estimators and our proposed estimator $\hat{\mu}$ in a one-sample missing response problem. The relative estimators

include full data sample mean $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$, complete-case estimator

$\hat{\mu}_{CC} = \sum_{i=1}^n D_i Y_i / \sum_{i=1}^n D_i$, regression estimator $\hat{\mu}_{REG} = n^{-1} \sum_{i=1}^n m(X_i, \hat{\beta})$, Horvitz-Thompson (HT) estimator [1]

$$\hat{\mu}_{HT} = n^{-1} \sum_{i=1}^n D_i Y_i / \pi(X_i, \hat{\gamma}),$$

inverse probability weighting (IPW) estimator [35]

$$\hat{\mu}_{IPW} = \frac{\sum_{i=1}^n D_i Y_i / \pi(X_i, \hat{\gamma})}{\sum_{i=1}^n D_i / \pi(X_i, \hat{\gamma})},$$

augmented inverse probability weighting (AIPW) estimator [2]

$$\hat{\mu}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{D_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} m(X_i, \hat{\beta}) \right\},$$

and the empirical-likelihood-based (EL) estimator [4]

$$\hat{\mu}_{EL} = \sum_{i=1}^n \frac{\hat{\theta} \pi^{-1}(X_i, \hat{\gamma})}{1 + \hat{\eta}^T r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})} D_i Y_i / \sum_{i=1}^n D_i,$$

where $\hat{\theta} = n^{-1} \sum_{i=1}^n \pi(X_i, \hat{\gamma})$, $\mathbf{a}(x) = (a_1(x), \dots, a_r(x))^T$ are r independent known functions,

$\hat{a} = (n^{-1} \sum_{i=1}^n a_1(X_i), \dots, n^{-1} \sum_{i=1}^n a_r(X_i))^T$, and $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2^T)^T$ is the solution of

$$\sum_{i=1}^{n_1} \frac{r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})}{1 + \eta^T r(X_i, \hat{\gamma}, \hat{\theta}, \hat{a})} = 0$$

with

$$r(x, \gamma, \theta, a) = \begin{pmatrix} 1 - \theta \pi^{-1}(x, \gamma) \\ \pi^{-1}(x, \gamma) \{a(x) - a\} \end{pmatrix}.$$

We generate data by the following process: $X \sim \text{Un}(-2.5, 2.5)$, $D|X = x \sim \text{Ber}\{\pi(x)\}$, and $Y|X = x \sim N\{m(x), 4x^2 + 2\}$, where $\pi(x) = 1 - \{1 + \exp(-0.1 + 0.5x - 0.3 \exp(x))\}^{-1}$ and $m(x) = 1 + 2x + 3x^2$ such that the missing rate is around 0.69 and $\mu_0 = 7.25$. The working propensity scores are $\pi_T(x, \gamma_T) = 1 - \{1 + \exp(\gamma_{T0} + \gamma_{T1}x + \gamma_{T2} \exp(x))\}^{-1}$ and $\pi_F(x, \gamma_F) = 1 - \exp\{-\exp(\gamma_{F0} + \gamma_{F1}x^2 + \gamma_{F2}x^4)\}$. The working regression models are $m_T(x, \beta_T) = \beta_{T0} + \beta_{T1}x + \beta_{T2}x^2$ and $m_F(x, \beta_F) = \beta_{F0} + \beta_{F1}x + \beta_{F2} \exp(x)$, or $a_T(x) = (x, x^2)^T$ and $a_F(x) = \{x, \exp(x)\}^T$ for μ_{EL} .

Let

$$d_X = \frac{\bar{X}_o - \bar{X}_m}{\sqrt{\frac{s_o^2 + s_m^2}{2}}} \quad (14)$$

denote the standardized difference between observed and missing groups of covariate X in one sample, where $\bar{X}_o$ and $\bar{X}_m$ denote the sample mean of covariate X in observed and missing groups, and s_o^2 and s_m^2 denote the sample variances, respectively.

Figure 1 presents the histograms with kernel density curves of $\bar{X}$ and $\overline{\exp(X)}$ in observed and missing groups from 5000 Monte Carlo samples with sample size 500. The left panel of Figure 1 shows that there is no significant difference between observed and missing groups in covariate X . Mean ($\pm$ SD) of $\bar{X}$ in observed and missing groups are -0.015 ± 0.101 and 0.004 ± 0.082 , respectively; the mean standardized difference between observed and missing groups in X is $\bar{d}_X = -0.014$ with 95% confidence interval -0.197 to 0.170 . In the meantime, the right panel of Figure 1 indicates a significant difference between the two groups in $\exp(X)$. Mean ($\pm$ SD) of $\overline{\exp(X)}$ in observed and missing groups are 1.97 ± 0.18 and 2.62 ± 0.17 , respectively; the mean standardized difference between observed and missing groups in X is $\bar{d}_{\exp(X)} = -0.23$ with 95% confidence interval -0.41 to -0.05 . In this case, the complete-case estimator $\hat{\mu}_{\text{CC}}$ will introduce biases, and therefore, debias methods should be applied.

5000 Monte Carlo simulations are then conducted with two sample sizes, 300 and 500, under four scenarios:

- a. both $\pi(x)$ and $m(x)$ are correctly modeled by $\pi_T(x, \gamma_T)$ and $m_T(x, \beta_T)$ or $a_T(x)$,
- b. $\pi(x)$ is correctly modeled by $\pi_T(x, \gamma_T)$, $m(x)$ is incorrectly modeled by $m_F(x, \beta_F)$ or $a_F(x)$,
- c. $m(x)$ is correctly modeled by $m_T(x, \beta_T)$ or $a_T(x)$, $\pi(x)$ is incorrectly modeled by $\pi_F(x, \gamma_F)$,
- d. both $\pi(x)$ and $m(x)$ are incorrectly modeled by $\pi_F(x, \gamma_F)$ and $m_F(x, \beta_F)$ or $a_F(x)$.

We examine and compare the bias and root-mean-squared error (RMSE) of each estimator. The results are shown in Table 1, Figure 2, and Figure 3 present the boxplots of 8 estimators from 5000 Monte Carlo samples of size 300 and 500, respectively.

Since $\bar{Y}$ is the sample mean with no missing data, it always performs the best, as expected, and we use it as a benchmark. Conversely, $\hat{\mu}_{CC}$ only uses information of the complete-case responses, because it always has the largest biases and RMSEs under MAR assumption, therefore, performs the worst under the four scenarios.

The first scenario considers the case that both $\pi(x)$ and $m(x)$ are correctly modeled. At $n = 300$ and $n = 500$, the bias of $\hat{\mu}_{REG}$, $\hat{\mu}_{HT}$, $\hat{\mu}_{IPW}$, $\hat{\mu}_{AIPW}$, $\hat{\mu}_{EL}$, and $\hat{\mu}$ are all small. Then, by comparing the RMSEs, $\hat{\mu}_{REG}$ performs the best. The RMSEs of $\hat{\mu}_{AIPW}$, $\hat{\mu}_{EL}$, and $\hat{\mu}$ are very close, which are smaller than the RMSEs of $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW}$.

The second scenario compares the performance of the estimators when $\pi(x)$ is correctly modeled, but $m(x)$ is incorrectly modeled. As expected, the bias of $\hat{\mu}_{REG}$ is large because of the misspecified working regression model. All estimators with working propensity score have small bias. Our proposed estimator is more efficient than other estimators, since the RMSE of $\hat{\mu}$ is smaller than other estimators, especially at $n = 300$, the RMSE of $\hat{\mu}$ is smaller than that of $\hat{\mu}_{EL}$ by 10%. The boxplots also show that $\hat{\mu}$ has fewer outliers compared to other estimators.

Under the third scenario, $m(x)$ is correctly modeled, but $\pi(x)$ is incorrectly modeled. Large biases demonstrate that $\hat{\mu}_{HT}$ and $\hat{\mu}_{IPW}$ are no longer consistent. The three doubly robust estimators have similar biases and RMSEs. However, the RMSE of $\hat{\mu}$ is still smaller than that of $\hat{\mu}_{EL}$.

The last scenario examines the case that both $\pi(x)$ and $m(x)$ are incorrectly modeled. All the estimators produce some biases to a certain extent, except for $\bar{Y}$; however, $\hat{\mu}_{EL}$ performs much better than other estimators by comparing the biases and RMSEs. The boxplots indicate that $\hat{\mu}_{EL}$ has fewer outliers than other estimators in this case. In addition, the performance of $\hat{\mu}_{AIPW}$ is the second worst except for $\hat{\mu}_{CC}$, which is consistent with the results given by Kang and Schafer [11].

4.2. Causal inference or two-sample missing response problem

The second simulation study is presented to compare the performance of relative estimators and our proposed estimator $\hat{\mu}$ in causal inference. Since causal inference can be viewed as a

two-sample missing response problem, the relative estimators can be directly derived from the methods used in one-sample missing response problem, defined as full data sample mean

difference $\bar{\Delta} = n^{-1} \sum_{i=1}^n \{Y_i(1) - Y_i(0)\}$, complete-case (CC) estimator

$\hat{\Delta}_{CC} = \sum_{i=1}^n D_i Y_i / \sum_{i=1}^n D_i - \sum_{i=1}^n (1-D_i) Y_i / \sum_{i=1}^n (1-D_i)$, regression estimator

$\hat{\Delta}_{REG} = n^{-1} \sum_{i=1}^n \{m_1(X_i, \hat{\beta}^1) - m_0(X_i, \hat{\beta}^0)\}$, Horvitz- Thompson (HT) estimator [1]

$\hat{\Delta}_{HT} = n^{-1} \sum_{i=1}^n [D_i Y_i / \pi(X_i, \hat{\gamma}) - (1-D_i) Y_i / \{1 - \pi(X_i, \hat{\gamma})\}]$, inverse probability weighting (IPW) estimator [35]

$$\hat{\Delta}_{IPW} = \frac{\sum_{i=1}^n D_i Y_i / \pi(X_i, \hat{\gamma})}{\sum_{i=1}^n D_i / \pi(X_i, \hat{\gamma})} - \frac{\sum_{i=1}^n (1-D_i) Y_i / \{1 - \pi(X_i, \hat{\gamma})\}}{\sum_{i=1}^n (1-D_i) / \{1 - \pi(X_i, \hat{\gamma})\}},$$

augmented inverse probability weighting (AIPW) estimator [2]

$$\hat{\Delta}_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{D_i Y_i}{\pi(X_i, \hat{\gamma})} - \frac{D_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} m_1(X_i, \hat{\beta}^1) \right\} - \frac{1}{n} \sum_{i=1}^n \left\{ \frac{(1-D_i) Y_i}{1 - \pi(X_i, \hat{\gamma})} + \frac{D_i - \pi(X_i, \hat{\gamma})}{\pi(X_i, \hat{\gamma})} m_0(X_i, \hat{\beta}^0) \right\},$$

and empirical-likelihood-based (EL) estimator [4]

$$\hat{\Delta}_{EL} = \sum_{i=1}^n \frac{\hat{\theta} \pi^{-1}(X_i, \hat{\gamma})}{1 + (\hat{\eta}^1)^T r_1(X_i, \hat{\gamma}, \hat{\theta}, \hat{a}_1)} D_i Y_i / \sum_{i=1}^n D_i - \sum_{i=1}^n \frac{(1-\hat{\theta}) \{1 - \pi(X_i, \hat{\gamma})\}^{-1}}{1 + (\hat{\eta}^0)^T r_0(X_i, \hat{\gamma}, \hat{\theta}, \hat{a}_0)} (1-D_i) Y_i / \sum_{i=1}^n (1-D_i),$$

where $\hat{\theta} = n^{-1} \sum_{i=1}^n \pi(X_i, \hat{\gamma})$; for $j = 0, 1$, $a_j(x) = (a_{j1}(x), \dots, a_{jr_j}(x))^T$ are r_j independent

known functions, $\hat{a}_j = (n^{-1} \sum_{i=1}^n a_{j1}(X_i), \dots, n^{-1} \sum_{i=1}^n a_{jr_j}(X_i))^T$, and $\hat{\eta}^j = (\hat{\eta}_1^j, (\hat{\eta}_2^j)^T)^T$ is the solution of

$$\sum_{i=1}^{n_j} \frac{r_j(X_i, \hat{\gamma}, \hat{\theta}, \hat{a}_j)}{1 + (\eta^j)^T r_j(X_i, \hat{\gamma}, \hat{\theta}, \hat{a}_j)} = 0$$

with

$$r_j(x, \gamma, \theta, a_j) = \begin{pmatrix} 1 - \theta^j (1 - \theta)^{1-j} \pi^{-j}(x, \gamma) \{1 - \pi(x, \gamma)\}^{-(1-j)} \\ \pi^{-j}(x, \gamma) \{1 - \pi(x, \gamma)\}^{-(1-j)} \{a_j(x) - a_j\} \end{pmatrix}.$$

We generate data by the following process: $Y(1) = 2 + 3X_1 + X_2 + \varepsilon_1$, $Y(0) = 1 + X_1 + 2X_2 + \varepsilon_0$, $X_1, X_2, \varepsilon_1, \varepsilon_0 \sim N(0, 1)$, $D|X = x \sim \text{Ber}\{\pi(x)\}$, where

$\pi(x) = 1 - \{1 + \exp(-1 + 0.5x_1 - 0.3x_2^2)\}^{-1}$ such that the average treatment rate is around 0.23

and $\pi_0 = 1$. The working propensity scores are

$\pi_T(x, \gamma_T) = 1 - \{1 + \exp(\gamma_{T0} + \gamma_{T1}x_1 + \gamma_{T2}x_2^2)\}^{-1}$ and $\pi_F(x, \gamma_F) = 1 - \{1 + \exp(\gamma_{F0} + \gamma_{F1}\exp(x_1) + \gamma_{F2}x_2)\}^{-1}$. The working regression models are

$m_{1T}(x, \beta_T^1) = \beta_{T0}^1 + \beta_{T1}^1 x_1 + \beta_{T2}^1 x_2$, $m_{0T}(x, \beta_T^0) = \beta_{T0}^0 + \beta_{T1}^0 x_1 + \beta_{T2}^0 x_2$ and $m_{1F}(x, \beta_F^1) = \beta_{F0}^1 + \beta_{F1}^1 x_1 + \beta_{F2}^1 x_2^2$, $m_{0F}(x, \beta_F^0) = \beta_{F0}^0 + \beta_{F1}^0 x_1^2 + \beta_{F2}^0 x_2$, or $a_{1T}(x) = (x_1, x_2)^T$, $a_{0T}(x) = (x_1, x_2)^T$ and $a_{1F}(x) = (x_1, x_2^2)^T$, $a_{0F}(x) = (x_1^2, x_2)^T$ for $\hat{\pi}_{EL}$.

Redefine d_X in (14) as the standardized difference between treated and control groups of covariate X in one sample. Figure 4 presents the histograms with kernel density curves of $\overline{X_1}$ and $\overline{X_2^2}$ in treated and control groups from 5000 Monte Carlo samples with sample size 500, which shows that there are significant differences between treated and control groups in both X_1 and X_2^2 . Mean ($\pm$ SD) of $\overline{X_1}$ in treated and control groups are 0.36 ± 0.09 and -0.11 ± 0.05 , respectively; the mean standardized difference between treated and control groups in X_1 is $\bar{d}_{X_1} = 0.48$ with 95% confidence interval 0.27 to 0.69; meanwhile, mean ($\pm$ SD) of $\overline{X_2^2}$ in treated and control groups are 0.68 ± 0.09 and 1.10 ± 0.08 , respectively; the mean standardized difference between treated and control groups in X_2^2 is $\bar{d}_{X_2^2} = -0.33$ with 95% confidence interval -0.52 to -0.15 . It implies that the strongly ignorable assumption [8] holds, and the sample mean difference estimation will be biased.

5000 Monte Carlo simulations are then conducted with two sample sizes, 300 and 500, under four scenarios:

- a. $\pi(x)$, $m_0(x)$, and $m_1(x)$ are all correctly modeled by $\pi_T(x, \gamma_T)$, $m_{0T}(x, \beta_T^0)$, and $m_{1T}(x, \beta_T^1)$ (or $\pi_T(x, \gamma_T)$, $a_{0T}(x)$, and $a_{1T}(x)$),
- b. $\pi(x)$ is correctly modeled by $\pi_T(x, \gamma_T)$, $m_0(x)$ and $m_1(x)$ are incorrectly modeled by $m_{0F}(x, \beta_F^0)$ and $m_{1F}(x, \beta_F^1)$ (or $a_{0F}(x)$ and $a_{1F}(x)$),
- c. $m_0(x)$ and $m_1(x)$ are correctly modeled by $m_{0T}(x, \beta_T^0)$ and $m_{1T}(x, \beta_T^1)$ (or $a_{0T}(x)$ and $a_{1T}(x)$), $\pi(x)$ is incorrectly modeled by $\pi_F(x, \gamma_F)$,
- d. $\pi(x)$, $m_0(x)$, and $m_1(x)$ are all incorrectly modeled by $\pi_F(x, \gamma_F)$, $m_{0F}(x, \beta_F^0)$, and $m_{1F}(x, \beta_F^1)$ (or $\pi_F(x, \gamma_F)$, $a_{0F}(x)$, and $a_{1F}(x)$).

Bias and root-mean-squared error (RMSE) of each estimator are given in Table 2, Figure 5, and Figure 6 present the boxplots of 8 estimators from 5000 Monte Carlo samples of size 300 and 500, respectively. Results in Table 2, Figure 5, and Figure 6 show that under the causal inference setting, the proposed estimator $\hat{\pi}$ performs better than other methods in most cases under four scenarios, except for $\hat{\pi}_{REG}$ when the working regression models are correctly specified. When the propensity score model is correctly specified and the regression models are misspecified, the RMSE of $\hat{\pi}$ is much less than that of $\hat{\pi}_{EL}$ (18% reduction when $n=300$). In this case, the boxplots also show that $\hat{\pi}$ has much fewer outliers than $\hat{\pi}_{EL}$ (Panel (b) in Figure 5 and Figure 6).

5. Application to the CORAL data

The CORAL study [5] was a prospective, international clinical trial of individuals with ARAS. Patients were randomly assigned to either a stenting plus medical therapy group or a medical therapy only group and were then followed to a maximum of 8 years; however, the study of the causal effect of smoking on renal function of patients with ARAS is an observational study, since smoking/non-smoking groups were not randomly assigned. As a result, differences of covariates in treated and control groups may cause biases if the sample average difference is used to estimate the ATE. Methods used in the simulation study are then applied to the CORAL baseline dataset to adjust for confounding factors. We examine the ATE of smoking on patients' renal function measured by cystatin C and CKD-EPI glomerular filtration rate (CKD-EPI GFR). Note that lower cystatin C and higher CKD-EPI GFR represent better renal function.

Of the 931 patients enrolled in the CORAL study, 277 (30%) were self-reported smokers. Smokers were significantly younger than non-smokers (63.3 ± 9.1 and 72.4 ± 7.8 years, respectively; $p < 0.001$), establishing age as a main confounder in the estimation of the smoking effect, since reduced renal function is related to advancing age [5]. The distributions of age in smokers and non-smokers are shown in Figure 7. Logistic regression and linear regression models are postulated as the working propensity score and the working regression models, respectively. After model selection, covariates X are determined to be Age, Gender, BMI, and Diabetes status. Also, we have Y =cystatin C or CKD-EPI GFR and D =Smoking status. Observe that of the 931 patients, 46 are missing CKD-EPI GFR, 45 are missing cystatin C, 4 are missing BMI, 13 are missing Diabetes status, and 10 are missing smoking status. It is plausible to assume that the missingness in these variables are completely at random; therefore, the patients with missing data are excluded. Based on complete-case data from the remaining 866 patients, results of the estimation of ATE are shown in Table 3.

The left half of Table 3 shows the ATE of smoking on cystatin C of the patients. Before the adjustment for confounding factors, sample mean comparison $\hat{\tau}_{CC} = -0.007$ with a P-value = 0.842, which implies that smoking has no effect on renal function of patients by means of cystatin C. Nevertheless, after adjusting for confounder, almost all estimators present a significant positive ATE at $\alpha = 0.05$, except for $\hat{\tau}_{HT}$, which implies smoking worsens renal function of patients by means of cystatin C.

Furthermore, the right half of Table 3 shows the ATE of smoking on CKD-EPI GFR of the patients. Before the adjustment for confounding factors, sample mean comparison $\hat{\tau}_{CC} = 5.68$ with a P-value = 0.002, which indicates smoking has a contradictory beneficial effect on renal function of patients by means of CKD-EPI GFR. However, after the adjustment for confounding factors, all estimators give an opposite sign of ATE; in addition, $\hat{\tau}_{REG}$, $\hat{\tau}_{AIPW}$, $\hat{\tau}_{EL}$ are significantly negative at $\alpha = 0.05$, which indicates smoking worsens renal function of patients by means of CKD-EPI GFR.

Finally, the higher cystatin C and lower CKD-EPI GFR for smokers demonstrated the negative effect of smoking on renal function for patients with ARAS.

6. Concluding remarks

In this paper, we have proposed an alternative empirical-likelihood-based approach to estimating mean response and causal effect under MAR assumption. In common with estimators proposed by Robins *et al.* [2] and Qin and Zhang [4], our proposed empirical-likelihood-based estimator also enjoys the double-robustness property and achieves the semiparametric efficiency lower bound when the regression model and the propensity score model are both correctly specified. Compared to Qin and Zhang [4], our approach postulates a working regression model instead of a set of known functions, which reduces calculation difficulty. Furthermore, our approach performs calibration only on the working regression function, whereas the approach of Qin and Zhang [4] performs calibration on both the working propensity score and the known functions $a(x)$. The difference in whether calibration is performed on the working propensity score entails that the estimation of the expected propensity score in our proposed method is likelihood-based, while the estimation of the expected propensity score in Qin and Zhang [4] is moment-based. Moreover, our proposed approach yields different asymptotic variances of the estimated mean response and ATE from those in Qin and Zhang [4]; neither dominates each other asymptotically. However, our simulations show favorable results and improvements of our approach under our simulation settings. Simulation results indicate that our proposed method is appreciably more efficient, in general, than its competitors when the working propensity score is correctly specified, especially when the working regression model is misspecified. It would be interesting to compare the two methods theoretically, and this provides a venue for further research.

Acknowledgments

The authors wish to thank the editor, the associate editor, and two reviewers for their thoughtful and helpful comments and suggestions. Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Numbers U01HL072734, U01HL072735, U01HL072736, and U01HL072737. Dr. Cooper and Ms. Brewster have received support from the National Heart, Lung, and Blood Institute (NHLBI), National Institutes of Health (5U01HL071556). Dr. Drummond is supported by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (1F32DK104615-01). Dr. Haller has received support from the National and Ohio Valley Affiliate of the American Heart Association (13POST16860035). Dr. Tian is supported by the NHLBI, the National Institutes of Health (1R01HL-105649). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Drugs for the study were provided by AstraZeneca, device support was provided by Cordis Corporation, and supplemental financial support was granted by both Cordis Corporation and Pfizer Inc.

References

1. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47(260):663–685.
2. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89(427):846–866.
3. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90(429):106–121.
4. Qin J, Zhang B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2007; 69(1):101–122.

5. Cooper CJ, Murphy TP, Cutlip DE, Jamerson K, Henrich W, Reid DM, Cohen DJ, Matsumoto AH, Steffes M, Jaff MR, et al. Stenting and medical therapy for atherosclerotic renal-artery stenosis. *New England Journal of Medicine*. 2014; 370(1):13–22. [PubMed: 24245566]
6. Little, RJ.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. Wiley; New York: 2002.
7. Cochran, WG. *Sampling techniques*. 3. Wiley; New York: 2007.
8. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70(1):41–55.
9. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics*. 2005; 61(4):962–973. [PubMed: 16401269]
10. Tan Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*. 2006; 101(476):1619–1637.
11. Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*. 2007; 22(4):523–539.
12. Qin J, Shao J, Zhang B. Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*. 2008; 103(482):797–810.
13. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*. 2009; 96(3):723–734. [PubMed: 20161511]
14. Zhang G, Little R. A comparative study of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*. 2011; 81(12):2039–2058.
15. Han P, Wang L. Estimation with missing data: beyond double robustness. *Biometrika*. 2013; 100(2):417–430.
16. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988; 75(2):237–249.
17. Owen AB. Empirical likelihood ratio confidence regions. *The Annals of Statistics*. 1990; 18(1):90–120.
18. Owen, AB. *Empirical likelihood*. Chapman and Hall/CRC; Boca Raton: 2001.
19. Qin J, Lawless J. Empirical likelihood and general estimating equations. *The Annals of Statistics*. 1994; 22(1):300–325.
20. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79(387):516–524.
21. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 1985; 39(1):33–38.
22. Hahn J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*. 1998; 66(2):315–332.
23. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003; 71(4):1161–1189.
24. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*. 2004; 23(19):2937–2960. [PubMed: 15351954]
25. Zhang B. Empirical likelihood in causal inference. *Econometric Reviews*. 2016; 35(2):201–231.
26. Stegmayr B. A study of patients with diabetes mellitus (type 1) and end-stage renal failure: tobacco usage may increase risk of nephropathy and death. *Journal of internal medicine*. 1990; 228(2):121–124. [PubMed: 2394962]
27. Orth SR, Stöckmann A, Conradt C, Ritz E, Ferroicw M, Kreusser W, Piccoli G, Rambašek M, Roccatello D, Schäfer K, et al. Smoking as a risk factor for end-stage renal failure in men with primary renal disease. *Kidney international*. 1998; 54(3):926–931. [PubMed: 9734618]
28. Shankar A, Klein R, Klein BE. The association among smoking, heavy drinking, and chronic kidney disease. *American journal of epidemiology*. 2006; 164(3):263–271. [PubMed: 16775042]
29. Obert DM, Hua P, Pilkerton ME, Feng W, Jaimes EA. Environmental tobacco smoke furthers progression of diabetic nephropathy. *The American journal of the medical sciences*. 2011; 341(2):126–130. [PubMed: 20924282]

30. Drummond CA, Brewster PS, He W, Ren K, Evans KL, Tuttle KR, Yu S, Dawson T, Haller ST, Jamerson K, et al. Cigarette smoking and cardio-renal events in patients with atherosclerotic renal artery stenosis. *Journal of the American Society of Hypertension*. 2015; 9(4):e53.
31. Vardi Y. Nonparametric estimation in the presence of length bias. *The Annals of Statistics*. 1982; 10(2):616–620.
32. Vardi Y. Empirical distributions in selection bias models. *The Annals of Statistics*. 1985; 13(1): 178–203.
33. White H. Maximum likelihood estimation of misspecified models. *Econometrica*. 1982; 50(1):1–25.
34. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*. 1995; 90(429):122–129.
35. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*. 2001; 2(3):259–278.

Appendix: Proof of Theorem 2

Based on the likelihood theory, $\hat{\gamma}$ is a solution of the score equation

$$U(\gamma) = \sum_{i=1}^n \frac{\{D_i - \pi(X_i, \gamma)\} \pi_1(X_i, \gamma)}{\pi(X_i, \gamma) \{1 - \pi(X_i, \gamma)\}} = 0$$

derived from the binomial likelihood (1). Taylor expansion of $U(\hat{\gamma})$ at γ_0 gives

$$\hat{\gamma} - \gamma_0 = \frac{1}{n} \{E(AA^T)\}^{-1} U(\gamma_0) + o_p(n^{-1/2}), \quad (15)$$

Write

$$C = E \left[\frac{\pi_1^T(X, \gamma_0)}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\} \right].$$

Then expanding the second equation of (5) at $(\theta_0, 0, \gamma_0, \beta_0^*)$ leads to

$$\begin{aligned} 0 &= \sum_{i=1}^{n_1} \frac{\theta_0}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} + \left[\sum_{i=1}^{n_1} \frac{1}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} \right] (\hat{\theta} - \theta_0) \\ &- \left[\sum_{i=1}^{n_1} \frac{\theta_0^2}{\pi^2(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\}^2 \right] \hat{\lambda} - \left[\sum_{i=1}^{n_1} \frac{\theta_0 \pi_1^T(X_i, \gamma_0)}{\pi^2(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} \right] (\hat{\gamma} - \gamma_0) \\ &+ \left[\sum_{i=1}^{n_1} \frac{\theta_0}{\pi(X_i, \gamma_0)} \left\{ \frac{\partial m(X_i, \beta_0^*)}{\partial \beta^T} - \frac{1}{n} \sum_{i=1}^n \frac{\partial m(X_i, \beta_0^*)}{\partial \beta^T} \right\} \right] (\hat{\beta} - \beta_0^*) + O_p(1) \\ &= \sum_{i=1}^n \frac{\theta_0 \{D_i - \pi(X_i, \gamma_0)\}}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\} - n\theta_0^2 B \hat{\lambda} - n\theta_0 C (\hat{\gamma} - \gamma_0) + o_p(n^{1/2}), \end{aligned}$$

which suggests that,

$$\hat{\lambda} = \frac{1}{n\theta_0} B^{-1} \sum_{i=1}^n \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\} - \frac{1}{\theta_0} B^{-1} C(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2}). \quad (16)$$

Next, based on the expansions (15), (16) and the result $n/n_1 = \theta_0^{-1} + o_p(1)$, we expand $\hat{\mu} - \mu_0$ at $(\theta_0, 0, \gamma_0, \beta_0^*)$, which gives

$$\begin{aligned} \hat{\mu} - \mu_0 &= \sum_{i=1}^{n_1} \hat{p}_i(Y_i - \mu_0) \\ &= \frac{1}{n_1} \sum_{i=1}^n \frac{1}{\hat{\theta}^{-1} \pi(X_i, \hat{\gamma}) + \hat{\lambda} \{m(X_i, \hat{\beta}) - \hat{m}(\hat{\beta})\}} D_i(Y_i - \mu_0) \\ &= \frac{1}{n_1} \sum_{i=1}^n \frac{\theta_0 D_i(Y_i - \mu_0)}{\pi(X_i, \gamma_0)} \\ &\quad + \frac{1}{n_1} \sum_{i=1}^n \frac{D_i(Y_i - \mu_0)}{\pi(X_i, \gamma_0)} (\hat{\theta} - \theta_0) \\ &\quad - \frac{1}{n_1} \sum_{i=1}^n \frac{\theta_0^2 D_i(Y_i - \mu_0)}{\pi^2(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - \hat{m}(\beta_0^*)\} \hat{\lambda} \\ &\quad - \frac{1}{n_1} \sum_{i=1}^n \frac{\theta_0 D_i(Y_i - \mu_0) \pi_1^T(X_i, \gamma_0)}{\pi^2(X_i, \gamma_0)} (\hat{\gamma} - \gamma_0) \\ &\quad + O_p(n^{-1}) \\ &= \frac{1}{n_1} \sum_{i=1}^n \left[\frac{D_i(Y_i - \mu_0)}{\pi(X_i, \gamma_0)} - GB^{-1} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\} \right] \\ &\quad - E \left[\frac{(Y_i - \mu_0) \pi_1^T(X, \gamma_0)}{\pi(X, \gamma_0)} - GB^{-1} \frac{\pi_1^T(X, \gamma_0)}{\pi(X, \gamma_0)} \{m(X, \beta_0^*) - m_0^*\} \right] (\hat{\gamma} \\ &\quad - \gamma_0) + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n [H_i - E(HA^T) \{E(AA^T)\}^{-1} A_i] \\ &\quad + o_p(n^{-1/2}) \end{aligned}$$

where

$$\begin{aligned} H_i &= \frac{D_i(Y_i - \mu_0)}{\pi(X_i, \gamma_0)} - GB^{-1} \frac{D_i - \pi(X_i, \gamma_0)}{\pi(X_i, \gamma_0)} \{m(X_i, \beta_0^*) - m_0^*\}, \\ A_i &= \frac{\{D_i - \pi(X_i, \gamma_0)\} \pi_1(X_i, \gamma_0)}{\pi(X_i, \gamma_0) \{1 - \pi(X_i, \gamma_0)\}}, \quad i = 1, \dots, n. \end{aligned}$$

Applying the central limit theorem, we obtain that $n^{1/2}(\hat{\mu} - \mu_0) \rightarrow N(0, \text{Var}(K))$ in distribution, where $K = H - E(HA^T) \{E(AA^T)\}^{-1} A$. In addition, by using the law of large number, $\hat{\mu} - \mu_0$ converge to 0 in probability, which proves the consistency of $\hat{\mu}$. The proof of Theorem 2 is complete.

Figure 1.

Histograms with kernel density curves of $\bar{X}$ (left) and $\overline{\exp(X)}$ (right) in observed and missing groups based on 5000 Monte Carlo simulations with $n = 500$

Figure 2. Estimates of $\bar{Y}$, $\hat{\mu}_{CC}$, $\hat{\mu}_{REG}$, $\hat{\mu}_{HT}$, $\hat{\mu}_{IPW}$, $\hat{\mu}_{AIPW}$, $\hat{\mu}_{EL}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations with sample size 300. Missing rate is about 69%

Figure 3.

Estimates of $\bar{Y}$, $\hat{\mu}_{CC}$, $\hat{\mu}_{REG}$, $\hat{\mu}_{HT}$, $\hat{\mu}_{IPW}$, $\hat{\mu}_{AIPW}$, $\hat{\mu}_{EL}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations with sample size 500. Missing rate is about 69%

Figure 4.

Histograms with kernel density curves of $\bar{X}_1$ (left) and $\bar{X}_2^2$ (right) in treated and control groups based on 5000 Monte Carlo simulations with $n = 500$

Figure 5.

Estimates of Δ , Δ_{CC} , Δ_{REG} , Δ_{HT} , Δ_{IPW} , Δ_{AIPW} , Δ_{EL} , and Δ based on 5000 Monte Carlo simulations with sample size 300. Average treatment rate is about 23%

Figure 6. Estimates of $\hat{\Delta}$, $\hat{\Delta}_{CC}$, $\hat{\Delta}_{REG}$, $\hat{\Delta}_{HT}$, $\hat{\Delta}_{IPW}$, $\hat{\Delta}_{AIPW}$, $\hat{\Delta}_{EL}$, and $\hat{\Delta}$ based on 5000 Monte Carlo simulations with sample size 500. Average treatment rate is about 23%

Figure 7.
Distributions of age in smokers and non-smokers (mean $\pm$ standard deviation)

Table 1

Biases and root-mean-squared errors (RMSE) of $\bar{Y}$, $\hat{\mu}_{CC}$, $\hat{\mu}_{REG}$, $\hat{\mu}_{HT}$, $\hat{\mu}_{IPW}$, $\hat{\mu}_{EL}$, and $\hat{\mu}$ based on 5000 Monte Carlo simulations. Missing rate is about 69%

Estimator	$n = 300$			$n = 500$			$n = 300$			$n = 500$		
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
<i>(a) Both correctly modeled</i>												
$\bar{Y}$	0.001	0.405	-0.000	0.317	0.003	0.412	0.000	0.321	0.000	0.321	0.000	0.321
$\hat{\mu}_{CC}$	-1.468	1.595	-1.483	1.560	-1.465	1.588	-1.476	1.550	-1.476	1.550	-1.476	1.550
$\hat{\mu}_{REG}$	0.007	0.531	-0.007	0.415	0.123	0.629	0.109	0.490	0.109	0.490	0.109	0.490
$\hat{\mu}_{HT}$	-0.022	0.718	-0.021	0.524	-0.020	0.722	-0.016	0.524	-0.016	0.524	-0.016	0.524
$\hat{\mu}_{IPW}$	-0.020	0.644	-0.019	0.484	-0.020	0.646	-0.016	0.484	-0.016	0.484	-0.016	0.484
$\hat{\mu}_{AIPW}$	0.005	0.549	-0.008	0.429	0.042	0.636	0.019	0.488	0.019	0.488	0.019	0.488
$\hat{\mu}_{EL}$	0.005	0.552	-0.008	0.430	0.020	0.668	-0.005	0.475	-0.005	0.475	-0.005	0.475
$\hat{\mu}$	0.005	0.547	-0.008	0.428	0.033	0.602	0.014	0.463	0.014	0.463	0.014	0.463
<i>(c) $m(x)$ correctly modeled</i>												
$\bar{Y}$	0.007	0.409	0.002	0.322	0.003	0.406	0.006	0.316	0.003	0.316	0.006	0.316
$\hat{\mu}_{CC}$	-1.449	1.575	-1.460	1.539	-1.457	1.581	-1.473	1.550	-1.473	1.550	-1.473	1.550
$\hat{\mu}_{REG}$	0.007	0.539	0.003	0.417	0.127	0.639	0.111	0.487	0.111	0.487	0.111	0.487
$\hat{\mu}_{HT}$	-0.122	0.658	-0.130	0.503	-0.136	0.652	-0.132	0.500	-0.132	0.500	-0.132	0.500
$\hat{\mu}_{IPW}$	-0.122	0.647	-0.130	0.498	-0.136	0.640	-0.132	0.496	-0.132	0.496	-0.132	0.496
$\hat{\mu}_{AIPW}$	0.007	0.551	0.003	0.423	0.303	0.729	0.297	0.578	0.297	0.578	0.297	0.578
$\hat{\mu}_{EL}$	0.005	0.557	0.002	0.427	0.068	0.587	0.049	0.441	0.049	0.441	0.049	0.441
$\hat{\mu}$	0.006	0.553	0.003	0.425	0.235	0.676	0.226	0.522	0.226	0.522	0.226	0.522
<i>(d) Both incorrectly modeled</i>												

$\bar{Y}$, $\hat{\mu}_{CC}$, $\hat{\mu}_{REG}$, $\hat{\mu}_{HT}$, $\hat{\mu}_{IPW}$, $\hat{\mu}_{EL}$ are corresponding to full data sample mean, complete-case estimator, regression estimator, Horvitz-Thompson estimator, inverse probability weighting estimator, augmented inverse probability weighting estimator, and empirical-likelihood-based estimator, respectively, defined in Section 4.1; $\hat{\mu}$ is corresponding to our proposed estimator, defined in Section 2.1.

Table 2

Biases and root-mean-squared errors (RMSE) of $\hat{\pi}$, $\hat{\pi}_{CC}$, $\hat{\pi}_{REG}$, $\hat{\pi}_{HT}$, $\hat{\pi}_{IPW}$, $\hat{\pi}_{AIPW}$, $\hat{\pi}_{EL}$, and $\hat{\pi}$ based on 5000 Monte Carlo simulations. Average treatment rate is about 23%

Estimator	$n = 300$			$n = 500$			$n = 300$			$n = 500$		
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
<i>(a) All correctly modeled</i>												
$\hat{\pi}$	0.000	0.151	-0.001	0.117	-0.003	0.154	-0.001	0.120				
$\hat{\pi}_{CC}$	1.201	1.272	1.191	1.234	1.191	1.263	1.191	1.234				
$\hat{\pi}_{REG}$	0.003	0.194	-0.001	0.150	0.110	0.281	0.111	0.227				
$\hat{\pi}_{HT}$	0.006	0.402	-0.001	0.288	-0.005	0.384	0.000	0.282				
$\hat{\pi}_{IPW}$	0.022	0.399	0.009	0.297	0.012	0.389	0.011	0.297				
$\hat{\pi}_{AIPW}$	0.002	0.206	-0.001	0.157	-0.002	0.270	-0.001	0.209				
$\hat{\pi}_{EL}$	0.003	0.204	-0.001	0.155	-0.007	0.323	-0.000	0.237				
$\hat{\pi}$	0.003	0.201	-0.001	0.154	-0.001	0.265	-0.001	0.205				
<i>(c) $m(x)$ correctly modeled</i>												
$\hat{\pi}$	-0.004	0.152	0.001	0.118	-0.000	0.152	-0.000	0.118				
$\hat{\pi}_{CC}$	1.200	1.271	1.201	1.244	1.191	1.262	1.193	1.236				
$\hat{\pi}_{REG}$	-0.005	0.193	-0.000	0.151	0.112	0.276	0.114	0.225				
$\hat{\pi}_{HT}$	0.417	0.621	0.418	0.593	0.423	0.542	0.410	0.572				
$\hat{\pi}_{IPW}$	0.467	0.570	0.472	0.541	0.468	0.570	0.464	0.536				
$\hat{\pi}_{AIPW}$	-0.004	0.207	0.000	0.157	-0.009	0.281	-0.031	0.431				
$\hat{\pi}_{EL}$	-0.005	0.198	0.000	0.153	0.022	0.252	0.023	0.198				
$\hat{\pi}$	-0.005	0.195	0.000	0.152	0.009	0.231	0.002	0.188				
<i>(d) All incorrectly modeled</i>												

$\hat{\pi}$, $\hat{\pi}_{CC}$, $\hat{\pi}_{REG}$, $\hat{\pi}_{HT}$, $\hat{\pi}_{IPW}$, $\hat{\pi}_{AIPW}$, and $\hat{\pi}_{EL}$ are corresponding to full data sample mean difference, complete-case estimator, regression estimator, Horvitz-Thompson estimator, inverse probability weighting estimator, augmented inverse probability weighting estimator, and empirical-likelihood-based estimator, respectively, defined in Section 4.2; $\hat{\pi}$ is corresponding to our proposed estimator, defined in Section 2.2.

Average treatment effect (ATE) of smoking on renal function of patients with ARAS measured by cystatin C (mg/L) and CKD-EPI GFR (ml/min per 1.73 m²). ATE, standard error (SE), 95% confidence interval and P-value of $\hat{\alpha}_{CC}$, $\hat{\alpha}_{REG}$, $\hat{\alpha}_{HT}$, $\hat{\alpha}_{IPW}$, $\hat{\alpha}_{AIPW}$, $\hat{\alpha}_{EL}$, $\hat{\alpha}$

Table 3

Estimator	Cystatin C				CKD-EPI GFR			
	ATE	SE	95% CI	P-value	ATE	SE	95% CI	P-value
$\hat{\alpha}_{CC}$	-0.007	0.036	[-0.078, 0.064]	0.842	5.68	1.82	[2.11, 9.25]	0.002
$\hat{\alpha}_{REG}$	0.149	0.043	[0.064, 0.234]	0.001	-6.08	1.92	[-9.83, -2.32]	0.002
$\hat{\alpha}_{HT}$	0.087	0.062	[-0.035, 0.209]	0.062	-4.66	2.69	[-9.92, 0.60]	0.083
$\hat{\alpha}_{IPW}$	0.117	0.045	[0.029, 0.204]	0.045	-3.53	1.95	[-7.36, 0.29]	0.070
$\hat{\alpha}_{AIPW}$	0.140	0.044	[0.053, 0.227]	0.002	-5.24	1.95	[-9.06, -1.42]	0.007
$\hat{\alpha}_{EL}$	0.138	0.044	[0.052, 0.224]	0.002	-4.47	1.92	[-8.22, -0.71]	0.020
$\hat{\alpha}$	0.140	0.044	[0.054, 0.227]	0.001	-4.94	1.92	[-8.71, -1.17]	0.010