# Comparing the conservative Neyman variance estimator to the heteroskedastic robust variance estimator

Cyrus Samii

Consider an experiment on $N$ units in which $0 < M < N$ units are randomly assigned to treatment. Denote treatment status for unit $i$ with the random variable, $X_i \in \{0,1\}$, for which $\Pr(X_i = 1) = M/N$. Then, assign indices such that all the treated group come first, $X_1, ..., X_M = 1$ and the control group come after, $X_{M+1}, ..., X_N = 0$. We observe $Y_i = X_i y_{1i} + (1 - X_i) y_{0i}$, where $y_{1i}$ and $y_{0i}$ are unit $i$'s fixed "potential outcomes" under treatment and control, respectively. (The lower case emphasizes that they are fixed.) We want to estimate the average treatment effect, $\beta$, for this fixed population,

$$\beta = \frac{1}{N} \sum_{i=1}^{N} (y_{1i} - y_{0i})$$

It is well known that in this setting we can estimate $\beta$ without bias via the simple difference in treated versus control means, or via a regression of the $Y_i$'s on a constant and the $X_i$'s. The two average treatment effect estimators are algebraically equivalent. Call this estimator $\hat{\beta}$.

The so-called "Neymann conservative" estimator for the variance of $\hat{\beta}$ is given by,

$$\hat{V}_N(\hat{\beta}) = \frac{1}{M} \frac{1}{M-1} \sum_{i=1}^{M} e_i^2 + \frac{1}{N-M} \frac{1}{N-M-1} \sum_{M+1}^{N} e_i^2$$

where $e_i$ refers to the regression residual. The regression residual is algebraically equivalent to $Y_i - \frac{1}{M} \sum_{j=1}^{M} Y_j$ if $i \leq M$ (i.e., treated), and $Y_i - \frac{1}{N-M} \sum_{j=M+1}^{N} Y_j$ if $i > M$ (i.e., control). $V_N$ is known as a conservative estimator because it ignores the (unobserved) covariance between potential outcomes, and is therefore guaranteed to be larger than the true variance of $\hat{\beta}$. (Refer to the Freedman, Pisani, and Purves (1998) textbook for more on this.)

The so-called heteroskedastic robust regression estimator for the variance of $\hat{\beta}$ (as implemented in Stata, with the finite sample adjustment) reduces to,

$$\begin{aligned} \hat{V}_{HR}(\hat{\beta}) &= \tfrac{N}{N-2} \tfrac{1}{(N-M)^2} \left[ \left( \tfrac{N}{M} - 1 \right)^2 \textstyle\sum_{i=1}^{M} e_i^2 + \sum_{i=M+1}^{N} e_i^2 \right] \\ &= \tfrac{N}{N-2} \left( \tfrac{1}{M^2} \textstyle\sum_{i=1}^{M} e_i^2 + \tfrac{1}{(N-M)^2} \sum_{i=M+1}^{N} e_i^2 \right) \end{aligned}$$

The two variance estimators are algebraically equivalent when $N = 2M$. When such is not the case, they are not equivalent, though rather close.