Statistica Sinica: Supplement

Semiparametric Inference of Causal Effect With Nonignorable Missing Confounders

Zhaohan Sun[1] and Lan Liu[2]

[1] *University of Waterloo, Waterloo, Ontario, Canada*

[2] *School of Statistics, University of Minnesota at Twin Cities, Minneapolis, Minnesota, USA*

Supplementary Material

Section S1 contains the proof of Lemma 1. Section S2 provides the proof of Proposition 1. Section S3 presents the asymptotic normality and variance for the IPW estimator. Section S4 presents the proof of Lemma 2. Section S5 shows the derivation of the equations for the regression estimator. Section S6 shows the proof of Proposition 2. Section S7 presents the proof of Proposition 3. Sections S8 and S9 include the extensions of Lemmas, Propositions and semiparametric estimators to the multiple missing patterns setting and to the average treatment effect on the treated as the parameter of interest setting. Section S10 includes additional table and figure.

S1. Proof of Lemma 1

$$
\begin{aligned}
E\left\{\frac{1(A=a)RY}{\Pr(A,R|X)}\right\} &= E\left\{\frac{1(A=a)RY}{\Pr(R|A,X)\Pr(A|X)}\right\} \\
&= E\left\{\frac{1(A=a)RY}{\Pr(R|A,X,Y)\Pr(A|X)}\right\} \\
&= E\left[E\left[E\left\{\frac{1(A=a)RY}{\Pr(R|A,X,Y)}|A,X,Y\right\}\frac{1}{\Pr(A|X)}\middle|X\right]\right] \\
&= E\left[E\left\{\frac{1(A=a)Y}{\Pr(A|X)}\middle|X\right\}\right] = E\{E(Y_a|X)\} = E(Y_a),
\end{aligned}
$$

where the second equation also follows from the outcome-independent missingness assumption.

S2. Proof of Proposition 1

We impose the following regularity condition for the choice of the user specified function $l(A,Y)$ in equation (3.2).

Condition 1. Assume

$$
E\left[\frac{\partial}{\partial\gamma^T}\left\{\frac{R}{\Pr(R|A,X;\gamma)}-1\right\}l(A,Y)\right] \quad \text{is invertible.}
$$

The regularity condition 1 is sufficient for the local uniqueness of an estimator $\hat{\gamma}^{ipw}$ for parameter γ obtained from equation (3.2).

Next, we show that (3.2) holds at the true parameter. We have

$$\begin{aligned} & E\left[\left\{\frac{R}{\Pr(R|A,X;\gamma)}-1\right\}l(A,Y)\right] \\ = \; & E\left[\left\{\frac{R}{\Pr(R|A,X,Y;\gamma)}-1\right\}l(A,Y)\right] \\ = \; & E\left[E\left\{\frac{R}{\Pr(R|A,X,Y;\gamma)}-1\Big|A,X,Y\right\}l(A,Y)\right] \\ = \; & 0, \end{aligned}$$

where the first step uses the outcome-independent missingness assumption.

The consistency of $\hat{\mu}^{ipw}$ follows from the law of large numbers and the derivation of lemma 1,

$$\begin{aligned} \hat{\mu}^{ipw} \quad & \xrightarrow{p} \quad E\left\{\frac{1(A=a)RY}{\Pr(A,R|X)}\right\} \\ & = \quad E(Y_a)=\mu_a. \end{aligned}$$

S3. Asymptotic normality and variance

The asymptotic normality and the variance of the IPW estimator can be derived using standard M-estimation theory (van der Vaart, A., 1998). Specifically, let $G_{\mu}^{ipw}(O) = \hat{\mu}_a^{ipw} - \mu$ and let $G_{\alpha}^{ipw}(O)$ denote the score function for α. Then, the true parameter value $\theta = (\mu_a, \alpha, \gamma)$ is the solution to $\int G^{ipw}(o;\theta)dF(o;\theta) = 0$, where $G^{ipw}(O) = \{G_{\mu}^{ipw}(O), G_{\alpha}^{ipw}(O), G_{\gamma}^{ipw}(O)\}$ and $G_{\gamma}^{ipw}(O) = \{R/\Pr(R|A,X;\gamma)-1\}l(A,Y)$. Thus, $\hat{\theta}^{ipw} = (\hat{\mu}_a^{ipw}, \hat{\alpha}, \hat{\gamma})$ is the solution to $\sum_{i=1}^{n} G^{ipw}(o_i;\theta) = 0$. By the M-estimation theory, under

regularity conditions, $\sqrt{n}(\hat{\theta}^{ipw} - \theta)$ converges in distribution to $N(0, \Sigma^{ipw})$ when the sample size n goes to infinity, where $\Sigma^{ipw} = U^{-1}VU^{-T}$, $U = -E\{\partial G^{ipw}(O_i;\theta)/\partial\theta\}$ and $V = E\{G_\alpha^{ipw}(O_i;\theta)^{\otimes 2}\}$. A consistent estimator of the asymptotic variance of $\hat{\mu}_a^{ipw}$ can be constructed by replacing expectations with their empirical counterparts and by replacing the parameters with their estimates. Similar derivations could be carried out for the regression and the DR estimators.

S4. Proof of Lemma 2

$$
\begin{aligned}
& E\Big[(1-R)E\{g_a(X)|A,R=0\} + RE\{g_a(X)|A,R=1\}\Big] \\
= \;& E\Big[E\Big[(1-R)E\{g_a(X)|A,R=0\} + RE\{g_a(X)|A,R=1\}|A\Big]\Big] \\
= \;& E\Big[E\Big[(1-R)E\{g_a(X)|A,R\} + RE\{g_a(X)|A,R\}|A\Big]\Big] \\
= \;& E\Big[E\Big\{(1-R)g_a(X) + Rg_a(X)|A\Big\}\Big] \\
= \;& E\Big[E\Big\{g_a(X)|A\Big\}\Big] \\
= \;& E\{g_a(X)\} \\
= \;& E\{E(Y|R=1,A=a,X)\} \\
= \;& E(Y_a),
\end{aligned}
$$

S5. Proof of Lemmas 3–4

We first derive the alternative representations of $\eta(r=0,r_0,x,x_0|a)$,

$$\begin{aligned}
\eta(r=0,r_0,x,x_0|a) &= \frac{\Pr(r=0|a,x)\Pr(r=1|a,x=0)}{\Pr(r=1|a,x)\Pr(r=0|a,x=0)} \\
&= \frac{\Pr(r=0|a,x,y)\Pr(r=1|a,x=0,y=0)}{\Pr(r=1|a,x,y)\Pr(r=0|a,x=0,y=0)} \\
&= \frac{\Pr(r=0,a,x,y)\Pr(r=1,a,x=0,y=0)}{\Pr(r=1,a,x,y)\Pr(r=0,a,x=0,y=0)} \\
&= \frac{f(x,y|a,r=0)f(x=0,y=0|a,r=1)}{f(x,y|a,r=1)f(x=0,y=0|a,r=0)}.
\end{aligned}$$

Hence, we have the following representation for $f(x,y|a,r=0)$,

$$\begin{aligned}
f(x,y|a,r=0) &= \frac{f(x,y|a,r=0)}{\iint f(x,y|a,r=0)\,dx\,dy} \\
&= \frac{\dfrac{f(x,y|a,r=0)f(x=0,y=0|a,r=1)}{f(x=0,y=0|a,r=0)}}{\iint \dfrac{f(x,y|a,r=0)f(x=0,y=0|a,r=1)}{f(x=0,y=0|a,r=0)}\,dx\,dy} \\
&= \frac{\eta(r=0,r_0,x,x_0|a)f(x,y|a,r=1)}{E\{\eta(r=0,r_0,X,x_0|a)|a,r=1\}}.
\end{aligned}$$

Finally, we have

$$
\begin{aligned}
E\{l(a,Y)|a,r=0\} &= \iint l(a,y)f(x,y|a,r=0)\,dx\,dy \\
&= \frac{\iint \eta(r=0,r_0,x,x_0|a)l(a,y)f(x,y|a,r=1)\,dx\,dy}{E\{\eta(r=0,r_0,X,x_0|a)|a,r=1\}} \\
&= \frac{E\{\eta(r=0,r_0,X,x_0|a)l(a,Y)|a,r=1\}}{E\{\eta(r=0,r_0,X,x_0|a)|a,r=1\}}.
\end{aligned}
$$

S6. Proof of Proposition 2

We impose the following regularity condition for the user specified function $l(A,Y)$ in the equation (3.3).

Condition 2. Assume

$$
E\left[(1-R)\left\{\frac{\partial}{\partial\xi^T}\frac{E\{\eta(r=0,r_0,X,x_0|a;\xi)l(a,Y)|a,r=1;\beta\}}{E\{\eta(r=0,r_0,X,x_0|a;\xi)|a,r=1;\beta\}}\right\}\right] \quad \text{is invertible.}
$$

The regularity condition 2 is sufficient for local uniqueness of an estimator for ξ obtained from equation (3.3).

It is straightforward to see the following estimating equation holds at the true parameter values for β and ξ, when $f(x,y|a,r=1;\beta)$ and $\eta(r=0,r_0,x,x_0;\xi)$ are correctly specified.

$$
E\left[(1-R)\left\{l(A,Y)-E\{l(A,Y)|A,R=0;\beta,\xi\}\right\}\right]=0.
$$

By the law of large numbers and Lemma 2, we have

$$\begin{aligned}\hat{\mu}^{reg} &\xrightarrow{p} E\Big[(1-R)E\{g_a(X)|A,R=0\}+RE\{g_a(X)|A,R=1\}\Big]\\ &= E(Y_a)=\mu_a,\end{aligned}$$

proving the consistency of $\hat{\mu}_a^{reg}$ for μ_a.

S7. Proof of Proposition 3

First, we impose the following regularity condition for the choice of user specified function $l(A,Y)$ in equation (3.5).

Condition 3. Assume

$$E\Bigg[\frac{\partial}{\partial\xi^T}\Bigg[\frac{Rl(A,Y)}{\Pr(R|A,X;\delta,\xi)}-\Bigg\{\frac{R}{\Pr(R|A,X;\delta,\xi)}-1\Bigg\}\times\Bigg\{\frac{E\{\eta(r=0,r_0,X,x_0|a;\xi)l(a,Y)|a,r=1;\beta\}}{E\{\eta(r=0,r_0,X,x_0|a;\xi)|a,r=1;\beta\}}\Bigg\}\Bigg]\Bigg]$$

is invertible.

The regularity condition 3 is sufficient for the local uniqueness of estimators for the parameters δ and ξ obtained from equation (3.5).

Proof: To prove the DR property of the DR estimators, assume there exists α^* and δ^* such that $\hat{\alpha} \xrightarrow{p} \alpha^*$ and $\hat{\delta} \xrightarrow{p} \delta^*$ as $n \to \infty$, that is, the estimators $\hat{\alpha}$ and $\hat{\delta}$ will converge in probability to some constants regardless of whether the propensity score models are correctly specified or

not. Similarly, assume there exists $\beta^\dagger$ such that $\hat{\beta} \xrightarrow{p} \beta^\dagger$ as $n \to \infty$. With slightly abuse of notations, we use α, β, δ and ξ to denote the true parameter values and α^*, δ^* and $\beta^\dagger$ to denote the possibly misspecified parameter values. Let $h_a^*(A,X,Y) = h_a(A,X,Y;\alpha^*,\beta,\delta^*,\xi)$ and $h_a^\dagger(A,X,Y) = h_a(A,X,Y;\alpha,\beta^\dagger,\delta,\xi)$, where $h_a(A,X,Y;\alpha,\beta,\delta,\xi) = 1(A = a)\{Y - g_a(X;\beta)\}/\Pr(A|X;\alpha,\delta,\xi) + g_a(X;\beta)$. We first prove that if (a) $\Pr(r = 1|a, x = 0;\delta)$ is correctly specified or if (b) $f(x,y|a,r = 1;\beta)$ is correctly specified, equation (3.5) holds at the true parameter ξ, hence $\hat{\xi}^{dr}$ is doubly robust.

If $\Pr(r = 1|a, x = 0;\delta)$ is correctly specified, since we also assume $\eta(r = 0, r_0, x, x_0|a;\xi)$ is correctly specified, then $\Pr(r|a,x;\delta,\xi)$ is correctly specified. Also, by the outcome-independent missingness assumption $R \perp\!\!\!\perp Y|A,X$, we have

$$
\begin{aligned}
& E\left[\{\frac{R}{\Pr(R=1|A,X;\delta,\xi)} - 1\}\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta^\dagger,\xi\}\right\}\right] \\
= \;& E\left[E\left[\{\frac{R}{\Pr(R=1|A,X;\delta,\xi)} - 1\}\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta^\dagger,\xi\}\right\}\middle| A,X\right]\right] \\
= \;& E\left[E\left\{\frac{R}{\Pr(R=1|A,X;\delta,\xi)} - 1\middle| A,X\right\}E\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta^\dagger,\xi\}\right\}\right] \\
= \;& 0.
\end{aligned}
$$

If $\Pr(x, y|a; r = 1, \beta)$ is correctly specified, then

$$\begin{aligned} & E\left[\left\{\frac{R}{\Pr(R=1|A,X;\delta^*,\xi)} - 1\right\}\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta,\xi\}\right\}\right] \\ = \; & E\left[R\left\{\frac{1}{\Pr(R=1|A,X;\delta^*,\xi)} - 1\right\}\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta,\xi\}\right\}\right] \\ & -E\left[(1-R)\left\{l(A,Y) - E\{l(A,Y)|A,R=0;\beta,\xi\}\right\}\right]. \end{aligned}$$

It is straightforward to see the second term of the right hand side equals to zero when $\Pr(x, y|a; r = 1, \beta)$ is correctly specified. Thus, we only need to prove the first term also equals zero. By the definition of the odds ratio $\eta(r = 0, r_0, x, x_0|a)$, we have

$$r\{\frac{1}{\Pr(r=1|a,x)} - 1\} = r\eta(r=0,r_0,x,x_0|a)\frac{\Pr(r=0|a,x=0)}{\Pr(r=1|a,x=0)}.$$

By equation (4)

$$E\big\{R\eta(r=0,r_0,X,x_0|a)l(a,Y)|a\big\} = E\big\{R\eta(r=0,r_0,X,x_0|a)|a\big\}E\big\{l(a,Y)|a,r=0\big\},$$

and thus, $E\big[R\eta(r=0,r_0,X,x_0|a)\big\{l(a,Y) - E\{l(a,Y)|a,r=0\}\big\}\big|a\big] = 0.$

Hence,

$$E\left[\frac{\Pr(R=0|A,X=0)}{\Pr(R=1|A,X=0)}R\eta(r=0,r_0,X,x_0|A)\left\{l(A,Y) - E\{l(A,Y)|A,R=0\}\right\}\right] = 0,$$

proving the DR property of $\hat{\xi}^{dr}$.

Next, we show the DR property of $\hat{\mu}_a^{dr}$. If $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ are correctly specified,

$$
\begin{aligned}
& \hat{\mu}_a^{dr} \\
\xrightarrow{p} \; & E\Bigg[\frac{R}{\Pr(R|A, X; \delta, \xi)}\Big\{h_a^\dagger(A, X, Y) - E\{h_a^\dagger(A, X, Y)|A, R = 0; \beta^\dagger, \xi\}\Big\} \\
& + E\{h_a^\dagger(A, X, Y)|R = 0, A; \beta^\dagger, \xi\}\Bigg] \\
= \; & E\Bigg[\frac{1(A = a)RY}{\Pr(R|A, X; \delta, \xi)\Pr(A|X; \alpha, \delta, \xi)} + \frac{R}{\Pr(R|A, X; \delta, \xi)}\Big\{1 - \frac{1(A = a)}{\Pr(A|X; \alpha, \delta, \xi)}\Big\}g_a(X; \beta^\dagger) \\
& + \Big\{1 - \frac{R}{\Pr(R|A, X; \delta, \xi)}\Big\}E\{h_a^\dagger(A, X, Y)|A, R = 0; \beta^\dagger, \xi\}\Bigg] \\
= \; & E\Bigg[\frac{1(A = a)RY}{\Pr(R|A, X; \delta, \xi)\Pr(A|X; \alpha, \delta, \xi)}\Bigg] \\
= \; & \mu_a.
\end{aligned}
$$

If the baseline regression model $f(x, y|a, r = 1)$ is correctly specified, for any function $t(a, x, y)$, we have

$$
\begin{aligned}
E\big\{t(a, X, Y)|a, r = 0\big\} &= \frac{\iint \eta(r = 0, r_0, x, x_0|a) f(x, y|a, r = 1) t(a, x, y)\, dx\, dy}{E\{\eta(r = 0, r_0, x, x_0|a)|a, r = 1\}} \\
&= \frac{E\big\{\eta(r = 0, r_0, X, x_0|a) t(a, X, Y)|a, r = 1\big\}}{E\big\{\eta(r = 0, r_0, X, x_0|a)|a, r = 1\big\}} \\
&= \frac{E\big\{R\eta(r = 0, r_0, X, x_0|a) t(a, X, Y)|a\big\}}{E\big\{R\eta(r = 0, r_0, X, x_0|a)|a\big\}}.
\end{aligned}
$$

Hence, we have

$$E\big\{R\eta(r=0,r_0,X,x_0|a)t(A,X,Y)|A\big\} = E\big\{R\eta(r=0,r_0,X,x_0|A)|A\big\}E\big\{t(A,X,Y)|A,R=0\big\},$$

and thus, $E\big[R\eta(r=0,r_0,X,x_0|A)\big\{t(A,X,Y)-E\{t(A,X,Y)|A,R=0\big\}|A\big] =$

0. Therefore, set $t(A,X,Y) = h_a(A,X,Y)$ and we have

$$
\begin{aligned}
& \hat{\mu}_a^{dr} \\
\xrightarrow{p} \; & E\Bigg[\frac{R}{\Pr(R|A,X;\delta^*,\xi)}\Big\{h_a^*(A,X,Y) - E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Big\} \\
& + E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Bigg] \\
= \; & E\Bigg[R\Big\{\frac{1}{\Pr(R|A,X;\delta^*,\xi)} - 1\Big\}\Big\{h_a^*(A,X,Y) - E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Big\} \\
& + R\Big\{h_a^*(A,X,Y) - E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Big\} + E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Bigg] \\
= \; & E\Bigg[R\Big\{\frac{1}{\Pr(R|A,X;\delta^*,\xi)} - 1\Big\}\Big\{h_a^*(A,X,Y) - E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Big\} \\
& + Rh_a^*(A,X,Y) + (1-R)E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Bigg] \\
= \; & E\Bigg[Rh_a^*(A,X,Y) + (1-R)E\{h_a^*(A,X,Y)|A,R=0;\beta,\xi\}\Bigg] \\
= \; & E\Big\{h_a^*(A,X,Y)\Big\} \\
= \; & E\Bigg[1(A=a)\{Y - g_a(X;\beta)\}/\Pr(A|X;\alpha^*,\delta^*,\xi) + g_a(X;\beta)\Bigg] \\
= \; & E\Bigg[1(A=a)\{Y - E(Y|a,X,R=1;\beta)\}/\Pr(A|X;\alpha^*,\delta^*,\xi) + E(Y|a,X,R=1;\beta)\Bigg] \\
= \; & E\Bigg[1(A=a)\{Y - E(Y|a,X;\beta)\}/\Pr(A|X;\alpha^*,\delta^*,\xi) + E(Y|a,X;\beta)\Bigg] \\
= \; & E\Big\{E(Y|a,X;\beta)\Big\} = \mu_a.
\end{aligned}
$$

S8. Multiple Missing Patterns

In the manuscript, we only consider one missing pattern, i.e., we do not distinguish which components of the confounders are missing. However,

as mentioned by one reviewer, it is more reasonable to consider multiple missing patterns. Let $R = 1$ denote the confounders are fully observed, and $R = 2, \ldots, K$ denote different missing patterns, i.e., different combinations of confounders are missing. For p confounders, there are at most $2^p - 1$ patterns for at least one confounder missing, i.e., $K \leq 2^p$. We use the binary indicator $1(R = k)$ to denote if the missingness is k^{th} pattern for $k = 1, \ldots, K$. In this section, we show that our previous methods can be easily extended to the multiple missing pattern setting.

Under the multiple missingness setting described as above, Lemma 3 has the following extension and the derivation can be obtained from the definition of $\eta(r = k, r_0, x, x_0|a)$ given below.

Lemma 3*.

$$f(x, y|a, r = k) = \frac{\eta(r = k, r_0 = 1, x, x_0 = 0|a) f(x, y|a, r = 1)}{E\{\eta(r = k, r_0 = 1, X, x_0|a)|a, r = 1\}},$$

where $r_0 = 1$ *and* $x_0 = 0$ *and*

$$\eta(r = k, r_0, x, x_0|a) = \frac{\Pr(r = k|a, x) \Pr(r_0|a, x_0)}{\Pr(r_0|a, x) \Pr(r = k|a, x_0)}, \text{ for } k = 1, \ldots, K.$$

We use ξ_k to denote the parameter in the model $\eta(r = k, r_0 = 1, x, x_0 = 0|a; \xi_k)$ and let $\xi = (\xi_2, \ldots, \xi_K)$.

S8.1 IPW

With multiple missing patterns, the IPW estimator has the following representation

$$\hat{\mu}_a = \mathbb{P}_n\left\{1(A=a)1(R=1)Y/\Pr(A,R|X;\hat{\delta},\hat{\xi})\right\}.$$

Although the IPW estimator is in the same form as the IPW estimator in the binary missing types setting, the semi-parametric odds ratio representation for the joint distribution of A and R given X is modified as

$$\Pr(a,r|x) = \frac{\psi(a,a_0=1,r,r_0=1|x)\Pr(r|a_0=1,x)\Pr(a|r_0=1,x)}{\sum_{r=1}^{K}\sum_{a=0}^{1}\psi(a,a_0=1,r,r_0=1|x)\Pr(r|a_0=1,x)\Pr(a|r_0=1,x)}.$$

Following a similar argument as Propostion 1, the IPW estimator is consistent if $\Pr(a|r=1,x;\delta)$ and $\Pr(r=k|a,x;\xi_k)$ are correctly specified for all $k=1,\ldots,K$.

S8.2 Regression

With multiple missingness patterns, Lemma 4 can be extended as follows

Lemma 4*.

$$E\{l(a,Y)|a,r=k\} = \frac{E\{\eta(r=k,r_0,X,x_0|a)l(a,Y)|a,r=1\}}{E\{\eta(r=k,r_0,X,x_0|a)|a,r=1\}},\ \textit{for } k=2,\ldots,K.$$

The estimation equation 3.3 has the following expression

$$\mathbb{P}_n\left[1(R=k)\left\{l(A,Y)-E\{l(A,Y)|A,R=0;\hat{\beta},\hat{\xi}_k\}\right\}\right]=0,\ \text{for } k=2,\ldots,K. \tag{3.3*}$$

Hence, with estimated parameters $\hat{\beta}$, $\hat{\xi}_k$, $k=2,\ldots,K$, the regression estimator can be constructed as follows

$$\hat{\mu}_a^{reg}=\mathbb{P}_n\left[\sum_{k=2}^{K}1(R=k)E\{g_a(X;\hat{\beta})|a,r=k;\hat{\beta},\hat{\xi}_k^{reg}\}+1(R=1)g_a(X;\hat{\beta})\right].$$

S8.3 DR

Lemma 5 has the following extension

Lemma 5*.

$$\Pr(r=1|a,x) \;=\; \frac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x=0)+\sum_{k=2}^{K}\eta(r=k,r_0,x,x_0|a)\Pr(r=k|a,x=0)}.$$

Proof.

$$\begin{aligned}
\Pr(r=1|a,x) &= \frac{\Pr(r=1|a,x=0)}{\left\{\dfrac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x)}\right\}} \\
&= \frac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x=0)\cdot\dfrac{\Pr(r=1|a,x)+\sum_{k=2}^{K}\Pr(r=k|a,x)}{\Pr(r=1|a,x)}} \\
&= \frac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x=0)+\displaystyle\sum_{k=2}^{K}\Pr(r=k|a,x)\cdot\frac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x)}} \\
&= \frac{\Pr(r=1|a,x=0)}{\Pr(r=1|a,x=0)+\displaystyle\sum_{k=2}^{K}\eta(r=k,r_0,x,x_0|a)\Pr(r=k|a,x=0)}.
\end{aligned}$$

□

The estimation equation 3.5 for ξ can be extended as follows

$$\mathbb{P}_n\left[1(R=1)\{\frac{\Pr(R=k|A,X;\delta,\xi_k)}{\Pr(R=1|A,X;\delta,\xi)}\}\left\{l(A,X,Y)-E\{l(A,X,Y)|A,R=k;\beta,\xi_k\}\right\}\right]=0, \tag{3.5*}$$

for $k=2,\ldots,K$. The DR estimator can be constructed as

$$\begin{aligned}
\hat{\mu}_a^{dr} =&\mathbb{P}_n\Bigg[\sum_{k=2}^{K}1(R=1)\frac{\Pr(R=k|A,X;\hat{\delta},\hat{\xi}_k)}{\Pr(R=1|A,X;\hat{\delta},\hat{\xi})}\{\hat{h}_a(A,X,Y)-E\{\hat{h}_a(A,X,Y)|A,R=k;\hat{\beta},\hat{\xi}_k\}\Bigg] \\
&+1(R=1)\hat{h}_a(A,X,Y)+\sum_{k=2}^{K}1(R=k)E\{\hat{h}_a(A,X,Y)|A,R=k;\hat{\beta},\hat{\xi}_k\}\Bigg].
\end{aligned} \tag{3.6*}$$

Proposition 3*. Assume $\eta(r = k, r_0 = 1, x, x_0 = 0|a; \xi_k)$ $k = 1, \ldots, K$ are correctly specified, if either (i) $\Pr(r = k|a, x = 0; \delta)$, for all $k = 1, \ldots, K$ and $\Pr(a|r = 1, x; \alpha)$ or (ii) $f(x, y|a, r = 1; \beta)$ is correctly specified; then $\hat{\xi}^{dr}$ and $\hat{\mu}_a^{dr}$ are consistent for ξ and μ_a, where $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β, and $\hat{\xi}^{dr}$ and $\hat{\delta}$ are obtained by (3.5*) and $\hat{\mu}_a^{dr}$ is given in (3.6*).

The doubly robustness of $\hat{\xi}^{dr}$ can be verified same as before: the estimation equations (3.5*) hold if either $\Pr(r = k|a, x = 0; \delta)$, for $k = 1, \ldots, K$ or $f(x, y|a, r = 1; \beta)$ is correctly specified. Now we prove the consistency of $\hat{\mu}_a^{dr}$ assuming all the odds ratios $\eta(r = k, r_0 = 1, x, x_0 = 0|a; \xi_k)$ have been correctly specified with unknown parameter ξ_k for $k = 2, \ldots, K$. If $\Pr(r = k|a, x = 0; \delta)$ are correctly specified for $k = 2, \ldots, K$,

$$
\begin{aligned}
& \hat{\mu}_a^{dr} \\
\xrightarrow{p} \; & E\Bigg[\sum_{k=2}^{K} 1(R=1)\frac{\Pr(R=k|A,X;\delta,\xi_k)}{\Pr(R=1|A,X;\delta,\xi)}\{h_a^\dagger(A,X,Y) - E\{h_a^\dagger(A,X,Y)|A,R=k;\beta^\dagger,\xi_k\}\Bigg] \\
& +1(R=1)h_a^\dagger(A,X,Y) + \sum_{k=2}^{K} 1(R=k)E\{h_a^\dagger(A,X,Y)|A,R=k;\beta^\dagger,\xi_k\}\Bigg] \\
= \; & E\Bigg[\sum_{k=2}^{K}\Bigg[\{1(R=k) - \frac{1(R=1)\Pr(R=k|A,X;\delta,\xi_k)}{\Pr(R=1|A,X,\delta,\xi)}\}E\{h_a^\dagger(A,X,Y)|A,R=k;\beta^\dagger,\xi_k\}\Bigg] \\
& +\frac{1(R=1)h_a^\dagger(A,X,Y)}{\Pr(R=1|A,X;\delta,\xi)}\Bigg] \\
= \; & E\Bigg\{h_a^\dagger(A,X,Y)\Bigg\} \\
= \; & \mu_a.
\end{aligned}
$$

If $f(x,y|a,r=1,\beta)$ has been correctly specified,

$$
\begin{aligned}
& \hat{\mu}_a^{dr} \\
\xrightarrow{p} \; & E\Bigg[\sum_{k=2}^{K} 1(R=1)\frac{\Pr(R=k|A,X;\delta^*,\xi_k)}{\Pr(R=1|A,X;\delta^*,\xi)}\{h_a^*(A,X,Y) - E\{h_a^*(A,X,Y)|A,R=k;\beta,\xi_k\}\Bigg] \\
& +1(R=1)h_a^*(A,X,Y) + \sum_{k=2}^{K} 1(R=k)E\{h_a^*(A,X,Y)|A,R=k;\beta,\xi_k\}\Bigg] \\
= \; & E\Bigg[1(R=1)h_a^*(A,X,Y) + \sum_{k=2}^{K} 1(R=k)E\{h_a^*(A,X,Y)|A,R=k;\beta,\xi_k\}\Bigg] \\
= \; & E\Bigg\{h_a^*(A,X,Y)\Bigg\} \\
= \; & E\Bigg\{E(Y|a,X;\beta)\Bigg\} = \mu_a.
\end{aligned}
$$

S9. Semiparametric Inference for ATT

In this section, we extend our semiparametric estimators when the parameter of interest is the average treatment effect on the treated (ATT) $E(Y_1 - Y_0|A = 1)$. Since $E(Y_1|A = 1)$ is easy to estimate, we only extend our estimators for $\tau = E(Y_0|A = 1)$.

S9.1 IPW estimator

The IPW estimator for τ is

$$\hat{\tau}^{ipw} = \mathbb{P}_n\left[\frac{(1-A)RY\,\Pr(A=1|X;\hat{\alpha},\hat{\delta},\hat{\xi})}{\Pr(A=0,R=1|X;\hat{\alpha},\hat{\delta},\hat{\xi})\hat{\Pr}(A=1)}\right],$$

where $\hat{\alpha}$, $\hat{\delta}$ and $\hat{\xi}$ are defined in the Section 7 of the supplement. The consistency of this estimator follows from

$$\begin{aligned}
\hat{\tau}^{ipw} &\to E\left\{\frac{(1-A)RY\,\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0,R=1|X;\alpha,\delta,\xi)\,\Pr(A=1)}\right\} \\
&= E\left\{\frac{(1-A)RE(Y|A=0,X)\,\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0,R=1|X;\alpha,\delta,\xi)\,\Pr(A=1)}\right\} \\
&= E\left\{\frac{E(Y_0|X)\,\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=1)}\right\} \\
&= E\left\{\frac{E(Y_0|X)A}{\Pr(A=1)}\right\} \\
&= E\left\{\frac{Y_0A}{\Pr(A=1)}\right\} = E(Y_0|A=1).
\end{aligned}$$

S9.2 Regression estimator

Let $g_0(x) = E(Y|a = 0, x, r = 1; \beta)$, the regression estimator can be constructed as

$$\hat{\tau}^{reg} = \mathbb{P}_n\left[\frac{A}{\hat{\Pr}(A=1)}\left[(1-R)E\{g_0(X;\hat{\beta})|A, R=0; \hat{\beta}, \hat{\xi}\} + Rg_0(X;\hat{\beta})\right]\right].$$

When $f(x, y|a, r; \beta)$ is correctly specified, the proof of the consistency of the regression estimator is given as below

$$\begin{aligned}
\hat{\tau}^{reg} &\xrightarrow{p} E\left[\frac{A}{\Pr(A=1)}\left[(1-R)E\{g_0(X;\beta)|A, R=0\} + Rg_0(X;\beta)\right]\right] \\
&= E\left[\frac{A}{\Pr(A=1)}E\left[(1-R)E\{g_0(X)|A, R=0\} + Rg_0(X;\beta)\middle|A\right]\right] \\
&= E\left[\frac{A}{\Pr(A=1)}E\left\{(1-R)g_0(X;\beta) + Rg_0(X;\beta)\middle|A\right\}\right] \\
&= E\left\{\frac{A}{\Pr(A=1)}g_0(X;\beta)\right\} \\
&= E\left\{\frac{AY_0}{\Pr(A=1)}\right\} = E(Y_0|A=1).
\end{aligned}$$

S9.3 DR estimator

Let $h(A, X, Y; \alpha, \delta, \xi, \beta) = [(1-A)\Pr(A=1|X;\alpha,\delta,\xi)\{Y - g_0(X;\beta)\}]/\{\Pr(A = 0|X;\alpha,\delta,\xi)\Pr(A=1)\} + \{Ag_0(X;\beta)\}/\Pr(A=1)$. The DR estimator can be constructed as

$$\begin{aligned}
\hat{\tau}^{dr} &= \mathbb{P}_n\left[\frac{R}{\Pr(R=1|A, X; \hat{\delta}, \hat{\xi}^{dr})}\left[h(A, X, Y) - E\{h(A, X, Y)|A, R=0; \hat{\beta}, \hat{\xi}^{dr}\}\right]\right. \\
&\quad \left. + E\{h(A, X, Y)|A, R=0; \hat{\beta}, \hat{\xi}^{dr}\}\right].
\end{aligned}$$

Where α, β, δ and ξ are defined same as those in the Section 7 of the supplement.

Proposition 4*. Assume $\eta(r = 0, r_0 = 1, x, x_0 = 0|a; \xi)$ is correctly specified, if either (i) $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ or (ii) $f(x, y|a, r = 1; \beta)$ is correctly specified; then $\hat{\tau}^{dr}$ is consistent for $E(Y_0|A = 1)$, where $\hat{\alpha}$, $\hat{\beta}$ are MLE of α and β, $\hat{\xi}$ and $\hat{\delta}$ are obtained from the estimation equation (3.5).

The proof of the Proposition 4* is given as below

Proof. If $\Pr(r = 1|a, x = 0; \delta)$ and $\Pr(a|r = 1, x; \alpha)$ are correctly specified,

i.e., (i) holds,

$$
\begin{aligned}
& \hat{\tau}^{dr} \\
\xrightarrow{p} \; & E\left[\frac{R}{\Pr(R=1|A,X;\delta,\xi)}h^{\dagger}(A,X,Y)\right. \\
& \left.+\left\{1-\frac{R}{\Pr(R=1|A,X;\delta,\xi)}\right\}E\{h^{\dagger}(A,X,Y)|A,R=0;\beta^{\dagger},\xi\}\right] \\
= \; & E\left[h^{\dagger}(A,X,Y)E\left[\frac{R}{\Pr(R=1|A,X;\delta,\xi)}\middle|A,X\right]\right] \\
& +E\left[E\{h^{\dagger}(A,X,Y)|A,R=0;\beta^{\dagger},\xi\}E\left[\left\{1-\frac{R}{\Pr(R=1|A,X;\delta,\xi)}\right\}\middle|A,X\right]\right] \\
= \; & E\{h^{\dagger}(A,X,Y)\} \\
= \; & E\left[\frac{(1-A)\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0|X;\alpha,\delta,\xi)\Pr(A=1)}\{Y-g_0(X;\beta^{\dagger})\}+\frac{A}{\Pr(A=1)}g_0(X;\beta^{\dagger})\right] \\
= \; & E\left[\frac{(1-A)Y\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0|X;\alpha,\delta,\xi)\Pr(A=1)}\right. \\
& \left.+\left\{\frac{A}{\Pr(A=1)}-\frac{(1-A)\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0|X;\alpha,\delta,\xi)\Pr(A=1)}\right\}g_0(X;\beta^{\dagger})\right] \\
= \; & E\left[\frac{(1-A)Y_0\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0|X;\alpha,\delta,\xi)\Pr(A=1)}\right] \\
& +E\left[g_0(X;\beta^{\dagger})E\left[\left\{\frac{A}{\Pr(A=1)}-\frac{(1-A)\Pr(A=1|X;\alpha,\delta,\xi)}{\Pr(A=0|X;\alpha,\delta,\xi)\Pr(A=1)}\right\}\middle|X\right]\right] \\
= \; & E\left\{\frac{AY_0}{\Pr(A=1)}\right\} \\
= \; & E(Y_0|A=1).
\end{aligned}
$$

If $f(x,y|a,r=1;\beta)$ is correctly specified, i.e., (ii) holds,

$$
\begin{aligned}
& \hat{\mu}_a^{dr} \\
\xrightarrow{p} \; & E\left[\frac{R}{\Pr(R|A,X;\delta^*,\xi)}\left\{h^*(A,X,Y) - E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right\}\right. \\
& \left. + E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right] \\
= \; & E\left[R\left\{\frac{1}{\Pr(R|A,X;\delta^*,\xi)} - 1\right\}\left\{h^*(A,X,Y) - E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right\}\right. \\
& \left. + R\left\{h^*(A,X,Y) - E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right\} + E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right] \\
= \; & E\left[R\left\{\frac{1}{\Pr(R|A,X;\delta^*,\xi)} - 1\right\}\left\{h^*(A,X,Y) - E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right\}\right. \\
& \left. + Rh^*(A,X,Y) + (1-R)E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right] \\
= \; & E\left[Rh^*(A,X,Y) + (1-R)E\{h^*(A,X,Y)|A,R=0;\beta,\xi\}\right] \\
= \; & E\left\{h^*(A,X,Y)\right\} \\
= \; & E\left[\frac{(1-A)\Pr(A=1|X;\alpha^*,\delta^*,\xi)}{\Pr(A=0|X;\alpha^*,\delta^*,\xi)\Pr(A=1)}\{Y - g_0(X;\beta)\} + \frac{A}{\Pr(A=1)}g_0(X;\beta)\right] \\
= \; & E\left[\frac{(1-A)\Pr(A=1|X;\alpha^*,\delta^*,\xi)}{\Pr(A=0|X;\alpha^*,\delta^*,\xi)\Pr(A=1)}\{Y_0 - g_0(X;\beta)\} + \frac{A}{\Pr(A=1)}g_0(X;\beta)\right] \\
= \; & E\left[E\left[\frac{(1-A)\Pr(A=1|X;\alpha^*,\delta^*,\xi)}{\Pr(A=0|X;\alpha^*,\delta^*,\xi)\Pr(A=1)}\{Y_0 - g_0(X;\beta)\}\middle|X\right]\right] \\
& + E\left\{\frac{A}{\Pr(A=1)}g_0(X;\beta)\right\} \\
= \; & E\left[\frac{(1-A)\Pr(A=1|X;\alpha^*,\delta^*,\xi)}{\Pr(A=0|X;\alpha^*,\delta^*,\xi)\Pr(A=1)}E\left[\{Y_0 - g_0(X;\beta)\}\middle|X\right]\right] \\
& + E\left\{\frac{A}{\Pr(A=1)}g_0(X;\beta)\right\} \\
= \; & E\left\{\frac{A}{\Pr(A=1)}g_0(X;\beta)\right\} \\
= \; & E\left\{\frac{AY_0}{\Pr(A=1)}\right\} = E(Y_0|A=1).
\end{aligned}
$$

□

S10. Additional Table and Figure

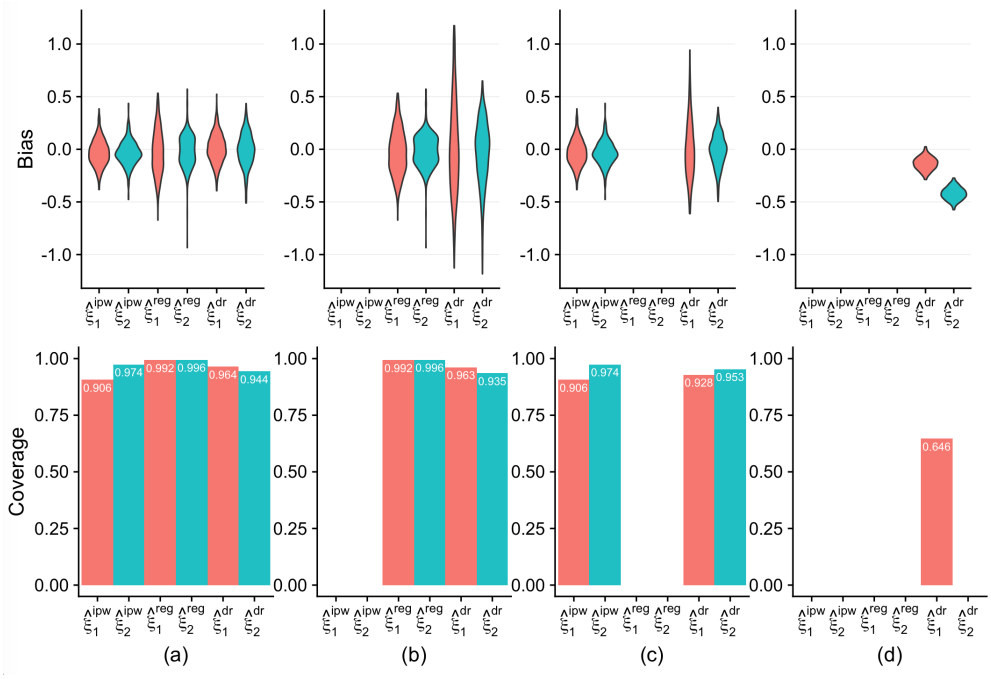
Table 1: Bias, coverage of the 95% Wald type confidence intervals of the IPW, the regression and the DR estimators for μ_a when (a) both baseline propensity and outcome models are correctly specified, (b) only baseline outcome model is correctly specified, (c) only baseline propensity score models are correctly specified, and (d) none of the baseline models is correctly specified

		IPW		Regression		DR	
		μ_0	μ_1	μ_0	μ_1	μ_0	μ_1
Bias	(a)	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
	(b)	0.18	0.05	<0.01	<0.01	<0.01	<0.01
	(c)	<0.01	<0.01	0.60	0.66	<0.01	<0.01
	(d)	0.18	0.05	0.60	0.66	0.08	0.03
Coverage	(a)	0.95	0.95	0.96	0.95	0.94	0.92
	(b)	<0.01	0.50	0.96	0.95	0.95	0.95
	(c)	0.95	0.95	0	0	0.99	0.98
	(d)	<0.01	0.50	0	0	0.29	0.88

Table 2: Point estimates and standard deviation [in brackets] for models in SO_2 emission data.

	IPW Propensity	Regression	DR Propensity
α_1	-3.51[0.13]		-3.51[0.14]
α_2	-1.54[0.24]		-1.54[0.24]
α_3	1.00[0.02]		1.00[0.03]
α_4	3.55[0.19]		3.55[0.21]
δ_1	-0.17[0.29]		-0.02[0.36]
δ_2	5.80[0.09]		7.74[0.36]
β_1		-394.94[25.58]	-394.83[25.57]
β_2		383.80[46.20]	384.11[46.20]
β_3		377.82[14.08]	377.87[14.09]
β_4		1177.64[58.24]	1177.08[58.25]
β_5		-928.10[25.03]	-928.38[25.03]
ξ_1	1.42[0.26]	1.85[0.39]	4.40[5.43]
ξ_2	1.00[1.76]	-4.49[0.44]	-1.00[0.41]
ξ_3	0.69[1.46]	8.45[1.27]	1.76[7.20]
μ_1	435.35[19.05]	265.52[4.80]	513.67[29.49]
μ_0	1029.95[14.47]	914.50[9.64]	1024.95[15.43]
$\mu_1 - \mu_0$	-594.60[26.36]	-648.99[4.85]	-511.28[32.45]

Figure 2: Bias, coverage of 95% Wald type confidence intervals of parameters ξ in the selection bias function $\eta(r = 0, r_0, x, x_0; \xi)$ when (a) both the baseline propensity score and outcome models are correctly specified, (b) only the baseline outcome model is correctly specified, (c) only the baseline propensity score models are correctly specified, and (d) none of the baseline models is correctly specified.



References

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.