A note on optimal estimation under non-monotone missingness by design

Jae-Kwang Kim

January 18, 2024

- Define
$$\delta_{i1} = \begin{cases} 1 & \text{if } Y_{1i} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$
and, similarly, we can define δ_{2i}.

- We define $A_{01} = \{i; \delta_{1i} = 0 \text{ and } \delta_{2i} = 1\}$.

- We are interested in estimating $\theta = E(Y_2)$. Note that, if we ignore A_2, then the missingness pattern is monotone and the following three-phase regression estimator can be used.
$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} g_1^*(x_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}}{\pi_{1i}} \{g_2^*(x_i, y_{1i}) - g_1^*(x_i)\} + \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{1i}\delta_{2i}}{\pi_{12i}} \{y_{2i} - g_2^*(x_i, y_i)\}$$
where $g_1^*(x) = E(Y_2 \mid x)$ and $g_2^*(x, y_1) = E(Y_2 \mid x, y_1)$.

- Now, to incorporate the information in A_{01}, we consider the following two-phase regression estimator
$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} g_1^*(x_i) + \frac{1}{n}\sum_{i=1}^{n} \frac{(1-\delta_{1i})\delta_{2i}}{\pi_{2i} - \pi_{12i}} \{y_{2i} - g_1^*(x_i)\}$$

- The two estimators are both unbiased. We can consider
$$\hat{\theta}_\alpha = \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2$$
and obtain the optimal choice of α that minimizes the variance.

- Please check Section 11.4 of the sampling book that I wrote.