Small sample correction methods for cluster-robust variance estimators and hypothesis tests

James E. Pustejovsky*
Department of Educational Psychology
University of Texas at Austin

and

Elizabeth Tipton
Department of Human Development
Teachers College, Columbia University

August 11, 2015

Abstract

The text of your abstract. 200 or fewer words.

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

Cluster-robust variance estimators (CRVE) and hypothesis tests based upon such estimators are ubiquitous in applied econometric work. Nearly every respectable paper in the past 15 years uses cluster-robust variance estimators because to do otherwise would be to risk being seen as insufficiently rigorous (or anti-conservative....ughh....how gauche!).

There's been a lot of fretting recently that even CRVE may actually not be rigorous enough. Cite the following people so as not to get their ire up:

- Brewer et al. (2013)
- Cameron et al. (2008)
- Cameron & Miller (2015)
- Carter et al. (2013)
- Ibragimov & Müller (2010)
- Imbens & Kolesar (2012)
- Kezdi (2004)
- McCaffrey et al. (2001)
- McCaffrey & Bell (2006)
- Webb & MacKinnon (2013)
- Kline & Santos (2012)

1.1 Econometric framework

We will consider linear regression models in which the errors within a cluster have an unknown variance structure. The model is

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\epsilon}_j, \tag{1}$$

for $j = 1, ..., m$, where $\mathbf{Y}_j$ is $n_j \times 1$, $\mathbf{X}_j$ is an $n_j \times p$ matrix of regressors for cluster j, $\boldsymbol{\beta}$ is a $p \times 1$ vector, and $\boldsymbol{\epsilon}_j$ is an $n_j \times 1$ vector of errors. Assume that $\mathrm{E}(\boldsymbol{\epsilon}_j | \mathbf{X}_j) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}_j | \mathbf{X}_j) = \boldsymbol{\Sigma}_j$, for $j = 1, ..., m$, where $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown, and the errors are independent across clusters. Let $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_m')'$ and $\boldsymbol{\Sigma} = \bigoplus_{j=1}^m \boldsymbol{\Sigma}_j$.

The vector of regression coefficients is estimated by weighted least squares (WLS). Given a set of m symmetric weighting matrices $\mathbf{W}_1, ..., \mathbf{W}_m$, the WLS estimator is

$$\hat{\boldsymbol{\beta}} = \mathbf{M} \sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{Y}_j, \tag{2}$$

where $\mathbf{M} = \left(\sum_{j=1}^m \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j \right)^{-1}$. Let $\mathbf{W} = \bigoplus_{j=1}^m \mathbf{W}_j$.

Common choices for weighting include the unweighted case, in which $\mathbf{W}_j$ is an identity matrix of dimension $n_j \times n_j$, and inverse-variance weighting under a working model. In the latter case, the errors are assumed to follow some known structure, $\mathrm{Var}(\mathbf{e}_j | \mathbf{X}_j) = \boldsymbol{\Phi}_j$, where $\boldsymbol{\Phi}_j$ is a known function of a low-dimensional parameter. The weighting matrices are then taken to be $\mathbf{W}_j = \hat{\boldsymbol{\Phi}}_j^{-1}$, where the $\hat{\boldsymbol{\Phi}}_j$ are constructed from estimates of the variance parameters.

The WLS estimator also encompasses the estimator proposed by Ibragimov & Müller (2010) for clustered data. Assuming that $\mathbf{X}_j$ has rank p, their proposed approach involves estimating $\boldsymbol{\beta}$ separately within each cluster and taking the simple average of these estimates. The resulting average is equivalent to the WLS estimator with weights $\mathbf{W}_j = \mathbf{X}_j \left(\mathbf{X}_j' \mathbf{X}_j \right)^{-2} \mathbf{X}_j$.

2 Cluster-robust variance estimators

The variance of the WLS estimator is

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{M} \left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \boldsymbol{\Sigma}_j \mathbf{W}_j \mathbf{X}_j \right) \mathbf{M}, \tag{3}$$

which includes the unknown variance matrices. One approach to estimating this variance would be to posit a working model–potentially the same working model used to construct weights–and substitute estimates of the working variance structure in place of $\boldsymbol{\Sigma}$. Under a

working model $\mathbf{\Phi}$, denote this "model-based" variance estimator as

$$\mathbf{V}^M = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\hat{\mathbf{\Phi}}_j\mathbf{W}_j\mathbf{X}_j\right)\mathbf{M}. \tag{4}$$

If $\boldsymbol{\beta}$ is estimated using inverse-variance weights defined under the same working model, then the model-based variance estimator simplifies to $\mathbf{V}^M = \mathbf{M}$.

Cluster-robust variance estimators provide a means of estimating Var $\left(\hat{\boldsymbol{\beta}}\right)$ and test hypotheses regarding $\hat{\boldsymbol{\beta}}$ in the absence of a valid working model for the error structure, or if the working variance model used to develop weights is mis-specified. They are thus a generalization of heteroskedasticity-consistent (HC) variance estimators. Like the HC estimators, several different variants have been proposed, with different rationales and different finite-sample properties

The most widely used estimator is

$$\mathbf{V}^{CR0} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{W}_j\mathbf{X}_j\right)\mathbf{M}, \tag{5}$$

where $\mathbf{e}_j = \mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}$. Following the naming conventions used by Cameron & Miller (2015), we will refer to this estimator as CR0. Note that CR0 is constructed by substituting $\mathbf{e}_j\mathbf{e}_j'$ in place of $\mathbf{\Sigma}_j$ in (3). Although the individual squared residuals provide only very crude estimates of the unknown variance matrices, the resulting estimator is asymptotically consistent for the variance of $\hat{\boldsymbol{\beta}}$ as m increases (CITE). However, CR0 is known to have a downward bias when the number of independent clusters is small (CITE).

The small-sample bias in CR0 can be seen as analogous to that arising by estimating the variance of a sample of m observations using a denominator of m rather than $m-1$. One approach to correcting this bias is to scale CR0 by a factor of $m/(m-1)$. Thus, define the CR1 estimator as $\mathbf{V}^{CR1} = [m/(m-1)]\mathbf{V}^{CR0}$. Some software implementations use a slightly different correction factor. For example, the Stata command `regress` scales CR0 by the factor $m(N-1)/[(m-1)(N-p)]$, where $N = \sum_{j=1}^{m} n_j$.

2.1 Correction based on a working-model

McCaffrey et al. (2001) proposed to correct the small-sample bias of CR0 so that it is exactly unbiased under a specified working model for the variance structure. In their

implementation, the residuals from each cluster are multiplied by an $n_j \times n_j$ adjustment matrix $\mathbf{A}_j$ that creates the unbiasedness property. The variance estimator, which we will call CR2a, is then

$$\mathbf{V}^{CR2a} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j'\mathbf{W}_j\mathbf{X}_j\right)\mathbf{M}, \tag{6}$$

The adjustment matrices are chosen to satisfy

$$\mathbf{A}_j'\left(\mathbf{I}-\mathbf{H}\right)_j\mathbf{\Phi}\left(\mathbf{I}-\mathbf{H}\right)_j'\mathbf{A}_j = \mathbf{\Phi}_j, \tag{7}$$

where $\mathbf{\Phi} = \bigoplus_{j=1}^{m}\mathbf{\Phi}_j$, $\mathbf{H} = \mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{W}$, and $\left(\mathbf{I}-\mathbf{H}\right)_j$ denotes the rows of $\mathbf{I}-\mathbf{H}$ corresponding to cluster j. However, this criterion does not uniquely define $\mathbf{A}_j$. Based on extensive simulations, McCaffrey et al. (2001) found that a symmetric solution worked well, with

$$\mathbf{A}_j = \left(\mathbf{\Phi}_j^C\right)'\mathbf{B}_j^{-1/2}\mathbf{\Phi}_j^C,$$

where $\mathbf{\Phi}_j^C$ is the upper triangular Cholesky factorization of $\mathbf{\Phi}_j$,

$$\mathbf{B}_j = \mathbf{\Phi}_j^C\left(\mathbf{I}-\mathbf{H}\right)_j\mathbf{\Phi}\left(\mathbf{I}-\mathbf{H}\right)_j'\left(\mathbf{\Phi}_j^C\right)',$$

and $\mathbf{B}_j^{-1/2}$ is the inverse of the symmetric square root of $\mathbf{B}$. If ordinary (unweighted) least squares is used to estimate $\hat{\boldsymbol{\beta}}$ and the working variance model posits that the errors are all independent and homoskedastic, then $\mathbf{W} = \mathbf{\Phi} = \mathbf{I}_N$ and $\mathbf{A}_j = \left(\mathbf{I}_j - \mathbf{H}_{jj}\right)^{-1/2}$, where $\mathbf{I}_j$ is an $n_j \times n_j$ identity matrix and $\mathbf{H}_{jj} = \mathbf{X}_j\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_j'$.

Two difficulties arise in the implementation of CR2a. First, the matrices $\mathbf{B}_1, ..., \mathbf{B}_m$ may not be positive definite, so that $\mathbf{B}_j^{-1/2}$ cannot be calculated for every cluster. This occurs, for instance, in balanced panel models when the specification includes fixed effects for each unit and each timepoint and clustering is over the units (Angrist & Pischke 2009, p. ??). However, this problem is easily overcome by using a generalized inverse of $\mathbf{B}_j$. A second, computational difficulty with CR2a is that it requires the inversion (or pseudo-inversion) of m matrices, each of dimension $n_j \times n_j$. Consequently, computation of CR2a will be slow if some clusters contain a large number of of individual units.

2.2 Another working-model correction

The criterion (7) is not the only way to obtain a variance estimator that is precisely unbiased under a working model. An alternative approach, which to our knowledge is novel, is to

use

$$\mathbf{V}^{CR2b} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{D}_j \mathbf{X}_j' \mathbf{W}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{W}_j \mathbf{X}_j \mathbf{D}_j'\right) \mathbf{M}, \tag{8}$$

where the adjustment matrices $\mathbf{D}_1, ..., \mathbf{D}_m$ are chosen so that

$$\mathbf{D}_j \mathbf{X}_j' \mathbf{W}_j \left(\mathbf{I} - \mathbf{H}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{W}_j \mathbf{X}_j \mathbf{D}_j' = \mathbf{X}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \mathbf{X}_j. \tag{9}$$

Just as with CR2a, there are several different forms of adjustment matrices that satisfy (9). A symmetric solution is to take

$$\mathbf{D}_j = \left(\mathbf{F}_j^C\right)' \left[\mathbf{F}_j^C \mathbf{G}_j \left(\mathbf{F}_j^C\right)'\right]^{-1/2} \mathbf{F}_j^C \tag{10}$$

where $\mathbf{F}_j = \mathbf{X}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \mathbf{X}_j$ and $\mathbf{G}_j = \mathbf{X}_j' \mathbf{W}_j \left(\mathbf{I} - \mathbf{H}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{W}_j \mathbf{X}_j$. Note that $\mathbf{F}_j$ and $\mathbf{G}_j$ might not be positive definite, and so a generalized form of the Cholesky decomposition and symmetric inverse-square root must be used. The $\mathbf{D}_j$ correction matrices will be defined (I think) so long as $\text{rank}(\mathbf{F}_j) \geq \text{rank}(\mathbf{G}_j)$ for $j = 1, ..., m$. In datasets where clusters have a large number of individual units, CR2b will be less computationally intensive than CR2a because the adjustment matrices are all of dimension $p \times p$.

References

Angrist, J. D. & Pischke, J. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton University Press, Princeton, NJ.

Brewer, M., Crossley, T. F. & Joyce, R. (2013), Inference with difference-in-differences revisited.

Cameron, A. C., Gelbach, J. B. & Miller, D. (2008), 'Bootstrap-based improvements for inference with clustered errors', *The Review of Economics and Statistics* **90**(3), 414–427.

Cameron, A. C. & Miller, D. L. (2015), A practitioner's guide to cluster-robust inference.

Carter, A. V., Schnepel, K. T. & Steigerwald, D. G. (2013), 'Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity', pp. 1–32.

Ibragimov, R. & Müller, U. K. (2010), 't-Statistic based correlation and heterogeneity robust inference', *Journal of Business & Economic Statistics* **28**(4), 453–468.

Imbens, G. W. & Kolesar, M. (2012), Robust standard errors in small samples: some practical advice.
URL: *http://www.nber.org/papers/w18478*

Kezdi, G. (2004), Robust standard error estimation in fixed-effects panel models.
URL: *http://papers.ssrn.com/sol3/Delivery.cfm?abstractid=596988*

Kline, P. & Santos, A. (2012), 'A score based approach to wild bootstrap inference', *Journal of Econometric Methods* **1**(1).

McCaffrey, D. F. & Bell, R. M. (2006), 'Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters.', *Statistics in medicine* **25**(23), 4081–98.

McCaffrey, D. F., Bell, R. M. & Botts, C. H. (2001), Generalizations of biased reduced linearization, *in* 'Proceedings of the Annual Meeting of the American Statistical Association', number 1994.

Webb, M. & MacKinnon, J. G. (2013), Wild Bootstrap Inference for Wildly Different Cluster Sizes.