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

James E. Pustejovsky *
University of Wisconsin - Madison
and
Elizabeth Tipton †
Northwestern University

September 28, 2022

Abstract

In panel data models and other regressions with unobserved effects, fixed effects estimation is often paired with cluster-robust variance estimation (CRVE) in order to account for heteroskedasticity and un-modeled dependence among the errors. Although asymptotically consistent, CRVE can be biased downward when the number of clusters is small, leading to hypothesis tests with rejection rates that are too high. More accurate tests can be constructed using bias-reduced linearization (BRL), which corrects the CRVE based on a working model, in conjunction with a Satterthwaite approximation for t-tests. We propose a generalization of BRL that can be applied in models with arbitrary sets of fixed effects, where the original BRL method is undefined, and describe how to apply the method when the regression is estimated after absorbing the fixed effects. We also propose a small-sample test for multiple-parameter hypotheses, which generalizes the Satterthwaite approximation for t-tests. In simulations covering a wide range of scenarios, we find that the conventional cluster-robust Wald test can severely over-reject while the proposed small-sample test maintains Type I error close to nominal levels. The proposed methods are implemented in an R package called clubSandwich.

Keywords: cluster dependence, fixed effects, robust standard errors, small samples

*Department of Educational Psychology, University of Wisconsin - Madison, 1025 West Johnson Street, Madison, WI 53706. Email: pustejovsky@wisc.edu

†Department of Statistics, Northwestern University. Email: tipton@northwestern.edu

1 INTRODUCTION

In many economic analyses, interest centers on the parameters of a linear regression model estimated by ordinary or weighted least squares from data exhibiting within-group dependence. Such dependence can arise from sampling or random assignment of aggregate units (e.g., counties, districts, villages), each of which contains multiple observations; from repeated measurement of an outcome on a common set of units, as in panel data; or from model misspecification, as in analysis of regression discontinuity designs (e.g., Lee & Card 2008). A common approach to inference in these settings is to use a cluster-robust variance estimator (CRVE, Arellano 1987, Liang & Zeger 1986, White 1984). The advantage of the CRVE is that it produces consistent standard errors and test statistics without imposing strong parametric assumptions about the correlation structure of the errors in the model. Instead, the method relies on the weaker assumption that units can be grouped into clusters that are mutually independent. In the past decade, use of CRVEs has become standard practice for micro-economic researchers, as evidenced by coverage in major textbooks and review articles (e.g., Wooldridge 2010, Angrist & Pischke 2009, Cameron & Miller 2015).

As a leading example, consider a difference-in-differences analysis of state-by-year panel data, where the goal is to understand the effects on employment outcomes of several state-level policy shifts. Each policy effect would be parameterized as a dummy variable in a regression model, which might also include other demographic controls. It is also common to include fixed effects for states and time-points in order to control for unobserved confounding in each dimension. The model could be estimated by least squares with the fixed effects included as dummy variables (or what we will call the LSDV estimator). More commonly, the effects of the policy indicators would be estimated after absorbing the fixed effects, a computational technique that is also known as the fixed effects estimator or "within transformation" (Wooldridge 2010). Standard errors would then be clustered by state to account for residual dependence in the errors from a given state, and these clustered standard errors would be used to test hypotheses about the set of policies. The need to cluster the standard errors by state, even when including state fixed effects, was highlighted by Bertrand et al. (2004), who showed that to do otherwise can lead to inappropriately small standard errors and hypothesis tests with incorrect rejection rates.

The consistency property of CRVEs is asymptotic in the number of independent clusters (Wooldridge 2003). Recent methodological work has demonstrated that CRVEs can be biased downward and associated hypothesis tests can have Type-I error rates considerably in excess of nominal levels when based on a small or moderate number of clusters (e.g., MacKinnon & Webb 2016). Cameron & Miller (2015) provided an extensive review of this literature, including a discussion of current practice, possible solutions, and open problems. In particular, they demonstrated that small-sample corrections for t-tests implemented in common software packages such as Stata and SAS do not provide adequate control of Type-I error.

Bell & McCaffrey (2002) proposed a method that improves the small-sample properties of CRVEs (see also McCaffrey et al. 2001). The method, called bias-reduced linearization (BRL), entails adjusting the 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 by 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 small-sample corrections to single-parameter hypothesis tests using the BRL variance estimator, based on Satterthwaite (Bell & McCaffrey 2002) or saddlepoint approximations (McCaffrey & Bell 2006). In an analysis of a longitudinal cluster-randomized trial with 35 clusters, Angrist & Lavy (2009) observed that the BRL correction makes a difference for inferences.

Despite a growing body of evidence that BRL performs well (e.g., Imbens & Kolesar 2016), several problems with the method hinder its wider application. First, Angrist & Pischke (2009) noted that the BRL correction is undefined in some commonly used models, such as state-by-year panels that include fixed effects for states and for years (see also Young 2016). Second, in models with fixed effects, the magnitude of the BRL adjustment depends on whether it is computed based on the full design matrix (i.e., the LSDV estimator) or after absorbing the fixed effects. Third, extant methods for hypothesis testing based on BRL are limited to single-parameter constraints (Bell & McCaffrey 2002, McCaffrey & Bell 2006) and small-sample methods for multiple-parameter hypothesis tests remain lacking.

This paper addresses each of these concerns in turn, with the aim of extending the

BRL method so that is suitable for general application. First, we describe a simple modification of the BRL adjustment that remains well-defined in models with arbitrary sets of fixed effects, where existing BRL adjustments break down. Second, we demonstrate how to calculate the BRL adjustments based on the fixed effects estimator and identify conditions under which first-stage absorption of the fixed effects can be ignored. Finally, we propose a procedure for testing multiple-parameter hypotheses by approximating the sampling distribution of the Wald statistic using Hotelling's T^2 distribution with estimated degrees of freedom. The method is a generalization of the Satterthwaite correction proposed by Bell & McCaffrey (2002) for single parameter constraints. The proposed methods are implemented in the R package `clubSandwich`, which is available on the Comprehensive R Archive Network.

Our work is related to a stream of recent literature that has examined methods for cluster-robust inference with a small number of clusters. Conley & Taber (2011) proposed methods for hypothesis testing in a difference-in-differences setting where the number of treated units is small and fixed, while the number of untreated units increases asymptotically. Ibragimov and Müller (2010; 2016) proposed cluster-robust t-tests that maintain the nominal Type-I error rate by re-weighting within-cluster estimators of the target parameter. Young (2016) proposed a Satterthwaite correction for t-tests based on a different type of bias correction to the CRVE, where the bias correction term is derived under a working model. Cameron et al. (2008) investigated a range of bootstrapping procedures that provide improved Type-I error control in small samples, finding that a cluster wild-bootstrap technique was particularly accurate in small samples. Nearly all of this work has focused on single-parameter hypothesis tests only. For multiple-parameter constraints, Cameron & Miller (2015) suggested an ad hoc degrees of freedom adjustment and noted, as an alternative, that bootstrapping techniques can in principle be applied to multiple-parameter tests. However, little methodological work has examined the accuracy of multiple-parameter tests.

The paper is organized as follows. The remainder of this section introduces our econometric framework and reviews standard CRVE methods, as implemented in most software applications. Section 2 reviews the original BRL correction and describes modifications that make it possible to implement BRL in a broad class of models with fixed effects. Sec-

tion ?? discusses hypothesis tests based on the BRL-adjusted CRVE. Section ?? reports a simulation study examining the null rejection rates of multiple-parameter hypothesis tests, where we find that the small-sample test offers drastic improvements over commonly implemented alternatives. Section ?? illustrates the use of the proposed hypothesis tests in two applications. Section ?? concludes and discusses avenues for future work.

1.1 Econometric framework

We consider a linear regression model of the form,

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

where there are a total of m clusters; cluster i contains n_i units; $\mathbf{y}_i$ is a vector of the n_i values of the outcome for units in cluster i; $\mathbf{R}_i$ is an $n_i \times r$ matrix containing predictors of primary interest (e.g., policy variables) and any additional controls; $\mathbf{S}_i$ is an $n_i \times s$ matrix describing fixed effects that are identified across multiple clusters; and $\mathbf{T}_i$ is an $n_i \times t$ matrix describing cluster-specific fixed effects, which must satisfy $\mathbf{T}_h\mathbf{T}_i' = \mathbf{0}$ for $h \neq i$. Note that the distinction between the covariates $\mathbf{R}_i$ versus the fixed effects $\mathbf{S}_i$ is arbitrary and depends on the analyst's inferential goals. In a fixed effects model for state-by-year panel data, $\mathbf{R}_i$ would include variables describing policy changes, as well as additional demographic controls; $\mathbf{S}_i$ would include year fixed effects; and $\mathbf{T}_i$ would indicate state fixed effects (and perhaps also state-specific time trends). Interest would center on testing hypotheses regarding the coefficients in $\boldsymbol{\beta}$ that correspond to the policy indicators, while $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ would be treated as incidental.

We shall assume that $\mathrm{E}\left(\boldsymbol{\epsilon}_i \,|\mathbf{R}_i, \mathbf{S}_i, \mathbf{T}_i\right) = \mathbf{0}$ and $\mathrm{Var}\left(\boldsymbol{\epsilon}_i \,|\mathbf{R}_i, \mathbf{S}_i, \mathbf{T}_i\right) = \boldsymbol{\Sigma}_i$, for $i = 1, ..., m$, where the form of $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m$ may be unknown but the errors are independent across clusters. Let $\mathbf{U}_i = [\mathbf{R}_i\ \mathbf{S}_i]$ denote the set of predictors that are identified across multiple clusters, $\mathbf{X}_i = [\mathbf{R}_i\ \mathbf{S}_i\ \mathbf{T}_i]$ denote the full set of predictors, $\boldsymbol{\alpha} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\mu}')'$, and $p = r + s + t$. Let $N = \sum_{i=1}^{m} n_i$ denote the total number of observations. Let $\mathbf{y}$, $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, $\mathbf{X}$, and $\boldsymbol{\epsilon}$ denote the matrices obtained by stacking their corresponding components, as in $\mathbf{R} = (\mathbf{R}_1'\ \mathbf{R}_2'\ \cdots\ \mathbf{R}_m')'$.

We assume that $\boldsymbol{\beta}$ is estimated by weighted least squares (WLS) using symmetric, full rank weighting matrices $\mathbf{W}_1, ..., \mathbf{W}_m$. Clearly, the WLS estimator includes ordinary least

squares (OLS) as a special case. More generally, the WLS estimator encompasses feasible generalized least squares, where it is assumed that $\text{Var}\left(\mathbf{e}_i \,|\mathbf{X}_i\right) = \boldsymbol{\Phi}_i$, a known function of a low-dimensional parameter. For example, an auto-regressive error structure might be posited to describe repeated measures on an individual over time. The weighting matrices are then taken to be $\mathbf{W}_i = \hat{\boldsymbol{\Phi}}_i^{-1}$, where the $\hat{\boldsymbol{\Phi}}_i$ are constructed from estimates of the variance parameter. Finally, for analysis of data from complex survey designs, WLS may be used with sampling weights in order to account for unequal selection probabilities.

1.2 Absorption

The goal of most analyses is to estimate and test hypotheses regarding the parameters in $\boldsymbol{\beta}$, while the fixed effects $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$ are not of inferential interest. Furthermore, LSDV estimation becomes computationally intensive and numerically inaccurate if the model includes a large number of fixed effects (i.e., $s+t$ large). A commonly implemented alternative to LSDV is to first absorb the fixed effects, which leaves only the r parameters in $\boldsymbol{\beta}$ to be estimated. Because Section 2 examines the implications of absorption for application of the BRL adjustment, we now formalize this procedure. 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 $p \times r$ 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 $\mathbf{M_Z} = \left(\mathbf{Z}'\mathbf{WZ}\right)^{-1}$ and $\mathbf{H_Z} = \mathbf{ZM_Z Z}'\mathbf{W}$.

The absorption technique involves obtaining the residuals from the regression of $\mathbf{y}$ on $\mathbf{T}$ and from the multivariate regression of $[\mathbf{R}\ \mathbf{S}]$ 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 $\ddot{\mathbf{S}} = \left(\mathbf{I} - \mathbf{H_T}\right)\mathbf{S}$, $\ddot{\mathbf{R}} = \left(\mathbf{I} - \mathbf{H_{\ddot{S}}}\right)\left(\mathbf{I} - \mathbf{H_T}\right)\mathbf{R}$, and $\ddot{\mathbf{y}} = \left(\mathbf{I} - \mathbf{H_{\ddot{S}}}\right)\left(\mathbf{I} - \mathbf{H_T}\right)\mathbf{y}$. 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

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

This estimator is algebraically identical to the LSDV estimator, $\hat{\boldsymbol{\beta}} = \mathbf{K}'\mathbf{M_X X}'\mathbf{Wy}$, but avoids the need to solve a system of p linear equations. For further details on sequential

absorption, see Davis (2002). In the remainder, we assume that fixed effects are absorbed before estimation of $\boldsymbol{\beta}$.

1.3 Standard CRVE

The WLS estimator $\hat{\boldsymbol{\beta}}$, has true variance

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

which depends upon the unknown variance matrices $\boldsymbol{\Sigma}_i$.

The CRVE involves estimating $\text{Var}\left(\hat{\boldsymbol{\beta}}\right)$ empirically, without imposing structural assumptions on $\boldsymbol{\Sigma}_i$. There are several versions of this approach, all of which can be written as

$$\mathbf{V}^{CR} = \mathbf{M}_{\ddot{\mathbf{R}}}\left(\sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{A}_i \mathbf{e}_i \mathbf{e}_i' \mathbf{A}_i' \mathbf{W}_i \ddot{\mathbf{R}}_i\right) \mathbf{M}_{\ddot{\mathbf{R}}}, \tag{4}$$

where $\mathbf{e}_i = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ is the vector of residuals from cluster i and $\mathbf{A}_i$ is some n_i by n_i adjustment matrix.

The form of the adjustment matrices parallels those of the heteroskedasticity-consistent variance estimators proposed by MacKinnon & White (1985). The original CRVE, described by Liang & Zeger (1986), uses $\mathbf{A}_i = \mathbf{I}_i$, an $n_i \times n_i$ identity matrix. Following Cameron & Miller (2015), we refer to this estimator as CR0. This estimator is biased towards zero because the cross-product of the residuals $\mathbf{e}_i\mathbf{e}_i'$ tends to under-estimate the true variance $\boldsymbol{\Sigma}_i$ in cluster i. A rough bias adjustment is to take $\mathbf{A}_i = c\mathbf{I}_i$, where $c = \sqrt{(m/(m-1))}$; we denote this adjusted estimator as CR1. Some functions in Stata use a slightly different correction factor $c_S = \sqrt{(mN)/[(m-1)(N-p)]}$; we will refer to the adjusted estimator using c_S as CR1S. When $N >> p$, $c_S \approx \sqrt{m/(m-1)}$ and so CR1 and CR1S will be very similar. However, CR1 and CR1S can differ quite substantially for models with a large number of fixed effects and small within-cluster sample size; recent guidance emphasizes that CR1S is not appropriate for this scenario Cameron & Miller (2015). The CR1 and CR1S estimators are commonly used in empirical applications.

Use of these adjustments still tends to under-estimate the true variance of $\hat{\boldsymbol{\beta}}$ because the degree of bias depends not only on the number of clusters m, but also on skewness of the

covariates and unbalance across clusters (Carter et al. 2013, MacKinnon 2013, Cameron & Miller 2015, Young 2016). A more principled approach to bias correction would take into account the features of the covariates in $\mathbf{X}$. One such estimator uses adjustment matrices given by $\mathbf{A}_i = \left(\mathbf{I} - \ddot{\mathbf{R}}_i \mathbf{M}_{\ddot{\mathbf{R}}} \ddot{\mathbf{R}}_i' \mathbf{W}_i\right)^{-1}$. This estimator, denoted CR3, closely approximates the jackknife re-sampling estimator (Bell & McCaffrey 2002, Mancl & DeRouen 2001). However, CR3 tends to over-correct the bias of CR0, while the CR1 estimator tends to under-correct. The next section describes in detail the BRL approach, which makes adjustments that are intermediate in magnitude between CR1 and CR3.

2 BIAS REDUCED LINEARIZATION

The BRL correction is premised on a "working" model for the structure of the errors, which must be specified by the analyst. Under a given working model, adjustment matrices $\mathbf{A}_i$ are defined so that the variance estimator is exactly unbiased. We refer to this correction as CR2 because it extends the HC2 variance estimator for regressions with uncorrelated errors, which is exactly unbiased when the errors are homoskedastic (MacKinnon & White 1985). The idea of specifying a model may seem antithetical to the purpose of using CRVE, yet extensive simulation studies have demonstrated that the method performs well in small samples even when the working model is incorrect (Tipton 2015, Bell & McCaffrey 2002, Cameron & Miller 2015, Imbens & Kolesar 2016). Although the CR2 estimator might not be exactly unbiased when the working model is misspecified, its bias still tends to be greatly reduced compared to CR1 or CR0 (thus the name "bias reduced linearization"). Furthermore, as the number of clusters increases, reliance on the working model diminishes.

Let $\mathbf{\Phi} = \text{diag}\left(\mathbf{\Phi}_1, ..., \mathbf{\Phi}_m\right)$ denote a working model for the covariance structure (up to a scalar constant). For example, we might assume that the errors are uncorrelated and homoskedastic, with $\mathbf{\Phi}_i = \mathbf{I}_i$ for $i = 1, ..., m$. Alternatively, Imbens & Kolesar (2016) suggested using a random effects (i.e., compound symmetric) structure, in which $\mathbf{\Phi}_i$ has unit diagonal entries and off-diagonal entries of ρ, with ρ estimated using the OLS residuals.

In the original formulation of Bell & McCaffrey (2002), the BRL adjustment matrices

are chosen to satisfy the criterion

$$\mathbf{A}_i \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i' \mathbf{A}_i' = \mathbf{\Phi}_i \tag{5}$$

for a given working model, where $\left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i$ denotes the rows of $\mathbf{I} - \mathbf{H}_\mathbf{X}$ corresponding to cluster i. If the working model and weight matrices are both taken to be identity matrices, then the adjustment matrices simplify to $\mathbf{A}_i = \left(\mathbf{I}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i'\right)^{-1/2}$, where $\mathbf{Z}^{-1/2}$ denotes the symmetric square-root of the matrix $\mathbf{Z}$.

2.1 A more general BRL criterion

The original formulation of $\mathbf{A}_i$ is problematic because, for some fixed effects models that are common in economic applications, Equation 5 has no solution. Angrist & Pischke (2009) note that this problem occurs in balanced state-by-year panel models that include fixed effects for states and for years, where $\mathbf{I}_i - \ddot{\mathbf{U}}_i \mathbf{M}_{\ddot{\mathbf{U}}} \ddot{\mathbf{U}}_i'$ is not of full rank. Young (2016) reported that this problem occurred frequently when applying BRL to a large corpus of fitted regression models drawn from published studies.

This issue can be solved by using an alternative criterion to define the adjustment matrices, for which a solution always exists. Instead of (5), we propose to use adjustment matrices $\mathbf{A}_i$ that satisfy:

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

A variance estimator that uses such adjustment matrices will be exactly unbiased when the working model is correctly specified.

A symmetric solution to Equation (6) is given by

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

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

$$\mathbf{B}_i = \mathbf{D}_i \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_\mathbf{X}\right)_i' \mathbf{D}_i', \tag{8}$$

and $\mathbf{B}_i^{+1/2}$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_i$. The Moore-Penrose inverse of $\mathbf{B}_i$ is well-defined and unique (Banerjee & Roy 2014, Thm. 9.18). In

contrast, the original BRL adjustment matrices involve the symmetric square root of the regular inverse of $\mathbf{B}_i$, which does not exist when $\mathbf{B}_i$ is rank-deficient. If $\mathbf{B}_i$ is of full rank, then our adjustment matrices reduce to the original formulation described by Bell & McCaffrey (2002).

The adjustment matrices given by (7) and (8) satisfy criterion (6), as stated in the following theorem.

Theorem 1. *Let* $\mathbf{L}_i = \left(\ddot{\mathbf{U}}'\mathbf{W}\ddot{\mathbf{U}} - \ddot{\mathbf{U}}_i'\mathbf{W}_i\ddot{\mathbf{U}}_i\right)$*, where* $\ddot{\mathbf{U}} = (\mathbf{I} - \mathbf{H_T})\,\mathbf{U}$*, and assume that* $\mathbf{L}_1, ..., \mathbf{L}_m$ *have full rank* $r + s$*. Further assume that* $Var(\boldsymbol{\epsilon}_i \,|\mathbf{X}_i) = \sigma^2\boldsymbol{\Phi}_i$*, for* $i = 1, ..., m$*. Then the adjustment matrix* $\mathbf{A}_i$ *defined in (7) and (8) satisfies criterion (6) and the CR2 variance estimator is exactly unbiased.*

Proof is given in the supplementary materials. The main implication of Theorem 1 is that, under our more general definition, the CR2 variance estimator remains well-defined even in models with large sets of fixed effects.

2.2 Absorption and LSDV Equivalence

In fixed effects regression models, a problem with the original definition of BRL is that it can result in a different estimator depending upon which design matrix is used. If $\boldsymbol{\beta}$ is estimated using LSDV, then it is natural to calculate the CR2 adjustment matrices based on the full covariate design matrix, $\mathbf{X}$. However, if $\boldsymbol{\beta}$ is estimated after absorbing the fixed effects, the analyst might choose to calculate the CR2 correction based on the absorbed covariate matrix $\ddot{\mathbf{R}}$—that is, by substituting $\mathbf{H}_{\ddot{\mathbf{R}}}$ for $\mathbf{H}_{\mathbf{X}}$ in (8)—in order to avoid calculating the full projection matrix $\mathbf{H}_{\mathbf{X}}$. This approach can lead to different adjustment matrices because it is based on a subtly different working model. Essentially, calculating CR2 based on $\mathbf{H}_{\ddot{\mathbf{R}}}$ amounts to assuming that the working model $\boldsymbol{\Phi}$ applies not to the model errors $\boldsymbol{\epsilon}$, but rather to the errors from the final-stage regression of $\ddot{\mathbf{y}}$ on $\ddot{\mathbf{R}}$. Because the CR2 adjustment is relatively insensitive to the working model, the difference between accounting for or ignoring absorption will in many instances be small. We investigate the differences between the approaches as part of the simulation study in Section **??**.

When based on the full regression model, a drawback of using the CR2 adjustment matrices is that it entails calculating the projection matrix $\mathbf{H}_{\mathbf{X}}$ for the full set of p covariates

(i.e., including fixed effect indicators). Given that the entire advantage of using absorption to calculate $\hat{\boldsymbol{\beta}}$ is to avoid computations involving large, sparse matrices, it is of interest to find methods for more efficiently calculating the CR2 adjustment matrices. Some computational efficiency can be gained by using the fact that the residual projection matrix $\mathbf{I} - \mathbf{H}_{\mathbf{X}}$ can be factored into components as $(\mathbf{I} - \mathbf{H}_{\mathbf{X}})_i = (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}})_i (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}}) (\mathbf{I} - \mathbf{H}_{\mathbf{T}})$.

In certain circumstances, further computational efficiency can be achieved by computing the adjustment matrices after absorbing the within-cluster fixed effects $\mathbf{T}$ (but not the between-cluster fixed effects $\mathbf{S}$). Specifically, if the weights used for WLS estimation are the inverses of the working covariance model, so that $\mathbf{W}_i = \boldsymbol{\Phi}_i^{-1}$ for $i = 1, ..., m$, then the adjustment matrices can be calculated without accounting for the within-cluster fixed effects. This result is formalized in the following theorem.

Theorem 2. *Let* $\tilde{\mathbf{A}}_i = \mathbf{D}_i' \tilde{\mathbf{B}}_i^{+1/2} \mathbf{D}_i$, *where*

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

If $\mathbf{W} = \boldsymbol{\Phi}^{-1}$ *and* $\mathbf{T}_i \mathbf{T}_k' = \mathbf{0}$ *for* $i \neq k$, *then* $\mathbf{A}_i = \tilde{\mathbf{A}}_i$.

Proof is given in the supplementary materials. The main implication of Theorem 2 is that the more computationally tractable formula $\tilde{\mathbf{B}}_i$ can be used in the common case that the weighting matrices are the inverse of the working covariance model. Following the working model suggested by Bell & McCaffrey (2002), in which $\boldsymbol{\Phi} = \mathbf{I}$, the theorem shows that the adjustment method is invariant to the choice of estimator so long as the model is estimated by OLS (i.e., with $\mathbf{W} = \mathbf{I}$). %In this case, the CR2 adjustment matrices then simplify further to $\mathbf{A}_i = \left(\mathbf{I}_i - \ddot{\mathbf{U}}_i \left(\ddot{\mathbf{U}}' \ddot{\mathbf{U}} \right)^{-1} \ddot{\mathbf{U}}_i' \right)^{+1/2}$. In contrast, if the working model proposed by Imbens & Kolesar (2016) is instead used (while still using OLS), then the the CR2 adjustments might differ depending on whether LSDV or the fixed effects estimator is used.

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. (1987), 'Computing robust standard errors for within-groups estimators', *Oxford Bulletin of Economics and Statistics* **49**(4), 431–434.

Banerjee, S. & Roy, A. (2014), *Linear Algebra and Matrix Analysis for Statistics*, Taylor & Francis, Boca Raton, FL.

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', *Journal of Human Resources* **50**(2), 317–372.

Carter, A. V., Schnepel, K. T. & Steigerwald, D. G. (2013), Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity.

Conley, T. G. & Taber, C. R. (2011), 'Inference with "Difference in Differences" with a Small Number of Policy Changes', *Review of Economics and Statistics* **93**(1), 113–125.

Davis, P. (2002), 'Estimating multi-way error components models with unbalanced data structures', *Journal of Econometrics* **106**, 67–95.

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

Ibragimov, R. & Müller, U. K. (2016), 'Inference with few heterogeneous clusters', *Review of Economics and Statistics* **98**(1), 83–96.

Imbens, G. W. & Kolesar, M. (2016), 'Robust Standard Errors in Small Samples: Some Practical Advice', *Review of Economics and Statistics* **forthcoming**.

Lee, D. S. & Card, D. (2008), 'Regression discontinuity inference with specification error', *Journal of Econometrics* **142**(2), 655–674.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**(1), 13–22.

MacKinnon, J. G. (2013), Thirty years of heteroskedasticity-robust inference, *in* X. Chen & N. R. Swanson, eds, 'Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis', Springer New York, New York, NY.

MacKinnon, J. G. & Webb, M. D. (2016), 'Wild bootstrap inference for wildly different cluster sizes', *Journal of Applied Econometrics* **forthcoming**.

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

Mancl, L. A. & DeRouen, T. A. (2001), 'A covariance estimator for GEE with improved small-sample properties', *Biometrics* **57**(1), 126–134.

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.

Tipton, E. (2015), 'Small sample adjustments for robust variance estimation with meta-regression.', *Psychological Methods* **20**(3), 375–393.

White, H. (1984), *Asymptotic theory for econometricians*, Academic Press, Inc., Orlando, FL.

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

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2nd edn, MIT Press, Cambridge, MA.

Young, A. (2016), Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections.
URL: *http://personal.lse.ac.uk/YoungA/Improved.pdf*