Doubly robust estimation under MNAR model

Proposal

- Recall that the density ratio function

$$r(\mathbf{x}, y) = \frac{f(\mathbf{x}, y \mid \delta = 0)}{f(\mathbf{x}, y \mid \delta = 1)} := \frac{f_0(\mathbf{x}, y)}{f_1(\mathbf{x}, y)}$$

  satisfies

$$\frac{1}{\pi(\mathbf{x}, y)} = 1 + c \cdot r(\mathbf{x}, y),$$

  where $c = p^{-1} - 1$ and $p = P(\delta = 1)$.

- We wish to consider the following class of estimator

$$\hat{\theta}_{PS} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\pi(\mathbf{x}_i, y_i)} y_i + \frac{1}{n} \sum_{i=1}^{n} \left\{ 1 - \frac{\delta_i}{\pi(x_i, y_i)} \right\} m(\mathbf{x}_i)$$

- Note that

$$\hat{\theta}_{PS} - \hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \{\delta_i \omega(x_i, y_i) - 1\} \{y_i - m(x_i)\}$$

where $\omega(x, y) = \{\pi(x, y)\}^{-1}$.

- Let's compute the conditional expectation of the above difference wrt the outcome model.

$$\begin{aligned} & E\left(\hat{\theta}_{PS} - \hat{\theta}_n \mid X, \delta\right) \\ = \; & \delta E(\omega Y \mid X, \delta = 1) - \delta E(y \mid x, \delta = 1) - (1 - \delta)E(y \mid x, \delta = 0) \\ & -\delta m(x) \cdot E(\omega \mid X, \delta = 1) + m(x) \end{aligned}$$

- Writing $\tilde{\pi}(x) = P(\delta = 1 \mid x)$, we obtain

$$\begin{aligned} E(\hat{\theta}_{PS} - \hat{\theta}_n \mid X) &= [\tilde{\pi}(x)E(\omega Y \mid X, \delta = 1) - E(Y \mid x)] \\ &\quad + [1 - \tilde{\pi}(x) \cdot E(\omega \mid x, \delta = 1)]\, m(x) \\ &:= A(x) + B(x) \end{aligned}$$

- If $\pi(x, y) = \{\omega(x, y)\}^{-1}$ is correctly specified, we have

$$E(\omega \mid X, \delta = 1) = \frac{1}{P(\delta = 1 \mid x)} = \frac{1}{\tilde{\pi}(x)} \tag{1}$$

and

$$\frac{E(\omega Y \mid X, \delta = 1)}{E(\omega \mid X, \delta = 1)} = E(Y \mid x). \tag{2}$$

- By (1), we have $B(x) = 0$, regardless of the choice of $m(x)$. Also, by (2), we have $A(x) = 0$.

- Now, we wish to impose conditions on $m(x)$ such that a modest violation of the propensity score model can be allowed.
- If we choose $m(x) = E(Y \mid x)$, we can impose

$$\frac{E(\omega Y \mid X, \delta = 1)}{E(\omega \mid X, \delta = 1)} = m(x) \tag{3}$$

  as a constraint for (doubly) robust estimation.
- Given $\pi(x, y) = \{\omega(x, y)\}^{-1}$, we can compute $\tilde{\pi}(x)$ using $\tilde{\pi}(x) = \{\int \omega(x, y) f_1(y \mid x) dy\}^{-1}$. We can decompose

$$m(x) = \tilde{\pi}(x) m_1(x) + (1 - \tilde{\pi}(x))\, m_0(x).$$

- Thus, (3) can be written as

$$E(\omega Y \mid X, \delta = 1) = m_1(x) + \frac{1 - \tilde{\pi}(x)}{\tilde{\pi}(x)} m_0(x). \tag{4}$$

Outcome regression model

- To achieve the doubly robustness, we assume the following working outcome regression (OR) model

$$f_0(y \mid x) = \frac{\exp(\beta y) f_1(y \mid x)}{\int \exp(\beta y) f_1(y \mid x) dy} \tag{5}$$

where β is an unknown parameter.

- For given $\tilde{\pi}(x)$, we can express

$$m(x; \beta) = \tilde{\pi}(x) m_1(x) + \{1 - \tilde{\pi}(x)\} m_0(x; \beta)$$

and

$$m_0(x; \beta) = \frac{\int \exp(\beta y) y f_1(y \mid x) dy}{\int \exp(\beta y) f_1(y \mid x) dy}.$$

- Parameter β can be estimated from the calibration using (4). See the following example.

Example

- To explain the computational details, assume that the response model is a logistic regression model of the form

$$\pi(x, y) = \frac{\exp(\phi_0 + \phi_1 x_1 + \phi_2 y)}{1 + \exp(\phi_0 + \phi_1 x_1 + \phi_2 y)} \tag{6}$$

- Thus, the original calibration equation should be

$$\sum_{i=1}^{n} \delta_i \omega_i (1, x_{1i}, y_i) = \sum_{i=1}^{n} (1, x_{1i}, y_i). \tag{7}$$

Here, we assume that $x = (x_1, x_2)$ and x_2 is the nonresponse instrumental variable.

- Now, the smoothing weights $\tilde{\omega}(x)$ should utilize the calibration equation in (7). That is, we should have

$$\sum_{i=1}^{n} \delta_i \tilde{\omega}(x_i)(1, x_{1i}) = \sum_{i=1}^{n} (1, x_{1i}) \tag{8}$$

and

$$\sum_{i=1}^{n} \delta_i E_1(\omega_i y_i \mid \mathbf{x}_i) = \sum_{i=1}^{n} \{\delta_i m_1(x_i) + (1 - \delta_i) m_0(x_i)\}. \tag{9}$$

- Note that, by (4), we can express (9) as

$$\sum_{i=1}^{n} \delta_i \{\tilde{\omega}(x_i) - 1\} m_0(x_i) = \sum_{i=1}^{n} (1 - \delta_i) m_0(x_i). \tag{10}$$

This equation will be used to estimate β in (5).

- To compute $\tilde{\omega}(x_i)$, first estimate an estimator of $\phi = (\phi_0, \phi_1, \phi_2)$ under the working PS model in (6) by solving

$$\sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi(x_i, y_i; \phi)} - 1 \right\} \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

  We can use the optimal estimator of Morikawa and Kim (2021) for more efficient estimation, if necessary.
- We assume a Gaussian model for $f_1(y \mid \mathbf{x})$.

$$\hat{f}_1(y \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left\{ -\frac{1}{2\hat{\sigma}^2} \left(y - \mathbf{x}'\hat{\alpha}\right)^2 \right\}. \tag{11}$$

- Since $\omega(\mathbf{x}, y; \phi) = 1 + \exp(-\phi_0 - \phi_1 x_1 - \phi_2 y)$, we can use the moment generating function formula of Gaussian distribution to compute

$$\int \exp\left(-\phi_2 y\right) \hat{f}_1\left(y \mid \mathbf{x}\right) dy = \exp\left(-\phi_2 \mathbf{x}'\hat{\alpha} + \frac{1}{2}\phi_2^2 \hat{\sigma}^2 \right). \tag{12}$$

- Thus, we obtain

$$\tilde{\omega}(\mathbf{x}_i;\hat{\phi}) = \left\{1 + \exp\left(-\hat{\phi}_0 - \hat{\phi}_1 x_{1i} - \hat{\phi}_2 \mathbf{x}_i'\hat{\alpha} + \frac{1}{2}\hat{\phi}_2^2\hat{\sigma}^2\right)\right\}.$$

- The $\tilde{\omega}_i = \tilde{\omega}(x_i)$ is used to estimate β using the estimating equation in (10).
- Once all the parameters are estimated, then we can use

$$\hat{\theta}_{PS} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(\mathbf{x}_i, y_i)}y_i + \frac{1}{n}\sum_{i=1}^{n}\left\{1 - \frac{\delta_i}{\hat{\pi}(x_i, y_i)}\right\}\hat{m}(\mathbf{x}_i) \qquad (13)$$

as the final proposed estimator of $\theta = E(Y)$, where

$$\hat{m}(x_i) = \tilde{\pi}(x_i)\hat{m}_1(x_i) + (1 - \tilde{\pi}(x_i))\hat{m}_0(x_i)$$

- By construction, $\hat{\omega}(x, y) = \{\hat{\pi}(x, y)\}^{-1}$ satisfies

$$E\left\{\delta\hat{\omega}(y - m(x)) \mid \mathbf{x}\right\} = 0.$$

Thus, as long as the OR model is correct, $\hat{\theta}_{\mathrm{PS}}$ is unbiased.

PS model using information projection

- Recall that

$$
\begin{aligned}
\frac{1}{\pi(\mathbf{x}, y)} &= 1 + c \cdot \frac{f_0(x, y)}{f_1(x, y)} \\
&= 1 + c \cdot \frac{f_0(x)}{f_1(x)} \cdot \frac{f_0(y \mid x)}{f_1(y \mid x)} \\
&= 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \cdot \frac{f_0(y \mid x)}{f_1(y \mid x)}
\end{aligned}
$$

- If $\pi(\mathbf{x}, y)$ is correct, then the covariate balancing property holds automatically in the population level. That is, $\omega(\mathbf{x}, y) = \{\pi(\mathbf{x}, y)\}^{-1}$ satisfies

$$
E\{\delta\omega(\mathbf{x}, Y)h(x, Y) \mid x\} = E\{h(x, Y) \mid x\}. \tag{14}
$$

Thus, there is no need to add this constraint to the PS model if it is really correct.

- In practice, we do not know the true probability of response, and $\pi^{(0)}(\mathbf{x}, y)$ is a working PS model.
- In this case, we can express

$$\frac{1}{\pi^{(0)}(\mathbf{x}, y)} = 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \cdot \frac{f_0^{(0)}(y \mid x)}{f_1(y \mid x)}, \tag{15}$$

where $f_0^{(0)}(y \mid x)$ is the baseline density for $f_0(y \mid x)$ derived from $\pi^{(0)}(\mathbf{x}, y)$.

- Note that $\omega^{(0)}(\mathbf{x}, y) = \{\pi^{(0)}(\mathbf{x}, y)\}^{-1}$ does not necessarily satisfy (14).

- Note that

$$
\begin{aligned}
E\{\delta\omega(\mathbf{x}, Y)h(x, Y) \mid x\} &= E[\delta E\{\omega(\mathbf{x}, Y)h(x, Y) \mid x, \delta = 1\} \mid x] \\
&= \tilde{\pi}(x) \cdot \int \omega(\mathbf{x}, y)h(x, y)f_1(y \mid x)dy \\
&= \tilde{\pi}(x) \cdot \int h(x, y)f_1(y \mid x)dy \\
&+ \{1 - \tilde{\pi}(x)\} \cdot \int h(x, y)f_0(y \mid x)dy.
\end{aligned}
$$

- Thus, constraint (14) can be understood as a constraint for $f_0(y \mid x)$:

$$
\begin{aligned}
& \tilde{\pi}(x) \cdot \int h(x, y)f_1(y \mid x)dy + \{1 - \tilde{\pi}(x)\} \cdot \int h(x, y)\color{red}{f_0(y \mid x)}\color{black}dy \\
= \ & E\{h(x, Y) \mid x\} \qquad (16)
\end{aligned}
$$

- Let's use $h(x, y) = y$. In this case, (16) reduces to

$$
\tilde{\pi}(x) \cdot E_1(Y \mid x) + \{1 - \tilde{\pi}(x)\} \cdot \int y\color{red}{f_0(y \mid x)}\color{black}dy = E(Y \mid x). \qquad (17)
$$

- Because f_0 and $f_0^{(0)}$ are densities, a natural choice for the distance function is the Kullback-Leibler divergence measure.

$$D_{\mathrm{KL}}\left(f_0 \parallel f_0^{(0)}\right) = \int f_0(y \mid x) \log\left(\frac{f_0(y \mid x)}{f_0^{(0)}(y \mid x)}\right) d\mu(y). \tag{18}$$

- Thus, for a given $f_0^{(0)}$, we wish to find the minimizer of (18) subject to (17). The solution is given by

$$f_0^*(y \mid x) = f_0^{(0)}(y \mid x) \cdot \frac{\exp(\lambda' y)}{\int \exp(\lambda' y)\, f_0^{(0)}(y \mid x) d\mu(y)}, \tag{19}$$

where λ is the Lagrange multiplier for the constraint (17).

- The solution in (19) is obtained by the information projection technique.

- Using (19) and (15), we can obtain the following.

$$
\begin{aligned}
& \{\pi^*(\mathbf{x}, y; \lambda)\}^{-1} \\
= \quad & 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \cdot \frac{f_0^*(y \mid x)}{f_1(y \mid x)} \\
= \quad & 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \cdot \frac{f_0^{(0)}(y \mid x)}{f_1(y \mid x)} \cdot \frac{\exp(\lambda' y)}{\int \exp(\lambda' y)\, f_0^{(0)}(y \mid x) d\mu(y)} \\
= \quad & 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \cdot \frac{\exp(\lambda' y)\, O^{(0)}(x_1, y)}{\int \exp(\lambda' y)\, O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)}, \qquad (20)
\end{aligned}
$$

where

$$
O^{(0)}(x_1, y) = \frac{1}{\pi^{(0)}(x_1, y)} - 1
$$

and λ is the parameter satisfying

$$
E\left[\delta\{\pi^*(x, y)\}^{-1} y \mid x\right] = E(Y \mid x). \qquad (21)
$$

- Now, note that

$$E\left(\delta Y \mid x\right) \quad = \quad \tilde{\pi}(x)E_1(Y \mid x)$$

and

$$\begin{aligned} & E\left[\delta Y \cdot \frac{\exp\left(\lambda' y\right) O^{(0)}(x_1, y)}{\int \exp\left(\lambda' y\right) O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)} \mid x\right] \\ = \quad & \tilde{\pi}(x) \cdot \frac{\int y \exp\left(\lambda' y\right) O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)}{\int \exp\left(\lambda' y\right) O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)}. \end{aligned}$$

- Thus,

$$E\left[\delta\{\pi^*(x, Y)\}^{-1} Y \mid x\right] = \tilde{\pi}(x)E_1(Y \mid x) + \{1 - \tilde{\pi}(x)\}\, m_0(x) \quad (22)$$

where

$$m_0(x) = \frac{\int y \exp\left(\lambda' y\right) O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)}{\int \exp\left(\lambda' y\right) O^{(0)}(x_1, y) f_1(y \mid x) d\mu(y)}.$$

- Let

$$\hat{m}_0(x;\lambda) = \frac{\int y \exp(\lambda' y)\, O^{(0)}(x_1, y)\hat{f}_1(y \mid x) d\mu(y)}{\int \exp(\lambda' y)\, O^{(0)}(x_1, y)\hat{f}_1(y \mid x) d\mu(y)}.$$

- In the spirit of (10), we can estimate λ by solving

$$\sum_{i=1}^{n} \delta_i \left\{\hat{\omega}(x_i;\lambda) - 1\right\} \hat{m}_0(x_i;\lambda) = \sum_{i=1}^{n} (1-\delta_i)\hat{m}_0(x_i;\lambda)$$

where

$$\hat{\omega}(x;\lambda) = \int \left\{\pi^*(x, y;\lambda)\right\}^{-1} \hat{f}_1(y \mid x) d\mu(y),$$

and $\pi^*(x, y;\lambda)$ is defined in (20).

- Note that we use

$$f_0(y \mid x) \propto \exp(\lambda y) O^{(0)}(x_1, y) f_1(y \mid x),$$

which is equivalent to the previous working OR model below when we use (6) or $\mathrm{logit}(\pi) = f(x) + \phi y$ for any $f(\cdot)$.

$$f_0(y \mid x) \propto \exp(\beta y) f_1(y \mid x)$$

- Also, we have

$$\begin{aligned} \hat{\omega}(x;\lambda) &= \int \{\pi^*(x, y; \lambda)\}^{-1} \hat{f}_1(y \mid x) d\mu(y) \\ &= 1 + \left\{ \frac{1}{\tilde{\pi}(x)} - 1 \right\} \int \frac{\exp(\lambda y) O^{(0)}(x_1, y)}{\int \exp(\lambda y) O^{(0)}(x_1, y) \hat{f}_1(y \mid x) d\mu(y)} \hat{f}_1(y \mid x) d\mu(y) \\ &= \frac{1}{\tilde{\pi}(x)} \overset{let}{=} \tilde{\omega}(x) \end{aligned}$$

- Thus, the estimating equation for λ is equivalent to the estimating equation (10) for β.

Result II - Estimation of $E(Y)$

Table: Monte Carlo biases, variance, and MSE of the estimators $\hat{\theta}_{Full}, \hat{\theta}_{ps}$ and $\hat{\theta}_R$ across 1,000 MC samples

Working RP	Method	Bias	Var	MSE
RP1	$\hat{\theta}_{Full}$	-0.0007	0.0019	0.0019
	$\hat{\theta}_{ps}$	-0.0005	0.0026	0.0026
	$\hat{\theta}_R$	-0.0015	0.0027	0.0027
	$\hat{\theta}_{I.proj}$	-0.0015	0.0025	0.0025
RP2	$\hat{\theta}_{Full}$	-0.0007	0.0019	0.0019
	$\hat{\theta}_{ps}$	0.2699	0.0035	0.0763
	$\hat{\theta}_R$	0.2712	0.0038	0.0773
	$\hat{\theta}_{I.proj}$	0.2728	0.0037	0.0781

NOTE: $\hat{\theta}_R$ is the proposed robust estimator, and $\hat{\theta}_{I.proj}$ is the proposed estimator based on the PS model using I-projection.

Discussion

- The proposed estimator is doubly robust in the sense that it is unbiased either the PS model (6) or the OR model for $f_0(y \mid x)$ in (5) is correctly specified.
- The OR model $f_0(y \mid x)$ can be more generalized as in Franks et al. (2022).
- The respondent's model $f_1(y \mid x)$ is assumed to be correctly specified.
- We can extend it to nonparmaetric estimation of $f_1(y \mid x)$.

Reference

Franks, A. M., E. M. Airoldi and D. B. Rubin (2022), 'Nonstandard conditionally specified models for nonignoable missing data', *Proceedings of the National Academy of Science* **117**, 19045–19053.

Morikawa, Kosuke and Jae Kwang Kim (2021), 'Semiparametric optimal estimation with nonignorable nonresponse data', *The Annals of Statistics* **45**, 2991–3014.