Setup

Consider the following setup. Let $(X, Y_1, Y_2, \delta) \overset{ind}{\sim} F$ for some distribution F that is unknown. We define $Z = (X, Y_1, Y_2$ and we observe the following:

Table 1: This table shows some of our notation and some of the corresponding notation from [1].

Segments	X	Y_1	Y_2	Prob. Element in Segment	δ	C	$G_C(Z)$
A_{00}	✓			π_{00}	δ_{00}	1	$\{X\}$
A_{10}	✓	✓		π_{10}	δ_{10}	2	$\{X, Y_1\}$
A_{01}	✓		✓	π_{01}	δ_{01}	3	$\{X, Y_2\}$
A_{11}	✓	✓	✓	π_{11}	δ_{11}	∞	$\{X, Y_1, Y_2\}$

Define $\varpi(r, Z) = \Pr(C = r \mid Z)$. For now, we assume that $\varpi(r, Z)$ is known. Notice that $\varpi(\infty, Z) = \pi_1 1$, $\varpi(3, Z) = \pi_{01}$, $\varpi(2, Z) = \pi_{10}$, and $\varpi(1, Z) = \pi_{00}$. Since $\pi_{00} + \pi_{10} + \pi_{01} + \pi_{11} = 1$, we only need to define three inclusion probabilities.

Suppose that we want to estimate $\theta = E[g(X, Y_1, Y_2)]$ for a known function g. The proposed estimator is

$$\hat{\theta}_{prop} = \tag{1}$$
$$n^{-1}\sum_{i=1}^n E[g_i \mid X_i] + n^{-1}\sum_{i=1}^n \frac{\delta_{10}}{\pi_{10}}(E[g_i \mid X_i, Y_{1i}] - E[g_i \mid X_i]) + n^{-1}\sum_{i=1}^n \frac{\delta_{01}}{\pi_{01}}(E[g_i \mid X_i, Y_{2i}] - E[g_i \mid X_i])$$
$$+ n^{-1}\sum_{i=1}^n \frac{\delta_{11}}{\pi_{11}}(g_i - E[g_i \mid X_i, Y_{1i}] - E[g_i \mid X_i, Y_{2i}] + E[g_i \mid X_i]).$$

The goal is the show that over all functions $b_1(X, Y_1)$ and $b_2(X, Y_2)$, $\hat{\theta}_{prop}$ is the optimal estimator in the form:

$$\hat{\theta} = \tag{2}$$
$$n^{-1}\sum_{i=1}^n E[g_i \mid X_i] + n^{-1}\sum_{i=1}^n \frac{\delta_{10}}{\pi_{10}}(b_1(X_i, Y_{1i}) - E[g_i \mid X_i]) + n^{-1}\sum_{i=1}^n \frac{\delta_{01}}{\pi_{01}}(b_2(X_i, Y_{2i}) - E[g_i \mid X_i])$$
$$+ n^{-1}\sum_{i=1}^n \frac{\delta_{11}}{\pi_{11}}(g_i - b_1(X_i, Y_{1i}) - b_2(X_i, Y_{2i}) + E[g_i \mid X_i])$$

Semiparametric Inference

We know from Theorem 7.2 of [1] that if $\Pr(C = \infty \mid Z) = \pi_{11} > 0$ then the semiparametric influence function has the form (see page 20 of notes):

$$\frac{I(C=\infty)g(Z)}{\varpi(\infty, Z)} + \frac{I(C=\infty)}{\varpi(\infty, Z)} \left(\sum_{r \neq \infty} \varpi(r, G_r(Z)) L_{2r}(G_r(Z)) \right) - \sum_{r \neq \infty} I(C = \infty) L_{2r}(G_r(Z))$$

where $L_{2r}(G_r(Z))$ is an arbitrary function of $G_r(Z)$. Notice, that this form does not identify *the* optimal estimator but a class of semiparametric functions. A reasonable choice of an estimator for $L_{2r}(G_r(Z))$ is

$$L_{2r}(G_r(Z)) = E[g(Z) \mid G_r(Z)].$$

It turns out that $\hat{\theta}_{prop}$ from Equation 1 is contained within the previous class of estimators with

$$L_{2r}(G_r(Z)) = \varpi(r, G_r(Z)) E[g(Z) \mid G_r(Z)].$$

(See pages 81-82 of notes for proof.)

Linear Expectations

To simplify this problem, we consider the following estimator[1]:

$$\begin{aligned} &\hat{\theta}_c \\ &= \frac{\delta_{11}}{\pi_{11}} g(Z) + \left(1 - \left(\frac{\delta_{10} + \delta_{11}}{\pi_{10} + \pi_{11}}\right) - \left(\frac{\delta_{01} + \delta_{11}}{\pi_{01} + \pi_{11}}\right) + \frac{\delta_{11}}{\pi_{11}}\right) c_0 E[g \mid X] \\ &+ \left(\left(\frac{\delta_{10} + \delta_{11}}{\pi_{10} + \pi_{11}}\right) - \frac{\delta_{11}}{\pi_{11}}\right) c_1 E[g \mid X, Y_1] + \left(\left(\frac{\delta_{01} + \delta_{11}}{\pi_{01} + \pi_{11}}\right) - \frac{\delta_{11}}{\pi_{11}}\right) c_2 E[g \mid X, Y_2] \end{aligned} \tag{3}$$

Projection onto Nuisance Tangent Space

The goal is now to find the coefficients c_0, c_1, and c_2 such that $\langle \hat{\theta}_c, L_2 \rangle \equiv E[\hat{\theta}_c L_2] = 0$ for all $L_2 \in \Lambda_2$ (see [1] for definition of Λ_2). If we can find such coefficients that the estimator $\hat{\theta}_c$ will be orthogonal to Λ_2 and hence by Theorem 10.1 of [1] semiparametrically optimal. The good news is that we know (CITEME) that any element $L_2 \in \Lambda_2$ has a form:

[1] This estimator has slightly different coefficients compared to the initial estimator.

$$
L_2 = \left(\frac{\delta_{11}}{\pi_{11}}\pi_{00} - \delta_{00}\right) L_{20}(X) + \left(\frac{\delta_{11}}{\pi_{11}}\pi_{10} - \delta_{10}\right) L_{21}(X, Y_1) + \left(\frac{\delta_{11}}{\pi_{11}}\pi_{01} - \delta_{01}\right) L_{22}(X, Y_2). \tag{4}
$$

Then expanding and solving $E[\hat{\theta}_c L_2] = 0$ yields:

$$
\begin{aligned}
0 &= E[\hat{\theta}_c L_2] \\
& E\Bigg[\left(\frac{\pi_{00}}{\pi_{11}} + \left(\frac{\pi_{10}}{\pi_{10}+\pi_{11}}\right)\frac{\pi_{00}c_1}{\pi_{11}} + \left(\frac{\pi_{01}}{\pi_{01}+\pi_{11}}\right)\frac{\pi_{00}c_2}{\pi_{11}} + \frac{\pi_{00}(\pi_{10}\pi_{01} - \pi_{11}^2)c_0}{(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})\pi_{11}}\right) E[g \mid X] L_{20}(X) \\
& + \frac{\pi_{10}}{\pi_{11}}\left(E[g(Z)L_{21}(X,Y_1) \mid X] - c_1 E[E[g(Z) \mid X, Y_1] L_{21}(X,Y_1) \mid X] + \frac{\pi_{01}c_2}{\pi_{10}+\pi_{11}} E[E[g \mid X, Y_2] L_{21}(X,Y_1) \mid X] + \frac{\pi_{10}\pi_{01}}{\pi_{11}(\pi_{01}+\pi_{11})} E[g \mid X] E[L_{21}(X,Y_1) \mid X] c_0\right) \\
& + \frac{\pi_{01}}{\pi_{11}}\left(E[g(Z)L_{22}(X,Y_2) \mid X] + \frac{\pi_{10}c_1}{\pi_{10}+\pi_{11}} E[E[g(Z) \mid X, Y_1] L_{22}(X,Y_2) \mid X] - c_2 E[E[g \mid X, Y_2] L_{22}(X,Y_2) \mid X] + \frac{\pi_{10}}{(\pi_{01}+\pi_{11})} E[g \mid X] E[L_{22}(X,Y_2) \mid X] c_0\right)\Bigg]
\end{aligned}
$$

To solve for c_0, c_1, and c_2 we need the following to hold for any $L_{21}(X, Y_1)$ and $L_{22}(X, Y_2)$:

$$
\begin{aligned}
& 1 + c_0 \frac{\pi_{01}\pi_{10} - \pi_{11}^2}{(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})} + c_1 \frac{\pi_{10}}{\pi_{01}+\pi_{11}} + c_2 \frac{\pi_{01}}{\pi_{01}+\pi_{11}} = 0 \\
& E\left[\left(g(Z) + c_0 \frac{\pi_{01}}{\pi_{01}+\pi_{11}} E[g(Z) \mid X] - c_1 E[g(Z) \mid X, Y_1] + c_2 \frac{\pi_{01}}{\pi_{10}+\pi_{11}} E[g(Z) \mid X, Y_2]\right) L_{21}(X, Y_1) \mid X\right] = 0 \\
& E\left[\left(g(Z) + c_0 \frac{\pi_{10}}{\pi_{10}+\pi_{11}} E[g(Z) \mid X] + c_1 \frac{\pi_{10}}{\pi_{10}+\pi_{11}} E[g(Z) \mid X, Y_1] - c_2 E[g(Z) \mid X, Y_2]\right) L_{22}(X, Y_2) \mid X\right] = 0
\end{aligned}
$$

Optimal Normal Model

An alternative optimization problem is to find coefficients c_0, c_1, and c_2, that minimize the variance of $\hat{\theta}_c$. After expanding the variance and differentiating by each c_i, the optimal parameters are

$$
\begin{bmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \hat{c}_2 \end{bmatrix} = -\begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}^{-1} \times \begin{bmatrix} E[E[g \mid X]^2]\left(1 + \frac{\pi_{10}\pi_{01} - \pi_{11}^2}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}\right) \\ E[E[g \mid X, Y_1]^2]\left(\frac{-\pi_{10}}{\pi_{11}(\pi_{10}+\pi_{11})}\right) \\ E[E[g \mid X, Y_2]^2]\left(\frac{-\pi_{01}}{\pi_{11}(\pi_{01}+\pi_{11})}\right) \end{bmatrix}
$$

where

$$
\begin{aligned}
M_{11} &= E[E[g \mid X]^2]\left(\frac{\pi_{11}^2 + \pi_{10}\pi_{01}}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})} - 1\right)\\
M_{12} &= E[E[g \mid X]^2]\left(\frac{-\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}\right)\\
M_{13} &= E[E[g \mid X]^2]\left(\frac{-\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}\right)\\
M_{22} &= E[V(E[g \mid X, Y_1] \mid X)]\left(\frac{\pi_{10}}{\pi_{11}(\pi_{10}+\pi_{11})}\right)\\
M_{23} &= E[E[g \mid X, Y_1]E[g \mid X, Y_2]]\left(\frac{\pi_{10}\pi_{01}}{\pi_{11}(\pi_{10}+\pi_{11})(\pi_{01}+\pi_{11})}\right)\\
M_{33} &= E[V(E[g \mid X, Y_2] \mid X)]\left(\frac{\pi_{01}}{\pi_{11}(\pi_{01}+\pi_{11})}\right)
\end{aligned}
$$

References

[1] Anastasios A Tsiatis. “Semiparametric theory and missing data”. In: (2006).