Department of Economics
Working Paper Series

Bootstrap-Based Improvements for Inference with Clustered Errors

A. Colin Cameron
University of California, Davis

Douglas Miller
University of California, Davis

Jonah B. Gelbach
Department of Economics, University of Maryland and College of Law, Florida State University

July 14, 2006

Paper # 06-21

Microeconometrics researchers have increasingly realized the essential need to account for any within-group dependence in estimating standard errors of regression parameter estimates. The typical preferred solution is to calculate cluster-robust or sandwich standard errors that permit quite general heteroskedasticity and within-cluster error correlation, but presume that the number of clusters is large. In applications with few (5-30) clusters, standard asymptotic tests can over-reject considerably. We investigate more accurate inference using cluster bootstrap-t procedures that provide asymptotic refinement. These procedures are evaluated using Monte Carlos, including the much-cited differences-in-differences example of Bertrand, Mullainathan and Duflo (2004). In situations where standard methods lead to rejection rates in excess of ten percent (or more) for tests of nominal size 0.05, our methods can reduce this to five percent. In principle a pairs cluster bootstrap should work well, but in practice a Wild cluster bootstrap performs better.

UCDAVIS

Department of Economics
One Shields Avenue
Davis, CA 95616
(530)752-0741

http://www.econ.ucdavis.edu/working_search.cfm

Bootstrap-Based Improvements for Inference with Clustered Errors

A.Colin Cameron*, Jonah B. Gelbach† and Douglas L. Miller‡

June 14, 2006

Abstract

Microeconometrics researchers have increasingly realized the essential need to account for any within-group dependence in estimating standard errors of regression parameter estimates. The typical preferred solution is to calculate cluster-robust or sandwich standard errors that permit quite general heteroskedasticity and within-cluster error correlation, but presume that the number of clusters is large. In applications with few (5-30) clusters, standard asymptotic tests can overreject considerably. We investigate more accurate inference using cluster bootstrap-t procedures that provide asymptotic refinement. These procedures are evaluated using Monte Carlos, including the much-cited differences-in-differences example of Bertrand, Mullainathan and Duflo (2004). In situations where standard methods lead to rejection rates in excess of ten percent (or more) for tests of nominal size 0.05, our methods can reduce this to five percent. In principle a pairs cluster bootstrap should work well, but in practice a Wild cluster bootstrap performs better.

Keywords: clustered errors; random effects; cluster robust; sandwich; bootstrap; bootstrap-t; clustered bootstrap; pairs bootstrap; wild bootstrap.

JEL Classification: C15, C12, C21.

*Department of Economics, University of California - Davis.

†Department of Economics, University of Maryland and College of Law, Florida State University.

‡Department of Economics, University of California - Davis.

1 Introduction

Microeconometrics researchers have increasingly realized the essential need to account for any within-group dependence in estimating standard errors of regression parameter estimates. A leading example involves using ordinary least squares (OLS) and U.S. household survey data to estimate the effects of changes in state policies on wages or labor supply. The policy regressor is perfectly correlated (invariant) within state and, even after control for many additional regressors, the error likely will be correlated within states given that states are thought of as distinct labor markets.

In such settings the default OLS standard errors that ignore such clustering can greatly underestimate the true OLS standard errors, as emphasized by Moulton (1986, 1990).

A common correction is to compute cluster-robust standard errors that generalize the White (1980) heteroskedastic-consistent estimate of OLS standard errors to the clustered setting. This permits both error heteroskedasticity and quite flexible error correlation within cluster, unlike a much more restrictive random effects or error components model. In econometrics this adjustment was proposed by White (1984) and Arellano (1987), and it is implemented in STATA, for example, using the cluster option. In the statistics literature these are called sandwich standard errors, proposed by Liang and Zeger (1986) for generalized estimating equations, and they are implemented in SAS, for example, within the GENMOD procedure. A recent brief survey is given in Wooldridge (2003).

Not all empirical studies use appropriate corrections for clustering. In particular, for fixed effects panel models the errors are usually correlated even after control for fixed effects, yet many studies either provide no control for serial correlation or erroneously cluster at too fine a level. Kezdi (2004) demonstrated the usefulness of cluster robust standard errors in this setting and contrasted these with other standard errors based on stronger distributional assumptions. Bertrand, Duflo, and Mullainathan (2004), henceforth BDM (2004), focused on implications for difference-in-difference (DID) studies using data on individuals across states and years. Then the regressor of interest is an indicator variable that is highly correlated within cluster (state) so there is great need to correct standard errors for clustering. The clustering should be on state, rather than on state-year.

A practical limitation of inference with cluster-robust standard errors is that the asymptotic justification assumes that the number of clusters goes to infinity. Yet in some applications there may be few clusters. For example, this happens if clustering is on region and there are few regions. With a small number of clusters the cluster-robust standard errors are downwards biased. Bias corrections have been proposed in the statistics literature; see Kauermann and Carroll (2001), Mancl and DeRouen (2001), and Bell and McCaffrey (2002). But even after appropriate bias correction, with few clusters the usual Wald statistics for hypothesis testing with asymptotic standard normal or chi-square critical values over-reject. BDM (2004) demonstrate through a Monte Carlo experiment that the Wald test based on (unadjusted) cluster-robust standard errors over-rejects if standard normal critical values are used. Donald and Lang (2004) also demonstrate this and propose, for DID studies with policy invariant within state, an alternative two-step GLS estimator that leads to T-distributed Wald tests in some special circumstances. Angrist and Lavy (2002) in an applied study find that bias adjustment of cluster-robust standard errors can make quite a difference.

In this paper we investigate whether bootstrapping to obtain asymptotic refinement leads to improved inference for OLS estimation with cluster-robust standard errors when there are few clusters. We focus on cluster bootstrap-t procedures that are generalizations of those proposed for regression with heteroskedastic errors in the nonclustered case. Previous studies have also investigated the performance of some cluster bootstraps, an early simulation study being that by Sherman and le Cressie (1997). But there have been relatively few such studies, and each study focuses on one particular bootstrap method.

Several features of our bootstraps are worth emphasizing. First, the bootstraps involve resampling entire clusters. Second, our goal is to use variants of the bootstrap that provide asymptotic refinement, whereas most econometric applications use the bootstrap only to obtain consistent estimates of standard errors. Third, we consider several different cluster resampling schemes: pairs bootstrap, residuals bootstrap and wild bootstrap. Fourth, we consider examples with as few as five clusters, as it is not unusual in a clustered setting to have so few clusters yet still have parameter estimates sufficiently precise that coefficients may be statistically significant

at conventional significance levels. Fifth, we evaluate our bootstrap procedures in a number of settings including examples of others that were used to demonstrate the deficiencies of standard cluster-robust methods.

The paper is organized as follows. Section 2 provides a summary of standard asymptotic methods of inference for OLS with clustered data, and presents small-sample corrections to cluster-robust standard errors that have been recently proposed in the statistics literature. Section 3 presents various possible bootstraps for clustered data, with additional details relegated to an Appendix. Sections 4 to 6 present, respectively, a Monte Carlo experiment using generated data, a Monte Carlo experiment using data from BDM (2004), and an application using data from Gruber and Poterba (1994).

The primary contribution of this paper is to offer methods for more accurate cluster-robust inference. These methods are fairly simple to implement and matter substantively in both our Monte Carlo experiments and our replications.

A second important contribution of this paper is to offer a careful and precise description of the various bootstraps a researcher might perform, and the similarities and differences between our proposed methods and several commonly-applied methods. Our primary motivation for presenting this description is to be precise about our methods. However, it also offers empiricists a clearer understanding of the menu of bootstrap choices and their consequences.

2 Cluster-Robust Inference

Before considering the bootstrap we present results on inference with clustered errors.

2.1 OLS with Clustered Errors

The model we consider is one with G clusters (subscripted by g), and with N_g observations (subscripted by i) within each cluster. Errors are independent across clusters but correlated within clusters. The model can be written at various levels of aggregation as

$$\begin{array}{rll} y_{ig} & = \mathbf{x}_{ig}'\boldsymbol{\beta} + u_{ig}, & i = 1, ..., N_g,\ g = 1, ..., G, \\ \mathbf{y}_g & = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{u}_g, & g = 1, ..., G, \\ \mathbf{y} & = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, & \end{array} \tag{1}$$

where $\boldsymbol{\beta}$ is $k \times 1$, $\mathbf{x}_{ig}$ is $k \times 1$, $\mathbf{X}_g$ is $N_g \times k$, $\mathbf{X}$ is $N \times k$, $N = \sum_g N_g$, y_{ig} and u_{ig} are scalar, $\mathbf{y}_g$ and $\mathbf{u}_g$ are $N_g \times 1$ vectors and $\mathbf{y}$ and $\mathbf{u}$ are $N \times 1$ vectors.

Interest lies in inference for the OLS estimator

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}\mathbf{x}_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{x}_{ig}y_{ig}\right) \\ &= \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned} \tag{2}$$

Under the assumptions that data are independent over g but errors are correlated within cluster, with $\mathrm{E}[\mathbf{u}_g] = \mathbf{0}$, $\mathrm{E}[\mathbf{u}_g\mathbf{u}_g'] = \boldsymbol{\Sigma}_g$, and $\mathrm{E}[\mathbf{u}_g\mathbf{u}_h'] = \mathbf{0}$ for cluster $h \neq g$, we have $\widehat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}[\boldsymbol{\beta}, \mathrm{V}[\widehat{\boldsymbol{\beta}}]]$ where

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g\boldsymbol{\Sigma}_g\mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{3}$$

This differs from and is usually larger than the specialization

$$\mathrm{V}[\widehat{\boldsymbol{\beta}}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1} \tag{4}$$

that is based on the assumption of iid errors and leads to the **default OLS variance estimate** when σ_u^2 is estimated by $s^2 = \widehat{\mathbf{u}}'\widehat{\mathbf{u}}/(N-k)$. In the case where the primary source of clustering is due to group-level common shocks, a useful approximation is that for the j^{th} regressor the default OLS variance estimate should be inflated by the **variance inflation multiplier**

$$\tau_j \simeq 1 + \rho_{x_j}\rho_u(\bar{N}_g - 1), \tag{5}$$

where ρ_{x_j} is the within cluster correlation of x_j, ρ_u is the within cluster error correlation, and $\bar{N}_g$ is the average cluster size. This result is exact if cluster sizes are identical, all regressors are invariant within cluster (so $\rho_{x_j} = 1$) and the error follows a random effects model, defined below. In practice it provides a useful guide in a range of settings.[1] As one example, if there are 81 observations per cluster, the regressor is invariant within

[1] Kloek (1981) provides the special case. Scott and Holt (1982) and Greenwald (1983) give more general results for unbalanced clustering and within-cluster regressor and error correlation that varies across clusters. Similar variance inflation arises in survey design, for example two-stage sampling, and in that literature the multiplier is called the design effect.

cluster, and $\rho_u = 0.10$, then default OLS standard errors should be inflated by $\sqrt{1 + 0.1 \times 80} = \sqrt{9} = 3$.

Clearly it can be important to control for clustering, a point emphasized by Moulton (1986, 1990) and more recently by BDM (2004). The underestimation bias is increasing in (1) cluster size; (2) within-cluster correlation of the regressor; and (3) within-cluster correlation of the error. The challenge for inference is that $\mathbf{\Sigma}_g$ is unknown.

2.2 Moulton-type Standard Errors

One approach is to model $\mathbf{\Sigma}_g$ to depend on unknown parameters, say $\mathbf{\Sigma}_g = \mathbf{\Sigma}_g(\boldsymbol{\gamma})$, and then use estimate $\widehat{\mathbf{\Sigma}}_g = \mathbf{\Sigma}_g(\widehat{\boldsymbol{\gamma}})$. The **random effects (RE) model** specifies

$$u_{ig} = \alpha_g + \varepsilon_{ig},\ \alpha_g \sim iid\ [0, \sigma^2_\alpha],\ \varepsilon_{ig} \sim iid\ [0, \sigma^2_\varepsilon]. \tag{6}$$

Then $\mathrm{V}[u_{ig}] = \sigma^2_u = \sigma^2_\varepsilon + \sigma^2_\alpha$, $\mathrm{Cov}[u_{ig}, u_{jg}] = \sigma^2_\alpha$, and $\rho_u = \mathrm{Cor}[u_{ig}, u_{ig'}] = \sigma^2_\alpha/(\sigma^2_\varepsilon + \sigma^2_\alpha)$ is called the error intraclass correlation coefficient. This model yields cluster error covariance matrix $\mathbf{\Sigma}_g = \sigma^2_\varepsilon \mathbf{I}_{N_g} + \sigma^2_\alpha \mathbf{e}_{N_g}\mathbf{e}'_{N_g}$, where $\mathbf{e}_{N_g}$ is an $N_g \times 1$ vector of ones, and we use variance estimate

$$\widehat{\mathrm{V}}_{\mathrm{RE}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum\nolimits_{g=1}^{G} \mathbf{X}_g \widehat{\mathbf{\Sigma}}_g \mathbf{X}'_g\right)(\mathbf{X}'\mathbf{X})^{-1}, \tag{7}$$

where $\widehat{\mathbf{\Sigma}}_g = \widehat{\sigma}^2_\varepsilon \mathbf{I}_{N_g} + \widehat{\sigma}^2_\alpha \mathbf{e}_{N_g}\mathbf{e}'_{N_g}$, and $\widehat{\sigma}^2_\varepsilon$ and $\widehat{\sigma}^2_\alpha$ are consistent estimates for σ^2_ε and σ^2_α. We call the resulting standard errors **Moulton-type standard errors**.

2.3 Cluster-Robust Variance Estimates

The RE model places restrictions of homoskedasticity and equicorrelation within cluster. A less parametrically restrictive approach is to use the **cluster-robust variance estimator (CRVE)**

$$\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum\nolimits_{g=1}^{G} \mathbf{X}_g \widetilde{\mathbf{u}}_g \widetilde{\mathbf{u}}'_g \mathbf{X}'_g\right)(\mathbf{X}'\mathbf{X})^{-1}, \tag{8}$$

which is consistent if $\operatorname{plim} G^{-1}\sum_{g=1}^{G} \mathbf{X}_g \widetilde{\mathbf{u}}_g \widetilde{\mathbf{u}}_g{}' \mathbf{X}'_g = \operatorname{plim} G^{-1}\sum_{g=1}^{G} \mathbf{X}_g \mathbf{\Sigma}_g \mathbf{X}'_g$. Standard errors based on $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}]$ are called **cluster-robust standard errors**. In the simplest case OLS residuals are used, so $\widetilde{\mathbf{u}}_g = \widehat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\widehat{\boldsymbol{\beta}}$.

The CRVE controls for both error heteroskedasticity across clusters and quite general correlation and heteroskedasticity within cluster, at the expense of requiring that the number of clusters $G \to \infty$. It is implemented for many STATA regression commands using the cluster option (which uses $\widetilde{\mathbf{u}}_g = \sqrt{c}\widehat{\mathbf{u}}_g$ where $c = \frac{G}{G-1}\frac{N-1}{N-k} \simeq \frac{G}{G-1}$ with large N), and is used in SAS in the GENMOD procedure (which uses $\widetilde{\mathbf{u}}_g = \widehat{\mathbf{u}}_g$).

The CRVE estimate is an immediate extension of the **heteroskedastic consistent covariance matrix estimator (HCCME)** of White (1980) for nonclustered data, i.e. in the special case $N_g = 1$. Many of the methods we consider below can be viewed as extensions to the clustered case to proposed adaptations of the nonclustered HCCME case.

2.4 Unbiased Cluster-Robust Variance Estimates

A weakness of the standard CRVE with $\widetilde{\mathbf{u}}_g = \widehat{\mathbf{u}}_g$ is that it is biased, since $\mathrm{E}[\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'] \neq \mathbf{\Sigma}_g = \mathrm{E}[\mathbf{u}_g\mathbf{u}_g']$. The bias depends on the form of $\mathbf{\Sigma}_g$ but will usually be downwards.[2] Several corrected residuals $\widetilde{\mathbf{u}}_g$ for (8) have been proposed. The simplest, already mentioned, is to use $\widetilde{\mathbf{u}}_g = \sqrt{G/(G-1)}\widehat{\mathbf{u}}_g$.

Kauermann and Carroll (2001) and Bell and McCaffrey (2002) use

$$\widetilde{\mathbf{u}}_g = [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1/2}\widehat{\mathbf{u}}_g, \tag{9}$$

where $\mathbf{H}_{gg} = \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$. This transformed residual leads to unbiased CRVE in (8) in the special case that $\mathbf{\Sigma}_g = \sigma^2\mathbf{I}$, though in simulations the authors find the correction works quite well even if $\mathbf{\Sigma}_g \neq \sigma^2\mathbf{I}$.[3] Bell and McCaffrey (2002) also use

$$\widetilde{\mathbf{u}}_g = \sqrt{\frac{G-1}{G}}[\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g. \tag{10}$$

Then $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}]$ in (8) with $\widetilde{\mathbf{u}}_g$ calculated using (10) can be shown to equal the jackknife estimate of the variance of the OLS estimator.[4] This jackknife

[2] For example, Kezdi (2004) uses $\widetilde{\mathbf{u}}_g = \widehat{\mathbf{u}}_g$ and finds in his simulations that with $G = 10$ the donwards bias is between 9 and 16 percent.

[3] These papers, and that by Bell, McCaffrey and Botts (2001), also propose corrections for the more general case that $\mathbf{\Sigma}_g \neq \sigma^2\mathbf{I}$, provided $\mathbf{\Sigma}_g$ has a known parameterization, and for extension to generalized linear models.

[4] The jackknife drops in turn each observation, here a cluster, computes the leave-one-out estimate $\widehat{\boldsymbol{\beta}}_{(g)}$, $g = 1, ..., G$, and then uses variance estimate $\frac{G-1}{G}\sum_g(\widehat{\boldsymbol{\beta}}_{(g)} - \widehat{\boldsymbol{\beta}})$.

correction does not make an assumption about $\mathbf{\Sigma}_g$, and leads to downwards-biased CRVE if in fact $\mathbf{\Sigma}_g = \sigma^2\mathbf{I}$. Angrist and Lavy (2002) apply the correction (9) in an application with $G = 30$ to 40 and find that the correction increases cluster-robust standard errors by between 10 and 50 percent.

Similar corrections have been well-studied for the HCCME in the non-clustered case. The correction (9) generalizes the HC2 measure of MacKinnon and White (1985) that sets $\widetilde{u}_g = (1 - h_{gg})^{-1/2}$ where h_{gg} is the gth diagonal entry of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and is called the leverage measure in the literature on influential data points.[5] The correction (10) generalizes the HC3 measure (jackknife) of MacKinnon and White (1985). Chesher and Jewitt (1987) obtain HCCME bias results under the more general assumption that errors are heteroskedastic rather than iid, and show that the performance of various HCCME corrections depends crucially on the leverage measures, with greater problems the larger is the maximum leverage. Chesher and Austin (1991) emphasize that different design matrices $\mathbf{X}$ can lead to different assessments of HCCMEs, a result that can be expected to carry over to the clustered case.

We refer to the residual correction (10) as the CR3 variance estimator, since it is a cluster extension of the HC3 procedure. An additional complication in the clustered setting is that for some configurations of the regressor design matrix $\mathbf{I}_{N_g} - \mathbf{H}_{gg}$ need not be full rank, leading to problems in implementing these corrections.

2.5 Cluster-Robust Wald Tests

We consider two-sided Wald tests of $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$ where β_1 is a scalar component of $\boldsymbol{\beta}$.[6] We use the **Wald test statistic**

$$w = \frac{\widehat{\beta}_1 - \beta_1^0}{s_{\widehat{\beta}_1}}, \tag{11}$$

Applied to the OLS estimator, some algebra yields jackknife estimate of variance (8) where $\widetilde{\mathbf{u}}_g = \sqrt{(G-1)/G}\ [\mathbf{I}_{N_g} - \mathbf{H}_{gg}]^{-1}\widehat{\mathbf{u}}_g$. This is a multiple of the related measure proposed by Mancl and DeRouen (2001) in the more general setting of GEE.

[5] The motivation is that $\widehat{\mathbf{u}} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so that for $\mathrm{E}[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I}$, $\mathrm{E}[\widehat{\mathbf{u}}\widehat{\mathbf{u}}'] = \mathbf{M}\sigma^2\mathbf{I}\mathbf{M} = \sigma^2\mathbf{M}$. It follows that $\mathrm{E}[\widehat{u}_g^2] = \sigma^2\mathbf{M}_{gg}$, so $\mathrm{E}[(\mathbf{M}_{gg}^{-1/2}\widehat{u}_g)^2] = \sigma^2$.

[6] The generalization to single hypothesis $\mathbf{c}'\boldsymbol{\beta} - r = 0$ where $\mathbf{c}$ is a $k \times 1$ vector is trivial. For multiple hypotheses $\mathbf{C}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}$ the Wald asympototic chisquare test would be used.

where $s_{\widehat{\beta}_1}$ is the square root of the appropriate diagonal entry in $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}]$. This "t" test statistic is asymptotically normal under H_0, and we reject H_0 at significance level α if $|w| > z_{\alpha/2}$, where $z_{\alpha/2}$ is a standard normal critical value. Often $\alpha = 0.05$, in which case $z_{\alpha/2} = 1.960$.

Under standard assumptions the Wald test is of correct size as the number of clusters $G \to \infty$. The problem we focus on in this paper is that with few clusters the asymptotic normal critical values can provide a poor approximation to the correct, finite-G critical values for w, even if an unbiased variance matrix estimator is used in calculating $s_{\widehat{\beta}}$.

Small sample results are not possible even if the (clustered) errors are normally distributed. Some intuition can be gained, however, by considering balanced clusters ($N_g = N_G$) with all regressors invariant within cluster ($\mathbf{x}_{ig} = \mathbf{x}_g$) and errors u_{ig} that are iid $\mathcal{N}[0, \sigma^2]$. OLS estimation at either the individual level or the aggregate level will lead to the same OLS estimator. But the estimate of σ^2 differs, leading to t-tests with quite different degrees of freedom.

Using individual-level data, $y_{ig} = \mathbf{x}_g'\boldsymbol{\beta} + u_{ig}$, the Wald test statistic using default OLS standard errors is distributed $\mathrm{T}(N-k)$, since the errors are iid normal. Suppose we instead use aggregated data, $\bar{y}_g = \mathbf{x}_g'\boldsymbol{\beta} + \bar{u}_g$. Then the OLS estimator equals that from the individual-level regression, as clusters are balanced. Since $\bar{u}_g$ are iid $\mathcal{N}[0, \sigma^2/N_G]$ the Wald statistic using default OLS standard errors is $\mathrm{T}(G-k)$ distributed. The very large loss of degrees of freedom is due to aggregation leading to a different estimate of σ^2: here $s^2 = (G-k)^{-1}\sum_g \widehat{\bar{u}}_g^2$ rather than $s^2 = (N-k)^{-1}\sum_g \sum_i \widehat{u}_{ig}^2$.

Now suppose we instead use the CRVE in forming the Wald statistic for OLS estimation with individual-level data. This robust Wald statistic is no longer T distributed, even if the errors are iid normal. Some algebra yields $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\boldsymbol{\beta}}] = (\sum_g \mathbf{x}_g\mathbf{x}_g')^{-1}(\sum_g \overline{\widehat{u}}_g^2\mathbf{x}_g\mathbf{x}_g')(\sum_g \mathbf{x}_g\mathbf{x}_g')^{-1}$, where $\overline{\widehat{u}}_g = N_g^{-1}\sum_i \widehat{u}_{ig}$. This is a more variable variance estimator than that used with aggregated data and default OLS standard errors, which can be expanded as $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}] = (\sum_g \mathbf{x}_g\mathbf{x}_g')^{-1}(\sum_g s^2\mathbf{x}_g\mathbf{x}_g')(\sum_g \mathbf{x}_g\mathbf{x}_g')^{-1}$ with $s^2 = (G-k)^{-1}\sum_h \overline{\widehat{u}}_h^2$. Given this more variable estimate of $s_{\widehat{\beta}_1}$, the Wald statistic using cluster robust standard errors will have fatter tails than those of a $T(G-k)$ distribution. For more detailed discussion of a related result for the HCCME, see Kauermann and Carroll (2001).[7]

[7] Kezdi (2004) finds in his simulations with iid errors that CRVE standard errors have

In practice, as a small sample correction some programs use a T distribution to form critical values and p-values. STATA uses the $T(G-1)$ distribution, which the preceding example suggests is better than the standard normal, but may still not be conservative enough to avoid over-rejection. Bell and McCaffrey (2002) and Pan and Wall (2002) propose instead using a T distribution with degrees of freedom determined using an approximation method due to Satterthwaite (1941).

Rather than use OLS, Donald and Lang (2004) propose an alternative two-step estimator that leads to a Wald test that in some special cases is $T(G-k_1)$ distributed where k_1 is the number of regressors that are invariant within cluster and often $k_1 = 2$ (the intercept and the clustered regressor of interest).

We instead continue to use the standard OLS estimator with CRVE, and bootstrap to obtain bootstrap critical values that provide an asymptotic refinement and may work better than other inference methods for OLS when there are few clusters.

3 Cluster Bootstraps

Bootstrap methods generate a number of pseudo-samples from the original sample, for each pseudo-sample calculate the statistic of interest, and use the distribution of this statistic across pseudo-samples to infer the distribution of the original sample statistic. This is very similar to a Monte Carlo simulation, except the pseudo-samples are generated using fewer distributional assumptions.[8]

There is no single bootstrap as there are different statistics that we may be interested in, different ways to form pseudo-samples, and even for a given statistic and resampling method there are different ways to use results for statistical inference.

We provide considerable discussion here and in the appendix because in later sections we use a range of bootstrap methods that may be unfamiliar

standard deviation 2 to 3 times that for default standard errors. He does not consider the implications for Wald tests.

[8] The bootstrap was introduced by Efron (1979). Standard book treatments are Hall (1992), Efron and Tibsharani (1993), and Davison and Hinkley (1997). In econometrics see Horowitz (2001), MacKinnon (2002), and the texts by Davidson and MacKinnon (2004) and Cameron and Trivedi (2005).

to applied microeconometricians. The statistic considered is the Wald test statistic w defined in (11), for two-sided test of $H_0 : \beta_1 = \beta_1^0$ against $H_a : \beta_1 \neq \beta_1^0$.

The data are clustered into G independent groups, so the resampling method should be one that assumes independence across clusters but preserves within cluster features such as correlation.

3.1 Pairs Cluster Bootstrap-T

The obvious method is to resample the clusters with replacement from the original sample $\{(\mathbf{y}_1, \mathbf{X}_1), ..., (\mathbf{y}_G, \mathbf{X}_G)\}$. The following procedure provides a starting point.

Pairs Cluster Bootstrap-T Procedure

1. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample.

   (b) Calculate the Wald test statistic

   $$w_b^* = \frac{\widehat{\beta}_{1,b}^* - \widehat{\beta}_1}{s_{\widehat{\beta}_{1,b}^*}},$$

   where $\widehat{\beta}_{1,b}^*$ is obtained from OLS estimation using the b^{th} pseudo-sample, its standard error $s_{\widehat{\beta}_{1,b}^*}$ is estimated using the same cluster-robust method as used in calculating w, and $\widehat{\beta}_1$ is the original sample OLS estimate.

2. Use the empirical distribution of $w_1^*, ..., w_B^*$ to determine critical values and p-values. Thus reject H_0 at level α if and only if

   $$w < w_{[\alpha/2]}^* \text{ or } w > w_{[1-\alpha/2]}^*,$$

   where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.[9]

[9] For B realizations $x_1, ..., x_B$, the q^{th} quantile $x_{[q]}$ is the smallest value of x such that a proportion q of $x_1, ..., x_B$ are less than or equal to $x_{[q]}$. More formally $x_{[q]}$ is the smallest value of x such that $\sum_{b=1}^{B} \mathbf{1}(x_b \leq x_{[q]}) = qB$.

As an example, if $w = -2.10$, $w^*_{[.025]} = -2.32$ and $w^*_{[.975]} = 1.87$ then we would not reject H_0 at significance level $\alpha = 0.05$ since -2.10 does not fall outside $(-2.32, 1.87)$. This is called a **nonsymmetric test**, as the critical values in the two tails differ. We could also do a **symmetric** two-tailed test, in which case in step 2 we would reject H_0 if $|w| > |w^*|_{[\alpha]}$, where $|w^*|_{[q]}$ denotes the q^{th} quantile of $|w^*_1|, ..., |w^*_B|$.

This resampling method is called a **pairs cluster bootstrap** because, unlike other methods presented below, the pair $(\mathbf{y}, \mathbf{X})$ is jointly resampled. Alternative names used in the literature include **cluster bootstrap**, **case bootstrap**, **nonparametric bootstrap**, and **nonoverlapping block bootstrap**. Because this resampling method views the original sample as the dgp (with $\beta_1 = \widehat{\beta}_1$), the pseudo-sample statistics w^*_b are centered on this dgp value $\widehat{\beta}_1$.

The **bootstrap-t procedure**, proposed by Efron (1981) for confidence intervals, is also called a **percentile-t procedure**, because the "t" test statistic w is bootstrapped, and a **studentized bootstrap**, since the Wald test statistic is a studentized statistic. The bootstrap-t procedure has advantages that we now present.

3.2 Asymptotic Refinement of Standard Bootstrap Methods

The bootstrap-t procedure was preceded by more obvious bootstrap procedures that bootstrap the parameter estimate $\widehat{\beta}_1$, giving B estimates $\widehat{\beta}^*_{1,1}, ..., \widehat{\beta}^*_{1,B}$, rather than the test statistic.

The **percentile procedure** simply rejects H_0 if the original sample estimate $\widehat{\beta}_1$ falls outside $(\widehat{\beta}^*_{1,[\alpha/2]}, \widehat{\beta}^*_{1,[1-\alpha/2]})$ where $\widehat{\beta}^*_{1,[q]}$ denotes the q^{th} quantile of $(\widehat{\beta}^*_{1,1}, ..., \widehat{\beta}^*_{1,B})$.

The **bootstrap-se procedure** uses these estimates to form the **bootstrap estimate of standard error**

$$s_{\widehat{\beta}_1,\mathrm{B}} = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\beta}^*_{1b} - \overline{\widehat{\beta}^*_1})^2 \right)^{1/2}, \tag{12}$$

where $\overline{\widehat{\beta}^*_1} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}^*_{1b}$, uses $s_{\widehat{\beta}_1,\mathrm{B}}$ to calculate the Wald test, so

$$w_{\mathrm{BSE}} = \frac{\widehat{\beta}_1 - \beta^0_1}{s_{\widehat{\beta}_1,\mathrm{B}}}, \tag{13}$$

and rejects H_0 at significance level α if and only if $|w_{\text{BSE}}| > z_{\alpha/2}$. Here the bootstrap is used merely to obtain an estimate of the standard error of $\widehat{\beta}_1$, an advantage if this is difficult to do using conventional methods.

The bootstrap-t, percentile and bootstrap-se procedures are all asymptotically valid tests under standard modelling assumptions. In particular, a test at nominal (or desired) significance level 0.05 will have true size of 0.05 (and power of 1.00 against a fixed alternative) as $G \to \infty$.

But for small G all three procedures will have true size different from 0.05 as they rely on asymptotic approximation. Consider a test with nominal significance level or **nominal size** α. An asymptotic approximation yields an actual rejection rate or **true size** $\alpha + O(G^{-j/2})$, where $O(G^{-j/2})$ means is of order $G^{-j/2}$ and G is the number of clusters. Then the true size goes to α as $G \to \infty$, provided $j > 0$. Larger j is preferred, however, as then convergence to α is faster. A bootstrap provides **asymptotic refinement** if it leads to j larger than that for conventional (first-order) asymptotic methods.

Asymptotic refinement is more likely to occur if the bootstrap is applied to an **asymptotically pivotal statistic**, meaning one with asymptotic distribution that does not depend on unknown parameters; see Appendix A.1 for a more complete discussion. The cluster-robust Wald statistic is asymptotically pivotal as it is standard normal (no unknown parameters), so the bootstrap-t procedure provides asymptotic refinement. By contrast the percentile and bootstrap-se methods bootstrap the OLS estimator, which is not asymptotically pivotal as the asymptotic variance is dependent on unknown parameters.

The bootstrap-t procedure is not the only way to obtain an asymptotic refinement. In particular, the **bias-corrected accelerated (BCA) procedure**, defined in Appendix A.2, is a variation of the percentile method that leads to asymptotic refinement of the same order as the percentile-t method.

3.3 Wild Cluster Bootstrap-T

For a regression model with additive error, resampling methods other than pairs cluster can be used. In particular, one can hold regressors $\mathbf{X}$ constant throughout the pseudo-samples, while resampling the residuals which can be then used to construct new values of the dependent variable $\mathbf{y}$.

The obvious method is a **residual cluster bootstrap** that resamples with replacement from the original sample residual vectors to give residuals $\{\widehat{\mathbf{u}}_1^*, ..., \widehat{\mathbf{u}}_G^*\}$ and hence pseudo-sample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_g)\}$ where $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}_g^*$.

This resampling scheme has two weaknesses. First, it assumes that the regression error vectors $\mathbf{u}_g$ are iid, whereas in Section 2 we were specifically concerned that the variance matrix $\boldsymbol{\Sigma}_g$ will differ across clusters. Second, it presumes a balanced data set where all clusters are the same size.

The next bootstrap relaxes both these restrictions.

Wild Cluster Bootstrap-T Procedure

1. Obtain the OLS estimator $\widehat{\boldsymbol{\beta}}$ and the associated OLS residuals $\widehat{\mathbf{u}}_g$, $g = 1, ..., G$.

2. Do B iterations of this step. On the b^{th} iteration:

   (a) For each cluster $g = 1, ..., G$, form $\widehat{\mathbf{u}}_g^* = \widehat{\mathbf{u}}_g$ with probability 0.5 or $\widehat{\mathbf{u}}_g^* = -\widehat{\mathbf{u}}_g$ with probability 0.5, and hence $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}_g^*$. This yields wild cluster bootstrap resample $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G^*)\}$.

   (b) Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ and given these form the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$.

3. Reject H_0 at level α if and only if

$$w < w_{[\alpha/2]}^* \text{ or } w > w_{[1-\alpha/2]}^*,$$

   where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

The wild bootstrap was proposed by Wu (1986) for regression in the nonclustered case. Its asymptotic validity and asymptotic refinement were proven by Liu (1988) and Mammen (1993). Horowitz (1997, 2001) provides a Monte Carlo demonstrating good size properties. We use a version, called Rademacher weights, that offers asymptotic refinement if $\widehat{\boldsymbol{\beta}}$ is symmetrically distributed, the case if errors are symmetric. If $\widehat{\boldsymbol{\beta}}$ is asymmetrically distributed, our version is still asymptotically valid, but a different version provides asymptotic refinement. Davidson and Flachaire (2001) provide theory and simulation to support using Rademacher weights even in the asymmetric

case. See Appendix Section A.3 for further discussion. Here we have extended the wild bootstrap to a clustered setting. The only study to do so that we are aware of is the brief application by Brownstone and Valletta (2001).

Several authors, particularly Davidson and MacKinnon (1999), advocate use of bootstrap resampling methods that impose the null hypothesis. This is possible using both residual and Wild bootstraps. Thus we present results based on bootstraps that impose the null, in which case the bootstrap Wald statistics are centered on β_1^0 rather than $\widehat{\beta}_1$, and the residuals bootstrapped are those from the restricted OLS estimator $\widetilde{\boldsymbol{\beta}}$ that imposes $H_0 : \beta_1 = \beta_1^0$. For details see Appendix A.2.

3.4 Bootstraps with Few Clusters

With few clusters the bootstrap resampling methods produce a distinctly finite number of possible pseudo-samples, so the bootstrap distribution $w_1^*, ..., w_q^*$ will not be smooth even for large B. Furthermore, in some pseudo-samples $\widehat{\beta}_1$ or $s_{\widehat{\beta}_1}$ may be inestimable.

First, consider pairs cluster resampling. Each bootstrap resample contains G clusters. Some of the original sample clusters will appear more than once, while others will not appear at all. For example, with $G = 5$ we might obtain bootstrap resample $\{(\mathbf{y}_3, \mathbf{X}_3), (\mathbf{y}_5, \mathbf{X}_5), (\mathbf{y}_2, \mathbf{X}_2), (\mathbf{y}_3, \mathbf{X}_3), (\mathbf{y}_1, \mathbf{X}_1)\}$. In general, this bootstrap has $\binom{2G-1}{G-1}$ possible unordered recombinations of the data; see Hall (1992, p.283). There are many possible combinations even when there are few clusters: 126 combinations for $G = 5$ and $92,378$ combinations for $G = 10$. Nonetheless, implementation problems can arise. For example, if the k^{th} regressor is binary and is always invariant within cluster (so always 0 or always 1 for given g) then with few clusters some bootstrap resamples may have all clusters with the k^{th} regressor taking only value 0 (or value 1), so that $\widehat{\beta}_k$ is inestimable. This issue does not arise when regressors and dependent variables take many different values, such as in the Section 4 Monte Carlos. But it does arise in our application to the BDM (2004) and Gruber and Poterba (1994) differences-in-differences example, because the regressors of interest in those cases are indicator variables.

Second, consider the wild cluster bootstrap. For each cluster there are two possible resample values, so with G clusters there are 2^G possible recom-

binations: 32 combinations for $G = 5$ and $1,024$ combinations for $G = 10$. In general this is much less than for pairs cluster. However, a binary regressor invariant within cluster will not cause a problem as we are not resampling the regressors. And, as clear from Appendix A.2, the Wild bootstrap need not be restricted to a two-point distribution, though we do not pursue this.

3.5 Bootstrap Discussion

Clearly there are many ways to bootstrap when data are clustered. The standard applied procedure is to use pairs cluster resampling and the bootstrap-se procedure which offers no asymptotic refinement.[10] There are just a few studies that we are aware of that consider asymptotic refinement.

In the statistics literature, Sherman and le Cessie (1997) conduct simulations for OLS with as few as ten clusters. For 90 percent confidence intervals, they find that the pairs cluster bootstrap-t undercovers by considerably less than the percentile method which in turn is better than using conventional robust confidence intervals based on cluster-robust standard errors. The bootstrap-t intervals are wider than the others, however, and occasionally have very large end-points. The authors also study logit models, and they find occasionally that bootstrap resamples are inestimable due to all dependent variables taking value zero (or one).

Flynn and Peters (2004) consider cluster randomized trials where a pairs cluster bootstrap draws G clusters by separately resampling from the $G/2$ treatment clusters and the $G/2$ control clusters. For skewed data and few clusters they find that pairs cluster BCA confidence intervals have considerable undercoverage, even more than conventional robust confidence intervals, though in their Monte Carlo design the robust intervals do remarkably well. The authors also consider a second-stage of resampling within each cluster, using a method for hierarchical data given in Davison and Hinkley (1997) that is applicable if the random effects model (6) is assumed.

In the econometrics literature, BDM (2004) apply a pairs cluster bootstrap using the bootstrap-t procedure. BDM use default OLS standard errors, however, rather than cluster-robust standard errors, in computing both the original data and the resampled data Wald statistics. Because of this

[10] Stata, for example, offers this as an estimation option. Additionally it has a cluster pairs bootstrap procedure that does provide BCA confidence intervals in addition to bootstrap-se confidence intervals, but these are rarely used.

their method will in general not yield tests of correct size. It can nonetheless lead to considerable size improvement compared to using default standard errors, since it controls for any inconsistent estimation of standard errors up to a constant scale factor. Specifically, if w is computed using $a \times s_{\widehat{\beta}_1}$ and w_b^* is computed using $a \times s^*_{\widehat{\beta}_1}$, for constant $a > 0$ that does not vary across bootstrap samples, then we obtain the same rejection rate regardless of the value of a and asymptotic refinement is obtained.[11] For the BDM example the scale factor is data-dependent, however, and varies across bootstrap resamples. The authors find that their bootstrap does better than using default OLS standard errors and standard normal critical values, yet surprisingly does worse than using cluster robust standard errors with standard normal critical values.

The only study we know of that uses wild bootstraps in a clustered setting is the brief application by Brownstone and Valletta (2001).

3.6 Test Methods used in this Paper

In the remainder of the paper we implement the Wald test using nine bootstrap procedures, as well as four non-bootstrap procedures. Table 1 provides a summary.

Our first four tests do not use the bootstrap and differ only in the method from section 2 used to calculate $\widehat{\mathrm{V}}[\widehat{\boldsymbol{\beta}}]$. They use, respectively, the default variance estimate (4), the Moulton-type estimate (7), the cluster-robust estimate (8), and the cluster-robust estimate with jackknife corrected residuals (10). Method 1 is invalid if there is clustering, method 2 is invalid unless the clustering follows a random effects model, while methods 3 to 4 are asymptotically valid provided clusters are independent.

Methods 5 to 7 use the bootstrap-se procedure, with bootstrap standard error computed using (12) and Wald statistic using (13). We use three different cluster bootstrap resampling methods, respectively, the pairs cluster bootstrap, the residual clusters bootstrap with H_0 imposed, and the wild bootstrap with H_0 imposed. For details see Appendix A.2.3. Methods 5-7 do not provide asymptotic refinement, and method 6 is valid only if cluster error vectors are iid.

[11] This fact follows since the cdf of a random variable is invariant to monotonic transformations of the random variable.

Method 8 uses the BCA bootstrap with pairs cluster resampling, see Appendix A.2.2, to provide an asymptotic refinement.

Methods 9 to 13 use the bootstrap-t procedure. The first three of these methods use pairs cluster resampling with different standard error estimates. Method 9 is the already discussed method of BDM that uses default standard errors rather than CRVE standard errors. Methods 10 and 11 use different variants of the CRVE defined in (8), respectively, the standard CRVE and the CR3 correction. In each case the same variance matrix estimation method is used for both the original sample and bootstrap resamples. Methods 12 and 13 use, respectively, residual and wild bootstraps, and both use the standard CRVE estimate and impose H_0. Method 12 is valid only if cluster error vectors are iid. For details see Appendix A.2.1.

The bootstrap-t and BCA procedures should set the number of bootstrap replications B so that $(B+1) \times \alpha$ is an integer. For an explanation see Davidson and MacKinnon (2004, p.164), for example. Additionally for low B there will be nontrivial simulation error in the bootstrap. A common choice is $B = 999$, since $1{,}000 \times \alpha$ is an integer for common values of α such as $\alpha = 0.05$.

4 Monte Carlo Simulations

To examine the finite-sample properties of our methods we conducted several Monte Carlo exercises for dgp a linear model with intercept and single regressor. The error is clustered according to a random effects model, with either constant correlation within cluster or departures from this induced by heteroskedasticity. This design is relevant to a cross-section study of individuals with clustering at the state level, for example. The regressor and dependent variable are continuous and take distinct values across clusters and (usually) within clusters, so that even with few clusters it is unlikely that a pairs cluster bootstrap sample will be inestimable.

Then

$$y_{ig} = \beta_0 + \beta_1 \mathbf{x}_{ig} + u_{ig}, \tag{14}$$

so $\boldsymbol{\beta} = [\beta_0 \ \beta_1]'$, with different generating processes for $\mathbf{x}_{ig}$ and u_{ig} used in subsequent subsections. Estimation is by OLS, as in (1). Since $\beta_1 = 1$ in the dgp, we set $\beta_1^0 = 1$ and the Wald test statistic is $w = (\widehat{\beta}_1 - 1)/s_{\widehat{\beta}_1}$.

We perform R replications, where each replication yields a new draw of data from the dgp, and leads to reject or nonrejection of H_0 depending on whether or not $|w| > 1.96$. In each replication there are G groups ($g = 1, ..., G$), with N_G individuals ($i = 1, ..., N_G$) in each group. We varied the number of groups G from 5 to 30 and usually set $N_G = 30$. The various methods used down each column of Tables 2-4 are then applied to the same generated data. For bootstraps we used $B = 399$ bootstraps rather than the recommended $B = 999$ or higher. This lower value is fine for a Monte Carlo exercise, since the bootstrap simulation error will cancel out across Monte Carlo replications.

We estimate the **actual rejection rate** a, by $\widehat{a}$, the fraction of the R replications for which H_0 is rejected. This is an estimate of the true size of the test which should be 0.05. With a finite number of replications a may be differ from the true size due to simulation error. The simulation standard error is $s_{\widehat{a}} = \sqrt{\widehat{a}(1-\widehat{a})/(R-1)}$. For $R = 1000$ replications, $s_{\widehat{a}} \simeq 0.007$ for $\widehat{a} = 0.05$ and $s_{\widehat{a}} \simeq 0.009$ for $a = 0.10$. We can reject that true size is the desired 0.05 at the 95% significance level when $|\widehat{a} - 0.05| > 1.96 \times s_{\widehat{a}}$. With $R = 1000$ a value of $\widehat{a} = 0.07$ will be statistically significantly different from 0.05, while a value of $\widehat{a} = 0.06$ will not.[12]

4.1 Simulations with Homoskedastic Clustered Errors

In the first simulation exercise both regressors and errors are correlated within group, with errors homoskedastic. Data were generated according to:

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} = \beta_0 + \beta_1 (z_g + z_{ig}) + (\varepsilon_g + \varepsilon_{ig}), \qquad (15)$$

with z_g, z_{ig}, ε_g, and ε_{ig} each an independent $\mathcal{N}[0, 1]$ draw, and $\beta_0 = 0$ and $\beta_1 = 1$. Here the components z_g and ε_g that are common to individuals within a group induce within group correlation of both regressors and errors with $\rho_x = 0.5$ and $\rho_u = 0.5$. The simulation is based on $R = 1000$ Monte Carlo replications.

Our first results appear in Table 2. Each column gives results for the various number of groups ($G = 5$, 10, 15, 20, 25, 30) and throughout $N_G = 30$. The first entry is the estimated true size of the test, the proportion

[12] Values of R higher than $1,000$ would be better, but each entry in a table already requires $RB \simeq 400,000$ separate estimations.

of times the null hypothesis is rejected. The Monte Carlo standard error is given in parentheses. Each row presents a different method, detailed in section 3.6. For comparison, we also show the rejection rate that would hold if we used the asymptotic normal critical value of 1.96, but the Wald statistic actually had a T distribution with $G-k = G-2$ degrees of freedom. This rejection rate is $\Pr[|T| > 1.96|T \sim T_{G-2}]$, though recall from section 2.5 that even with normal errors the finite distribution of the various Wald statistics is unknown.

We begin with conventional (nonbootstrap) Wald tests using different estimators of standard errors. The default OLS standard errors that assume iid errors do poorly here, with rejection rates given in row 1 of 0.43 to 0.50 that are much higher than 0.05.[13] This illustrates the need to correct standard errors for clustering. The Moulton-type estimate for standard errors should work well here since it this takes advantage of correct knowledge of the dgp. The rejection rates in row 2 are considerably higher than 0.05, especially for low G, though are similar to those expected if the Wald test statistic is actually $\mathrm{T}(G-2)$ distributed (see the bottom row). The cluster-robust standard errors lead to rejection rates much better than those using default standard errors, though still over-reject with rejection rates that are 0.01 to 0.06 greater than those using Moulton-type standard errors. The CR3 correction to cluster-robust standard errors leads to larger standard errors and to rejection rates that from row 4 are much closer to 0.05, though still significantly different from 0.05.

We then consider using the bootstrap to compute standard errors. The pairs cluster bootstrap-se method yields rejection rates in row 5 that are very similar to the cluster-robust method, except for $G = 5$. The residual cluster bootstrap-se method leads to rejection rates in row 6 that are close to 0.05. From row 7, the wild cluster bootstrap-se method under-rejects for $G \leq 10$, and rejects at level close to 0.05 for $G > 10$. The closeness to 0.05 of the latter two bootstrap methods is surprising given that they do not offer an asymptotic refinement.

The BCA bootstrap with pairs cluster resampling should provide an

[13] These rejections rates are consistent with theory. The dgp is a classic balanced two-way random effect error, so (5) applies yielding standard error inflation $\sqrt{1+(30-1)0.5\times 0.5} = \sqrt{8.25} = 2.87$. Using the underestimated standard errors leads to a Wald statistic with actual asymptotic size equal to $\Pr[|z| > 1.96/2.87] \simeq 0.50$. This underestimation depends on group size rather than number of groups.

asymptotic refinement, yet from row 8 it has rejection rates similar to those using CR standard errors (row 3), aside from small improvement for $G = 5$ and $G = 10$ where the rejection rates are nonetheless still in excess of 0.10.

The remainder of the Table use the theoretically preferred bootstrap-t procedure with various resampling methods. Even though it uses default standard errors, the BDM bootstrap (row 9) does better than using CR standard errors and is a great improvement compared to not bootstrapping (row 1). The pairs cluster bootstrap-t has rejection rates in row 10 of 0.08 that are much closer to 0.05 than tests without bootstrap, though they are still significantly different from 0.05. The CR3 correction makes little difference. Apparently while it leads to considerable decrease in the value of the Wald statistic (see the difference in rejection rates between rows 3 and 4), comparable decreases occur in the bootstrap resamples leading to similar rejection rates. Both the residual cluster bootstrap-t and wild cluster bootstrap-t rejection rates are not statistically different from 0.05 (with the exception of the residual bootstrap with $G = 5$).

In summary, Table 2 demonstrates that all the bootstrap-t methods are an improvement on the usual cluster-robust method with standard normal critical values; the BCA method provides no improvement; and the residual cluster bootstrap-se also performs well.

4.2 Simulations with Heteroskedastic Clustered Errors

The second simulation brings in the additional complication of heteroskedastic errors. Then the Moulton-type correction and the residual bootstrap are no longer valid theoretically.

We generated data according to the process:

$$y_{ig} = \alpha + \beta x_{ig} + u_{ig} = \alpha + \beta\left(z_g + z_{ig}\right) + \left(\varepsilon_g + \varepsilon_{ig}\right), \tag{16}$$

with z_g, z_{ig}, ε_g and ε_g again independent $\mathcal{N}[0,1]$ draws, but now $\varepsilon_{ig} \sim \mathcal{N}[0, 9 * (z_g + z_{ig})^2]$. The dgp sets $\alpha = 1$ and $\beta = 1$. Here $\rho_x = 0.5$ again, while the error is heteroskedastic both within and across clusters and has correlation coefficient less than the Table 2 case of 0.5.

Results appear in Table 3. Default OLS standard errors again do poorly, with rejection rates hovering around 0.30, though these results are better than those in Table 2 due to the lower error correlation in the present design.

The Moulton-type correction breaks down given the heteroskedasticity, as expected, with rejection rates in row 2 of $0.17 - 0.26$. The cluster-robust methods do a little better than in the preceding table, but in rows 3 and 4 still generally exceed 0.05.

The residual cluster bootstrap-se method now breaks down due to heteroskedasticity, with rejection rates in row 6 in excess of 0.15. The pairs cluster bootstrap-se and wild cluster bootstrap-se methods (rows 5 and 7) perform similarly to Table 2.

The BCA bootstrap again has rejection rates in row 8 similar to those using CR standard errors (row 3), aside from small improvement for $G = 5$ and $G = 10$.

The results for the bootstrap-t methods in rows 9 to 13 are similar to those in Table 2. The incorrect BDM bootstrap-t (row 9) has similar high rejection rates to those in Table 2, aside from marked deterioration for $G = 5$. The remaining bootstrap-t methods all yield rejection rates less than 0.08, with the residual cluster bootstrap-t and wild cluster bootstrap-t doing best. The good performance of the residual cluster bootstrap-t is surprising given that errors are heteroskedastic across clusters.

In summary, the Table 3 results for inference with heteroskedastic clustered errors are similar to those for homoskedastic clustered errors except that, as expected, the Moulton-type correction and residual cluster bootstrap-se methods now perform very poorly. The bootstrap-t methods are an improvement on the usual cluster-robust method with standard normal critical values, while the BCA method provides no improvement.

4.3 Alternative Critical Values, Cluster Sizes and Regressor Design

We perform a third set of Monte Carlo experiments to examine how the different estimators perform under varying assumptions. These simulations are presented in Table 4 with each simulation based on $G = 10$ groups.

Column 1 of Table 4 provides a baseline against which the other results are compared. It uses the same dgp as that of Table 2, though the simulations begin from a different seed so that actual rejection rates differ from the $G = 10$ column in Table 2 due to simulation variability.

Tables 2 and 3 used asymptotic normal critical values in performing the Wald test using methods 1 to 7. In Table 4 column 2 we instead use critical

values from a T distribution with 8 degrees of freedom, an ad hoc finite sample correction, so that we reject H_0 if $|w| > 2.306$ rather than $|w| > 1.960$. Then the Moulton-type estimator and the CR3 correction lead to rejection rates not statistically significant from 0.05. The CR standard errors and pairs cluster bootstrap-se still lead to over-rejection, though by not as much. And the residual cluster bootstrap-se and wild cluster bootstrap-se, which seem to do very well when asymptotic normal critical values are used, now lead to great under-rejection.

In columns 3 to 5 of Table 4 we consider alternative cluster sizes of, respectively, 2, 10 and 100 observations, whereas Tables 2 and 3 always used $N_G = 30$ observations per cluster. For method 1, the effect of ignoring clustering altogether increases greatly with cluster size, as predicted by (5). Once clustering is accounted for, by any of methods 2-13, rejection rates do not vary significantly with cluster size.

In column 6 of Table 3 we examine the performance of the various testing methods when there are there are three additional regressors, each with no clustering component, and we continue to test the first regressor. The four regressors are scaled down by a factor of 1/2, so that the sum of their variances will equal the variance of the single regressor used in the dgp of column 1. The only significant change in rejection rates is an increase in the already high rejection rate for method 1 which neglects clustering.

All preceding regression designs set the intraclass correlation ρ_x of the regressor of interest to be 0.5. In column 7 we increase ρ_x to $\rho_x = 1$ (cluster-invariant regressor with $x_{ig} = z_g$) and in column 8 we decrease $\rho_x = 0$ in column 8 (iid regressor with $x_{ig} = z_{ig}$). In both cases the regressor is scaled up by $\sqrt{2}$ to keep $\mathrm{V}[x_{ig}]$ unchanged.

With cluster-invariant regressor (column 7) the failure to control for clustering is magnified and the row 1 rejection rate is largest than that in the benchmark column 1, as expected; the other nonbootstrap methods that attempt to control for clustering also have higher rejection rates. For bootstrap-se and bootstrap-t there is little change in rejection rates, except that for reasons unknown the pairs cluster bootstrap (both bootstrap-se and bootstrap-t) now has rejection rates not statistically significantly different from 0.05.

With iid regressor (column 8) and the current random effects dgp for the errors formula (5) predicts that the default OLS standard errors are

consistent and the rejection rate in row 1 is close to 0.05. The Moulton-type and CR estimators also have rejection rates much closer to 0.05. The various bootstrap procedures lead to rejection rates that are all within 0.03 of those in column 1, with no obvious pattern.

Finally, in column 9 we change the dgp to examine an unbalanced setting, so that one half of the clusters are small (with group size $N_G = 10$) and half of the clusters are large (with group size $N_G = 50$). The residual cluster bootstrap requires equal cluster sizes, so it cannot be used in this design. The remaining methods yield results qualitatively similar to those in column 1, with the main change being that the standard CRVE leads to much larger over-rejection in row 3.[14]

In summary, all the bootstrap-t methods are an improvement on the usual cluster-robust method with standard normal critical values; the BCA method provides no improvement; and the residual cluster bootstrap-se also performs well.

Table 4 indicates that when non-bootstrap methods are used to control for clustering, it is better to use critical values from a T distribution than from a standard normal, results vary little with cluster size, and controlling for clustering is more difficult as the regressors become more highly correlated within cluster. Rejection rates for bootstrap-t methods varied little across the various designs, except that the pairs cluster and pairs CR3 bootstrap-t methods performed better when regressors are cluster invariant.

The remaining conclusions are similar to those from Tables 2 and 3. Again the bootstrap-t methods are an improvement on the usual cluster-robust method, while the BCA method provides no improvement. And again for the bootstrap-t with pairs cluster resampling there is no advantage to making finite sample corrections to the CRVE.

[14] For the pairs cluster bootstrap we select with replacement ten clusters, so that with unbalanced cluster sizes the number of observations in each bootstrap pseudo-sample will generally differ from N. By contrast the wild cluster bootstrap pseudo-samples will all be of size N.

5 Bertrand, Duflo and Mullainathan (2004) Simulations

To enable a more practically familiar application of our methods, we now consider the differences-in-differences setup explored in Bertrand, Duflo, and Mullainathan (2004).[15]

The data set is of U.S. states over time. The dependent variable is the state-by-year average log wage level (after partialling out certain individual characteristics). For such a variable, the error term within cluster is serially correlated, even if state and year fixed effects are included as regressors. By contrast the random effects model implies equicorrelation for the error term within cluster. The regressor of interest is a state policy dummy variable that is binary, making it more likely that with few clusters a pairs cluster bootstrap sample will be inestimable.

The original data is CPS data on many individuals over time and states. Most of the BDM (2004) study uses a smaller data set that aggregates individual observations to the state-year level. We begin with these data, which have the advantage of being balanced and relatively small, before moving to the individual data.

5.1 Aggregated State-year Data

Using our choice of subscripts, the ig^{th} observation is for the i^{th} year in the g^{th} state. There are fifty states and twenty-one years, so $G = 50$ and $N_G = 21$. The model estimated is

$$y_{ig} = \alpha_g + \gamma_i + \beta_1 I_{ig} + u_{ig},$$

where y_{ig} is a year-state measure of excess earnings (after control for age and education), the regressors are state dummies, year dummies, and a policy change indicator I_{ig}.[16]

[15] We extracted individual-level data from the relevant CPS data sets and, when appropriate, aggregated these data using the method presented in BDM (2004). This gave data similar to that in BDM (2004). We thank these authors for sharing some of their data with us, enabling this comparison.

[16] We retain our notation for consistency with the rest of our discussion. However, more obvious subscripts for this problem are i for individual, s for state and t for year. The underlying model is $y_{ist} = \alpha_s + \gamma_t + \mathbf{x}'_{ist}\boldsymbol{\delta} + \beta I_{st} + u_{ist}$, where y_{ist} is individual log-earnings for women aged 25-50 years, and $\mathbf{x}_{ist}$ is age and education. BDM use a two-step OLS

If a policy change occurs in state g at time i^*, then $I_{ig} = 0$ for $i < i^*$ and $I_{ig} = 1$ for $i \geq i^*$. BDM's experiments randomly assign a policy change to occur in half the states, and when it does occur it occurs somewhere between the sixth and fifteenth year. Since policy changes happen only once and are not reversed, the regressor I_{ig} will be highly correlated over i for state g. In each simulation a different draw of G states with replacement is made from the original 50 states.

The Wald statistic studied is $w = \widehat{\beta}_1 / s_{\widehat{\beta}_1}$. BDM investigate size properties by letting the policy change be a "placebo" regressor that has no effect on y_{ig}. So for two-sided tests of $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ at significance level 0.05 we should reject H_0 in five percent of the simulations. They also investigate the power against the alternative power $H_a : \beta_1 = 0.02$ by actually increasing y_{ig} by 0.02 when $I_{ig} = 1$.

BDM consider the effect of having a small number of clusters by letting $G = 6$, 10, 20, and 50. The key tables we consider are BDM's Table 5, which uses their version of the block bootstrap, and Table 8, which uses the cluster-robust method. They find that (1) default standard errors do poorly, leading to actual rejection rates between 0.4 and 0.5; (2) cluster-robust standard errors do well for all but $G = 6$ (which has simulated rejection rate 0.12); and (3) their bootstrap, which they call a block bootstrap and is discussed in our section 3.5, does poorly for low numbers of clusters, with actual rejection rates 0.44, 0.23 and 0.13 for $G = 6$, 10 and 20, respectively.

The first row of our Table 5 shows high rejection rates when default standard errors are used. Our results are similar to rows 1, 3, 5 and 7 of BDM's Tables 5 and 8, though in some cases they are as many as three simulation standard errors different. The Moulton-type estimator gives rejection rates in row 2 that show little improvement. This is a consequence of errors being serially correlated rather than equicorrelated within cluster. The third row uses the cluster-robust variance estimator, and gives results very close to the comparable rows 2, 4, 6, and 8 of BDM's Table 8.

Rows 4 to 6 of Table 5 give rejection rates when the Wald statistic is calculated using bootstrap standard error estimates. These generally lead to tests with actual size between 0.04 and 0.09. The one notable exception is that the cluster-pairs standard error bootstrap (row 4) produces severe

procedure: (1) regress y_{ist} on $\mathbf{x}_{ist}$ yielding OLS residual $\widehat{u}_{ist}$; (2) regress $\overline{\widehat{u}}_{st} = N_{st}^{-1} \sum_i \widehat{u}_{ist}$ on state dummies, year dummies, and I_{st}. Thus our y_{ij} is their $\overline{\widehat{u}}_{st}$.

under-rejection (0.001) with $G = 6$. Informal experimentation suggests to us that this is due to the fact that many bootstrap replications (with only a couple of states sampled) sample only one "treatment" or "control" state. For these replications, the treatment dummy (or constant) is fit perfectly, and so has zero estimated residuals. When these "zero" residuals are plugged into the CRVE formula (8) the resulting $\widehat{\mathrm{V}}_{\mathrm{CR}}[\widehat{\beta}_1^*]$ is unreasonably small, leading to Wald statistics in some bootstrap resamples that are too large to consistently represent the Wald statistic's true distribution. This in turn results in the severe under-rejection.[17]

The BCA method with pairs cluster resampling in general leads to greater over-rejection in row 7 than when CR standard errors are used (row 3).

The remaining rows 8 to 11 of Table 5 give rejection rates for various bootstrap-t procedures. From row 8 we find that the BDM bootstrap performs similarly to cluster-robust standard errors. This is a substantial improvement compared to row 1, but still there is over-rejection, especially for $G = 6$. For reasons we cannot explain the rejection rates we obtain are considerably lower than those given in BDM Table 5. The pairs cluster bootstrap-t under-rejects appreciably for both $G = 6$ and $G = 10$ for reasons already discussed for pairs bluster bootstrap-se. The residual and wild cluster bootstrap-t methods (rows 9 and 10) do very well with actual rejection rates approximately equal to 0.05, even for $G = 6$. Both these bootstraps have the advantage that only residuals are resampled – the regressors are not resampled.

The discussion so far has focused on size. The even columns in Table 5 report power against a fixed alternative. As expected, power increases as the number of clusters increases, permitting greater precision in estimation. In almost all cases in Table 5, for given G a test with higher actual size has higher power, so the various testing procedures appear to have similar power once we control for size.

5.2 Individual-Level Data

For completeness we additionally consider regression using individual-level data. Recall that we are using g to denote the clustering unit and i to denote

[17] We thank Doug Staiger for suggesting this mechanism to us.

year, so we use n to denote individual. Then the model is

$$y_{nig} = \alpha_g + \gamma_i + \mathbf{x}'_{nig}\boldsymbol{\delta} + \beta_1 I_{ig} + u_{nig},$$

where the individual-level regressors $\mathbf{x}_{nig}$ are a quartic in age and three education dummies. The placebo law binary indicator I_{ig} is generated as before.

Table 6 reports the results of $R = 250$ simulations with $B = 199$ replications used for the bootstrap. These low numbers are chosen as the individual-level data set is very large.[18] We consider cases $G = 6$ and $G = 10$ as we are most interested in inference with few clusters.

The first row of Table 6 reports high rejection rates of 0.44 when we use the CRVE but erroneously cluster on state-year combinations. These are essentially the same as those given in BDM Table 2.

In the second row of Table 6 we see that using the CRVE and correctly clustering on state only considerably reduces the rejection rate. But at 0.15 for $G = 6$ and 0.10 for $G = 10$, the rejection rate is still much too high. (Because BDM only present corrections for clustering using aggregate data, we cannot directly compare these results to any of theirs).

The third row of Table 6 shows that the bootstrap-t procedure using wild cluster resampling (with clustering on the state) leads to rejection rates not statistically significantly different from 0.05. Because the individual level data are unbalanced we cannot use the residual cluster bootstrap. We did not pursue a pairs cluster bootstrap-t as it may encounter similar problems to those for aggregate data, albeit to a lesser extent.

In summary, using both aggregate and micro data, the wild cluster bootstrap-t leads to rejection rates of 0.05. The pairs cluster bootstrap-t works fine for $G \geq 20$, but for $G \leq 10$ can fail due to problems posed by the binary regressor.

6 Gruber and Poterba (1994) Application

Gruber and Poterba (1994, henceforth GP) examine the impact of tax incentives on the decision to purchase health insurance. They analyze differential changes for self-employed and business-employed in the after-tax price of

[18] The sample sizes vary from one replication to another due to drawing differently sized states. They average about 66,000 observations for $G = 6$ and 108,000 for $G = 10$.

health insurance due to the Tax Reform Act of 1986 (TRA86). The TRA86 extended the tax subsidy for health insurance to self-employed individuals; individuals employed by a business had a tax subsidy both before and after TRA86, and so can serve as a comparison group.

The dependent variable y is whether or not an employed person has private health insurance. Like GP we focus on individuals 25-54 years of age. The policy variable I is whether an employed person is eligible to receive tax subsidy for the insurance. This depends on the type of employer. Those employed by a business were eligible throughout 1982-89, while those self-employed were ineligible from 1982-86 and eligible from 1987-89.

While BDM include a full set of year effects that allows the intercept to vary by year, GP use a pre-post design that restricts the intercept to take the same value within the sets of pre-reform and post-reform years. Given the structure of the binary policy variable the model can be rewritten as

$$y_{ijt} = \alpha_1 + \alpha_2 \text{SELF}_{ijt} + \alpha_3 \text{POST}_{ijt} + \beta_1 \text{SELF}_{ijt} \times \text{POST}_{ijt} + u_{jt},$$

where i denotes individual, j denotes employer type, t denotes year, $\text{SELF}_{ijt} = 1$ if individual i is self-employed at time t, and $\text{POST}_{ijt} = 1$ if the year is 1987, 1988 or 1999.

We perform difference-in-difference analysis, controlling for potential clustering of errors of a form considered by Donald and Lang (2004). Like Donald and Lang, we ignore additional regressors (the bulk of Gruber and Poterba's paper examines subtle interactions between pre-tax income, employment status, and the TRA86). Unlike Donald and Lang we constructed our own data set from the March CPS, one that closely matches Table IV of GP, permitting analysis using both aggregated data and individual-level data.

6.1 Aggregated Employment-Year Data

In their preliminary analysis, GP report in Table IV average insurance rates by year and employer-type for March CPS data on eight years (five before the TRA86 and three after), leading to an aggregated data set with sixteen observations. Our simple difference-in-difference estimate is 0.055, with a standard error of 0.0044. This is more precise than and differs a little from the GP Table VI estimate of 0.067, with a standard error of 0.008, as GP compared only the two years 1985-86 with the two years 1988-89.

We now consider possible clustering. The analog of the type of clustering considered by BDM (2004) and in our Section 4 is to cluster by employer-type. But then there are just two clusters and no real chance to control for clustering. We instead follow Donald and Lang (2004), and treat years as clusters, so that there $G = 8$ clusters in our analysis. When we cluster on year, the cluster-robust standard error obtained using (8) is 0.0074, compared to 0.0044 without control for clustering. The regressor is highly statistically significant, with $\widehat{\beta}_1/s_{\widehat{\beta}_1} = 7.46$ and very low p-value.

To enable more meaningful analysis we test $H_0 : \beta_1 = 0.040$ against $H_0 : \beta_1 \neq 0.040$. Then $w = (0.055 - 0.040)/0.0074 = 2.02$ with p-value of 0.043 using standard normal critical values and 0.090 using the T distribution with $G - 2 = 6$ degrees of freedom.

If we instead bootstrap this Wald statistic with $B = 999$ replications, the pairs cluster bootstrap-t yields $p = 0.209$, the residual cluster bootstrap-t gives $p = 0.112$, and the wild cluster bootstrap-t gives a p-value of 0.070. We believe that the p-value for the pairs cluster bootstrap is implausibly large, for reasons discussed in the BDM replication, while the other two bootstraps lead to plausible p-values that, as expected, are larger than those obtained by using asymptotic normal critical values.

6.2 Individual-Level Data

We then apply our methods to the micro-level data that we extracted, with estimation of the binary outcome by a linear probability model. Then $\widehat{\beta}_1 = 0.055$ with cluster-robust standard error obtained using (8) of 0.0066, and $w = (0.055 - 0.040)/0.0066 = 2.27$. The p-value is 0.023 if we use standard normal critical values and 0.064 if we use the T(6) distribution.

Because the year-defined groups are not equally sized, we are unable to do the residual bootstrap with the micro data. For a bootstrap with $B = 999$ replications, the pairs cluster bootstrap-t yields $p = 0.212$ and the wild cluster bootstrap-t gives a p-value of 0.071.

These bootstrap results are essentially the same as those using aggregated data. The pairs cluster bootstrap-t runs into problems due to binary regressors that are invariant within cluster. Eight years are drawn with replacement and the model is inestimable if, for example, only data from the five initial years 1982-86 is drawn. The wild cluster bootstrap-t uses a different resampling method that will not encounter this problem, and appears

to do well.

7 Conclusion

Many microeconometrics studies use clustered data, with regression errors correlated within cluster and regressors correlated within cluster. Then it is essential that one control for clustering. A good starting point is to use Wald tests (or "t" tests) that use cluster-robust standard errors, provided the appropriate level of clustering is chosen. As made clear in section 2 of BDM (2004), too many studies fail to do even this much.

In this paper we are concerned with the additional complication of having few clusters. Then the use of appropriate cluster-robust standard errors still leads to nontrivial over-rejection by Wald tests. Our Monte Carlo simulations reveal that at the very least one should provide some small sample correction of standard errors, such as magnifying the residuals in (8) by a factor $\sqrt{G/(G-1)}$ and using a T distribution with G or fewer degrees of freedom (we arbitrarily used $G-2$ in Table 3).

The primary contribution of this paper is to use bootstrap procedures to obtain more accurate cluster-robust inference when there are few clusters. Our discussion and implementations of the bootstrap make it clear that there are many possible variations on a bootstrap. Section 3 presented a range of bootstrap procedures and resampling methods that may be used. And for replicable results it can be necessary to precisely define the way the bootstrap is implemented, as done in the Appendix.

The usual way that the bootstrap is used, to obtain an estimate of the standard error, does not lead to improved inference with few clusters as it does not provide an asymptotic refinement.

The BCA method is a method that in theory should do better, as it does provide an asymptotic refinement. But we find in our simulations that it provides very modest improvement.

We focus on the bootstrap-t procedure, the method most emphasized by theoretical econometricians and statisticians. This provides asymptotic refinement because it bootstraps the Wald statistic, rather than the parameter estimates, and the Wald statistic is asymptotically pivotal provided it is calculated using standard errors corrected for clustering. We find that the bootstrap-t procedure can lead to considerable improvement. And its

performance is little affected by what method is used to obtain standard errors (such as CRVE or CR3) provided the method is asymptotically valid and the same method is used in calculating the Wald statistic in the original sample and in the bootstrap resamples.

But these improvements depend on the resampling method used and on the discreteness of the data being resampled. The standard method for resampling that preserves the within-cluster features of the error is a pairs cluster bootstrap that resamples at the cluster level, so that if the g^{th} cluster is selected then all data (dependent and regressor variables) in that cluster appear in the resample. This bootstrap can lead to inestimable models or nearly inestimable models in some bootstrap pseudo-samples when there are few clusters and regressors take a very limited range of values. While not all applications will encounter this problem, it does arise when interest lies in a binary policy variable that is invariant or relatively invariant within cluster.

We find that an alternative cluster bootstrap, the wild cluster bootstrap does especially well. This bootstrap is a cluster generalization of the wild bootstrap for heteroskedastic models. Even when analysis is restricted to a wild cluster bootstrap, several different variations are possible. The variation we use is one that uses equal weights and probability, and uses residuals from OLS estimation that imposes the null hypothesis. This bootstrap works well in our own simulation exercise and when applied to the data of BDM (2004).

The BDM (2004) study is one of the highest profile papers highlighting the importance of cluster robust inference. One important conclusion of BDM (2004) is that for few (six) clusters the cluster-robust estimator performs poorly, and for moderate (ten and twenty) number of clusters their bootstrap based method also does poorly. We perform a re-analysis of their exercise, and come to much more optimistic conclusions. Using the wild cluster bootstrap-t method our empirical rejection rates are extremely close to the theoretical values, even with as few as six clusters, and there is no noticeable loss of power after accounting for size. Our results offer not only theoretical improvements, but practical ones as well. We hope researchers will take advantage of these improvements in the plentiful cases when clustering among a relatively small number of groups is a real concern.

A Appendix

Appendix A.1 presents a general discussion of the bootstrap and why it is asymptotically better to bootstrap an asymptotically pivotal statistic (bootstrap-t method). Appendix A.2 details the various bootstraps summarized in Table 1.

A.1 Asymptotic Refinement for Bootstrap-t

The theory draws heavily on Hall (1982) and Horowitz (2001). Cameron and Trivedi (2005) provide a more introductory discussion.

A.1.1 General Bootstrap Procedure

We use the generic notation $T_N = T_N(\mathcal{S}_N)$ to denote the statistic of interest, calculated on the basis of a sample $\mathcal{S}_N$ of size N.[19] We focus on inference for a single regression coefficient β_1 from multivariate OLS regression. Then leading examples are $T_N = \widehat{\beta}_1$, and $T_N = (\widehat{\beta}_1 - \beta_1^0)/s_{\widehat{\beta}_1}$, where we recall that β_1^0 is given by the null hypothesis.

We wish to approximate the finite sample cdf of T_N, $H_N(t) = \Pr[T_N \leq t]$. The bootstrap does this by obtaining B resamples of the original sample $\mathcal{S}_N$, using methods given in the subsequent subsection. The b^{th} resample is denoted $\mathcal{S}_{Nb}^*$ and is used to form a statistic $T_{Nb}^* = T_N^*(\mathcal{S}_{Nb}^*)$. The empirical distribution of T_{Nb}^*, $b = 1, .., B$, is used to estimate the distribution of T_N, so $\Pr[T_N \leq t]$ is estimated by the fraction of the realized values of $T_{N1}^*, ..., T_{Nb}^*$ that are less than t, denoted

$$\widehat{H}_N(t) = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_{Nb}^* \leq t). \tag{17}$$

This distribution can be used to compute moments such as variance, and also to compute test critical values and p-values.

General Bootstrap Procedure for a Statistic T_N

1. Do B iterations of this step. On the b^{th} iteration:

[19] In the cluster setting $N = \sum_{g=1}^{G} N_g$, and $N \to \infty$ because $G \to \infty$.

(a) Re-sample the data from $\mathcal{S}_N$ using one of the procedures presented in Appendix A.2. Call the resulting re-sample $\mathcal{S}^*_{Nb}$.

(b) Use the bootstrap re-sample form $T^*_{Nb} = T^*_N(\mathcal{S}^*_{Nb})$, where in some but not all cases $T^*_N(\cdot) = T_N(\cdot)$.

2. Conduct inference using $\widehat{H}_N(t) = B^{-1}\sum_{b=1}^{B} \mathbf{1}(T^*_{Nb} \leq t)$ to find critical values for the test statistic T_N based on the original data. See Appendix A.2 for further details.

The bootstrap-t method directly approximates the distribution of $T_N = (\widehat{\beta}_1 - \beta^0_1)/s_{\widehat{\beta}_1}$. If the bootstrap resampling method imposes H_0 then $T^*_{Nb} = (\widehat{\beta}^*_{1b} - \beta^0_1)/s_{\widehat{\beta}^*_{1b}}$, where $\widehat{\beta}^*_{1b}$ is the estimator of β_1 and $s_{\widehat{\beta}^*_{1b}}$ is the standard error from re-sample $\mathcal{S}^*_{Nb}$. Note that we center T^*_{Nb} on β^0_1 since the resampling dgp has $\beta_1 = \beta^0_1$. If instead the bootstrap resampling method does not impose H_0, the case necessarily for pairs cluster, then $T^*_{Nb} = (\widehat{\beta}^*_{1b} - \widehat{\beta}_1)/s_{\widehat{\beta}^*_{1b}}$. The centering is on $\widehat{\beta}_1$ and the bootstrap views the original sample as the population. That is, implicitly we impose $\beta_1 = \widehat{\beta}_1$, and the bootstrap resamples are viewed as B samples from the implied population.

By contrast the bootstrap-se, percentile and BCA methods bootstrap $T_N = \widehat{\beta}_1$. Then $T^*_{Nb} = \widehat{\beta}^*_{1b}$, where $\widehat{\beta}^*_{1b}$ is the estimator of β_1 from re-sample $\mathcal{S}^*_{Nb}$.

A.1.2 Asymptotic Refinement

For notational simplicity drop the subscript N, so $T_N(S_N) = T$ has small-sample cdf denoted $H(t|F) = \Pr[T \leq t|F]$ where F is the true cdf generating the underlying data in sample S_N. The distribution H usually is analytically intractable. The usual first-order asymptotic theory replaces it with the asymptotic distribution of the test-statistic. The bootstrap instead replaces H with $\widehat{H}(t|\widehat{F}) = \Pr[T^* \leq t|\widehat{F}]$ where $\widehat{F}$ denotes the cdf used to obtain bootstrap resamples. We are concerned with how good an estimate $\widehat{H}(t|\widehat{F})$ is of $H(t|F)$.

The bootstrap leads to consistent estimates and hypothesis tests under relatively weak assumptions. Because the bootstrap should be based on a distribution $\widehat{F}$ that is consistent for F, one must take care to choose the resampling method so as to mimic the properties of F. For consistency,

the bootstrap requires smoothness and continuity in F and in $\widehat{H}$. These assumptions are satisfied for our application for the OLS estimator with clustered errors.

A consistent bootstrap need not have asymptotic refinement, however. A key requirement is that we work with an asymptotically pivotal statistic, as now explained.

To begin with assume that T is standardized to have mean 0 and variance 1. The usual asymptotic approximation $T \overset{a}{\sim} \mathcal{N}[0,1]$ is

$$\Pr[T \leq t|F] = \Phi(t) + O(N^{-1/2}),$$

where $\Phi(\cdot)$ is the standard normal cdf and N is sample size. When one uses the standard normal critical values with a "t" statistic, this is the approximation on which one relies. The Edgeworth expansion gives a better asymptotic approximation

$$\Pr[T \leq t|F] = \Phi(t) + N^{-1/2}a(t)\phi(t) + O(N^{-1}),$$

where $\phi(\cdot)$ is the standard normal density and $a(\cdot)$ is an even quadratic polynomial with coefficients that depend on the low-order cumulants (or moments) of the underlying data. One can directly use the preceding result, but computation of the polynomial coefficients in $a(t)$ is theoretically demanding. The bootstrap provides an alternative.

The bootstrap version of T is the statistic T^*, which has Edgeworth expansion

$$\Pr[T^* \leq t|\widehat{F}] = \Phi(t) + N^{-1/2}\widehat{a}(t)\phi(t) + O_p(N^{-1}),$$

where $\widehat{F}$ is the empirical distribution function of the sample. If $\widehat{a}(t) = a(t) + O_p(N^{-1/2})$, which is often the case, then

$$\Pr[T \leq t|F] = \Pr[T^* \leq t|\widehat{F}] + O_p(N^{-1}). \tag{18}$$

This statement means that the bootstrap cdf $\Pr[T^* \leq t|\widehat{F}]$ is within $O_p(N^{-1})$ of the unknown true cdf $\Pr[T \leq t|F]$, which is a better approximation than one gets using $\Phi(t)$, since the standard normal cdf is within $O(N^{-1/2})$ and $\Pr[O(N^{-1/2}) - O_p(N^{-1}) > 0]$ gets arbitrarily close to 1 for sufficiently large N.

What if we use a nonpivotal statistic T? Suppose $T \overset{a}{\sim} \mathcal{N}[0, \sigma^2]$ so that $T/s \overset{a}{\sim} \mathcal{N}[0,1]$ where s is a consistent estimate of the standard error. Then Edgeworth expansions still apply, but now

$$\Pr[T \leq t|F] = \Phi(t/\sigma) + N^{-1/2} b(t/\sigma)\phi(t/\sigma) + O(N^{-1}),$$

for some quadratic function $b(\cdot) \neq a(\cdot)$, and similarly for the bootstrap estimates

$$\Pr[T^* \leq t|\widehat{F}] = \Phi(t/s) + N^{-1/2}\widehat{b}(t/s)\phi(t/s) + O_p(N^{-1}).$$

Now, even if $\widehat{b}(\cdot) = b(\cdot) + O_p(N^{-1/2})$, these functions are evaluated at t/s where usually $s = \sigma + O_p(N^{-1/2})$. It follows for nonpivotal T that

$$\Pr[T \leq t|F] = \Pr[T^* \leq t|\widehat{F}] + O_p(N^{-1/2}), \tag{19}$$

so there is no asymptotic refinement. Thus nonpivotal statistics bring no improvement in the convergence rate relative to using first-order asymptotic theory.

The main requirement for the asymptotic refinement (18) is that an asymptotically pivotal statistic is the object being bootstrapped. The bootstrap-t procedure does this.

The preceding analysis shows that for tests of nominal size α the true size is $\alpha + O(N^{-j/2})$ where $j = 2$ using the bootstrap-t procedure, while $j = 1$ using the usual asymptotic normal approximation and the percentile and bootstrap-se procedures. These results for are for a one-sided test or a nonsymmetric two-sided test. For a two-sided symmetric test, cancellation occurs because $a(t)$ is an even function, so one further term in the Edgeworth expansion can be used. Then $j = 3$ using the bootstrap-t procedure and $j = 2$ using the other procedures.

A.2 Bootstrap Procedures

A.2.1 Bootstrap-T Procedures

We begin with the preferred bootstrap-t procedures using three bootstrap sampling schemes - pairs cluster, residual cluster and wild cluster - that are generalizations of pairs, clusters and wild resampling for nonclustered data.

Pairs Cluster Bootstrap-t

1. From the original sample form $w = (\widehat{\beta}_1 - \beta_0)/s_{\widehat{\beta}_1}$, where $s_{\widehat{\beta}_1}$ is obtained using the CRVE in (8) with $\widetilde{\mathbf{u}}_g = (G/(G-1))\widehat{\mathbf{u}}_g$.

2. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters.

   (b) Calculate the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \widehat{\beta}_1)/s_{\widehat{\beta}_{1,b}^*}$, where $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ are obtained from OLS estimation using the b^{th} pseudo-sample, $s_{\widehat{\beta}_{1,b}^*}$ is obtained using the same method as that in step 1, and $\widehat{\beta}_1$ is the original OLS estimate.

3. Reject H_0 at level α if and only if $w < w_{[\alpha/2]}^*$ or $w > w_{[1-\alpha/2]}^*$, where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

This procedure is a straightforward generalization of the pairs bootstrap that assumes $(y_{ig}, \mathbf{x}_{ig})$ are iid across g and within i for given g. Here we instead assume only that $(\mathbf{y}_g, \mathbf{X}_g)$ are iid.

We consider two variations of this procedure that use alternative estimators of $s_{\widehat{\beta}}$ in both step 1 and in step 2b. First, the pairs cluster CR3 bootstrap-T uses the CRVE in (8) with $\widetilde{\mathbf{u}}_g$ calculated using the CR3 correction in (10). Second, the pairs cluster BDM bootstrap-T uses default OLS standard errors and is a symmetric version of the Wald test, following BDM (2004).

The remaining bootstrap-t procedures use residual cluster and wild cluster resampling schemes that take advantage of the ability to resample with the null hypothesis $\beta_1 = \beta_1^0$ imposed.

Cluster Residual Bootstrap-t with H_0 imposed

1. From OLS estimation on the original sample form $w = (\widehat{\beta}_1 - \beta_0)/s_{\widehat{\beta}_1}$, where $s_{\widehat{\beta}_1}$ is obtained using the CRVE in (8) with $\widetilde{\mathbf{u}}_g = (G/(G-1))\widehat{\mathbf{u}}_g$. Also obtain the restricted OLS estimator $\widehat{\boldsymbol{\beta}}^R$ that imposes $H_0 : \beta_1 = \beta_1^0$, and the associated restricted OLS residuals $\{\widehat{\mathbf{u}}_1^R, ..., \widehat{\mathbf{u}}_G^R\}$.[20]

[20] The restricted estimator can be obtained by regressing $y_{ig} - \beta_1^0 x_{1,ig}$ on a constant and all regressors other than $x_{1,ig}$.

2. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$ by resampling with replacement G times from $\{\widehat{\mathbf{u}}_1^R, ..., \widehat{\mathbf{u}}_G^R\}$ to give $\{\widehat{\mathbf{u}}_1^{R*}, ..., \widehat{\mathbf{u}}_G^{R*}\}$ and then forming $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}}^R + \widehat{\mathbf{u}}_g^{R*}$, $g = 1, ..., G$.

   (b) Calculate the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \beta_1^0)/s_{\widehat{\beta}_{1,b}^*}$, where $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ are obtained from unrestricted OLS estimation using the b^{th} pseudo-sample, with $s_{\widehat{\beta}_{1,b}^*}$ computed using the same method as that in step 1.

3. Reject H_0 at level α if and only if $w < w_{[\alpha/2]}^*$ or $w > w_{[1-\alpha/2]}^*$, where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

Hall (1992, pp.184-191) provides theoretical justification for the residual bootstrap for clustered errors. This bootstrap is used as a benchmark in Monte Carlo simulations for the other bootstraps. In practice it is too restrictive as it assumes that $\mathbf{u}_g$ are iid, ruling out heteroskedasticity across clusters, and that clusters are balanced.

Wild Cluster bootstrap with H_0 imposed

1. From OLS estimation on the original sample form $w = (\widehat{\beta}_1 - \beta_0)/s_{\widehat{\beta}}$, where $s_{\widehat{\beta}}$ is obtained using the CRVE in (8) with $\widetilde{\mathbf{u}}_g = (G/(G-1))\widehat{\mathbf{u}}_g$. Also obtain the restricted OLS estimator $\widehat{\boldsymbol{\beta}}^R$ that imposes $H_0 : \beta_1 = \beta_1^0$, and the associated restricted OLS residuals $\{\widehat{\mathbf{u}}_1^R, ..., \widehat{\mathbf{u}}_G^R\}$.[21]

2. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\widehat{\mathbf{y}}_1^*, \mathbf{X}_1), ..., (\widehat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$ by the following method. For each cluster $g = 1, ..., G$, form either $\widehat{\mathbf{u}}_g^{R*} = \widehat{\mathbf{u}}_g^R$ with probability 0.5 or $\widehat{\mathbf{u}}_g^{R*} = -\widehat{\mathbf{u}}_g^R$ with probability 0.5 and then form $\widehat{\mathbf{y}}_g^* = \mathbf{X}_g'\widehat{\boldsymbol{\beta}}^R + \widehat{\mathbf{u}}_g^{R*}$, $g = 1, ..., G$.

[21] The restricted estimator can be obtained by regressing $y_{ig} - \beta_1^0 x_{1,ig}$ on a constant and all regressors other than $x_{1,ig}$.

(b) Calculate the Wald test statistic $w_b^* = (\widehat{\beta}_{1,b}^* - \beta_1^0)/s_{\widehat{\beta}_{1,b}^*}$, where $\widehat{\beta}_{1,b}^*$ and its standard error $s_{\widehat{\beta}_{1,b}^*}$ are obtained from unrestricted OLS estimation using the b^{th} pseudo-sample, with $s_{\widehat{\beta}_{1,b}^*}$ computed using the same method as that in step 1.

3. Reject H_0 at level α if and only if $w < w_{[\alpha/2]}^*$ or $w > w_{[1-\alpha/2]}^*$, where $w_{[q]}^*$ denotes the q^{th} quantile of $w_1^*, ..., w_B^*$.

The wild bootstrap has the benefit that it imposes the constraint that the bootstrap error $\widehat{\mathbf{u}}_g^*$ has conditional mean zero, a constraint not imposed by a pairs bootstrap. Specifically, consider the more general bootstrap residual

$$\widehat{\mathbf{u}}_g^* = a_g \widehat{\mathbf{u}}_g, \tag{20}$$

where a_g are drawings independent over g (and not dependent on the data) that satisfy $\mathrm{E}[a_g] = 0$ and $\mathrm{E}[a_g^2] = 1$. Then $\mathrm{E}[\widehat{\mathbf{u}}_g^*] = \mathrm{E}[a_g \widehat{\mathbf{u}}_g] = \mathrm{E}[\widehat{\mathbf{u}}_g \mathrm{E}[a_g | \widehat{\mathbf{u}}_g]] = 0$. Additionally $\mathrm{E}[\widehat{\mathbf{u}}_g^* \widehat{\mathbf{u}}_g^{*\prime}] = \mathrm{E}[\widehat{\mathbf{u}}_g \widehat{\mathbf{u}}_g']$ so the variance is preserved.

A variety of weights a_g have been proposed. The ones we use, with $a_g = 1$ with probability 0.5 and $a_g = -1$ with probability 0.5 are called Rademacher weights. Mammen (1993) actually proposed an alternative set of weights: $a_g = (1 - \sqrt{5})/2 \simeq -0.6180$ with probability $(1 + \sqrt{5})/2\sqrt{5} \simeq 0.7236$ and $a_g = 1 - (1 - \sqrt{5})/2$ with probability $1 - (1 + \sqrt{5})/2\sqrt{5}$. These weights are the only two-point distribution that satisfy the constraints $\mathrm{E}[a_g] = 0$ and $\mathrm{E}[a_g^2] = 1$ and the additional constraint $\mathrm{E}[a_g^3] = 1$, which is necessary to achieve asymptotic refinement if $\widehat{\boldsymbol{\beta}}$ is asymmetrically distributed.

Our implementation of the Wild bootstrap follows results of Davidson and Flachaire (2001) for the nonclustered case. They provide theoretical reasons and Monte Carlo evidence that favor using the simpler Rademacher variables as weights, even if $\widehat{\boldsymbol{\beta}}$ is asymmetrically distributed. And they favor using restricted OLS residuals.

We do not consider yet another residual bootstrap, a parametric bootstrap that makes draws of errors from an assumed error distribution.

A.2.2 BCA Bootstrap

We consider the BCA method for pairs cluster resampling only, though it can be applied to other resampling methods (that do not impose H_0).

Pairs Cluster Bias-corrected Accelerated (BCA) Bootstrap

1. From the original sample form $\widehat{\beta}_1$.

2. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample.

   (b) Calculate the OLS estimate $\widehat{\beta}_{1b}^*$.

3. Define the q-quantile $\widehat{\beta}_{1,[q]}^*$ to be the smallest of the B estimates $\widehat{\beta}_{1b}^*$ such that at least a fraction q of the B estimates $\widehat{\beta}_{1b}^*$ are less than $\widehat{\beta}_{1,[q]}^*$. Reject H_0 if and only if $\beta_1^0 < \widehat{\beta}_{1,[\alpha_1]}^*$ or $\beta_1^0 > \widehat{\beta}_{1,[\alpha_2]}^*$ where

$$\begin{aligned} \alpha_1 &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 - z_{\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{\alpha/2})}\right) \\ \alpha_2 &= \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{\alpha/2}}{1 - \widehat{a}(\widehat{z}_0 + z_{\alpha/2})}\right), \end{aligned}$$

   where $\Phi(\cdot)$ is the standard normal cdf, $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal, $\widehat{z}_0 = \Phi^{-1}(B^{-1}\sum_{b=1}^{B} \mathbf{1}(\widehat{\beta}_{1b}^* < \widehat{\beta}_1))$, and for clustered bootstraps

$$a = \frac{\sum_{g=1}^{G}(\overline{\beta} - \widehat{\beta}_{1,(g)})^3}{[\sum_{g=1}^{G}(\overline{\beta} - \widehat{\beta}_{1,(g)})^2]^{3/2}},$$

   where $\widehat{\beta}_{1,(g)}$ is the estimator of β_1 obtained from the original sample if the g^{th} cluster is excluded and $\overline{\beta} = \sum_{g=1}^{G} \widehat{\beta}_{1,(g)}/G$.

Note that this testing procedure does not require estimation of the standard error of the estimator, either using the original sample or any bootstrap resample, though the computation of a does use jackknife estimates of the variance and third moment of $\widehat{\beta}_1$. The BCA method, unlike bootstrap-t, is transformation respecting so that, for example, if it yields confidence interval (c_1, c_2) for $\widehat{\beta}_1$, then it yields confidence interval (c_1^2, c_2^2) for $\widehat{\beta}_1^2$. Efron (1987) and Hall (1992, pp. 128-141) provide details and discussion in the nonclustered case.

A.2.3 Bootstrap-se Methods

We present the bootstrap-se for pairs cluster resampling.

Pairs Cluster Bootstrap-se

1. From the original sample form $w = (\widehat{\beta}_1 - \beta_0)/s_{\widehat{\beta}}$, where $s_{\widehat{\beta}}$ is obtained using the CRVE in (8) with $\widetilde{\mathbf{u}}_g = (G/(G-1))\widehat{\mathbf{u}}_g$.

2. Do B iterations of this step. On the b^{th} iteration:

   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), ..., (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample.

   (b) Calculate the OLS estimate $\widehat{\beta}_{1,b}^*$.

3. Reject H_0 at level α if and only if $|w_{\text{BSE}}| > z_{\alpha/2}$, where

$$w_{\text{BSE}} = \frac{\widehat{\beta}_1 - \beta_1^0}{s_{\widehat{\beta}_1,\text{B}}},$$

   $s_{\widehat{\beta}_1,\text{B}}$ is the bootstrap estimate of the standard error

$$s_{\widehat{\beta}_1,\text{B}} = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\beta}_{1b}^* - \overline{\widehat{\beta}_1^*})^2 \right)^{1/2},$$

   and $\overline{\widehat{\beta}_1^*} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\beta}_{1b}^*$.

This method is easily adapted to the other resampling schemes by appropriately amending steps 1 and 2(a).

B References

Angrist and Lavy (2002), "The Effect of High School Matriculation Awards: Evidence from Randomized Trials", NBER Working Paper Number 9389.

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Group Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431-434.

Baker, M. and N.M. Fortin (2001), "Occupational Gender Composition and Wages in Canada, 1987-1988," *Canadian Journal of Economics*, 343-376.

Bell, R.M. and D.F. McCaffrey (2002), "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, 169-179.

Bertrand, M., E. Duflo and S. Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, 119, 249-275.

Brownstone, D. and Valletta (2001), "The Bootstrap and Multiple Imputations: Harnessing Increased Computing Power for Improved Statistical Tests, *Journal of Economic Perspectives*, 15(4), 129-141.

Cameron, A.C. and P.K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge: Cambridge University Press.

Chesher, A. and G. Austin (1991), "The Finite-Sample Distributions of Heteroskedasticity Robust Wald Statistics," *Journal of Econometrics*, Vol. 47, 153-173.

Chesher, A. and I. Jewitt (1987), "The Bias of A Heteroskedasticity Consistent Covariance Matrix Estimator," *Econometrica*, Vol. 55, 1217-1222.

Davidson, R. and E. Flachaire (2001), "The Wild Bootstrap, Tamed at Last," unpublished manuscript.

Davidson, R. and J.G. MacKinnon (1999), "The Size Distortion of the Bootstrap," *Econometric Theory*, 15, 361-376.

Davidson, R. and J.G. MacKinnon (2004), *Econometric Theory and Methods*, Oxford, Oxford University Press.

Davison, A.C. and D.V. Hinkley (1997), *Bootstrap Methods and their Application*, New York, Cambridge University Press.

Donald, S. G. and Lang, K. (2004), "Inference with Differences in Differences and Other Panel Data", unpublished manuscript.

Efron, B. (1979), "Bootstrapping Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1-26.

Efron, B. (1981), "Nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.

Efron, B. (1987), "Better Bootstrap Confidence Intervals (with Discussion)," *Journal of the American Statistical Association*, 82, 171-200.

Efron, B., and J. Tibsharani (1993), *An Introduction to the Bootstrap*, London, Chapman and Hall.

Flynn, T.N., and Peters, T.J. (2004), "Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results," *BMC Health Services Research*, 4:33.

Greenwald, B.C. (1983), "A General Analysis of Bias in the Estimated Standard Errors of Least Squares Coefficients," *Journal of Econometrics*, Vol. 22, 323-338.

Gruber, J. and J. Poterba (1994), "Tax Incentives and the Decision to Purchase Health Insurance: Evidence from the Self-Employed," *Quarterly Journal of Economics*, 109, 701-733.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Horowitz, J.L. (1997), "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in *Advances in Economics and Econometrics: Theory and Applications*, *Seventh World Congress*, D.M. Kreps and K.F.Wallis (Eds.), Volume 3, 188-222, Cambridge, UK, Cambridge University Press.

Horowitz, J.L. (2001), "The Bootstrap," in *Handbook of Econometrics*, Volume 5, J.J. Heckman and E. Leamer (Eds.), 3159-3228, Amsterdam, North-Holland.

Kauermann, G. and R.J. Carroll (2001), "A Note on the Efficiency of Sandwich Covariance Matrix Estimation," *Journal of the American Statistical Association*, 96, 1387-1396.

Kézdi, G. (2004), "Robust Standard Error Estimation in Fixed-Effects Models," Robust Standard Error Estimation in Fixed-Effects Panel Models," Hungarian Statistical Review, Special Number 9, 95-116.

Kish, L. and Frankel (1974), "Inference from Complex Surveys with Discussion", *Journal of the Royal Statistical Society*, Series B, 36, 1-37.

Kloek, T. (1981), "OLS Estimation in a Model where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated," *Econometrica*, 49, 205-07.

Liang, K.-Y., and S.L. Zeger (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

Liu, R.Y. (1988), "Bootstrap Procedures under Some Non-iid Models," *Annals of Statistics*, 16, 1696-1708.

MacKinnon, J.G. (2002), "Bootstrap Inference in Econometrics," *Canadian Journal of Economics*, 35, 615-645.

MacKinnon, J.G., and H. White (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

MacPherson, D.A. and Barry T. Hirsch (1995), "Wages and Gender Occupation: Why Do Women's Jobs Pay Less," *Journal of Labor Economics*, 426-471.

Mancl, L.A. and T.A. DeRouen, "A Covariance Estimator for GEE with Improved Finite- Sample Properties," *Biometrics*, 57, 126-134.

Mammen, E. (1993), "Bootstrap and Wild Bootstrap for High Dimensional Linear Models," *Annals of Statistics*, 21, 255-285.

McCaffrey, D.F., Bell, R.M., and C.H. Botts (2001), "Generalizations of bias Reduced Linearization," Proceedings of the Survey Research Methods Section, American Statistical Association.

Moulton, B.R. (1986), "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics*, 32, 385-397.

Moulton, B.R. (1990), "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72, 334-38.

Pan, W. and M. Wall (2002), "Small-sample Adjustments in Using the Sandwich Variance Estimator in Generalized Estimating Equation," *Statistics in Medicine*, 21, 1429-1441.

Rothenberg, T. (1988), "Approximate Power Functions for Some Robust Tests of Regression Coefficients," *Econometrica*, Vol. 56, 997-1019.

Satterthwaite, F.F. (1941), "Synthesis of Variance," *Psychometrika*, 6, 309-316.

Scott, A.J., and D. Holt (1982), "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods," *Journal of the American Statistical Association*, 77, 848-854.

Sherman, M. and S. le Cressie (1997), "A Comparison Between Bootstrap Methods and Generalized Estimating Equations for Correlated Outcomes in Generalized Linear Models," *Communications in Statistics*,

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, H. (1984), *Asymptotic Theory for Econometricians*, San Diego: Academic Press.

Wooldridge, J.M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, 133-138.

Wu, C.F.G. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261-1295.

Table 1: Different Methods for Wald Test

Method	Bootstrap?	Refinement?	H_0 imposed?
Conventional Wald			
1. Default (iid errors)	No	-	-
2. Moulton type	No	-	-
3. Cluster-robust	No	-	-
4. Cluster-robust CR3	No	-	-
Wald bootstrap-se			
5. Pairs cluster	Yes	No	-
6. Residuals cluster H_0	Yes	No	-
7. Wild cluster H_0	Yes	No	-
BCA test			
8. Pairs cluster	Yes	Yes	
Wald bootstrap-t			
9. BDM	Yes	No	No
10. Pairs cluster	Yes	Yes	No
11. Pairs CR3 cluster	Yes	Yes	No
12. Residuals cluster H_0	Yes	Yes	Yes
13. Wild cluster H_0	Yes	Yes	Yes

Table 2: 1,000 simulations from dgp with group level random errors.
Rejection rates for tests of nominal size 0.05 with simulation se's in parentheses.

Estimator		Number of Groups (G)					
#	Method	5	10	15	20	25	30
1	Assume iid	0.426	0.479	0.489	0.490	0.504	0.472
		(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
2	Moulton-type estimator	0.130	0.084	0.086	0.074	0.080	0.052
		(0.011)	(0.009)	(0.009)	(0.008)	(0.009)	(0.007)
3	Cluster Robust	0.195	0.132	0.096	0.093	0.095	0.069
		(0.013)	(0.011)	(0.009)	(0.009)	(0.009)	(0.008)
4	CR3 Residual Correction	0.088	0.084	0.065	0.072	0.067	0.057
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
5	Pairs Cluster Bootstrap - se	0.152	0.122	0.095	0.096	0.100	0.072
		(0.011)	(0.010)	(0.009)	(0.009)	(0.009)	(0.008)
6	Residual Cluster Bootstrap - se	0.047	0.049	0.063	0.062	0.066	0.043
		(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.006)
7	Wild Cluster Bootstrap - se	0.012	0.031	0.039	0.041	0.056	0.040
		(0.003)	(0.005)	(0.006)	(0.006)	(0.007)	(0.006)
8	Pairs Cluster Bootstrap - BCA	0.161	0.106	0.101	0.087	0.094	0.068
		(0.012)	(0.010)	(0.010)	(0.009)	(0.009)	(0.008)
9	BDM Bootstrap - t	0.117	0.109	0.094	0.094	0.095	0.068
		(0.010)	(0.010)	(0.009)	(0.009)	(0.009)	(0.008)
10	Pairs Cluster Bootstrap - t	0.081	0.082	0.075	0.073	0.070	0.054
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
11	Pairs CR3 Bootstrap - t	0.081	0.085	0.070	0.072	0.069	0.051
		(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.007)
12	Residual Cluster Bootstrap - t	0.034	0.052	0.049	0.044	0.056	0.050
		(0.006)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
13	Wild Cluster Bootstrap - t	0.054	0.062	0.056	0.045	0.060	0.045
		(0.007)	(0.008)	(0.007)	(0.007)	(0.008)	(0.007)
	T_distribution(G-k)	0.145	0.086	0.072	0.066	0.062	0.060

Table 3: 1,000 simulations from dgp with group level random errors and heteroskedasticity. Rejection rates for tests of nominal size 0.05 with simulation se's in parentheses.

Estimator		Number of Groups (G)					
#	Method	5	10	15	20	25	30
1	Assume iid	0.302	0.288	0.307	0.295	0.287	0.297
		(0.015)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
2	Moulton-type estimator	0.261	0.214	0.206	0.175	0.174	0.180
		(0.014)	(0.013)	(0.013)	(0.012)	(0.012)	(0.012)
3	Cluster Robust	0.208	0.118	0.110	0.081	0.072	0.068
		(0.013)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
4	CR3 Residual Correction	0.138	0.092	0.086	0.070	0.062	0.062
		(0.011)	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)
5	Pairs Cluster Bootstrap - se	0.174	0.111	0.109	0.085	0.074	0.070
		(0.012)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
6	Residual Cluster Bootstrap - se	0.181	0.169	0.183	0.157	0.149	0.163
		(0.012)	(0.012)	(0.012)	(0.012)	(0.011)	(0.012)
7	Wild Cluster Bootstrap - se	0.019	0.041	0.057	0.040	0.038	0.043
		(0.004)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
8	Pairs Cluster Bootstrap - BCA	0.183	0.103	0.099	0.082	0.070	0.064
		(0.012)	(0.010)	(0.009)	(0.009)	(0.008)	(0.008)
9	BDM Bootstrap - t	0.181	0.108	0.110	0.090	0.070	0.068
		(0.012)	(0.010)	(0.010)	(0.009)	(0.008)	(0.008)
10	Pairs Cluster Bootstrap - t	0.079	0.067	0.074	0.058	0.054	0.053
		(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
11	Pairs CR3 Bootstrap - t	0.064	0.062	0.072	0.057	0.050	0.048
		(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
12	Residual Cluster Bootstrap - t	0.066	0.057	0.066	0.049	0.043	0.047
		(0.008)	(0.007)	(0.008)	(0.007)	(0.006)	(0.007)
13	Wild Cluster Bootstrap - t	0.053	0.056	0.058	0.048	0.041	0.044
		(0.007)	(0.007)	(0.007)	(0.007)	(0.006)	(0.006)
	T_distribution(G-k)	0.145	0.086	0.072	0.066	0.062	0.060

Table 4: 1,000 simulations from different dgps (see text) and G = 10 groups.
Rejection rates for tests of nominal size 0.05 with simulation se's in parentheses.

Estimator #	Method	Column number	Main - from Table 2 *1*	Reject based on T (8 dof) *2*	Cluster size = 2 *3*	Cluster size = 10 *4*	Cluster size = 100 *5*	4 RHS variables *6*	X's are constant within group *7*	X's are iid *8*	Unbalanced group sizes (10,50) *9*
1	Assume iid		0.491		0.106	0.268	0.679	0.687	0.770	0.054	0.524
			(0.016)		(0.010)	(0.014)	(0.015)	(0.015)	(0.013)	(0.007)	(0.016)
2	Moulton-type estimator		0.092	0.044	0.095	0.098	0.088	0.089	0.125	0.061	0.129
			(0.009)	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)	(0.008)	(0.011)
3	Cluster Robust		0.129	0.082	0.137	0.126	0.115	0.129	0.183	0.103	0.183
			(0.010)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.013)	(0.010)	(0.012)
4	CR3 Residual Correction		0.090	0.054	0.094	0.086	0.077	0.080	0.090	0.086	0.091
			(0.009)	(0.007)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
5	Pairs Cluster Bootstrap - se		0.120	0.071	0.100	0.114	0.120	0.128	0.063	0.122	0.138
			(0.010)	(0.008)	(0.009)	(0.010)	(0.010)	(0.010)	(0.008)	(0.010)	(0.011)
6	Residual Cluster Bootstrap - se		0.058	0.013	0.069	0.068	0.060	0.057	0.054	0.080	
			(0.007)	(0.004)	(0.008)	(0.008)	(0.008)	(0.007)	(0.007)	(0.009)	
7	Wild Cluster Bootstrap - se		0.028	0.006	0.048	0.044	0.032	0.030	0.036	0.053	0.019
			(0.005)	(0.002)	(0.007)	(0.006)	(0.006)	(0.005)	(0.006)	(0.007)	(0.004)
8	Pairs Cluster Bootstrap - BCA		0.111		0.125	0.112	0.109	0.112	0.100	0.134	0.140
			(0.010)		(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.011)	(0.011)
9	BDM Bootstrap - t		0.119		0.086	0.115	0.112	0.119	0.121	0.097	0.128
			(0.010)		(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)	(0.011)
10	Pairs Cluster Bootstrap - t		0.096		0.085	0.083	0.086	0.090	0.066	0.079	0.120
			(0.009)		(0.009)	(0.009)	(0.009)	(0.009)	(0.008)	(0.009)	(0.010)
11	Pairs CR3 Bootstrap - t		0.090		0.075	0.077	0.081	0.084	0.050	0.082	0.110
			(0.009)		(0.008)	(0.008)	(0.009)	(0.009)	(0.007)	(0.009)	(0.010)
12	Residual Cluster Bootstrap - t		0.055		0.052	0.056	0.050	0.043	0.043	0.065	
			(0.007)		(0.007)	(0.007)	(0.007)	(0.006)	(0.006)	(0.008)	
13	Wild Cluster Bootstrap - t		0.055		0.064	0.056	0.048	0.052	0.045	0.064	0.061
			(0.007)		(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
	T_distribution(8)		0.086								

Table 5: 1,000 simulations from BDM (2004) design.
Rejection rates for tests of nominal size 0.05 with simulation se's in parentheses.
Size column measures size and power column measures power.

Estimator		Number of States (G) 6 Size	10 Size	20 Size	50 Size	6 Power	10 Power	20 Power	50 Power
#	Method								
1	Assume iid	0.459	0.438	0.461	0.439	0.515	0.506	0.574	0.692
		(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)
2	Moulton-type estimator	0.449	0.428	0.454	0.429	0.510	0.490	0.565	0.686
		(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)
3	Cluster Robust	0.109	0.088	0.049	0.048	0.165	0.110	0.142	0.254
		(0.010)	(0.009)	(0.007)	(0.007)	(0.012)	(0.010)	(0.011)	(0.014)
4	Pairs Cluster Bootstrap - se	0.001	0.087	0.060	0.058	0.001	0.103	0.161	0.275
		(0.001)	(0.009)	(0.008)	(0.007)	(0.001)	(0.010)	(0.012)	(0.014)
5	Residual Cluster Bootstrap - se	0.043	0.055	0.045	0.048	0.079	0.069	0.127	0.260
		(0.006)	(0.007)	(0.007)	(0.007)	(0.009)	(0.008)	(0.011)	(0.014)
6	Wild Cluster Bootstrap - se	0.043	0.056	0.046	0.047	0.076	0.075	0.134	0.262
		(0.006)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.011)	(0.014)
7	Pairs Cluster Bootstrap - BCA	0.087	0.111	0.067	0.061	0.147	0.134	0.166	0.276
		(0.009)	(0.010)	(0.008)	(0.008)	(0.011)	(0.011)	(0.012)	(0.014)
8	BDM Bootstrap - t	0.111	0.086	0.053	0.054	0.161	0.113	0.153	0.270
		(0.010)	(0.009)	(0.007)	(0.007)	(0.012)	(0.010)	(0.011)	(0.014)
9	Pairs Cluster Bootstrap - t	0.006	0.022	0.043	0.061	0.007	0.033	0.112	0.255
		(0.002)	(0.005)	(0.006)	(0.008)	(0.003)	(0.006)	(0.010)	(0.014)
10	Residual Cluster Bootstrap - t	0.046	0.051	0.039	0.044	0.081	0.065	0.118	0.256
		(0.007)	(0.007)	(0.006)	(0.006)	(0.009)	(0.008)	(0.010)	(0.014)
11	Wild Cluster Bootstrap - t	0.067	0.053	0.041	0.045	0.110	0.078	0.124	0.247
		(0.008)	(0.007)	(0.006)	(0.007)	(0.010)	(0.008)	(0.010)	(0.014)

Table 6: 250 simulations from BDM (2004) design using micro data.
Rejection rates for tests of nominal size 0.05 with simulation se's in parentheses.

Estimator		Number of States (G) 6	10
#	Method	Size	Size
1	Pairs cluster on state-year	0.440	0.444
		(0.031)	(0.031)
3	Pairs cluster on state	0.148	0.100
		(0.023)	(0.019)
11	Cluster on state, Wild Bootstrap-t	0.080	0.048
		(0.017)	(0.014)

Note: Micro regressions control for a quartic in age, three education dummies, and state and year fixed effects. Number of Monte Carlo replications R=250. Nummber of boostrap replications B=199.