Estimating the Covariance Matrix

Caleb Leedy

3 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

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

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

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

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

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

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

One potential solution would be to directly estimate V and use some sort of $\hat{V}$. However, this problem has a unique structure in that we get to choose the functional form both g and $\hat{g}$. Therefore, I propose the following: choose functions that after estimation are approximately independent from each other with a standard variance. If this is true, the covariance matrix V is approximately the identity matrix. Not only do we not have to estimate V, but we do not even need to invert it! I now give a more detailed explanation of this method.

Suppose that we observe variables $Z = (X_{m_1}, Y_{m_2})$ such that the variables are subject to missingness and 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.

Once we have the original $f_1, \dots, f_K$ we orthogonalize them using a Gram-Schmidt process. Let $g_1 = f_1 / \hat{\text{Var}}(f_1)$. Then, $g_2 = f_2 - \frac{\hat{\text{Cov}}(f_2, g_1)}{\hat{\text{Var}}(f_2)} g_1$, and we can continue using the sequence,

$$\tilde{g}_k = f_k - \sum_{i=1}^{k} \frac{\hat{\text{Cov}}(f_k, g_i)}{\hat{\text{Var}}(f_k)} g_i \text{ followed by } g_k = \tilde{g}_k / \hat{\text{Var}}(\tilde{g}_k).$$

However, the parameters for $\tilde{g}_k$ and $k > 1$ are all regression coefficients and the multiplier for g_k is simply computing a variance. Then to achieve efficiency, we propose estimating the regression coefficients by running the coresponding regression on all of the data points in A_k as well as computing the variance of $\tilde{g}_k$ within A_k.

Simulation

Conclusion

Did it work?