Introduction

This document reports the optimality of the proposed non-monotone estimator. It turns out that the proposed estimator is not optimal. We show a better estimator and present a simulation study.

Setup

- Let $U = \{1, \dots N\}$ be a finite population.
- Consider the random variables $(X_i, Y_{1i}, Y_{2i}, \pi_i) \overset{\text{ind}}{\sim} F$ for some unknown F for all $i \in U$.
- Let $I_i \overset{\text{ind}}{\sim} \pi_i$ for $i \in U$ be the sample inclusion indicator.
- We assume that x_i is observed throughout the finite population.
- However, (y_{1i}, y_{2i}, π_i) is only observed if $I_i = 1$.
- The goal is to estimate $\theta = E[g(X, Y_1, Y_2)]$ in the population.
- The proposed estimator is

$$\begin{aligned}
\hat{\theta}_{\text{eff}} = n^{-1} &\sum_{i=1}^{n} E[g_i \mid X_i] \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}}{\pi_{1+}(X_i)} (E[g_i \mid X_i, Y_{1i}] - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{2i}}{\pi_{2+}(X_i)} (E[g_i \mid X_i, Y_{2i}] - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}R_{2i}}{\pi_{11}(X_i)} (g_i - E[g_i \mid X_i, Y_{1i}] - E[g_i \mid X_i, Y_{2i}] + E[g_i \mid X_i])
\end{aligned} \tag{1}$$

- The goal is to show that this estimator is optimal within the class of estimators:

$$\begin{aligned}
\hat{\theta}_{\text{eff}} = n^{-1} &\sum_{i=1}^{n} E[g_i \mid X_i] \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}}{\pi_{1+}(X_i)} (b(X_i, Y_{1i}) - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{2i}}{\pi_{2+}(X_i)} (a(X_i, Y_{2i}) - E[g_i \mid X_i]) \\
&+ n^{-1} \sum_{i=1}^{n} \frac{R_{1i}R_{2i}}{\pi_{11}(X_i)} (g_i - b_i - a_i + E[g_i \mid X_i]).
\end{aligned}$$

Results

I have done a lot to explore this area but I have several points of confusion. The first part recaps what Dr. Fuller and I discussed this past week.

- Let $g = E[Y_2]$, and suppose that we know the distribution of X and the covariance structure of $[X, Y_1, Y_2, \pi]$. Then instead of modeling the relationship between X and Y_k, we can use difference estimators: $W_1 \equiv Y_1 - b_1\tilde{X}$ and $W_2 =\equiv Y_2 - b_2\tilde{X}$, where $\tilde{X} = X - E[X]$. To minimize the variance of W_k we can choose the optimal value of b_k, which is

$$b_k = \frac{\text{Cov}(Y_k, X)}{\text{Var}(X)}.$$

  Since these values are known, b_k is known. This means that we now have the following table:

	W_1	W_2
A_{11}	✓	✓
A_{10}	✓	
A_{01}		✓
A_{00}		

  Because we assumed that the distribution of X is known, the section A_{00} contains no additional information about Y_1 or Y_2. Let $\mu_k = E[Y_k]$.

- We now consider a normal model:

$$\begin{pmatrix} x_i \\ e_{1i} \\ e_{2i} \end{pmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma_{11} & 0 \\ 0 & 0 & \sigma_{22} \end{bmatrix}, \right)$$

  and define $y_{1i} = x_i + e_{1i}$ and $y_{2i} = x_i + e_{2i}$. Then $b_1 = b_2 = 1$. We define $\bar{w}_k^{(ij)}$ as the mean of y_k in segment A_{ij}. This means that we have means $\bar{w}_1^{(11)}$, $\bar{w}_2^{(11)}$, $\bar{w}_1^{(10)}$, and $\bar{w}_2^{(01)}$. Let $W = [\bar{w}_1^{(11)}, \bar{w}_2^{(11)}, \bar{w}_1^{(10)}, \bar{w}_2^{(01)}]'$, then for $n_{ij} = |A_{ij}|$, we have

$$W - M\mu \sim N(\vec{0}, V)$$

  where

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } V = \begin{bmatrix} \frac{\sigma_{11}}{n_{11}} & 0 & 0 & 0 \\ 0 & \frac{\sigma_{22}}{n_{11}} & 0 & 0 \\ 0 & 0 & \frac{\sigma_{11}}{n_{10}} & 0 \\ 0 & 0 & 0 & \frac{\sigma_{22}}{n_{01}} \end{bmatrix}.$$

Thus, the BLUE for $\mu = [\mu_1, \mu_2]'$ is

$$\hat{\mu} = (M'V^{-1}M)^{-1}M'V^{-1}W.$$

- If we do the matrix algebra, $V^{-1} = \text{diag}\left(\frac{n_{11}}{\sigma_{11}}, \frac{n_{11}}{\sigma_{22}}, \frac{n_{10}}{\sigma_{11}}, \frac{n_{01}}{\sigma_{22}}\right)$ and hence,

$$\hat{\mu} = \begin{bmatrix} \frac{n_{11}\bar{w}_1^{(11)} + n_{10}\bar{w}_1^{(10)}}{n_{11}+n_{10}} \\ \frac{n_{11}\bar{w}_2^{(11)} + n_{01}\bar{w}_2^{(01)}}{n_{11}+n_{01}} \end{bmatrix} \tag{2}$$

- Since this is the BLUE, we would expect it to be at least as good as the proposed estimator. And if the proposed estimator is optimal it should be equivalent to the BLUE in the case that X, Y_1 and Y_2 are normal. So I ran a simulation study to test this. For $i = \{1, \dots, n = 1000\}$,

$$\begin{pmatrix} x_i \\ e_{1i} \\ e_{2i} \end{pmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \right) \text{ and } y_{1i} = x_i + e_{1i}, y_{2i} = x_i + e_{2i}.$$

Each observation i was then assigned a segment A_{11}, A_{10}, A_{01} or A_{00} independently with each draw having probability $p_{11} = 0.4$, $p_{10} = 0.2$, $p_{01} = 0.2$ and $p_{00} = 0.2$ respectively. This means that $\pi_{11} = 0.4$ and $\pi_{1+} = 0.6 = \pi_{2+}$. We let $g = E[Y_2]$ and we test the following estimators:

  - Oracle: This computes the average value of Y_2 in all observations (even when Y_2 is not supposed to be observed).

$$\hat{\theta} = n^{-1}\sum_{i=1}^{n} y_{2i}.$$

  - CC: This computes the average value of Y_2 in all observations in which Y_2 is observed.

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\delta_{2i}y_{2i}}{\sum_{i=1}^{n}\delta_{2i}}.$$

  - IPW: This computes the survey weighted average of Y_2 when both Y_1 and Y_2 are observed.

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\delta_{1i}\delta_{2i}y_{2i}}{\sum_{i=1}^{n}\delta_{1i}\delta_{2i}}.$$

  - Proposed: This is the proposed estimator from Equation 1.
  - WLS: This is the weighted linear estimator from Equation 2. (Note, since $g = E[Y_2]$ this only contains the second element from Equation 2.)

The results are shown in the table below. This simulation was run with the number of observations $n = 1000$ and the Monte Carlo sample size of $B = 1000$.

Table 1: This table shows the estimators of $\theta = E[Y_2]$. The true value of θ is 0. The bias column shows the average bias of the estimator and the actual value of $\theta = 0$ across the $B = 1000$ simulations. The SD column shows the average standard deviation across the $B = 1000$ simulations for each algorithm. The Tstat column displays the t-statistic of a t-test comparing the estimator to the actual value. The value of this column is computed via $\frac{\bar{\hat{\theta}}-\theta}{\sqrt{\text{Var}\hat{\theta}/B}}$. The Pval column displays the p-value of the t-statistic.

Algorithm	Bias	SD	Tstat	Pval
Oracle	-0.001	0.045	-0.953	0.170
CC	0.000	0.056	-0.059	0.476
IPW	0.000	0.071	-0.183	0.428
Prop	0.000	0.051	-0.131	0.448
WLS	-0.001	0.042	-0.470	0.319

- There are a couple things of note from Table 1. First, the WLS estimator has a smaller standard error than the Oracle estimator. This seems incorrect. (Please see the next section for questions.) Second, the proposed estimator did not have a similar standard error as the WLS estimator. This suggests that it is not optimal.

Questions for Dr. Fuller

1. For the linear estimator I don't think that I am incorporating sampling weights. I am a bit confused about how to do this. Should I compute the means $\bar{w}_k^{(ij)}$ using weights according to the sample weights?

2. Is it ok for the linear estimator to outperform the oracle estimator? This seems odd to me because the oracle estimator seems to contain more information.

3. On Tuesday, you asked me to compute the linear estimator in terms of means of Y and estimators of 0. I think that the WLS estimator is the estimator in terms of estimators of 0 because $y_k - b_k x$ is an estimator of 0. To transform the estimator into an estimator in terms of Y, should I write the form in terms of Y instead of W?