Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effect models

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

and

Elizabeth Tipton
Department of Human Development
Teachers College, Columbia University

September 12, 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 thank Dan Knopf for helpful discussions about the linear algebra behind the cluster-robust variance estimator. Coady Wing,...

1 INTRODUCTION

2 STANDARD CLUSTER-ROBUST VARIANCE ESTIMATION

2.1 Econometric framework

We consider a generic fixed effects model in which

$$\mathbf{y}_j = \mathbf{R}_j\boldsymbol{\beta} + \mathbf{S}_j\boldsymbol{\gamma} + \mathbf{T}_j\boldsymbol{\delta} + \boldsymbol{\epsilon}_j, \tag{1}$$

where $\mathbf{R}_j$ is an $n_j \times r$ matrix of covariates, $\mathbf{S}_j$ is an $n_j \times s$ matrix describing fixed effects that vary across clusters, and $\mathbf{T}_j$ is an $n_j \times t$ matrix describing fixed effects that are identified only within clusters. We assume that $\mathrm{E}(\boldsymbol{\epsilon}_j | \mathbf{R}_j, \mathbf{S}_j, \mathbf{T}_j) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}_j | \mathbf{R}_j, \mathbf{S}_j, \mathbf{T}_j) = \boldsymbol{\Sigma}_j$, for $j = 1, ..., m$, where the form of $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown but the errors are independent across clusters. For notational convenience, let $\mathbf{U}_j = [\mathbf{R}_j \ \mathbf{S}_j]$, $\mathbf{X}_j = [\mathbf{U}_j \ \mathbf{T}_j]$, $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\delta}')'$, and $x = r + s + t$. Denote the total number of individual observations by $N = \sum_{j=1}^{m} n_j$. Let $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, and $\mathbf{X}$ denote the matrices obtained by stacking their corresponding components, as in $\mathbf{R} = (\mathbf{R}_1' \ \mathbf{R}_2' \ \cdots \ \mathbf{R}_m')'$.

In this model, inferential interest is confined to $\boldsymbol{\beta}$ and the fixed effects $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are treated as nuisance parameters. The distinction between the covariates $\mathbf{R}_j$ versus the fixed effects $[\mathbf{S}_j \ \mathbf{T}_j]$ thus depends on context and the analyst's inferential goals. The distinction between the two fixed effect matrices $\mathbf{S}_j$ and $\mathbf{T}_j$ is less ambiguous, in that the within-cluster fixed effects satisfy $\mathbf{T}_j\mathbf{T}_k' = \mathbf{0}$ for $j \neq k$. We further assume that $\left(\mathbf{U}'\mathbf{U} - \mathbf{U}_j'\mathbf{U}_j\right)$ is of full rank for $j = 1, ..., m$.

We shall consider weighted least-squares (WLS) estimation of $\boldsymbol{\beta}$. For each cluster j, let $\mathbf{W}_j$ be a symmetric, $n_j \times n_j$ weighting matrix of full rank. 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 and Müller (2010) for clustered data. Assuming that $\mathbf{X}_j$ has rank p for $j = 1, ..., m$, their proposed approach involves estimating

Several approaches computing the WLS estimator are possible. One possibility is to calculate WLS estimates of the full parameter vector $\boldsymbol{\alpha}$ directly. However, this method can be computationally intensive and numerically inaccurate if the fixed effects specification is large (i.e., $s + t$ large). An alternative is to first absorb the fixed effect specification. We shall describe the latter approach because it is more efficient and numerically accurate.

Denote the full block-diagonal weighting matrix as $\mathbf{W} = \text{diag}\left(\mathbf{W}_1, ..., \mathbf{W}_m\right)$. Let $\mathbf{K}$ be the $x \times r$ matrix that selects the covariates of interest, so that $\mathbf{XK} = \mathbf{R}$ and $\mathbf{K}'\boldsymbol{\alpha} = \boldsymbol{\beta}$. For a generic matrix $\mathbf{Z}$ of full column rank, let $\mathbf{M}_\mathbf{Z} = \left(\mathbf{Z}'\mathbf{W}\mathbf{Z}\right)^{-1}$ and $\mathbf{H}_\mathbf{Z} = \mathbf{Z}\mathbf{M}_\mathbf{Z}\mathbf{Z}'\mathbf{W}$.

The absorption technique involves obtaining the residuals from the regression of $\mathbf{y}$ on $\mathbf{T}$ and from the multivariate regressions of $\mathbf{U} = [\mathbf{R}\ \mathbf{S}]$ on $\mathbf{T}$. The $\mathbf{y}$ residuals and $\mathbf{R}$ residuals are then regressed on the $\mathbf{S}$ residuals. Finally, these twice-regressed $\mathbf{y}$ residuals are regressed on the twice-regressed $\mathbf{R}$ residuals to obtain the WLS estimates of $\boldsymbol{\beta}$. Let $\ddot{\mathbf{S}} = \left(\mathbf{I} - \mathbf{H}_\mathbf{T}\right)\mathbf{S}$, $\ddot{\mathbf{R}} = \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right)\left(\mathbf{I} - \mathbf{H}_\mathbf{T}\right)\mathbf{R}$, and $\ddot{\mathbf{y}} = \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right)\left(\mathbf{I} - \mathbf{H}_\mathbf{T}\right)\mathbf{y}$. In what follows, subscripts on $\ddot{\mathbf{R}}$, $\ddot{\mathbf{S}}$, $\ddot{\mathbf{U}}$, and $\ddot{\mathbf{y}}$ refer to the rows of these matrices corresponding to a specific cluster. The WLS estimator of $\boldsymbol{\beta}$ can then be written as

$$\hat{\boldsymbol{\beta}} = \mathbf{M}_{\ddot{\mathbf{R}}} \sum_{j=1}^{m} \ddot{\mathbf{R}}_j' \mathbf{W}_j \ddot{\mathbf{y}}_j. \tag{2}$$

This estimator is algebraically identical to the direct WLS estimator based on the full set of predictors,

$$\hat{\boldsymbol{\beta}} = \mathbf{K}'\mathbf{M}_\mathbf{X} \sum_{j=1}^{m} \mathbf{X}_j' \mathbf{W}_j \mathbf{y}_j,$$

but avoids the need to solve a system of x linear equations.

The variance of the WLS estimator is

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

which depends upon the unknown variance matrices $\boldsymbol{\Sigma}_j$. One approach to estimating this variance is based on a parametric model for the error structure. In this approach, it is assumed that $\text{Var}\left(\mathbf{e}_j \,|\mathbf{X}_j\right) = \boldsymbol{\Phi}_j$, where $\boldsymbol{\Phi}_j$ is a known function of a low-dimensional parameter. For example, an auto-regressive error structure might be posited to describe repeated

$\boldsymbol{\beta}$ separately within each cluster and taking the simple average of these estimates. The resulting average is equivalent to the WLS estimator with weights $\mathbf{W}_j = \mathbf{X}_j \left(\mathbf{X}_j'\mathbf{X}_j\right)^{-2} \mathbf{X}_j$.

measures on an individual over time. If this approach is used, each $\mathbf{\Sigma}_j$ is substituted with an estimate $\hat{\mathbf{\Phi}}_j$, producing the model-based variance estimator

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

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 Standard CRVE

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

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

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 and White (1985). Setting $\mathbf{A}_j = \mathbf{I}_j$, an $n_j \times n_j$ identity matrix, results in the original CRVE estimator. Following Cameron and Miller (2015), we refer to this estimator as $\mathbf{V}^{CR0}$. Setting $\mathbf{A}_j = c\mathbf{I}_j$, where $c = \sqrt{(m/(m-1))(N/(N-p))}$, results in the CRV1 estimator, denoted $\mathbf{V}^{CR1}$. Note that when $N >> p$, $c \approx \sqrt{m/(m-1)}$, and some software uses the latter approximation. 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.

There are several alternative small-sample corrections that are used with CRVE. The BRL approach will be described in the next section. Because it is an extension of the HC2 estimator for regressions with heteroskedastic but uncorrelated errors, we refer to it as CR2. Finally, a further alternative is to use a jack-knife resampling estimator; the CR3 estimator closely approximates the jack-knife approach, by taking $\mathbf{A}_j = \left(\mathbf{I} - \ddot{\mathbf{R}}_j \mathbf{M}_{\ddot{\mathbf{R}}} \ddot{\mathbf{R}}_j' \mathbf{W}_j\right)^{-1}$.

3 BIAS REDUCED LINEARIZATION

The BRL approach chooses adjustment matrices so that the variance estimator is exactly unbiased under a specific working model for the data. It is therefore directly analogous to the HC2 heteroskedasticity-robust estimator, which is exactly unbiased under homoskedasticity. Bell and McCaffrey (2002) developed the BRL estimator for linear regression models with errors having unknown dependence structure within clusters. However, their implementation is not applicable to many fixed effect models, where the adjustment matrices may be undefined. Furthermore, the form of their adjustment matrices varies depending on whether fixed effects are absorbed or estimated directly by WLS, which is undesirable. Our implementation of BRL addresses both of these issues and can be implemented in models with quite general fixed effects specifications. It reduces to Bell and McCaffrey's implementation for models without fixed effects.

Let $\mathbf{\Phi}_j$ be a working model for the covariance of the errors in cluster j, and denote $\mathbf{\Phi} = \text{diag}\left(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_m\right)$. Consider adjustment matrices satisfying the following criterion:

$$\ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{A}_j \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_j' \mathbf{A}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j = \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \ddot{\mathbf{R}}_j, \tag{6}$$

where $\left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_j$ denotes the rows of $\mathbf{I} - \mathbf{H}_\mathbf{X}$ corresponding to cluster j. A variance estimator that uses such adjustment matrices will be exactly unbiased when the working model is correctly specified. [2] When the working model deviates from the true covariance $\mathbf{\Sigma}_j$, the variance estimator remains biased. However, Bell and McCaffrey (2002) showed that the CR2 estimator still greatly reduces the bias compared to the more basic CR0 and CR1 estimators (thus the name "bias reduced linearization"). Extensive simulation results indicate that the remaining bias is typically minimal, even for large deviations from the assumed structure (CITE). Furthermore, as the number of clusters increases, the reliance on the working model diminishes. One way to understand this approach is that it provides necessary scaffolding in the small sample case, which falls away when there is sufficient data.

[2] Note that this criterion differs from the criterion used by Bell and McCaffrey (2002) in that it pre- and post-multiplies both sides by $\mathbf{W}_j \ddot{\mathbf{R}}_j$. As will be seen, this modification permits the use of generalized matrix inverses in calculating the adjustment matrices, thus avoiding rank-deficiency problems that would otherwise leave them undefined.

Criterion (6) does not uniquely define $\mathbf{A}_j$. Following McCaffrey, Bell and Botts (2001), we propose to use a symmetric solution in which

$$\mathbf{A}_j = \mathbf{D}_j' \mathbf{B}_j^{+1/2} \mathbf{D}_j, \tag{7}$$

where $\mathbf{D}_j$ is the upper-right triangular Cholesky factorization of $\hat{\mathbf{\Phi}}_j$,

$$\mathbf{B}_j = \mathbf{D}_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right) \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \mathbf{D}_j', \tag{8}$$

and $\mathbf{B}_j^{+1/2}$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_j$. The Moore-Penrose inverse is well-defined even when $\mathbf{B}_j$ is not of full rank. Theorem 1 in Appendix A shows that the adjustment matrices given by (7) and (8) satisfy criterion (6) and are invariant to whether the model is estimated by direct WLS estimation or after absorbing some or all of the fixed effects.

In many applications, it will make sense to choose weighting matrices that are the inverses of the working covariance model, so that $\mathbf{W}_j = \mathbf{\Phi}_j^{-1}$. In this case, the adjustment matrices can be calculated using $\tilde{\mathbf{B}}_j$ in place of $\mathbf{B}_j$, where

$$\tilde{\mathbf{B}}_j = \mathbf{D}_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right) \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j' \mathbf{D}_j'. \tag{9}$$

See Theorem 2 in Appendix A. In the simple case of ordinary (unweighted) least squares, in which the working variance model posits that the errors are all independent and homoskedastic and $\mathbf{W} = \mathbf{\Phi} = \mathbf{I}$, the adjustment matrices simplify further to

$$\mathbf{A}_j = \left(\mathbf{I}_j - \ddot{\mathbf{U}}_j \left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}}\right)^{-1} \ddot{\mathbf{U}}_j'\right)^{+1/2},$$

where $\ddot{\mathbf{U}} = \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{U}$.

In the remainder of this paper, we will focus on this BRL approach, using the $\mathbf{V}^{CR2}$ estimator throughout.

4 HYPOTHESIS TESTING

4.1 Small-sample corrections for t-tests

4.2 Small-sample corrections for F-tests

5 SIMULATION EVIDENCE

6 EXAMPLES

6.1 Tennessee STAR class-size experiment.

6.2 Heterogeneous treatment impacts

6.3 Robust Hausmann test

7 DISCUSSION

A BRL adjustment matrices

This appendix states and provides proof of two theorems regarding the BRL adjustment matrices.

Theorem 1. *Let* $\mathbf{L} = \left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}} - \ddot{\mathbf{U}}_j'\ddot{\mathbf{U}}_j\right)$ *and assume that* $\mathbf{L}$ *has full rank* $r+s$*, so that its inverse exists. Then the adjustment matrices* $\mathbf{A}_j$ *defined in (7) and (8) satisfy criterion (6) and* $\mathbf{V}^{CR2}$ *is exactly unbiased when the working covariance model* $\mathbf{\Phi}$ *is correctly specified.*

Proof. The Moore-Penrose inverse of $\mathbf{B}_j$ can be computed from its eigen-decomposition. Let $b \leq n_j$ denote the rank of $\mathbf{B}_j$. Let $\mathbf{\Lambda}$ be the $b \times b$ diagonal matrix of the positive eigenvalues of $\mathbf{B}_j$ and $\mathbf{V}$ be the $n_j \times b$ matrix of corresponding eigen-vectors, so that $\mathbf{B}_j = \mathbf{V}\mathbf{V}'$. Then $\mathbf{B}_j^{+} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}'$ and $\mathbf{B}_j^{+1/2} = \mathbf{V}\mathbf{\Lambda}^{-1/2}\mathbf{V}'$.

Now, observe that $\left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}\right) \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) = \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$. Thus,

$$\begin{aligned}\ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{A}_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j' \mathbf{A}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j &= \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{D}_j \mathbf{B}_j^{+1/2} \mathbf{B}_j \mathbf{B}_j^{+1/2} \mathbf{D}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j \\ &= \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{D}_j \mathbf{V}\mathbf{V}' \mathbf{D}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j. \end{aligned} \tag{10}$$

Because $\mathbf{D}_j$, and $\mathbf{\Phi}$ are positive definite and $\mathbf{B}_j$ is symmetric, the eigenvectors $\mathbf{V}$ define an orthogonal basis for the column span of $\left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_j$.

Cite Bannerjee and Roy

We now show that $\ddot{\mathbf{U}}_j$ is in the column space of $\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$. Let $\mathbf{Z}_j$ be an $n_j \times (r+s)$ matrix of zeros. Let $\mathbf{Z}_k = -\ddot{\mathbf{U}}_k \mathbf{L}^{-1} \mathbf{M}_{\ddot{\mathbf{U}}}^{-1}$, for $k \neq j$ and take $\mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_m')'$. Now note that $\left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{Z} = \mathbf{Z}$. It follows that

$$\begin{aligned}\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j \mathbf{Z} &= \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{Z} = \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_j \mathbf{Z} \\ &= \mathbf{Z}_j - \ddot{\mathbf{U}}_j \mathbf{M}_{\ddot{\mathbf{U}}} \sum_{k=1}^{m} \ddot{\mathbf{U}}_k' \mathbf{W}_k \mathbf{Z}_k = \ddot{\mathbf{U}}_j \mathbf{M}_{\ddot{\mathbf{U}}} \left(\sum_{k \neq j} \ddot{\mathbf{U}}_k' \mathbf{W}_k \ddot{\mathbf{U}} \right) \mathbf{L}^{-1} \mathbf{M}_{\ddot{\mathbf{U}}}^{-1} \\ &= \ddot{\mathbf{U}}_j. \end{aligned}$$

Thus, there exists an $N \times (r+s)$ matrix $\mathbf{Z}$ such that $\left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)_j \mathbf{Z} = \ddot{\mathbf{U}}_j$, i.e., $\ddot{\mathbf{U}}_j$ is in the column span of $\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$. Because $\mathbf{D}_j \mathbf{W}_j$ is positive definite and $\ddot{\mathbf{R}}_j$ is a sub-matrix of $\ddot{\mathbf{U}}_j$, $\mathbf{D}_j \mathbf{W}_j \ddot{\mathbf{R}}_j$ is also in the column span of $\left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j$. It follows that

$$\ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{D}_j \mathbf{V}\mathbf{V}' \mathbf{D}_j' \mathbf{W}_j \ddot{\mathbf{R}}_j = \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \ddot{\mathbf{R}}_j. \tag{11}$$

Substituting (11) into (10) demonstrates that $\mathbf{A}_j$ satisfies criterion (6).

Under the working model, the residuals from cluster j have mean $\mathbf{0}$ and variance

$$\text{Var}\left(\ddot{\mathbf{e}}_j\right) = \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j',$$

It follows that

$$\begin{aligned}\text{E}\left(\mathbf{V}^{CR2}\right) &= \mathbf{M}_{\ddot{\mathbf{R}}} \left[\sum_{j=1}^{m} \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{A}_j \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\mathbf{X}}\right)_j' \mathbf{A}_j \mathbf{W}_j \ddot{\mathbf{R}}_j \right] \mathbf{M}_{\ddot{\mathbf{R}}} \\ &= \mathbf{M}_{\ddot{\mathbf{R}}} \left[\sum_{j=1}^{m} \ddot{\mathbf{R}}_j' \mathbf{W}_j \mathbf{\Phi}_j \mathbf{W}_j \ddot{\mathbf{R}}_j \right] \mathbf{M}_{\ddot{\mathbf{R}}} \\ &= \text{Var}\left(\hat{\boldsymbol{\beta}}\right) \end{aligned}$$

□

Theorem 2. *Let* $\tilde{\mathbf{A}}_j = \mathbf{D}_j'\tilde{\mathbf{B}}_j^{+1/2}\mathbf{D}_j$*, where* $\tilde{\mathbf{B}}_j$ *is given in (9). If* $\mathbf{T}_j\mathbf{T}_k' = \mathbf{0}$ *for* $j \neq k$ *and* $\mathbf{W} = \mathbf{\Phi}$*, then* $\mathbf{A}_j = \tilde{\mathbf{A}}_j$*.*

Proof. From the fact that $\ddot{\mathbf{U}}_j'\mathbf{W}_j\mathbf{T}_j = \mathbf{0}$ for $j = 1, ..., m$, it follows that

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

and

$$\mathbf{B}_j^{-1} = \left(\hat{\mathbf{\Phi}}_j^{C\prime}\right)^{-1} \left(\mathbf{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j' - \mathbf{S}_j\mathbf{M}_{\mathbf{S}}\mathbf{S}_j'\right)^{-1} \left(\hat{\mathbf{\Phi}}_j^{C}\right)^{-1}. \tag{12}$$

Let $\mathbf{U}_j = \left(\mathbf{\Phi}_j - \ddot{\mathbf{R}}_j\mathbf{M}_{\ddot{\mathbf{R}}}\ddot{\mathbf{R}}_j'\right)^{-1}$. Using a generalized Woodbury identity (Henderson and 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 (12), we then have that $\mathbf{B}_j^{-1} = \left(\hat{\mathbf{\Phi}}_j^{C\prime}\right)^{-1} \mathbf{U}_j \left(\hat{\mathbf{\Phi}}_j^{C}\right)^{-1}$. Now, $\ddot{\mathbf{B}}_j = \hat{\mathbf{\Phi}}_j^C\mathbf{U}_j^{-1}\hat{\mathbf{\Phi}}_j^{C\prime}$ and so $\ddot{\mathbf{B}}_j^{-1} = \mathbf{B}_j^1$. It follows that $\ddot{\mathbf{A}}_j = \mathbf{A}_j$ for $j = 1, ..., m$. □

B 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 and 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

$$\mathrm{E}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) = \sum_{j=1}^{m}\mathbf{t}_{1j}'\boldsymbol{\Sigma}\mathbf{t}_{2j} \tag{13}$$

$$\mathrm{Var}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2\right) = \sum_{i=1}^{m}\sum_{j=1}^{m}\left(\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{2j}\right)^2 + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{1j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{2j} \tag{14}$$

$$\mathrm{Cov}\left(\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2, \mathbf{c}_3'\mathbf{V}^{CR}\mathbf{c}_4\right) = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{3j} + \mathbf{t}_{1i}'\boldsymbol{\Sigma}\mathbf{t}_{3j}\mathbf{t}_{2i}'\boldsymbol{\Sigma}\mathbf{t}_{4j}. \tag{15}$$

Furthermore, the distribution of $\mathbf{c}_1'\mathbf{V}^{CR2}\mathbf{c}_2$ 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

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

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

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

MacKinnon, J. G. and 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. and Botts, C. H. (2001), Generalizations of biased reduced linearization, *in* 'Proceedings of the Annual Meeting of the American Statistical Association', number 1994.