Harvard University

Harvard University Biostatistics Working Paper Series

Year 2015 *Paper* 191

Doubly Robust Estimation of a Marginal Average Effect of Treatment on the Treated With an Instrumental Variable

Lan Liu[*] Wang Miao[†] Baoluo Sun[‡]

James M. Robins[**] Eric J. Tchetgen Tchetgen[††]

[*] Harvard University, lanliu@hsph.harvard.edu

[†] Peking University, mwfy@pku.edu.cn

[‡] Harvard University, baoluosun@fas.harvard.edu

[**] Harvard University, robins@hsph.harvard.edu

[††] Harvard University, etchetge@hsph.harvard.edu

Doubly Robust Estimation of a Marginal Average Effect of Treatment on the Treated With an Instrumental Variable

Lan Liu, Wang Miao, Baoluo Sun,

James Robins, and Eric Tchetgen Tchetgen

Author's Footnote:

Lan Liu is Postdoctoral fellow (Email: lanliu@hsph.harvard.edu) in the department of Biostatistics at Harvard University. Wang Miao is Ph.D. Candidate (Email: mwfy@pku.edu.cn) in Beijing International Center for Mathematical Research at Peking University. Baoluo Sun is Ph.D. Candidate in the department of Biostatistics of Harvard University. James Robins (Email: robins@hsph.harvard.edu) and Eric Tchetgen Tchetgen (Email: etchetge@hsph.harvard.edu) are Professor in the department of Biostatistics at Harvard University. The authors were partially supported by NIH grants R01 AI032475, R21 AI113251, R01 ES020337, R01 AI104459. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The authors thank Professor Andrea Rotnitzky for helpful discussion. Wang Miao's research was supported by the China Scholarship Council.

Abstract

The objective of many studies in health and social sciences is to evaluate the causal effect of a treatment or exposure on a specific outcome using observational data. In such studies, the exposure is typically not randomized and therefore confounding bias can rarely be ruled out with certainty. The instrumental variable (IV) design plays the role of a quasi-experimental handle since the IV is associated with the treatment and only affects the outcome through the treatment. A valid IV can be used to obtain a test of the null hypothesis of no treatment effect with nominal type 1 error rate. Beyond testing for the causal null, one may wish to obtain an accurate estimate of the treatment causal effect. In this paper, we present a novel framework for identification and estimation using an IV of the marginal average causal effect of treatment amongst the treated (ETT) in the presence of unmeasured confounding. For inference, we propose three different semiparametric strategies: (i) inverse probability weighting (IPW), (ii) outcome regression, and (iii) doubly robust (DR) estimation which is consistent if either (i) or (ii) is consistent, but not necessarily both. An extensive simulation study is carried out to investigate the finite sample performance of the proposed estimators. The methods are further illustrated in a well known application of the impact of participation in a 401(k) retirement programs on savings.

Keywords: Counterfactuals; Double robustness; Instrumental variable; Unmeasured confounding; Effect of treatment on the treated.

1. INTRODUCTION

A major interest of social and epidemiology studies lies in evaluating the effects of a treatment or exposure. For practical reasons, the average treatment effect among treated individuals (ETT) is sometimes of greater interest than the treatment effect in the population. For example, in epidemiology studies concerning the toxic effects of a new drug or in sociology studies evaluating the effects of a policy among those who the policy is applied to, then ETT is the parameter of interest. In observational or randomized studies with non-compliance, a primary challenge is the presence of unobserved confounding, i.e. that treatment groups may differ for reasons other than treatment that may be related to the outcome. Thus, a comparison of outcomes between treatment groups may not only reflect the treatment effect, but also differences due to the process of treatment selection. Instrumental variables (IV) are useful in addressing unmeasured confounding.

An IV is a variable that is associated with the treatment and that affects the outcome only through the treatment. The key idea of the IV method is to extract exogenous variation in the treatment that is unconfounded with the outcome and to take advantage of this bias-free component to make causal inference about the treatment effect (Okui et al., 2012; Angrist and Krueger, 2001).

The development of the IV approach can be traced back to Wright (1928) and Goldberger (1972) under linear structural equations in econometrics. Imbens and Angrist (1994), Angrist et al. (1996) and Heckman (1997) formalized the IV approach within the framework of potential outcomes or counterfactuals. Imbens and Angrist (1994) and Angrist et al. (1996) defined the effect of treatment on individuals who would comply to their assigned treatment. Under a monotonicity assumption about the effect of the IV on exposure, the complier average treatment effect can be identified. Further research along these lines include fully parametric estimation strategies (Tan, 2006; Barnard et al., 2003; Frangakis et al., 2004) as well as semiparametric methods (Abadie, 2003; Abadie et al., 2002; Tan, 2006; Ogburn et al., 2014).

Alternatively, Robins (1989) and Robins (1994) evaluated the ETT conditional on the IV and observed covariates under additive and multiplicative structural nested models (SNMs). Identification is achieved by assuming a certain degree of homogeneity with regard to the IV in an SNM of a conditional ETT (Hernán and Robins, 2006). Mainly, the assumption states that the magnitude of the ETT does not vary with the IV. This is also referred to as the no current treatment value interaction assumption. A major advantage of this identification strategy is that it is guaranteed to hold under the null hypothesis of no causal effect. Similar identification results are developed by Joffe and Brensinger (2003) in the context of a structural distribution model. Vansteelandt and Goetghebeur (2003), Robins and Rotnitzky (2004), Tan (2010), Clarke et al. (2014) and Matsouaka and Tchetgen Tchetgen (2014) investigated estimation of this conditional causal effect under a similar identifying assumption using additive, multiplicative and logistic SNMs.

The interpretation and identification conditions for the complier effect and the ETT are somewhat distinct and each have their own appeal and limitations. On the one hand, the so-called compliers are themselves not individually identified since only one out of the two potential treatments defining compliers can be observed. Also, the definition of "compliers" is instrument-dependent (Pearl, 2011). However, the monotonicity assumption only places restrictions on the effect of the IV on the treatment but not on the treatment effect on the outcome. On the other hand, the population of interest is easy to target for the ETT. However, the literature mentioned above identifies ETT by specifying the functional form of the treatment causal effect. This is unattractive since it places constraints directly on the main parameter of inference. To address these limitations, Tchetgen Tchetgen and Vansteelandt (2013) developed new nonparametric identification results of ETT under an alternative assumption that restricts the degrees of confounding bias across IV values while leaving the causal effect unrestricted. Therefore, the approach of Tchetgen Tchetgen and Vansteelandt (2013) may be particularly valuable to obtain an accurate estimate of the causal effect, especially when an ITT analysis rejects the null hypothesis of no causal effect. However, it may be sensitive to model misspecification of the selection bias function. Tchetgen Tchetgen and Vansteelandt (2013) developed a variety of semiparametric estimators of the conditional treatment effect on the treated as a function of the IV and covariates on the additive scale.

Our work is developed in the same vein by allowing the causal effect to remain unrestricted. Therefore, similar to the approach of Tchetgen Tchetgen and Vansteelandt (2013), our methods will be particularly valuable when the primary goal is to obtain an accurate estimate of the treatment effect. However, we draw a number of specific distinctions between our approach and that of Tchetgen Tchetgen and Vansteelandt (2013). The first distinction lies in the parameter of interest. Although IV methods for the average treatment effect among treated individuals in an IV- and covariate-specific subpopulation are well established, IV methods that directly target the marginal average treatment effect amongst all treated individuals have received little attention. Due to high dimensionality of covariates, in order to obtain a marginal causal effect, Tchetgen Tchetgen and Vansteelandt (2013) must first decide a functional form for the conditional ETT before marginalizing over covariates and IV. Clearly the validity of such an estimator is subject to misspecification of such functional form for the conditional ETT (Tan, 2010). Furthermore, the marginal effect for the treated may be particularly relevant for policy making since it applies to the entire treated subpopulation instead of just a subgroup of the individuals with specific covariate values. Here we propose a novel approach which sidesteps a conditional ETT model and directly targets the marginal ETT. The second distinction lies in the identification assumption and its scale dependency. Tchetgen Tchetgen and Vansteelandt (2013) assumed the absence of an interaction between the IV and the potential outcome in the treatment selection bias function on the additive scale. This approach is only applicable for a continuous outcome and may be overly restrictive about the structure of selection bias due to confounding. Here, we propose a new identification strategy which is applicable for any type of outcome, and provides necessary and sufficient global identification conditions instead of sufficient local identification conditions. The last distinction lies in the specific estimators for ETT. Although Tchetgen Tchetgen and Vansteelandt (2013) proposed a generalized version of G-estimation, regression approach and doubly robust (DR) estimator, they required a correct model for the treatment propensity score conditional on IV and covariates in all three strategies. We circumvent the dependence of the regression estimator on the propensity score and replace it with an assumption of correct specification of a selection bias model.

The outline for the paper is as follows. In Section 2, we introduce the notation and state the main assumptions. Nonparametric identification of ETT is studied in Section 3. We introduce inverse probability weighting (IPW), regression based estimators as well as DR estimators in Section 4. The performance of various estimators is assessed in a simulation study in Section 5. In Section 6, the methods are further illustrated with a study concerning the impact of participation in a 401(k) retirement programs on savings. We conclude with a brief discussion in Section 7.

2. PRELIMINARY RESULTS

Suppose that we observe independently and identically distributed data $O = (A, Y, Z, C)$, where A is a binary treatment, Y is the observed outcome of interest and (Z, C) are observed pre-exposure variables. Let Y_a denote the counterfactual outcome under treatment a for $a = 0, 1$. We make the consistency assumption $Y = AY_1 + (1 - A)Y_0$ almost surely. The marginal effect of treatment on the treated is ETT $= E(Y_1 - Y_0|A = 1)$. Since $E(Y_1|A = 1) = E(Y|A = 1)$ can be consistently estimated from the observed average outcome of treated individuals, throughout, we may focus on making inferences about ψ where

$$\psi = E(Y_0|A = 1).$$

Suppose there exists unmeasured variables denoted by U such that controlling for (U, Z, C) suffices to control for confounding, i.e. $Y_0 \perp\!\!\!\perp A|(Z, C, U)$, however,

$$Y_0 \not\perp\!\!\!\perp A|(Z, C), \tag{1}$$

where $\perp\!\!\!\perp$ denotes statistical independence. As pointed out by Robins et al. (2000), counterfactual outcomes can be viewed as the ultimate unmeasured confounder. This is because by the consistency assumption, the observed outcome Y is a deterministic function of the treatment and the counterfactuals. Thus, given (Y_0, Y_1), U does not contain any further information about Y. To make explicit use of (1), we define the propensity score $\pi(Y_0, Z, C) = \Pr(A = 1|Y_0, Z, C)$ as a function of Y_0.

Thus, the magnitude of confounding can be encoded by the associational measure on a scale defined by the link function κ, such as logit or probit link. We define $\alpha(Y_0, Z, C) = \kappa\{\Pr(A = 1|Y_0, Z, C)\} - \kappa\{\Pr(A = 1|Y_0 = 0, Z, C)\}$, where $Y_0 = 0$ is a reference value Y_0 can take such that $\alpha(0, Z, C) = 0$. With such a definition, $\alpha(Y_0, Z, C)$ encodes on the κ scale, an association between the exposure-free potential outcome and the probability of being exposed within levels of Z and C. Throughout, we refer to α as the treatment selection bias function. Alternatively, one can define $\tilde{\alpha}(A, Z, C)$ as an association measure on the λ scale between the observed treatment and the mean exposure free outcome conditional on Z and C, that is $\tilde{\alpha}(A, Z, C) = \lambda\{E(Y_0|A, Z, C)\} - \lambda\{E(Y_0|A = 0, Z, C)\}$ where λ is any link function, which shall be referred to as an outcome selection bias function. Note $\alpha(Y_0, Z, C) = \tilde{\alpha}(A, Z, C) = 0$ recovers the situation of no unmeasured confounding. Tchetgen Tchetgen and Vansteelandt (2013) previously considered the outcome selection bias function on the additive scale, i.e. λ is the identity link $\lambda(x) = x$.

Note that these two definitions of a selection bias function coincide when the outcome is binary and κ and λ are both the logit link. To be more specific, $\alpha(Y_0, Z, C) = \log OR(A = 1, Y_0|Z, C)$ where we let

$$OR(X_1, X_2|X_3) = \frac{f(X_1, X_2|X_3)f(X_1 = x_{10}, X_2 = x_{20}|X_3)}{f(X_1 = x_{10}, X_2|X_3)f(X_1, X_2 = x_{20}|X_3)},$$

denote the odds ratio between X_1 and X_2 given X_3. The values, x_{10} and x_{20} are baseline values that X_1 and X_2 can take. Throughout the paper, we let f (sometimes with subscript) denote the density function of corresponding random variables. Let Y_{az} denote the potential outcome if A and Z are set to a and z respectively. We formalize the IV assumptions introduced earlier using potential outcomes:

(IV.1) Stochastic exclusion restriction:

$$Y_{az} = Y_a \text{ almost surely for all } a \text{ and } z;$$

(IV.2) Unconfounded IV-outcome relation:

$$f_{Y_0|Z,C}(y|z,c) = f_{Y_0|C}(y|c) \text{ for all } z \text{ and } c;$$

(IV.3) IV relevance:

$$\Pr(A = 1|Z = z, C = c) \neq \Pr(A = 1|Z = 0, C = c) \text{ for all } z \neq 0 \text{ and } c.$$

Assumption (IV.1) states that Z does not have a direct effect on the outcome Y. Assumption (IV.2) is ensured under physical randomization but will hold more generally if one includes all common causes of Z and Y. Assumption (IV.3) states that A and Z have a non-null association conditional on C, even if the association is not causal. If assumption (IV.1)–(IV.3) are satisfied, Z is said to be a valid IV.

3. NONPARAMETRIC IDENTIFICATION

In this section, we first consider the case of binary variables, and note that under the exclusion restriction assumption the IV model is not identifiable in general. Here we give a necessary and sufficient identification condition of the joint distribution of (Y_0, A, Z, C). We also give a sufficient condition of the model which is easier to check in practice. These conditions are further illustrated with examples.

For simplicity, we consider the situation where covariates are omitted. For binary outcome and instrument, one can only identify the quantities $\Pr(Y_0, Z|A = 0)$, $\Pr(Z|A = 1)$, $\Pr(A = 0)$ from the observed data. These quantities are functions of the unknown parameters: $\Pr(Z = 1)$, $\Pr(Y_0 = 1)$, and $\Pr(A = 0|Y_0, Z)$. Without imposing an additional assumption, there are six unknown parameters (one for $\Pr(Z = 1)$, one for $\Pr(Y_0 = 1)$ and four for $\Pr(A = 1|Y_0, Z)$), however, only five degrees of freedom are available from the observed data (one for $\Pr(A = 0)$, one for $\Pr(Z|A = 0)$ and three for $\Pr(Y, Z|A = 0)$). As a result, the parameters are not fully identifiable. Particularly, ψ is not identifiable.

Additional assumptions, such as Robins' no current treatment value interaction assumption or the assumption of Tchetgen Tchetgen and Vansteelandt (2013), must be imposed to reduce the set of candidate models for the joint distribution $f(A, Y_0, Z, C)$. Below, we give a more general sufficient and necessary condition for identification. We restrict the candidates for the joint distribution to a smaller set, which is a subset of all distributions satisfying assumptions (IV.1)–(IV.3). Again allowing for covariates, let $\mathcal{P}_{A|Y_0,Z,C}$ and $\mathcal{P}_{Y_0|C}$ denote the collections of candidates for $\Pr(A=0|Y_0,Z,C)$ and $f(Y_0|C)$.

Condition 1. Any two elements $\Pr_1(A=0|Y_0,Z,C)$, $\Pr_2(A=0|Y_0,Z,C) \in \mathcal{P}_{A|Y_0,Z,C}$ and $f_1(Y_0|C)$, $f_2(Y_0|C) \in \mathcal{P}_{Y_0|C}$, satisfy the inequality:

$$\frac{\Pr_1(A=0|Y_0,Z,C)}{\Pr_2(A=0|Y_0,Z,C)} \neq \frac{f_2(Y_0|C)}{f_1(Y_0|C)}.$$

The following proposition states that condition 1 is a necessary and sufficient condition for identifiability of the joint distribution of (A, Y_0, Z, C).

Proposition 1. Assume (IV.1)–(IV.3), the joint distribution of (A, Y_0, Z, C) is identifiable if and only if condition 1 holds.

It is very convenient to check condition 1 in parametric models, but it may be hard for semiparametric and nonparametric models, since $\mathcal{P}_{A|Y_0,Z,C}$ and $\mathcal{P}_{Y_0|C}$ can be complicated. The following corollary gives a more convenient condition.

Corollary 1. Suppose that for any two candidates $\Pr_1(A=0|Y_0,Z,C)$, $\Pr_2(A=0|Y_0,Z,C) \in \mathcal{P}_{A|Y_0,Z,C}$, $\Pr_1(A=0|Y_0,Z,C)/\Pr_2(A=0|Y_0,Z,C)$ is either a constant or varies with Z. Then the joint distribution of (A, Y_0, Z, C) is identifiable.

Although the condition provided in Corollary 1 is a sufficient condition, it allows identification of a large class of models. We further illustrate Proposition 1 and Corollary 1 with several examples. For simplicity, we omit covariates, however, we note that these can easily be taken into consideration. We first consider the case of binary outcome with binary instrument.

Example 1. We consider $\mathcal{P}_{A|Y_0,Z} = \{\Pr(A=0|Y_0,Z) : \text{logit } \Pr(A=0|Y_0,Z;\theta_1,\theta_2,\eta_1,\eta_2) = \theta_1 + \theta_2 Z + \eta_1 Y_0 + \eta_2 Y_0 Z, \theta_1,\theta_2,\eta_1,\eta_2 \in \mathcal{R}\}$. The model is saturated since $\mathcal{P}_{A|Y_0,Z}$ contains all possible treatment mechanisms. It can be shown that neither the joint distribution nor

ψ is identifiable even under assumption (IV.1)–(IV.3). However, if we assume the following separable treatment mechanism, the joint distribution and thus ψ are both identifiable:

$$\mathcal{P}_{A|Y_0,Z} = \{\Pr(A = 0|Y_0, Z) : \text{logit} \quad \Pr(A = 0|Y_0, Z; \theta_1, \theta_2, \eta_1) = \theta_1 + \theta_2 Z + \eta_1 Y_0; \theta_1, \theta_2, \eta_1 \in \mathcal{R}\}.$$

The model excludes an interaction between Y_0 and Z in the treatment mechanism, and satisfies condition 1. It also agrees with the intuition that we have one less parameter than the saturated model. Under the assumed model, we have five unknown parameters and five available degrees of freedom from the empirical distribution of the observed data.

The IV model with separable treatment mechanism is also identifiable for continuous outcome with continuous instrument.

Example 2. Assume the Logistic separable treatment mechanism: $\mathcal{P}_{A|Y_0,Z} = \{\Pr(A = 0|Y_0, Z) : \text{logit } \Pr(A = 0|Y_0, Z) = q(Z) + h(Y_0)\}$, where q and h are unknown differentiable functions. It can be shown that $\mathcal{P}_{A|Y_0,Z}$ satisfies condition 1 and thus the joint distribution is identifiable under (IV.1)–(IV.3).

The above examples show that the joint density $f(A, Y_0, Z)$ is not identifiable when the treatment selection mechanism is left unrestricted (example 1), but it is identifiable in the submodel of separable selection (example 2). We present in the Supplementary Materials the proofs for the above examples, and additional examples, such as the case of continuous outcome with binary instrument, and a separable treatment mechanism. We also provide a simple data generating mechanism of unmeasured confounding which may be used to motivate the separable model of example 2. The results also readily extend in the presence of covariates C, for instance by allowing both q and h to depend on C in example 2.

4. ESTIMATION

While nonparametric identification conditions are provided in the previous section, such conditions will seldom suffice for reliable statistical inference when, as will typically be the case in observational studies, the set of covariates C is too large for nonparametric inference, due to the curse of dimensionality (Robins and Ritov, 1997). To make progress, we posit parametric models for various nuisance parameters, and provide three possible strategies of semiparametric inference that depend on different subsets of models. In this Section, we describe an IPW estimator, an outcome regression-based estimator and a doubly robust (DR) estimator of ETT under assumptions (IV.1)–(IV.3) and condition 1. Throughout, we assume a model for the selection bias function $\alpha(Y_0, Z, C; \eta)$ is correctly specified, with $\alpha(Y_0 = 0, Z, C; \eta) = \alpha(Y_0, Z, C; 0) = 0$. Also throughout, we posit a parametric model $f_{Z|C}(z|c) = \Pr(Z = z|C = c; \rho)$ for Z. We let $\hat{\rho}$ denote the MLE of ρ. Let $\mathbb{P}_n$ be the empirical measure, that is $\mathbb{P}_n f(O) = n^{-1} \sum_{i=1}^{n} f(O_i)$. Let $\hat{E}$ denote the expectation taken under the empirical distribution of C.

4.1 IPW estimator

For estimation, we first propose an IPW IV approach which extends standard IPW estimation of ETT to a setting with unobserved confounding and an IV. We make the positivity assumption that for all values of Y_0, Z and C the probability of not being treated is bounded away from 0. Let $\beta(Z, C) = \kappa\{\Pr(A = 1|Y_0 = 0, Z, C)\}$ thus $\kappa\{\Pr(A = 1|Y_0, Z, C)\} = \alpha(Y_0, Z, C) + \beta(Z, C)$, whereas before κ is any link function. Suppose we also posit a model $\beta(Z, C; \theta)$ for $\beta(Z, C)$. The propensity score can then be written $\kappa\{\pi(Y_0, Z, C; \gamma)\} = \alpha(Y_0, Z, C; \eta) + \beta(Z, C; \theta)$ where $\gamma = (\eta, \theta)$. The IPW approach relies on the crucial assumption that the propensity score model $\pi(Y_0, Z, C)$ is correctly specified and the following representation of ETT,

$$E(Y_0|A = 1) = E\left\{\frac{\pi(Y, Z, C)Y(1 - A)}{\Pr(A = 1)\{1 - \pi(Y, Z, C)\}}\right\}. \tag{2}$$

We prove the above equation in the Supplementary Materials. Then we solve the following equations to obtain an estimator of γ:

$$\mathbb{P}_n\{\frac{1 - A}{1 - \pi(Y, Z, C; \hat{\gamma})} - 1\} = 0, \tag{3}$$

$$\mathbb{P}_n\Big[\frac{1 - A}{1 - \pi(Y, Z, C; \hat{\gamma})}\{h_1(Z, C) - E(h_1(Z, C)|C; \hat{\rho})\}\Big] = 0, \tag{4}$$

$$\mathbb{P}_n\Big[\frac{1 - A}{1 - \pi(Y, Z, C; \hat{\gamma})}\{h_2(C) - \hat{E}(h_2(C))\}\Big] = 0, \tag{5}$$

$$\mathbb{P}_n\Big[\frac{1 - A}{1 - \pi(Y, Z, C; \hat{\gamma})}t(Y, C)\{l(Z, C) - E(l(Z, C)|C; \hat{\rho})\}\Big] = 0, \tag{6}$$

where $(h_1^T, h_2^T, l^T)^T$ satisfies the regularity condition (A.1) described in the Supplementary Materials. Equations (4) and (5) identify the association between (Z, C) and A, i.e. $\pi(0, Z, C)$, while by the exclusion restriction (IV.1), equation (6) identifies the selection bias function α. By equation (2), a propensity score estimate leads to an estimator for ψ. We have the following result:

Proposition 2. Under (IV.1)-(IV.3) and condition 1, suppose the propensity score model $\pi(Y, Z, C; \gamma)$ and $f_{Z|C}(z|c; \rho)$ are correctly specified, then the IPW estimator

$$\hat{\psi}^{ipw} = \mathbb{P}_n \frac{\pi(Y, Z, C; \hat{\gamma})Y(1-A)}{\hat{\Pr}(A=1)\{1-\pi(Y, Z, C; \hat{\gamma})\}},$$

is consistent.

We emphasize that κ can be any well defined link function (e.g., logit, probit), and Proposition 2 still holds. Condition (A.1) is imposed so that the population expectation value of the derivative of the vector of equations (3)–(6) is invertible when evaluated at the true parameter value. The functions h_1, h_2, t and l can be chosen depending on the model one posits for the propensity score. For example, assuming logit $\pi(Y_0, Z, C) = \theta_0 + \theta_1 Z + \theta_2 C + \eta Y_0$ where the dimension of $\tilde{\eta} = (\theta_1, \theta_2, \eta)$ is k then (h_1, h_2, t) forms a vector of dimension k and can be chosen as $(h_1, h_2, t) = \partial \text{logit } \pi(Y_0, Z, C)/\partial\tilde{\eta} = (Z, C, Y_0)$ and l can be chosen as any function of (Z, C), e.g., $l(Z, C) = Z$. Thus we have exactly $k + 1$ estimating equations. The choice of h_1, h_2, t and l will generally impact efficiency but should not affect consistency as long as the identification conditions hold and model misspecification is absent.

4.2 Regression and doubly robust estimators

Analogous to the propensity score, the outcome regression could in principle be modeled on an arbitrary scale defined by a link function λ, and the selection bias would then be defined for that scale. To see this, let $\delta(Z, C) = \lambda\{E(Y_0|A = 0, Z, C)\}$, then $E(Y_0|A = 1, Z, C) = \lambda^{-1}\{\tilde{\alpha}(1, Z, C) + \delta(Z, C)\}$. Parameters in $\delta(Z, C)$ can be estimated by maximum likelihood estimation (MLE), an estimation equation can be constructed to estimate parameters in $\tilde{\alpha}(1, Z, C)$ and a consistent regression estimator for ψ can thus be constructed (see Supplementary Materials for more details).

However, such parameterization does not permit the construction of a DR estimator unless the outcome is binary and both κ and λ are logit links. Alternatively, we give a representation of $E(Y_0|A = 1, Z, C)$ in terms of $\alpha(Y_0, Z, C)$ and $f(Y|A = 0, Z, C)$ for any type of outcome. Note that

$$\alpha(Y_0, Z, C) = \log \frac{f(Y_0|A = 1, Z, C)f(Y_0 = 0|A = 0, Z, C)}{f(Y_0|A = 0, Z, C)f(Y_0 = 0|A = 1, Z, C)}.$$

Let g be any function of Y_0 and C, we have the following representation

$$E[g(Y_0, C)|A = 1, Z, C] = \frac{E[\exp\{\alpha(Y, Z, C)\}g(Y, C)|A = 0, Z, C]}{E[\exp\{\alpha(Y, Z, C)\}|A = 0, Z, C]}. \tag{7}$$

We prove the equation in the Supplementary Materials.

As before, let η denote the parameter indexing a model for the selection bias function α. For estimation of η, we let $f(Y|A = 0, Z, C; \xi)$ denote a parametric model for the outcome, and let $\hat{\xi}$ denote the MLE of ξ obtained using only data among the unexposed. We obtain an estimator for η by solving:

$$\mathbb{P}_n\left[\Big\{w(Z, C) - E(w(Z, C)|C; \hat{\rho})\Big\}\Big\{AE[g(Y_0, C)|A = 1, Z, C; \eta, \hat{\xi}] + (1 - A)g(Y, C)\Big\}\right] = 0, \tag{8}$$

for any choice of functions w and g such that the regularity condition (A.2) stated in the Supplementary Materials holds. Based on equation (7), one can construct an estimator for ψ based on $\alpha(Y, Z, C; \hat{\eta})$, $\hat{\xi}$ and $\hat{\rho}$.

Proposition 3. Under (IV.1)-(IV.3) and condition 1, suppose $\alpha(Y_0, Z, C; \eta)$, $f_{Z|C}(z|c; \rho)$ and $f(Y|A = 0, Z, C; \xi)$ are correctly specified, then the outcome regression estimator

$$\hat{\psi}^{reg} = \mathbb{P}_n \frac{A}{\hat{\Pr}(A = 1)} \frac{E[\exp\{\alpha(Y, Z, C; \hat{\eta})\}Y|A = 0, Z, C; \hat{\xi}]}{E[\exp\{\alpha(Y, Z, C; \hat{\eta})\}|A = 0, Z, C; \hat{\xi}]},$$

is consistent.

Functions g and ω in equation (8) may be chosen depending on the model we posit for $\alpha(Y_0, Z, C)$. For example, assuming

$$\alpha(Y_0, Z, C; \eta) = \eta Y_0, \tag{9}$$

g can be chosen as $g = \partial\alpha(Y_0, Z, C; \eta)/\partial\eta = Y_0$ and ω can be chosen as any scalar function of (Z, C), e.g., $\omega = Z$. The choice of g and ω may impact efficiency but does not affect consistency as long as the identification conditions hold.

Note that the proposed regression-based estimator is closely related to the regression estimator proposed by Vansteelandt and Goetghebeur (2003) when Y is binary. Thus, equation (8) can be re-expressed as

$$\mathbb{P}_n\bigg\{\big(w(Z,C)-E(w(Z,C)|C;\hat{\rho})\big)\bigg(A\text{expit}\{\delta(Z,C;\hat{\xi})+\alpha(1,Z,C;\eta)\}+(1-A)Y\bigg)\bigg\}=0, \tag{10}$$

where $\text{expit}(x) = \exp(x)/\{1 + \exp(x)\}$. Vansteelandt and Goetghebeur (2003) developed a two-stage logistic estimator which combines a logistic SMM at the first stage and a logistic regression association model at the second stage. Let $\zeta(Z, C) = \text{logit } \Pr(Y_1 = 1|A = 1, Z, C) - \text{logit } \Pr(Y_0 = 1|A = 1, Z, C)$ which encodes the conditional ETT given Z and C. Vansteelandt and Goetghebeur (2003) posited a parametric model for $\delta(Z, C)$ which they aim to estimate with the IV Z. Let $\vartheta(Z, C; \varrho) = \text{logit } \Pr(Y = 1|A = 1, Z, C; \varrho)$, then the estimating equation they proposed to estimate $\zeta(Z, C; \nu)$ can be expressed as

$$\mathbb{P}_n\bigg\{\big(w(Z,C)-E(w(Z,C)|C;\hat{\rho})\big)\bigg(A\text{expit}\{\vartheta(Z,C;\hat{\varrho})-\zeta(Z,C;\nu)\}+(1-A)Y\bigg)\bigg\}=0. \tag{11}$$

Comparing (10) and (11), they mainly differ in the way $\Pr(Y_0 = 1|A = 1, Z, C)$ is estimated. More specifically, (10) obtained $\Pr(Y_0 = 1|A = 1, Z, C)$ using $\Pr(Y_0 = 1|A = 0, Z, C)$ as a baseline risk for the model while (11) uses $\Pr(Y_1 = 1|A = 1, Z, C)$ as baseline risk. This difference is important since Vansteelandt and Goetghebeur (2003) failed to obtain a DR estimator of $\zeta(Z, C)$ while as we show next, out choice of parameterization yields a DR estimator of the marginal ETT.

Heretofore , we have constructed estimators following two possible strategies. Both strategies assume correct models for $\alpha(Y_0, Z, C; \eta)$ and $f_{Z|C}(z|c; \rho)$, however, IPW further relies on a consistent estimator of $\pi(Y_0, Z, C)$ and outcome regression further relies on a consistent estimator of $E(Y_0|A = 1, Z, C)$. Define $\mathcal{M}_a$ as the collection of laws with parametric models $f_{Z|C}(z|c; \rho)$, $\alpha(Y_0, Z, C; \eta)$ and $\beta(Z, C; \theta)$ while $f(Y|A = 0, Z, C)$ is unrestricted. Likewise, define $\mathcal{M}_y$ as the collection of laws with parametric models $f_{Z|C}(z|c; \rho)$, $\alpha(Y_0, Z, C; \eta)$ and $f(Y|A = 0, Z, C; \xi)$ while $\beta(Z, C)$ is unrestricted. The main appeal of a doubly robust estimator is that it remains consistent if either $\beta(Z, C; \theta)$ or $f(Y|A = 0, Z, C; \xi)$ is correctly specified. To derive a DR estimator for ψ in the union space $\mathcal{M}_a \cup \mathcal{M}_y$, we first propose a DR estimator for the selection bias function. For notational convenience, let

$$\begin{aligned} Q_g(Y, A, Z, C; \gamma, \xi) &= \frac{(1 - A)\pi(Y, Z, C; \gamma)}{1 - \pi(Y, Z, C; \gamma)}\Big[g(Y, C) - E\{g(Y_0, C)|A = 1, Z, C; \eta, \xi\}\Big] \\ &\quad +AE\{g(Y_0, C)|A = 1, Z, C; \eta, \xi\}. \end{aligned} \qquad (12)$$

Equation (12) is key to obtaining a DR estimation of the selection bias function and thus of ETT. Specifically, consider the estimating equation for the selection bias parameter $\tilde{\eta}$

$$\mathbb{P}_n\big[\omega(Z, C) - E\{\omega(Z, C)|C; \hat{\rho}\}\big]\Big\{Q_g(Y, A, Z, C; \tilde{\gamma}, \hat{\xi}) + (1 - A)g(Y, C)\Big\} = 0. \qquad (13)$$

We solve equations (3)–(5) and (13) with $\hat{\gamma}$ replaced by $\tilde{\gamma} = (\tilde{\eta}, \tilde{\theta})$. The choice of h_1, h_2, g and w can be decided similarly as in Section 4.1 and 4.2.

Proposition 4. Under (IV.1)-(IV.3) and condition 1, $\tilde{\eta}$ is consistent for η, and $\hat{\psi}^{DR}$ is consistent for ψ in the union model $\mathcal{M}_a \cup \mathcal{M}_y$, where $\hat{\psi}^{DR} = \mathbb{P}_n Q_{\tilde{g}}(Y, A, Z, C; \tilde{\gamma}, \hat{\xi}) / \Pr(A = 1)$ and $\tilde{g}(Y, C) = Y$.

The DR estimator proposed here is closely related to the DR estimator proposed by Vansteelandt et al. (2007) in a missing data context in the sense that equation (12) is a conterfactural version of an analogous DR equation derived in their paper. Vansteelandt et al. (2007) considered identification and estimation in the context of non-ignorable missing data without an IV. The identification condition they investigated requires a priori knowledge of the selection bias function and the DR estimator they proposed is consistent in the submodel of $\mathcal{M}_a \cup \mathcal{M}_y$ where α is assumed to be known. In contrast, with a valid IV, the selection bias can now be identified under the condition from Section 3 and a DR estimator can be obtained in the larger union model $\mathcal{M}_a \cup \mathcal{M}_y$. To the best of our knowledge, equation (13) is new to the literature and so is our DR estimator of ETT.

5. SIMULATIONS

Simulations for both binary and continuous outcomes were conducted to evaluate the finite sample performance of the causal effect estimators derived in Sections 4.1 and 4.2. Simulations were conducted under four scenarios where the parametric models are in: (i) $\mathcal{M}_a \cap \mathcal{M}_y$, (ii) $\mathcal{M}_a \cap \mathcal{M}_y^c$, (iii) $\mathcal{M}_a^c \cap \mathcal{M}_y$ and (iv) $\mathcal{M}_a^c \cap \mathcal{M}_y^c$, where $\mathcal{M}_a^c$ is defined to be the complement space of $\mathcal{M}_a$ and likewise define $\mathcal{M}_y^c$. Thus in (i) both outcome regression and propensity score are correctly specified, in (ii) only the propensity score is correct, in (iii) only the outcome regression model is correct and in (iv) neither model is correct.

Simulations were first carried out only for a binary outcome. For scenario (i), the simulation study was conducted in the following steps:

Step 1: A hypothetical study population of size n was generated and each individual had baseline covariates C_1 and C_2 generated independently from Bernoulli distributions with probability 0.4 and 0.6 respectively. Then the IV was generated from the model: logit $\Pr(Z=1|C) = 0.2+0.4C_1-0.5C_2$ and potential outcomes from models logit $\Pr(Y_0 = 1|Z, C) = 0.6+0.8C_1-2C_2$ and logit $\Pr(Y_1 = 1|Z, C) = 0.7-0.3C_1$. The treatment variable A was generated from logit $\Pr(A = 1|C, Z, Y_0) = 0.4 + 2Z + 0.8C_1 - 0.6Y_0 - 1.6C_1Z$, and finally the observed outcome was $Y = Y_0(1 - A) + Y_1A$.

Step 2: The following propensity score model was fit to the data and the parameters $\gamma = (\theta_1, \theta_2, \theta_3, \theta_4, \eta)$ in model (14) were estimated using estimating equations (3)–(6) with $h_1(Z, C) = (Z, C_1Z)^T$, $h_2(C) = C_1$, $t(Y, C) = Y$ and $l(Z, C) = Z$ and finally $\hat{\psi}^{ipw}$ was calculated.

$$\text{logit } \Pr(A = 1|C, Z, Y_0; \gamma) = \theta_1 + \theta_2 Z + \theta_3 C_1 + \theta_4 C_1 Z + \eta Y_0. \tag{14}$$

Step 3: The selection bias function was correctly specified as (9), the regression outcome model (15) was fit to the exposed and ξ was estimated by MLE, and α was estimated by solving equation (8) with $\omega(Z, C) = Z$ and $g(Y, C) = Y$ and finally $\hat{\psi}^{reg}$ was calculated.

$$\text{logit } E(Y|A = 0, Z, C; \xi) = \xi_1 + \xi_2 C_1 + \xi_3 C_2 + \xi_4 Z + \xi_5 C_1 Z. \tag{15}$$

Step 4: The selection bias function was correctly specified as $\alpha(Y, Z, C; \eta) = \eta Y$, ξ in equation (15) was estimated by MLE, parameters γ in (14) was jointly estimated in the estimating equations (3)–(5) and (13) where h, t, l, ω, g are chosen the same as in Step 2 and Step 3 and finally $\hat{\psi}^{DR}$ was calculated.

Step 5: Steps 1–4 were repeated 1000 times.

Under the data generating mechanism described in Step 1, the exclusion restriction assumption was guaranteed to be satisfied for both $a = 0, 1$. As shown in example 1, ψ is identifiable from the observed data since the treatment mechanism is a separable logit model. Also note that the model for $E(Y|A = 0, C, Z)$ was saturated in C_1 and Z but only contains a linear term for C_2. It is easily verified that this specific model is guaranteed to contain the true data generating mechanism (see the Supplementary Materials). Simulations for scenario (ii) were similar to scenario (i) except that (14) was replaced with a misspecified propensity score model

$$\text{logit } \Pr(A = 1|C, Z, Y_0; \gamma) = \theta_1 + \theta_2 Z + \theta_3 C_1 + \eta Y_0, \tag{16}$$

and $h_2(C) = C_1$, which is misspecified if $\theta_4 \neq 0$ in equation (14). For scenario (iii), the potential outcome model (15) was replaced with

$$\text{logit } E(Y|A = 0, Z, C; \xi) = \xi_1 + \xi_2 C_1 + \xi_4 Z, \tag{17}$$

which was misspecified if $\xi_3 \neq 0$ and $\xi_5 \neq 0$ in equation (15). For scenario (iv), both the propensity score model (14) and the outcome model (15) were replaced with the misspecified models (16) and (17) respectively. Note that fewer covariates were included in the misspecified model but for notational convenience, the parameter subscripts were maintained as in the correctly specified model. The R package BB (Varadhan and Gilbert, 2009) was used to solve the nonlinear estimating equations. The bias, Monte Carlo standard error (MCSE) and average estimated standard error (ASE) of 1000 Monte Carlo simulated samples are reported in Table 1. At the sample size of $n = 5000$, when only the propensity model was misspecified, the IPW estimator had a bias of 0.116 while the outcome regression estimator had negligible bias equal to 0.010. When only the outcome regression model was misspecified, the regression estimator had a bias of 0.085 while the IPW estimator had negligible bias equal to 0.003. The DR estimator provides consistent results when either model was correct with biases of $5e^{-4}$ and 0.006 respectively in the above situation and 0.001 when both models were correctly specified.

Simulations for a continuous outcome were conducted similarly as for the binary outcome in the following steps.

Step 1*: Covariates C_1 and C_2 were generated same as in Step 1 and IV was generated from model logit $\Pr(Z = 1|C) = 0.7 + 0.8C_1 - C_2$, and potential outcomes from models $Y_0|Z, C \sim N(0.5+C_1+3C_2, 1)$ and $Y_1|Z, C \sim N(1.1-1.3C_1, 1)$. The treatment variable A was generated from logit $\Pr(A = 1|C, Z, Y_0) = -0.2 - 3Z - 3C_1 + 0.3Y_0 + 4C_1Z$, and finally the observed outcome was $Y = Y_0(1 - A) + Y_1A$.

Step 2*: Same as Step 2.

Step 3*: Same as Step 3 except the following regression outcome models were fit to the data.

$$E\{Y \exp(\eta Y)|A = 0, Z, C; \xi\} = \xi_1+\xi_2C_1+\xi_3C_2+\xi_4Z+\xi_5C_1Z+\xi_6C_2Z+\xi_7C_1C_2+\xi_8C_1C_2Z. \tag{18}$$

$$E\{\exp(\eta Y)|A = 0, Z, C; \xi\} = \xi_9+\xi_{10}C_1+\xi_{11}C_2+\xi_{12}Z+\xi_{13}C_1Z+\xi_{14}C_2Z+\xi_{15}C_1C_2+\xi_{16}C_1C_2Z. \tag{19}$$

Step 4*: Same as Step 4 except that (15) was replaced by (18) and (19).

Step 5*: Same as Step 5.

Simulation for a continuous outcome under scenario (ii) was carried out similarly as that for scenario (i) except that (14) was replaced by (16). For scenario (iii), the potential outcome models (18) and (19) was replaced with the linear models

$$E\{Y\exp(\eta Y)|A=0,Z,C;\xi\}=\xi_1+\xi_2C_1+\xi_4Z. \tag{20}$$

$$E\{\exp(\eta Y)|A=0,Z,C;\xi\}=\xi_9+\xi_{10}C_1+\xi_{12}Z. \tag{21}$$

For scenario (iv), both the propensity score model (14) and the potential outcome models (18) and (19) were replaced with the misspecified models (16) and (20) and (21) respectively. The R package nleqslv (Hasselman, 2014) was used to solve the nonlinear estimating equations.

Example A.2 of the supplementary materials shows that ψ is identifiable from the observed data since condition 1 is satisfied. The results of 1000 Monte Carlo simulated samples are provided in Table 2. At sample size $n = 1000$, when only the propensity score model was misspecified, the IPW estimator suffered from a bias of 0.806, while the outcome regression estimator had negligible bias equal to 0.009. When the outcome regression model was misspecified, the regression estimator had a bias of 0.738 while the IPW estimator had negligible bias equal to 0.024. The DR estimator was consistent in the union model with bias comparable to the consistent estimator under a given scenario.

6. APPLICATION

Since the 1980s, tax-deferred programs such as individual Retirement Accounts (IRAs) and the 401(k) plan have played an important role as a channel for personal savings in the United States. Aiming to encourage investment for future retirement, the 401(k) plan offers tax deductions on deposits into retirement accounts and tax-free accrual of interest. Moreover, it imposes penalties on early withdrawal of assets from corresponding accounts. The 401(k) plan shares certain similarities with IRAs in that both are deferred compensation plans for wage earners but the 401(k) plan is only provided by employers. The study includes 9275 people and once offered the 401(k) plan, individuals decide whether or not to participate in the program. However, participants usually have a stronger preference for savings which suggests the presence of selection bias. This was addressed as individual heterogeneity by Abadie (2003) and it has been pointed out that a simple comparison of personal savings between participants and non-participants may yield results that were biased upward. It also has been noted that given income, the 401(k) eligibility could be unrelated to the individual preferences for savings thus can be used as an instrument for participation in 401(k) program (Poterba and Venti, 1994; Poterba et al., 1995). The complier causal effect for the 401(k) plan was studied by Abadie (2003). Here, we reanalyze these data to illustrate the proposed estimators of the marginal ETT.

We illustrate the methods in the context of a dichotomous outcome defined as the indicator that a person falls in the lowest quartile of net savings of the observed sample (equal to $-\$500$). The treatment variable is a binary indicator of participation in a 401(k) plan and the IV is a binary indicator of 401(k) eligibility. The covariates are standardized log family income ($\log_{10}(\text{income}) - 4.5$), standardized age ($\text{age} - 41$) and its square, marital status and family size. Age ranged from 25 to 64 years, marital status is binary indicator variable and family size ranges from 1 to 13 people. These covariates are thought to be associated with unobserved preferences for savings. Let $\psi = E(Y_0|A = 1)$ denote for a family that actually participated in the 401(k) program, the probability that they would have had net financial assets above the first quartile, had possibly contrary to fact, they been forced not to participate in the program. The ETT $= E(Y_1 - Y_0|A = 1)$ is the effect of 401(k) plan on the difference scale for the probability of family net financial assets above the lowest quartile among participants. Equivalently, ETT can also be interpreted as an effect of the intervention in reducing a person's risk for poor savings performance as measured by falling below the first quartile of the empirical distribution of savings for the sample. Before implementing our IV estimators, we first obtained a standard IPW estimator of the ETT under an assumption of no unobserved confounding, i.e. $\hat{\psi}_0^{ipw}$ defined as $\hat{\psi}^{ipw}$ with $\alpha = 0$. Thus, the propensity score was modeled as:

$$\text{logit } \Pr(A = 1|Z, C) = 1 + Z + \log(\text{income}) + \text{married} + \text{age} + \text{fsize} + \text{age}^2,$$

and estimated by standard maximum likelihood. Note the IPW estimate of ψ was $\hat{\psi}_0^{ipw} = 0.688$ (se = 0.014), the standard error (se) was calculated using the sandwich estimator accounting for all sources of variabilities. In comparison, the estimator based on the empirical estimate of $E(Y|A=1)$ was 0.883 (se = 0.006). Thus an estimate of ETT was $\widehat{\text{ETT}} = 0.194$ (se = 0.016), which suggests the 401(k) plan may have a significant effect on increasing the family net financial assets among participants.

However, this result may be spurious due to the suspicion that even after controlling for observed covariates, there may be still be unmeasured factors that confound the relationship between 401(k) plan and the family net financial assets. Assuming as in Abadie (2003) the IV satisfies assumptions (IV.1)–(IV.3), we applied the methods proposed in Section 4 to estimate the ETT in the presence of unmeasured confounders. The following parametric models were considered:

$$\text{logit } \Pr(Z=1|C) = 1 + \log(\text{income}) + \text{married} + \text{age} + \text{fsize} + \text{age}^2,$$

$$\text{logit } \Pr(Y=1|A=0,Z,C) = 1 + Z + \log(\text{income}) + \text{married} + \text{age} + \text{fsize} + \text{age}^2,$$

We further assumed condition 1 and specified the selection bias function as in (9). Thus the selection bias function was assumed to depend on Y_0 linearly, which is reasonable if the odds ratio function relating a person's underlying preference U and Y does not depend on Z and the residual of U on $(A=0,Y,Z,C)$ is independent of (Y,Z), and there is no interaction between U and Z in the propensity score $\Pr(A=1|U,Z,C)$ (see example A.1 in the Supplementary Materials). We posited the following parametric model for propensity score which satisfies identifying condition 1 as a submodel of the separable model:

$$\text{logit } \Pr(A=1|Y_0,Z,C) = 1 + Z + Y_0 + \log(\text{income}) + \text{married} + \text{age} + \text{fsize} + \text{age}^2,$$

Table 3 reports the point estimates and estimated standard errors for the IV, propensity score and the outcome regression models. Note that although the DR estimator involves both propensity score and the outcome regression models, the outcome regression model is the same as required for the regression estimator thus these estimates are only repeated once.

The instrument is strongly associated with family income ($\log \mathrm{OR} = 2.823$, se $= 0.106$), age ($\log \mathrm{OR} = 0.007$, se $= 0.002$) and age square ($\log \mathrm{OR} = -0.002$, se $= 2e^{-4}$) but not strongly correlated with other covariates. The IV is strongly related with the outcome and is significant in the propensity score models for both IPW and DR estimators. All three estimators agree with each other. The selection bias parameter was estimated to be 0.320 (se = 0.115) by IPW, 0.385 (se = 0.135) by outcome regression and 0.280 (se = 0.101) by DR estimation. This provided strong evidence that unmeasured confounding may be present and the stronger saving preference one has, the more likely one would be to participate in the 401(k) plan. The ETT is still significant across all three estimators but with a smaller Z-score value than when the selection bias is ignored: for example, the IPW estimator suggests $\widehat{\mathrm{ETT}} = 0.132$ (se = 0.013). Thus we may conclude that even after adjustment for unobserved preferences for savings, the 401(k) plan still has a significant effect on net financial assets among participants.

These findings roughly agree with results obtained by Abadie in the sense that the IV estimate corrects the observational estimate towards the null although it may be difficult to directly compare our findings to those of Abadie who reported effect estimates for the average effect of the intervention only among the compliers under a monotonicity assumption of the IV-exposure relationship, and assuming no unobserved confounding of this first stage relation. The proposed approach relies on neither assumption, but instead relies for identification on a condition 1 encoded in the functional form of the propensity score model used for the analysis. In order to assess the robustness of the selection bias model, additional functional forms were explored. We considered adding to α an interaction between Y_0 and each of the covariates: log income, marriage status, family size. There was no evidence in favor of any such interaction.

7. DISCUSSION

In this paper, we establish that access to an IV allows for identification of an association between exposure to the treatment and the potential outcome when unexposed, which directly encodes the magnitude of selection bias into treatment due to confounding. We propose IPW, outcome regression as well as DR estimators for the treatment effect amongst treated individuals. Unlike Robins' SNMs, our proposed estimators will fail to be consistent when condition 1 fails even under the null hypothesis of no ETT. Therefore, the identification and inference approaches we have proposed may be particularly valuable when an ITT analysis indicates a non-null treatment effect and thus Robins' identification assumption of no current treatment value interaction may be violated.

The proposed methods assume the treatment is binary. They can be generalized without much effort to categorical treatment. However, when the treatment is continuous (for example, A is treatment dose), then a parametric model for the treatment effect as well as a model for the density of A may be unavoidable for estimation. We leave this as a topic for future research.

SUPPLEMENTARY MATERIALS

Appendix A contains proofs of the propositions. Appendix B presents proofs of the examples in the main text, and more examples about identification of the models. Appendix C presents more derivations mentioned in the main text.

REFERENCES

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113:231–263.

Abadie, A., Angrist, J., and Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70:91–117.

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91:444–455.

Angrist, J. and Krueger, A. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives*, 15:69–85.

Barnard, J., Frangakis, C., Hill, J., and Rubin, D. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the American Statistical Association*, 98:299–323.

Clarke, P., Palmer, T., and Windmeijer, F. (2014). Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Technical report*.

Frangakis, C., Brookmeyer, R., Varadhan, R., Safaeian, M., Vlahov, D., and Strathdee, S. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, 99:239–249.

Goldberger, A. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.

Hasselman, B. (2014). *nleqslv: Solve systems of non linear equations*. R package version 2.1.1.

Heckman, J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, 32:441–62.

Hernán, M. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17:360–372.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica: Journal of the Econometric Society*, pages 467–475.

Joffe, M. and Brensinger, C. (2003). Weighting in instrumental variables and g-estimation. *Statistics in medicine*, 22:1285–1303.

Matsouaka, R. A. and Tchetgen Tchetgen, E. J. (2014). Likelihood based estimation of logistic structural nested mean models with an instrumental variable.

Ogburn, E., Rotnitzky, A., and Robins, J. (2014). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society, in press.*

Okui, R., Small, D., Tan, Z., and Robins, J. (2012). Doubly robust instrumental variable regression. *Statistica Sinica*, 22:173–205.

Pearl, J. (2011). Principal stratificationa goal or a tool? *The International Journal of Biostatistics*, 7.

Poterba, J. and Venti, S. (1994). 401 (k) plans and tax-deferred saving. In *Studies in the Economics of Aging*, pages 105–142. University of Chicago Press.

Poterba, J., Venti, S., and Wise, D. (1995). Do 401 (k) contributions crowd out other personal saving? *Journal of Public Economics*, 58:1–32.

Robins, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113:159.

Robins, J. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23:2379–2412.

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

Robins, J. and Rotnitzky, A. (2004). Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91:763–783.

Robins, J., Rotnitzky, A., and Scharfstein, D. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, volume 116, pages 1–94. Springer.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101:1607–1618.

Tan, Z. (2010). Marginal and nested structural models using instrumental variables. *Journal of the American Statistical Association*, 105:157–169.

Tchetgen Tchetgen, E. and Vansteelandt, S. (2013). Alternative identification and inference of the effect of treatment on the treated with an instrumental variable. *Harvard University Biostatistics Working Paper Series*.

Vansteelandt, S. and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:817–835.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94:841–860.

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

Wright, S. (1928). Appendix to the tariff on animal and vegetable oils. *New York: MacMillan.(1934)," The Method of Path Coefficients," Annals of Mathematical Statistics*, 5:161–215.

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

Wu, M. and Follmann, D. (1999). Use of summary measures to adjust for informative missingness in repeated measures data with random effects. *Biometrics*, 55:75–84.

Table 1: Comparison of empirical bias, Monte Carlo standard error (MCSE) and average estimated standard error (ASE) for IPW, regression and DR estimators, 1000 Monte Carlo samples with different sample size n and binary outcomes.

	$n = 1000$			$n = 5000$		
$\pi\_tru\ \mu\_mis$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{ipw}$	0.050	0.184	0.205	0.003	0.100	0.104
$\hat{\psi}^{reg}$	0.034	0.140	0.151	0.085	0.062	0.064
$\hat{\psi}^{DR}$	0.029	0.181	0.209	0.006	0.105	0.110
$\pi\_mis\ \mu\_tru$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{ipw}$	0.068	0.154	0.162	0.116	0.067	0.069
$\hat{\psi}^{reg}$	0.054	0.155	0.174	0.010	0.086	0.089
$\hat{\psi}^{DR}$	0.044	0.177	0.197	5e-4	0.097	0.100
$\pi\_tru\ \mu\_tru$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{DR}$	0.037	0.175	0.196	0.001	0.097	0.100

Table 2: Comparison of empirical bias, Monte Carlo standard error (MCSE) and average estimated standard error (ASE) for IPW, regression and DR estimators, 1000 Monte Carlo samples with different sample size n and continuous outcomes.

	$n = 1000$			$n = 5000$		
$\pi\_tru\ \mu\_mis$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{ipw}$	0.024	0.324	0.324	0.011	0.145	0.141
$\hat{\psi}^{reg}$	0.738	0.326	0.328	0.725	0.145	0.144
$\hat{\psi}^{DR}$	0.013	0.293	0.295	0.002	0.131	0.128
$\pi\_mis\ \mu\_tru$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{ipw}$	0.806	0.376	0.370	0.783	0.166	0.162
$\hat{\psi}^{reg}$	0.009	0.304	0.314	5e-04	0.128	0.129
$\hat{\psi}^{DR}$	0.013	0.309	0.317	3e-04	0.128	0.129
$\pi\_tru\ \mu\_tru$						
	Bias	MCSE	ASE	Bias	MCSE	ASE
$\hat{\psi}^{DR}$	0.012	0.308	0.314	0.001	0.128	0.129

Table 3: Point estimates and estimated se [in bracket] of IPW, regression and DR estimators for ETT of 401(k) plan as well as the parameters for IV, propensity score and outcome regression outcome models required by those estimators.

	IV model	IPW propensity	regression	DR propensity
Intercept	-0.180 [0.058]	-8.685 [1.832]	1.307 [0.073]	-8.629 [1.796]
linc	2.695 [0.107]	1.626 [0.210]	0.618 [0.128]	1.633 [0.209]
age	0.007 [0.002]	-0.009 [0.005]	0.035 [0.003]	-0.009 [0.005]
fsize	-0.037 [0.019]	-0.004 [0.033]	-0.127 [0.022]	-0.005 [0.033]
marr	-0.145 [0.063]	-0.032 [0.108]	-0.133 [0.075]	-0.031 [0.108]
age^2	-0.002 [2e-04]	0.001 [4e-04]	6e-04 [3e-04]	0.001 [4e-04]
Z		9.150 [1.820]	-0.210 [0.074]	9.126 [1.781]
α		0.320 [0.115]	0.385 [0.135]	0.280 [0.101]
$\psi = E(Y_0 \mid A = 1)$		0.749 [0.012]	0.746 [0.012]	0.750 [0.012]
ETT		0.134 [0.013]	0.137 [0.014]	0.132 [0.014]

Online Supplementary Materials for "Doubly Robust Estimation of a Marginal Average Effect of Treatment on the Treated With an Instrumental Variable"

Appendix A contains proofs of the propositions. Appendix B presents proofs of the examples in the main text, and more examples about identification of the models. Appendix C presents more derivations mentioned in the main text.

APPENDIX A. **PROOFS OF PROPOSITIONS**

Proof of Proposition 1

Proof. We prove by contradiction. Suppose we have two candidates $\mathrm{Pr}_1(A, Z, Y_0, C)$ and $\mathrm{Pr}_2(A, Z, Y_0, C)$ satisfying the same observed density:

$$\mathrm{Pr}_1(A = 0, Z, Y_0, C) = \mathrm{Pr}_2(A = 0, Z, Y_0, C).$$

By the exclusion restriction assumption, we have the decomposition for the joint distribution:

$$f_j(A, Z, Y_0, C) = f_j(C) f_j(Z|C) f_j(Y_0|C) f_j(A|Y_0, Z, C) \text{ for } j = 1, 2.$$

Since $f(C)$ and $f(Z|C)$ can be identified from the observed data, we have $f_1(C) = f_2(C)$ and $f_1(Z|C) = f_2(Z|C)$. Thus,

$$f_1(Y_0|C)\mathrm{Pr}_1(A = 0|Y_0, Z, C) = f_2(Y_0|C)\mathrm{Pr}_2(A = 0|Y_0, Z, C),$$

and equivalently

$$\frac{\mathrm{Pr}_1(A = 0|Y_0, Z, C)}{\mathrm{Pr}_2(A = 0|Y_0, Z, C)} = \frac{f_2(Y_0|C)}{f_1(Y_0|C)}.$$

The equation contradicts the condition that we require the ratios unequal. So, that the ratios are not equal is equivalent to the impossibility of two sets of candidates satisfying the same observed quantities, i.e. the identifiability of the joint distribution. □

Proof of Proposition 2

Proof. We first prove equation (2). Note that

$$\begin{aligned} & E\left\{\frac{\pi(Y_0,Z,C)Y_0(1-A)}{\Pr(A=1)(1-\pi(Y_0,Z,C))}\right\} \\ = \;& E\left\{\frac{\pi(Y_0,Z,C)Y_0}{\Pr(A=1)}\right\} \\ = \;& E\left\{\frac{AY_0}{\Pr(A=1)}\right\} \\ = \;& E(Y_0|A=1) \\ = \;& \psi. \end{aligned}$$

Thus, equation (2) is proved.

We show that if $\pi(Y,Z,C)$ is correctly specified, the equations (3)–(6) hold at the true value γ thus they are indeed estimating equations for γ. The equality is easy to show for (3)–(5) by the law of iterated expectations. For (6), note that by the exclusion restriction assumption, we have $Y_0 \perp\!\!\!\perp Z|C$, thus

$$\begin{aligned} & E\big[\frac{1-A}{1-\pi(Y,Z,C)}t(Y,C)\{l(Z,C)-E(l(Z,C)|C)\}\big] \\ = \;& E\big[\frac{1-A}{1-\pi(Y_0,Z,C)}t(Y_0,C)\{l(Z,C)-E(l(Z,C)|C)\}\big] \\ = \;& E\big[t(Y_0,C)\{l(Z,C)-E(l(Z,C)|C)\}\big] \\ = \;& E\big[E(t(Y_0,C)|C)\{E(l(Z,C)|C)-E(l(Z,C)|C)\}\big] \\ = \;& 0. \end{aligned}$$

Thus, by equation (2), $\hat{\psi}^{ipw}$ is consistent for ψ.

Note that condition (A.1) is sufficient for local uniqueness of nuisance parameter estimates obtained from equations (3)–(6) and thus ψ is identified from the observed data.

$$E\frac{\partial}{\partial\gamma^T}\frac{1-A}{1-\pi(Y,Z,C;\gamma)}\begin{pmatrix} 1 \\ h_1(Z,C)-E(h_1(Z,C)|C) \\ h_2(C)-E(h_2(C)) \\ t(Y,C)\{l(Z,C)-E(l(Z,C)|C)\} \end{pmatrix} \text{ is invertible} \qquad \text{(A.1)}$$

□

Proof of Proposition 3

Proof. We first prove equation (7). Note that

$$\frac{\Pr(A=1|Y_0,Z,C)}{\Pr(A=0|Y_0,Z,C)} = \exp\{\alpha(Y_0,Z,C)+\beta(Z,C)\},$$

and

$$\frac{\Pr(A=1|Y_0=0,Z,C)}{\Pr(A=0|Y_0=0,Z,C)} = \exp\{\beta(Z,C)\}.$$

Thus,

$$\begin{aligned} & \frac{f(Y_0|A=1,Z,C)f(Y_0=0|A=0,Z,C)}{f(Y_0|A=0,Z,C)f(Y_0=0|A=1,Z,C)} \\ = \; & \frac{\Pr(A=1|Y_0,Z,C)\Pr(A=0|Y_0=0,Z,C)}{\Pr(A=0|Y_0,Z,C)\Pr(A=1|Y_0=0,Z,C)} \\ = \; & \exp\{\alpha(Y_0,Z,C)\}. \end{aligned}$$

Hence,

$$\begin{aligned} & \frac{E[\exp\{\alpha(Y,Z,C)\}g(Y,C)|A=0,Z,C]}{E[\exp\{\alpha(Y,Z,C)\}|A=0,Z,C]} \\ = \; & \frac{E[\exp\{\alpha(Y_0,Z,C)\}g(Y_0,C)|A=0,Z,C]}{E[\exp\{\alpha(Y_0,Z,C)\}|A=0,Z,C]} \\ = \; & E\Big[\frac{f(Y_0|A=1,Z,C)f(Y_0=0|A=0,Z,C)}{f(Y_0|A=0,Z,C)f(Y_0=0|A=1,Z,C)}g(Y,C)|A=0,Z,C\Big]/ \\ & E\Big[\frac{f(Y_0|A=1,Z,C)f(Y_0=0|A=0,Z,C)}{f(Y_0|A=0,Z,C)f(Y_0=0|A=1,Z,C)}|A=0,Z,C\Big] \\ = \; & E\Big[\frac{f(Y_0|A=1,Z,C)}{f(Y_0|A=0,Z,C)}g(Y,C)|A=0,Z,C\Big]/E\Big[\frac{f(Y_0|A=1,Z,C)}{f(Y_0|A=0,Z,C)}|A=0,Z,C\Big] \\ = \; & E(g(Y_0,C)|A=1,Z,C)/1 \\ = \; & E(g(Y_0,C)|A=1,Z,C). \end{aligned}$$

We then show that equation (8) holds at the true value of ξ and η and thus are indeed estimating equation for η. Note that by (IV.2), we have $Y_0 \perp\!\!\!\perp Z|C$, thus

$$\begin{aligned} & E\big\{\big(w(Z,C) - E(w(Z,C)|C)\big)\big(AE(g(Y_0,C)|A=1,Z,C) + (1-A)g(Y,C)\big)\big\} \\ = \;& E\big\{\big(w(Z,C) - E(w(Z,C)|C)\big)\big(Ag(Y_0,C) + (1-A)g(Y_0,C)\big)\big\} \\ = \;& E\big\{\big(w(Z,C) - E(w(Z,C)|C)\big)g(Y_0,C)\big\} \\ = \;& E\big\{\big(E(w(Z,C)|C) - E(w(Z,C)|C)\big)E(g(Y_0,C)|C)\big\} \\ = \;& 0. \end{aligned}$$

Consistency of the regression estimator follows from equation (8). Note that the following condition (A.2) is sufficient for local uniqueness of estimates for η obtained from equation (8).

$$E\{\{\omega(Z,C) - E(\omega(Z,C)|C)\}A\frac{\partial}{\partial\eta}\frac{E(\exp\{\alpha(Y,Z,C;\eta)\}g(Y,C)|A=0,Z,C)}{E(\exp\{\alpha(Y,Z,C;\eta)\}|A=0,Z,C)}\} \text{ is invertible.} \tag{A.2}$$

To see the relationship between (A.2) and the first derivative of (8), note that

$$\begin{aligned} & \frac{\partial}{\partial\eta}E[\{\omega(Z,C) - E(\omega(Z,C)|C)\}\{AE(g(Y_0,C)|A=1,Z,C;\eta) + (1-A)g(Y,C)\}] \\ = \;& E[\{\omega(Z,C) - E(\omega(Z,C)|C)\}\{A\frac{\partial E(g(Y_0,C)|A=1,Z,C;\eta)}{\partial\eta} + (1-A)g(Y,C)\}] \\ = \;& E[\{\omega(Z,C) - E(\omega(Z,C)|C)\}\{A\frac{\partial}{\partial\eta}\frac{E(\exp\{\alpha(Y,Z,C;\eta)\}g(Y,C)|A=0,Z,C)}{E(\exp\{\alpha(Y,Z,C;\eta)\}|A=0,Z,C)}\}]. \end{aligned}$$

□

Proof of Proposition 4

Proof. We use the superscript $*$ to denote a misspecified model. Otherwise, an expectation or a model is always evaluated at the true value of parameters. Note that by the definition of IV, we have $Y_0 \perp\!\!\!\perp Z|C$. Thus, if parametric models lie in $\mathcal{M}_a$:

$$\begin{aligned}
& E\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)\Big\{Q_g(Y,A,Z,C;\tilde{\gamma},\hat{\xi}) + (1-A)g(Y,C)\Big\} \\
\xrightarrow{p} \quad & E\Big[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)\big\{\pi(Y_0,Z,C)(g(Y_0,C) - E(g(Y_0,C)|A=1,Z,C)) + \\
& \qquad \pi(Y_0,Z,C)E(g(Y_0,C)|A=1,Z,C) + (1-\pi(Y_0,Z,C))g(Y_0,C)\big\}\Big] \\
= \quad & E[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)g(Y_0,C)] \\
= \quad & E[\big(E(\omega(Z,C)|C) - E(\omega(Z,C)|C)\big)g(Y_0,C)] = 0.
\end{aligned}$$

Also,

$$\begin{aligned}
& \hat{\psi}^{DR} \\
\xrightarrow{p} \quad & E\Big\{\frac{\pi(Y_0,Z,C)\{Y_0 - E^*(Y_0|A=1,Z,C)\}}{\Pr(A=1)} + \frac{E^*(Y_0|A=1,Z,C)\pi(Y_0,Z,C)}{\Pr(A=1)}\Big\} \\
= \quad & E(\frac{\pi(Y_0,Z,C)Y_0}{\Pr(A=1)}) \\
= \quad & E(\frac{AY_0}{\Pr(A=1)}) \\
= \quad & E(Y_0|A=1) = \psi.
\end{aligned}$$

Thus, if parametric models lie in $\mathcal{M}_y$:

$$\begin{aligned}
& E\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)\Big\{Q_g(Y,A,Z,C;\tilde{\gamma},\hat{\xi}) + (1-A)g(Y,C)\Big\} \\
\xrightarrow{p} \quad & E\Big[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big) \\
& \Big\{(1-A)\exp(\alpha(Y_0,Z,C) + \beta^*(Z,C))(g(Y_0,C) - \frac{E(g(Y_0,C)\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}{E(\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}) \\
& + AE(g(Y_0,C)|A=1,Z,C) + (1-A)g(Y_0,C)\Big\}\Big] \\
= \quad & E\Big[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)\Big\{AE(g(Y_0,C)|A=1,Z,C) + (1-A)g(Y_0,C)\Big\}\Big] \\
= \quad & E[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)\{\Pr(A=1|Z,C)E(g(Y_0,C)|A=1,Z,C) \\
& \qquad + \Pr(A=0|Z,C)E(g(Y_0,C)|A=0,Z,C)\}] \\
= \quad & E[\big(\omega(Z,C) - E(\omega(Z,C)|C)\big)E(g(Y_0,C)|Z,C)] \\
= \quad & E[\big(E(\omega(Z,C)|C) - E(\omega(Z,C)|C)\big)E(g(Y_0,C)|C)] \\
= \quad & 0.
\end{aligned}$$

Also,

$$\begin{aligned}
& \hat{\psi}^{DR} \\
\xrightarrow{p} \quad & E\Bigg[\frac{1-A}{\Pr(A=1)}\frac{\pi(Y_0,Z,C)}{1-\pi(Y_0,Z,C)}\left\{Y_0-\frac{E(Y_0\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}{E(\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}\right\} \\
& +\frac{AE(Y_0|A=1,Z,C)}{\Pr(A=1)}\Bigg] \\
= \quad & E\Bigg[\frac{1-A}{\Pr(A=1)}\exp\{\alpha(Y_0,Z,C)+\beta^*(Z,C)\}\left\{Y_0-\frac{E(Y_0\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}{E(\exp(\alpha(Y_0,Z,C))|A=0,Z,C)}\right\} \\
& +\frac{E(Y_0A)}{\Pr(A=1)}\Bigg] \\
= \quad & E(\frac{AY_0}{\Pr(A=1)}) \\
= \quad & E(Y_0|A=1)=\psi.
\end{aligned}$$

Thus, $\tilde{\eta}$ and $\hat{\psi}^{DR}$ are DR for η and ψ respectively. □

APPENDIX B. **PROOFS FOR EXAMPLES IN SECTION 3**

Proof of example 1

Proof. Let $\Pr(A=0|Y_0,Z,C;\theta_1,\theta_2,\eta_1,\eta_2)=\text{expit}(\theta_1+\theta_2Z+\eta_1Y_0+\eta_2Y_0Z)$ and $\Pr(Y_0=1|C;\tau)=\exp(\tau)$. We show that for any $(\theta_1,\theta_2,\eta_1,\eta_2,\tau)$, there exists $(\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\eta}_1,\tilde{\eta}_2,\tilde{\tau})\neq(\theta_1,\theta_2,\eta_1,\eta_2,\tau)$ such that

$$\Pr(A=0|Y_0,Z,C;\theta_1,\theta_2,\eta_1,\eta_2)\Pr(Y_0|C;\tau)=\Pr(A=0|Y_0,Z,C;\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\eta}_1,\tilde{\eta}_2)\Pr(Y_0|C;\tilde{\tau}). \tag{A.3}$$

Suppose there exists $\rho_1\neq0$ such that $\Pr(Y_0=0|C;\tilde{\tau})/\Pr(Y_0=0|C;\tau)=\exp(\rho_1)$, thus, (A.3) is equivalent to

$$\frac{\Pr(A=0|Y_0,Z,C;\theta_1,\theta_2,\eta_1,\eta_2)}{\Pr(A=0|Y_0,Z,C;\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\eta}_1,\tilde{\eta}_2)}=\frac{\Pr(Y_0|C;\tilde{\tau})}{\Pr(Y_0|C;\tau)}=\exp(\rho_1+\rho_2Y_0), \tag{A.4}$$

where $\rho_2=\log[\exp(-\rho_1-\tau)+\{\exp(\tau)-1\}/\exp(\tau)]$. Note that two different sets of parameters would lead to the same observed data distribution by properly choosing ρ_1 and choosing $\tilde{\theta}_1=\theta_1-\rho_1-\log\varpi_1$, $\tilde{\theta}_2=\theta_2+\log\varpi_1-\log\varpi_2$, $\tilde{\eta}_1=\eta_1-\rho_2+\log\varpi_1-\log\varpi_3$, $\tilde{\eta}_2=$

$\eta_2+\log\varpi_2+\log\varpi_3-\log\varpi_1-\log\varpi_4$ and $\tilde{\tau}=\tau+\rho_1+\rho_2$, where $\varpi_1 = 1+\exp(\theta_1)-\exp(\theta_1-\rho_1)$, $\varpi_2 = 1+\exp(\theta_1+\theta_2)-\exp(\theta_1+\theta_2-\rho_1)$, $\varpi_3 = 1+\exp(\theta_1+\eta_1)-\exp(\theta_1+\eta_1-\rho_1-\rho_2)$ and $\varpi_4 = 1+\exp(\theta_1+\theta_2+\eta_1+\eta_2)-\exp(\theta_1+\theta_2+\eta_1+\eta_2-\rho_1-\rho_2)$. For example, choose $\rho_1 = 0.3$, $\rho_2 = -0.38$, $(\theta_1,\theta_2,\eta_1,\eta_2,\tau_1) = (0.3, 0.6, 0.1, 0.7, -0.2)$ and $(\tilde{\theta}_1,\tilde{\theta}_2,\tilde{\eta}_1,\tilde{\eta}_2,\tilde{\tau}) = (-0.3, 0.41, 0.91, 1.37, -0.28)$, it is easy to verify they lead to the same observed distribution.

Next, we prove identifiability of the separable treatment mechanism:

$$\mathcal{P}_{A|Y_0,Z} = \{\Pr(A=0|Y_0,Z) : \text{logit} \;\; P(A=0|Y_0,Z,\theta_1,\theta_2,\eta_1) = \theta_1+\theta_2 Z+\eta_1 Y_0, \theta_1,\theta_2,\eta_1\in\mathcal{R}\}.$$

Under such treatment mechanism, we have $\eta_2=\tilde{\eta}_2=0$, and thus $\varpi_2\varpi_3=\varpi_1\varpi_4$, i.e. $\{1+\exp(\theta_1+\theta_2)-\exp(\theta_1+\theta_2-\rho_1)\}\{1+\exp(\theta_1+\eta_1)-\exp(\theta_1+\eta_1-\rho_1-\rho_2)\} = \{1+\exp(\theta_1)-\exp(\theta_1-\rho_1)\}\{1+\exp(\theta_1+\theta_2+\eta_1)-\exp(\theta_1+\theta_2+\eta_1-\rho_1-\rho_2)\}$ which indicates

$$\exp(\rho_2) = \frac{\exp(\eta_1)}{1+\exp(\eta_1+\rho_1)-\exp(\rho_1)}. \tag{A.5}$$

Since in (A.4), $\exp(\rho_1+\rho_2 Y_0)$ is the ratio of two densities for Y_0, we have ρ_1 and $\rho_1+\rho_2$ should be of the opposite sign. From equation (A.5), if $\rho_1>0$, then $\exp(\rho_1)>1$ and $\exp(\rho_1+\rho_2)>1$. Similarly, if $\rho_1<0$, then $\exp(\rho_1)<1$ and $\exp(\rho_1+\rho_2)<1$. Thus, we conclude that $\rho_1=\rho_2=0$, i.e. the separable treatment mechanism is identifiable for binary case. □

Proof of example 2

Proof. Suppose there exist two densities that make the ratios equal,

$$\frac{\text{expit}\{q_1(Z)+h_1(Y_0)\}}{\text{expit}\{q_2(Z)+h_2(Y_0)\}} = \frac{f_2(Y_0)}{f_1(Y_0)}. \tag{A.6}$$

We first take derivatives over Z on both sides, and we have

$$\frac{\partial\text{expit}\{q_1(Z)+h_1(Y_0)\}/\partial Z}{\text{expit}\{q_1(Z)+h_1(Y_0)\}} = \frac{\partial\text{expit}\{q_2(Z)+h_2(Y_0)\}/\partial Z}{\text{expit}\{q_2(Z)+h_2(Y_0)\}},$$

expand the expit functions and simplify the equation, and we have

$$\frac{\partial q_1(Z)/\partial Z}{\partial q_2(Z)/\partial Z}[1+\exp\{q_2(Z)+h_2(Y_0)\}] = 1+\exp\{q_1(Z)+h_1(Y_0)\}. \tag{A.7}$$

Next, we take derivatives over Y_0 on both sides of the above equation, and we have

$$\frac{\partial q_1(Z)/\partial Z}{\partial q_2(Z)/\partial Z}\frac{\partial h_2(Y_0)}{\partial Y_0}\exp\{q_2(Z)+h_2(Y_0)\} = \frac{\partial h_1(Y_0)}{\partial Y_0}\exp\{q_1(Z)+h_1(Y_0)\},$$

and equivalently,

$$\frac{\partial q_1(Z)/\partial Z}{\partial q_2(Z)/\partial Z}\exp\{q_2(Z)-q_1(Z)\} = \frac{\partial h_1(Y_0)/\partial Y_0}{\partial h_2(Y_0)/\partial Y_0}\exp\{h_1(Y_0)-h_2(Y_0)\}.$$

The left hand side of the above equation is a function of Z, but the right hand side is a function of Y_0. So we must have

$$\frac{\partial q_1(Z)/\partial Z}{\partial q_2(Z)/\partial Z}\exp\{q_2(Z)-q_1(Z)\} = c_1,$$

for some constant c_1. We multiply both sides of equation (A.7) by $\exp\{-q_1(Z)\}$, and we have

$$c_1[\exp\{-q_2(Z)\}+\exp\{h_2(Y_0)\}] = \exp\{-q_1(Z)\}+\exp\{h_1(Y_0)\},$$

and thus for some constant c_2,

$$c_1\exp\{-q_2(Z)\}+c_2 = \exp\{-q_1(Z)\}, \quad c_1\exp\{h_2(Y_0)\}-c_2 = \exp\{h_1(Y_0)\}.$$

We substitute $q_2(Z)$ and $h_2(Y_0)$ in equation (A.6) with the expressions above to obtain

$$\exp\{h_1(Y_0)\}+c_2 = \exp\{h_1(Y_0)\}\frac{f_1(Y_0)}{f_2(Y_0)},$$

and thus

$$\frac{f_1(Y_0)}{f_2(Y_0)} = 1+c_2\exp\{-h_1(Y_0)\}.$$

Note that $1+c_2\exp\{-h_1(Y_0)\}>1$ for $c_2>0$, and $1+c_2\exp\{-h_1(Y_0)\}<1$ for $c_2<0$. This cannot be true for the ratio of two densities. So we must have $c_2=0$, and thus $f_1(Y_0)/f_2(Y_0)=1$. As a result, the joint distribution is identifiable. □

The following example provides a simple data generating mechanism which results in the true distribution in the model specified in example 2.

Example A.1. Suppose that (A, Y, Z, C, U) satisfies (1) logit $\Pr(A = 1|U, Z, C) = \gamma_0 + \gamma_1 Z + \gamma_2 C + \gamma_3 U$, (2) $OR(Y, U|A = 0, Z, C) = \exp(wYU)$ and (3) $U|A = 0, Y, Z, C \overset{d}{=} E(U|A = 0, Y, Z, C) + \varepsilon$ where $\varepsilon \perp\!\!\!\perp (Y, Z)|A = 0, C$, then there is no odds ratio interaction between Y_0 and Z in the propensity score model $\Pr(A = 1|Y_0, Z, C)$.

Proof.

$$\begin{aligned}
&\frac{\Pr(A = 1|Y_0, Z, C)}{\Pr(A = 0|Y_0, Z, C)} \\
= \ & \int \frac{\Pr(A = 1|U, Y_0, Z, C)}{\Pr(A = 0|U, Y_0, Z, C)} \Pr(U|Y_0, A = 0, Z, C) dU \\
= \ & \int \frac{\Pr(A = 1|U, Z, C)}{\Pr(A = 0|U, Z, C)} \Pr(U|Y_0, A = 0, Z, C) dU \\
= \ & \int \exp\{\gamma_0 + \gamma_1 Z + \gamma_2 C + \gamma_3 U\} \Pr(U|Y_0, A = 0, Z, C) dU \\
= \ & \exp(\gamma_0 + \gamma_1 Z + \gamma_2 C) E[\exp(\gamma_3 U)|Y_0, A = 0, Z, C] \\
= \ & \exp(\gamma_0 + \gamma_1 Z + \gamma_2 C) E[\exp\{\gamma_3(E[U|Y_0, A = 0, Z, C] + \varepsilon)\}|Y_0, A = 0, Z, C] \\
= \ & \exp(\gamma_0 + \gamma_1 Z + \gamma_2 C) \exp\{\gamma_3 E[U|Y_0, A = 0, Z, C]\} E[\exp\{\gamma_3 \varepsilon\}|Y_0, A = 0, Z, C] \\
= \ & \kappa^{\gamma_3} \exp(\gamma_0 + \gamma_1 Z + \gamma_2 C) \exp\{\gamma_3 E(U|Y, A = 0, Z, C)\},
\end{aligned}$$

where $\kappa = E\{\exp(\varepsilon)|A = 0, C\}$. Note that

$$\begin{aligned}
&E(U|Y, A = 0, Z, C) \\
= \ & [E(U|Y = 1, A = 0, Z, C) - E(U|Y = 0, A = 0, Z, C)]Y + E(U|Y = 0, A = 0, Z, C),
\end{aligned}$$

and that

$$\begin{aligned}
&E(U|Y = 1, A = 0, Z, C) - E(U|Y = 0, A = 0, Z, C) \\
= \ & \frac{E[\exp(wU)U|Y = 0, A = 0, Z, C]}{E[\exp(wU)|Y = 0, A = 0, Z, C)]} - E[U|Y = 0, A = 0, Z, C)] \\
= \ & \frac{\partial}{\partial w} \log E[\exp(wU)|Y = 0, A = 0, Z, C)] - E[U|Y = 0, A = 0, Z, C)] \\
= \ & \frac{\partial}{\partial w} \{w(E(U|Y = 0, A = 0, Z, C) + \log \kappa)\} - E(U|Y = 0, A = 0, Z, C) \\
= \ & \log \kappa.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \frac{\Pr(A=1|Y_0,Z,C)}{\Pr(A=0|Y_0,Z,C)} \\
= \; & \kappa^{\gamma_3}\exp\{\gamma_0+\gamma_1 Z+\gamma_2 C\}\exp\{\gamma_3\log\kappa Y+\gamma_3 E(U|Y=0,A=0,Z,C)\},
\end{aligned}$$

which does not involve an interaction term between Y and Z. □

The data generating mechanism described in example A.1 is a generalization of the shared parameter model (Wu and Carroll, 1988; Wu and Follmann, 1999), which is semiparametric in that the distribution of U satisfies (3) of example A.1, but is otherwise unrestricted. More specifically, if the odds ratio function relating Y and U does not depend on Z and the residual ε is independent of (Y,Z), then the absence of an interaction between U and Z in the propensity score $\Pr(A=1|U,Z,C)$ implies no interaction between Y_0 and Z in the propensity score $\Pr(A=1|Y_0,Z,C)$.

The separable treatment mechanisms are also identifiable for continuous outcome with binary instrument.

Example A.2. Consider the case of continuous outcome with binary instrument. Assume the Logistic separable treatment mechanism: $\mathcal{P}_{A|Y_0,Z}=\{\Pr(A=0|Y_0,Z):\text{logit }\Pr(A=0|Y_0,Z)=\theta Z+h(Y_0)\}$, where h is a known or unknown function. It can be shown that $\mathcal{P}_{A|Y_0,Z}$ satisfies the condition 1 and thus the joint distribution is identifiable.

Proof of example A.2

Proof. Suppose there exist two sets of densities make the ratios equal,

$$\frac{\text{expit}\{\theta_1 Z+h_1(Y_0)\}}{\text{expit}\{\theta_2 Z+h_2(Y_0)\}}=\frac{f_2(Y_0)}{f_1(Y_0)}. \tag{A.8}$$

The above equation holds for both $Z=0,1$, so we have

$$\frac{\text{expit}\{h_1(Y_0)\}}{\text{expit}\{h_2(Y_0)\}}=\frac{\text{expit}\{\theta_1+h_1(Y_0)\}}{\text{expit}\{\theta_2+h_2(Y_0)\}}.$$

Simplifying the equation, we have

$$\exp\{h_1(Y_0)\}=\frac{\exp(\theta_2)-\exp(\theta_1)+\{\exp(\theta_2)-\exp(\theta_1+\theta_2)\}\exp\{h_2(Y_0)\}}{\exp(\theta_1)-\exp(\theta_1+\theta_2)}.$$

Substituting $\exp\{h_1(Y_0)\}$ with the above expression in equation (A.8), we have

$$\frac{f_2(Y_0)}{f_1(Y_0)} = 1 + \frac{\exp(\theta_2) - \exp(\theta_1)}{\exp(\theta_2) - \exp(\theta_1 + \theta_2)} \exp\{-h_2(Y_0)\}.$$

If $\theta_1 \neq \theta_2$, we must have $f_2(Y_0)/f_1(Y_0) < 1$ for any Y_0, or $f_2(Y_0)/f_1(Y_0) > 1$ for any Y_0. This cannot be true for the ratio of two densities. So we must have $\theta_1 = \theta_2$, and thus $f_1(Y_0)/f_2(Y_0) = 1$. As a result, the joint distribution is identifiable. □

Example A.3. Assume the Probit separable treatment mechanism: $\mathcal{P}_{A|Y_0,Z} = \{\Pr(A = 0|Y_0, Z) : \Pr(A = 0|Y_0, Z) = \Phi\{q(Z) + h(Y_0)\}\}$, where Φ is the standard normal distribution function, q and h are known or unknown functions, and q is differentiable. Then the joint distribution of A, Y_0, Z is identifiable.

Proof of example A.3

Proof. Suppose two sets of parameters make the ratio being a function of Y_0, i.e. for some function s,

$$\Phi\{q_1(Z) + h_1(Y_0)\} = \Phi\{q_2(Z) + h_2(Y_0)\}s(Y_0).$$

By taking derivatives over Z on both sides, we have

$$\frac{\partial q_1(Z)}{\partial Z}\phi\{q_1(Z) + h_1(Y_0)\} = \frac{\partial q_2(Z)}{\partial Z}\phi\{q_2(Z) + h_2(Y_0)\}s(Y_0),$$

where ϕ is the standard normal density function. And equivalently

$$\log\frac{\phi\{q_1(Z) + h_1(Y_0)\}}{\phi\{q_2(Z) + h_2(Y_0)\}} = \log\frac{\partial q_2(Z)/\partial Z}{\partial q_1(Z)/\partial Z} + \log s(Y_0),$$

which implies that

$$\{q_2(Z) + h_2(Y_0)\}^2 - \{q_1(Z) + h_1(Y_0)\}^2 = 2\left\{\log\frac{\partial q_2(Z)/\partial Z}{\partial q_1(Z)/\partial Z} + \log s(Y_0)\right\}. \qquad \text{(A.9)}$$

Note that the right hand side does not include an interaction term of Z and Y_0, we have

$$q_1(Z)h_1(Y_0) = q_2(Z)h_2(Y_0),$$

and thus

$$\frac{q_1(Z)}{q_2(Z)} = \frac{h_2(Y_0)}{h_1(Y_0)}.$$

So the only possible case is when $q_1(Z) = cq_2(Z)$ and $h_2(Y_0) = ch_1(Y_0)$ for some positive constant c. Substituting q_2 and h_2 with $1/cq_1$ and ch_1 in equation (A.9), we have

$$\left(\frac{1}{c^2} - 1\right) q_1^2(Z) + (c^2 - 1)h_1(Y_0)^2 = 2\{-\log c + \log s(Y_0)\}.$$

Note that the right hand side does not vary with Z, we must have $c = 1$, and thus $q_1(Z) = q_2(Z)$ and $h_2(Y_0) = h_1(Y_0)$. □

APPENDIX C. **ADDITIONAL RESULTS MENTIONED IN THIS PAPER**

Regression estimator using any link function λ

Let $\delta(Z, C) = \lambda\{E(Y_0|A = 0, Z, C)\}$, then $E(Y_0|A = 1, Z, C) = \lambda^{-1}\{\tilde{\alpha}(1, Z, C) + \delta(Z, C)\}$. We let $\delta(Z, C; \xi)$ denote a parametric model for $\delta(Z, C)$ and let $\hat{\xi}$ denote the MLE of ξ using only data among the unexposed. Although in the main text η is used to denote the parameter in α, here we use it to denote the parameters in $\tilde{\alpha}$. We obtain an estimator for η by solving:

$$\mathbb{P}_n\left[\{w(Z, C) - E(w(Z, C)|C; \hat{\rho})\}\left\{A\lambda^{-1}\{\tilde{\alpha}(1, Z, C; \eta) + \delta(Z, C; \hat{\xi})\} + (1 - A)Y\right\}\right] = 0, \tag{A.10}$$

We have the following proposition for the outcome regression estimator with any link function λ,

Proposition A.1. Suppose $\tilde{\alpha}(A, Z, C; \eta)$, $f_{Z|C}(z|c; \rho)$ and $\delta(Z, C; \xi)$ are correctly specified, then the outcome regression estimator

$$\hat{\psi}^{reg} = \mathbb{P}_n \frac{A}{\hat{\Pr}(A = 1)} \lambda^{-1}\{\tilde{\alpha}(1, Z, C; \hat{\eta}) + \delta(Z, C; \hat{\xi})\},$$

is consistent.

Relationship between outcome models

The following result was used in the simulations for binary outcome. For binary Y, we derive the relationship between the regression model $\Pr(Y_0 = 1|A = 1, Z, C)$ and data generating model $\Pr(Y_0 = 1|C)$ and this casts light on how to control the degree of misspecification of the regression model through the data generation model:

$$\begin{aligned}
& \text{logit } \Pr(Y_0 = 1|A, Z, C) \\
= \;& \log \frac{\Pr(Y_0 = 1|A, Z, C)}{\Pr(Y_0 = 0|A, Z, C)} \\
= \;& \log\Big\{\frac{\Pr(Y_0 = 1|A, Z, C)}{\Pr(Y_0 = 0|A, Z, C)}\Big/\frac{\Pr(Y_0 = 1|A = 0, Z, C)}{\Pr(Y_0 = 0|A = 0, Z, C)}\Big\} \\
& - \log\Big\{\frac{\Pr(Y_0 = 1|Z, C)}{\Pr(Y_0 = 0|Z, C)}\Big/\frac{\Pr(Y_0 = 1|A = 0, Z, C)}{\Pr(Y_0 = 0|A = 0, Z, C)}\Big\} + \log \frac{\Pr(Y_0 = 1|Z, C)}{\Pr(Y_0 = 0|Z, C)}.
\end{aligned}$$

Note that

$$\begin{aligned}
& \sum_a \frac{\Pr(Y_0 = 1|A = a, Z, C)}{\Pr(Y_0 = 0|A = a, Z, C)} \Pr(A = a|Y_0 = 0, Z, C) \\
= \;& \sum_a \frac{\Pr(Y_0 = 1, A = a|Z, C)}{\Pr(Y_0 = 0, A = a|Z, C)} \Pr(A = a|Y_0 = 0, Z, C) \\
= \;& \sum_a \frac{\Pr(Y_0 = 1, A = a|Z, C)}{\Pr(Y_0 = 0|Z, C) \Pr(A = a|Y_0 = 0, Z, C)} \Pr(A = a|Y_0 = 0, Z, C) \\
= \;& \sum_a \frac{\Pr(Y_0 = 1, A = a|Z, C)}{\Pr(Y_0 = 0|Z, C)} = \frac{\Pr(Y_0 = 1|Z, C)}{\Pr(Y_0 = 0|Z, C)},
\end{aligned}$$

i.e. we can marginalize the ratio $\Pr(Y_0 = 1|A, Z, C)/\Pr(Y_0 = 0|A, Z, C)$ using the probability

$\Pr(A|Y_0=0,Z,C)$ to get the marginalized ratio $\Pr(Y_0=1|Z,C)/\Pr(Y_0=0|Z,C)$. Thus,

$$\begin{aligned}
& \text{logit}\ \Pr(Y_0=1|A,Z,C) \\
= \;& \log\Big\{\frac{\Pr(Y_0=1|A,Z,C)}{\Pr(Y_0=0|A,Z,C)}\Big/\frac{\Pr(Y_0=1|A=0,Z,C)}{\Pr(Y_0=0|A=0,Z,C)}\Big\} \\
& -\log\Big\{\frac{\Pr(Y_0=1|Z,C)}{\Pr(Y_0=0|Z,C)}\Big/\frac{\Pr(Y_0=1|A=0,Z,C)}{\Pr(Y_0=0|A=0,Z,C)}\Big\}+\log\frac{\Pr(Y_0=1|Z,C)}{\Pr(Y_0=0|Z,C)} \\
= \;& \log\Big\{\frac{\Pr(Y_0=1|A,Z,C)}{\Pr(Y_0=0|A,Z,C)}\Big/\frac{\Pr(Y_0=1|A=0,Z,C)}{\Pr(Y_0=0|A=0,Z,C)}\Big\} \\
& -\log\Big\{\sum_a\frac{\Pr(Y_0=1|A,Z,C)}{\Pr(Y_0=0|A,Z,C)}\Big/\frac{\Pr(Y_0=1|A=0,Z,C)}{\Pr(Y_0=0|A=0,Z,C)}\Pr(A|Y_0=0,Z,C)\Big\} \\
& +\log\frac{\Pr(Y_0=1|Z,C)}{\Pr(Y_0=0|Z,C)} \\
= \;& \alpha(1,Z,C)A-\log\{\exp\{\alpha(1,Z,C)\}\Pr(A=1|Y_0=0,Z,C)+\Pr(A=0|Y_0=0,Z,C)\} \\
& +\text{logit}\ \Pr(Y_0=1|C).
\end{aligned}$$

Note that in our simulation $\alpha(1,Z,C)=\eta$ and $\Pr(A|Y_0,Z,C)=\Pr(A|Y_0,C_1,Z)$, thus

$$\begin{aligned}
& \text{logit}\ \Pr(Y_0=1|A,Z,C) \\
= \;& \alpha(1,Z,C)A-g(Z,C_1)+\text{logit}\ \Pr(Y_0=1|C_1,C_2),
\end{aligned}$$

where $g(Z,C_1)=\log\{\exp\{\alpha(1,Z,C)\}\Pr(A=1|Y_0=0,Z,C)+\Pr(A=0|Y_0=0,Z,C)\}$. Thus we can control the effect of C_2 in the model $\Pr(Y_0=1|A,Z,C)$ through $\Pr(Y_0=1|C)$.

Table A.4: Convergence failure rate for IPW, regression and DR estimators out of 1000 Monte Carlo samples with different sample size n and binary outcome. The convergence criteria is the residual of the square average estimating equation component being smaller than 1e-5.

	500	1000	5000
$\pi\_tru\ \mu\_mis$			
$\hat{\psi}^{ipw}$	10	2	2
$\hat{\psi}^{reg}$	251	120	4
$\hat{\psi}^{DR}$	18	3	0
$\pi\_mis\ \mu\_tru$			
$\hat{\psi}^{ipw}$	43	54	18
$\hat{\psi}^{reg}$	192	61	1
$\hat{\psi}^{DR}$	49	11	0
$\pi\_tru\ \mu\_tru$			
$\hat{\psi}^{DR}$	42	10	0
$\pi\_mis\ \mu\_mis$			
$\hat{\psi}^{DR}$	20	3	2

Table A.5: Convergence failure rate for IPW, regression and DR estimators out of 1000 Monte Carlo samples with different sample size n and continuous outcome. The convergence criteria is the residual of the square average estimating equation component being smaller than 1e-5.

	500	1000	5000
$\pi\_tru\ \mu\_mis$			
$\hat{\psi}^{ipw}$	22	3	0
$\hat{\psi}^{reg}$	0	0	0
$\hat{\psi}^{DR}$	86	24	2
$\pi\_mis\ \mu\_tru$			
$\hat{\psi}^{ipw}$	12	2	0
$\hat{\psi}^{reg}$	0	0	0
$\hat{\psi}^{DR}$	29	10	0
$\pi\_tru\ \mu\_tru$			
$\hat{\psi}^{DR}$	35	1	0
$\pi\_mis\ \mu\_mis$			
$\hat{\psi}^{DR}$	74	27	2