On Doubly Robust Estimation with Nonignorable Missing Data Using Instrumental Variables

Baoluo Sun[1], Wang Miao[2], and Deshanee S. Wickramarachchi[1]

[1]*Department of Statistics and Data Science, National University of Singapore*

[2]*Department of Probability and Statistics, Peking University*

Supplementary Material

This Supplementary Material contains detailed proofs of propositions, a further discussion on local efficiency and additional simulation results.

S1 Identification

For any given $u \in \mathcal{U}$, consider the following model for the complete data distribution under assumptions 1 and 2,

$$p(r, y, z; \varphi) = \pi(y, z; \tau)^r \{1 - \pi(y, z; \tau)\}^{1-r} p(y; \psi) p(z; \beta),$$

where $\varphi = (\tau, \psi, \beta)$. Suppose two candidate values φ_1 and φ_2 of φ yields the same observed data distribution,

$$p(z; \varphi_1) = p(z; \varphi_2), \quad p(R = 1, y \mid z; \varphi_1) = p(R = 1, y \mid z; \varphi_2),$$

which characterizes all values of φ to be ruled out for identification. This yields the result below.

Lemma 1. *For any given $u \in \mathcal{U}$, the parameter φ is identified if and only if for any two values φ_1 and φ_2 of φ,*

(a) $p(z;\varphi_1) \neq p(z;\varphi_2)$*; or*

(b) $p(R=1,y \mid z;\varphi_1) \neq p(R=1,y \mid z;\varphi_2)$.

Because for any given $u \in \mathcal{U}$, $p(z)$ can be uniquely determined from the observed data, (a) can be checked based on observed data, but (b) involves the missingness process. An equivalent statement of (b) is $\frac{\pi(y,z;\tau_1)}{\pi(y,z;\tau_2)} \neq \frac{p(y;\psi_2)}{p(y;\psi_1)}$. For any two candidate values φ_1 and φ_2 of φ such that $p(z;\varphi_1) = p(z;\varphi_2)$, the ratio $\frac{p(y;\psi_2)}{p(y;\psi_1)}$ must vary with y. A straightforward corollary follows if the ratio $\frac{\pi(y,z;\tau_1)}{\pi(y,z;\tau_2)}$ is a constant or varies with z.

Corollary 1. *For any given $u \in \mathcal{U}$, the parameter φ is identified if $\frac{\pi(y,z;\tau_1)}{\pi(y,z;\tau_2)}$ is either a constant or varies with z for any two values τ_1 and τ_2 of τ.*

Note that the condition in corollary 1 does not restrict the models $p(y;\psi)$ and $p(z;\beta)$. We can check the condition in corollary 1 for specific semi-parametric or parametric models; examples may be found in Sun et al. (2018).

S2 Proof of Proposition 1

We first prove the following result.

Lemma 2. *Suppose the exclusion restriction $Z \perp Y \mid U$ holds for the full data distribution. Then the following mean zero condition holds,*

$$\mathbb{E}\left\{f(W) - f^{\dagger}\left(W; \bar{\beta}, \bar{\psi}\right)\right\} = 0,$$

for arbitrary measurable and square-integrable function $f(W)$ of the full data W, if either (a) $p(z \mid u; \bar{\beta}) = p(z \mid u)$ or (b) $p(y \mid u; \bar{\psi}) = p(y \mid u)$.

Proof. If (a) holds, then by the Law of Iterated Expectations,

$$\mathbb{E}\left[f(W) - E\{f(W) \mid V\} - \mathbb{E}\left\{f(W) \mid X; \bar{\psi}\right\} + \mathbb{E}\left\{\mathbb{E}\left(f(W) \mid X; \bar{\psi}\right) \mid U\right\}\right] = 0,$$

where $V = (Y, U)$ and $\mathbb{E}\left\{f(W) \mid X; \bar{\psi}\right\}$ denotes expectation taken under a possibly misspecified model for $p(y \mid u)$. Similarly, if (b) holds,

$$\mathbb{E}\left[f(W) - \mathbb{E}\left\{f(W) \mid X\right\} - \mathbb{E}\left\{f(W) \mid V; \bar{\beta}\right\} + \mathbb{E}\left\{\mathbb{E}\left(f(W) \mid V; \bar{\beta}\right) \mid U\right\}\right] = 0,$$

where $\mathbb{E}\left\{f(W) \mid V; \bar{\beta}\right\}$ denotes expectation taken under a possibly misspecified model for $p(z \mid u)$. □

We assume that the regularity conditions of Newey and McFadden (1994, Theorem 3.4) hold for the moment function $\{g^{\mathrm{T}}(O; \phi, \theta, c, d), m^{\mathrm{T}}(O; \gamma, \theta)\}^{\mathrm{T}}$. Then the probability limit of the empirical moment condition

$$\mathbb{P}_n\{g^{\mathrm{T}}(O; \phi, \theta, c, d), m^{\mathrm{T}}(O; \gamma, \theta)\}^{\mathrm{T}} = 0,$$

has a unique solution $(\bar{\phi}^{\mathrm{T}}, \bar{\theta}^{\mathrm{T}})^{\mathrm{T}}$, and by standard Taylor expansion,

$$\begin{aligned}
0 =& n^{-1/2}\sum_{i=1}^{n} g(O_i;\bar{\phi},\bar{\theta},c,d) + \left[\mathbb{E}\left\{\frac{\partial}{\partial\phi}g(O;\phi,\bar{\theta},c,d)\right\}\Big|_{\phi=\bar{\phi}}\right. \\
&- \mathbb{E}\left\{\frac{\partial}{\partial\theta}g(O;\bar{\phi},\theta,c,d)\right\}\Big|_{\theta=\bar{\theta}} \\
&\times \mathbb{E}\left\{\frac{\partial}{\partial\theta}m(O;\bar{\gamma},\theta)\right\}^{-1}\Big|_{\theta=\bar{\theta}} \left.\mathbb{E}\left\{\frac{\partial}{\partial\phi}m(O;\gamma,\bar{\theta})\right\}\Big|_{\phi=\bar{\phi}}\right] n^{1/2}(\widehat{\phi}(c,d)-\bar{\phi}) \\
&- \mathbb{E}\left\{\frac{\partial}{\partial\theta}g(O;\bar{\phi},\theta,c,d)\right\}\Big|_{\theta=\bar{\theta}} \mathbb{E}\left\{\frac{\partial}{\partial\theta}m(O;\bar{\gamma},\theta)\right\}^{-1}\Big|_{\theta=\bar{\theta}} m(O;\bar{\gamma},\bar{\theta}) + o_p(1) \\
=& n^{-1/2}\sum_{i=1}^{n} G(O;\bar{\phi},\bar{\theta},c,d) + \mathbb{E}\left\{\frac{\partial}{\partial\phi}G(O;\phi,\bar{\theta},c,d)\right\}\Big|_{\phi=\bar{\phi}} n^{1/2}(\widehat{\phi}(c,d)-\bar{\phi}) + o_p(1),
\end{aligned}$$

where

$$\begin{aligned}
G(O;\phi,\bar{\theta},c,d) =& \, g(O;\phi,\bar{\theta},c,d) \\
&- \mathbb{E}\left\{\frac{\partial}{\partial\theta}g(O;\bar{\phi},\theta,c,d)\right\}\Big|_{\theta=\bar{\theta}} \mathbb{E}\left\{\frac{\partial}{\partial\theta}m(O;\bar{\gamma},\theta)\right\}^{-1}\Big|_{\theta=\bar{\theta}} m(O;\gamma,\bar{\theta}).
\end{aligned}$$

It follows by Slutsky's Theorem and the Central Limit Theorem that

$$\begin{aligned}
n^{1/2}(\widehat{\phi}(c,d)-\bar{\phi}) =& -n^{1/2}\left[\mathbb{E}\left\{\frac{\partial}{\partial\phi}G(O;\phi,\bar{\theta},c,d)\right\}\Big|_{\phi=\bar{\phi}}\right]^{-1} \mathbb{P}_n\{G(O;\bar{\phi},\bar{\theta},c,d)\} \\
&+ o_p(1).
\end{aligned}$$

In the union semiparametric model $\cup_{j=1,2}\mathcal{M}_j$,

$$\mathbb{E}\{g(O;\phi_0,\bar{\theta},c,d)\} = \mathbb{E}\{\mathbb{E}(g(O;\phi_0,\bar{\theta},c,d)\mid W)\} = \mathbb{E}\{q(W;\mu_0,\bar{\beta},\bar{\psi},c,d)\} = 0,$$

where the last equality holds due to lemma 2. The second part of the proposition follows from Neyman orthogonality (Chernozhukov et al., 2018,

2022). At the intersection submodel $\cap_{j=1,2}\mathcal{M}_j$,

$$\mathbb{E}\left\{ \frac{\partial}{\partial\theta} g(O;\phi_0,\theta,c,d) \right\}\Big|_{\theta=\bar{\theta}} = 0.$$

S3 Proof of Proposition 2

The proof is similar to that of proposition 1 and is thus omitted.

S4 Local efficiency

The efficient influence function can be computed as the orthogonal projection of any influence function onto the tangent space of the model $\mathcal{M}$ (van der Laan and Robins, 2003; Tsiatis, 2007). A closed form expression for this projection is available when both Y and Z are general discrete variables taking values in $\{0,1,...,\ell_z\}$ and $\{0,1,...,\ell_y\}$, respectively (Sun et al., 2018). Specifically, let $v_1(Y) = \{I(Y=1),...,I(Y=\ell_y)\}^{\mathrm{T}}$ and $v_2(Z) = \{I(Z=1),...,I(Z=\ell_z)\}^{\mathrm{T}}$, where $I(\cdot)$ is the indicator function. Then for any function of the full data W, $f(W) - f^{\dagger}(W) = f_{yz}(U)v(W)$ for some conformable $f_{yz}(U)$, where $v(W) = \{v_1(Y) - \mathbb{E}(v_1(Y) \mid U)\} \otimes \{v_2(Z) - \mathbb{E}(v_2(Z) \mid U)\}$.

To characterize the optimal choice of the index function, let

$$g_1(O;\phi) = \frac{R(Y-\mu)}{\pi(W;\gamma)} + \left\{1 - \frac{R}{\pi(W;\gamma)}\right\} \mathbb{E}\{(Y-\mu) \mid R=0, X;\gamma\};$$
$$g_2(O;\gamma) = \frac{Rv(W)}{\pi(W;\gamma)} + \left\{1 - \frac{R}{\pi(W;\gamma)}\right\} \mathbb{E}\{v(W) \mid R=0, X;\gamma\}.$$

Then the efficient influence function is indexed by $c^*(W) = c_{yz}(U)v(W)$ and $d^*(W) = d_{yz}(U)v(W)$, where

$$c_{yz}^{\mathrm{T}}(U) = \mathbb{E}\{g_2(O;\gamma_0)g_2^{\mathrm{T}}(O;\gamma_0) \mid U\}^{-1}\mathbb{E}\{g_2(O;\gamma_0)g_1(O;\phi_0) \mid U\},$$
$$d_{yz}^{\mathrm{T}}(U) = \mathbb{E}\{g_2(O;\gamma_0)g_2^{\mathrm{T}}(O;\gamma_0) \mid U\}^{-1}\mathbb{E}\{\partial g_2(O;\gamma)/\partial\gamma \mid_{\gamma=\gamma_0} \mid U\}.$$

(Sun et al., 2018). In practice, these optimal index functions are estimated. Let $\hat{c}^*(W) = c^*(W;\hat{\phi},\hat{\theta})$ and $\hat{d}^*(W) = d^*(W;\hat{\phi},\hat{\theta})$, where $\hat{\phi}$ is any initial doubly robust estimator of ϕ_0. A doubly robust and locally efficient estimator of ϕ_0 is given by the first $(p+1)$ elements of the joint solution to the empirical moment condition

$$\mathbb{P}_n\{g^{\mathrm{T}}(O;\phi,\theta,\hat{c}^*,\hat{d}^*), m^{\mathrm{T}}(O;\gamma,\theta)\}^{\mathrm{T}} = 0.$$

S5 Additional simulation results

S5.1 Violations of exclusion restriction

To evaluate the finite-sample properties of the proposed estimator under departures from the exclusion restriction in assumption 2, the baseline co-

variate $U = (U_1, U_2)^{\mathrm{T}}$ is generated from a bivariate normal distribution $N(0, \Sigma)$, where the elements of Σ are $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{12} = 0.2$. Conditional on U, (R, Y, Z) is generated from the following generalized linear models,

$$
\begin{aligned}
Z \mid U &\sim \text{Bernoulli}\{p_1 = \text{expit}(1 + 2U_1 - U_2 - 0.8U_1U_2)\}, \\
Y \mid Z, U &\sim \text{Bernoulli}\{p_2 = \text{expit}(0.5 - 2U_1 + U_2 - \rho Z)\}, \\
R \mid Y, X &\sim \text{Bernoulli}\{\pi = \text{expit}(2 - 3Z + 0.8U_1 + U_2 + \gamma Y)\},
\end{aligned}
$$

where $\gamma = 2$, and ρ encodes the log odds ratio between Y and Z conditional on U. The estimators are implemented in the same way as described in the main manuscript, but with data generated under $\rho = -0.1$ or -0.2. The results of 1000 simulation replicates of sample size $n = 500$, 1000 or 5000 are summarized in Tables S1 and S2. The bias and undercoverage of the estimators $\tilde{\mu}$ and $\hat{\mu}_{\text{dr}}$ become noticeable when $\rho = -0.2$, which is in agreement with theory. The conclusions are otherwise qualitatively similar to those obtained under exclusion restriction.

Table S1: Summary of results for estimation of the outcome mean when $\rho = -0.1$.

			(C1)		(C2)		(C3)		(C4)		(C5)	
			All correct		mis $p(z \mid u)$		mis $p(y \mid u)$		mis $\eta(x)$		All mis	
	$\hat{\mu}_{\text{cc}}$	$\hat{\mu}_{\text{full}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$
						$n = 500^{\dagger}$						
|Bias|	.138	.002	.009	.016	.012	.100	.007	.016	.054	.042	.105	.112
$\sqrt{\text{Var}}$	.025	.021	.056	.053	.057	.046	.059	.053	.049	.098	.047	.040
$\sqrt{\text{EVar}}$	.026	.022	.055	.139	.324	6.110	.065	.107	.071	.042	.048	.045
Cov95	.000	.958	.919	.936	.930	.939	.916	.947	.788	.813	.400	.295
						$n = 1000$						
|Bias|	.138	.001	.009	.011	.012	.098	.008	.012	.054	.029	.103	.110
$\sqrt{\text{Var}}$	.017	.016	.040	.041	.040	.032	.041	.037	.034	.035	.033	.028
$\sqrt{\text{EVar}}$	.019	.016	.038	.039	.039	.838	.039	.039	.034	.029	.034	.032
Cov95	.000	.945	.932	.934	.937	.937	.926	.942	.614	.812	.148	.040
						$n = 5000$						
|Bias|	.139	.001	.011	.011	.012	.100	.011	.010	.056	.025	.104	.111
$\sqrt{\text{Var}}$	.008	.007	.018	.016	.017	.014	.018	.016	.015	.015	.015	.013
$\sqrt{\text{EVar}}$	.008	.007	.018	.016	.018	.055	.018	.017	.015	.013	.015	.014
Cov95	.000	.952	.915	.924	.916	.409	.919	.934	.043	.529	.000	.000

Note: †The results for $\hat{\mu}_{\text{dr}}$ excluded 6 simulation replicates due to convergence failure at $n = 500$. |Bias| and $\sqrt{\text{Var}}$ are the Monte Carlo absolute bias and standard deviation of the point estimates, $\sqrt{\text{EVar}}$ is the square root of the mean of the variance estimates and Cov95 is the coverage proportion of the 95% confidence intervals, based on 1000 repeated simulations. Zeros denote values smaller than .0005.

Table S2: Summary of results for estimation of the outcome mean when $\rho = -0.2$.

			(C1)		(C2)		(C3)		(C4)		(C5)	
			All correct		mis $p(z \mid u)$		mis $p(y \mid u)$		mis $\eta(x)$		All mis	
	$\hat{\mu}_{\text{cc}}$	$\hat{\mu}_{\text{full}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$	$\tilde{\mu}$	$\hat{\mu}_{\text{dr}}$
						$n = 500^{\dagger}$						
|Bias|	.141	.002	.022	.026	.026	.110	.020	.026	.066	.050	.116	.121
$\sqrt{\text{Var}}$	.025	.021	.057	.051	.058	.046	.060	.053	.049	.064	.047	.040
$\sqrt{\text{EVar}}$	.026	.022	.057	.084	.120	4.817	.612	.071	.049	.044	.048	.046
Cov95	.000	.960	.919	.937	.927	.934	.908	.939	.700	.767	.315	.216
						$n = 1000$						
|Bias|	.140	.002	.022	.024	.026	.108	.021	.023	.066	.041	.115	.119
$\sqrt{\text{Var}}$	.018	.016	.041	.038	.042	.033	.042	.039	.034	.035	.033	.028
$\sqrt{\text{EVar}}$	.019	.016	.038	.040	.040	0.730	.040	.040	.034	.030	.033	.032
Cov95	.000	.958	.908	.915	.906	.936	.898	.923	.465	.725	.087	.020
						$n = 5000$						
|Bias|	.142	.001	.024	.022	.026	.111	.024	.022	.068	.037	.116	.121
$\sqrt{\text{Var}}$	.008	.007	.018	.016	.018	.015	.019	.016	.015	.016	.015	.013
$\sqrt{\text{EVar}}$	.008	.007	.018	.016	.018	.050	.018	.017	.015	.014	.015	.014
Cov95	.000	.953	.729	.742	.702	.251	.751	.791	.005	.246	.000	.000

Note: †The results for $\hat{\mu}_{\text{dr}}$ excluded 10 simulation replicates due to convergence failure at $n = 500$. |Bias| and $\sqrt{\text{Var}}$ are the Monte Carlo absolute bias and standard deviation of the point estimates, $\sqrt{\text{EVar}}$ is the square root of the mean of the variance estimates and Cov95 is the coverage proportion of the 95% confidence intervals, based on 1000 repeated simulations. Zeros denote values smaller than .0005.

S5.2 Continuous outcome

We perform additional Monte Carlo simulations with continuous Y. The baseline covariate U_j is generated independently from the truncated normal distribution in the interval $(-1, 1)$, for $j = 1, 2$. Conditional on $U = (U_1, U_2)^{\mathrm{T}}$, (R, Y, Z) is generated from the following process consistent with assumptions 1–3,

$$\begin{aligned} Z \mid U &\sim \text{Bernoulli}\{p = \text{expit}(1 + 2U_1 - U_2 - 0.8U_1U_2)\}, \\ Y \mid Z, U &\sim \text{Normal}(0.5 - 2U_1 + U_2, \quad \sigma^2 = 0.25), \\ R \mid Y, X &\sim \text{Bernoulli}\{\pi = \text{expit}(2 - 4Z + 0.8U_1 + U_2 + \gamma Y)\}, \end{aligned}$$

where $\gamma = 2$. We implement the proposed estimator $\tilde{\phi}(c, d) = (\tilde{\mu}, \tilde{\gamma})^{\mathrm{T}}$ with $c(w) = 0$, $d(w) = yz$ and the models

$$\begin{aligned} \pi(w; \xi, \gamma) &= \text{expit}\{(1, z, u_1, u_2)\xi + \gamma y\}, \\ p(y \mid u; \psi) &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y - h_1(u)\lambda)^2}{(2\sigma^2)}\right\}, \quad \psi = (\sigma, \lambda^{\mathrm{T}})^{\mathrm{T}}, \\ p(Z = 1 \mid u; \beta) &= \text{expit}\{h_2(u)\beta\}. \end{aligned}$$

To investigate its performance under possible model misspecifications, we implement the doubly robust estimator based on the following specifications, (TT) $h_1(u) = (1, u_1, u_2)$, $h_2(u) = (1, u_1, u_2, u_1u_2)$, (TF) $h_1(u) = (1, u_1, u_2)$, $h_2(u) = (1, u_1, u_1^2)$, (FT) $h_1(u) = (1, u_1, u_1^2)$, $h_2(u) = (1, u_1, u_2, u_1u_2)$,

and (FF) $h_1(u) = (1, u_1, u_1^2)$, $h_2(u) = (1, u_1, u_1^2)$. We also implement the complete-case estimator $\hat{\mu}_{\text{cc}} = \mathbb{P}_n\{RY\}$, and the infeasible full-data estimator $\hat{\mu}_{\text{full}} = n^{-1}\sum_{i=1}^{n} Y_i$ as performance benchmark. For inference, we construct 95% Wald confidence intervals based on the sandwich estimator of asymptotic variance.

We perform 1000 simulation replicates under each specification and sample size $n = 1000, 5000$. Convergence failure occurred in a small number of these replicates when solving the nonlinear moment condition for $\tilde{\phi}$, and the convergence failure rates are tabulated in Table S3. The simulation results for estimation of the outcome mean based on the replicates with successful convergence are summarized in Table S4. The complete-case estimator $\hat{\mu}_{\text{cc}}$ shows severe bias and undercoverage. The proposed doubly robust estimator $\tilde{\mu}$ performs well in terms of bias and coverage across the simulation scenarios TT, TF and, FT, but exhibits bias and undercoverage in scenario FF, which is in agreement with theory.

Table S3: Convergence failure rate for $\tilde{\phi}$ out of 1000 Monte Carlo replicates.

n	TT	TF	FT	FF
1000	0.007	0.049	0.008	0.001
5000	0.000	0.001	0.000	0.000

Note: The convergence criteria is the residual of the square average estimating equation component being less than 1e-7.

Table S4: Summary of results for estimation of the mean of a continuous outcome.

	$\hat{\mu}_{\text{cc}}$	$\hat{\mu}_{\text{full}}$	$\tilde{\mu}$			
			TT	TF	FT	FF
|Bias|	0.742	0.001	0.001	0.008	0.003	0.208
	0.743	0.000	0.002	0.001	0.002	0.206
$\sqrt{\text{Var}}$	0.053	0.042	0.066	0.080	0.067	0.072
	0.023	0.018	0.030	0.036	0.030	0.031
$\sqrt{\text{EVar}}$	0.053	0.041	0.063	0.075	0.064	0.071
	0.024	0.018	0.029	0.036	0.031	0.031
Cov95	0.000	0.942	0.932	0.912	0.932	0.173
	0.000	0.961	0.939	0.949	0.943	0.000

Note: The results are presented in two rows, of which the first stands for sample size $n = 1000$, and the second for $n = 5000$. |Bias| and $\sqrt{\text{Var}}$ are the Monte Carlo absolute bias and standard deviation of the point estimates, $\sqrt{\text{EVar}}$ is the square root of the mean of the variance estimates and Cov95 is the coverage proportion of the 95% confidence intervals, based on replicates with successful convergence in 1000 repeated simulations. Zeros denote values smaller than .0005.

Bibliography

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *Economet. J. 21*(1), C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022). Locally robust semiparametric estimation. *Econometrica 90*(4), 1501–1535.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier.

Sun, B., L. Liu, W. Miao, K. Wirth, J. Robins, and E. J. Tchetgen Tchetgen (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statist. Sinica 28*(4), 1965–1983.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.