Estimating the Covariance Matrix

Caleb Leedy

5 February 2024

Summary

In this document, we

1. Propose a new technique for estimating a covariance matrix, and
2. Show via simulation studies that it works.

Problem and Proposal

Problem

Previously, have a model $\hat{g} = Zg + e$ where $\hat{g}$ is defined by

$$g_1^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_1(x_i)$$
$$g_2^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_2(y_{1i})$$
$$g_3^{(11)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{11}}{\pi_{11}} g_3(y_{2i})$$
$$g_1^{(10)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{10}}{\pi_{10}} g_1(x_i)$$
$$g_2^{(10)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{10}}{\pi_{10}} g_2(y_{1i})$$
$$g_1^{(01)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{01}}{\pi_{01}} g_1(x_i)$$
$$g_3^{(01)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{01}}{\pi_{01}} g_2(y_{2i})$$
$$g_1^{(00)} = n^{-1} \sum_{i=1}^{n} \frac{\delta_{00}}{\pi_{00}} g_1(x_i)$$

and

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

We also have

$$V_{11} = \begin{bmatrix} \frac{1}{\pi_{11}}E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{11}}E[g_1g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{11}}E[g_1g_3] - E[g_1]E[g_3] \\ \frac{1}{\pi_{11}}E[g_1g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{11}}E[g_2^2] - E[g_2]^2 & \frac{1}{\pi_{11}}E[g_2g_3] - E[g_2]E[g_3] \\ \frac{1}{\pi_{11}}E[g_1g_3] - E[g_1]E[g_3] & \frac{1}{\pi_{11}}E[g_2g_3] - E[g_2]E[g_3] & \frac{1}{\pi_{11}}E[g_3^2] - E[g_3]^2 \end{bmatrix},$$

$$V_{10} = \begin{bmatrix} \frac{1}{\pi_{10}}E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{10}}E[g_1g_2] - E[g_1]E[g_2] \\ \frac{1}{\pi_{10}}E[g_1g_2] - E[g_1]E[g_2] & \frac{1}{\pi_{10}}E[g_2^2] - E[g_2]^2 \end{bmatrix},$$

$$V_{01} = \begin{bmatrix} \frac{1}{\pi_{01}}E[g_1^2] - E[g_1]^2 & \frac{1}{\pi_{01}}E[g_1g_3] - E[g_1]E[g_3] \\ \frac{1}{\pi_{01}}E[g_1g_3] - E[g_1]E[g_3] & \frac{1}{\pi_{01}}E[g_3^2] - E[g_3]^2 \end{bmatrix}, \text{ and } V_{00} = \left[\frac{1}{\pi_{00}}E[g_1^2] - E[g_1]^2\right].$$

However, this challenge with actually using this model is that we have to use the matrix V, which we assume to be *known*. In this write up, we will not make this assumption.

Proposal

Instead of using the known covariance matrix V, we estimate it instead. Suppose that we observe variables $Z = (X_{m_1}, Y_{m_2})$ with an objective of estimating $\theta = E[g(Y_{m_2})$ for some known function g such that the variables Z are subject to missingness. We assume that there are R unique combinations of observed variables including a fully observed case. We can index the combinations of observed variables by r and assume that the fully observed case occurs at $r = 1$. Let $G_r(Z)$ be the variables that are observed at a particular value of r. We can choose a sequence of functions $f_1, \dots, f_K$ that we want to estimate. Each f_k is assumed to be a function of a subset of Z and it makes sense to assume (since these are chosen by the analyst) that there is at least one f_k for each combination of observed variables $G_r(Z)$. Let A_k be the sets of observed variable combinations that can be evaluate by the function f_k. If f_k can be evaluated by the observed variables $G_r(Z)$, this will consist of all of the combinations of variables r' such that $G_r(Z) \subseteq G'_r(Z)$. Because we assume that $G_1(Z) = Z$, A_k is always non-empty.

To estimate the covariance matrix V, we estimate the covariance between f_{k_1} and f_{k_2} directly by computing the estimated covariance on $A_{k_1} \cap A_{k_2}$.

Simulation

Non-Monotone Case

First, we consider the following non-monotone simulation setup.

$$
\begin{bmatrix} x_1 \\ e_1 \\ e_2 \end{bmatrix} \overset{ind}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} \right)
$$

$$
\begin{aligned}
x_2 &= x + e_1 \\
y &= \mu + x + e_2
\end{aligned}
$$

This yields outcome variables X_1 and X_2 that are correlated both with Y and additionally with each other. To generate the missingness pattern, we select 250 observations into the four segments independently using simple random sampling. The goal of this estimation problem is to estimate $\theta = E[Y]$.

Table 1: Results from simulations study with independent equally sized segments A_{11}, A_{10}, A_{01}, and A_{00} 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 estimated covariance matrix $\widehat{V}$ with f-functions $f_1 = X_1$, $f_2 = X_2$ and $f_3 = Y$.

Algorithm	Bias	SD	Tstat	Pval
Oracle	-0.002	0.044	-1.124	0.131
CC	-0.005	0.088	-1.820	0.035
IPW	-0.005	0.088	-1.820	0.035
GLS	-0.002	0.053	-1.286	0.099
GLSEstVar	-0.002	0.053	-1.271	0.102

Optimal choice of g-functions

TODO for the next set of simulations.

Conclusion

Did it work? Yes, estimating the covariance matrix did not produce a loss in efficiency.