Non-Monotone GLS Simulation: No A_{00}

Caleb Leedy

1 February 2024

Introduction

The goal of this report is to conduct a simulation study to show the validity of using a GLS estimator when intermediate models do not also estimate the parameter in question. This setup is different from the previous setup is a couple of ways:

1. We do not use $E[g \mid G_r(Z)]$ as the g-functions. Instead, we have $g = (g_1, g_2, g_3)' = (X, Y_1, Y_2)'$.
2. We have fewer comparison estimators. Since we changed the g-functions it makes less sense to compare then to other estimators that are using different intermediate estimators.
3. We use a simple random sample (SRS) instead of a Poisson sample. For this setup, we have each segment being totally independent of each other. Each segment also has a fixed sample size of 250 instead of having segments with random sample sizes with a total observation count of 1000.
4. The GLS estimator is now only estimating $\theta = E[Y_2] = E[g_3]$.

Notation and Setup

Let $Z = (X, Y_1, Y_2)'$. We want to estimate the parameter $\theta = E[Y_2]$ where we may not always observe Y_1 and Y_2. Define segments that contain observations in which the same variables are observed as in Table 1.

Table 1: This table identifies which variables are observed in each segment. Since X is always observed, the subscript for each segment identifies which of variables Y_1 and Y_2 are in the segment based on the position of a 1.

Segment	Variables Observed
A_{00}	X
A_{10}	X, Y_1
A_{01}	X, Y_2
A_{11}	X, Y_1, Y_2

Let δ_{i,j_1,j_2} be the sample inclusion indicator for observation i in segment A_{j_1,j_2}, and let π_{j_1,j_2} be the probability of selecting an element into A_{j_1,j_2}.

We consider the vector $g(Z) = (g_1(X), g_2(Y_1), g_3(Y_2))'$ and *for this simulation setup*, let g_1, g_2, and g_3 all be the identity function $I()$. This means that $g(Z) = (X, Y_1, Y_2)'$. Notice, that we have $\theta = E[Y_2] = E[g_3]$. In each segment A_{j_1,j_2} we can obtain estimators of some of the g elements. We have the following:

$$g_1^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_1(x_i)$$

$$g_2^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_2(y_{1i})$$

$$g_3^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_3(y_{2i})$$

$$g_1^{(10)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{10}}{\pi_{10}} g_1(x_i)$$

$$g_2^{(10)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{10}}{\pi_{10}} g_2(y_{1i})$$

$$g_1^{(01)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{01}}{\pi_{01}} g_1(x_i)$$

$$g_3^{(01)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{01}}{\pi_{01}} g_2(y_{2i})$$

$$g_1^{(00)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{00}}{\pi_{00}} g_1(x_i)$$

This yields the following linear estimator,

$$\hat{g} = Zg + e$$

where

$$\hat{g} = \begin{bmatrix} g_1^{(11)} \\ g_2^{(11)} \\ g_3^{(11)} \\ g_1^{(10)} \\ g_2^{(10)} \\ g_1^{(01)} \\ g_3^{(01)} \\ g_1^{(00)} \end{bmatrix}, Z = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, E[e] = 0, \text{ and } \mathrm{Var}(e) = n^{-1} \begin{bmatrix} V_{11} & 0 & 0 & 0 \\ 0 & V_{10} & 0 & 0 \\ 0 & 0 & V_{01} & 0 \\ 0 & 0 & 0 & V_{00} \end{bmatrix}.$$

Here, we also have

$$V_{11} = \begin{bmatrix} \frac{1}{\pi_{11}} E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{11}} E[g_1 g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{11}} E[g_1 g_3] - E[g_1]E[g_3] \\ \frac{1}{\pi_{11}} E[g_1 g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{11}} E[g_2^2] - E[g_2]^2 & \frac{1}{\pi_{11}} E[g_2 g_3] - E[g_2]E[g_3] \\ \frac{1}{\pi_{11}} E[g_1 g_3] - E[g_1]E[g_3] & \frac{1}{\pi_{11}} E[g_2 g_3] - E[g_2]E[g_3] & \frac{1}{\pi_{11}} E[g_3^2] - E[g_3]^2 \end{bmatrix},$$

$$V_{10} = \begin{bmatrix} \frac{1}{\pi_{10}} E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{10}} E[g_1 g_2] - E[g_1]E[g_2] \\ \frac{1}{\pi_{10}} E[g_1 g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{10}} E[g_2^2] - E[g_2]^2 \end{bmatrix},$$

$$V_{01} = \begin{bmatrix} \frac{1}{\pi_{01}} E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{01}} E[g_1 g_3] - E[g_1]E[g_3] \\ \frac{1}{\pi_{01}} E[g_1 g_3] - E[g_1]E[g_3] & \frac{1}{\pi_{01}} E[g_3^2] - E[g_3]^2 \end{bmatrix}, \text{ and } V_{00} = \left[\frac{1}{\pi_{00}} E[g_1^2] - E[g_1]^2 \right].$$

Simulation

We use the following simulation setup

$$
\begin{aligned}
\begin{bmatrix} x \\ e_1 \\ e_2 \end{bmatrix} &\overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} \right) \\
y_1 &= x + e_1 \\
y_2 &= \mu + x + e_2
\end{aligned}
$$

This yields outcome variables Y_1 and Y_2 that are correlated both with X and additionally with each other. To generate the missingness pattern, we select 250 observations into the four segments independently.

Table 2: Results from simulations study with independent equally sized segments A_{11}, A_{10}, and A_{01} all of size $n = 250$. In this simulation we have the true mean of Y_2 equal to $\mu = 5$ and the covariance between e_1 and e_2 is $\rho = 0.5$. The goal is to estimate $E[Y_2] = \mu$. For the GLS estimation, we use the true known covariance matrix (using the true values of μ and ρ), and we use g-functions $g_1 = X$, $g_2 = Y_1$ and $g_3 = Y_2$.

Algorithm	Bias	SD	Tstat	Pval
Oracle	-0.001	0.049	-0.818	0.207
CC	-0.004	0.089	-1.317	0.094
IPW	-0.004	0.089	-1.317	0.094
GLS	-0.002	0.053	-1.490	0.068

A quick comparison between the standard deviation of the GLS estimator and the standard deviation of the GLS estimator in `gls_sim.qmd` shows that this standard deviation is slightly larger, which means that we do lose efficiency by not including A_{00}.