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 18, 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), Bell & McCaffrey (2002)
- 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 $\boldsymbol{\Sigma} = \bigoplus_{j=1}^m \boldsymbol{\Sigma}_j$. Additionally, let $N = \sum_{j=1}^m n_j$, let $\mathbf{I}$ denote an $N \times N$ identity matrix, and let $\mathbf{I}_j$ denote an $n_j \times n_j$ identity matrix.

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 $\mathbf{W}_j = \mathbf{I}_j$ for $j = 1, ..., m$, 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 and $\boldsymbol{\Phi} = \bigoplus_{j=1}^m \boldsymbol{\Phi}_j$. 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 parameter.

The WLS estimator also encompasses the estimator proposed by Ibragimov & Müller (2010) for clustered data. Assuming that $\mathbf{X}_j$ has rank p for $j = 1, ..., m$, 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 depends upon the unknown variance matrices. One approach to estimating this variance would be to posit a working model—typically the same working model used to

construct weights—and substitute estimates of the working variance structure in place of $\boldsymbol{\Sigma}$. Under working model $\boldsymbol{\Phi}$, denote this "model-based" variance estimator as

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

If $\boldsymbol{\beta}$ is estimated using inverse-variance weights defined under the same working model, then $\mathbf{W}_j = \hat{\boldsymbol{\Phi}}_j^{-1}$ and 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 testing hypotheses regarding $\hat{\boldsymbol{\beta}}$ in the absence of a valid working model for the error structure, or when the working variance model used to develop weights is mis-specified. They are thus a generalization of heteroskedasticity-consistent (HC) variance estimators (MacKinnon & White 1985). 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 $\boldsymbol{\Sigma}_j$ 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 $m(N-1)/[(m-1)(N-p)]$.

2.1 Correction based on a working-model

McCaffrey et al. (2001, see also Bell & McCaffrey 2002) proposed to correct the small-

sample bias of CR0 so that it is exactly unbiased under a specified working model. In their implementation, the residuals from each cluster are multiplied by adjustment matrices $\mathbf{A}_1, ..., \mathbf{A}_m$ that are chosen to lead to 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 matrix $\mathbf{A}_j$ is of dimension $n_j \times n_j$ and satisfies

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

where $\mathbf{H} = \mathbf{X}\mathbf{M}\mathbf{X}'\mathbf{W}$, and $\left(\mathbf{I}-\mathbf{H}\right)_j$ denotes the rows of $\mathbf{I}-\mathbf{H}$ corresponding to cluster j. The criterion (7) does not uniquely define $\mathbf{A}_j$. Based on extensive simulations, McCaffrey et al. (2001) found that a symmetric solution worked well, with

$$\mathbf{A}_j = \left(\hat{\mathbf{\Phi}}_j^C\right)'\mathbf{B}_j^{-1/2}\hat{\mathbf{\Phi}}_j^C, \tag{8}$$

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

$$\mathbf{B}_j = \hat{\mathbf{\Phi}}_j^C\left(\mathbf{I}-\mathbf{H}\right)_j\hat{\mathbf{\Phi}}\left(\mathbf{I}-\mathbf{H}\right)_j'\left(\hat{\mathbf{\Phi}}_j^C\right)', \tag{9}$$

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

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. 320). However, this problem can be 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

$$\mathbf{V}^{CR2b} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{D}_j \mathbf{X}_j' \mathbf{W}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{W}_j \mathbf{X}_j \mathbf{D}_j'\right) \mathbf{M}, \tag{10}$$

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

$$\mathbf{D}_j \mathbf{X}_j' \mathbf{W}_j \left(\mathbf{I} - \mathbf{H}\right)_j \hat{\mathbf{\Phi}} \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{W}_j \mathbf{X}_j \mathbf{D}_j' = \mathbf{X}_j' \mathbf{W}_j \hat{\mathbf{\Phi}}_j \mathbf{W}_j \mathbf{X}_j. \tag{11}$$

Just as with CR2a, there are several different forms of adjustment matrices that satisfy (11). A symmetric solution is to take

$$\mathbf{D}_j = \left(\mathbf{F}_j^C\right)' \left[\mathbf{F}_j^C \mathbf{G}_j \left(\mathbf{F}_j^C\right)'\right]^{-1/2} \mathbf{F}_j^C \tag{12}$$

where $\mathbf{F}_j = \mathbf{X}_j' \mathbf{W}_j \hat{\mathbf{\Phi}}_j \mathbf{W}_j \mathbf{X}_j$ and $\mathbf{G}_j = \mathbf{X}_j' \mathbf{W}_j \left(\mathbf{I} - \mathbf{H}\right)_j \hat{\mathbf{\Phi}} \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{W}_j \mathbf{X}_j$. Note that $\mathbf{F}_j$ and $\mathbf{G}_j$ might not be positive definite, and so generalized forms of the Cholesky decomposition and symmetric inverse-square root must be used. 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 $p \times p$.

2.3 Jackknife correction

The Jackknife is an alternative, general-purpose approach to estimating Var $\left(\hat{\boldsymbol{\beta}}\right)$ under unknown variance structures (CITE). The jackknife variance estimator involves re-estimating the vector of regression coefficients m times, each time omitting a single cluster of data. Let $\hat{\boldsymbol{\beta}}_{(j)}$ denote the WLS estimator of $\boldsymbol{\beta}$ based on omitting cluster j. One form of the jackknife estimator is then

$$\mathbf{V}^{JK} = \frac{m-1}{m} \sum_{j=1}^{m} \left(\hat{\boldsymbol{\beta}}_{(j)} - \hat{\boldsymbol{\beta}}\right) \left(\hat{\boldsymbol{\beta}}_{(j)} - \hat{\boldsymbol{\beta}}\right)'. \tag{13}$$

This estimator can be expressed in the form of a sandwich estimator with adjusted residuals, and therefore generalizes the HC3 estimator in the heteroskedastic case. For simplicity, we omit the correction factor $(m-1)/m$ and define the CR3 estimator as

$$\mathbf{V}^{CR3} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \left(\mathbf{I}_j - \mathbf{H}_{jj}\right)^{-1} \mathbf{e}_j \mathbf{e}_j' \left(\mathbf{I}_j - \mathbf{H}_{jj}'\right)^{-1} \mathbf{W}_j \mathbf{X}_j\right) \mathbf{M}.$$

Bell & McCaffrey (2002) show that when $\left(\mathbf{I}_j - \mathbf{H}_{jj}\right)^{-1}$ exists for each $j = 1, ..., m$, then $\mathbf{V}^{JK} = [(m-1)/m]\mathbf{V}^{CR3}$.

2.4 Considerations with panel models

CRVEs are often used in connection with fixed effects panel data models. In such models, clusters correspond to repeated measures on individual units (e.g., yearly data describing each of the states in the U.S.), and the regression specification includes separate intercepts for each unit. One common model is

$$y_{jt} = \mathbf{r}_{jt}\boldsymbol{\alpha} + \gamma_j + \epsilon_{jt}$$

for $j = 1, ...m$ and $t = 1, ..., n_j$, where $\mathbf{r}_{ij}$ is an $r \times 1$ row vector of covariates. If the number and timing of the measurements is identical across cases, then the panel is balanced. Another common specification for balanced panels includes additional effects for each unique measurement occassion:

$$y_{jt} = \mathbf{r}_{jt}\boldsymbol{\alpha} + \gamma_j + \nu_t + \epsilon_{jt}$$

for $j = 1, ..., m$ and $t = 1, ..., n$. In what follows, we consider a generic fixed effects model in which

$$\mathbf{y}_j = \mathbf{R}_j\boldsymbol{\alpha} + \mathbf{S}_j\boldsymbol{\gamma} + \boldsymbol{\epsilon}_j, \tag{14}$$

where $\mathbf{R}_j$ is an $n_j \times r$ matrix of covariates, $\mathbf{S}_j$ is an $n_j \times s$ matrix describing the fixed effects specification, $\mathbf{X}_j = [\mathbf{R}_j\ \mathbf{S}_j]$, $\boldsymbol{\beta} = (\boldsymbol{\alpha}', \boldsymbol{\gamma}')'$, and $p = r + s$.

In fixed effects panel models, inferential interest is confined to $\boldsymbol{\alpha}$ and the fixed effects are treated as nuisance parameters. If the dimension of the fixed effects specification is large, it is computationally inefficient (and can be numerically inaccurate) to estimate $\boldsymbol{\beta}$ by ordinary or weighted least squares. Instead, it is useful to first absorb the fixed effects and then estimate $\boldsymbol{\alpha}$ on the reduced covariate vector. Although both approaches yield algebraically equivalent estimators of $\boldsymbol{\alpha}$, the small-sample adjustments to the CRVEs can differ depending on whether they are calculated based on the full covariate matrix or after absorbing the fixed effects. We view absorption is a computational device, rather than a distinct approach to estimation, and so it is useful to describe how to calculate the CRVEs when $\boldsymbol{\alpha}$ is estimated using absorption.

Let $\mathbf{H_S} = \mathbf{S}\left(\mathbf{S}'\mathbf{W}\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{W}$, $\ddot{\mathbf{Y}} = \left(\mathbf{I} - \mathbf{H_S}\right)\mathbf{Y}$, $\ddot{\mathbf{R}} = \left(\mathbf{I} - \mathbf{H_S}\right)\mathbf{R}$, $\mathbf{M}_{\ddot{\mathbf{R}}} = \left(\ddot{\mathbf{R}}'\mathbf{W}\ddot{\mathbf{R}}\right)^{-1}$, and $\mathbf{H}_{\ddot{\mathbf{R}}} = \ddot{\mathbf{R}}\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}'\mathbf{W}$. Using absorption, the WLS estimator of $\boldsymbol{\alpha}$ can be calculated as

$$\hat{\boldsymbol{\alpha}} = \mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}'\mathbf{W}\ddot{\mathbf{Y}}.$$

This estimator is algebraically equivalent to the corresponding sub-vector of $\hat{\boldsymbol{\beta}}$ calculated as in (2), based on the full covariate matrix $\mathbf{X}$. Furthermore, the residuals can be calculated from the absorbed model using $\mathbf{e} = \ddot{\mathbf{y}} - \ddot{\mathbf{R}}\hat{\boldsymbol{\alpha}}$. Let $\ddot{\mathbf{V}}^{CR0}$ denote the CR0 estimator calculated using $\ddot{\mathbf{R}}$ in place of $\mathbf{X}$, $\mathbf{M}_{\ddot{\mathbf{R}}}$ in place of $\mathbf{M}$, and $\ddot{\mathbf{e}} =$ in place of $\mathbf{e}$. It can be shown that $\ddot{\mathbf{V}}^{CR0}$ is algebraically equivalent to $\mathbf{V}^{CR0}$ calculated based on the full covariate matrix, as in (5). Because CR1 differs from CR0 by the constant factor $m/(m-1)$, it too is invariant to how $\hat{\boldsymbol{\alpha}}$ is calculated.

In contrast to CR0 and CR1, the CR2a estimator will differ depending on whether it is calculated based on the quantities from the absorbed model or those from the full WLS model. It is thus useful to define CR2a in such a way that the calculations based on the absorbed model yield algrebraically identical results to the calculations from the full WLS model. This can be accomplished by ensuring that the adjustment matrices given in Equation (8) are calculated based on the full covariate matrix $\mathbf{X}$. Specifically, in models with fixed effects, the adjustment matrices are calculated as

$$\mathbf{A}_j = \left(\hat{\boldsymbol{\Phi}}_j^C\right)' \left[\hat{\boldsymbol{\Phi}}_j^C \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right) \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \left(\hat{\boldsymbol{\Phi}}_j^C\right)'\right]^{-1/2} \hat{\boldsymbol{\Phi}}_j^C.$$

This formula avoids the need to calculate $\mathbf{H}$, which would involve inverting a $p \times p$ matrix.

Comment on whether this matters when fixed effects include only cluster indicators.

Like CR2a, the CR2b estimator is affected by whether it is calculated based on the absorbed model or the full WLS model. However, the structure of the adjustment matrices is such that their dimension can be reduced by focusing on the subset of covariates that are of inferential interest: instead of using $p \times p$ adjustment matrices, one can reduce their dimension to $r \times r$. Thus, in models with fixed effects, we propose to calculate CR2b as

$$\ddot{\mathbf{V}}^{CR2b} = \mathbf{M}_{\ddot{\mathbf{R}}} \left(\sum_{j=1}^{m} \ddot{\mathbf{D}}_j \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j \ddot{\mathbf{D}}_j'\right) \mathbf{M}_{\ddot{\mathbf{R}}}, \tag{15}$$

where the adjustment matrices are given by

$$\ddot{\mathbf{D}}_j = \left(\ddot{\mathbf{F}}_j^C\right)' \left[\ddot{\mathbf{F}}_j^C \ddot{\mathbf{G}}_j \left(\ddot{\mathbf{F}}_j^C\right)'\right]^{-1/2} \ddot{\mathbf{F}}_j^C, \tag{16}$$

$\ddot{\mathbf{F}}_j = \ddot{\mathbf{R}}_j' \mathbf{W}_j \hat{\boldsymbol{\Phi}}_j \mathbf{W}_j \ddot{\mathbf{R}}_j$ and $\ddot{\mathbf{G}}_j = \ddot{\mathbf{R}}_j' \mathbf{W}_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{s}}\right) \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\mathbf{s}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \mathbf{W}_j \ddot{\mathbf{R}}_j$. The CR2b estimator then has the property that $\mathrm{E}\left(\ddot{\mathbf{V}}^{CR2b}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\alpha}}\right)$ when the working model is correct.

In panel models that include fixed effects for each cluster, the CR3 variance estimator does not work when the model is estimated by WLS because $\mathbf{I}_j - \mathbf{H}_{jj}$ is not of full rank. However, the jackknife estimator of Var $(\hat{\boldsymbol{\alpha}})$ remains well-defined, as

$$\mathbf{V}^{JK} = \frac{m-1}{m} \sum_{j=1}^{m} \left(\hat{\boldsymbol{\alpha}}_{(j)} - \hat{\boldsymbol{\alpha}}\right) \left(\hat{\boldsymbol{\alpha}}_{(j)} - \hat{\boldsymbol{\alpha}}\right)', \tag{17}$$

where $\hat{\boldsymbol{\alpha}}_{(j)}$ is calculated by omitting cluster j. If the fixed effects specification consists entirely of cluster-level effects, or more generally, if $\mathbf{S}_j'\mathbf{S}_k = \mathbf{0}$ when $j \neq k$ for all $j, k = 1, ..., m$, then $\mathbf{V}^{JK} = [(m-1)/m]\ddot{\mathbf{V}}^{\mathbf{CR3}}$, where

$$\ddot{\mathbf{V}}^{CR3} = \mathbf{M}_{\ddot{\mathbf{R}}} \left(\sum_{j=1}^{m} \ddot{\mathbf{R}}_j' \mathbf{W}_j \left(\mathbf{I}_j - \mathbf{H}_{\ddot{\mathbf{R}}jj}\right)^{-1} \mathbf{e}_j \mathbf{e}_j' \left(\mathbf{I}_j - \mathbf{H}_{\ddot{\mathbf{R}}jj}'\right)^{-1} \mathbf{W}_j \ddot{\mathbf{R}}_j \right) \mathbf{M}_{\ddot{\mathbf{R}}} \tag{18}$$

and $\mathbf{H}_{\ddot{\mathbf{R}}jj} = \ddot{\mathbf{R}}_j \mathbf{M}_{\ddot{\mathbf{R}}} \ddot{\mathbf{R}}_j \mathbf{W}_j$. If the fixed effects specification includes terms that are not strictly nested within clusters, then $\ddot{\mathbf{V}}^{CR3}$ will not be exactly equivalent to $\mathbf{V}^{JK}$, and the former will depend to some extent on which terms are absorbed.

So we would recommend what?

3 Single-constraint hypothesis tests

Wald-type test statistics based on CRVEs are often used to test hypotheses regarding and construct confidence intervals for the coefficients in the regression specification. Such procedures are justified based on the asymptotic behavior of robust Wald statistics as the number of clusters grows large (i.e., $m \to \infty$). However, evidence from a wide variety of contexts indicates that the asymptotic results can be a very poor approximation when the number of clusters is small, even when small-sample corrections such as CR2a or CR3 are used (Bell & McCaffrey 2002, Bertrand et al. 2004, Cameron et al. 2008). Furthermore, the accuracy of asymptotic approximations depends on design features such as the degree of imbalance in the covariates, skewness of the covariates, and similarity of cluster sizes (McCaffrey et al. 2001, Tipton & Pustejovsky forthcoming, Webb & MacKinnon 2013). Consequently, no simple rule-of-thumb exists for what constitutes an adequate sample size to trust the asymptotic test.

We first consider testing single linear constraints (i.e., t-tests) on the parameter $\boldsymbol{\beta}$, in which the null hypothesis has the form $H_0 : \mathbf{c}'\boldsymbol{\beta} = d$ for fixed $p \times 1$ vector $\mathbf{c}$ and scalar

constant d. The Wald test statistic is then of the form

$$Z = \left(\mathbf{c}'\hat{\boldsymbol{\beta}} - d\right) / \sqrt{\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}}, \tag{19}$$

where $\mathbf{V}^{CR}$ is one of the CRVEs described in the previous section. An asymptotically valid test rejects H_0 at level α if $|Z|$ exceeds the $\alpha/2$ critical value of a standard normal distribution. However, this test tends to have actual rejection rates higher than α when m is not large.

3.1 Small-sample corrections

Four approaches to small-sample correction have been proposed for Wald-type t-tests. The first and surely most common approach is to compare $|Z|$ to the appropriate critical value from a t distribution with $m-1$ degrees of freedom. Hansen (2007) provided one justification for the use of a $t(m-1)$ reference distribution by identifying conditions under which Z converges in distribution to $t(m-1)$ as the within-cluster sample sizes grow large, with m fixed (see also Donald & Lang 2007). Ibragimov & Müller (2010) proposed a weighting technique derived so that that $t(m-1)$ critical values would be conservative (leading to rejection rates less than or equal to α). However, both of these arguments require that $\mathbf{c}'\boldsymbol{\beta}$ be separately identified within each cluster. Outside of these circumstances, using $t(m-1)$ critical values can still lead to over-rejection (Cameron & Miller 2015). Furthermore, this correction does not take into account that the distribution of $\mathbf{V}^{CR}$ is affected by the structure of the covariate matrix.

A second approach, proposed by McCaffrey et al. (2001), is to use a Satterthwaite approximation (Satterthwaite 1946) to the distribution of Z. This approach compares Z to a t reference distribution, with degrees of freedom ν that are estimated from the data. Theoretically, the degrees of freedom should be

$$\nu = \frac{2\left[\mathrm{E}\left(\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}\right)\right]^2}{\mathrm{Var}\left(\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}\right)}. \tag{20}$$

Expressions for the first two moments of $\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}$ can be derived under the assumption that the errors $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are normally distributed; see Appendix A. In practice, both moments involve the variance structure $\boldsymbol{\Sigma}$, which is unknown. McCaffrey et al. (2001) proposed

to estimate the moments based on a working variance model; for the CR2a and CR2b estimators, the same working model, $\mathbf{\Phi}$, is used to derive the adjustment matrices for the CRVE and to estimate the moments of that estimator. A "model-based" estimate of the degrees of freedom is then calculated as

$$\nu_M = \frac{\left(\sum_{j=1}^{m} \mathbf{t}_j' \hat{\mathbf{\Phi}} \mathbf{t}_j\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\mathbf{t}_i' \hat{\mathbf{\Phi}} \mathbf{t}_j\right)^2}, \tag{21}$$

where $\mathbf{t}_j = (\mathbf{I} - \mathbf{H})_j' \mathbf{A}_j' \mathbf{W}_j \mathbf{X}_j \mathbf{M}\mathbf{c}$ for CR2a or $\mathbf{t}_j = (\mathbf{I} - \mathbf{H})_j' \mathbf{W}_j \mathbf{X}_j \mathbf{D}_j' \mathbf{M}\mathbf{c}$ for CR2b. Alternately, for any of the CRVEs one could instead use an empirical, "plug-in" estimate of the degrees of freedom, constructed by substituting $\mathbf{e}_j \mathbf{e}_j'$ in place of $\mathbf{\Sigma}_j$. (This is similar to the Welch-Satterthwaite degrees of freedom estimate typically used for the two-sample t-test with unequal variances (Satterthwaite 1946).) However, Bell & McCaffrey (2002) found using simulation that the plug-in degrees of freedom estimate produced very conservative rejection rates. As a more refined empirical degrees of freedom estimate, we consider using $\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}$ as an estimate of its own expectation and estimating its variance by scaling the squares of the residual cross-products to account for their kurtosis. Specifically, we propose to use

$$\nu_E = \frac{\left(\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}\right)^2}{\sum_{h,i,j,k=1}^{m} a_{jk} \mathbf{t}_{hj}' \mathbf{e}_j \mathbf{e}_j' \mathbf{t}_{ij} \mathbf{t}_{hk} \mathbf{e}_k \mathbf{e}_k' \mathbf{t}_{ik}} - 2, \tag{22}$$

where $a_{jk} = (1 + 2(j = k))^{-1}$ and $\mathbf{t}_{hj}$ is the $n_j \times 1$ sub-vector of $\mathbf{t}_h$ corresponding to cluster j.

Note that Pan and Wall use a different degrees-of-freedom estimator, based on the sample variance of the variance contribution from each cluster. This makes sense in the GEE context because it avoids relying on normality of the errors.

Third, McCaffrey & Bell (2006) proposed to use a saddlepoint approximation to the distribution of Z. Like the Satterthwaite approximation, the saddlepoint approximation is derived under the assumption that the errors are normally distributed. Rather than using the moments of $\mathbf{c}'\mathbf{V}^{CR}\mathbf{c}$, the saddlepoint instead uses the fact that it is distributed as a weighted sum of χ_1^2 random variables. The weights depend on $\mathbf{\Sigma}$, and so must be estimated. McCaffrey & Bell (2006) did so based on a working model for the variance, in which case the weights are given by the eigen-values of the $m \times m$ matrix with $(i,j)^{th}$ entry $\mathbf{t}_i' \hat{\mathbf{\Phi}} \mathbf{t}_j$. An empirical approach could also be considered, in which case the weights are given by the eigen-values of the $m \times m$ matrix with $(i,j)^{th}$ entry $\sum_{k=1}^{m} \mathbf{t}_{ik}' \mathbf{e}_k \mathbf{e}_k' \mathbf{t}_{ij}$.

A final approach is to use a bootstrap re-sampling technique that leads to small-sample

refinements in the test rejection rates. Not all bootstrap re-sampling methods work well in small samples. Among the alternatives, Webb & MacKinnon (2013) describe a wild boostrap procedure that performs well even when m is very small and when clusters are of unequal size.

3.2 Simulation evidence

4 Multiple-constraint hypothesis tests

While t-tests of single coefficients are surely more common, tests of multiple constraints are also of interest for empirical data analysis. Examples of such tests include robust Hausmann-type endogeneity tests (Arellano 1993). We will consider linear constraints on $\boldsymbol{\beta}$, where the null hypothesis has the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ for fixed $q \times p$ matrix $\mathbf{C}$ and $q \times 1$ vector $\mathbf{d}$. The Wald statistic is then

Expand

$$Q = \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)' \left(\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'\right)^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right),$$

where $\mathbf{V}^{CR}$ is one of the CRVEs described in the previous section. The asymptotically valid Wald test rejects H_0 at level α if Q exceeds $\chi^2(\alpha; q)$, the α critical value from a chi-squared distribution with q degrees of freedom.

Citations to evidence that asymptotic test is way too liberal

4.1 Small-sample correction

Compared to single-constraint tests involving t, fewer approaches to small-sample correction are available for multiple-constraint tests. The saddlepoint approximation is not applicable due to the more complex structure of Q, which involves the matrix inverse of $\mathbf{V}^{CR}$. A simple correction, analogous to the first approach for t-tests, would be to compare Q/q to an $F(q, m-1)$ reference distribution. The wild bootstrap for clustered data (Webb & MacKinnon 2013) is also directly applicable to multiple-constraint tests, though to our knowledge its small-sample performance has not been assessed.

Worth mentioning the Cameron and Miller ad hoc approximation?

Several small-sample corrections for multiple-constraint Wald tests have been proposed that involve an F reference distribution with denominator degrees of freedom that are determined from the data. These approximations can thus be seen as generalizations

(loosely speaking) of the Satterthwaite approximation. Working in the context of CRVE for generalized estimating equations, Pan & Wall (2002) proposed to approximate the distribution of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ by a multiple of a Wishart distribution, from which it follows that Q approximately follows a multiple of an F distribution. Specifically, if $\eta\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ approximately follows a Wishart distribution with η degrees of freedom and scale matrix $\mathbf{C}\text{Var}\left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)\mathbf{C}'$, then

$$\left(\frac{\eta - q + 1}{\eta q}\right) Q \dot{\sim} F(q, \eta - q + 1). \tag{23}$$

We will refer to this as the approximate Hotelling's T^2 (AHT) test.

Just as in the Satterthwaite approximation, the degrees of freedom of the Wishart distribution are chosen to match the mean and variance of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$. However, when $q > 1$ it is not possible to exactly match both moments. Pan & Wall (2002) propose to use as degrees of freedom the value that minimizes the squared differences between the covariances among the entries of $\eta\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ and the covariances of the Wishart distribution with η degrees of freedom and scale matrix $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$. Zhang (2012*a*,*b*, 2013) proposed a simpler method in the context of heteroskedastic and multivariate analysis of variance models, which is a special case of the linear regression model considered here. The simpler approach involves matching the mean and total variance of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ (i.e., the sum of the variances of its entries), which avoids the need to calculate any covariances. Let $\mathbf{c}_1, ..., \mathbf{c}_q$ denote the $p \times 1$ row-vectors of $\mathbf{C}$. Denote the entries of $\mathbf{C}\mathbf{V}^{CR}\mathbf{C}'$ as $v_{st} = \mathbf{c}_s'\mathbf{V}^{CR}\mathbf{c}_t$, for $s, t = 1, ..., q$. The degrees of freedom are then given by

$$\eta = \frac{\sum_{s,t=1}^{q}\left[\text{E}^2\left(v_{st}\right) + \text{E}\left(v_{ss}\right)\text{E}\left(v_{tt}\right)\right]}{\sum_{s,t=1}^{q}\text{Var}\left(v_{st}\right)}, \tag{24}$$

which reduces to (20) if $q = 1$.

In practice, the moments of v_{st} must be estimated. As with single-case tests, both model-based and empirical estimates can be considered. Assuming that the errors $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are normally distributed with variance $\boldsymbol{\Phi}_j$, the model-based degrees of freedom are given by substituting $\mathbf{c}_s'\mathbf{V}^M\mathbf{c}_t$ for $\text{E}\left(v_{st}\right)$ and evaluating $\text{Var}\left(v_{st}\right)$ using Equation (26) with $\mathbf{u}_1 = \mathbf{c}_s$, $\mathbf{u}_2 = \mathbf{c}_t$, and $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Phi}}$. Still assuming normality of the errors, an empirical degrees of freedom estimate is given by using v_{st} as an estimate of $\text{E}\left(v_{st}\right)$ and estimating $\text{Var}\left(v_{st}\right)$ as

$$\sum_{h,i,j,k=1}^{m} a_{jk}\mathbf{t}_{hj}'\mathbf{e}_j\mathbf{e}_j'\mathbf{t}_{ij}\mathbf{t}_{hk}\mathbf{e}_k\mathbf{e}_k'\mathbf{t}_{ik}$$

4.2 Simulation evidence

5 Examples

6 Discussion

A Distribution theory for $\mathbf{V}^{CR}$

The small-sample approximations for t-tests and F-tests both involve the distribution of the entries of $\mathbf{V}^{CR}$. This section explains the relevant distribution theory.

First, note that any of the CRVEs can be written in the form $\mathbf{V}^{CR} = \sum_{j=1}^{M} \mathbf{T}_j \mathbf{e}_j \mathbf{e}_j' \mathbf{T}_j'$ for some $p \times n_j$ matrices $\mathbf{T}_j$. The form of the $\mathbf{T}_j$ matrices depends on which variance estimator is used: $\mathbf{T}_j = \mathbf{M}\mathbf{X}_j'\mathbf{W}_j$ for CR0, $\mathbf{T}_j = \mathbf{M}\mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j$ for CR2a, $\mathbf{T}_j = \mathbf{M}\mathbf{D}_j\mathbf{X}_j'\mathbf{W}_j$ for CR2b, and $\mathbf{T}_j = \mathbf{M}\mathbf{X}_j'\mathbf{W}_j\left(\mathbf{I}_j - \mathbf{H}_{jj}\right)^{-1}$ for CR3.

Next, let $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4$ be fixed, $p \times 1$ vectors and consider the linear combination $\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2$. Bell & McCaffrey (2002, Theorem 4) show that the linear combination is a quadratic form in $\mathbf{Y}$:

$$\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2 = \mathbf{Y}'\left(\sum_{j=1}^{m} \mathbf{t}_{2j}\mathbf{t}_{1j}'\right)\mathbf{Y},$$

for $N \times 1$ vectors $\mathbf{t}_{xj} = \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{T}_j'\mathbf{u}_x$, $x = 1, ..., 4$, and $j = 1, ..., m$.

Standard results regarding quadratic forms can be used to derive the moments of the linear combination. We now assume that $\boldsymbol{\epsilon}_1, ..., \boldsymbol{\epsilon}_m$ are multivariate normal with zero mean and variance $\boldsymbol{\Sigma}$. It follows that

$$\mathrm{E}\left(\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2\right) = \sum_{j=1}^{m} \mathbf{t}_{1j}'\boldsymbol{\Sigma}\mathbf{t}_{2j} \tag{25}$$

$$\mathrm{Var}\left(\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{2j}\right)^2 + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{1j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{2j} \tag{26}$$

$$\mathrm{Cov}\left(\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2, \mathbf{u}_3'\mathbf{V}^{CR}\mathbf{u}_4\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{3j} + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{3j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}. \tag{27}$$

Furthermore, the distribution of $\mathbf{u}_1'\mathbf{V}^{CR}\mathbf{u}_2$ can be expressed as a weighted sum of χ_1^2 distributions, with weights given by the eigen-values of the $m \times m$ matrix with $(i,j)^{th}$ entry $\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{2j}$, $i, j = 1, ..., m$.

References

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

Arellano, M. (1993), 'On the testing of correlated effects with panel data', *Journal of Econometrics* **59**(1-2), 87–97.

Bell, R. M. & McCaffrey, D. F. (2002), 'Bias reduction in standard errors for linear regression with multi-stage samples', *Survey Methodology* **28**(2), 169–181.

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics* **119**(1), 249–275.

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.

Donald, S. G. & Lang, K. (2007), 'Inference with difference-in-differences and other panel data', *Review of Economics and Statistics* **89**(2), 221–233.

Hansen, C. B. (2007), 'Asymptotic properties of a robust variance matrix estimator for panel data when T is large', *Journal of Econometrics* **141**, 597–620.

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).

MacKinnon, J. G. & White, H. (1985), 'Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**, 305–325.

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.

Pan, W. & Wall, M. M. (2002), 'Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations.', *Statistics in medicine* **21**(10), 1429–41.

Satterthwaite, F. E. (1946), 'An approximate distribution of estimates of variance components', *Biometrics bulletin* **2**(6), 110–114.

Tipton, E. & Pustejovsky, J. E. (forthcoming), 'Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression', *Journal of Educational and Behavioral Statistics* .

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

Zhang, J.-T. (2012*a*), 'An approximate degrees of freedom test for heteroscedastic two-way ANOVA', *Journal of Statistical Planning and Inference* **142**(1), 336–346.

Zhang, J.-T. (2012*b*), 'An approximate Hotelling T2 -test for heteroscedastic one-way MANOVA', *Open Journal of Statistics* **2**, 1–11.

Zhang, J.-T. (2013), 'Tests of linear hypotheses in the ANOVA under heteroscedasticity', *International Journal of Advanced Statistics and Probability* **1**(2), 9–24.