BOOTSTRAP-BASED IMPROVEMENTS FOR INFERENCE WITH CLUSTERED ERRORS

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

Abstract—Researchers have increasingly realized the need to account for within-group dependence in estimating standard errors of regression parameter estimates. The usual solution is to calculate cluster-robust standard errors that permit heteroskedasticity and within-cluster error correlation, but presume that the number of clusters is large. Standard asymptotic tests can over-reject, however, with few (five to thirty) clusters. We investigate inference using cluster bootstrap-t procedures that provide asymptotic refinement. These procedures are evaluated using Monte Carlos, including the example of Bertrand, Duflo, and Mullainathan (2004). Rejection rates of 10% using standard methods can be reduced to the nominal size of 5% using our methods.

I. Introduction

MICROECONOMETRICS researchers have increasingly realized the essential need to account for any within-group dependence in estimating standard errors of regression parameter estimates. In many 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. Kézdi (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 variation 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). Angrist and Lavy (2002) in an applied study find that bias adjustment of cluster-robust standard errors can make quite a difference. 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 (2007) 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. Ibragimov and Muller (2007) propose an alternate approach based on separate estimation within each group. They separate the data into independent groups, estimate the model within each group, average the separate estimates, and divide by the sample standard deviation of these estimates, and then compare against critical values from a T distribution. This approach holds promise for settings with few groups and where model identification and a central limit theorem holds within each group. Our proposed method does not require the latter two conditions, can be used to test multiple hypotheses, and is based on the parameter estimator commonly used in practice.

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.

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 many empirical

Received for publication June 15, 2006. Revision accepted for publication May 4, 2007.

* Department of Economics, University of California–Davis; Department of Economics, University of Arizona; and Department of Economics, University of California–Davis, respectively.

We thank an anonymous referee and seminar participants at the Australian National University, Dartmouth College, Florida State University, Indiana University, U.C. Berkeley, and U.C. Riverside for useful comments. Miller acknowledges funding from the National Institute on Aging, through grant number T32-AG00186 to the NBER.

The Review of Economics and Statistics, August 2008, 90(3): 414–427

studies 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. 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 II 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 III presents various possible bootstraps for clustered data, with additional details relegated to an appendix. Sections IV to VI 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. It also offers empiricists a clearer understanding of the menu of bootstrap choices and their consequences.

II. Cluster-Robust Inference

Before considering the bootstrap we present results on inference with clustered errors.[1]

A. 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

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

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.

[1] For this and subsequent sections, additional details and explanation are provided in the working-paper version of this paper (Cameron, Gelbach, & Miller, 2006).

Interest lies in inference for the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Under the assumptions that data are independent over g but errors are correlated within cluster, with $E[\mathbf{u}_g] = \mathbf{0}$, $E[\mathbf{u}_g\mathbf{u}_g'] = \boldsymbol{\Sigma}_g$, and $E[\mathbf{u}_g\mathbf{u}_h'] = \mathbf{0}$ for cluster $h \neq g$, we have $\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} \mathcal{N}[\mathbf{0}, NV[\hat{\boldsymbol{\beta}}]]$ where

$$V[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G} \mathbf{X}_g\boldsymbol{\Sigma}_g\mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{2}$$

This differs from and is usually larger than the specialization $V[\hat{\boldsymbol{\beta}}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$ that is based on the assumption of i.i.d. errors and leads to the default OLS variance estimate when σ_u^2 is estimated by $s^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(N - k)$. The underestimation bias is typically increasing in (a) cluster size; (b) within-cluster correlation of the regressor; and (c) within-cluster correlation of the error; see Kloek (1981). The bias can be very large (Moulton, 1986, 1990; BDM, 2004).

One approach to correcting this bias is to model $\boldsymbol{\Sigma}_g$ to depend on unknown parameters, say $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_g(\gamma)$, and then use estimate $\hat{\boldsymbol{\Sigma}}_g = \boldsymbol{\Sigma}_g(\hat{\gamma})$. The random-effects (RE) model assumes that there are cluster-specific i.i.d. random effects, estimates the variance of these effects and of the i.i.d. individual shocks, and uses these for $\hat{\boldsymbol{\Sigma}}_g$. We call the resulting standard errors Moulton-type standard errors.

B. Cluster-Robust Variance Estimates

The RE model places restrictions of homoskedasticity and equicorrelation within cluster, and assumes knowledge of the functional form $\boldsymbol{\Sigma}_g(\gamma)$. A less parametrically restrictive approach is to use the cluster-robust variance estimator (CRVE)

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

In the simplest case OLS residuals are used, so $\tilde{\mathbf{u}}_g = \hat{\mathbf{u}}_g = \mathbf{y}_g - \mathbf{X}_g\hat{\boldsymbol{\beta}}$.

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 $\tilde{\mathbf{u}}_g = \sqrt{c}\hat{\mathbf{u}}_g$ where $c = (G/(G - 1))((N - 1)/(N - k)) \simeq G/(G - 1)$ with large N), and is used in SAS in the GENMOD procedure (which uses $\tilde{\mathbf{u}}_g = \hat{\mathbf{u}}_g$).

A weakness of the standard CRVE with $\tilde{\mathbf{u}}_g = \hat{\mathbf{u}}_g$ is that it is biased, since $E[\hat{\mathbf{u}}_g\hat{\mathbf{u}}_g'] \neq \boldsymbol{\Sigma}_g = E[\mathbf{u}_g\mathbf{u}_g']$. The bias depends on the form of $\boldsymbol{\Sigma}_g$ but will usually be downwards.[2] Several corrected residuals $\tilde{\mathbf{u}}_g$ for equation (3) have been proposed by Kauermann and Carroll (2001) and Bell and McCaffrey

[2] For example, Kézdi (2004) uses $\tilde{\mathbf{u}}_g = \hat{\mathbf{u}}_g$ and finds in his simulations that with $G = 10$ the downwards bias is between 9% and 16%.

(2002). In our simulations we examine a correction proposed by Bell and McCaffrey (2002) that is equivalent to the jackknife estimate of the variance of the OLS estimator. This correction generalizes the HC3 measure (jackknife) of MacKinnon and White (1985), and so we refer to this correction as the CR3 variance estimator.[3] For a more detailed discussion of the computation of this estimator, see Cameron, Gelbach, and Miller (2006). Angrist and Lavy (2002) apply a related (but distinct) correction (which is a cluster generalization of the HC2 measure of MacKinnon & White, 1985) in an application with $G = 30$ to 40 and find that the correction increases cluster-robust standard errors by between 10% and 50%.

C. 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 $\boldsymbol{\beta}$.[4] We use the Wald test statistic

$$w = (\hat{\beta}_1 - \beta_1^0)/s_{\hat{\beta}_1}, \tag{4}$$

where $s_{\hat{\beta}_1}$ is the square root of the appropriate diagonal entry in $\hat{V}_{\mathrm{CR}}[\hat{\boldsymbol{\beta}}]$. 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.

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 $s_{\hat{\beta}}$.

General small-sample results are not possible even if the (clustered) errors are normally distributed. 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 may be 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 (2007) 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.

III. 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.[5]

There is no single bootstrap as there are different statistics that we may be interested in, different ways to form pseudo-samples, and different ways to use results for statistical inference. In this section we discuss several different bootstraps that are examined in our simulations. We provide greater detail on the bootstrap algorithms in appendix B, and in our working paper (Cameron, Gelbach, & Miller, 2006).

Choices that need to be made when bootstrapping include the following: what observational units to sample (individual observations or clusters); what objects to sample in generating bootstrap sample ($(\mathbf{y}, \mathbf{X})$ pairs, residuals drawn from the sample residuals, or residuals based on transformations of sample residuals); what statistics to calculate in each bootstrap replication; how to use the resulting bootstrap distribution of the statistics; and whether to impose the null hypothesis in generating the bootstrap samples. Some combinations of these choices provide asymptotic refinement; others do not. Some choices in principle provide valid tests, but in fact perform poorly with few clusters and commonly occurring empirical settings.

The statistic considered is the Wald test statistic w defined above. The data are clustered into G independent groups, so the resampling method should be one that assumes independence across clusters but preserves correlation within clusters.

A. Pairs Cluster Bootstrap-se and Bootstrap-t

The obvious method is to resample the clusters with replacement from the original sample $\{(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_G, \mathbf{X}_G)\}$. This resampling method is called a pairs cluster bootstrap.[6] A commonly used bootstrap in the empirical literature is to use a pairs bootstrap with the bootstrap-se procedure. The bootstrap-se procedure uses the bootstrap estimates of $\hat{\beta}_1$, denoted $\hat{\beta}_{1b}^*$, to form the bootstrap estimate of standard error

$$s_{\hat{\beta}_1,B} = \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\beta}_{1b}^* - \overline{\hat{\beta}_1^*})^2 \right)^{1/2}, \tag{5}$$

[3] The jackknife drops in turn each observation, here a cluster, computes the leave-one-out estimate $\hat{\boldsymbol{\beta}}_{(g)}$, $g = 1, \ldots, G$, and then uses variance estimate $(G - 1)/G \sum_g (\hat{\boldsymbol{\beta}}_{(g)} - \hat{\boldsymbol{\beta}})$. The CR3 measure for OLS is a multiple of the related measure proposed by Mancl and DeRouen (2001) in the more general setting of GEE.

[4] 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 asymptotic *chi*-square test would be used.

[5] 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).

[6] Alternative names used in the literature include cluster bootstrap, case bootstrap, nonparametric bootstrap, and nonoverlapping block bootstrap.

where $\overline{\hat{\beta}_1^*} = (1/B)\sum_{b=1}^{B}\hat{\beta}_{1b}^*$. This estimated standard error is then used in a typical Wald test. This is popular in the applied literature as it enables an estimate of the standard error when the analytic formula for the standard error is difficult to compute. However, in contrast to the bootstrap-t procedure, it does not offer asymptotic refinement, and so may perform worse with few clusters.

The bootstrap-t procedure, proposed by Efron (1981) for confidence intervals, computes the following Wald statistic for each bootstrap replication

$$w_b^* = (\hat{\beta}_{1,b}^* - \hat{\beta}_1)/s_{\hat{\beta}_{1,b}^*}, \quad b = 1, \ldots, B,$$

where $s_{\hat{\beta}_{1,b}^*}$ is a cluster-robust standard error for $\hat{\beta}_{1,b}^*$. Note that w_b^* is centered on $\hat{\beta}_1$. This centering is changed to β_1^0 if the resampling method imposes H_0. The resulting distribution of $w_1^*, \ldots, w_B^*$ is then used to form inference on the original Wald statistic in equation (4).[7] We offer more details in appendix B.

Both bootstrap-se and bootstrap-t procedures are asymptotically valid. For small number of clusters G, however, the true size will differ from the nominal significance level α. Furthermore, the true size will also differ across the two procedures. An asymptotic approximation yields an actual rejection rate or true size $\alpha + O(G^{-j/2})$. 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. Bootstrap-t procedures provide an asymptotic refinement, while bootstrap-se procedures do not. Further, as we show in our simulations below, this can matter in actual data settings with few clusters.

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 for a more complete discussion. The bootstrap-t procedure directly bootstraps w, which is asymptotically pivotal since the standard normal has no unknown parameters.

An alternative method with asymptotic refinement is the bias-corrected accelerated (BCA) procedure, defined in Efron (1987), Hall (1992, pp. 128–141), and our working paper (Cameron, Gelbach, & Miller, 2006). This bootstraps $\hat{\beta}_1$, which is asymptotically nonpivotal as its asymptotic distribution depends on unknown $\sigma^2_{\hat{\beta}_1}$, but then provides adjustment for bias and asymmetry. This is a popular method for confidence intervals—STATA reports BCA rather than percentile-t confidence intervals. We adapt BCA to testing by rejecting H_0 if w is outside the confidence interval, and include a cluster version of BCA in our simulations.

[7] The bootstrap-t procedure 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.

There are just a few studies that we are aware of that consider asymptotic refinement. Sherman and le Cessie (1997) conduct simulations for OLS with as few as ten clusters. For 90% confidence intervals, they find that the pairs cluster bootstrap-t undercovers by considerably less than confidence intervals based on CRVE. 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 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 their method may not yield asymptotic refinement. 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.

B. Residual and 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 $\{\hat{\mathbf{u}}_1^*, \ldots, \hat{\mathbf{u}}_G^*\}$ and hence pseudo-sample $\{(\hat{\mathbf{y}}_1^*, \mathbf{X}_1), \ldots, (\hat{\mathbf{y}}_G^*, \mathbf{X}_g)\}$ where $\hat{\mathbf{y}}_g^* = \mathbf{X}_g'\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_g^*$.

This resampling scheme has two weaknesses. First, it assumes that the regression error vectors $\mathbf{u}_g$ are i.i.d., whereas in section II 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 wild bootstrap relaxes both these restrictions. This procedure creates pseudo-samples based on $\hat{\mathbf{u}}_g^* = \hat{\mathbf{u}}_g$ with probability 0.5 and $\hat{\mathbf{u}}_g^* = -\hat{\mathbf{u}}_g$ with probability 0.5, with this assignment at the cluster level. The wild bootstrap was proposed for regression in the nonclustered case (and with different weights) by Wu (1986). 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. The weights we

use (+1 with probability 0.5 and −1 with probability 0.5) are called Rademacher weights.[8] Here we have extended the wild bootstrap to the clustered setting. The only other 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 $\hat{\beta}_1$, and the residuals bootstrapped are those from the restricted OLS estimator $\tilde{\boldsymbol{\beta}}$ that imposes $H_0 : \beta_1 = \beta_1^0$. For details see appendix B.

C. Bootstraps with Few Clusters

With few clusters the bootstrap resampling methods produce a distinctly finite number of possible pseudo-samples, so the bootstrap distribution $w_1^*, \ldots, w_q^*$ will not be smooth even with many bootstrap replications. Furthermore, in some pseudo-samples $\hat{\beta}_1$ or $s_{\hat{\beta}_1}$ may be inestimable. This is likely to be a problem with pairs cluster bootstrap when there is a binary regressor that is invariant within cluster (so always 0 or always 1 for given g). Then if there are few clusters some bootstrap resamples may have all clusters with the regressor taking only value 0 (or value 1), so that $\hat{\beta}_k$ is inestimable. This issue does not arise when regressors and dependent variables take several different values, such as in the section IV Monte Carlos. But it does arise in our application to the BDM (2004) and Gruber and Poterba (1994) differences-in-differences examples, because the regressors of interest in those cases are indicator variables. The residual or wild cluster bootstraps do not encounter these problems, as we are not resampling the regressors. In our BDM simulations below we see this issue is problematic for $G \leq 10$.

D. 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 nonbootstrap procedures. Table 1 provides a summary.

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

Table 1.—Different Methods for Wald Test

Method	Bootstrap?	Refinement?	H_0 imposed?
Conventional Wald			
1. Default (i.i.d. 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

Methods 5 to 7 use the bootstrap-se procedure. 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 B2. Methods 5–7 do not provide asymptotic refinement, and method 6 is valid only if cluster error vectors are i.i.d.

Method 8 uses the BCA bootstrap with pairs cluster resampling.

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 equation (3), 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 the 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 i.i.d. For details see Appendix B1.

IV. 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

[8] These weights lead to asymptotic refinement if $\hat{\boldsymbol{\beta}}$ is symmetrically distributed, which is the case if errors are symmetric. If $\hat{\boldsymbol{\beta}}$ is asymmetrically distributed, our version is still asymptotically valid, but different weights provide asymptotic refinement. Davidson and Flachaire (2001) provide theory and simulation to nonetheless support using Rademacher weights even in the asymmetric case.

TABLE 2.—1,000 SIMULATIONS FROM DGP WITH GROUP-LEVEL RANDOM ERRORS
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

Estimator #	Method	Number of Groups (G)					
		5	10	15	20	25	30
1	Assume i.i.d.	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-2)	0.145	0.086	0.072	0.066	0.062	0.060

unlikely that a pairs cluster bootstrap sample will be inestimable.

Then

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig}, \tag{6}$$

with different generating processes for x_{ig} and u_{ig} used in subsequent subsections. Since $\beta_1 = 1$ in the dgp, the Wald test statistic is $w = (\hat{\beta}_1 - 1)/s_{\hat{\beta}_1}$.

We perform R replications, where each replication yields a new draw of data from the dgp, and leads to rejection or nonrejection of H_0. In each replication there are G groups ($g = 1, \ldots, G$), with N_G individuals ($i = 1, \ldots, N_G$) in each group. We varied the number of groups G from 5 to 30 and usually set $N_G = 30$. The various methods given in each row 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 $\hat{a}$, the fraction of the R replications for which H_0 is rejected. This is an estimate of the true size of the test. With a finite number of replications a may differ from the true size because of simulation error. The simulation standard error is $s_{\hat{a}} = \sqrt{\hat{a}(1 - \hat{a})/(R - 1)}$. For example, $s_{\hat{a}} = 0.007$ for $\hat{a} = 0.05$ and $R = 1000$.

A. *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

$$\begin{aligned} y_{ig} &= \beta_0 + \beta_1 x_{ig} + u_{ig} \\ &= \beta_0 + \beta_1(z_g + z_{ig}) + (\varepsilon_g + \varepsilon_{ig}), \end{aligned} \tag{7}$$

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. The simulation is based on $R = 1000$ Monte Carlo replications.

Our first results appear in table 2. Each column gives results for the various numbers 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 Monte Carlo standard error is given in parentheses. Each row presents a different method, detailed in section IIID. 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 $G - 2$ degrees of freedom, $\Pr[|T| > 1.96 | T \sim T_{G-2}]$.

We begin with conventional (nonbootstrap) Wald tests using different estimators of standard errors. The default OLS standard errors that assume i.i.d. errors do poorly here, with rejection rates given in row 1 of 0.43 to 0.50. This

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 standard errors in parentheses)

Estimator #	Method	Number of Groups (G)					
		5	10	15	20	25	30
1	Assume i.i.d.	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-2)	0.145	0.086	0.072	0.066	0.062	0.060

illustrates the need to correct standard errors for clustering. The Moulton-type estimate for standard errors should work well here since 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 $T(G - 2)$ distributed. The CRVE is much better than default standard errors, though still over-rejects compared with Moulton-type standard errors. The CR3 correction leads to rejection rates much closer to (but significantly different from) 0.05.

The pairs cluster bootstrap-se method yields rejection rates in row 5 that are very similar to the CRVE, 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 a 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 asymptotic refinement, yet from row 8 it has rejection rates similar to those using CRVE.

The remainder of table 2 uses 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 CRVE and is a great improvement compared with not bootstrapping (row 1). The pairs cluster bootstrap-t has rejection rates in row 10 of 0.08 that are much closer to (but significantly different from) 0.05. The CR3 correction makes little difference. 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 on CRVE; and the residual cluster bootstrap-se also performs well.

B. 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 following process:

$$\begin{aligned} y_{ig} &= \beta_0 + \beta_1 x_{ig} + u_{ig} \\ &= \beta_0 + \beta_1(z_g + z_{ig}) + (\varepsilon_g + \varepsilon_{ig}), \end{aligned} \tag{8}$$

with z_g, z_{ig}, and ε_g again independent $\mathcal{N}[0, 1]$ draws, but now $\varepsilon_{ig} \sim \mathcal{N}[0, 9 \times (z_g + z_{ig})^2]$. The dgp sets $\beta_0 = 1$ and $\beta_1 = 1$.

Results appear in table 3. Default OLS standard errors again do poorly, with rejection rates around 0.30. The Moulton-type correction breaks down given the heteroskedasticity, as expected. The cluster-robust methods do a little better than in the preceding table, but rejection rates in rows 3 and 4 still generally exceed 0.05. The residual cluster

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 standard errors in parentheses)

Estimator #	Method	Column Number	Main—from Table 2	Reject based on T (8 dof)	Cluster Size = 2	Cluster Size = 10	Cluster Size = 100	4 RHS Variables	Xs are Constant Within Group	Xs Are i.i.d.	Unbalanced Group Sizes (10, 50)
			1	2	3	4	5	6	7	8	9
1	Assume i.i.d.		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								

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 CRVE (row 3).

The results for the bootstrap-t methods in rows 9 to 13 are similar to those in table 2. The 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.

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.

C. *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. In column 2, for tests without asymptotic refinement we 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 CRVE 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, two, ten, and one hundred observations. For method 1, the rejection rates increase with cluster size. 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 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 one-half, 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 it to $\rho_x = 0$ (i.i.d. regressor with $x_{ig} = z_{ig}$). In both cases the regressor is scaled up by $\sqrt{2}$ to keep $V[x_{ig}]$ unchanged.

With cluster-invariant regressor (column 7) the failure to control for clustering is magnified and the rejection rates in rows 1 to 3 are larger than in the benchmark column 1. For bootstrap-se and bootstrap-t there is little change in rejection rates, except that for reasons unknown the pairs cluster bootstraps (both bootstrap-se and bootstrap-t) now have rejection rates not statistically significantly different from 0.05.

With i.i.d. regressor (column 8) the default OLS standard errors are consistent and the rejection rate in row 1 is close to 0.05. The Moulton-type and CRVE also have rejection rates much closer to 0.05. The various bootstrap procedures lead to rejection rates that are all close to those in column 1.

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.

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 also indicates that when nonbootstrap methods are used to control for clustering, it is better to use critical values from a $T(G - 2)$ distribution than from a standard normal.

V. 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). The main results of this section are that the residual and wild cluster bootstrap methods perform well in cases with as few as six clusters. These results stand in contrast to the more pessimistic conclusion about cluster bootstrapping in BDM (2004).

The data set is of U.S. states over time. The dependent variable is the state-by-year average log wage level (after partialing 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. The regressor of interest is a state policy dummy variable.

The original data are 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.[9]

A. Aggregated State-Year Data

Using our choice of subscripts, the igth observation is for the ith year in the gth state. There are 50 states and 21 years. The aggregate 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, and the regressors are state dummies, year dummies, and a policy change indicator I_{ig}.[10]

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. In each simulation a different draw of G states with replacement is made from the original 50 states.

The Wald statistic studied is $w = \hat{\beta}_1 / s_{\hat{\beta}_1}$. BDM investigate size properties by letting the policy change be a "placebo" regressor that has no effect on y_{ig}. They also investigate the power against the alternative $H_a: \beta_1 = 0.02$ by actually increasing y_{ig} by 0.02 when $I_{ig} = 1$. They find that (a) default standard errors do poorly; (b) cluster-robust standard errors do well for all but $G = 6$; and (c) their bootstrap, which we discuss in our section IIIA, 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 two rows of table 5 show that the default standard errors and Moulton-type estimator lead to high rejection rates. The third row uses the cluster-robust variance estimator, and gives results very close to those in BDM's table 8.

Rows 4 to 6 of our 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

[9] 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.

[10] 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 procedure: (a) regress y_{ist} on $\mathbf{x}_{ist}$ yielding OLS residual $\hat{u}_{ist}$; (b) regress $\bar{\hat{u}}_{st} = N_{st}^{-1} \sum_i \hat{u}_{ist}$ on state dummies, year dummies, and I_{st}. Thus our y_{ig} is their $\bar{\hat{u}}_{st}$.

TABLE 5.—1,000 SIMULATIONS FROM BDM (2004) DESIGN
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses; size column measures size and power column measures power)

Estimator #	Method	Number of States (G) 6 Size	10 Size	20 Size	50 Size	6 Power	10 Power	20 Power	50 Power
1	Assume i.i.d.	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)

cluster-pairs standard error bootstrap (row 4) produces severe under-rejection (0.001) with $G = 6$. Informal experimentation suggests to us that this is because 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 (3) the resulting $\hat{V}_{CR}[\hat{\beta}_1^*]$ 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.[11]

The BCA method with pairs cluster resampling in general leads to greater over-rejection than when CRVE is used.

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. 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 discussed above. The residual and wild cluster bootstrap-t methods (rows 10 and 11) do very well with actual rejection rates approximately equal to 0.05, even for $G = 6$.

The discussion so far has focused on size. The last 4 columns in table 5 report power against a fixed alternative. As expected, power increases as the number of clusters increases, since there is then greater precision in estimation.

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

B. *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 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. I_{ig} is generated as before.

Table 6 reports the results of $R = 250$ simulations with $B = 199$ replications used for the bootstrap. We consider cases $G = 6$ and $G = 10$. The first row reports high rejection rates when we use the CRVE but erroneously cluster on state-year combinations. In the second row of table 6 we see that using the CRVE and correctly clustering

TABLE 6.—250 SIMULATIONS FROM BDM (2004) DESIGN USING MICRODATA
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

Estimator #	Method	Number of States (G) 6 Size	10 Size
1	CRVE cluster on state-year	0.440	0.444
		(0.031)	(0.031)
3	CRVE cluster on state	0.148	0.100
		(0.023)	(0.019)
11	Wild bootstrap-t cluster on state	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 $R = 250$. Number of bootstrap replications $B = 199$.

on state considerably reduces the rejection rates, but they are still much too high. 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.

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 because of problems posed by the binary regressor.

VI. 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 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 model can be written 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 (2007). Like Donald and Lang, we ignore additional regressors (GP examine subtle interactions between pretax income, employment status, and the TRA86).

In their preliminary analysis, GP report in their 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.

We next follow Donald and Lang (2007), and treat years as clusters, so that there are $G = 8$ clusters in our analysis. When we cluster on year, the cluster-robust standard error obtained using equation (3) is 0.0074. The regressor is highly statistically significant, with $\hat{\beta}_1/s_{\hat{\beta}_1} = 7.46$ and very low p-value.

To enable more meaningful analysis we test $H_0 : \beta_1 = 0.040$ against $H_a : \beta_1 \neq 0.040$. 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.[12] 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.

We have also estimated this model on individual-level data (see Cameron, Gelbach, & Miller, 2006), with results very similar to those reported here.

VII. Conclusion

Many microeconometric studies use clustered data, with regression errors 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 equation (3) 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. 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.

We focus on the bootstrap-t procedure, the method most emphasized by theoretical econometricians and statisticians, and which provides asymptotic refinement. We find that the bootstrap-t procedure can lead to considerable improvement, provided 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 gth cluster is selected then all data (dependent and

[12] The results reported use Mammen weights and do not impose the null hypothesis. Similar results were obtained using Rademacher weights and imposing the null hypothesis.

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 (conditional on other regressors) 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 a moderate (ten and twenty) number of clusters their bootstrap-based method also does poorly. We perform a reanalysis 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.

REFERENCES

Angrist, J., and V. Lavy, "The Effect of High School Matriculation Awards: Evidence from Randomized Trials," NBER working paper no. 9389 (2002).

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

Bell, R. M., and D. F. McCaffrey, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology* 28:2 (2002), 169–179.

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

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

Cameron, A. C., J. G. Gelbach, and D. L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," U.C. Davis working paper #06-21 (2006).

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

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

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

——— *Econometric Theory and Methods* (Oxford: Oxford University Press, 2004).

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

Donald, S. G., and K. Lang, "Inference with Difference-in-Differences and Other Panel Data," this REVIEW 89:2 (2007), 221–233.

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

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

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

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

Flynn, T. N., and T. J. Peters, "Use of the Bootstrap in Analysing Cost Data from Cluster Randomised Trials: Some Simulation Results," *BMC Health Services Research* 4:33 (2004).

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

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

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

——— "The Bootstrap" (pp. 3159–3228), in J. J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics,* Vol. 5 (Amsterdam: North-Holland, 2001).

Ibragimov, Rustam, and Ulrich K. Muller, "*t*-Statistic Based Correlation and Heterogeneity Robust Inference," Harvard Institute of Economic Research discussion paper no. 2129 (February 2007).

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

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

Kloek, T., "OLS Estimation in a Model where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated," *Econometrica* 49 (1981), 205–207.

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

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

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

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

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

Mancl, L. A., and T. A. DeRouen, "A Covariance Estimator for GEE with Improved Finite-Sample Properties," *Biometrics* 57 (2001), 126–134.

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

——— "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," this REVIEW 72 (1990), 334–338.

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

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

Sherman, M., and S. le Cessie, "A Comparison Between Bootstrap Methods and Generalized Estimating Equations for Correlated Outcomes in Generalized Linear Models," *Communications in Statistics—Simulation and Communication* 26 (1997), 901–925.

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

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

Wooldridge, J. M., "Cluster-Sample Methods in Applied Econometrics," *American Economic Review* 93 (2003), 133–138.
Wu, C. F. G., "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis," *Annals of Statistics* 14 (1986), 1261–1295.

APPENDIX A

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

Asymptotic Refinement for Bootstrap-t

The theory draws heavily on Hall (1992) and Horowitz (2001). Cameron and Trivedi (2005) provide a more introductory discussion.

1. General bootstrap procedure

We use the generic notation $T_N = T_N(\mathbf{S}_N)$ to denote the statistic of interest, calculated on the basis of a sample $\mathbf{S}_N$ of size N. We focus on inference for a single regression coefficient β_1 from multivariate OLS regression. Then leading examples are $T_N = \hat{\beta}_1$, and $T_N = (\hat{\beta}_1 - \beta_1^0)/s_{\hat{\beta}_1}$, 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 $\mathbf{S}_N$, using methods given in the subsequent subsection. The bth resample is denoted $\mathbf{S}_{Nb}^*$ and is used to form a statistic $T_{Nb}^* = T_N^*(\mathbf{S}_{Nb}^*)$. The empirical distribution of T_{Nb}^*, $b = 1, \ldots, B$, 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 $T_{N1}^*, \ldots, T_{Nb}^*$ that are less than t, denoted

$$\hat{H}_N(t) = B^{-1} \sum_{b=1}^{B} \mathbf{1}(T_{Nb}^* \leq t), \tag{A1}$$

where $\mathbf{1}(\cdot)$ is the indicator function. 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 bth iteration:
   (a) Resample the data from $\mathbf{S}_N$ using one of the procedures presented in appendix B. Call the resulting resample $\mathbf{S}_{Nb}^*$.
   (b) Use the bootstrap resample form $T_{Nb}^* = T_N^*(\mathbf{S}_{Nb}^*)$, where in some but not all cases $T_N^*(\cdot) = T_N(\cdot)$.
2. Conduct inference using $\hat{H}_N(t)$. See appendix B for further details.

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

By contrast the bootstrap-se, percentile, and BCA methods bootstrap $T_N = \hat{\beta}_1$. Then $T_{Nb}^* = \hat{\beta}_{1b}^*$, where $\hat{\beta}_{1b}^*$ is the estimator of β_1 from resample $\mathbf{S}_{Nb}^*$.

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 $\hat{H}(t|\hat{F}) = \Pr[T^* \leq t|\hat{F}]$ where $\hat{F}$ denotes the cdf used to obtain bootstrap resamples. We are concerned with how good an estimate $\hat{H}(t|\hat{F})$ 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 $\hat{F}$ 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 $\hat{H}$. 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

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

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

$$\Pr[T \leq t|F] = \Pr[T^* \leq t|\hat{F}] + O_p(N^{-1}). \tag{A2}$$

This statement means that the bootstrap cdf $\Pr[T^* \leq t|\hat{F}]$ 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

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

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

$$\Pr[T \leq t|F] = \Pr[T^* \leq t|\hat{F}] + O_p(N^{-1/2}), \tag{A3}$$

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 (A2) 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 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.

APPENDIX B

Bootstrap Procedures

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, residual, and wild resampling for nonclustered data.

Pairs cluster bootstrap-t:

1. From the original sample form $w = (\hat{\beta}_1 - \beta_0)/s_{\hat{\beta}_1}$, where $s_{\hat{\beta}_1}$ is obtained using the CRVE in equation (3) with $\tilde{\mathbf{u}}_g = (G/(G-1))\mathbf{u}_g$.
2. Do B iterations of this step. On the bth iteration:
   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), \ldots, (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample of clusters.
   (b) Calculate the Wald test statistic $w_b^* = (\hat{\beta}_{1,b}^* - \hat{\beta}_1)/s_{\hat{\beta}_{1,b}^*}$, where $\hat{\beta}_{1,b}^*$ and its standard error $s_{\hat{\beta}_{1,b}^*}$ are obtained from OLS estimation using the bth pseudo-sample, $s_{\hat{\beta}_{1,b}^*}$ is obtained using the same method as that in step 1, and $\hat{\beta}_1$ 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 qth quantile of $w_1^*, \ldots, w_B^*$.

We consider two variations of this procedure that use alternative estimators of $s_{\hat{\beta}}$ both in step 1 and in step 2b. First, the pairs cluster CR3 bootstrap-t uses the CRVE in equation (3) with $\tilde{\mathbf{u}}_g$ calculated using the CR3 correction. 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 $w = (\hat{\beta}_1 - \beta_0)/s_{\hat{\beta}_1}$, where $s_{\hat{\beta}_1}$ is obtained using the CRVE in equation (3) with $\tilde{\mathbf{u}}_g = (G/(G-1))\hat{\mathbf{u}}_g$. Also obtain the restricted OLS estimator $\hat{\boldsymbol{\beta}}^R$ that imposes $H_0 : \beta_1 = \beta_1^0$, and the associated restricted OLS residuals $\{\hat{\mathbf{u}}_1^R, \ldots, \hat{\mathbf{u}}_G^R\}$.[13]
2. Do B iterations of this step. On the bth iteration:
   (a) Form a sample of G clusters $\{(\hat{\mathbf{y}}_1^*, \mathbf{X}_1), \ldots, (\hat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$ by resampling with replacement G times from $\{\hat{\mathbf{u}}_1^R, \ldots, \hat{\mathbf{u}}_G^R\}$ to give $\{\hat{\mathbf{u}}_1^{R*}, \ldots, \hat{\mathbf{u}}_G^{R*}\}$ and then forming $\hat{\mathbf{y}}_g^* = \mathbf{X}_g'\hat{\boldsymbol{\beta}}^R + \hat{\mathbf{u}}_g^{R*}$, $g = 1, \ldots, G$.
   (b) Calculate the Wald test statistic $w_b^* = (\hat{\beta}_{1,b}^* - \beta_1^0)/s_{\hat{\beta}_{1,b}^*}$ where $\hat{\beta}_{1,b}^*$ and its standard error $s_{\hat{\beta}_{1,b}^*}$ are obtained from unrestricted OLS estimation using the bth pseudo-sample, with $s_{\hat{\beta}_{1,b}^*}$ 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 qth quantile of $w_1^*, \ldots, w_B^*$.

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 i.i.d., ruling out heteroskedasticity across clusters, and that clusters are balanced.

Wild cluster bootstrap-t with H_0 imposed:

The wild cluster bootstrap-t with H_0 imposed follows the same steps as the cluster residual bootstrap-t with H_0 imposed, except that step 2a is replaced as follows:

2a. Form a sample of G clusters $\{(\hat{\mathbf{y}}_1^*, \mathbf{X}_1), \ldots, (\hat{\mathbf{y}}_G^*, \mathbf{X}_G)\}$ by the following method. For each cluster $g = 1, \ldots, G$, form either $\hat{\mathbf{u}}_g^{R*} = \hat{\mathbf{u}}_g^R$ with probability 0.5 or $\hat{\mathbf{u}}_g^{R*} = -\hat{\mathbf{u}}_g^R$ with probability 0.5 and then form $\hat{\mathbf{y}}_g^* = \mathbf{X}_g'\hat{\boldsymbol{\beta}}^R + \hat{\mathbf{u}}_g^{R*}$, $g = 1, \ldots, G$.

A variety of weights a_g have been proposed for the wild bootstrap. 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 $E[a_g] = 0$ and $E[a_g^2] = 1$ and the additional constraint $E[a_g^3] = 1$, which is necessary to achieve asymptotic refinement if $\hat{\boldsymbol{\beta}}$ is asymmetrically distributed.

2. Bootstrap-se Methods

We present the bootstrap-se for pairs cluster resampling.

Pairs cluster bootstrap-se:

1. From the original sample form $\hat{\beta}_1$.
2. Do B iterations of this step. On the bth iteration:
   (a) Form a sample of G clusters $\{(\mathbf{y}_1^*, \mathbf{X}_1^*), \ldots, (\mathbf{y}_G^*, \mathbf{X}_G^*)\}$ by resampling with replacement G times from the original sample.
   (b) Calculate the OLS estimate $\hat{\beta}_{1,b}^*$.
3. Reject H_0 at level α if and only if $|w_{\text{BSE}}| > z_{\alpha/2}$, where

$$w_{\text{BSE}} = \frac{\hat{\beta}_1 - \beta_1^0}{s_{\hat{\beta}_{1,B}}},$$

$s_{\hat{\beta}_{1,B}}$ is the bootstrap estimate of the standard error

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

and $\overline{\hat{\beta}_1^*} = (1/B) \sum_{b=1}^{B} \hat{\beta}_{1b}^*$.

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

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