Finding Optimal f-Functions

Caleb Leedy

8 February 2024

Summary

This document takes the case of observing three variables (X_1, X_2, Y) under nonmonotone missingness and discusses the optimal function f_1, f_2, and f_3 under a simple random sampling design.

Notation and Problem Setup

Consider the case where $Z = (X_1, X_2, Y)$ and we want to estimate the parameter $\theta = E[Y]$. Suppose that we have a finite population of size N. Instead of observing the entire data set we observe the segments in Table 1. Each of the segments is a simple random sample of size n and independent from the other segments. Hence this means that the first order selection probability $\pi_i = \pi = n/N$ for every i.

Table 1: This table identifies which variables are observed in each segment. Since X_1 is always observed, the subscript for each segment identifies which of variables X_2 and Y are in the segment based on the position of a 1.

Segment	Variables Observed
A_{00}	X_1
A_{10}	X_1, X_2
A_{01}	X_1, Y
A_{11}	X_1, X_2, Y

As an analyst, we can choose functions f_1, f_2, f_3 such that $g = Zf + e$ where

$$g_1^{(11)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{11}}{\pi_{11}}f_1(x_{1i})$$
$$g_2^{(11)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{11}}{\pi_{11}}f_2(x_{2i})$$
$$g_3^{(11)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{11}}{\pi_{11}}f_3(y_i)$$
$$g_1^{(10)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{10}}{\pi_{10}}f_1(x_{1i})$$
$$g_2^{(10)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{10}}{\pi_{10}}f_2(x_{2i})$$
$$g_1^{(01)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{01}}{\pi_{01}}f_1(x_{1i})$$
$$g_3^{(01)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{01}}{\pi_{01}}f_2(y_i)$$
$$g_1^{(00)} = n^{-1}\sum_{i=1}^{n}\frac{\delta_{00}}{\pi_{00}}f_1(x_{1i})$$

where

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

For now assume that V is known. Due to the construction of g, we know that

$$V = \begin{bmatrix} V_{11} & 0 & 0 & 0 \\ 0 & V_{10} & 0 & 0 \\ 0 & 0 & V_{01} & 0 \\ 0 & 0 & 0 & V_{00} \end{bmatrix}$$

where

$$V_{11} = N^{-1}\pi^{-1}\mathrm{Cov}(f,f), V_{10} = N^{-1}\pi^{-1}\begin{bmatrix} \mathrm{Cov}(f_1,f_1) & \mathrm{Cov}(f_1,f_2) \\ \mathrm{Cov}(f_1,f_2) & \mathrm{Cov}(f_2,f_2) \end{bmatrix},$$

$$V_{01} = N^{-1}\pi^{-1}\begin{bmatrix} \mathrm{Cov}(f_1,f_1) & \mathrm{Cov}(f_1,f_3) \\ \mathrm{Cov}(f_1,f_3) & \mathrm{Cov}(f_3,f_3) \end{bmatrix}, \text{ and } V_{00} = N^{-1}\pi^{-1}\mathrm{Cov}(f_3,f_3).$$

Previously, we have shown that the optimal estimator is the GLS estimator $\hat{f} = (Z'V^{-1}Z)^{-1}Z'V^{-1}g$ and by linear model theory $\mathrm{Var}(\hat{f}) = (Z'V^{-1}Z)^{-1}$ since $\mathrm{Var}(g) = V$ by construction.