Small sample hypothesis testing using cluster-robust variance estimation

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

and

Elizabeth Tipton
Department of Human Development
Teachers College, Columbia University

August 27, 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

While the focus of much economics research is on understanding the causes and correlates of the behaviors of individuals, the data encountered in empirical applications is often clustered. For example, individuals are often clustered by countries, regions, or states; by firms, organizations, or schools; or by time-periods or follow-up waves. This clustering is typically accounted for in analyses through the use of cluster robust variance estimation (CRVE), an analog to the heteroscedasticity robust standard errors developed by Huber (1967), Eicker (1967), and White (1980) to account for non-constant variance in ordinary least squares. The use of CRVE is widespread, as evidenced by the large number of citations to key articles in the field (e.g., 849 cites for Wooldridge 2003), the large number of citations overall (i.e., over 11,000 for ”clustered standard errors” in Google Scholar), and the large number of articles employing the methods in economics journals (i.e., over 500 citations).

Are we trying to reference the panel data case here? Clustering by time period doesn't seem quite right.

Add citation

CRVE is routinely used in order to test hypotheses involving either individual coefficients or sets of multiple constraints on the regression specification. The theory behind CRVE is asymptotic in the number of clusters, and recently, researchers have turned attention to the performance of these tests in small and moderate samples. Cameron & Miller (2015) provide a thorough review of this literature, including a discussion of current practice, possible solutions, and open problems. Among their principle conclusions are that conventional CRVE have downward bias in small samples and that hypothesis tests based on CRVEs can have Type-I error rates that are far from the nominal level of the test. Moreover, they argue that the small-sample corrections for t-tests typically found in software such as Stata and SAS are inadequate. In the course of reviewing a variety of recent proposals for addressing these problems, Cameron and Miller highlight a potentially promising method called bias-reduced linearization method (BRL), introduced by McCaffrey et al. (2001) and Bell & McCaffrey (2002). BRL entails correcting the downward bias of the most common CRVE so that it is exactly unbiased under a working model specified by the analyst, while also remaining asymptotically consistent under arbitrary true variance structures. Simulations reported Bell & McCaffrey (2002) demonstrate that the BRL correction serves to reduce the bias of the CRVE even when the working model is misspecified. The same authors also proposed and studied small-sample corrections to t-tests based on the BRL variance estimator, based on Satterthwaite Bell & McCaffrey (2002) or saddlepoint approximations (McCaffrey & Bell 2006).

Despite promising simulation evidence that BRL performs well (e.g., Imbens & Kolesar 2012), several problems arise in implementing the method in practice. First, the BRL adjustment breaks down for some common use-cases that arise in analysis of panel data (Angrist & Pischke 2009). Second, casual application of BRL for analysis of panel data can b

Specifically....

Explain Angrist and Pischke problem, panel data problem, F-tests

The goal of this paper is to fully articulatethe BRL methodology so that it is suitable for everyday econometric practice. We make three principle contributions. First, we demonstrate that using generalized inverses to calculate the BRL adjustment matrices address the rank-deficiency problem, while retaining the property that CRVEs based on the adjustment matrices are unbiased under a working model. Second, we describe how to apply BRL when fitting models that include fixed effects, which are absorbed prior to parameter estimation. Third, we propose a method of testing multiple-constraint hypotheses (i.e., F-tests) based on CRVE with the BRL adjustments. Such tests are commonly used in the analysis of experiments—for example, for testing baseline equivalence, when testing multiple outcomes by seemingly unrelated regression (SUR), and when there are multiple treatment groups—as well as in analysis of panel data (e.g., Hausman tests). In support of this third aim, we provide simulation evidence that the proposed small-sample F-test offers drastic improvements over commonly implemented alternatives and performs comparably with current state-of-the-art methods such as the cluster-wild bootstrap procedure described by Cameron et al. (2008) and Webb & MacKinnon (2013). To date, the Wild bootstrap (and other resampling methods) are the 'best practice' with small samples, and we show that the BRL method performs just as well statistically.

”flesh out?”

Does it?

The organization of the paper is as follows.... We conclude the paper with a set of three examples comparing results from these three approaches to illustrate the breadth of application, and a discussion of important considerations for practice.

2 CLUSTER-ROBUST VARIANCE ESTIMATION

2.1 Econometric framework

We will consider linear regression models in which the errors within a cluster have an unknown variance structure. Suppose that there are $j = 1, ..., m$ clusters, each with n_j observations. In cluster j and observation i, assume that the outcome y_{ij} is related to a vector of p covariates $\mathbf{x}_{ij}$ by

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + \boldsymbol{\epsilon}_{ij}. \tag{1}$$

By stacking the outcomes, covariate vectors, and errors, the model can be written more compactly as,

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta} + \boldsymbol{\epsilon}_j, \tag{2}$$

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. Importantly, the covariate matrix $\mathbf{X}_j$ can include a wide variety of covariate forms, including those that vary at the cluster or observation level, as well as fixed effects for each cluster (or groups within each cluster).

We assume that $\mathrm{E}\left(\boldsymbol{\epsilon}_j \left| \mathbf{X}_j \right.\right) = \mathbf{0}$ and $\mathrm{Var}\left(\boldsymbol{\epsilon}_j \left| \mathbf{X}_j \right.\right) = \boldsymbol{\Sigma}_j$, for $j = 1, ..., m$, where the form of $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown but the errors are independent across clusters. In many cases, the errors are assumed to follow some known structure, $\mathrm{Var}\left(\mathbf{e}_j \left| \mathbf{X}_j \right.\right) = \boldsymbol{\Phi}_j$, where $\boldsymbol{\Phi}_j$ is a known function of a low-dimensional parameter.

Is this the right place to introduce the working model?

We shall consider estimating the vector of regression coefficients β using weighted least squares (WLS). For each cluster j, let $\mathbf{W}_j$ be a symmetric, $n_j \times n_j$ weighting matrix. The WLS estimate can then be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{M} \sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{Y}_j, \tag{3}$$

where $\mathbf{M} = \left(\sum_{j=1}^m \mathbf{X}_j' \mathbf{W}_j \mathbf{X}_j\right)^{-1}$. This WLS framework includes the unweighted case (where $\mathbf{W}_j = \mathbf{I}_j$, the identity matrix), as well as feasible GLS. In the latter case, 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.[1]

[1] The WLS estimator also encompasses the estimator proposed by Ibragimov & Müller (2010) for clus-

The variance of the WLS estimator is

$$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\boldsymbol{\Sigma}_j\mathbf{W}_j\mathbf{X}_j\right)\mathbf{M}, \tag{4}$$

which depends upon the unknown variance matrices $\boldsymbol{\Sigma}_j$. One approach to estimating this variance is model-based. In this approach, it is assumed that $\text{Var}\left(\mathbf{e}_j\left|\mathbf{X}_j\right.\right) = \boldsymbol{\Phi}_j$, where $\boldsymbol{\Phi}_j$ is a known function of a low-dimensional parameter, which is then estimated. For example, a hierarchical error structure is common, wherein observations in the same cluster share a random effect. If this approach is used, each $\boldsymbol{\Sigma}_j$ is substituted with the estimate $\hat{\boldsymbol{\Phi}}_j$. If additionally $\mathbf{W}_j = \hat{\boldsymbol{\Phi}}_j^{-1}$, the model-based variance estimator can be shown to simplify to $\mathbf{V}^M = \mathbf{M}$. However, if the working model is mis-specified, the model-based variance estimator will be inconsistent and inferences based upon it will be invalid.

2.2 Extant CRVEs

Cluster-robust variance estimators provide a means of estimating $\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$ 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. Each of these are of the form

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

for some n_j by n_j adjustment matrix $\mathbf{A}_j$. The form of these adjustments parallels those of the heteroscedastity-consistent (HC) variance estimators proposed by MacKinnon & White (1985). Letting $\mathbf{A}_j = \mathbf{I}_j$, the identity matrix, results in the original CRVE estimator; following Cameron and Miller (2015), we refer to this estimator as $\mathbf{V}^{CR0}$. If instead, we set $\mathbf{A}_j = c\mathbf{I}_j$, where $c = \sqrt{(m/(m-1))(N/(N-p))}$, where $N = \sum_{j=1}^m n_j$, this results in the CRV1 estimator, $\mathbf{V}^{CR1}$. Note that when N is large, here $c \approx \sqrt{m/(m-1)}$; this correction

tered 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 $\mathbf{W}_j = \mathbf{X}_j\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-2}\mathbf{X}_j$.

is the most commonly implemented in practice (e.g., including Stata, SAS). Importantly, this correction does not depend on $\mathbf{X}_j$ and is the same for all hypotheses tested. Like the CR0 estimator, however, this estimator often under-estimates the true variance.

An alternative correction, akin to MacKinnon and White's CR2 estimator, is the BRL method provided by Bell and McCaffrey (2002). The $\mathbf{V}^{CR2}$ estimator defines $\mathbf{A}_j$ as the matrix that satisfies,

$$\mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j'\left(\mathbf{I}-\mathbf{H}\right)_j\boldsymbol{\Sigma}\left(\mathbf{I}-\mathbf{H}\right)_j'\mathbf{A}_j\mathbf{W}_j\mathbf{X}_j = \mathbf{X}_j'\mathbf{W}_j\boldsymbol{\Sigma}_j\mathbf{W}_j\mathbf{X}_j, \tag{6}$$

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.

.

Need to define X,W

Importantly, determining $\mathbf{A}_j$ depends on knowledge of $\boldsymbol{\Sigma}_j$, which is unknown (and thus the reason for using the CRVE approach). In order to make progress, Bell and McCaffrey proposed to define $\mathbf{A}_j$ under an assumed structure to $\boldsymbol{\Sigma}_j$, known as a ”working” model. When this working model (which we now call $\boldsymbol{\Phi}_j$) is correct and $\mathbf{A}_j$ is defined following (Eqn), then it can be shown that the $\mathbf{V}^{CR2}$ estimator is unbiased for $\mathbf{V}$ (see Eqn X of BM 2002). When the assumed structure deviates from the true covariance $\boldsymbol{\Sigma}_j$, the estimator remains biased, though Bell and McCaffrey show that the bias is greatly reduced (thus the name ”bias reduced linearization”). Extensive simulation results indicate that this bias is typically minimal, even for large deviations from the assumed structure (CITE).

Following previous notation, this focus on a working model means we can write $\boldsymbol{\Sigma}_j = \boldsymbol{\Phi}_j$, which is a low-level function of variance parameters that can be estimated. Bell and McCaffrey further note that the criterion (10) does not uniquely define $\mathbf{A}_j$. Based on extensive simulations, McCaffrey et al. (2001) found that a symmetric solution worked well, with

$$\mathbf{A}_j = \left(\hat{\boldsymbol{\Phi}}_j^C\right)'\mathbf{B}_j^{-1/2}\hat{\boldsymbol{\Phi}}_j^C, \tag{7}$$

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

$$\mathbf{B}_j = \hat{\boldsymbol{\Phi}}_j^C\left(\mathbf{I}-\mathbf{H}\right)_j\hat{\boldsymbol{\Phi}}\left(\mathbf{I}-\mathbf{H}\right)_j'\left(\hat{\boldsymbol{\Phi}}_j^C\right)', \tag{8}$$

and $\mathbf{B}_j^{-1/2}$ is the inverse of the symmetric square root of $\mathbf{B}_j$. To be more concrete, in the simplest case of ordinary (unweighted) least squares in which the working variance model posits that the errors are all independent and homoskedastic, then we can show that $\mathbf{W} = \mathbf{\Phi} = \mathbf{I}$ and $\mathbf{A}_j = \left(\mathbf{I}_j - \mathbf{X}_j \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}_j'\right)^{-1/2}$. In the remainder of this paper, we will focus on this BRL approach, using the $\mathbf{V}^{CR2}$ estimator throughout.

Two difficulties arise in the implementation of CR2. 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 CR2 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.

Expand on this.

Finally, note that a third estimator, CR3, is also available; this estimator corresponds to the jacknife, and has been shown (both analytically and through extensive simulations) to over-estimate the variance. The BRL approach thus sits between the CR1 and CR3 estimators, providing a nearly unbiased method for estimating the variance. These adjustments to the CRVE estimator, however, do not wholely address the small-sample hypothesis testing problem. In the next sections, we review a degrees of freedom estimation strategy for t-tests, originally provided by Bell and McCaffrey (2002), and then introduce a similar strategy for F-tests. The work presented for F-tests is new, and we argue, together with the t-test case provides a unified framework for using the BRL method in practice.

3 Bias-reduced linearization

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 CR2, is then

$$\mathbf{V}^{CR2} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j'\mathbf{W}_j\mathbf{X}_j\right)\mathbf{M}, \tag{9}$$

The adjustment matrix $\mathbf{A}_j$ is of dimension $n_j \times n_j$ and satisfies

$$\mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j'\left(\mathbf{I}-\mathbf{H}\right)_j\hat{\mathbf{\Phi}}\left(\mathbf{I}-\mathbf{H}\right)_j'\mathbf{A}_j\mathbf{W}_j\mathbf{X}_j = \mathbf{X}_j'\mathbf{W}_j\hat{\mathbf{\Phi}}_j\mathbf{W}_j\mathbf{X}_j, \tag{10}$$

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 (10) does not uniquely define $\mathbf{A}_j$. Based on extensive simulations, McCaffrey et al. (2001) found that a symmetric solution worked well, with

$$\mathbf{A}_j = \left(\hat{\mathbf{\Phi}}_j^C\right)'\mathbf{B}_j^{-1/2}\hat{\mathbf{\Phi}}_j^C, \tag{11}$$

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

$$\mathbf{B}_j = \hat{\mathbf{\Phi}}_j^C\left(\mathbf{I}-\mathbf{H}\right)_j\hat{\mathbf{\Phi}}\left(\mathbf{I}-\mathbf{H}\right)_j'\left(\hat{\mathbf{\Phi}}_j^C\right)', \tag{12}$$

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 $\mathbf{A}_j = \left(\mathbf{I}_j - \mathbf{X}_j\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_j'\right)^{-1/2}$.

3.1 Implementation of CR2 in panel models

The t-test and F-test developed here have wide application in economic analyses. One area of application is in panel data models, wherein repeated measures on individual units are often captured (e.g., yearly data describing each of the states in the U.S.). Work by XXXXX highlighted that in these situations, best practice is to account for the clusters both through their inclusion in the model (i.e., either cluster fixed effects) and through use of CRVE. As Cameron and Miller (2015) highlight, however, an important question is how the inclusion of these cluster fixed effects might affect the small sample properties of the test statistics of main interest in small samples.

In the panel-data model, 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

$$\mathbf{y}_j = \mathbf{R}_j\boldsymbol{\alpha} + \mathbf{S}_j\boldsymbol{\gamma} + \boldsymbol{\epsilon}_j, \tag{13}$$

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 (or "demean" the data) and then estimate $\boldsymbol{\alpha}$ on the reduced covariate vector. While both approaches yield algebraically equivalent estimators of $\boldsymbol{\alpha}$, absorption is computationally less intensive and is therefore the standard method in many software programs (e.g., Stata). Cameron and Miller (2015) note, however, that using the standard small-sample adjustments to CRVE (i.e., CR1), including the clusters as fixed effects or absorbing them can lead different standard errors. To see why, recall that in CR1 adjustments, $\mathbf{A}_j = \sqrt{((m/(m-1))(N/(N-p)))}$. Following this approach, the adjustment depends on p, which is larger when the estimates are includes as fixed effects and smaller when instead absorbtion is used. In cases in which the number of observations per cluster is small the differences can be quite large. For example, as Cameron and Miller indicate, when $n_j = 2$ for all clusters, this can result in (CR1 based) standard errors over twice as large when using cluster fixed effects versus absorbtion.

As we will show here, a benefit of using the BRL approach is that the CR2 estimator is not affected by the inclusion of clusters as fixed effects or through absorption. To see how, 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 (3), 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 CITE.

In contrast to CR0, it is possible that the CR2 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 it 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 (11) are calculated based on the full covariate matrix $\mathbf{X}$. Specifically, in models with fixed effects, the adjustment matrices are calculated as in (11), but with

$$\mathbf{B}_j = \hat{\boldsymbol{\Phi}}_j^C \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right) \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \left(\hat{\boldsymbol{\Phi}}_j^C\right)'. \tag{14}$$

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

It is unnecessary to account for absorption of fixed effects under certain commonly occurring circumstances. Specifically, if the model is estimated using weighted least-squares with working inverse-variance weights, and if absorption is performed only for fixed effects that are equivalent to or nested within the units on which clusters are defined, then the adjustment matrices can be calculated directly from Equations (11) and (12), using $\mathbf{H}_{\ddot{\mathbf{R}}}$ in place of $\mathbf{H}$. This result is formalized in the following theorem:

Theorem. Consider model (13) and let $\ddot{\mathbf{V}}^{CR2}$ be the CR2 matrix calculated based on the absorbed model, i.e.,

$$\ddot{\mathbf{V}}^{CR2} = \mathbf{M}_{\ddot{\mathbf{R}}} \left(\sum_{j=1}^m \ddot{\mathbf{R}}_j' \mathbf{W}_j \ddot{\mathbf{A}}_j \mathbf{e}_j \mathbf{e}_j' \ddot{\mathbf{A}}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j\right) \mathbf{M}_{\ddot{\mathbf{R}}},$$

where $\ddot{\mathbf{A}}_j = \hat{\boldsymbol{\Phi}}_j^{C'} \ddot{\mathbf{B}}_j^{-1/2} \hat{\boldsymbol{\Phi}}_j^C$ and $\ddot{\mathbf{B}}_j = \hat{\boldsymbol{\Phi}}_j^C \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \hat{\boldsymbol{\Phi}}_j^{C'}$. Let $\mathbf{J}$ be the $p \times r$ matrix that selects the covariates of interest, i.e., $\mathbf{XJ} = \mathbf{R}$ and $\mathbf{J}'\boldsymbol{\beta} = \boldsymbol{\alpha}$. Assume that $\mathbf{W}_j = \hat{\boldsymbol{\Phi}}_j^{-1}$ for $j = 1, ..., m$ and that $\mathbf{S}_i \mathbf{M}_{\mathbf{S}} \mathbf{S}_j' \mathbf{W}_j = \mathbf{0}$ for every $i \neq j$. Then $\ddot{\mathbf{V}}^{CR2} = \mathbf{J}'\mathbf{V}^{CR2}\mathbf{J}$.

Appendix B provides a proof. When the necessary conditions hold, this approach is preferable for reasons of numerical precision.

4 HYPOTHESIS TESTING

Wald-type test statistics based on CRVEs are often used to test hypotheses regarding 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 CR2 are employed (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 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 based on CR2 is then

$$Q = \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right)' \left(\mathbf{C}\mathbf{V}^{CR2}\mathbf{C}'\right)^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}\right).$$

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 corrections for t-tests

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

$$\nu = \frac{2\left[\mathrm{E}\left(\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}\right)\right]^2}{\mathrm{Var}\left(\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}\right)}. \tag{15}$$

Expressions for the first two moments of $\mathbf{c}'\mathbf{V}^{CR2}\mathbf{c}$ 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 the same working model as used to derive the adjustment matrices. A "model-based" estimate of the degrees of freedom is then calculated as

$$\nu_M = \frac{\left(\sum_{j=1}^m \mathbf{s}_j'\hat{\boldsymbol{\Phi}}\mathbf{s}_j\right)^2}{\sum_{i=1}^m \sum_{j=1}^m \left(\mathbf{s}_i'\hat{\boldsymbol{\Phi}}\mathbf{s}_j\right)^2}, \tag{16}$$

where $\mathbf{s}_j = \left(\mathbf{I} - \mathbf{H}\right)_j' \mathbf{A}_j'\mathbf{W}_j\mathbf{X}_j\mathbf{M}\mathbf{c}$. Alternately, for any of the CRVEs one could instead use an empirical estimate of the degrees of freedom, constructed by substituting $\mathbf{e}_j\mathbf{e}_j'$ in place of $\boldsymbol{\Sigma}_j$. However, Bell & McCaffrey (2002) found using simulation that the plug-in degrees of freedom estimate produced very conservative rejection rates.

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^2_1 random variables. The weights depend on $\boldsymbol{\Sigma}$, 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 $\mathbf{s}_i'\hat{\boldsymbol{\Phi}}\mathbf{s}_j$.

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.

Need to describe the bootstrap in more detail.

4.2 Small-sample corrections for F-tests

While t-tests of single coefficients are surely most 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), tests for non-linearities in exogeneous variables in OLS models, tests for pre-treatment balance on covariates in randomized experiments, and tests of parameter restrictions in seemingly unrelated regression (SUR). It is useful, therefore, to also have a small-sample F-test available that aligns with the BRL approach introduced in the previous section.

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?

Compared to single-constraint tests, fewer approaches to small-sample correction are available for multiple-constraint tests. A simple correction, analogous to the CR1 for t-tests, would be to compare Q/q to an $F(q, m-1)$ reference distribution. As we will show in our simulation study, like the t-test case, this test tends to be overly liberal.

The ideal adjustment, therefore, would be to determine empirically the degrees of freedom of the F distribution using an approach similar to that for the BRL t-test. In the broad literature, several small-sample corrections for multiple-constraint Wald tests of this form have been proposed. Working in the context of CRVE for generalized estimating equations, Pan & Wall (2002) proposed to approximate the distribution of $\mathbf{C}\mathbf{V}^{CR2}\mathbf{C}'$ by a multiple of a Wishart distribution, from which it follows that Q approximately follows a multiple of an F distribution. Specifically, if $\eta\mathbf{C}\mathbf{V}^{CR2}\mathbf{C}'$ 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

$$\left(\frac{\eta - q + 1}{\eta q}\right) Q \mathrel{\dot{\sim}} F(q, \eta - q + 1). \tag{17}$$

We will refer to this as the approximate Hotelling's T^2 (AHT) test, and the remainder of this section will develop this test in greater detail.

Just as in the Satterthwaite approximation, in this test, 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 $\mathbf{C}$. Let $\mathbf{t}_{sh} = (\mathbf{I} - \mathbf{H})_h' \mathbf{A}_h' \mathbf{W}_h \mathbf{X}_h \mathbf{M} \mathbf{c}_s$ for $s = 1, ..., q$ and $h = 1, ..., m$. The degrees of freedom are then estimated under the working model as

$$\eta_M = \frac{\sum_{s,t=1}^{q} \sum_{h,i=1}^{m} b_{st} \mathbf{t}_{sh}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{th} \mathbf{t}_{si}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{ti}}{\sum_{s,t=1}^{q} \sum_{h,i=1}^{m} \mathbf{t}_{sh}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{ti} \mathbf{t}_{sh}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{ti} + \mathbf{t}_{sh}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{si} \mathbf{t}_{th}' \hat{\boldsymbol{\Omega}} \mathbf{t}_{ti}}, \tag{18}$$

where $b_{st} = 1 + (s = t)$ for $s, t = 1, .., q$. Note that η_M reduces to ν_M if $q = 1$.

This F-test shares features with the t-test developed by Bell and McCaffrey. Like the t-test, the degrees of freedom of this F-test depend non only on the number of clusters, but also on features of the covariates being tested. Again, these degrees of freedom can be much smaller than $m - 1$, and are particularly smaller when the covariates being tested exhibit high imbalances or leverage. Unlike the t-test case, however, in multi-parameter case, it is often more difficult to diagnose the cause of these small degrees of freedom. In some situations, however, these are straightforward extensions to the findings in t-tests. For example, if the goal is to test if there are differences across a four-arm treatment study, the degrees of freedom are largest (and close to $m - 1$) when the treatment is allocated equally across the four groups within each cluster. When the proportion varies across clusters, these degrees of freedom fall, often leading to degrees of freedom in the "small sample" territory even when the number of clusters is large. In the next section, we will illustrate these principles in a simulation study.

5 Simulation evidence

6 EXAMPLES

In this section we examine three short examples of the use of CRVE with small samples, spanning a variety of applied contexts. In the first example, the effects of substantive interest are identified within each cluster. In the second example, the effects involve between-cluster contrasts. The third example involves a cluster-robust Hausmann test for differences between within- and across-cluster information. In each example, we illustrate how the proposed small-sample t- and F-tests can be used and how they can differ from both the standard CR1 and Wild bootstrap tests. R code and data files are available for each analysis as an online supplement.

6.0.1 Tennessee STAR class-size experiment.

The Tennessee STAR class size experiment is one of the most well studied interventions in education. In the experiment, K 3 students and teachers were randomized within each of 79 schools to one of three conditions: small class-size (targetted to have 13-17 students), regular class-size, or regular class-size with an aide (see Schazenbach, 2006 for a review). Analyses of the original study and follow up waves have found that being in a small class improves a variety of outcomes, including higher test scores (Schanzenbach 2006), increased likelihood of taking college entrance exams (Krueger & Whitmore 2001), and increased rates of home ownership and earnings (Chetty et al. 2011).

The class-size experiment consists of three treatment conditions and multiple, student-level outcomes of possible interest. The analytic model is

$$Y_{ijk} = \mathbf{z}_{jk}'\boldsymbol{\alpha}_i + \mathbf{x}_{jk}'\boldsymbol{\beta} + \gamma_k + \epsilon_{ijk} \tag{19}$$

For outcome i, student j is found in school k; $\mathbf{z}_{jk}$ includes dummies for the small-class and regular-plus-aide conditions; and the vector $\mathbf{x}_{jk}$ includes a set of student demographics (i.e., free or reduced lunch status; race; gender; age). Following Krueger (1999), we put the the reading, word recognition, and math scores on comparable scales by converting each outcome to percentile rankings based upon their distributions in the control condition.

We estimated the model in two ways. First, we estimated $\boldsymbol{\alpha}_i$ separately for each outcome i and tested the null hypothesis that $\boldsymbol{\alpha}_i = \mathbf{0}$. Second, we use the seemingly unrelated regression (SUR) framework to test for treatment effects across conditions, using a simultaneous test across outcomes. In the SUR model, separate treatment effects are estimated for each outcome, but the student demographic effects and school fixed effects are pooled across outcomes. An overall test of the differences between conditions thus amounts to testing the null hypothesis that $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \mathbf{0}$. In all models, we estimated $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}$ after absorbing the school fixed effects and clustered the errors by school.

6.0.2 Heterogeneous treatment impacts

Angrist & Lavy (2009) reported results from a randomized trial in Israel aimed at increasing matriculation certification for post-secondary education among low achievers. In the Achievement Awards demonstration, 40 non-vocational high schools with the lowest 1999 certification rates nationally were selected (but with a minimum threshold of 3%). This included 10 Arab and 10 Jewish religious schools and 20 Jewish secular schools. The 40 schools were then pair-matched based on the 1999 certification rates, and within each pair one school was randomized to receive a cash-transfer program. In these treatment schools, every student who completed certification was eligible for a payment. The total amount at stake for a student who passed all the milestones was just under $2,400.

Baseline data was collected in January 2001 with follow up data collected in June 2001 and 2002. Following Angrist & Lavy (2009), we focus on the number of certification tests taken as the outcome and report results separately for girls, for boys, and for the combined sample. Given that the program took place in three different types of schools, in this example we focus on determining if there is evidence of variation in treatment impacts across types of schools (i.e., Jewish secular, Jewish religious, and Arab). We use the analytic model:

$$Y_{ij} = \mathbf{z}_j'\boldsymbol{\alpha} + T_j\mathbf{z}_j\boldsymbol{\delta} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \epsilon_{ij} \tag{20}$$

In this model for student i in school j, $\mathbf{z}_j$ is a vector of dummies indicating school type; T_j is a treatment dummy indicating if school j was assigned to the treatment condition; and $\mathbf{x}_{ij}$ contains individual student demographics (i.e., mothers and fathers education; immigration status; number of siblings; and an indicator for the quartile of their pre-test achievement from previous years). The components of $\boldsymbol{\delta}$ represent the average treatment impacts in Jewish secular, Jewish religious, and Arab schools. We test the null hypothesis that $\delta_1 = \delta_2 = \delta_3$ to determine if the treatment impact differs across school types. In the second panel of Table 1 we provide the results of this test separately for boys and girls and by year. Importantly, note that the 2000 results are baseline tests, while the 2001 and 2002 results measure the effectiveness of the program.

Add note about program being discontinued in 2002

6.0.3 Robust Hausmann test

In this final example, we shift focus from analyses of experiments to panel data. Here we build off of an example first developed in Bertrand et al. (2004) using Current Population Survey (CPS) data to relate demographics to earnings. Following Cameron & Miller (2015), we aggregated the data from the individual level to the time period, producing a balanced panel with 36 time points within 51 states (including the District of Columbia). We focus on the model,

$$Y_{tj} = \mathbf{r}_{tj}'\boldsymbol{\alpha} + \gamma_j + \epsilon_{ij}. \tag{21}$$

In this model, time-point t is nested within state j; the outcome Y_{tj} is log-earnings, which are reported in 1999 dollars; $\mathbf{r}_{tj}$ includes a vector of demographic covariates specific to the time point (i.e., dummy variables for female and white; age and age-squared); and γ_j is a fixed effect for state j.

For sake of example, we focus here on determining whether to use a fixed effects (FE) estimator or a random effects (RE) estimator the four parameters in $\boldsymbol{\alpha}$, based on a Hausmann test. In an OLS model with uncorrelated, the Hausmann test directly compares the vectors of FE and RE estimates using a chi-squared test. However, this specification fails when cluster-robust standard errors are employed, and instead an artificial-Hausman test

(Arellano 1993) is typically used (Wooldridge 2002, pp. 290-291). This test instead amends the model to additionally include within-cluster deviations (or cluster aggregates) of the variables of interest. In our example, this becomes,

$$Y_{tj} = \mathbf{r}_{tj}'\boldsymbol{\alpha} + \ddot{\mathbf{r}}_{tj}\boldsymbol{\beta} + \gamma_j + \epsilon_{tj}, \tag{22}$$

where $\ddot{\mathbf{r}}_{tj}$ denotes the vector of within-cluster deviations of the covariates (i.e., $\ddot{\mathbf{r}}_{tj} = \mathbf{r}_{tj} - \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_{tj}$). The four parameters in $\boldsymbol{\beta}$ represent the differences between the within-panel and between-panel estimates of $\boldsymbol{\alpha}$. The artificial Hausmann test therefore reduces to testing the null hypothesis that $\boldsymbol{\beta} = \mathbf{0}$ using an F test with $q = 4$. We estimate the model using WLS with weights derived under the assumption that $\gamma_1, ..., \gamma_J$ are mutually independent, normally distributed, and independent of ϵ_{tj}.

7 DISCUSSION

While it's odd to think about using a working model in combination with CRVE, it does put a little bit more emphasis on attending to modeling assumptions, which is probably a good thing.

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 $\mathbf{V}^{CR2}$. This section explains the relevant distribution theory.

First, note that the CR2 estimator can be written in the form $\mathbf{V}^{CR2} = \sum_{j=1}^{M}\mathbf{T}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{T}_j'$ for $p \times n_j$ matrices $\mathbf{T}_j = \mathbf{M}\mathbf{X}_j'\mathbf{W}_j\mathbf{A}_j$. Let $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4$ be fixed, $p \times 1$ vectors and consider the linear combination $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$. Bell & McCaffrey (2002, Theorem 4) show that the linear combination is a quadratic form in $\mathbf{Y}$:

$$\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2 = \mathbf{Y}'\left(\sum_{j=1}^{m}\mathbf{t}_{2j}\mathbf{t}_{1j}'\right)\mathbf{Y},$$

for $N \times 1$ vectors $\mathbf{t}_{sh} = (\mathbf{I} - \mathbf{H})_h'\,\mathbf{T}_h'\mathbf{c}_s$, $s = 1, ..., 4$, and $h = 1, ..., m$.

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

$$\begin{aligned}
\mathrm{E}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) &= \sum_{j=1}^m \mathbf{t}_{1j}'\boldsymbol{\Sigma}\mathbf{t}_{2j} && (23)\\
\mathrm{Var}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) &= \sum_{i=1}^m \sum_{j=1}^m \left(\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{2j}\right)^2 + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{1j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{2j} && (24)\\
\mathrm{Cov}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2, \mathbf{c}_3'\mathbf{V}^{CR}\mathbf{c}_4\right) &= \sum_{i=1}^m \sum_{j=1}^m \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{3j} + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{3j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}. && (25)
\end{aligned}$$

Furthermore, the distribution of $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$ can be expressed as a weighted sum of χ^2_1 distributions, with weights given by the eigen-values of the $m \times m$ matrix with $(i,j)^{th}$ entry $\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{2j}$, $i,j = 1,...,m$.

B CR2 invariance

This appendix provides a theorem that identifies circumstances under which it is unnecessary to account for fixed effect absorption when calculating the adjustment matrices used in $\mathbf{V}^{CR2}$.

Formulas for the inverse of a partitioned matrix can be used to demonstrate that $\mathbf{X}_j\mathbf{M}\mathbf{J} = \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}$. Thus, equivalence of $\ddot{\mathbf{V}}^{CR2}$ and $\mathbf{J}'\mathbf{V}^{CR2}\mathbf{J}$ follows if $\mathbf{A}_j = \ddot{\mathbf{A}}_j$ for $j = 1,...,m$.

From the fact that $\ddot{\mathbf{R}}_j'\mathbf{W}_j\mathbf{S}_j = \mathbf{0}$ for $j = 1,...,m$, it follows that

$$\begin{aligned}
\mathbf{B}_j &= \hat{\boldsymbol{\Phi}}_j^C \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right) \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\mathbf{S}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \hat{\boldsymbol{\Phi}}_j^{C'}\\
&= \hat{\boldsymbol{\Phi}}_j^C \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}} - \mathbf{H}_{\mathbf{S}}\right)_j \hat{\boldsymbol{\Phi}} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}} - \mathbf{H}_{\mathbf{S}}\right)_j' \hat{\boldsymbol{\Phi}}_j^{C'}\\
&= \hat{\boldsymbol{\Phi}}_j^C \left(\boldsymbol{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j' - \mathbf{S}_j\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\right) \hat{\boldsymbol{\Phi}}_j^{C'}
\end{aligned}$$

and

$$\mathbf{B}_j^{-1} = \left(\hat{\boldsymbol{\Phi}}_j^{C'}\right)^{-1} \left(\boldsymbol{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j' - \mathbf{S}_j\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\right)^{-1} \left(\hat{\boldsymbol{\Phi}}_j^{C}\right)^{-1}. \qquad (26)$$

Let $\mathbf{U}_j = \left(\boldsymbol{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j'\right)^{-1}$. Using a generalized Woodbury identity (Henderson & Searle 1981),

$$\mathbf{U}_j = \mathbf{W}_j - \mathbf{W}_j\ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}} \left(\mathbf{M}_{\ddot{\mathbf{R}}} - \mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j\mathbf{W}_j\ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\right)^{-} \mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j'\mathbf{W}_j,$$

where M^{-} is a generalized inverse of $\mathbf{M}$. It follows that $\mathbf{U}_j\mathbf{S}_j = \mathbf{W}_j\mathbf{S}_j$. Another application of the generalized Woodbury identity gives

$$\begin{aligned}\left(\mathbf{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j' - \mathbf{S}_j\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\right)^{-1} &= \mathbf{U}_j - \mathbf{U}_j\mathbf{S}_j\mathbf{M}_{\mathbf{S}}\left(\mathbf{M}_{\mathbf{S}} - \mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\mathbf{U}_j\mathbf{S}_j\mathbf{M}_{\mathbf{S}}\right)^{-}\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\mathbf{U}_j \\ &= \mathbf{U}_j - \mathbf{W}_j\mathbf{S}_j\mathbf{M}_{\mathbf{S}}\left(\mathbf{M}_{\mathbf{S}} - \mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\mathbf{W}_j\mathbf{S}_j\mathbf{M}_{\mathbf{S}}\right)^{-}\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\mathbf{W}_j \\ &= \mathbf{U}_j.\end{aligned}$$

The last equality follows from the fact that $\mathbf{S}_j\mathbf{M}_{\mathbf{S}}\left(\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\mathbf{W}_j\mathbf{S}_j\mathbf{M}_{\mathbf{S}} - \mathbf{M}_{\mathbf{S}}\right)^{-}\mathbf{M}_{\mathbf{S}}\mathbf{S}_j' = \mathbf{0}$ because the fixed effects are nested within clusters. Substituting into (26), we then have that $\mathbf{B}_j^{-1} = \left(\hat{\mathbf{\Phi}}_j^{C'}\right)^{-1}\mathbf{U}_j\left(\hat{\mathbf{\Phi}}_j^{C}\right)^{-1}$. Now, $\ddot{\mathbf{B}}_j = \hat{\mathbf{\Phi}}_j^{C}\mathbf{U}_j^{-1}\hat{\mathbf{\Phi}}_j^{C'}$ and so $\ddot{\mathbf{B}}_j^{-1} = \mathbf{B}_j^{1}$. It follows that $\ddot{\mathbf{A}}_j = \mathbf{A}_j$ for $j = 1, ..., m$.

References

Angrist, J. D. & Lavy, V. (2009), 'The effects of high stakes high school achievement awards : Evidence from a randomized trial', *American Economic Review* **99**(4), 1384–1414.

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.

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.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. & Yagan, D. (2011), 'How does your kindergarten classroom affect your earnings? Evidence from Project STAR', *The Quarterly Journal of Economics* **126**(4), 1593–1660.

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

Eicker, F. (1967), Limit theorems for regressions with unequal and dependent errors, *in* 'Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, pp. 59–82.

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.

Henderson, H. V. & Searle, S. R. (1981), 'On deriving the inverse of a sum of matrices', *Siam Review* **23**(1), 53–60.

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*

Krueger, A. & Whitmore, D. (2001), 'The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR', *The Economic Journal* **111**(468), 1–28.

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.

Schanzenbach, D. W. (2006), ‘What have researchers learned from Project STAR?’, *Brookings Papers on Education Policy* **2006**(1), 205–228.

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.

White, H. (1980), ‘A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity’, *Econometrica* **48**(4), 817–838.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Wooldridge, J. M. (2003), ‘Cluster-sample methods in applied econometrics’, *The American Economic Review* **93**(2), 133–138.

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.