Estimation of local treatment effects under the binary instrumental variable model

BY LINBO WANG

Department of Statistical Sciences, University of Toronto, Toronto, Ontario M5S 3G3, Canada

linbo.wang@utoronto.ca

YUEXIA ZHANG

Department of Computer and Mathematical Sciences, University of Toronto, Toronto, Ontario M1C 1A4, Canada

yuexia.zhang@utoronto.ca

THOMAS S. RICHARDSON

Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.

thomasr@u.washington.edu

AND JAMES M. ROBINS

Department of Epidemiology, Harvard T. H. Chan School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.

robins@hsph.harvard.edu

Background

Unmeasured confounding occurs in observational studies as well as imperfect randomized controlled trials.

An instrumental variable is a pre-treatment covariate that is associated with the outcome only through its effect on the treatment.

Instrumental Variable methods

- average treatment effect

  relies on untestable homogeneity assumptions involving unmeasured confounders

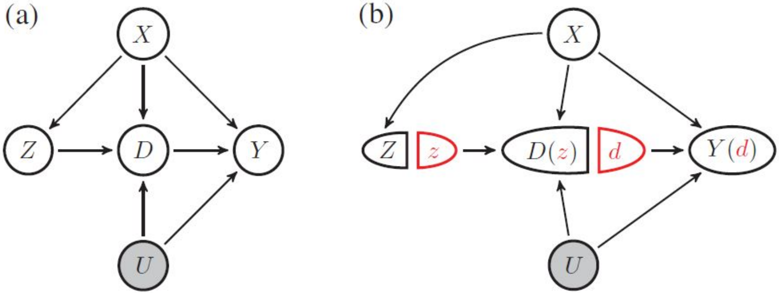
- local average treatment effect

  can be nonparametrically identified under a certain monotonicity assumption

Framework

Causal effect estimation with a binary exposure indicator D and a binary outcome Y

Suppose that the effect of D on Y is subject to confounding by observed variables X as well as unobserved variables U.



Assumptions for Z

- (Exclusion restriction). For all z and z', $Y(z, d) = Y(z', d) \equiv Y(d)$ almost surely.
- (Independence). We have that $Z \perp\!\!\!\perp (Y(d), D(z)) \mid X, d = 0, 1, z = 0, 1$.
- (Instrumental variable relevance). We have that $\mathrm{pr}\{D(1) = 1 \mid X\} \neq \mathrm{pr}\{D(0) = 1 \mid X\}$ almost surely.
- (Positivity). There exists $\sigma > 0$ such that $\sigma < \mathrm{pr}(Z = 1 \mid X) < 1 - \sigma$ almost surely.
- (Monotonicity). We have that $D(1) \geqslant D(0)$ almost surely.

Principal strata

Table 1. *Principal strata t_D based on* $\{D(1), D(0)\}$

$D(1)$	$D(0)$	Principal stratum	Abbreviation
1	1	Always taker	AT
1	0	Complier	CO
0	1	Defier	DE
0	0	Never taker	NT

Target

We are interested in estimating the conditional treatment effects in the complier stratum on the additive and multiplicative scales

$$\begin{aligned}\mathrm{LATE}(X) &= E\{Y(1) - Y(0) \mid D(1) > D(0), X\},\\ \mathrm{MLATE}(X) &= E\{Y(1) \mid D(1) > D(0), X\} / E\{Y(0) \mid D(1) > D(0), X\}.\end{aligned}$$

Abadie (2002, Lemma 2.1) showed that under Assumptions 1-5, the local average treatment effects are identifiable as

$$\begin{aligned}\mathrm{LATE}(X) &= \delta^{\mathrm{L}}(X) - \frac{E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X)}{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)},\\ \mathrm{MLATE}(X) &= \delta^{\mathrm{M}}(X) - \frac{E(YD \mid Z = 1, X) - E(YD \mid Z = 0, X)}{E\{Y(1 - D) \mid Z = 1, X\} - E\{Y(1 - D) \mid Z = 0, X\}}.\end{aligned}$$

Frölich (2007)

Define the conditional mean functions $m_z(x) = \mathrm{E}[Y \mid X = x, Z = z]$ and $\mu_z(x) = \mathrm{E}[D \mid X = x, Z = z]$ and let $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ be corresponding nonparametric regression estimators thereof. A nonparametric imputation estimator of γ is

$$\frac{\sum_i \left(\hat{m}_1\left(X_i\right) - \hat{m}_0\left(X_i\right)\right)}{\sum_i \left(\hat{\mu}_1\left(X_i\right) - \hat{\mu}_0\left(X_i\right)\right)},$$

where the expected values $\mathrm{E}[Y \mid X, Z]$ and $\mathrm{E}[D \mid X, Z]$ are imputed for each observation X_i. Using the observed values Y_i and D_i as estimates of $\mathrm{E}\left[Y_i \mid X_i, Z = z\right]$ and $\mathrm{E}\left[D_i \mid X_i, Z = z\right]$, whenever $z = Z_i$, gives the conditional LATE estimator $\hat{\gamma}$ as

$$\hat{\gamma} = \frac{\sum_{i:Z_i=1}\left(Y_i - \hat{m}_0\left(X_i\right)\right) - \sum_{i:Z_i=0}\left(Y_i - \hat{m}_1\left(X_i\right)\right)}{\sum_{i:Z_i=1}\left(D_i - \hat{\mu}_0\left(X_i\right)\right) - \sum_{i:Z_i=0}\left(D_i - \hat{\mu}_1\left(X_i\right)\right)}.$$

Abadie (2003)

Consider the function of (D, X) that is equal to $\mathrm{E}[\,Y_0 \mid X, D_1 > D_0\,]$ if $D = 0$, and is equal to $\mathrm{E}[\,Y_1 \mid X, D_1 > D_0\,]$ if $D = 1$. This function describes average treatment responses for any group of compliers defined by some value for the covariates. Abadie refers to this function as the Local Average Response Function (LARF).

(i) estimate a parameterization of the LARF by Least Squares (LS),

(ii) specify a parametric distribution for $p(\,Y \mid X, D, D_1 > D_0\,)$ and estimate the parameters of the LARF by Maximum Likelihood. (ML)

Ogburn et al. (2015)

Ogburn et al. (2015) proposed doubly robust estimators based on direct parameterization of the target functional $\delta^{\mathrm{L}}(X)$.

Given a correct model $\delta^{\mathrm{L}}(X;\alpha)$, their estimators are consistent and asymptotically normal for the parameter of interest α if either the instrumental density model $\mathrm{pr}(Z=1 \mid X;\gamma)$ or another nuisance model $E\big(Y - D \times \delta^{\mathrm{L}}(X) \mid X;\beta\big)$ is correctly specified.

However, the nuisance model $E\big(Y - D \times \delta^{\mathrm{L}}(X) \mid X;\beta\big)$ is variation-dependent on the target model $\delta^{\mathrm{L}}(X;\alpha)$. Yet it is often not possible for $\delta^{\mathrm{L}}(X;\alpha)$ and $E\big(Y - D \times \delta^{\mathrm{L}}(X) \mid X;\beta\big)$ to be correct simultaneously.

Wang & Tchetgen Tchetgen (2018)

Wang & Tchetgen Tchetgen (2018) studied rstimating $E_X\left\{\delta^{\mathrm{L}}(X)\right\}$. They proposed alternative nuisance models that are variation-independent of $\delta^{\mathrm{L}}(X;\alpha)$, including a model for $\delta^D(X) \equiv E(D \mid Z = 1, X) - E(D \mid Z = 0, X)$.

As long as the models for $\delta^{\mathrm{L}}(X)$ and $\delta^D(X)$ both lie in their respective parameter spaces,

$$E(Y \mid Z = 1, X) - E(Y \mid Z = 0, X) = \delta^{\mathrm{L}}(X) \times \delta^D(X)$$

also lies in its parameter space $[-1, 1]$.

They derived a maximum likelihood estimator and a truly doubly robust estimator for $\delta^{\mathrm{L}}(X)$.

Contributions

Propose novel estimating procedures for both the additive and the multiplicative local average treatment effects with a binary outcome.

- the posited models are variation-independent, and hence congenial to each other
- the resulting estimates lie in the natural nontrivial parameter space
- directly parameterize the local average treatment effect curves to improve interpretability and reduce the risk of model misspecification
- allow for efficient and truly doubly robust estimation of the causal parameter of interest

A novel parameterization

our goal: to find nuisance models such that

- they are variation-independent of each other
- they are variation-independent of $\delta_L(X;\alpha)$ and $\delta_M(X;\alpha)$
- there exists a bijection between the observed-data likelihood on $pr(D = d, Y = y \mid Z = z, X)$ and the combination of target and nuisance models

A novel parameterization

our goal: to find nuisance models such that

$$\Delta = \{p_x(d, y \mid z) \geqslant 0 : \sum_{d,y} p_x(d, y \mid z) = 1$$

$$p_x(1, y \mid 1) \geqslant p_x(1, y \mid 0), p_x(0, y \mid 1) \leqslant p_x(0, y \mid 0), y = 0, 1\}$$

$p_x(d, y \mid z), d, y, z = 0, 1$ are not variation-independent of each other.

We seek to model components $p(\mathrm{AT}; X), p(\mathrm{NT}; X)$ and $p(\mathrm{CO}; X)$

as well as $p(Y(1) \mid \mathrm{AT}; X)$, $p(Y(1) \mid \mathrm{CO}; X)$ and $p(Y(0) \mid \mathrm{CO}; X)$.

$$p(\mathrm{AT}; X) + p(\mathrm{NT}; X) + p(\mathrm{CO}; X) = 1.$$

Moreover, they do not contain our target function $\delta^{\mathrm{L}}(X)$ or $\delta^{\mathrm{M}}(X)$

Theorem 1

$$\phi_1(X) \equiv \mathrm{pr}\,(t_D = \mathrm{CO} \mid X) = \mathrm{pr}(D = 1 \mid Z = 1, X) - \mathrm{pr}(D = 1 \mid Z = 0, X),$$

$$\phi_2(X) \equiv \mathrm{pr}\,(t_D = \mathrm{AT} \mid t_D \in \{\mathrm{AT}, \mathrm{NT}\}, X) = \frac{\mathrm{pr}(D = 1 \mid Z = 0, X)}{\mathrm{pr}(D = 1 \mid Z = 0, X) + \mathrm{pr}(D = 0 \mid Z = 1, X)},$$

$$\phi_3(X) \equiv \mathrm{pr}\,(Y = 1 \mid t_D = \mathrm{NT}, X) = \mathrm{pr}(Y = 1 \mid D = 0, Z = 1, X),$$

$$\phi_4(X) \equiv \mathrm{pr}\,(Y = 1 \mid t_D = \mathrm{AT}, X) = \mathrm{pr}(Y = 1 \mid D = 1, Z = 0, X),$$

$$\mathrm{OP}^{CO}(X) \equiv \frac{E\{Y(1) \mid t_D = \mathrm{CO}, X\}\, E\{Y(0) \mid t_D = \mathrm{CO}, X\}}{[1 - E\{Y(1) \mid t_D = \mathrm{CO}, X\}]\,[1 - E\{Y(0) \mid t_D = \mathrm{CO}, X\}]},$$

Proof of Theorem 1

$$\{\mathrm{pr}(D=d, Y=y \mid Z=z, X), d, y, z \in \{0,1\}\}$$
$$\to \left\{\theta(X), \phi_1(X), \phi_2(X), \phi_3(X), \phi_4(X), \mathrm{OP}^{\mathrm{CO}}(X)\right\}$$

$$\phi_1(X) = 1 - p_X(0,0 \mid 1) - p_X(0,1 \mid 1) - p_X(1,0 \mid 0) - p_X(1,1 \mid 0)$$
$$\phi_2(X) = \{p_X(1,0 \mid 0) + p_X(1,1 \mid 0)\} / \{p_X(0,0 \mid 1) + p_X(0,1 \mid 1) + p_X(1,0 \mid 0) + p_X(1,1 \mid 0)\};$$
$$\phi_3(X) = p_X(0,1 \mid 1) / \{p_X(0,0 \mid 1) + p_X(0,1 \mid 1)\};$$
$$\phi_4(X) = p_X(1,1 \mid 0) / \{p_X(1,0 \mid 0) + p_X(1,1 \mid 0)\};$$
$$\begin{aligned} OP^{CO}(X) &= \frac{E\{Y(1)I(t_D = CO) \mid X\}\, E\{Y(0)I(t_D = CO) \mid X\}}{[p(CO;X) - E\{Y(1)I(t_D = CO) \mid X\}]\,[p(CO;X) - E\{Y(0)I(t_D = CO) \mid X\}]} \\ &= \frac{\{p(CO,HE;X) + p(CO,AR;X)\}\{p(CO,HU;X) + p(CO,AR;X)\}}{\{p(CO,NR;X) + p(CO,HU;X)\}\{p(CO,NR;X) + p(CO,HE;X)\}} \\ &= \frac{\{p_X(1,1 \mid 1) - p_X(1,1 \mid 0)\}\{p_X(0,1 \mid 0) - p_X(0,1 \mid 1)\}}{\{p_X(1,0 \mid 1) - p_X(1,0 \mid 0)\}\{p_X(0,0 \mid 0) - p_X(0,0 \mid 1)\}} \end{aligned}$$

Proof of Theorem 1

To show the map is a bijection, for each realization of X, let $c = (c_0, \ldots, c_5)$ be a vector in $\mathcal{D} \times [0,1]^4 \times [0, +\infty)$. We need to show there is one and only one $p \in \Delta$ such that

$$\{\phi_1(X), \ldots, \phi_4(X)\} = (c_1, \ldots, c_4)$$

and

$$\left\{\theta(X), OP^{CO}(X)\right\} = (c_0, c_5) .$$

Proof of Theorem 1

$$p(0,1 \mid 1) = (1-c_1)(1-c_2)c_3; p(0,0 \mid 1) = (1-c_1)(1-c_2)(1-c_3)$$
$$p(1,1 \mid 0) = (1-c_1)c_2c_4; p(1,0 \mid 0) = (1-c_1)c_2(1-c_4)$$

According to Richardson et al. (2017)

$$E\{Y(1) \mid t_D = CO\} = \frac{p(1,1 \mid 1) - p(1,1 \mid 0)}{p(CO)} = f_1(c_0, c_5)$$
$$E\{Y(0) \mid t_D = CO\} = \frac{p(0,1 \mid 0) - p(0,1 \mid 1)}{p(CO)} = f_0(c_0, c_5)$$

where $p(CO) = 1 - p(0,0 \mid 1) - p(0,1 \mid 1) - p(1,0 \mid 0) - p(1,1 \mid 0)$ and $f_d(c_0, c_5), d = 0, 1$ are known smooth functions of c_0, c_5 that take values between 0 and 1 .

Proof of Theorem 1

The functional form

$$f_0\left(c_0, c_5\right) = \begin{cases} \frac{1}{2(c_5-1)}\left[c_5\left(2-c_0\right)+c_0-\left\{c_0^2\left(c_5-1\right)^2+4c_5\right\}^{1/2}\right], & \theta=\delta^L \\ \frac{1}{2c_0(1-c_5)}\left[-\left(c_0+1\right)c_5+\left\{c_5^2\left(c_0-1\right)^2+4c_0c_5\right\}^{1/2}\right], & \theta=\delta^M \end{cases}$$

and

$$f_1\left(c_0, c_5\right) = \begin{cases} f_0\left(c_0, c_5\right)+c_0, & \theta=\delta^L \\ f_0\left(c_0, c_5\right)c_0, & \theta=\delta^M \end{cases}.$$

Proof of Theorem 1

$$
\begin{aligned}
p(0,1 \mid 1) = (1-c_1)(1-c_2)\,c_3; p(0,0 \mid 1) = (1-c_1)(1-c_2)(1-c_3)\,;\\
p(1,1 \mid 0) = (1-c_1)\,c_2 c_4; p(1,0 \mid 0) = (1-c_1)\,c_2\,(1-c_4)\,;\\
p(1,1 \mid 1) &= f_1(c_0, c_5)\,c_1 + p(1,1 \mid 0);\\
p(0,1 \mid 0) &= f_0(c_0, c_5)\,c_1 + p(0,1 \mid 1);\\
p(1,0 \mid 1) &= 1 - p(0,0 \mid 1) - p(0,1 \mid 1) - p(1,1 \mid 1);\\
p(0,0 \mid 0) &= 1 - p(0,1 \mid 0) - p(1,0 \mid 0) - p(1,1 \mid 0).
\end{aligned}
$$

Proof of Theorem 1

We now only need to show $\boldsymbol{p}$ lies in Δ.

$$p(1,1 \mid 1) = f_1(c_0, c_5)\, c_1 + p(1,1 \mid 0) \leq c_1 + (1 - c_1)\, c_2 c_4 \leq 1$$

$$\begin{aligned}
p(0,1 \mid 0) &= f_0(c_0, c_5)\, c_1 + p(0,1 \mid 1) \leq c_1 + (1 - c_1)(1 - c_2)\, c_3 \leq 1; \\
p(1,0 \mid 1) &= 1 - p(0,0 \mid 1) - p(0,1 \mid 1) - p(1,1 \mid 1) \\
&= 1 - (1 - c_1)(1 - c_2) - f_1(c_0, c_5)\, c_1 - p(1,1 \mid 0) \\
&\geq 1 - (1 - c_1)(1 - c_2) - c_1 - (1 - c_1)\, c_2 c_4 \\
&= (1 - c_1)\, c_2 (1 - c_4) \geq 0 \\
p(0,0 \mid 0) &= 1 - p(0,1 \mid 0) - p(1,0 \mid 0) - p(1,1 \mid 0) \\
&= 1 - f_0(c_0, c_5)\, c_1 - p(0,1 \mid 1) - (1 - c_1)\, c_2 \\
&\geq 1 - c_1 - (1 - c_1)(1 - c_2)\, c_3 - (1 - c_1)\, c_2 \\
&= (1 - c_1)(1 - c_2)(1 - c_3) \geq 0.
\end{aligned}$$

Proof of Theorem 1

$$
\begin{aligned}
p(1,1 \mid 1) &= f_1\left(c_0, c_5\right) c_1 + p(1,1 \mid 0) \geq p(1,1 \mid 0) \\
p(1,0 \mid 1) &\geq \left(1 - c_1\right) c_2 \left(1 - c_4\right) = p(1,0 \mid 0) \\
p(0,1 \mid 0) &= f_0\left(c_0, c_5\right) c_1 + p(0,1 \mid 1) \geq p(0,1 \mid 1) \\
p(0,0 \mid 0) &\geq \left(1 - c_1\right)\left(1 - c_2\right)\left(1 - c_3\right) = p(0,0 \mid 1)
\end{aligned}
$$

Doubly Robust Estimation

Let $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_4, \hat{\eta}$ and $\hat{\gamma}$ be the maximum likelihood estimators of $\alpha, \beta_1, \ldots, \beta_4, \eta$ and γ, respectively. Also let

$$H(Y, D, X; \alpha) = \begin{cases} Y - D\theta(X; \alpha), & \theta(X) = \delta^{\mathrm{L}}(X) \\ Y\theta(X; \alpha)^{-D}, & \theta(X) = \delta^{\mathrm{M}}(X) \end{cases}$$

Theorem 2

Let $\hat{\alpha}_{\mathrm{dr}}$ solve the estimating equation

$$\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \hat{\gamma})} [H(Y, D, X; \alpha) - \hat{E}\{H(Y, D, X; \alpha) \mid X\}] = 0$$

where $\mathbb{P}_n$ denotes the empirical mean operator, $\omega(X)$ is an arbitrary measurable function of X,

$$f(Z \mid X; \hat{\gamma}) = \{\mathrm{pr}(Z = 1 \mid X; \hat{\gamma})\}^Z \{1 - \mathrm{pr}(Z = 1 \mid X; \hat{\gamma})\}^{1-Z}$$

$$\hat{E}\{H(Y, D, X; \alpha) \mid X\}$$
$$= \begin{cases} \hat{f}_0 \hat{\phi}_1 + \left(1 - \hat{\phi}_1\right)\left(1 - \hat{\phi}_2\right)\hat{\phi}_3 + \left(1 - \hat{\phi}_1\right)\hat{\phi}_2 \hat{\phi}_4 - \theta \left(1 - \hat{\phi}_1\right)\hat{\phi}_2, & \theta(X) = \delta^{\mathrm{L}}(X), \\ \hat{f}_0 \hat{\phi}_1 + \left(1 - \hat{\phi}_1\right)\hat{\phi}_2 \hat{\phi}_4 \theta^{-1} + \left(1 - \hat{\phi}_1\right)\left(1 - \hat{\phi}_2\right)\hat{\phi}_3, & \theta(X) = \delta^{\mathrm{M}}(X), \end{cases}$$

with

$$\hat{f}_0 = \begin{cases} \frac{1}{2(\hat{OP}-1)} \left[\hat{OP}(2-\theta) + \theta - \left\{\theta^2(\hat{OP}-1)^2 + 4\hat{OP}\right\}^{1/2}\right], & \theta(X) = \delta^{\mathrm{L}}(X), \\ \frac{1}{2\theta(1-\hat{OP})} \left[-(\theta+1)\hat{OP} + \left\{\hat{OP}^2(\theta-1)^2 + 4\theta\hat{OP}\right\}^{1/2}\right], & \theta(X) = \delta^{\mathrm{M}}(X), \end{cases}$$

Proof of Theorem 2: $\theta(X) = \delta^L(X)$

In this case $H(Y, D, X) = Y - D\theta(X)$.

$$
\begin{aligned}
&E\{H(Y,D,X) \mid Z=1,X\} - E\{H(Y,D,X) \mid Z=0,X\} \\
&\quad = E(Y \mid Z=1,X) - E(D \mid Z=1,X)\theta(X) - E(Y \mid Z=0,X) + E(D \mid Z=0,X)\theta(X) \\
&\qquad = E(Y \mid Z=1,X) - E(Y \mid Z=0,X) \\
&\qquad - \{E(D \mid Z=1,X) - E(D \mid Z=0,X)\} \frac{E(Y \mid Z=1,X) - E(Y \mid Z=0,X)}{E(D \mid Z=1,X) - E(D \mid Z=0,X)} = 0
\end{aligned}
$$

Thus, $E\{H(Y,D,X) \mid X\} = E\{H(Y,D,X) \mid Z=0,X\}$.

Proof of Theorem 2: $\theta(X) = \delta^L(X)$

$$
\begin{aligned}
&E\{H(Y,D,X) \mid X\} = E\{H(Y,D,X) \mid Z=0,X\} \\
=&E(Y \mid Z=0,X) - E(D \mid Z=0,X)\theta(X) \\
=&P(Y=1 \mid Z=0,X) - P(D=1 \mid Z=0,X)\theta(X) \\
=&P(D=1, Y=1 \mid Z=0,X) + P(D=0, Y=1 \mid Z=0,X) - P(D=1 \mid Z=0,X)\theta(X) \\
=&p_X(1,1 \mid 0) + p_X(0,1 \mid 0) - \theta(X)\{1-\phi_1(X)\}\phi_2(X) \\
=&\{1-\phi_1(X)\}\phi_2(X)\phi_4(X) + f_0\left\{\theta(X), OP^{CO}(X)\right\}\phi_1(X) \\
&+\{1-\phi_1(X)\}\{1-\phi_2(X)\}\phi_3(X) - \theta(X)\{1-\phi_1(X)\}\phi_2(X).
\end{aligned}
$$

Proof of Theorem 2: $\theta(X) = \delta^M(X)$

In this case $H(Y, D, X) = Y\theta(X)^{-D}$.

$$
\begin{aligned}
& E\{H(Y,D,X) \mid Z=1,X\} - E\{H(Y,D,X) \mid Z=0,X\} \\
=& E\left\{Y\theta(X)^{-D} \mid Z=1,X\right\} - E\left\{Y\theta(X)^{-D} \mid Z=0,X\right\} \\
=& \theta(X)^{-1}P(Y=1,D=1 \mid Z=1,X) + P(Y=1,D=0 \mid Z=1,X) \\
& \ - \theta(X)^{-1}P(Y=1,D=1 \mid Z=0,X) - P(Y=1,D=0 \mid Z=0,X) \\
=& \theta(X)^{-1}\{P(Y=1,D=1 \mid Z=1,X) - P(Y=1,D=1 \mid Z=0,X)\} \\
& \ + P(Y=1,D=0 \mid Z=1,X) - P(Y=1,D=0 \mid Z=0,X) \\
=& -\frac{P(Y=1,D=0 \mid Z=1,X) - P(Y=1,D=0 \mid Z=0,X)}{P(Y=1,D=1 \mid Z=1,X) - P(Y=1,D=1 \mid Z=0,X)} \\
& \ \times \{P(Y=1,D=1 \mid Z=1,X) - P(Y=1,D=1 \mid Z=0,X)\} \\
& \ + P(Y=1,D=0 \mid Z=1,X) - P(Y=1,D=0 \mid Z=0,X) \\
=& 0
\end{aligned}
$$

Proof of Theorem 2: $\theta(X) = \delta^M(X)$

Thus, $E\{H(Y,D,X) \mid X\} = E\{H(Y,D,X) \mid Z=0,X\}$

$$
\begin{aligned}
E\{H(Y,D,X) \mid X\} &= E\{H(Y,D,X) \mid Z=0,X\} \\
=& E\left\{Y\theta(X)^{-D} \mid Z=0,X\right\} \\
=& \theta(X)^{-1} P(Y=1,D=1 \mid Z=0,X) + P(Y=1,D=0 \mid Z=0,X) \\
=& \{1-\phi_1(X)\}\,\phi_2(X)\phi_4(X)\theta(X)^{-1} \\
& + f_0\left\{\theta(X), OP^{CO}(X)\right\}\phi_1(X) + \{1-\phi_1(X)\}\,\{1-\phi_2(X)\}\,\phi_3(X)
\end{aligned}
$$

Proof of Theorem 2

If the model for $E\{H(Y,D,X;\alpha) \mid X\}$ is correctly specified, but the model for $f(Z \mid X;\gamma)$ may be mis-specified, then $\hat{\gamma} \xrightarrow{p} \gamma^*$ under some regularity conditions (White, 1982), where γ^* is not necessarily equal to γ. Furthermore,

$$
\begin{aligned}
& E\left(\mathbb{P}_n\omega(X)\frac{2Z-1}{f(Z \mid X;\gamma^*)}[H(Y,D,X;\alpha) - E\{H(Y,D,X;\alpha) \mid Z=0,X\}]\right) \\
&= E\left(E\left\{\mathbb{P}_n\omega(X)\frac{2Z-1}{f(Z \mid X;\gamma^*)}[H(Y,D,X;\alpha) - E\{H(Y,D,X;\alpha) \mid Z=0,X\}] \mid Z,X\right\}\right) \\
&= E\left(\mathbb{P}_n\omega(X)\frac{2Z-1}{f(Z \mid X;\gamma^*)}[E\{H(Y,D,X;\alpha) \mid Z,X\} - E\{H(Y,D,X;\alpha) \mid Z=0,X\}]\right) \\
&= 0
\end{aligned}
$$

Proof of Theorem 2

If the model for $f(Z \mid X; \gamma)$ is correctly specified, but the model for $E\{H(Y, D, X; \alpha) \mid X\}$ may be mis-specified. Assume $\widehat{E}\{H(Y, D, X; \alpha) \mid X\} - E^{*}\{H(Y, D, X; \alpha) \mid X\} \xrightarrow{p} 0$.

$$
\begin{aligned}
& E\left(\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} \left[H(Y, D, X; \alpha) - E^{*}\{H(Y, D, X; \alpha) \mid X\}\right]\right) \\
= & E\left\{\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} H(Y, D, X; \alpha)\right\} - E\left[\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} E^{*}\{H(Y, D, X; \alpha) \mid X\}\right] \\
= & E\left(E\left[\mathbb{P}_n \omega(X) \frac{2Z-1}{\{P(Z=1 \mid X; \gamma)\}^{Z}\{1-P(Z=1 \mid X; \gamma)\}^{1-Z}} H(Y, D, X; \alpha) \mid X\right]\right) \\
& - E\left(\mathbb{P}_n \omega(X) E^{*}\{H(Y, D, X; \alpha) \mid X\} E\left[\frac{2Z-1}{\{P(Z=1 \mid X; \gamma)\}^{Z}\{1-P(Z=1 \mid X; \gamma)\}^{1-Z}} \mid X\right]\right) \\
= & E\left(\mathbb{P}_n \omega(X)\left[E\{H(Y, D, X; \alpha) \mid Z=1, X\} - E\{H(Y, D, X; \alpha) \mid Z=0, X\}\right]\right) - 0
\end{aligned}
$$

Proof of Theorem 2

If the model for $f(Z \mid X; \gamma)$ is correctly specified, but the model for $E\{H(Y, D, X; \alpha) \mid X\}$ may be mis-specified. Assume $\widehat{E}\{H(Y, D, X; \alpha) \mid X\} - E^*\{H(Y, D, X; \alpha) \mid X\} \xrightarrow{p} 0$.

$$
\begin{aligned}
& E\left(\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} \left[H(Y, D, X; \alpha) - E^*\{H(Y, D, X; \alpha) \mid X\}\right]\right) \\
=& E\left\{\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} H(Y, D, X; \alpha)\right\} - E\left[\mathbb{P}_n \omega(X) \frac{2Z-1}{f(Z \mid X; \gamma)} E^*\{H(Y, D, X; \alpha) \mid X\}\right] \\
=& E\left(E\left[\mathbb{P}_n \omega(X) \frac{2Z-1}{\{P(Z=1 \mid X; \gamma)\}^Z \{1 - P(Z=1 \mid X; \gamma)\}^{1-Z}} H(Y, D, X; \alpha) \mid X\right]\right) \\
& - E\left(\mathbb{P}_n \omega(X) E^*\{H(Y, D, X; \alpha) \mid X\} E\left[\frac{2Z-1}{\{P(Z=1 \mid X; \gamma)\}^Z \{1 - P(Z=1 \mid X; \gamma)\}^{1-Z}} \mid X\right]\right) \\
=& E\left(\mathbb{P}_n \omega(X) [E\{H(Y, D, X; \alpha) \mid Z=1, X\} - E\{H(Y, D, X; \alpha) \mid Z=0, X\}]\right) - 0
\end{aligned}
$$

Simulation Studies

In simulation studies, we generate data from the following models:

$$\delta^{\mathrm{L}}(X) = \tanh\left(\alpha^{\mathrm{T}} X\right), \quad \delta^{\mathrm{M}}(X) = \exp\left(\alpha^{\mathrm{T}} X\right)$$
$$\phi_i(X) = \operatorname{expit}\left(\beta_i^{\mathrm{T}} X\right) \quad (i = 1, \ldots, 4), \quad \mathrm{OP}^{\mathrm{CO}}(X) = \exp\left(\eta^{\mathrm{T}} X\right)$$
$$\mathrm{pr}(Z = 1 \mid X) = \operatorname{expit}\left(\gamma^{\mathrm{T}} X\right)$$

where the covariates X include an intercept and a random variable generated from $\mathrm{Un}(-1, 1)$.

- $\alpha = (0, -1)^{\mathrm{T}}, \beta_i = (-0.4, 0.8)^{\mathrm{T}} (i = 1, \ldots, 4), \eta = (-0.4, 1)^{\mathrm{T}}$ and $\gamma = (0.1, -1)^{\mathrm{T}}$.
- he strength of the instrumental variable, $\Delta^D = E\{E(D \mid Z = 1, X) - E(D \mid Z = 0, X)\} = 0.406$.
- The sample size is 1000 .

Simulation Studies

We also consider scenarios in which the nuisance models are misspecified. Two misspecified covariates:

- $X^{\dagger}$ that include an intercept and an irrelevant covariate generated from an independent $Un(-1, 1)$.
- X' including

$$(\underbrace{1, \ldots, 1}_{0.5n}, \underbrace{0 \ldots, 0}_{0.5n})^{\mathrm{T}}, \quad (\underbrace{0, \ldots, 0}_{0.1n}, \underbrace{1, \ldots, 1}_{0.9n})^{\mathrm{T}}.$$

More specifically, we consider the following assumption of misspecification:

- We fits the model $\mathrm{pr}\left(Z = 1 \mid X^{\dagger}; \gamma\right)$ and/or $\phi_i\left(X'; \beta_i\right)$ $(i = 1, \ldots, 4)$ and $\mathrm{OP}^{\mathrm{CO}}\left(X'; \eta\right)$. Meanwhile, we still uses the correct functional form in these models.
- The target model $\theta(X; \alpha)$ is always correctly specified.

Simulation Studies

We repeat the simulation studies of the three methods proposed in this passage,

- mle: the proposed maximum likelihood estimator;
- drw: the proposed doubly robust estimator with the optimal weighting function;
- dru: the proposed doubly robust estimator with the identity weighting function.

and the following model misspecification scenarios:

- bth: X is used in all nuisance models;
- psc: X is used in the instrumental density model, and X' is used in other nuisance models;
- opc: $X^{\dagger}$ is used in the instrumental density model, and X is used in other nuisance models;
- bad: $X^{\dagger}$ is used in the instrumental density model, and X' is used in other nuisance models.

Simulation Studies: Results

	LATE		MLATE	
Method	α_0	α_1	α_0	α_1
mle.bth	4.55 (0.44)	9.31 (0.93)	13.81 (0.81)	18.33 (1.32)
mle.opc	5.86 (0.48)	11.94 (0.95)	16.47 (0.85)	21.11 (1.35)
mle.psc	5.86 (0.48)	11.94 (0.95)	16.47 (0.85)	21.11 (1.35)
mle.bad	19.15 (0.69)	2.80 (0.62)	41.28 (1.46)	0.53 (1.05)
dru.bth	0.95 (0.45)	6.16 (1.04)	3.58 (1.00)	9.75 (1.79)
dru.opc	1.21 (0.43)	4.55 (0.97)	0.02 (1.05)	10.78 (1.99)
dru.psc	1.21 (0.43)	4.55 (0.97)	0.02 (1.05)	10.78 (1.99)
dru.bad	15.25 (0.70)	29.99 (1.80)	25.23 (1.43)	17.30 (2.56)
drw.bth	0.29 (1.04)	1.55 (1.20)	3.04 (1.24)	4.75 (1.77)
drw.opc	3.71 (1.27)	0.63 (1.45)	0.30 (1.10)	6.39 (1.78)
drw.psc	3.71 (1.27)	0.63 (1.45)	0.30 (1.10)	6.39 (1.78)
drw.bad	10.97 (0.56)	13.68 (1.19)	29.54 (1.65)	4.18 (2.59)

Table: bias $\times$ 100 (standard error $\times$ 100) of parameter estimation under different scenarios

Simulation Studies: Results

Here are the results of simulation studies reported by the authors.

	$\theta(X) = \delta^{\mathrm{L}}(X)$				$\theta(X) = \delta^{\mathrm{M}}(X)$			
	α_0		α_1		α_0		α_1	
mle.bth	0.28	(0.35)	−3.5	(0.78)	−0.092	(0.71)	−3.0	(1.2)
mle.bad	−20	(0.42)	−15	(0.80)	−48	(1.2)	−18	(2.1)
drw.bth	0.55	(0.36)	−4.1	(0.82)	0.54	(0.77)	−5.6	(1.5)
drw.psc	0.060	(0.38)	−5.9	(1.0)	−0.38	(1.2)	−12	(2.7)
drw.opc	0.55	(0.36)	−3.9	(0.79)	0.49	(0.75)	−5.3	(1.4)
drw.bad	−10	(0.40)	−9.6	(1.1)	−28	(1.4)	25	(3.3)
dru.bth	1.3	(0.44)	−5.8	(1.0)	1.8	(0.84)	−8.1	(1.7)
reg.ogburn.bth	−5.7	(1.6)	−2.9	(3.1)	7.8	(2.0)	−1.1	(2.2)
reg.ogburn.bad	−9.0	(0.25)	100	(0.23)	140	(5.6)	93	(3.6)
drw.ogburn.bth	0.10	(0.46)	−4.2	(0.99)	3.2	(1.4)	−13	(2.5)
dru.ogburn.bth	1.3	(0.45)	−5.8	(1.1)	1.9	(0.85)	−8.2	(1.7)
dru.wang.bth	1.3	(0.45)	−5.8	(1.0)	—		—	
dru.simple.bth	1.3	(0.45)	−5.8	(1.0)	1.8	(0.84)	−8.0	(1.7)
dru.simple.psc	1.2	(0.44)	−6.2	(1.0)	1.9	(0.84)	−8.8	(1.7)
dru.simple.opc	4.5	(0.49)	−17	(1.2)	−0.15	(0.68)	11	(1.2)
dru.simple.bad	−16	(0.48)	−17	(1.3)	−34	(0.70)	18	(1.5)
ls.abadie.bth	−0.19	(0.37)	−4.1	(0.93)	0.42	(0.79)	−11	(1.6)
ls.abadie.bad	−23	(0.88)	22	(1.2)	−32	(1.9)	7.7	(3.6)
mle.crude	−2.8	(0.10)	60	(0.19)	0.36	(0.25)	51	(0.42)

401(k) Data

The 401(k) plan has become the most popular employer-sponsored retirement plan in the U.S.A. Economists have long been interested in whether 401(k) contributions represent additional savings or simply replace other retirement plans, such as Individual Retirement Accounts.

Our 401(k) dataset contains 9275 individuals and the following variables:

- e401k =1 if eligble for 401(k).
- inc annual income
- marr =1 if married
- male =1 if male respondent
- age in years
- fsize family size
- nettfa net total fin. assets, 1000
- p401k =1 if participate in 401(k)
- pira =1 if have IRA

401(k) Data

We choose

- e401(k) (eligble for 401(k)) as instrumental variable.
- p401(k) (participation in 401(k)) as treatment indicator.
- pira(k) (have IRA) as response variable.
- intercept, inc, inc^2, age, marr, fsize as covariates.

Our model assumptions are:

- Local treatment effect is modelled as $\delta^M(X) = \exp(\alpha' X)$.
- Since individuals may choose to participate in 401(k) plans, there are no defiers and or always takers. As a result, $\phi_2 \equiv 0, \phi_2\phi_4 =\equiv 0$, and the model of $P(D, Y|Z)$ can be determined by $\delta^M(X), \phi_1(X), \phi_3(X), OP^{CO}(X)$.

In our real data experiments, we use mle and drw without misspecification to estimate the parameters.

401(k) Data: Results

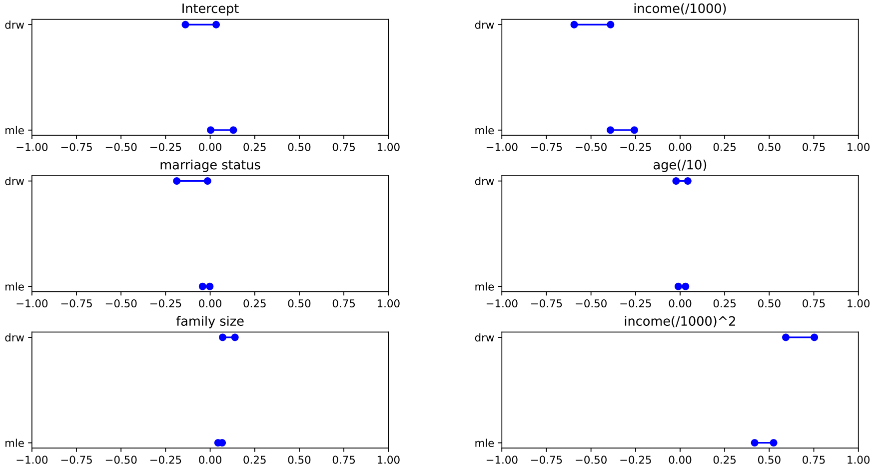


Figure: 95% confidence intervals of coefficients in MLATE using different methods

Here are the results reported by the authors:

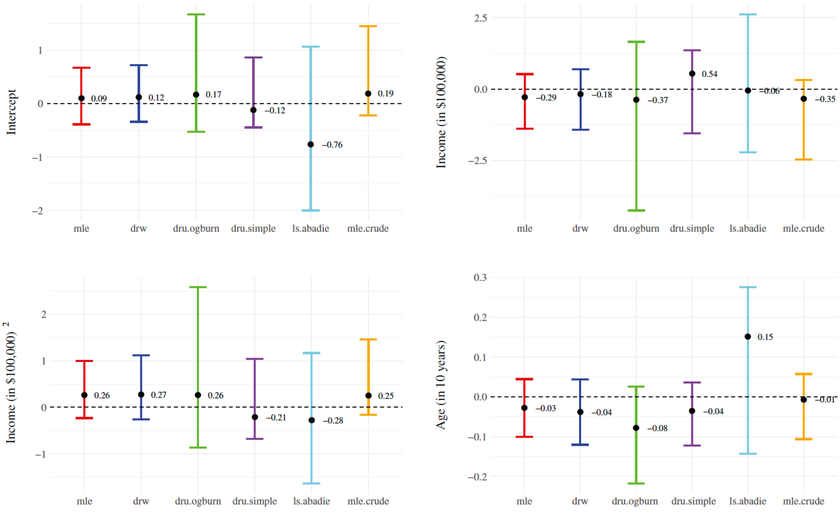


Figure: 95% confidence intervals of coefficients in MLATE using different methods (Part I)

401(k) Data: Results

Here are the results reported by the authors:

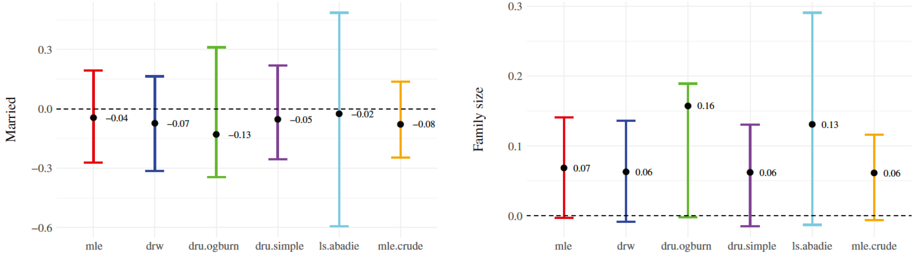


Figure: 95% confidence intervals of coefficients in MLATE using different methods (Part II)

401(k) Data: Results

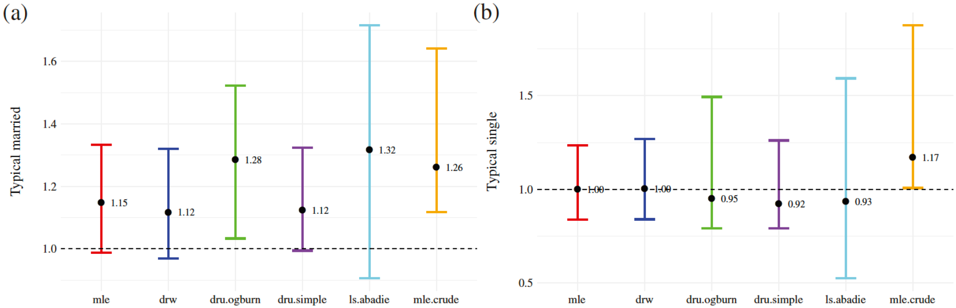


Figure: 95% confidence intervals of MLATE using different methods

We may find that the two proposed methods (mle and drw) tend to give a conservative of local treatment effect and enjoy smaller variance.