Queen's Economics Department Working Paper No. 1314

Wild Bootstrap Inference for Wildly Different Cluster Sizes

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

Corrected 2019-08

Wild Bootstrap Inference for Wildly Different Cluster Sizes *

James G. MacKinnon
jgm@econ.queensu.ca

Matthew D. Webb
matt.webb@carleton.ca

September 18, 2019

Abstract

The cluster robust variance estimator (CRVE) relies on the number of clusters being sufficiently large. Monte Carlo evidence suggests that the "rule of 42" is not true for unbalanced clusters. Rejection frequencies are higher for datasets with 50 clusters proportional to U.S. state populations than with 50 balanced clusters. Using critical values based on the wild cluster bootstrap performs much better. However, this procedure fails when a small number of clusters is treated. We explain why CRVE t statistics and the wild bootstrap fail in this case, study the "effective number" of clusters, and simulate placebo laws with dummy variable regressors.

Keywords: CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, bootstrap failure, difference in differences, effective number of clusters, placebo laws

This paper has been published. Please cite the published version:

James G. MacKinnon and Matthew D. Webb, "Wild bootstrap inference for wildly different cluster sizes," *Journal of Applied Econometrics*, 32, 2017, 233–254.

This working paper is being reissued to correct results that were affected by an algebraic error in programs used for some of the simulations. Programs that were supposed to generate data with an intra-cluster correlation of ρ actually generated data with an intra-cluster correlation of ρ^2. All figures and tables that have been corrected are so indicated.

*We are grateful to five referees and to seminar participants at Camp Econometrics, Ryerson University, the University of Calgary, Syracuse University, the Institute for Fiscal Studies, the Canadian Econometric Study Group, the Midwest Econometrics Group, Université du Québec à Montréal, Wilfrid Laurier University, Indiana University, Carleton University, and McMaster University for comments on earlier versions. We also thank Russell Davidson, Yulia Kotlyarova, Doug Steigerwald, Arthur Sweetman, and Yuanyuan Wan for helpful suggestions, and Andrew Carter, Kevin Schnepel, and Doug Steigerwald for their computer code. MacKinnon's research was supported, in part, by grant 410-2009-0194 from the Social Sciences and Humanities Research Council.

1 Introduction

Many empirical papers use data that are clustered or grouped. This clustering causes problems for inference whenever there is intra-cluster correlation, especially when there are independent variables that are constant within groups. This problem has been known since Kloek (1981) and Moulton (1990), and many procedures have been developed to deal with the tendency for intra-cluster correlation to bias standard errors downwards. The most common procedure is the cluster robust variance estimator (CRVE), which uses a formula (see Section 2) proposed in several papers, of which the earliest may be Liang and Zeger (1986). Stata uses this estimator when the `cluster` command is invoked.

The cluster robust variance estimator has been shown to work well when the number of clusters is large. However, several papers have pointed out problems with the estimator when the number of clusters is small. General results on covariance matrix estimation in White (1984) imply that the CRVE is consistent under three key assumptions:

A1. The number of clusters goes to infinity.

A2. The within-cluster error correlations are the same for all clusters.

A3. Each cluster contains an equal number of observations.

The limitations of the CRVE when assumption A1 is poor are now well-known; see, among others, Bertrand, Duflo and Mullainathan (2004), Donald and Lang (2007), and Brewer, Crossley and Joyce (2018). A wild bootstrap procedure that often works well when the number of clusters is not too small was proposed by Cameron, Gelbach and Miller (2008). It was modified to handle cases with twelve or fewer clusters by Webb (2014). Assumptions A2 and A3 were relaxed by Carter, Schnepel and Steigerwald (2017), which also showed how to calculate the "effective number" of clusters for cases with heterogeneous within-cluster correlation and unequal (unbalanced) cluster sizes. Cameron and Miller (2015) provides a thorough recent survey of cluster robust inference.

Assumption A3 is particularly important. Previous Monte Carlo experiments on the effectiveness of the CRVE, notably those in Bertrand, Duflo and Mullainathan (2004) and Cameron, Gelbach and Miller (2008), primarily use datasets with equal-sized clusters. Both papers also perform experiments with data from the Current Population Survey (CPS), as discussed in Section 7. Most of the simulations in the former paper use data averaged over state-year pairs. This imposes assumption A3, because each state has only one observation per year. Some simulations do involve micro data (unbalanced, clustered by state or state-year pair), but rejection rates with clustering at the state level are not reported.

Previous results have led to the rule of thumb that the CRVE works reasonably well when the number of clusters is sufficiently large. Angrist and Pischke (2008) suggests that 42 clusters are enough for reliable inference. However, we show that the "rule of 42" no longer holds when the assumption of equal-sized clusters is relaxed. Inference using CRVE standard errors can be quite unreliable even with 100 unbalanced clusters.

Many real-world datasets have wildly unequal cluster sizes. American datasets clustered at the state level are a prime example. A dataset with observations in clusters proportional to current state populations will have 12% of the sample from California. Eleven states will each contain less than 0.5% of the total sample, and the largest cluster will be roughly sixty times the size of the smallest one. This is a severe violation of the assumption of equal-sized

clusters.

The remainder of the paper is organized as follows. The next two sections briefly discuss the two methods that we investigate which promise improved inference with clustered data. Section 2 describes the wild cluster bootstrap, and Section 3 discusses the use of critical values for CRVE t statistics based on the effective number of clusters.

Section 4 presents Monte Carlo evidence using simulated datasets with a continuous test regressor and either equal cluster sizes or ones proportional to state populations. We show that inference based on CRVE t statistics can perform poorly in the latter case. Using critical values based on the effective number of clusters instead of the actual number usually improves matters, but it does not always yield reliable inferences. In contrast, the wild cluster bootstrap procedure always performs extremely well.

The remainder of the paper deals with estimating treatment effects, mainly in the context of difference-in-differences (DiD) estimates. For treatment effects, cluster sizes matter, but the number of treated clusters matters even more. The wild bootstrap works very well in most cases, but all the methods fail badly when the number of treated clusters is small or (in some cases) large. Section 5 presents Monte Carlo evidence for the DiD case using simulated datasets. Additional simulation results for the pure treatment case are presented in Section A.2 of the appendix.

Section 6 explains why inference based on CRVE t statistics, with or without the wild bootstrap, can fail with few treated clusters. The theoretical results of this section accord remarkably well with our simulation results. Section 7 extends the "placebo laws" Monte Carlo experiments of Bertrand, Duflo and Mullainathan (2004). Section 8 contains a brief empirical example based on Angrist and Kugler (2008), and Section 9 concludes.

2 The Wild Cluster Bootstrap

A linear regression model with clustered errors may be written as

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_G \end{bmatrix}, \tag{1}$$

where each cluster, indexed by g, has N_g observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ have $N = \sum_{g=1}^{G} N_g$ rows, $\boldsymbol{X}$ has k columns, and the parameter vector $\boldsymbol{\beta}$ has k rows. The covariance matrix of $\boldsymbol{\epsilon}$ is $\boldsymbol{\Omega}$, an $N \times N$ block-diagonal matrix with G diagonal blocks $\boldsymbol{\Omega}_g$, each of them $N_g \times N_g$, corresponding to the G clusters.[1] OLS estimation of equation (1) yields estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$. There are several cluster robust variance estimators. The most popular CRVE, which we investigate, appears to be

$$\frac{G(N-1)}{(G-1)(N-k)} (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2}$$

[1] The (excessively strong) conditions of White (1984) imply that the matrices $\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g$ must be identical across clusters, which is unlikely to hold in practice.

The first factor here is asymptotically negligible, but it always makes the CRVE larger when G and N are finite.

We wish to test the hypothesis that a single coefficient is zero. Without loss of generality, we let this be the last one, β_k. The procedure for using the (restricted) wild cluster bootstrap of Cameron, Gelbach and Miller (2008) to test the hypothesis that $\beta_k = 0$ is:

1. Estimate equation (1) by OLS.

2. Calculate t_k, the t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (2) as a cluster robust standard error.

3. Re-estimate the model (1) subject to the restriction that $\beta_k = 0$, so as to obtain the restricted residuals $\tilde{\boldsymbol{\epsilon}}$ and the restricted estimates $\tilde{\boldsymbol{\beta}}$.

4. For each of B bootstrap replications, indexed by j, generate a new set of bootstrap dependent variables y_{ig}^{*j} using the bootstrap DGP

$$y_{ig}^{*j} = \boldsymbol{X}_{ig}\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_{ig} v_g^{*j}, \tag{3}$$

   where y_{ig}^{*j} is an element of the vector $\boldsymbol{y}^{*j}$ of observations on the bootstrap dependent variable, $\boldsymbol{X}_{ig}$ is the corresponding row of $\boldsymbol{X}$, and so on. Here v_g^{*j} is a random variable that follows the Rademacher distribution; see Davidson and Flachaire (2008). It takes the values 1 and -1 with equal probability. Note that we would not want to use the Rademacher distribution if G were smaller than about 12; see Webb (2014), which proposes an alternative for such cases.

5. For each bootstrap replication, estimate (1) using $\boldsymbol{y}^{*j}$ as the regressand, and calculate t_k^{*j}, the bootstrap t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (2), with bootstrap residuals replacing OLS ones, as the standard error.

6. Calculate the symmetric bootstrap P value

$$\hat{p}_{\text{s}}^{*} = \frac{1}{B}\sum_{j=1}^{B} I\Big(|t_k^{*j}| > |t_k|\Big). \tag{4}$$

   It would also be valid to use an equal-tail P value. In our simulations, symmetric and equal-tail P values were always extremely close.

The wild cluster bootstrap procedure described here has two key features. The first is that the bootstrap error terms for every observation in cluster g depend on the same random variable v_g^{*j}. This ensures that, to the extent that the residuals $\tilde{\epsilon}_{ig}$ preserve the variances and within-cluster covariances of the error terms ϵ_{ig}, the bootstrap DGP also preserves these properties. However, this feature can cause serious problems in certain cases; see Section 6. The second key feature is that the bootstrap DGP (3) uses estimates under the null hypothesis. It would also be valid to use the unrestricted residuals $\hat{\epsilon}_{ig}$ instead of the restricted ones $\tilde{\epsilon}_{ig}$ in (3), and we discuss that variant in Section 6. The two variants of the wild cluster bootstrap can yield dramatically different results when the number of treated clusters is small.

3 The Effective Number of Clusters

The most obvious way to perform a test using a CRVE t statistic based on (2) is to compare it with the Student's t distribution with $N - k$ degrees of freedom. However, it is well known that this procedure almost always overrejects. It is generally much better to use the $t(G-1)$ distribution, as suggested by Donald and Lang (2007) and Bester, Conley and Hansen (2011). However, it may be possible to do even better if the degrees-of-freedom parameter is chosen in a more sophisticated way.

Carter, Schnepel and Steigerwald (2017), hereafter referred to as CSS, proposes a method for estimating the "effective number" of clusters, G^*. This number depends in a fairly complicated way on the $\boldsymbol{X}_g$ matrices, the cluster sizes N_g, $g = 1, \ldots, G$, and a parameter ρ that measures within-cluster correlation. CSS focuses on the use of G^* as a diagnostic, suggesting that inference may be unreliable when G and G^* differ substantially. However, the paper also mentions, but does not investigate, the possibility of using critical values from the $t(G^*)$ distribution together with conventional CRVE t statistics for inference. By analogy with the recommendation of Donald and Lang (2007), it seems more natural to use the $t(G^* - 1)$ distribution. We investigate both procedures.

CSS suggest setting $\rho = 1$. Alternatively, one can estimate ρ, which can be done in several ways. One of them is to use a method that is standard in the literature on panel data; see Davidson and MacKinnon (2004, Section 7.10). Consider the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{u}, \tag{5}$$

where $\boldsymbol{D}$ is a matrix of cluster dummy variables. If s^2 denotes the usual OLS estimate of the error variance and $\hat{\sigma}_\eta^2$ is the sample variance of the elements of $\hat{\boldsymbol{\eta}}$, then a natural (but biased) estimate of ρ is $\hat{\rho} = \hat{\sigma}_\eta^2/(s^2 + \hat{\sigma}_\eta^2)$. This estimator is based on the assumption that the error terms are equicorrelated within each cluster with correlation ρ.

An alternative way to make the degrees-of-freedom parameter a function of the data, originally proposed by Bell and McCaffrey (2002), was modified by Imbens and Kolesár (2016), and yet another degrees-of-freedom correction has very recently been proposed by Young (2016). Simulation results in Cameron and Miller (2015) suggest that the Imbens and Kolesár degrees-of-freedom parameter can often be very similar to the G^* parameter of CSS. Bell and McCaffrey (2002) also proposed a modified CRVE similar in spirit to the HC2 heteroskedasticity-consistent covariance matrix estimator of MacKinnon and White (1985). Unfortunately, calculating this CRVE involves finding the inverse symmetric square roots of $N_g \times N_g$ matrices for $g = 1, \ldots, G$. Calculating the degrees-of-freedom parameter requires even more extensive computations. In view of the large sample sizes in most of our experiments (extremely large for the placebo laws experiments of Section 7 and the empirical example of Section 8), it was not feasible to study these procedures.[2] Another approach to inference in model (1) was suggested by Ibragimov and Müller (2010), but it requires that the coefficient of interest be separately identifiable from the data for each cluster, which is not the case in most of our experiments.

[2]For the state-size experiments of Section 4, in which $N = 2000$ and the largest N_g equals 242, simply calculating the Bell-McCaffrey CRVE would have taken about 17 times as much CPU time as bootstrapping with $B = 399$.

4 Simulation Design – Continuous Regressors

In this section, we use Monte Carlo simulation experiments to explore the implications of assumptions A1 and A3 for clustered data when the regressors are continuous. We study conventional inference based on the CRVE, wild bootstrap inference, and inference where the critical values depend on G^*.

There are four sets of simulations, two with 50 clusters and two with 100 clusters. In each case, one set has equal-sized clusters, while the other has clusters with sizes proportional to the U.S. states without the District of Columbia. Because 50 clusters satisfies the "rule of 42," we would expect to see reliable inference in all cases if the rule actually held.

The model we estimate is

$$y_{ig} = \beta_1 + \beta_2 X_{ig} + \epsilon_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \tag{6}$$

where there are G clusters, and the g^{th} cluster has N_g observations. Both the X_{ig} and the ϵ_{ig} are standard normal and uncorrelated across clusters. The within-cluster correlation is ρ_x for the X_{ig} and ρ_ϵ for the ϵ_{ig}. We do not allow ρ_ϵ to equal 1, but we do allow $\rho_x = 1$. In that case, the regressor is constant within each cluster, a situation that is commonly encountered in practice. We vary both ρ_x and ρ_ϵ across the simulations. Each simulated dataset has 2000 observations, and each experiment involves 400,000 replications.[3] For the wild cluster bootstrap, we use 399 bootstrap samples.

Each simulation proceeds as follows:

1. Specify $\rho_x \in \{0, 0.2, \ldots, 0.8, 1\}$ and $\rho_\epsilon \in \{0, 0.1, \ldots, 0.8, 0.9\}$.
2. For each simulated sample, generate X_{ig} and ϵ_{ig}, use equation (6) to compute y_{ig}, with $\beta_2 = 0$, and then estimate equation (6) by OLS.
3. Test the hypothesis that $\beta_2 = 0$, using either t tests based on the CRVE with several different degrees-of-freedom parameters or a wild bootstrap test.
4. Repeat the two preceding steps $400,000$ times, and estimate the rejection frequency of each test at the .05 level.

We compare the CRVE t statistic with six different critical values: $t(N-2)$, $t(G-1)$, $t(G^*-1)$ for two values of G^*, and $t(G^*)$ for two values of G^*. The two values of G^* are based on $\rho = 0.99$ (because CSS suggest using $\rho = 1$, which sometimes caused numerical problems) and $\rho = \hat{\rho}$, which was defined just after equation (5). For reasons of space, however, we report results only for $t(G-1)$ and $t(G^*-1)$ based on $\hat{\rho}$. Using $t(N-2)$ critical values always led to more severe overrejection than using $t(G-1)$ ones. In most cases, basing G^* on $\rho = \hat{\rho}$ worked better than basing it on $\rho = 0.99$.

Table 1 presents results for samples of 2000 observations spread equally across 50 clusters, so that assumption A3 holds. The reliability of CRVE inference with $t(G-1)$ critical values depends on both ρ_ϵ and ρ_x. When ρ_x is close to 0, rejection rates are close to the desired .05 level. As ρ_x gets closer to 1, however, they increase, always exceeding 0.065 when $\rho_x = 1$. In general, increasing ρ_ϵ increases rejection rates slightly, although to a lesser degree than increasing ρ_x. The impact is most severe when ρ_x is large but less than 1.

[3] Results for samples of 1000 observations (20 per cluster instead of 40), not reported, were almost identical to the ones for $G = 50$ reported here. These and other experiments suggest that, for all the simulation designs in this paper, rejection frequencies for given G are essentially invariant to N.

Table 1: Rejection Frequencies with 50 Equal-Sized Clusters

		ρ_x					
ρ_ϵ		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**t(G−1)**	0.04978	0.05109	0.05280	0.05653	0.06004	0.06568
	t(G*−1)	0.04948	0.04961	0.04911	0.05001	0.05096	0.05357
	bootstrap	0.04946	0.05003	0.04989	0.05027	0.04999	0.04980
0.1	**t(G−1)**	0.05057	0.05211	0.05427	0.05723	0.06103	0.06655
	t(G*−1)	0.05027	0.05031	0.04982	0.05022	0.05121	0.05439
	bootstrap	0.05032	0.05014	0.05023	0.04995	0.04988	0.05043
0.2	**t(G−1)**	0.05020	0.05235	0.05606	0.06062	0.06388	0.06590
	t(G*−1)	0.04987	0.04965	0.05033	0.05188	0.05320	0.05390
	bootstrap	0.04967	0.04978	0.05008	0.05069	0.05046	0.04980
0.3	**t(G−1)**	0.05060	0.05292	0.05735	0.06124	0.06418	0.06581
	t(G*−1)	0.05016	0.04865	0.05022	0.05164	0.05302	0.05346
	bootstrap	0.04991	0.04941	0.05045	0.05003	0.05006	0.04955
0.4	**t(G−1)**	0.05102	0.05400	0.05797	0.06205	0.06467	0.06642
	t(G*−1)	0.05041	0.04816	0.04932	0.05166	0.05301	0.05395
	bootstrap	0.05014	0.04976	0.04980	0.05017	0.05003	0.05004
0.5	**t(G−1)**	0.05061	0.05534	0.05871	0.06251	0.06506	0.06558
	t(G*−1)	0.04961	0.04806	0.04915	0.05184	0.05334	0.05360
	bootstrap	0.05014	0.05084	0.05007	0.05009	0.05004	0.04963
0.6	**t(G−1)**	0.05087	0.05518	0.05965	0.06293	0.06535	0.06645
	t(G*−1)	0.04914	0.04677	0.04916	0.05158	0.05344	0.05412
	bootstrap	0.05004	0.05063	0.05019	0.04999	0.05012	0.05024
0.7	**t(G−1)**	0.05106	0.05469	0.05936	0.06274	0.06486	0.06606
	t(G*−1)	0.04825	0.04509	0.04824	0.05112	0.05285	0.05379
	bootstrap	0.05016	0.04989	0.04961	0.04984	0.04953	0.04961
0.8	**t(G−1)**	0.05101	0.05498	0.05982	0.06443	0.06479	0.06549
	t(G*−1)	0.04634	0.04446	0.04861	0.05235	0.05237	0.05326
	bootstrap	0.05009	0.04988	0.05011	0.05077	0.04948	0.04924
0.9	**t(G−1)**	0.05030	0.05525	0.05996	0.06353	0.06555	0.06620
	t(G*−1)	0.04322	0.04411	0.04840	0.05161	0.05349	0.05414
	bootstrap	0.04984	0.05018	0.05046	0.05020	0.05046	0.05041

Notes: Rejection frequencies are at the .05 level and are based on 400,000 replications. There are 50 equal-sized clusters with a total of 2000 observations. The effective number of clusters is $G^*(\hat{\rho})$. The wild bootstrap uses the Rademacher distribution with 399 bootstraps. `[corrected 2019-08]`

Using critical values from the $t(G^* - 1)$ distribution frequently results in more accurate inferences, but there is a tendency to overreject when ρ_x is large and to underreject when ρ_ϵ is large. Using $t(G^*)$ instead of $t(G^* - 1)$ makes the overrejection in the former case more severe and the underrejection in the latter case less severe. When $\rho_x = 1$, G^* is invariant to the value of ρ. Except in this case, setting $\rho = 0.99$ results in (often substantially) lower values of G^* than using $\hat{\rho}$, which tends to cause noticeable underrejection.

Table 2: Rejection Frequencies with 50 State-Sized Clusters

		ρ_x					
ρ_ϵ		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**t(G−1)**	0.05784	0.06019	0.06305	0.06710	0.07273	0.08191
	t(G*−1)	0.05035	0.03486	0.02875	0.02652	0.02567	0.02523
	bootstrap	0.04905	0.05001	0.05113	0.05080	0.05092	0.05060
0.1	**t(G−1)**	0.05903	0.06404	0.06996	0.07792	0.08497	0.09338
	t(G*−1)	0.05128	0.03559	0.03034	0.02881	0.02942	0.02820
	bootstrap	0.05016	0.05092	0.05085	0.05176	0.05176	0.05124
0.2	**t(G−1)**	0.05820	0.06962	0.07826	0.08665	0.09597	0.10113
	t(G*−1)	0.05063	0.03506	0.03083	0.03055	0.03204	0.02938
	bootstrap	0.04935	0.05150	0.05142	0.05162	0.05242	0.05106
0.3	**t(G−1)**	0.05892	0.07377	0.08341	0.09179	0.09926	0.10434
	t(G*−1)	0.05077	0.03280	0.02977	0.02964	0.03060	0.02944
	bootstrap	0.05006	0.05197	0.05158	0.05177	0.05164	0.05182
0.4	**t(G−1)**	0.05907	0.07510	0.08549	0.09506	0.10086	0.10506
	t(G*−1)	0.05089	0.02946	0.02770	0.02874	0.03003	0.02930
	bootstrap	0.05006	0.05133	0.05145	0.05236	0.05181	0.05125
0.5	**t(G−1)**	0.05937	0.07712	0.08704	0.09589	0.10226	0.10619
	t(G*−1)	0.05029	0.02705	0.02645	0.02770	0.02954	0.02904
	bootstrap	0.05057	0.05216	0.05189	0.05196	0.05204	0.05164
0.6	**t(G−1)**	0.05854	0.07778	0.08847	0.09750	0.10316	0.10606
	t(G*−1)	0.04843	0.02476	0.02579	0.02761	0.02932	0.02870
	bootstrap	0.05006	0.05203	0.05259	0.05219	0.05216	0.05110
0.7	**t(G−1)**	0.05876	0.07858	0.08983	0.09650	0.10387	0.10761
	t(G*−1)	0.04651	0.02313	0.02494	0.02697	0.02871	0.02872
	bootstrap	0.05098	0.05188	0.05267	0.05187	0.05220	0.05172
0.8	**t(G−1)**	0.05780	0.07869	0.09019	0.09789	0.10440	0.10776
	t(G*−1)	0.04285	0.02200	0.02433	0.02658	0.02854	0.02882
	bootstrap	0.05096	0.05232	0.05277	0.05240	0.05187	0.05198
0.9	**t(G−1)**	0.05668	0.07866	0.08857	0.09799	0.10474	0.10718
	t(G*−1)	0.03754	0.02103	0.02325	0.02612	0.02868	0.02826
	bootstrap	0.05133	0.05222	0.05178	0.05192	0.05219	0.05113

Notes: See notes to Table 1. There are 50 clusters proportional to US state populations with a total of 2000 observations. `[Corrected 2019-08]`

Even better results are obtained by using the wild bootstrap, which for all practical purposes performs perfectly. The smallest rejection frequency out of the 60 reported in Table 1 is 0.04924, and the largest is 0.05084. These numbers are close enough to .05 to be explained by chance. Since the standard error when the true rejection frequency is .05 is 0.000345, the largest implied t statistic for the hypothesis that the rejection frequency is 0.05 is just 2.43.

Table 3: Rejection Frequencies with 100 Equal-Sized Clusters

		ρ_x					
ρ_ϵ		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	**t(G−1)**	0.05029	0.05099	0.05184	0.05392	0.05581	0.05770
	t(G*−1)	0.05002	0.05009	0.04975	0.05067	0.05132	0.05230
	bootstrap	0.04990	0.05008	0.05020	0.05065	0.05032	0.04952
0.1	**t(G−1)**	0.05003	0.05110	0.05233	0.05393	0.05583	0.05893
	t(G*−1)	0.04977	0.05007	0.05017	0.05051	0.05123	0.05295
	bootstrap	0.04947	0.05006	0.05012	0.05019	0.05015	0.05038
0.2	**t(G−1)**	0.05057	0.05064	0.05315	0.05413	0.05577	0.05851
	t(G*−1)	0.05026	0.04934	0.05055	0.05031	0.05098	0.05269
	bootstrap	0.05004	0.04928	0.05020	0.04936	0.04928	0.05047
0.3	**t(G−1)**	0.05068	0.05202	0.05366	0.05590	0.05762	0.05921
	t(G*−1)	0.05028	0.05031	0.05052	0.05160	0.05246	0.05330
	bootstrap	0.05017	0.05035	0.05008	0.05047	0.05051	0.05051
0.4	**t(G−1)**	0.05104	0.05140	0.05394	0.05651	0.05717	0.05821
	t(G*−1)	0.05057	0.04914	0.05018	0.05155	0.05189	0.05207
	bootstrap	0.05062	0.04918	0.04997	0.05042	0.04984	0.04968
0.5	**t(G−1)**	0.05053	0.05201	0.05421	0.05676	0.05769	0.05814
	t(G*−1)	0.04986	0.04905	0.04996	0.05184	0.05231	0.05228
	bootstrap	0.05006	0.04939	0.04959	0.05046	0.05013	0.04984
0.6	**t(G−1)**	0.05095	0.05276	0.05426	0.05650	0.05786	0.05844
	t(G*−1)	0.05003	0.04917	0.04955	0.05134	0.05239	0.05264
	bootstrap	0.05039	0.05004	0.04944	0.04988	0.04999	0.04997
0.7	**t(G−1)**	0.05106	0.05241	0.05420	0.05708	0.05790	0.05783
	t(G*−1)	0.04936	0.04820	0.04907	0.05167	0.05220	0.05220
	bootstrap	0.05039	0.04974	0.04919	0.05021	0.04991	0.04951
0.8	**t(G−1)**	0.05044	0.05309	0.05565	0.05730	0.05795	0.05888
	t(G*−1)	0.04802	0.04836	0.05032	0.05162	0.05220	0.05300
	bootstrap	0.04974	0.05032	0.05060	0.05049	0.05009	0.05027
0.9	**t(G−1)**	0.05119	0.05291	0.05622	0.05664	0.05803	0.05797
	t(G*−1)	0.04763	0.04795	0.05057	0.05087	0.05231	0.05213
	bootstrap	0.05063	0.05015	0.05091	0.04953	0.05018	0.04981

Notes: See notes to Table 1. There are 100 equal-sized clusters with a total of 2000 observations. `[Corrected 2019-08]`

Since assumption A3 holds, it is not surprising that the "rule of 42" nearly holds in these simulations. Table 2 presents results from a second set of experiments in which that assumption is severely violated. Cluster sizes are now roughly proportional to U.S. state populations; the smallest clusters have just 4 observations, and the largest has 242. Even when ρ_ϵ and ρ_x are 0, the rejection rate is nearly 0.06. At the other extreme, when $\rho_\epsilon = 0.9$ and $\rho_x = 1$, the rejection rate is 0.1073. Increasing ρ_x leads to an increase in rejection rates.

Table 4: Rejection Frequencies with 100 State-Sized Clusters

		ρ_x					
ρ_ϵ		**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
0.0	$\mathbf{t(G-1)}$	0.05497	0.05614	0.05743	0.05943	0.06355	0.06864
	$\mathbf{t(G^*-1)}$	0.05101	0.03951	0.03370	0.03221	0.03190	0.03116
	bootstrap	0.05007	0.05076	0.05042	0.05005	0.05043	0.05083
0.1	$\mathbf{t(G-1)}$	0.05469	0.05695	0.06036	0.06431	0.06812	0.07321
	$\mathbf{t(G^*-1)}$	0.05071	0.03948	0.03563	0.03385	0.03375	0.03328
	bootstrap	0.04999	0.05028	0.05061	0.05066	0.05012	0.05008
0.2	$\mathbf{t(G-1)}$	0.05483	0.06022	0.06413	0.06948	0.07437	0.07857
	$\mathbf{t(G^*-1)}$	0.05088	0.04025	0.03620	0.03553	0.03603	0.03506
	bootstrap	0.05001	0.05080	0.05024	0.05009	0.05048	0.05053
0.3	$\mathbf{t(G-1)}$	0.05532	0.06187	0.06833	0.07327	0.07795	0.08153
	$\mathbf{t(G^*-1)}$	0.05119	0.03927	0.03656	0.03615	0.03674	0.03579
	bootstrap	0.05014	0.05035	0.05109	0.05073	0.05082	0.05059
0.4	$\mathbf{t(G-1)}$	0.05428	0.06359	0.06972	0.07556	0.07909	0.08266
	$\mathbf{t(G^*-1)}$	0.04983	0.03723	0.03580	0.03608	0.03642	0.03537
	bootstrap	0.04957	0.05038	0.05107	0.05097	0.05040	0.05035
0.5	$\mathbf{t(G-1)}$	0.05505	0.06413	0.07018	0.07640	0.07987	0.08279
	$\mathbf{t(G^*-1)}$	0.05018	0.03545	0.03411	0.03525	0.03595	0.03581
	bootstrap	0.05018	0.05046	0.05042	0.05124	0.05033	0.05013
0.6	$\mathbf{t(G-1)}$	0.05560	0.06512	0.07079	0.07676	0.08090	0.08438
	$\mathbf{t(G^*-1)}$	0.04981	0.03385	0.03269	0.03426	0.03589	0.03628
	bootstrap	0.05050	0.05042	0.04970	0.05034	0.05055	0.05105
0.7	$\mathbf{t(G-1)}$	0.05561	0.06444	0.07177	0.07702	0.08106	0.08368
	$\mathbf{t(G^*-1)}$	0.04898	0.03152	0.03204	0.03359	0.03531	0.03537
	bootstrap	0.05084	0.04993	0.05024	0.05014	0.05070	0.05059
0.8	$\mathbf{t(G-1)}$	0.05524	0.06568	0.07252	0.07775	0.08164	0.08393
	$\mathbf{t(G^*-1)}$	0.04717	0.03077	0.03155	0.03323	0.03460	0.03540
	bootstrap	0.05051	0.05018	0.05052	0.05076	0.05073	0.05042
0.9	$\mathbf{t(G-1)}$	0.05468	0.06609	0.07304	0.07749	0.08161	0.08442
	$\mathbf{t(G^*-1)}$	0.04433	0.02906	0.03108	0.03304	0.03433	0.03587
	bootstrap	0.05112	0.05047	0.05085	0.05099	0.05023	0.05093

Notes: See notes to Table 1. There are 100 clusters proportional to US state populations with a total of 2000 observations. `[Corrected 2019-08]`

So does increasing ρ_ϵ, except when $\rho_x = 0$. Thus, with even modest amounts of intra-cluster correlation, the "rule of 42" fails to hold in these experiments.

With state-sized clusters, using $t(G^* - 1)$ critical values generally results in underrejection, which is quite severe when ρ_x is large and generally becomes worse as ρ_ϵ gets larger. The underrejection is even more severe when G^* is based on $\rho = 0.99$ instead of $\rho = \hat{\rho}$. In the worst case ($\rho_\epsilon = 0$, $\rho_x = 0.6$), the rejection rate (not shown in the table) is just 0.0166. In this case, the average value of $G^*(\hat{\rho})$ is 9.43, while the average value of $G^*(0.99)$ is 5.97. Both these numbers seem to be unrealistically low. In cases such as these, using $t(G^*)$ instead of $t(G^* - 1)$ reduces the underrejection only modestly.

As before, much better inferences are obtained by using the wild bootstrap, although it does not work quite as well as with equal-sized clusters. Rejection frequencies range from 0.04905 to 0.05277. There is thus a very modest tendency to overreject in some cases, which would have been impossible to detect if we had not used such a large number of replications.

In order to investigate assumption A1, we repeated both sets of experiments using 100 clusters instead of 50, holding the sample size constant at 2000. Table 3 reports results for 100 equal-sized clusters. The wild bootstrap works perfectly, except for simulation error. The other methods work better than they did with 50 clusters, but they tend to overreject or underreject in the same cases. Table 4 reports results for 100 clusters that are roughly proportional to U.S. state populations, with each state appearing twice. The two smallest clusters have just 2 observations, and the two largest have 121. In effect, there are two Californias, two Delawares, and so on. All methods work better than they did in Table 2, but CRVE rejection frequencies can still be greater than 0.084. The wild cluster bootstrap rejection rate never exceeds 0.05124, although it does overreject more often than it underrejects.

These results demonstrate that inference based on the CRVE may not be reliable when cluster sizes differ substantially. Comparing G^* with G seems to provide valuable evidence that inference based on $t(G-1)$ may be unreliable, but using critical values from the $t(G^*-1)$ distribution does not always solve the problem. The most reliable approach in this setting, especially when G^* is small, is apparently to use the wild cluster bootstrap.

5 Simulation Design – Treatment Effects

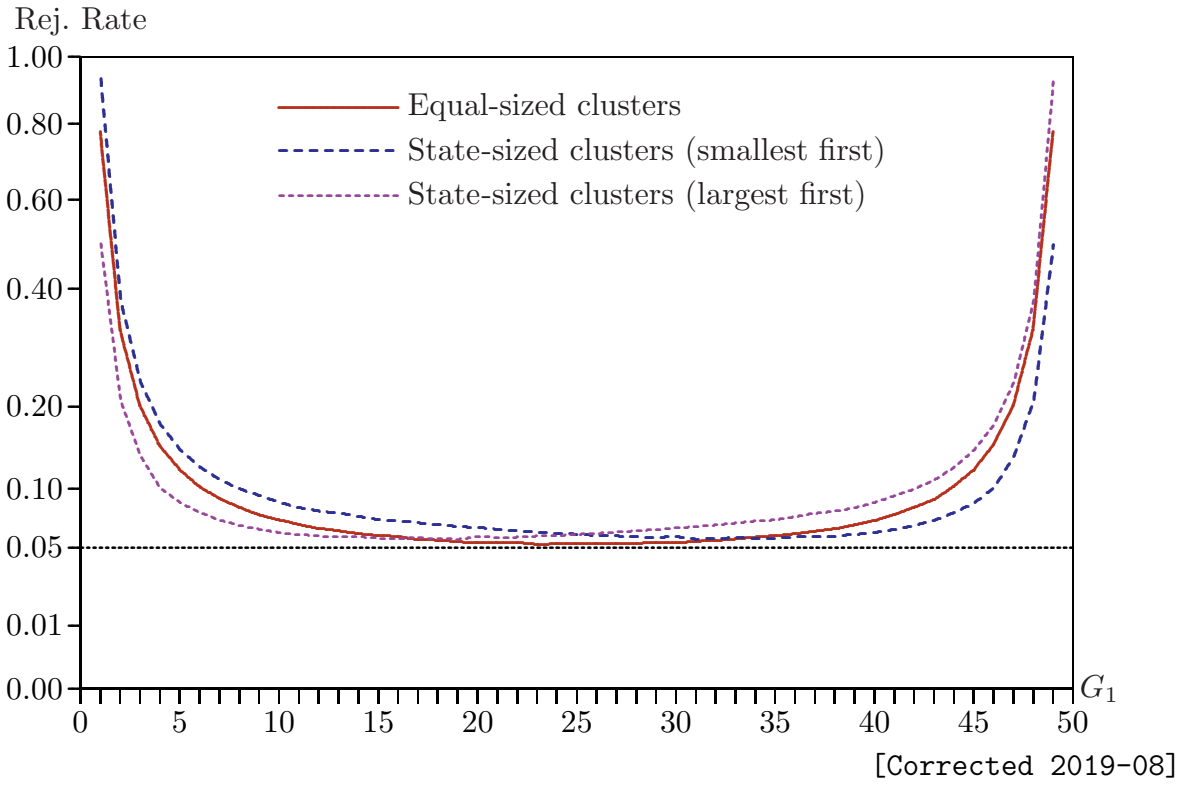
Many applications to clustered data involve treatment effects, either at the cluster level or by time period within some clusters. We conducted three sets of experiments to investigate this type of application. Results for one set are reported here, and results for a second set are reported in Section A.2 of the appendix. The third set produced results that are not reported because they were visually indistinguishable from those of the first set.

In many empirical studies, only some observations in some clusters are treated. If i indexes individuals, g indexes jurisdictions, such as states, and t indexes time periods, then a classic "difference in differences" (or "DiD") regression can be written as

$$y_{igt} = \beta_1 + \beta_2 \mathrm{GT}_{igt} + \beta_3 \mathrm{PT}_{igt} + \beta_4 \mathrm{GT}_{igt}\mathrm{PT}_{igt} + \epsilon_{igt}, \tag{7}$$

for $i = 1, \ldots, N_g$, $g = 1, \ldots, G$, and $t = 1, \ldots, T$. Here GT_{igt} is a "cluster treated" dummy that equals 1 if cluster g is ever treated, and PT_{igt} is a "period treated" dummy that equals 1 if treatment occurs in time period t. The coefficient of most interest is β_4, which shows the

Figure 1: Rejection rates and proportion treated, DiD, $t(G-1)$)



effect on treated clusters in periods when there is treatment. We let G_1 denote the number of treated clusters and M_1 denote the number of treated observations, that is, observations for which $\mathrm{GT}_{igt}\mathrm{PT}_{igt} = 1$.

In the DiD regression model, there is no role for ρ_x, and ρ_ϵ usually seems to have very little effect on rejection frequencies; we set the latter to 0.05 in all the experiments.[4] As we show in Section 6, what matters is G_1, the number of treated groups. In Figures 1, 2, and 3, we report results for 50 clusters with 2000 observations. The treatments are applied to state-sized clusters both from smallest to largest and from largest to smallest. The simulations used $400,000$ replications.

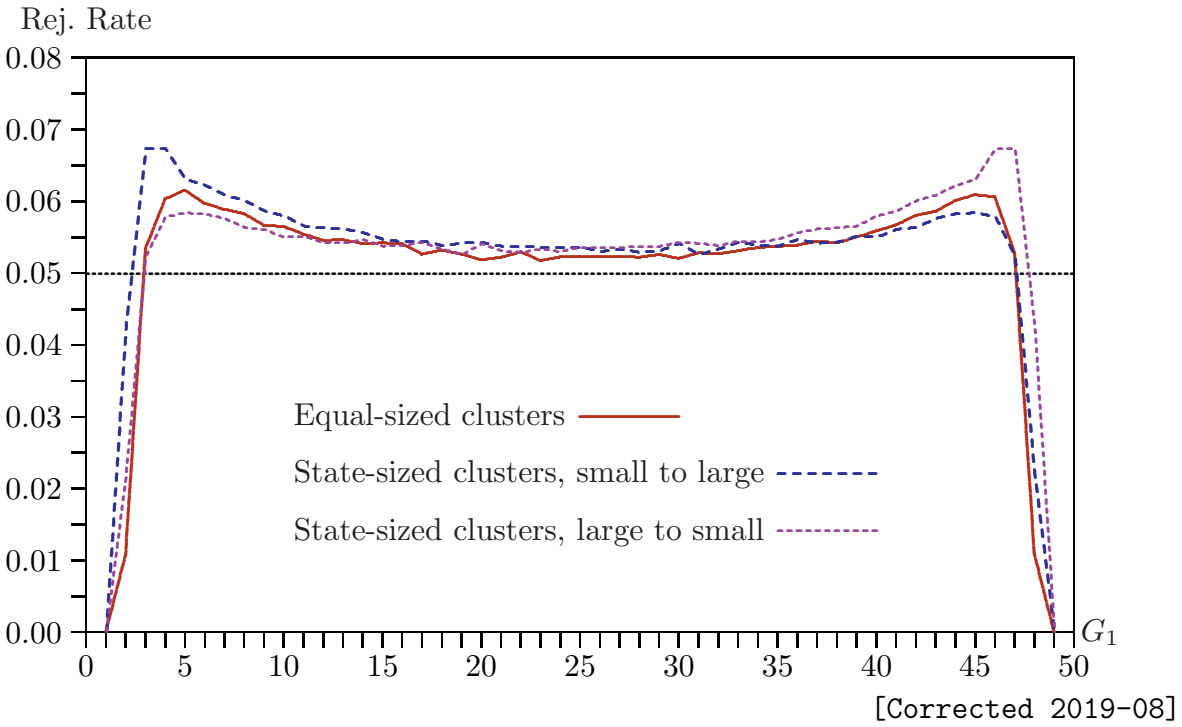
Figure 1 shows results for $t(G-1)$ critical values. It is based on equation (7) with either zero or half of the observations in each cluster treated.[5] Thus $\mathrm{PT}_{igt} = 1$ for half the observations in each cluster, while $\mathrm{GT}_{igt} = 1$ for G_1 clusters, with G_1 varying between 1 and 49. The vertical axis has been subjected to a square root transformation in order to accommodate both large and small rejection frequencies.

There is very severe overrejection in Figure 1 when either G_1 or $G-G_1$ is close to 0. This result is consistent with Monte Carlo results in Bell and McCaffrey (2002) and Conley and Taber (2011), and we explain the result in Section 6. The latter paper develops procedures for inference when there are just a few treated groups; see also MacKinnon and Webb (2020),

[4] In early versions of the paper, we set $\rho_\epsilon = 0.5$, a number that is unrealistically high.

[5] In exploratory experiments with fewer simulations, very similar results were obtained when either one quarter or three quarters of the observations in each cluster were treated.

Figure 2: Rejection rates and proportion treated, DiD, $t(G^* - 1)$



which studies those procedures and develops additional ones. A very different procedure for inference when there is only one treated group has been proposed by Abadie, Diamond and Hainmueller (2010) based on the idea of "synthetic controls."

With equal-sized clusters, rejection frequencies are very close to 0.05 for G_1 between 17 and 34. With state-sized clusters, they tend to be somewhat higher. The graph for equal-sized clusters is symmetric around $G_1 = 25$, while the ones for state-sized clusters are quite asymmetric. For small values of G_1, there is very severe overrejection when the smallest clusters are treated first. For large values of G_1, there is still serious overrejection, but it is considerably less severe.

The results when the largest clusters are treated first are the mirror image of the results when the smallest clusters are treated first. This must be the case, because the absolute value of the t statistic for $\beta_4 = 0$ in regression (7) is the same when the number of clusters treated is G_1 as it is when that number is $G - G_1$. We may conclude that overrejection tends to be most severe when $\min(G_1, G - G_1)$ is small and the observations that are in the minority are from the smallest clusters. These results will be explained in Section 6.

Figure 2 shows results for $t(G^* - 1)$ critical values. In all cases, G^* is invariant to ρ. Except when the number of treated clusters is very small or very large, the tests always overreject, but they never do so severely. There is extreme underrejection when $G_1 = 1$ and $G_1 = 49$, because G^* is not much greater than 1 in those cases, and the Student's t distribution has extremely long tails when the degrees of freedom parameter is very close to zero. Rejection frequencies are extremely sensitive to that parameter; using critical values based on $t(G^*)$ instead of $t(G^* - 1)$ can lead to severe overrejection.

Figure 3: Rejection rates and proportion treated, DiD, wild bootstrap

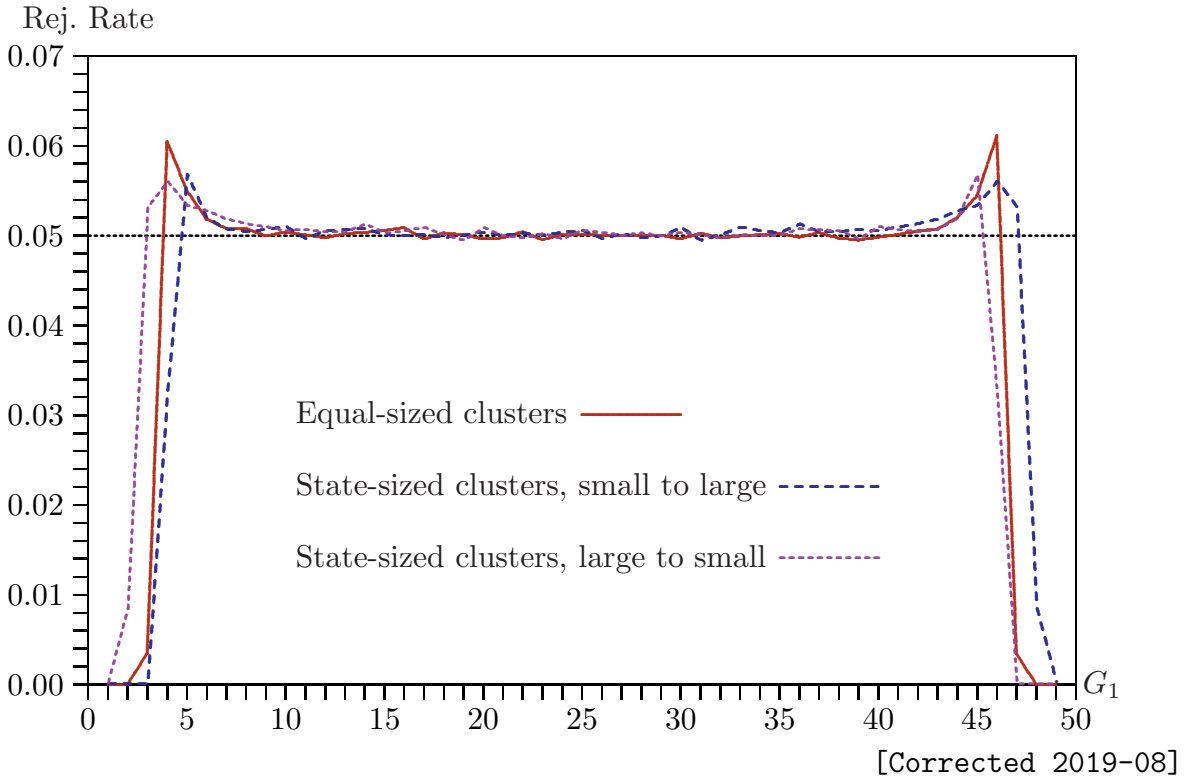


Figure 3 shows results for wild bootstrap tests based on simulations with 399 bootstraps. In all cases, there is severe underrejection when either G_1 or $G - G_1$ is very small. For equal-sized clusters, there is modest overrejection when G_1 and $G - G_1$ are both either 4 or 5, but the wild bootstrap tests work extremely well for G_1 between about 7 and 43. For state-sized clusters, the range of underrejection is narrower (wider) for small G_1, and wider (narrower) for large G_1, when the largest (smallest) clusters are treated first. All of these results will be explained in Section 6.

It is common to allow for cluster fixed effects in DiD models by dropping the constant term and the GT_{ig} variable and adding G dummy variables, one for each cluster. This is necessary if treatment occurs at different times for different clusters. We performed a second set of experiments for the fixed-effects DiD model and obtained results that were visually indistinguishable from those in Figures 1 through 3 and are therefore not reported.

Another possibility is a pure treatment model, where there is no time dimension and the test regressor is an indicator variable that equals 1 for some number of clusters. Thus, for each cluster, either all observations are treated or all are not treated. We performed a third set of experiments for this model and obtained results that are similar to, but do differ in some respects from, those in Figures 1 through 3. They are reported in Section A.2 of the appendix. We also report some additional simulation results for the DiD case with between 12 and 32 clusters and varying degrees of inequality in cluster sizes in Section A.6. Even with very small numbers of clusters, the (restricted) wild cluster bootstrap seems to perform well provided that neither G_1 nor $G - G_1$ is too small.

6 Why CRVE t Tests and Wild Bootstrap Tests Can Fail

As we have seen, t tests based on CRVE standard errors tend to overreject, often very severely, when the number of treated clusters is very small or very large, and wild bootstrap tests based on restricted residuals tend to underreject just as severely. In this section, we explain these phenomena. For simplicity, we focus on the pure treatment model

$$y_{ig} = \beta_1 + \beta_2 d_{ig} + \epsilon_{ig}, \tag{8}$$

where d_{ig} equals 1 for the first G_1 clusters and 0 for the remaining $G_0 = G - G_1$ clusters. Including additional regressors, or allowing only some observations within treated clusters to be treated, would not change the analysis in any fundamental way.

We make the simplifying assumption that G and N tend to infinity at the same rate, which implies that the N_g do not grow or shrink systematically as N increases. CRVE-based inference is known to be valid under weaker assumptions that allow G to grow more slowly than N, but not too slowly; see CSS. Since the focus of this section is on cases where CRVE-based inference fails despite a relatively strong assumption, there is no reason to relax that assumption. Moreover, many of our results do not require any asymptotic assumptions. They hold, at least qualitatively, in finite samples.

Standard asymptotic analysis is based on the assumption that the limit of $\phi \equiv G_1/G$ as $N \to \infty$ is strictly between 0 and 1. An alternative assumption, which is much more realistic when G_1 is small, is that G_1 is fixed as $N \to \infty$. This implies that $\phi \to 0$. This assumption allows us to explain all the simulation results in Figures 1 and 3, as well as ones reported later in this section, in Section 7, and in the appendix.

Equation (8) may be rewritten in vector notation as $\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{d} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$, $\boldsymbol{\iota}$, $\boldsymbol{d}$, and $\boldsymbol{\epsilon}$ are N-vectors with typical elements y_{ig}, 1, d_{ig}, and ϵ_{ig}, respectively, and i varies more rapidly than g. Then the OLS estimator of β_2 is

$$\hat{\beta}_2 = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{y}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})} = \beta_2 + \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}, \tag{9}$$

where $\bar{d}$ denotes the sample mean of the d_{ig}. For convenience, let M_1 denote the number of treated observations and $M_0 = N - M_1$ denote the number of untreated observations, so that $\bar{d} = M_1/N$. For any finite N, $\bar{d}$ will generally differ from ϕ unless $N_g = N/G$ for all g, but, like ϕ, it must be $O(N^{-1})$ when G_1 is fixed.

The estimator $\hat{\beta}_2$ is inconsistent under the assumption that G_1 is fixed. Since the denominator of $\hat{\beta}_2$ is $\bar{d}(1-\bar{d})N = M_1 M_0/N$, we can use equation (9) to obtain

$$\hat{\beta}_2 - \beta_2 = \frac{1}{M_1} \sum_{g=1}^{G_1} \sum_{i=1}^{N_g} \epsilon_{ig} - \frac{1}{M_0} \sum_{g=G_1+1}^{G} \sum_{i=1}^{N_g} \epsilon_{ig}. \tag{10}$$

The first term in (10) is $1/M_1$ times a summation of M_1 mean-zero random variables. Thus it is $O_p\big(M_1^{-1/2}\big) = O_p(1)$. The second term is $1/M_0$ times a summation of $M_0 = N - M_1$ mean-zero random variables. Thus it is $O_p(N^{-1/2})$. We conclude that the first term, even though it depends on just $M_1 = O(1)$ observations, must be larger asymptotically than the

second term. Thus a fundamental condition for consistency, namely, that the impact of any one observation on $\hat{\beta}_2$ must tend to zero as $N \to \infty$, is violated.

From expression (2), it is evident that the CRVE for $\hat{\beta}_2$ is proportional to

$$\frac{\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)}{\Big((\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})\Big)^2}. \tag{11}$$

The CRVE would provide a good estimate of $\text{Var}(\hat{\beta}_2)$ if its numerator provided a good estimate of $(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})$. However, it is impossible to estimate $\text{Var}(\hat{\beta}_2)$ consistently when G_1 is fixed; this is a special case of Corollary 1 in CSS.

Equations (9) and (11) imply that, under the null hypothesis, the t statistic for $\beta_2 = 0$ is

$$t_2 = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{\Big(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\Big)^{1/2}}, \tag{12}$$

where c is the square root of $\Big((G-1)(N-2)\Big)/\Big(G(N-1)\Big)$. This test statistic has variance very much greater than that of the $t(G-1)$ distribution when G_1 (or G_0) is small. The reason is that its denominator, which is the square root of the numerator of expression (11), is not only inconsistent but also severely biased downwards in that case.

In scalar notation, $1/c$ times the numerator of expression (11) can be written as

$$(1 - \bar{d})^2 \sum_{g=1}^{G_1}\left(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\right)^2 + \bar{d}^2 \sum_{g=G_1+1}^{G}\left(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\right)^2. \tag{13}$$

This expression is supposed to estimate the quantity

$$\begin{aligned}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota}) &= (1 - \bar{d})^2 \sum_{g=1}^{G_1}\boldsymbol{\iota}'\boldsymbol{\Omega}_g\boldsymbol{\iota} + \bar{d}^2 \sum_{g=G_1+1}^{G}\boldsymbol{\iota}'\boldsymbol{\Omega}_g\boldsymbol{\iota} \\ &= (1 - \bar{d})^2 \sum_{g=1}^{G_1}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\omega_{ij}^g + \bar{d}^2 \sum_{g=G_1+1}^{G}\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\omega_{ij}^g,\end{aligned} \tag{14}$$

where ω_{ij}^g denotes the ij^{th} element of the covariance matrix $\boldsymbol{\Omega}_g$. When $G_1 = 1$, the residuals for cluster 1 must sum to zero, because $d_{i1} = 1$ and $d_{ig} = 0$ for $g > 1$. This implies that the first term in expression (13) equals zero. In contrast, the first term in expression (14) is not zero. It will generally be quite large relative to the second term, because $\bar{d}^2$ must be very small when $M_1 \ll M_0$. We conclude that (13) will always underestimate (14) when $G_1 = 1$. For large G, the underestimation must be severe, because the first term in (14) is $O(1)$ and the second is $O(N^{-1})$. Because $\bar{d}$ depends on the size of the treated clusters, the overrejection will tend to be more severe when the smallest clusters are treated than when the largest clusters are treated, which is exactly what we see in Figure 1.

When two or more clusters are treated, the residuals for each treated cluster will not sum to zero, but they must sum to zero over all the treated clusters. Thus the sum of squared summations in the first term of (13) will always underestimate the corresponding double

summation in (14). In Section A.4 of the appendix, we show that, when the errors are IID, the expectation of the squared summation for the first treated cluster underestimates the corresponding true variance by a factor of $(M_1 - N_1)/M_1$. Thus the underestimation should go away as G_1 increases. In typical cases where N_1/M_1 tends to zero fairly rapidly, it does indeed go away quite quickly, as Figure 1 illustrates.

As Figure 3 shows, the restricted wild cluster bootstrap does not solve this problem. For the dummy variable regression (8), the restricted wild cluster bootstrap DGP (3) is

$$y_{ig}^{*j} = \tilde{\beta}_1 + \tilde{\epsilon}_{ig} v_g^{*j}, \tag{15}$$

where $\tilde{\beta}_1$ is the sample mean of the y_{ig}, and $\tilde{\epsilon}_{ig} = y_{ig} - \tilde{\beta}_1$. The OLS estimate $\hat{\beta}_2^{*j}$ for the j^{th} bootstrap sample is then given by equation (9), with $\boldsymbol{\epsilon}$ replaced by $\boldsymbol{\epsilon}^{*j}$, a vector with typical element $\tilde{\epsilon}_{ig} v_g^{*j}$, and the bootstrap t statistic is

$$t_2^{*j} = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})' \boldsymbol{\epsilon}^{*j}}{\left(\sum_{g=1}^{G} (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)' \hat{\boldsymbol{\epsilon}}_g^{*j} (\hat{\boldsymbol{\epsilon}}_g^{*j})' (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g) \right)^{1/2}}, \tag{16}$$

where $\hat{\boldsymbol{\epsilon}}_g^{*j}$ is the subvector of OLS residuals for cluster g and bootstrap sample j, and c is the square root of $\Big((G-1)(N-2)\Big)/\Big(G(N-1)\Big)$.

Now consider the extreme case in which $G_1 = 1$. The numerator of (16) becomes

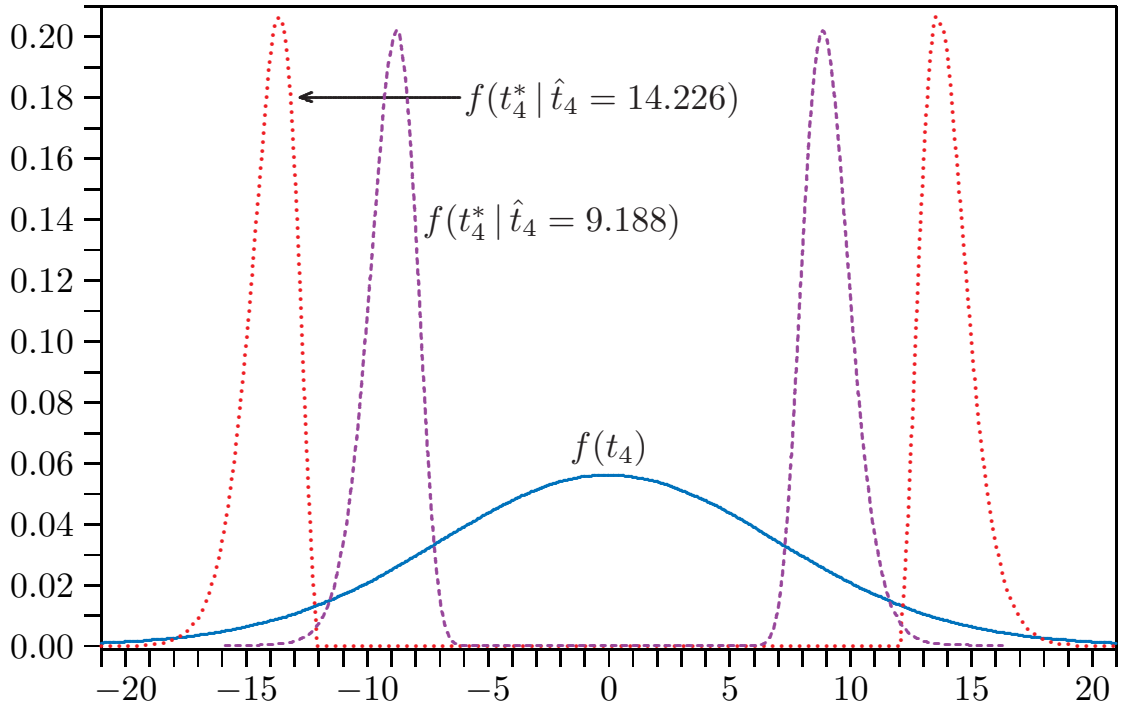
$$c(1 - \bar{d}) \sum_{i=1}^{N_1} \epsilon_{i1}^{*j} - c\bar{d} \sum_{g=2}^{G} \sum_{i=1}^{N_g} \epsilon_{ig}^{*j}. \tag{17}$$

Because $\bar{d} = N_1/N$, the first term in expression (17) must be the dominant one asymptotically, and also in finite samples unless N_1 is extraordinarily large. For the Rademacher distribution, the vectors of bootstrap error terms for $g = 1$ can have just two values, namely, $\tilde{\boldsymbol{\epsilon}}_1$ and $-\tilde{\boldsymbol{\epsilon}}_1$. Therefore, when $\boldsymbol{\epsilon}_1$ is such that $|t_2|$ is large, $\tilde{\boldsymbol{\epsilon}}_1^{*j}$ will probably be such that $|t_2^{*j}|$ is large. The distribution of the bootstrap t statistics t_2^{*j} is then bimodal, with half the realizations in the neighborhood of t_2 and the other half in the neighborhood of $-t_2$. Thus when $|t_2|$ is large, $|t_2^{*j}|$ tends to be even larger for a substantial number of bootstrap samples, making it rare to obtain a bootstrap P value below any specified level for a test. This explains why the wild cluster bootstrap underrejects severely when $G_1 = 1$.[6]

Figure 4 illustrates this phenomenon. The solid line is a kernel estimate based on five million replications of the empirical density of the statistic t_4 for the DiD model of equation (7) when $G_1 = 1$; recall that t_4 in the DiD model (7) plays the same role as t_2 in the pure treatment model (8). There are 2000 observations and 50 equal-sized clusters. This density is extremely dispersed. The 0.975 quantile is 14.226, whereas the one for the $t(49)$ distribution is 2.010. Not surprisingly, the CRVE $t(G-1)$ test rejects 77.75% of the time at the .05 level.

[6] This argument implicitly assumes that the distributions of t_2 and the t_2^{*j} are similar, which may not be true if the disturbances for the treated and untreated observations have very different variances. In such cases, the wild cluster bootstrap would almost certainly not work well, but it might not always underreject.

Figure 4: Densities of actual and bootstrap t statistics when $G_1 = 1$



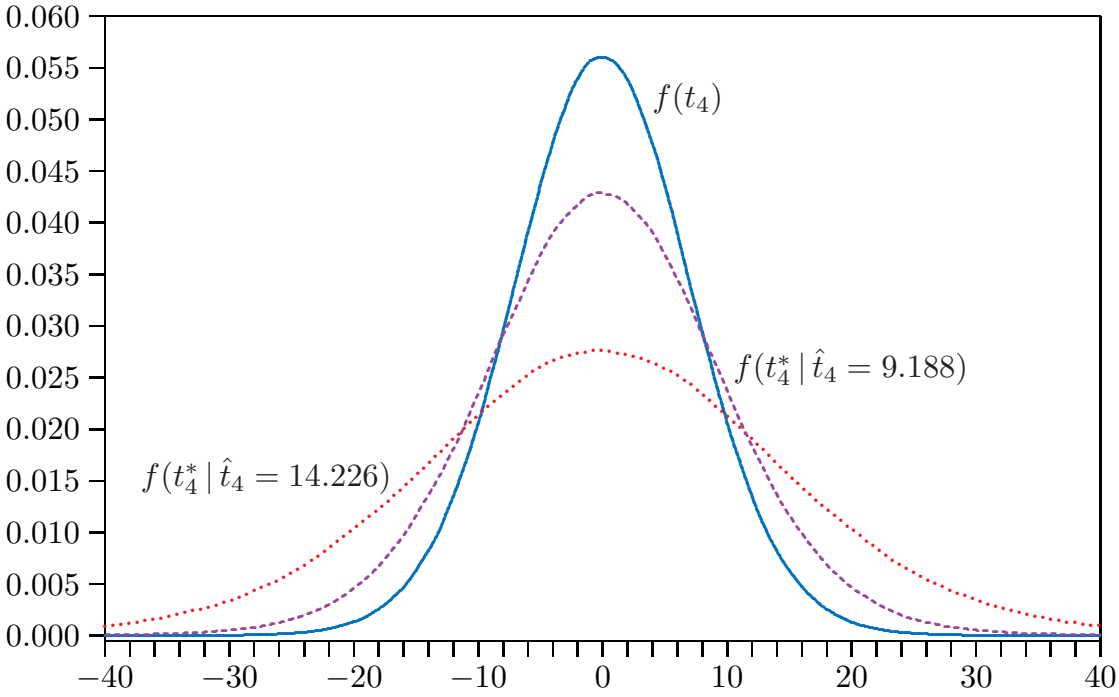
The dotted line in Figure 4 shows kernel density estimates of the bootstrap t statistics for a particular sample in which the realized value $\hat{t}_4 = 14.226$, the 0.975 quantile. This distribution is extremely bimodal, and each half of it is quite asymmetrical. The dashed line is similar, but it shows the bootstrap distribution for a particular sample in which $\hat{t}_4 = 9.188$, the 0.90 quantile. The two halves of the bootstrap distribution for $\hat{t}_4 = 14.226$ are much further apart than the two halves of the bootstrap distribution for $\hat{t}_4 = 9.188$. The 0.975 quantiles are 16.01 in the former case and 10.94 in the latter.

Although the Rademacher distribution is responsible for the bimodality of the bootstrap distributions in Figure 4, using another auxiliary distribution would not solve the problem. Figure 5 shows what happens when the standard normal distribution is used as the auxiliary distribution. The density of t_4 is the same as in Figure 4. The two bootstrap densities are, once again, conditional on $\hat{t}_4 = 14.226$ and $\hat{t}_4 = 9.188$. Although these densities are not bimodal, they are much too dispersed and consequently provide very poor approximations to the density of t_4. Moreover, they become more dispersed as $|t_4|$ increases.

The cases shown in Figure 4 are typical. As the test statistic becomes larger in absolute value, the bootstrap distributions tend to become more spread out. Thus it is not surprising that restricted wild bootstrap tests never reject at any conventional level of significance when $G_1 = 1$ in these experiments. No matter how large the absolute value of the test statistic, many of the realizations of the bootstrap statistic are even larger.

When $G_1 = 2$, there are four possible pairs of vectors of bootstrap error terms for the first two clusters. Two of these pairs are $[\tilde{\boldsymbol{\epsilon}}_1, \tilde{\boldsymbol{\epsilon}}_2]$ and $[-\tilde{\boldsymbol{\epsilon}}_1, -\tilde{\boldsymbol{\epsilon}}_2]$, which are very likely to yield large values of $|t_2^{*j}|$, while the other two are $[\tilde{\boldsymbol{\epsilon}}_1, -\tilde{\boldsymbol{\epsilon}}_2]$ and $[-\tilde{\boldsymbol{\epsilon}}_1, \tilde{\boldsymbol{\epsilon}}_2]$, which are not assured

Figure 5: Densities of t statistics for $G_1 = 1$ when auxiliary distribution is N(0, 1)



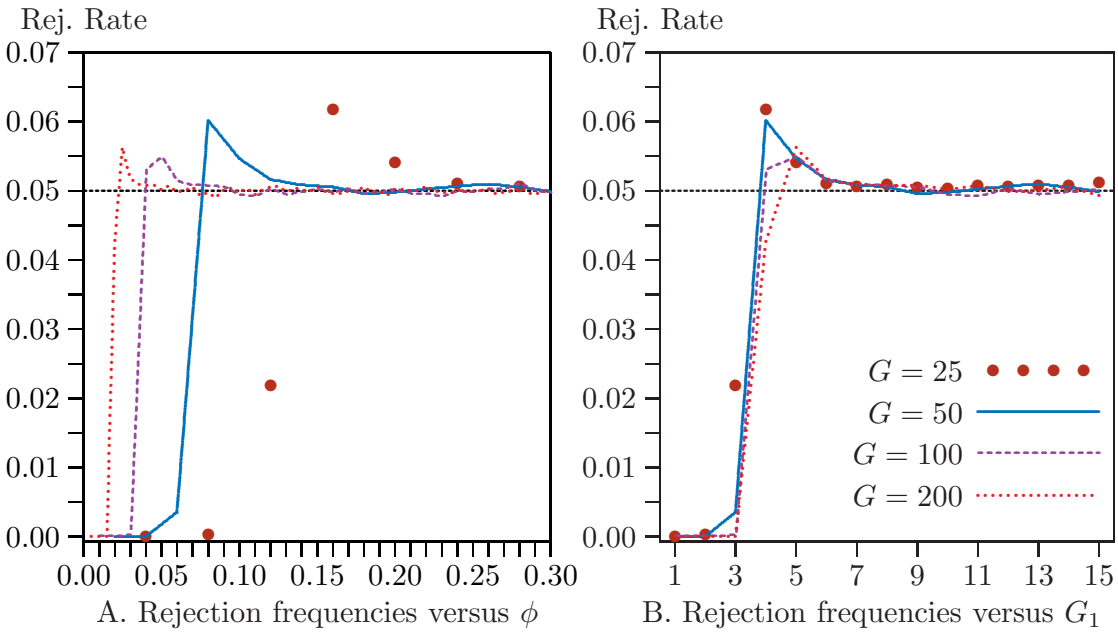
to do so. Thus the distribution of the t_2^{*j} for $G_1 = 2$ will still tend to have a large variance when $|t_2|$ is large, but not as large as when $G_1 = 1$. For the Rademacher distribution, the number of possible sets of bootstrap error terms for the treated observations is 2^{G_1}. As G_1 increases, the proportion of bootstrap samples for which $\tilde{\boldsymbol{\epsilon}}_g^{*j} = \hat{\boldsymbol{\epsilon}}_g$ or $\tilde{\boldsymbol{\epsilon}}_g^{*j} = -\hat{\boldsymbol{\epsilon}}_g$ for all $g \leq G_1$ rapidly declines. Once G_1 becomes large enough, the problem goes away; see Sections A.4 and A.6 of the appendix.

The foregoing analysis implies that the key parameter is G_1 rather than ϕ. The next two figures confirm this. Figure 6 shows what happens for the case of equal-sized clusters in a DiD model. Here G takes the values 25, 50, 100, and 200, with $N_g = 40$ in the first three cases and (to save computer time) $N_g = 20$ in the last one. The left-hand panel shows rejection frequencies as a function of ϕ. It demonstrates that the range of values of ϕ for which the wild bootstrap yields accurate inferences becomes wider as G increases. The right-hand panel shows rejection frequencies as a function of G_1. It suggests that the relationship between rejection frequencies and G_1 is almost invariant to G.

As we have seen, the numerator of the bootstrap test statistic under the null, $(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}^{*j}$, depends heavily on the restricted residuals for the treated clusters when G_1 is small. Thus it might seem that the failure of the wild cluster bootstrap for small G_1 could be avoided by using the unrestricted residuals $y_{ig} - \boldsymbol{X}_{ig}\hat{\boldsymbol{\beta}}$ rather than the restricted residuals $y_{ig} - \boldsymbol{X}_{ig}\tilde{\boldsymbol{\beta}}$ in the bootstrap DGP (3). Unfortunately, this is not so. Figure 7 deals with precisely the same cases as Figure 6, but the wild cluster bootstrap now uses unrestricted residuals. Instead of severe underrejection for small values of G_1, there is now very severe overrejection.

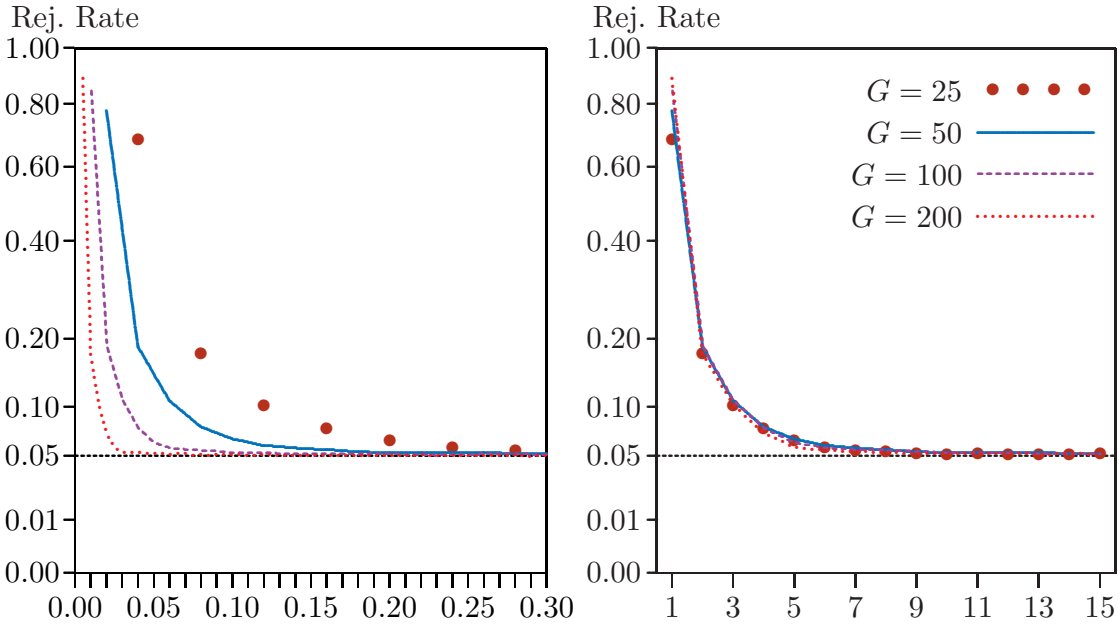
It is easy to explain the results in Figure 7. As before, the bootstrap test statistics

Figure 6: Restricted wild cluster bootstrap rejection frequencies



[Corrected 2019-08]

Figure 7: Unrestricted wild cluster bootstrap rejection frequencies



[Corrected 2019-08]

t_2^{*j} depend mainly on the bootstrap error terms for the G_1 treated clusters. But the OLS residuals used to compute them are orthogonal to the treatment dummy. Therefore, when $G_1 = 1$, the average of the bootstrap error terms over the observations in the treated cluster must be zero. Thus the first term in expression (17) is zero, and the distribution of the t_2^{*j} depends almost entirely on the second term. Since the value of t_2 depends principally on the first term, it is not surprising that the bootstrap distribution provides a terrible approximation when $G_1 = 1$. Things improve as G_1 increases, but they actually get worse as G increases, because that makes the first term larger relative to the second one. Observe that, in Figure 7, the overrejection for $G_1 = 1$ becomes more severe as G increases.

We conclude that, when G_1 is very small, using either CRVE t statistics or unrestricted wild cluster bootstrap tests is likely to result in severe overrejection, while using restricted wild cluster bootstrap tests is likely to result in severe underrejection. For given G_1, all three tests tend to perform worse as $\bar{d}$, the fraction of treated observations, declines. Since adding additional untreated clusters actually reduces $\bar{d}$, all three tests must fail asymptotically when G_1 is held fixed as the sample size increases. For the pure treatment model analyzed in this section, and also for the DiD model (7), all the results discussed in this paragraph also hold if G_1 is replaced by G_0 and $\bar{d}$ is replaced by $1 - \bar{d}$. However, this is not true for DiD models in which treatment occurs at different times in different clusters; see Section 7.

The results in Figures 6 and 7 suggest that, for the DiD case with equal-sized clusters, the wild cluster bootstrap (either restricted or unrestricted) can probably be used safely when $7 \leq G_1 \leq G - 7$, but it is likely to be seriously misleading for values less than 4 or greater than $G-4$. Of course, G_1 would probably need to be larger for the bootstrap to yield reliable results if the sizes of the treated clusters varied substantially; see Section A.6 of the appendix. It might well also need to be larger if there were pronounced heteroskedasticity that affected treated and untreated clusters differently, or if the model and dataset differed in other important ways from the ones in our experiments. Thus the exact value of G_1 needed for reliable inference is certain to be data-dependent.

7 Placebo Laws

An alternative way to study the reliability of inference using clustered data is to use real-world data and simulate the effect of "placebo laws." This ingenious approach was developed in Bertrand, Duflo and Mullainathan (2004), hereafter referred to as BDM, which uses data from the U.S. Current Population Survey. The dependent variable is the log of weekly earnings for women aged 25-50 from 1979 to 1999. The authors note that there is an issue with the modest number of clusters, but they do not mention the potential issues with clusters of varying sizes. In fact, they report only the mean cluster size.

The regression for y, the natural logarithm of women's earnings, is

$$y = \beta_{\text{treat}}\text{TREAT} + \text{YEARS}\,\boldsymbol{\beta}_{\text{years}} + \text{STATES}\,\boldsymbol{\beta}_{\text{states}} + \text{OTHERS}\,\boldsymbol{\beta}_{\text{other}} + \epsilon, \qquad (18)$$

where YEARS and STATES are full sets of fixed effects (so that there is no constant term), and OTHERS are age, age squared, and a set of education dummy variables. The treatment variable is analogous to the interaction term in a standard DiD equation, where it would be set to 1 for observations in the treatment states during the treatment periods and to 0 otherwise. In regression (18), the treatment variable is instead set to 1 randomly, so that

β_{treat} should be insignificantly different from zero. If the tests were working properly, we would expect β_{treat} to be significantly different from zero 5% of the time when testing at the .05 level.

The experiment in BDM is designed so that the treatment variable is randomly assigned to different states in each replication. For each replication, half the states are chosen at random to be treatment states, and half to be controls. Also, for each replication, a year between 1985 and 1995 is chosen at random to be the beginning of the treatment period. If this year is called $year^*$, then the treatment variable is

$$\text{TREAT} = I(\text{state} = treated)\, I(\text{year} \geq year^*),$$

where $I(\cdot)$ is the indicator function. Since these treatment variables are assigned at random, they should on average have no estimated impact on earnings.

Our simulations are similar to, but more extensive than, those in BDM. Instead of treating half the states, we perform 51 sets of simulations, according to the number of states treated. There are 51 states because we include the District of Columbia. We omit observations for which earnings $<$ \$20, which may be erroneous and are likely to have large residuals because of the log transformation, leaving us with a sample of 547,518 observations. In our principal experiments, we use micro data throughout, whereas most of the BDM simulations use data that are aggregated at the state-year level. In addition, the year that treatment starts is chosen at random for each treated state, rather than being the same for all treated states.

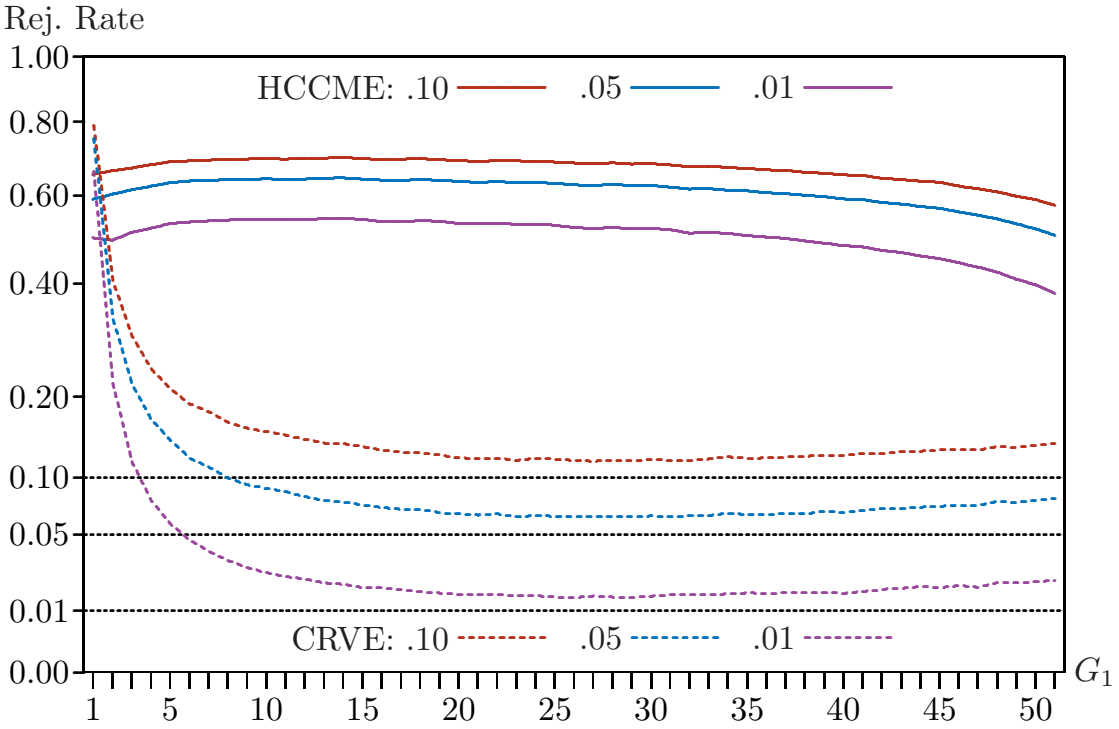
We estimate equation (18) for each replication and compute five different tests of the hypothesis that $\beta_{\text{treat}} = 0$. The first employs a t statistic based on a heteroskedasticity robust standard error and the $N(0,1)$ distribution. The second employs a t statistic based on a cluster robust standard error and the $t(G-1)$ distribution, as in the previous two sections. The third uses the same t statistic with critical values from the $t(G^*-1)$ distribution, with G^* based on an estimate of ρ. The fourth and fifth use P values computed by the restricted and unrestricted wild cluster bootstrap techniques.

Results for the first two tests are based on 100,000 replications, where the states and years to be treated are chosen randomly with replacement. However, the ones for $t(G^*-1)$ and the wild bootstrap tests are based on only 10,000 replications because of computational cost. When $G_1 = 1$, there are only $51 \times 11 = 561$ possible choices of state and years to treat. Therefore, results for that case are based on enumeration of those 561 cases.

The largest cluster (California) contains 42,625 observations. This makes it extremely expensive to compute G^*. A single replication using optimized code takes about 40 minutes. In order to make the experiments feasible, we compute G^* using only 1/100 of each sample. This reduces computational time by a factor of several thousand. Limited supporting experiments suggest that alternative approaches, such as using 1/50 samples or using several 1/100 samples and averaging, would have yielded almost identical results. The entire sample is used to estimate $\hat{\rho}$, which is always about 0.031 or 0.032.

Figure 8 plots rejection frequencies against G_1, the number of states treated, for t tests based on both HCCME and CRVE standard errors. Even though intra-state correlations seem to be very low, using t statistics based on HCCME standard errors results in very severe overrejection. Rejection frequencies at the .05 level always exceed 0.50. These results imply

Figure 8: HCCME and CRVE rejection frequencies for placebo laws



that the error terms are not simply equicorrelated within clusters, as in Moulton (1990). If they were, the state fixed-effect dummies would soak up any state-specific random effects, and the remaining errors would be uncorrelated.

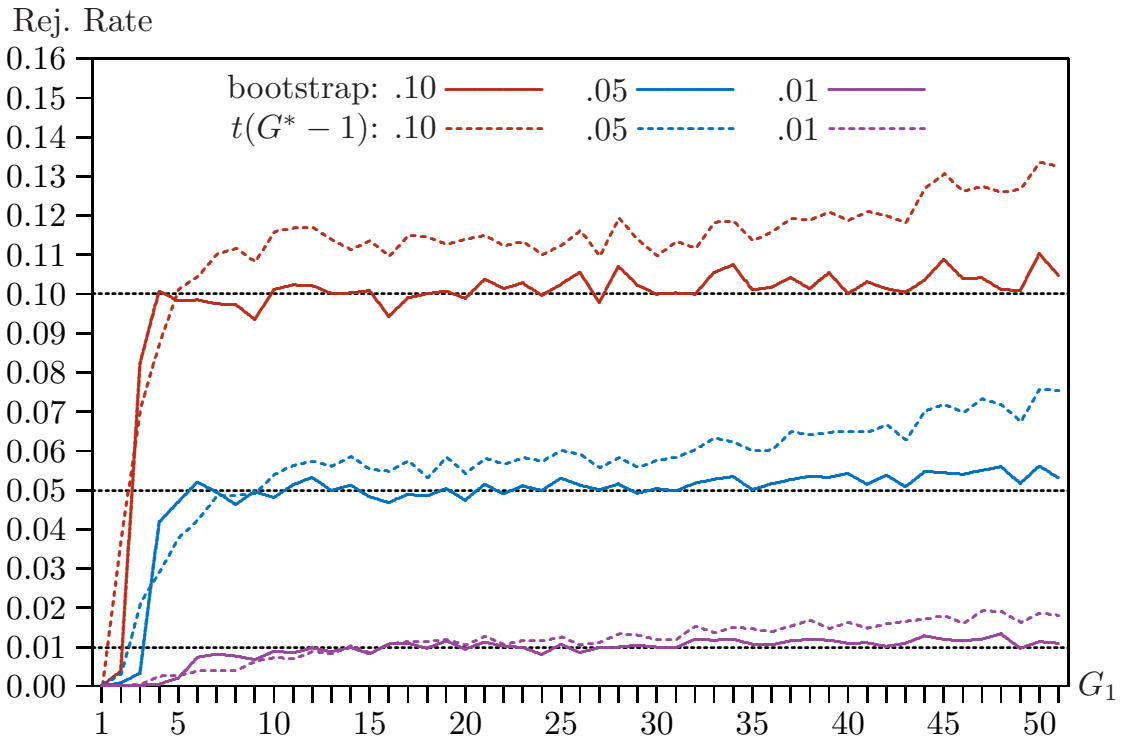
Using cluster robust standard errors and the $t(G-1)$ distribution works much better than using heteroskedasticity robust standard errors, except when only one state is treated. Nevertheless, as might have been expected from the results in Sections 5 and 6, there is severe overrejection when the number of treated states is small. For larger numbers of treated states, there is always overrejection, but it is fairly modest. Rejection frequencies at the .05 level always exceed 0.063, and they exceed 0.07 about half the time. In contrast to the results in Figure 1, the results for many treated states are very different from the results for few treated states. The two cases are no longer equivalent because the years in which treatment starts vary randomly across treated states.

Figure 9 plots rejection frequencies against G_1 for restricted wild cluster bootstrap tests and for $t(G^*-1)$ critical values. Both tests underreject severely when G_1 is very small. For larger values of G_1, the wild bootstrap performs extremely well, but using $t(G^*-1)$ critical values leads to moderate overrejection which increases with G_1. In Section A.3 of the appendix, we report results for unrestricted wild bootstrap tests. They always reject more often than the restricted ones, and they overreject severely when G_1 is small.

An alternative to using micro data is to average the values of all variables over all observations in each state-year pair, as BDM did in most of their experiments.[7] This results

[7]BDM took the average values of residuals, while we take average values of the actual data.

Figure 9: Bootstrap and $t(G^* - 1)$ rejection frequencies for placebo laws



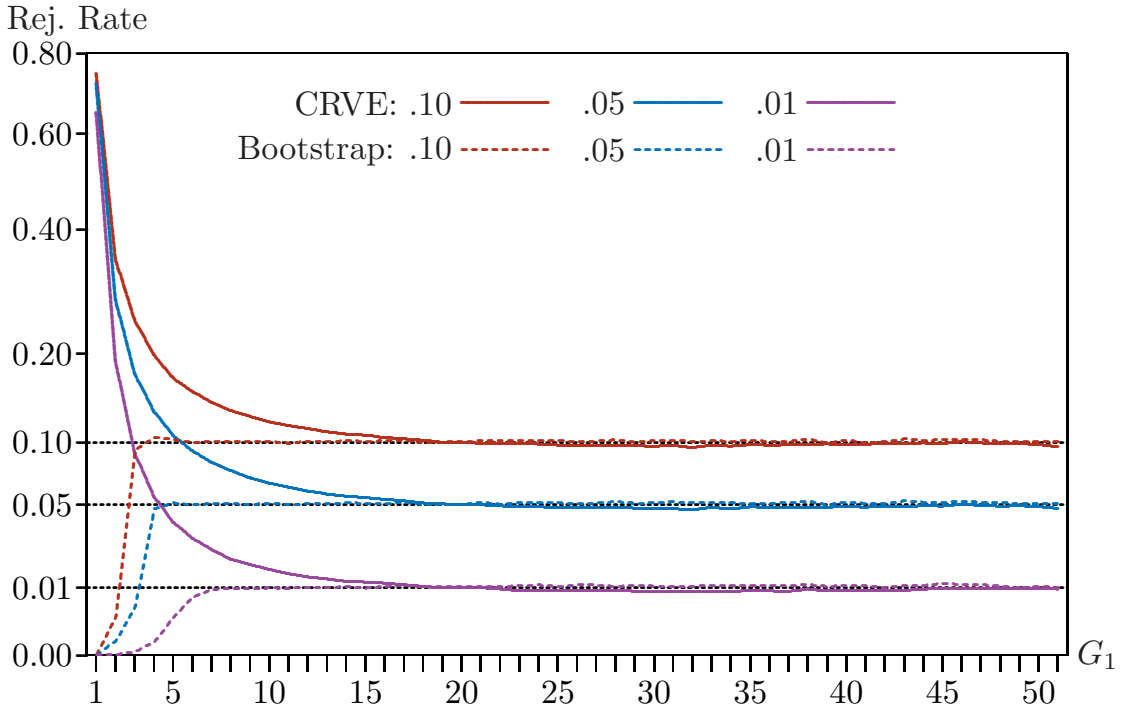
in $51 \times 21 = 1071$ observations and ensures that cluster sizes are perfectly balanced, with 21 observations in each. It also greatly reduces computational costs. Because the treatment variable in our experiments does not vary within state-year pairs, we might not expect much loss of power from this sort of aggregation. Evidence on this point is provided in Section A.8 of the appendix.

Figure 10 shows two sets of rejection frequencies as functions of G_1 for the aggregate placebo laws experiments. Additional results may be found in Section A.7. Results for cluster robust t statistics using the $t(G-1)$ distribution are based on 400,000 replications, and results for the wild cluster bootstrap are based on 100,000 replications. The CRVE t statistics overreject severely when G_1 is very small, but they perform very well for most values of G_1 and actually underreject slightly in most cases. As usual, the (restricted) wild cluster bootstrap underrejects severely when G_1 is very small, but it performs extremely well for $G_1 \geq 5$. The fact that both procedures perform better with aggregate data than with micro data reflects the fact that cluster sizes are always balanced in the former case.

8 Empirical Example

The results in Sections 4, 5, and 7 have shown that inference based on the standard CRVE coupled with the $t(G-1)$ distribution may be unreliable in situations with wildly different cluster sizes, but that other methods are more reliable. In this section, we illustrate how using either the wild cluster bootstrap or the effective number of clusters can affect inference

Figure 10: Rejection frequencies for placebo laws using aggregate data



in practice by replicating a few select results from Angrist and Kugler (2008).[8]

Angrist and Kugler (2008) investigates the impact of an exogenous shock in the price of coca within Columbia on economic and criminal activity. To estimate economic outcomes, the paper uses data on individuals who live in various Columbian departments. We replicate select results from the paper's Table 6, which uses a DiD methodology to estimate the impact of increased coca prices on log hours worked. In this specification, the rural departments are "treated" and the urban departments are "controls," because coca production was concentrated in the rural areas. The equation can be written as

$$\begin{aligned} y_{ijt} &= \boldsymbol{X}_i'\boldsymbol{\mu} + \beta_j + \delta_t + \alpha_{0,95-97}\, g_{jt,95-97} + \alpha_{0,98-00}\, g_{jt,98-00} \\ &\quad + \alpha_{1,95-97}\, d_{jt,95-97} + \alpha_{1,98-00}\, d_{jt,98-00} + \epsilon_{ijt}. \end{aligned} \tag{19}$$

Here $\boldsymbol{X}_i$ is a vector of control variables, β_j is a department dummy, δ_t is a year dummy, the α_0 terms are DiD coefficients for the rural growing areas, and the α_1 terms are DiD coefficients for the urban areas.

We replicate two sets of estimates from the paper's Table 6, namely, column 6 and column 9. The former explains log hours for men, and it excludes departments that are medium producers of coca from the analysis. The latter explains log hours for teenage boys, and it includes the medium producers. These regressions were chosen in part because the estimated P values implicitly reported for the test regressors are quite low. The sample for men has 181,882 observations, and the sample for teenage boys has 22,141 observations. In

[8]We thank Josh Angrist for making his data publicly available in his own data archive.

both cases, the cluster sizes are wildly different. For the adult men, there are 32 clusters; the largest has 25,775 observations, and the smallest has only 509. For the teenage boys, there are 38 clusters; the largest has 1,920 observations, and the smallest has only 42.

The results from these regressions can be found in Table 5. Panel A replicates the results for men (column 6 of the original table). We report six P values for each coefficient. The conventional ones use the $t(G-1)$ distribution. The symmetric bootstrap P values are based on 99,999 bootstrap samples (far more than would normally be needed). There are two G^* P values, one based on $t(G^*)$ and one based on $t(G^*-1)$, both estimated with $\hat{\rho} = 0.025$. All of the unconventional P values are larger than the conventional ones, and many of them are much larger. Only the rural coefficient for 1998-2000, $\alpha_{1,98-00}$ in equation (19), is significant at the 5% level according to the bootstrap.

Panel B replicates the results for teenage boys (column 9 of the original table). In this case, the bootstrap P values are very similar to the conventional ones, with both rural coefficients being significant at the 5% level. However, all of the $t(G^*-1)$ P values, and three of the four $t(G^*)$ P values, suggest that they are not significant. At 0.155, the estimate of ρ is much larger for the boys than for the men.

We also report estimates based on 288 aggregate observations at the department-year level for the men and 340 aggregate observations for the teenage boys. The estimates for the men, in Panel A, are of similar magnitude and sign to the micro estimates; however, all of the standard errors are significantly larger than before. The estimates for the boys, in Panel B, are in some cases quite different from the micro estimates. Two of the coefficients differ by a factor of almost three, and the standard errors are substantially larger.

Panel C calculates joint tests for the statistical significance of the two urban and the two rural coefficients. The reported statistics are quadratic forms in the 2-vectors of parameter estimates and the inverse of the appropriate 2×2 block of the CRVE. These Wald statistics are then divided by 2 so as to use the $F(2, G-1)$ distribution, as suggested in Cameron and Miller (2015), which also suggests computing wild cluster bootstrap P values. One of the statistics, for rural men, appears to be highly significant based on its $F(2, G-1)$ P value, but it is not significant at the 5% level according to the bootstrap. Because there is currently no way to calculate G^* for a joint test, we are unable to report results based on the effective number of clusters.

9 Conclusion

This paper identifies two circumstances in which inferences based on cluster robust standard errors should not be trusted, even when the sample size is large and the "rule of 42" is satisfied. There can be serious overrejection when clusters are of wildly unequal sizes. With dummy regressors for treatment, there can be extremely severe overrejection when the number of treated (or untreated) clusters is small, whether or not clusters are of equal sizes. This is true both for cluster-level treatments and for difference-in-differences regressions with and without cluster fixed effects. The problem is most severe when some or all of the treated clusters are small.

These results contrast with some of the earlier results of Bertrand, Duflo and Mullainathan (2004) and Cameron, Gelbach and Miller (2008), which suggest that cluster robust inference is reliable with 50 clusters. Those results are misleading because they are based

on equal-sized clusters and, in the former case, on aggregate data. Using aggregate data automatically solves the problem of unbalanced cluster sizes, but, as our empirical example illustrates, it can result in much larger standard errors.

Using critical values based on the effective number of clusters, as defined by Carter, Schnepel and Steigerwald (2017), instead of the actual number often improves matters substantially, although it does not always work as well as might be hoped. In contrast, with one notable exception, the restricted wild cluster bootstrap generally yields very reliable inferences for the cases we study. The exception is that, for reasons explained in Section 6, the restricted wild cluster bootstrap can underreject very severely when only a few clusters are treated or untreated. All the methods that we study perform very badly in this case.

Table 5: Empirical Example based on Angrist and Kugler (2008)

Panel A

Log Hours – Adult Men – No Medium Producers (32 departments)

	rural 95-97	rural 98-00	urban 95-97	urban 98-00
coef.	0.0581	0.1219	0.0405	0.0740
s.e.	0.0278	0.0359	0.0193	0.0395
t stat.	2.091	3.395	2.099	1.872
$G^*_{0.025}$	6.393	3.402	4.482	1.529
P values:				
$t(G-1)$	0.045	0.002	0.044	0.071
$t(G^*_{0.025})$	0.079	0.035	0.096	0.240
$t(G^*_{0.025}-1)$	0.087	0.059	0.114	0.447
bootstrap	0.186	0.028	0.092	0.090
Aggregate:				
coef.	0.0357	0.0975	0.0116	0.0563
s.e.	0.0621	0.0850	0.0455	0.0793
t stat.	0.575	1.147	0.255	0.710

Panel B

Log Hours – Teenage Boys – All Producers (38 departments)

	rural 95-97	rural 98-00	urban 95-97	urban 98-00
coef.	0.1185	0.2150	−0.0040	−0.0472
s.e.	0.0519	0.1052	0.0680	0.0904
t stat.	2.285	2.044	−0.058	−0.522
$G^*_{0.155}$	9.277	11.537	8.859	12.892
P values:				
$t(G-1)$	0.028	0.048	0.954	0.605
$t(G^*_{0.155})$	0.047	0.064	0.955	0.611
$t(G^*_{0.155}-1)$	0.051	0.067	0.955	0.611
bootstrap	0.030	0.050	0.958	0.628
Aggregate:				
coef.	0.0444	0.1432	−0.0490	−0.1139
s.e.	0.0664	0.1314	0.0899	0.1243
t stat.	0.669	1.090	−0.545	−0.916

Panel C

Joint Tests

	rural men	rural boys	urban men	urban boys
Test stat.	5.830	2.614	2.402	0.279
$F(2, G-1)$	0.007	0.087	0.107	0.758
bootstrap	0.091	0.140	0.226	0.796

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) 'Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program.' *Journal of the American Statistical Association* 105(490), 493–505

Angrist, Joshua D., and Adriana D. Kugler (2008) 'Rural windfall or a new resource curse? Coca, income, and civil conflict in Colombia.' *The Review of Economics and Statistics* 90(2), 191–215

Angrist, Joshua D., and Jorn-Steffen Pischke (2008) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton: Princeton University Press)

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

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), 249–275

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011) 'Inference with dependent data using cluster covariance estimators.' *Journal of Econometrics* 165(2), 137–151

Brewer, Mike, Thomas F. Crossley, and Robert Joyce (2018) 'Inference with difference-in-differences revisted.' *Journal of Econometric Methods* 7(1), 1–16

Cameron, A. Colin, and Douglas L. Miller (2015) 'A practitioner's guide to cluster robust inference.' *Journal of Human Resources* 50, 317–372

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90(3), 414–427

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017) 'Asymptotic behavior of a t-test robust to cluster heterogeneity.' *The Review of Economics and Statistics* 99(4), 698–709

Conley, Timothy G., and Christopher R. Taber (2011) 'Inference with "difference in differences" with a small number of policy changes.' *The Review of Economics and Statistics* 93(1), 113–125

Davidson, Russell, and Emmanuel Flachaire (2008) 'The wild bootstrap, tamed at last.' *Journal of Econometrics* 146(1), 162–169

Davidson, Russell, and James G. MacKinnon (1999) 'The size distortion of bootstrap tests.' *Econometric Theory* 15(3), 361–376

Davidson, Russell, and James G. MacKinnon (2004) *Econometric Theory and Methods* (Oxford: Oxford University Press)

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

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

Imbens, Guido W., and Michal Kolesár (2016) 'Robust standard errors in small samples: Some practical advice.' *Review of Economics and Statistics* 98(4), 701–712

Kloek, T. (1981) 'OLS estimation in a model where a microvariable is explained by aggregates and contemporaneous disturbances are equicorrelated.' *Econometrica* 49(1), 205–207

Liang, Kung-Yee, and Scott L. Zeger (1986) 'Longitudinal data analysis using generalized linear models.' *Biometrika* 73(1), 13–22

MacKinnon, James G., and Matthew D. Webb (2020) 'Randomization inference for difference-in-differences with few treated clusters.' *Journal of Econometrics* xxx, to appear

MacKinnon, James G., and Halbert White (1985) 'Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties.' *Journal of Econometrics* 29(3), 305–325

Moulton, Brent R. (1990) 'An illustration of a pitfall in estimating the effects of aggregate variables on micro units.' *Review of Economics & Statistics* 72(2), 334–338

Webb, Matthew D. (2014) 'Reworking wild bootstrap based inference for clustered errors.' Working Paper 1315, Queen's University, Department of Economics, August

White, Halbert (1984) *Asymptotic Theory for Econometricians* (Orlando: Academic Press)

Young, Alwyn (2016) 'Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections.' Technical Report, London School of Economics, January

Appendix

A.1 Introduction

Section A.2 presents simulation results for pure treatment effect models, where every observation in a cluster is treated if any observation is treated. Section A.3 briefly discusses the unrestricted wild cluster bootstrap and presents evidence on how well it performs for the placebo laws experiments. Section A.4 expands on the discussion in Section 6 and presents two additional figures which provide more intuition about why the wild bootstrap fails when the number of treated clusters is small.

Section A.5 briefly explains why inference based on CRVE t statistics and both variants of the wild cluster bootstrap are valid under standard asymptotics. Section A.6 provides evidence about the performance of the wild bootstrap when there are few clusters and cluster sizes are unbalanced, including the consequences of treated clusters being of wildly different sizes. Section A.7 discusses the use of aggregate data and presents results for placebo laws experiments using 1071 observations at the state-year level. Finally, Section A.8 discusses the power loss associated with aggregating micro data to the cluster-period level, creating perfectly balanced clusters.

A.2 Treatment Effects

Section 5 deals with treatment effects in the context of a difference-in-differences regression, where certain clusters are treated during certain time periods. Here we consider pure treatment effects with no time dimension, where the test regressor is an indicator variable that equals 1 for some proportion P of the clusters. Thus, for each cluster, either all observations are treated or all are not treated.

In Figures A.1, A.2, and A.3, we report results for 50 clusters with 2000 observations, $\rho_\epsilon = 0.05$, and G_1 that varies between 1 and 49.[9] The treatments are applied to equal-sized clusters and to state-sized clusters both from smallest to largest and from largest to smallest. The simulations use $400,000$ replications.

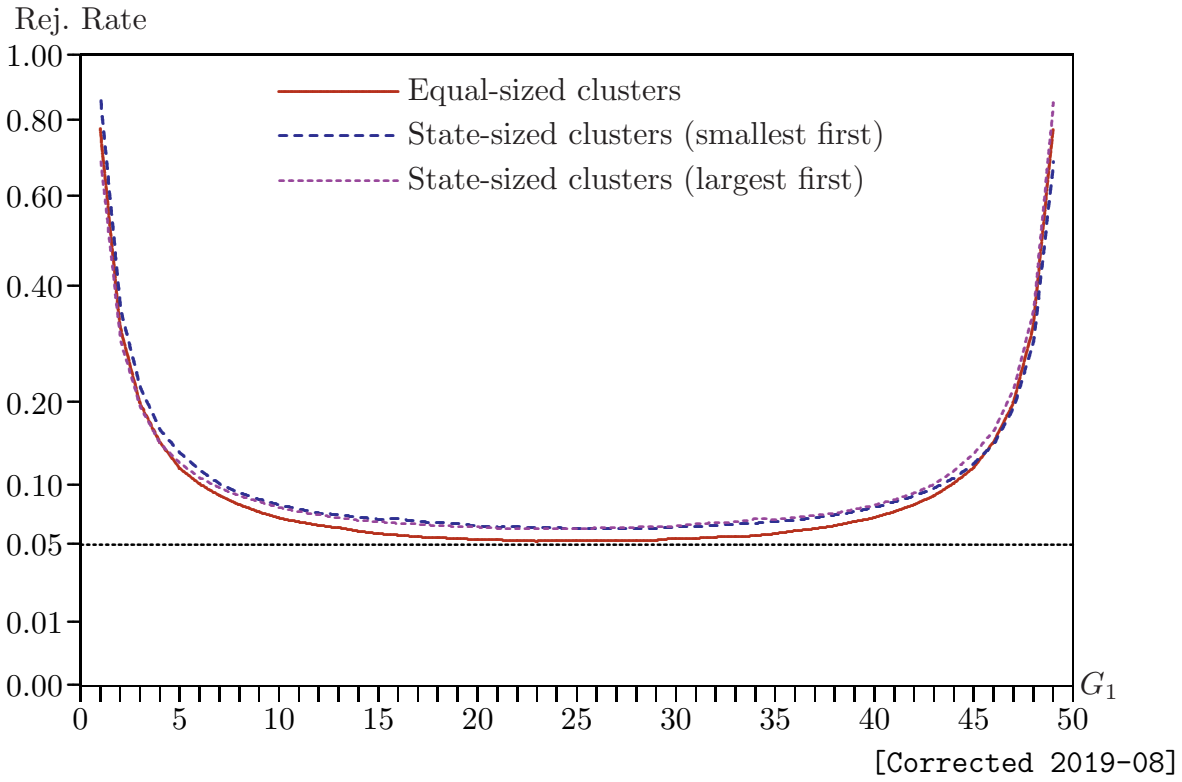
Figure A.1 shows results for tests based on CRVE standard errors and $t(49)$ critical values. As in the DiD case, there is very severe overrejection when either G_1 or $G-G_1$ is very small. Note that, because rejection frequencies vary so much, a square root transformation is applied to the vertical axis.

In Figure A.1, overrejection is quite modest when G_1 and $G - G_1$ are both far from 0. With equal-sized clusters, rejection frequencies are very close to 0.05 for G_1 between 17 and 33. With state-sized clusters, they are somewhat higher, never falling below 0.0616. The graph for equal-sized clusters is symmetric around $G_1 = 25$, while the ones for state-sized clusters are somewhat asymmetric. Overrejection is substantially more severe when only a few small clusters are either treated or not treated than when only a few large clusters are in that situation.

Figure A.2 shows results for tests based on $t(G^* - 1)$ critical values, where G^* is based

[9] In earlier versions of the paper, we set $\rho_\epsilon = 0.5$, a number that is unrealistically high. That is why Figures A.2 and A.3 look noticeably different from previous versions of the same figures.

Figure A.1: Rejection rates for pure treatment case, $t(G-1)$



on $\hat{\rho}$. There is extreme underrejection when $G_1 = 1$ and $G_1 = 49$. Rejection frequencies are very sensitive to the degrees of freedom parameter; using critical values based on $t(G^*)$ instead of $t(G^* - 1)$ leads to moderately severe overrejection. Away from the extremes, the tests can either underreject or overreject, although they always overreject for equal-sized clusters when $3 \leq G_1 \leq 47$. The rejection frequencies appear to be symmetric around $G_1 = 25$ for equal-sized clusters, but quite asymmetric for state-sized ones. In the latter case, there can be either noticeable overrejection or severe underrejection.

Figure A.3 shows results for wild bootstrap tests based on simulations with 399 bootstraps. In all cases, there is severe underrejection when either G_1 or $G - G_1$ is very small. In the most extreme cases, there are no rejections at all. For equal-sized clusters, there is modest overrejection when the number of treated or untreated clusters is between 4 and 6, but the wild bootstrap tests work extremely well for G_1 between about 7 and 43.

For state-sized clusters, the pattern is a bit more complicated. When the states are treated from smallest to largest, the bootstrap tests always underreject severely when G_1 is very close to 0 or 50, and they overreject quite severely when G_1 is between 44 and 48. When the states are treated from largest to smallest, the opposite problem occurs, with serious overrejection when G_1 is between 2 and 6, and severe underrejection when G_1 is very close to 0 or 50. In both cases, the peak overrejection occurs when there are 3 treated (or non-treated) clusters that together account for 26.6% of the observations.

Figures A.2 and A.3 are the only ones that changed substantially when the value of ρ_ϵ was increased from 0.0025 (due to a programming error) to the intended value of 0.05. The

Figure A.2: Rejection rates for pure treatment case, $t(G^* - 1)$

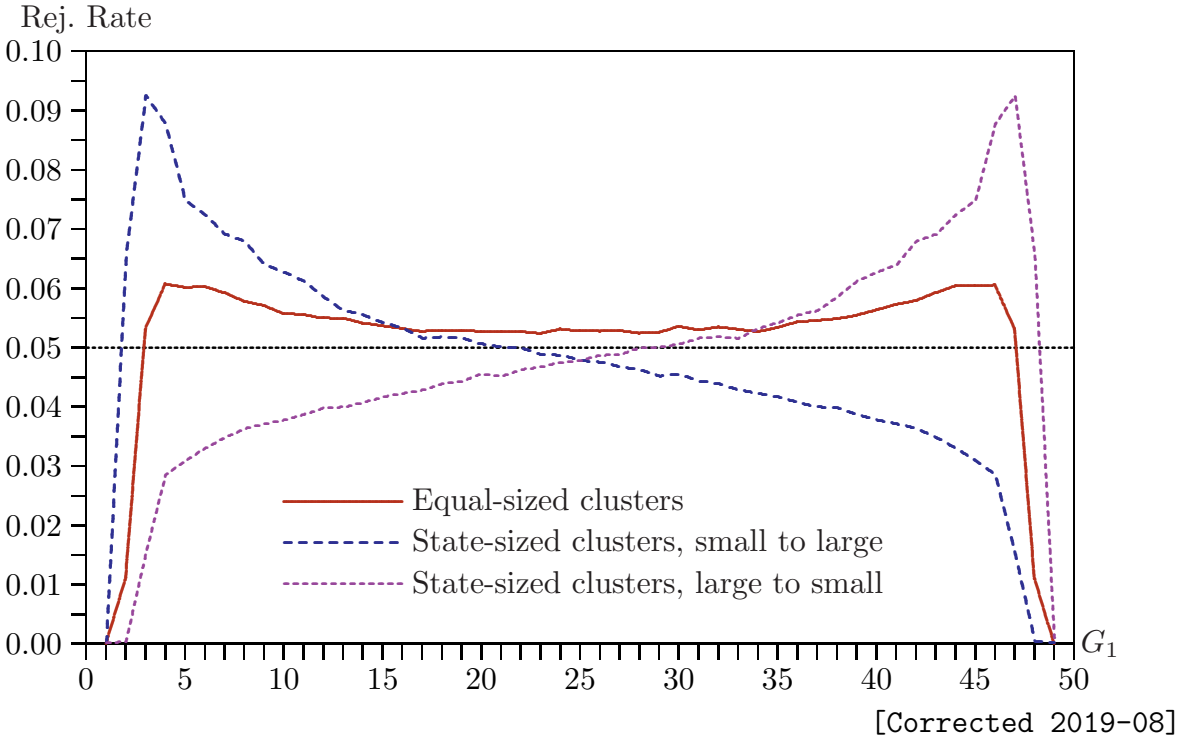
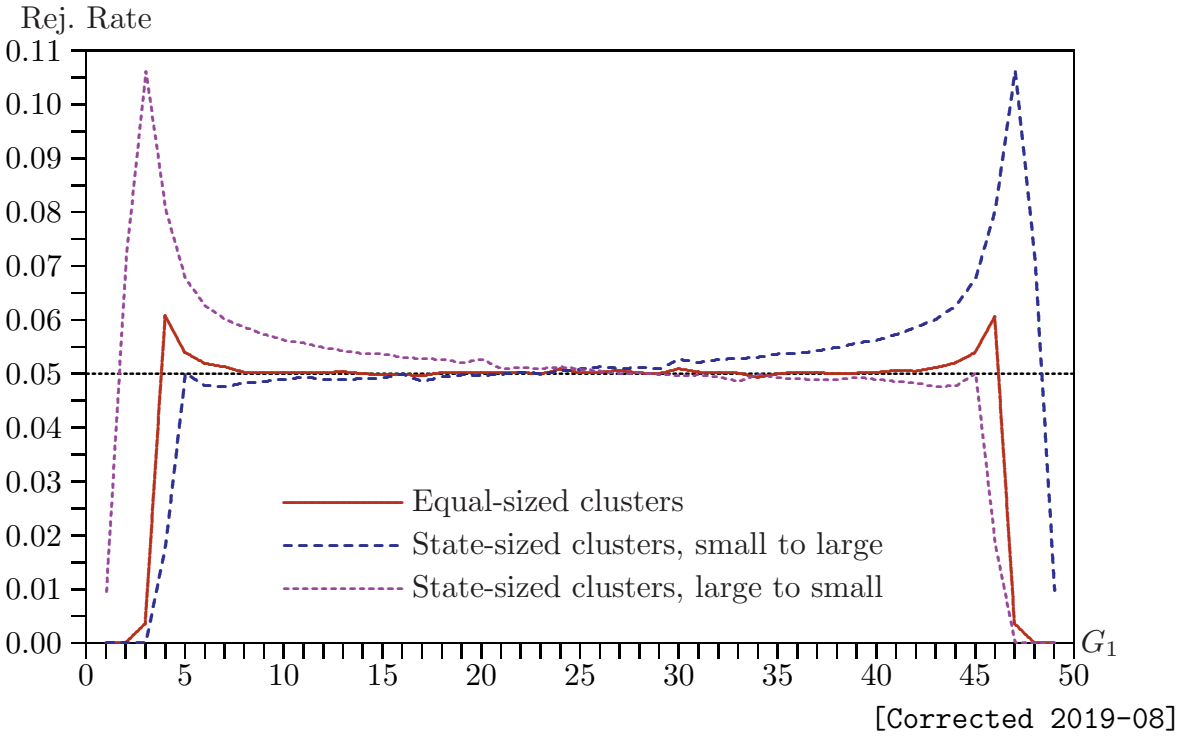


Figure A.3: Rejection rates for pure treatment case, wild bootstrap



overrejection by the WCR bootstrap when a few large clusters are treated became much more severe, as did the overrejection by tests based on $t(G^* - 1)$ when a few small clusters are treated. Why the value of ρ_ϵ matters so much more for the pure treatment case than for the DiD case (Figures 2 and 3 changed much less when ρ_ϵ was increased, although in qualitatively the same ways) is unknown.

A.3 The Unrestricted Wild Cluster Bootstrap

The wild cluster bootstrap is discussed in Section 2. The bootstrap DGP (3) uses restricted residuals, because it is generally best to base bootstrap DGPs on restricted estimates; see Davidson and MacKinnon (1999). However, it is also perfectly valid to use the unrestricted residuals $\hat{\epsilon}_{ig}$ instead of the restricted ones $\tilde{\epsilon}_{ig}$ in the bootstrap DGP. In addition, it would be possible to use the unrestricted estimates $\hat{\boldsymbol{\beta}}$ rather than the restricted estimates $\tilde{\boldsymbol{\beta}}$, provided the bootstrap test statistic were modified so as to test the null hypothesis that $\beta_k = \hat{\beta}_k$ rather than $\beta_k = 0$. The two variants of the unrestricted wild cluster bootstrap would yield identical results in all our experiments, because the $\boldsymbol{X}$ matrix does not depend on $\boldsymbol{\beta}$. We will refer to the wild cluster bootstrap based on restricted residuals as WCR and to the one based on unrestricted residuals as WCU.

Figure A.4: Placebo laws: Rejection rates for two wild bootstrap methods

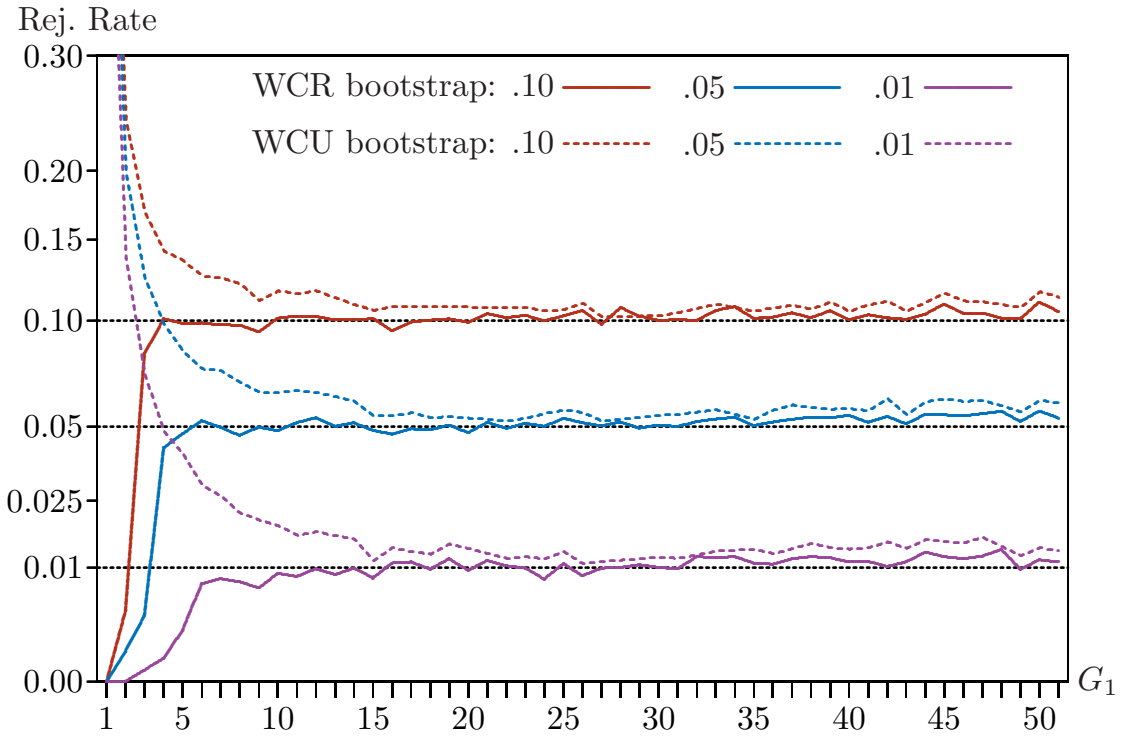


Figure A.4 shows rejection frequencies for the WCR and WCU bootstraps for the placebo laws experiments. As in Figure A.1, a square root transformation has been applied to the vertical axis. For clarity, the vertical axis only extends to 0.30, which means that results for

WCU with just one treated state cannot be shown. The rejection frequencies at the .10, .05, and .01 levels for this case are 0.779, 0.731, and 0.647, respectively. The results for WCR are the same as the ones in Figure 9. For very small values of G_1, WCR underrejects very severely, and WCU overrejects very severely. For somewhat larger values, WCR works well, but WCU overrejects noticeably. For values of G_1 greater than about 15, both procedures work quite well, although not perfectly. They both have a slight tendency to overreject in this region, with WCU always rejecting a bit more often than WCR, especially for very large values of G_1. Overall, WCR is clearly the procedure of choice.

The reason why WCU overrejects, often severely, for small values of G_1 was discussed in Section 6. Essentially, the problem is that the bootstrap DGP uses unrestricted residuals for the treated cluster(s) which sum to zero when $G_1 = 1$ and sum to numbers that are too small when $G_1 > 1$. This causes the variance of the bootstrap test statistics to be too small. Exactly the same problem causes the CRVE to underestimate the variance of the coefficient on the treatment dummy.

In Section 6, we discussed why the sum of the unrestricted residuals for just one treated cluster is equal to zero, but we did not formally discuss the case in which $G_1 > 1$. Consider once again the pure treatment model given by equation (8), where the first cluster and at least one other cluster is treated. The sum of the residuals for the first cluster is $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$, and we want to find the expectation of this quantity.

By a well-known result for OLS regression, $\hat{\boldsymbol{\epsilon}} = \big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{\epsilon}$. In the case of the pure treatment model,

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{\iota}'\boldsymbol{\iota} & \boldsymbol{\iota}'\boldsymbol{d} \\ \boldsymbol{d}'\boldsymbol{\iota} & \boldsymbol{d}'\boldsymbol{d} \end{bmatrix} = \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix},$$

and

$$\begin{bmatrix} \boldsymbol{\iota}'_{N_1} \\ \mathbf{0} \end{bmatrix}' \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} N_1 & N_1 \end{bmatrix},$$

so that

$$\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1 = \boldsymbol{\iota}'_{N_1}\boldsymbol{\epsilon} - \begin{bmatrix} N_1 & N_1 \end{bmatrix} \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\iota} & \boldsymbol{d} \end{bmatrix}' \boldsymbol{\epsilon}.$$

Recall that N_1 is the number of observations in cluster 1, and M_1 is the total number of observations in all the treated clusters.

In order to take the expectation of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1\hat{\boldsymbol{\epsilon}}'_1\boldsymbol{\iota}_{N_1}$, we need to make an assumption about how the vector $\boldsymbol{\epsilon}$ is distributed. The simplest assumption is that $\boldsymbol{\epsilon}$ has mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. In that case, it is easy to see that

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1) &= \sigma^2\boldsymbol{\iota}'_{N_1}\big(\mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\big)\boldsymbol{\iota}_{N_1} \\ &= N_1\sigma^2 - \begin{bmatrix} N_1 & N_1 \end{bmatrix} \begin{bmatrix} N & M_1 \\ M_1 & M_1 \end{bmatrix} \begin{bmatrix} N_1 \\ N_1 \end{bmatrix} \sigma^2. \end{aligned} \tag{A.1}$$

It is not difficult to verify that equation (A.1) can be rewritten as

$$\mathrm{Var}(\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1) = \sigma^2 \frac{N_1(M_1 - N_1)}{M_1}. \tag{A.2}$$

Thus the factor by which the variance of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$, the sum of the residuals for the first treated cluster, is shrunk relative to the variance of $\boldsymbol{\iota}'_{N_1}\boldsymbol{\epsilon}_1$ is simply $(M_1 - N_1)/M_1$. This factor is the ratio of the number of treated observations in clusters other than first one to the total number of treated observations. This result is consistent with what we saw in Section 6, namely, that when only one cluster is treated, the variance of $\boldsymbol{\iota}'_{N_1}\hat{\boldsymbol{\epsilon}}_1$ is zero.

The result (A.2) depends on the assumption of IID errors, which may not seem entirely reasonable in this case. However, any other assumption would lead to a much more complicated analysis, and it seems very unlikely that such an analysis would reverse the key implication of equation (A.2). This implication is that, as the fraction of treated observations that belong to a single cluster increases, the variance of the sum of the residuals for that cluster falls relative to the variance of the sum of the corresponding error terms. This strongly suggests that inference based on CRVE t statistics will be less reliable when the sizes of the treated clusters varies substantially than when all the treated clusters are approximately the same size. Section A.6 provides some evidence on this point.

A.4 Why the Wild Bootstrap Fails for Few Treated Clusters

In Section 6, we discussed why the wild bootstrap fails when the number of treated clusters is small. In particular, we considered the case in which $G_1 = 1$ and the bootstrap DGP uses the Rademacher distribution. We showed that, when the test statistic t_2 is large in absolute value, the distribution of the restricted wild cluster bootstrap t statistics tends to be bimodal, with half the realizations distributed around t_2 and the other half distributed around $-t_2$. This was illustrated in Figure 4.

As Section 6 of the paper explains, and as Figures 6 and 7 illustrate, the WCU and WCR bootstraps fail in very different ways when G_1 is small. When $G_1 = 1$ and all the clusters are of equal size, the distribution of the WCU bootstrap statistics seems to be very close to $t(G-1)$. This happens because the first term in expression (17) is identically zero. The numerator of the bootstrap statistic is therefore proportional to

$$\sum_{g=2}^{G} v_g^{*j} \left(\sum_{i=1}^{N_g} \hat{\epsilon}_{ig} \right). \tag{A.3}$$

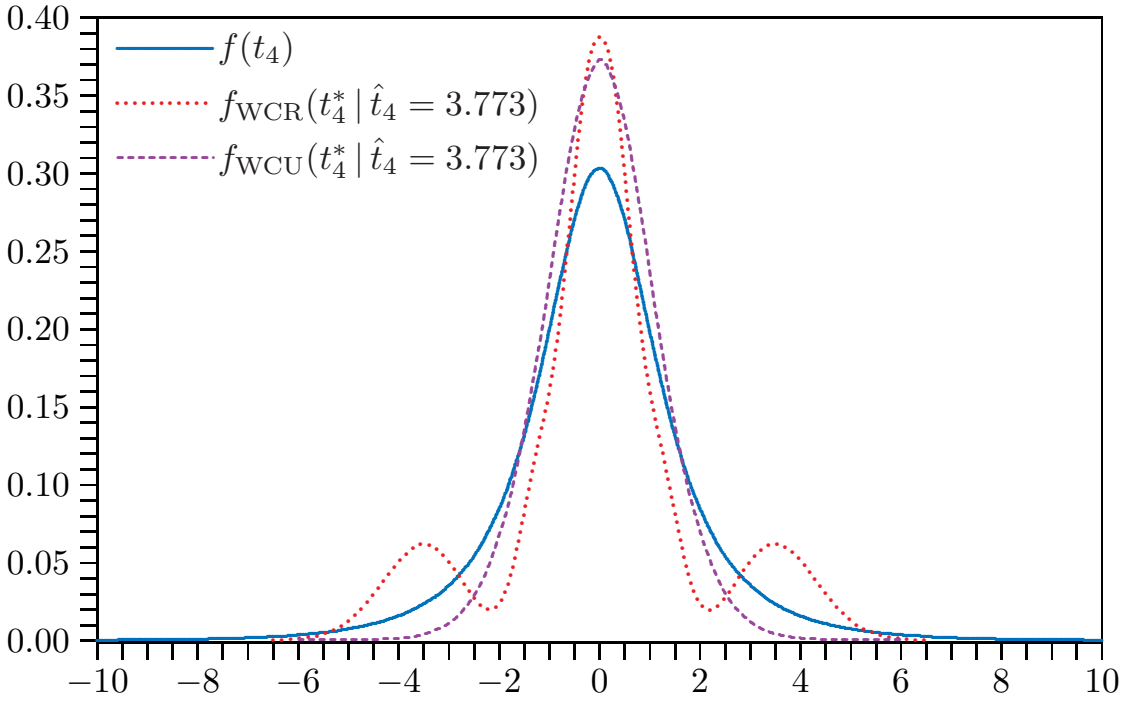
This is a summation of $G-1$ random variables, each of which has expectation zero. There is just one random variable for each of the untreated clusters, because each of the summations inside the parentheses is multiplied by a single auxiliary random variable v_g^{*j}.

As can be seen from equation (13), the denominator of the bootstrap statistic is in this case proportional to the square root of

$$\sum_{g=2}^{G} \left(\sum_{i=1}^{N_g} \hat{\epsilon}_{ig}^{*} \right)^2, \tag{A.4}$$

where the $\hat{\epsilon}_{ig}^{*}$ are the OLS residuals for the bootstrap data. These residuals (for the $G-1$ untreated clusters only) must be approximately equal to the corresponding bootstrap error terms, $v_g^{*j}\hat{\epsilon}_{ig}$. The summations of the $\hat{\epsilon}_{ig}^{*}$ over each cluster are not exactly normally distributed. Nevertheless, in our experiments, whatever distribution they do have when all the

Figure A.5: Densities of actual and bootstrap t statistics when $G_1 = 3$



N_g are equal seems to be close enough that $t(G-1)$ is a very good approximation to the distribution of the WCU bootstrap test statistics according to Kolmogorov-Smirnov tests. When cluster sizes are unbalanced, however, each of the $G-1$ summations in expression (A.4) has a different variance, and so we would not expect any t distribution to provide a good approximation. That is exactly what we find when the N_g vary substantially.

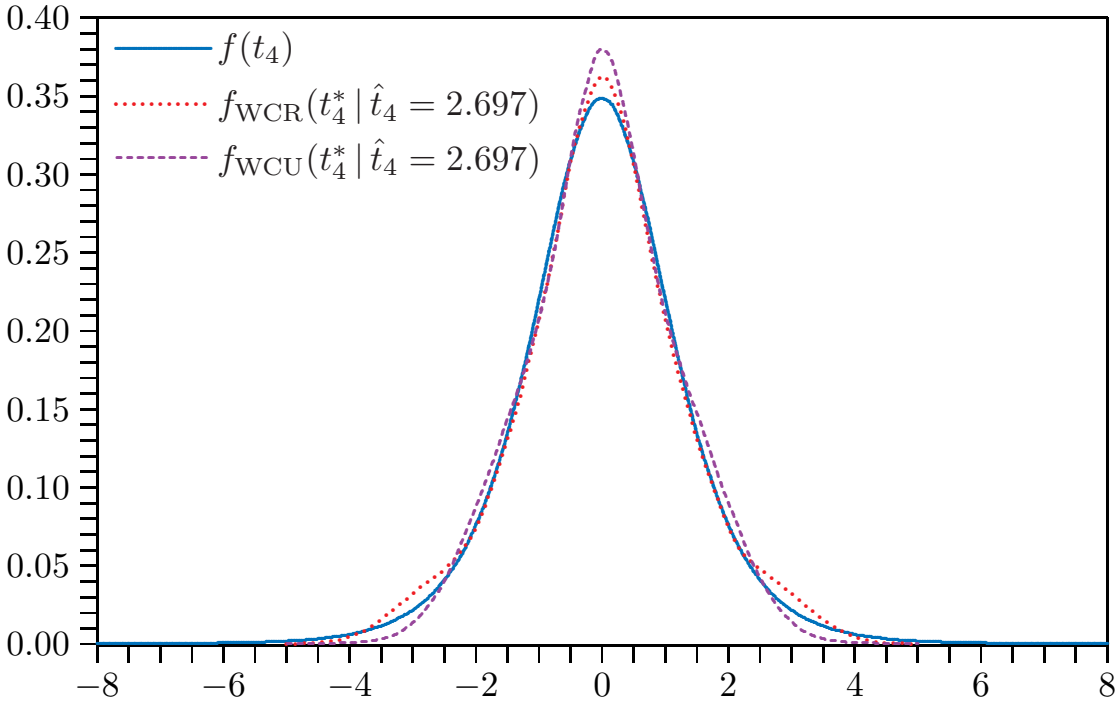
This explains why the rejection frequencies for WCU bootstrap tests and for tests based on the $t(G-1)$ distribution are so close when $G_1 = 1$ in many of our experiments. For example, for the placebo laws experiments with aggregate data, the former rejects 71.8% of the time at the .05 level, and the latter rejects 72.0% of the time. As expected, the correspondence is not quite as good with micro data. The WCU bootstrap tests reject 73.1% of the time, and tests based on the $t(G-1)$ distribution reject 75.0% of the time.

Of course, as G_1 increases, both variants of the wild cluster bootstrap improve rapidly. Figures A.5 and A.6 are similar to Figure 4, with 2000 observations and 50 equal-sized clusters. They show the densities of t_4 when $G_1 = 3$ and $G_1 = 5$, respectively, along with the densities of the WCR and WCU bootstrap statistics for the particular samples for which the realized test statistic $\hat{t}_4$ is the 0.975 quantile of the t_4 distribution.[10]

In Figure A.5, neither bootstrap density provides a good approximation, but both are appreciably better than those for $G_1 = 1$. Interestingly, due presumably to the use of the Rademacher distribution, the WCR bootstrap density is now trimodal. It underrejects because the two outer modes give it too much mass for $|t_4^*| > |t_4|$, and those modes move

[10]Figures A.5 and A.6, like Figures 4 and 5, are based on simulations that mistakenly set $\rho_\epsilon = 0.0025$. The value of ρ_ϵ has no effect on the key features of these figures.

Figure A.6: Densities of actual and bootstrap t statistics when $G_1 = 5$



away from the origin as $|t_4|$ increases. The outer modes are associated with realizations of the Rademacher random variables that are either 1 or -1 for all three treated clusters. The WCU bootstrap density has much thinner tails than the actual density, which explains why tests based on it overreject severely.

In Figure A.6, both bootstrap densities perform very much better than in Figure A.5. Moving from $G_1 = 3$ to $G_1 = 5$ evidently makes a big difference. The WCR bootstrap density is no longer trimodal, although it does have two bulges which presumably arise for the same reason. The WCU bootstrap density still has thinner tails than the actual density, but to a lesser extent.

A.5 Standard Asymptotics

In Section 6, we showed why inference based on CRVE t statistics fails asymptotically when G_1 is fixed and why the bootstrap fails to solve the problem. In this section, we briefly discuss why these issues do not occur under a standard asymptotic construction. Instead of holding G_1 fixed as $N \to \infty$, we make the more conventional assumption that $\phi = G_1/G$ tends to a constant that is strictly between 0 and 1. The simplest way to do this would be to consider only samples that are integer multiples of the original sample size, with regressor matrices that simply repeat the original one, so that ϕ and $\bar{d}$ are the same for every sample. Alternatively, we could allow both ϕ and $\bar{d}$ to vary somewhat as N increases, provided that $\bar{d} \to d_\infty$ and $\phi \to \phi_\infty$ as $N \to \infty$, with $0 < d_\infty < 1$ and $0 < \phi_\infty < 1$. We continue to assume, for simplicity, that G/N tends to a constant as $N \to \infty$.

Under the null hypothesis, the CRVE statistic is given by equation (12), which we rewrite here for convenience:

$$t_2 = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{\left(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\right)^{1/2}}. \tag{A.5}$$

Ignoring the scalar factor c, the numerator of this test statistic is

$$(1 - \bar{d})\sum_{g=1}^{G_1}\sum_{i=1}^{N_g}\epsilon_{ig} - \bar{d}\sum_{g=G_1+1}^{G}\sum_{i=1}^{N_g}\epsilon_{ig}. \tag{A.6}$$

Since $\bar{d}$ tends to a constant between 0 and 1, and the ϵ_{ig} have mean zero, both terms here are evidently $O_p(N^{1/2})$, and so is their weighted sum.

The square of the denominator of (A.5) is given by expression (13), which is

$$(1 - \bar{d})^2\sum_{g=1}^{G_1}\left(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\right)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}\left(\sum_{i=1}^{N_g}\hat{\epsilon}_{ig}\right)^2. \tag{A.7}$$

Both terms here are evidently $O_p(N)$, and so is their weighted sum. By the results of CSS, N^{-1} times expression (A.7) consistently estimates the variance of $N^{-1/2}$ times expression (A.6). Thus, provided we can apply a central limit theorem to the latter, we obtain the standard result that $t_2 \overset{a}{\sim} \mathrm{N}(0,1)$.

For the bootstrap to be asymptotically valid in this case, the distribution of the bootstrap test statistic t_2^{*j}, which is $\hat{\beta}_2^{*j}$ divided by the bootstrap CRVE standard error, must be asymptotically standard normal. This test statistic was given in (16) and is repeated here for convenience:

$$t_2^{*j} = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}^{*j}}{\left(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g^{*j}(\hat{\boldsymbol{\epsilon}}_g^{*j})'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\right)^{1/2}}. \tag{A.8}$$

Here $\boldsymbol{\epsilon}^{*j}$ is the vector of bootstrap error terms for bootstrap sample j, and $\hat{\boldsymbol{\epsilon}}_g^{*j}$ is the subvector of the bootstrap residual vector $\hat{\boldsymbol{\epsilon}}^{*j}$ corresponding to cluster g. These bootstrap residual vectors are obtained by regressing the $\boldsymbol{y}^{*j}$ on $\boldsymbol{\iota}$ and $\boldsymbol{d}$.

The arguments needed to show that t_2^{*j} is asymptotically standard normal are very similar to the ones for t_2 itself. Under the null hypothesis, the residual vectors $\tilde{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\epsilon}}$ both converge to the error vector $\boldsymbol{\epsilon}$ as $N \to \infty$. Since the bootstrap DGP (assume the Rademacher distribution for simplicity) merely changes the signs of all the residuals for each cluster with probability one-half, the bootstrap error vectors $\boldsymbol{\epsilon}^{*j}$ must asymptotically follow exactly the same distribution as $\boldsymbol{\epsilon}$ if that distribution is symmetric.[11] Similarly, the bootstrap residual vectors $\hat{\boldsymbol{\epsilon}}^{*j}$ must follow the same distributions asymptotically as the actual residual vectors $\hat{\boldsymbol{\epsilon}}$. Thus everything that is true asymptotically for the actual t statistics (A.5) is also true for the bootstrap t statistics (A.8). We conclude that, under the standard asymptotic construction of this section, t_2 and the t_2^{*j} both follow the standard normal distribution asymptotically.

[11] Even if the ϵ_{ig} were not symmetrically distributed, the wild cluster bootstrap using the Rademacher distribution would still be asymptotically valid, but the argument would be a bit more complicated.

The crucial feature of the standard asymptotic construction is that $\bar{d} \to d_\infty$ with $0 < d_\infty < 1$ as $N \to \infty$. If instead $\bar{d} \to 0$, as it does when G_1 is fixed, or $\bar{d} \to 1$, as it does when $G_0 = G - G_1$ is fixed, then the two terms in each of (A.6) and (A.7), and their bootstrap analogs, are of different orders. When G_1 is fixed, the first term in expression (A.6) is the product of two factors, each of which is $O_p(1)$, while the second term is $O_p(N^{-1})O_p(N^{1/2}) = O_p(N^{-1/2})$. Similarly, the first term in expression (A.7) is $O_p(1)$, while the second term is $O_p(N^{-2})O_p(N) = O_p(N^{-1})$. Thus standard asymptotic arguments break down. Indeed, as was shown in Section 6, $\hat{\beta}_2$ is actually inconsistent.

A.6 Few Unbalanced Clusters

In all but two of the experiments reported in the paper, the number of clusters G is at least 50, because in many of them cluster sizes are proportional to state populations. In applied work, however, the number of clusters is often substantially less than 50. In this section, we present some results for smaller values of G and for cluster sizes that vary in different ways. The results we report are just a few of the many we obtained, but they are fairly representative.

The model is the DiD regression given in equation (7). Half the observations in treated clusters are treated, and $\rho_\epsilon = 0.05$. Experiments with 400,000 replications and 399 bootstrap samples are performed for five values of G: 12, 16, 20, 25, and 32. There are $N = 50G$ observations. In order to allow for unbalanced cluster sizes, N_g is determined by a parameter $\gamma \geq 0$, as follows:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{A.9}$$

where $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{j=1}^{G-1} N_g$. When $\gamma = 0$, every N_g is equal to $N/G = 50$. As γ increases, cluster sizes become increasingly unbalanced.

Figure A.7 shows rejection frequencies at the .05 level for WCR and WCU bootstrap tests for three values of G (12, 20, and 32) when $\gamma = 3$ as functions of $\phi = G_1/G$. Cluster sizes are quite unbalanced. The smallest is always 8, and the largest is either 146, 155, or 167 for $G = 12$, 20, or 32, respectively. Clusters are treated from smallest to largest, and the pronounced asymmetry that is evident in the figure reflects this. As in Figure A.4, the vertical axis is subject to a square root transformation.

The WCR bootstrap results in Figure A.7 are quite similar to the ones in Figure 3. There is severe underrejection for extreme values of ϕ. As expected, the range of values of ϕ for which the bootstrap performs acceptably becomes wider as G increases. For $G = 20$ and $G = 32$, this range is $6 \leq G_1 \leq G - 5$. Whether there is any such range for $G = 12$ depends on how one defines "acceptably." The rejection frequencies fall between 0.053 and 0.055 when $5 \leq G_1 \leq 8$, and this might be considered acceptable.

The WCU bootstrap results in Figure A.7 are also as expected. There is always severe overrejection for extreme values of ϕ. For clarity, the vertical axis has been truncated at 0.20, and rejection frequencies for $G_1 = 1$ and $G_1 = G-1$ are always much greater than this. For $G = 12$, the WCU bootstrap always overrejects noticeably. For $G = 20$ and $G = 32$, it almost always rejects more often than the WCR bootstrap, but it performs acceptably for some intermediate values of ϕ.

Figure A.7: Rejection frequencies for bootstrap tests when G is small and $\gamma = 3$

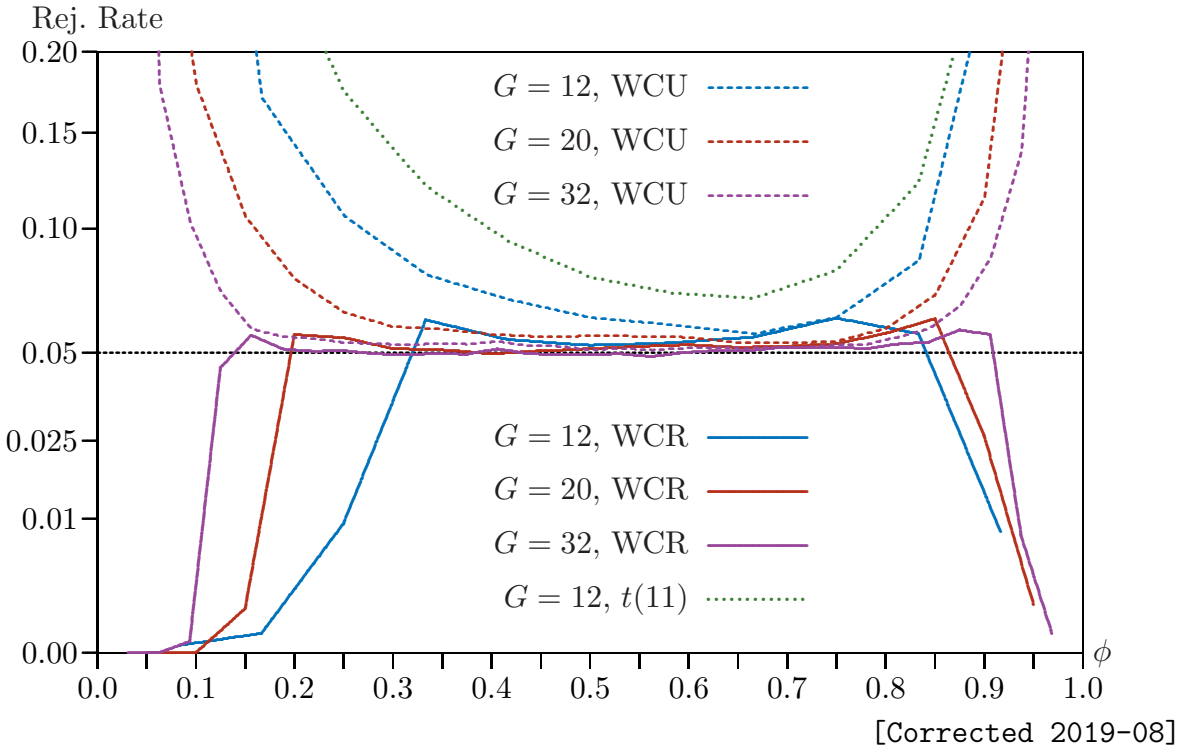
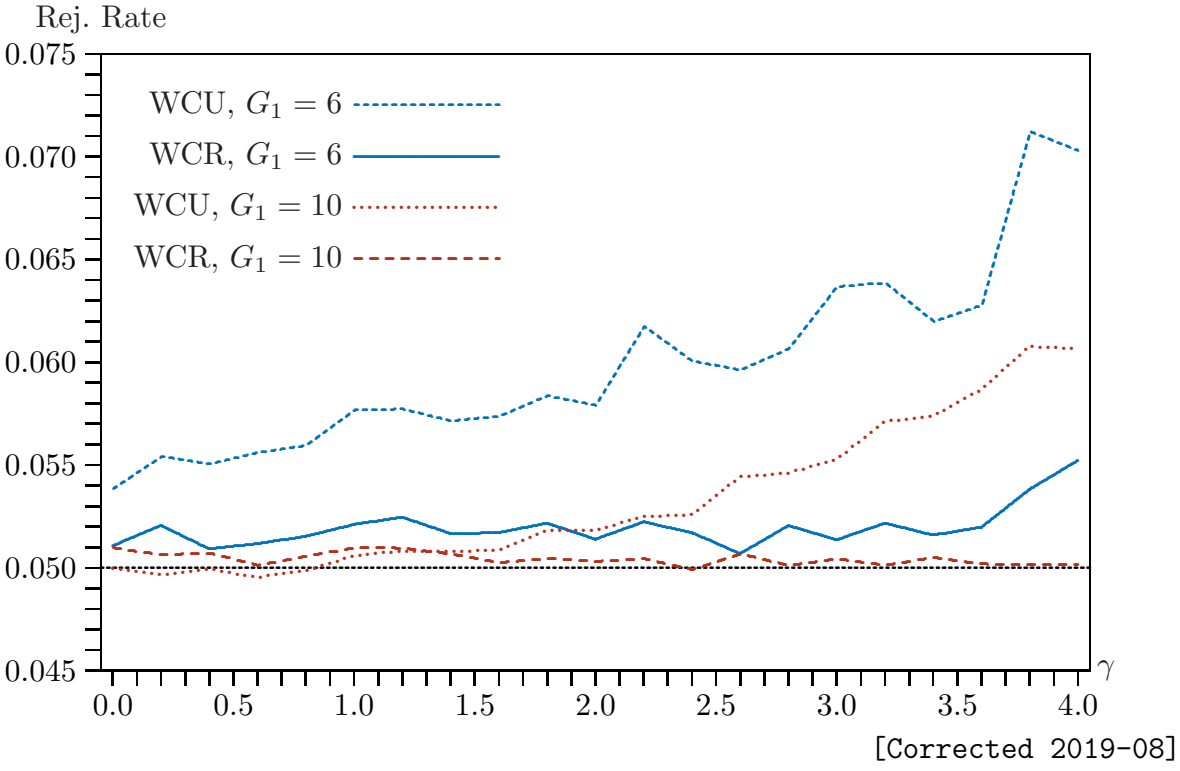


Figure A.8: Rejection frequencies for bootstrap tests as functions of γ for $G = 20$



Rejection frequencies for tests based on the $t(G-1)$ distribution have a "U" shape similar to the ones for the WCU bootstrap, but they are always larger (often much larger) than the latter. Results for $G = 12$ are shown in the figure. As the analysis in Section 6 predicts, overrejection is more severe for small values of ϕ than for large values, because clusters are being treated from smallest to largest.

Figure A.8 graphs rejection frequencies against γ for two cases with $G = 20$. In one case, $\phi = 0.3$, so that $G_1 = 6$. In the other, $\phi = 0.5$, so that $G_1 = 10$. The former is just at the edge of the region where the bootstrap seems to work acceptably, while the latter is well inside that region. In both cases, only 20% of the observations in the treated clusters are treated. Two patterns are evident from the figure. The WCR bootstrap almost always outperforms the WCU bootstrap, which overrejects in a large majority of cases. This overrejection becomes more severe as γ increases, probably because the sizes of the treated clusters become more dispersed. In contrast, the WCR bootstrap is much less sensitive to γ, although it clearly performs better for $\phi = 0.5$ than for $\phi = 0.3$.[12]

Equation (A.2) strongly suggests that inference based on CRVE t statistics, and probably also bootstrap inference, will become less reliable as the sizes of the treated clusters become more variable. In order to investigate this conjecture, we performed one more set of experiments. The model is still the DiD regression of equation (7), with $N = 2000$, $G = 40$, and $G_1 = 20$ in all cases. The 20 untreated clusters each have 50 observations. What varies across the experiments are the sizes of the 20 treated clusters. One cluster, numbered 1, has N_1 observations, with N_1 varying between 50 and 962, and the remaining 19 treated clusters each have $(1000 - N_1)/19$ observations. There are 100,000 replications.

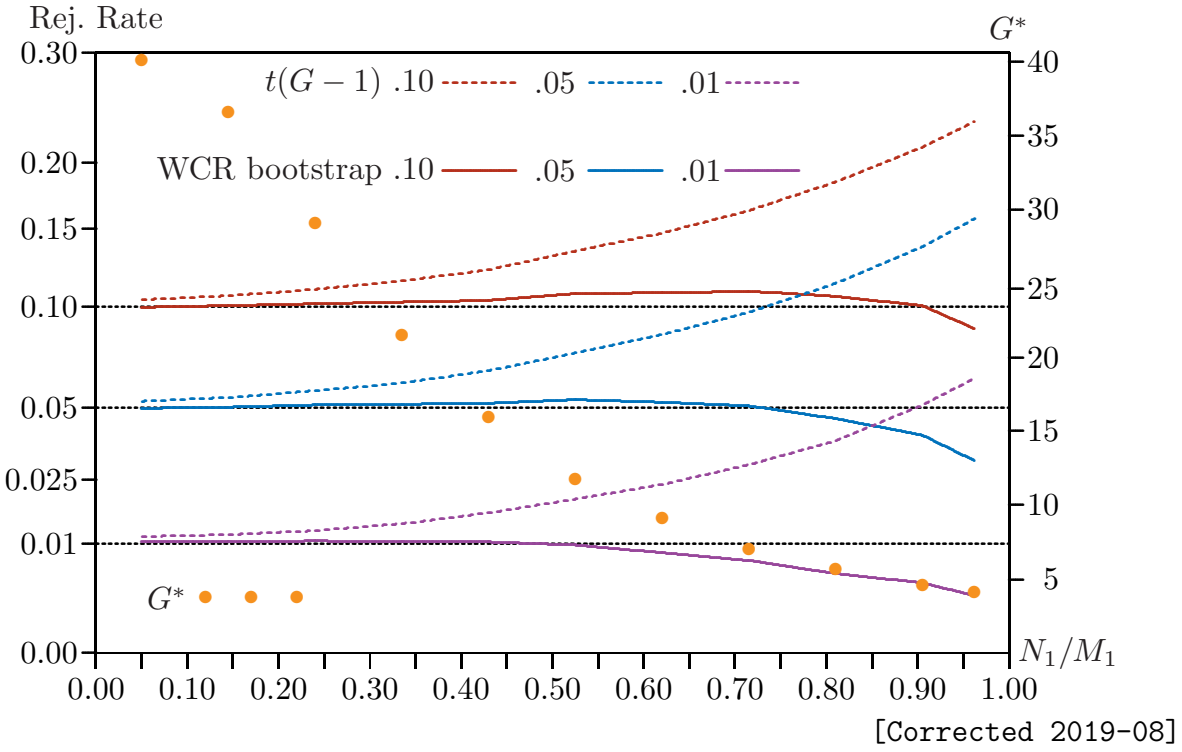
Figure A.9 shows rejection frequencies for CRVE t tests using $t(39)$ critical values and for WCR bootstrap tests at the .01, .05, and .10 levels. When N_1/M_1 is small or moderate in size, both tests perform well, as we would expect for such a large value of G_1. But as N_1/M_1 becomes larger, the performance of both tests deteriorates. The CRVE t tests overreject quite severely, and the WCR bootstrap tests may either overreject or underreject. The latter underrejects quite severely at the .01 level when $N_1/M_1 > 0.75$.

The figure also shows the values of G^*, calculated for the true value of $\rho = 0.05$, for each value of N_1/M_1; the scale is shown on the right-hand vertical axis. These values decline monotonically with N_1/M_1, eventually becoming very much smaller than $G_1 = 20$. These results, and others not reported, suggest that G^* is of considerable diagnostic value, with small values of G^* being associated with poor CRVE test performance.

The experiments of this section yield two principal results beyond the ones in the paper. The first is that wild cluster bootstrap tests perform very well indeed, even when G is very small and clusters are quite unbalanced, provided neither G_1 nor $G - G_1$ is too small and the sizes of the treated clusters are not extremely variable. The second is that the restricted wild cluster bootstrap (WCR) almost always outperforms the unrestricted one (WCU) when they both perform reasonably well. The latter usually rejects more often than the former.

[12] The somewhat ragged shapes of the rejection frequency curves in Figure A.8, especially when $G_1 = 6$, are not primarily due to experimental randomness. They arise because, as γ increases, the numbers of treated observations in clusters of various sizes change in a somewhat irregular fashion.

Figure A.9: Rejection frequencies as functions of N_1/M_1



A.7 Placebo Laws with Aggregate Data

Figure 10 presents some results for placebo laws experiments using aggregate data. It shows rejection frequencies for CRVE tests based on the $t(G-1)$ distribution and for restricted wild cluster bootstrap tests. However, because of the scale of the vertical axis, the WCR results are difficult to see clearly.

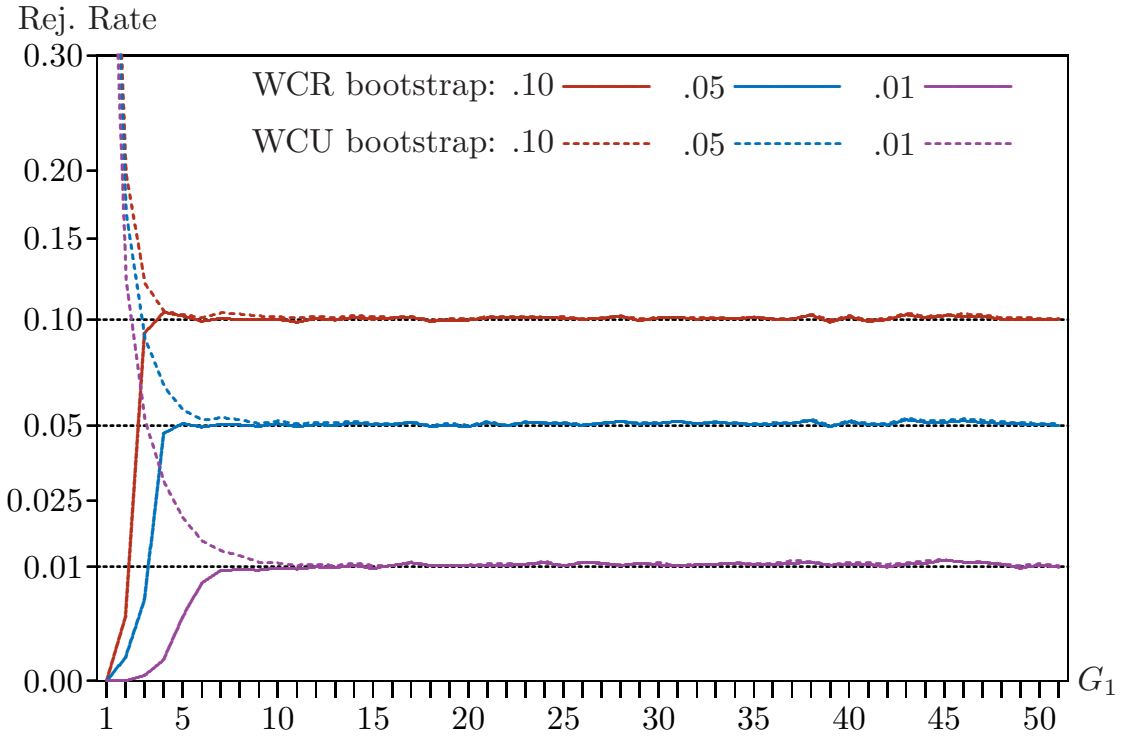
Figure A.10 is comparable to Figure A.4. Like the latter, it shows rejection frequencies for WCR and WCU bootstrap tests, but for aggregate rather than micro data. These are based on 100,000 replications with 399 bootstraps. It is evident that both bootstrap methods work extremely well for G_1 sufficiently large. The minimum value of G_1 that is needed seems to be larger for tests at the .01 level than for tests at higher levels, and perhaps slightly larger for WCU than for WCR.

As the analysis in Section 6 predicts, WCR underrejects severely, and WCU overrejects severely, when G_1 is small. For clarity, rejection frequencies for WCU when $G_1 = 1$ are not shown in the figure. They are 0.7522, 0.7184, and 0.6399 for tests at the .10, .05, and .01 levels, respectively. These are just a little bit lower than the rejection frequencies for CRVE t tests based on $t(G-1)$ critical values that are plotted in Figure 8.

A.8 Aggregation and Test Power

In Section 7, we showed that using aggregate data obtained by averaging over state-year pairs (or, in general, cluster-period pairs) can lead to better finite-sample performance under the null hypothesis. It does so by imposing perfectly balanced cluster sizes. This approach

Figure A.10: Rejection frequencies for placebo laws using aggregate data



could evidently lead to power loss if the estimates based on aggregate data were less efficient than the ones based on micro data. That seems to be what happens in the empirical example of Section 8. We discuss the issue of power loss in this section and present some simulation evidence.

Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where each observation is associated with a time period t as well as a cluster g. Let the number of time periods be T, and define $M \equiv GT$. Further, let M_m denote the number of observations for cell m, where $m = T(g-1)+t$ if the cells are ordered by cluster and then time period. Define the $M \times N$ averaging matrix $\boldsymbol{A}$ as the matrix with $1/M_m$ in every element of row m that corresponds to an observation belonging to that cell. Then $\boldsymbol{Ay}$ is an $M \times 1$ vector with typical element

$$\bar{y}_m = \frac{1}{M_m} \sum_{i=1}^{M_m} y_{gti},$$

where i indexes the observations associated with cell m. Similarly, $\boldsymbol{AX}$ is an $M \times k$ matrix with typical row $\bar{\boldsymbol{X}}_m$ consisting of averages of the $\boldsymbol{X}_{gti}$ over the observations in cell m.

The OLS estimator for the aggregate data is

$$\acute{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{Ay},$$

of which the covariance matrix is

$$(\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{A\Omega A}'\boldsymbol{AX}(\boldsymbol{X}'\boldsymbol{A}'\boldsymbol{AX})^{-1}. \tag{A.10}$$

This may be compared with the covariance matrix for the OLS estimator $\hat{\boldsymbol{\beta}}$ based on the original micro data:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{A.11}$$

Although it is easy to compare expressions (A.10) and (A.11) numerically, it is impossible to compare them analytically except in very special cases.

One such case is the dummy variable regression model discussed in Section 6 when every cell has the same number of elements. In this case,

$$\acute{\beta}_2 = \frac{(\boldsymbol{A}\boldsymbol{d} - \bar{d}\boldsymbol{A}\boldsymbol{\iota})'\boldsymbol{A}\boldsymbol{y}}{(\boldsymbol{A}\boldsymbol{d} - \bar{d}\boldsymbol{A}\boldsymbol{\iota})'(\boldsymbol{A}\boldsymbol{d} - \bar{d}\boldsymbol{A}\boldsymbol{\iota})} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}. \tag{A.12}$$

Note that $\bar{d}$ here is the sample mean of the elements of $\boldsymbol{A}\boldsymbol{d}$. It is the same as the $\bar{d}$ in equation (A.6), but that would not be true if every cell did not have the same number of elements. In this special case,

$$(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{y} = (M/N)(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{y},$$

and

$$(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota}) = (M/N)(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota}),$$

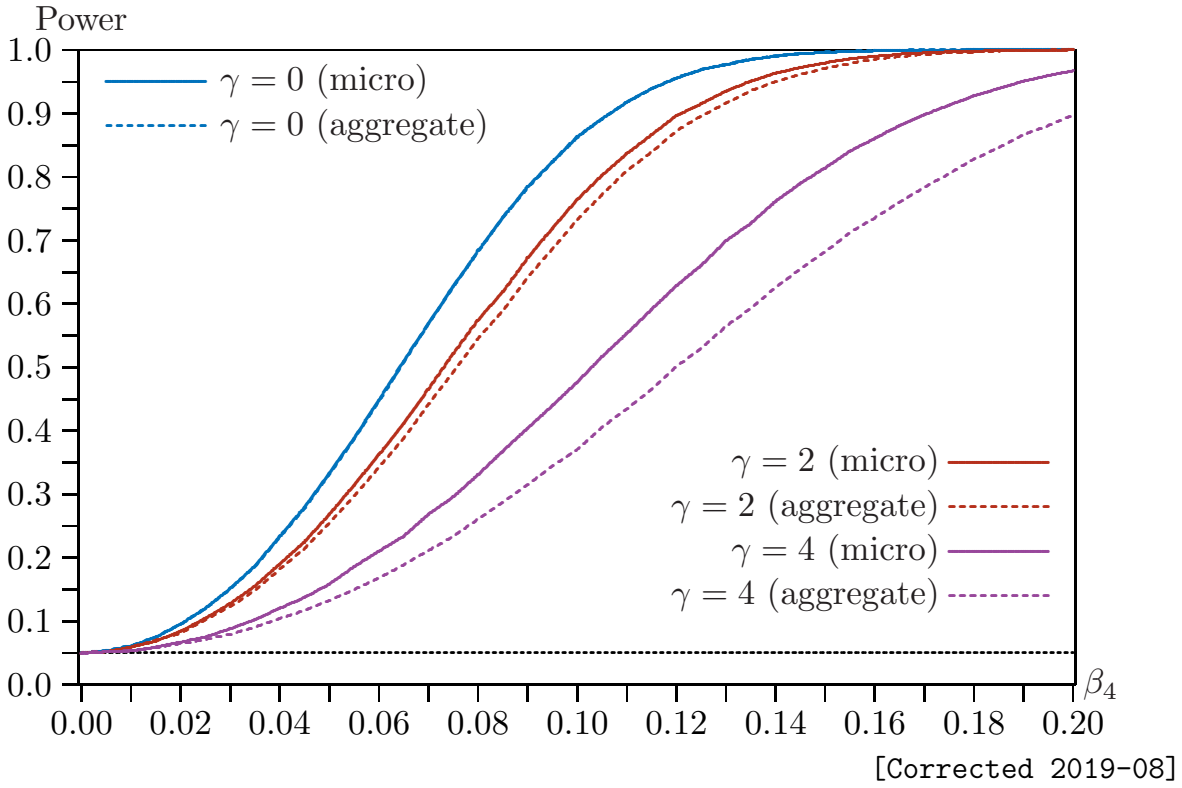
so that the aggregate estimator $\acute{\beta}_2$ given in equation (A.12) is equal to the micro estimator $\hat{\beta}_2$ given in equation (9).

The result that the two estimators are numerically identical is very special. It no longer holds if every cell does not have the same number of elements or if any regressor varies within cells. In fact, when regressors vary within cells, the efficiency loss from aggregation can be severe. For example, when the wage regression for the placebo laws data is estimated with no treatment variable using micro data, the coefficient on age is 0.02462 with a CRVE standard error of 0.00266. When the same regression is estimated using aggregate data, the coefficient is 0.02988 with a standard error of 0.04874. The aggregate standard error is 18.3 times larger than the micro one.

However, since treatment variables typically do not vary within cells, because they apply to every unit in a given cluster for a given time period, it is not clear whether we should expect significant power loss when using aggregate data for DiD regressions. Since neither $\hat{\boldsymbol{\beta}}$ nor $\acute{\boldsymbol{\beta}}$ is an efficient estimator, it is not even clear that the former will be more efficient than the latter.

In order to investigate this issue, we perform a series of experiments based on the DiD regression of equation (7), with 100,000 replications. Cluster sizes are determined by equation (A.9), which depends on a parameter $\gamma \geq 0$. In the experiments, $G = 50$, $N = 15,000$, $\rho_\epsilon = 0.05$, and every observation is assigned, with equal frequency, to one of 20 years. Half of the clusters are treated for 8 out of the 20 years. The micro regressions have 15,000 observations, and the aggregate ones have 1000 observations. The value of N is quite large here because we do not want there to be any empty cells. With $\gamma = 4$, there would be empty cells if N were much smaller than $15,000$. When $\gamma = 0$, every cluster has 300 observations, so each cell is an average of 15 observations. As γ increases, cluster sizes become more variable, and so do the numbers of observations per cell. For the largest value of γ that we

Figure A.11: Bootstrap power functions for several values of γ



investigate ($\gamma = 4$), some cells are averages of just one observation, and others are averages of 60.

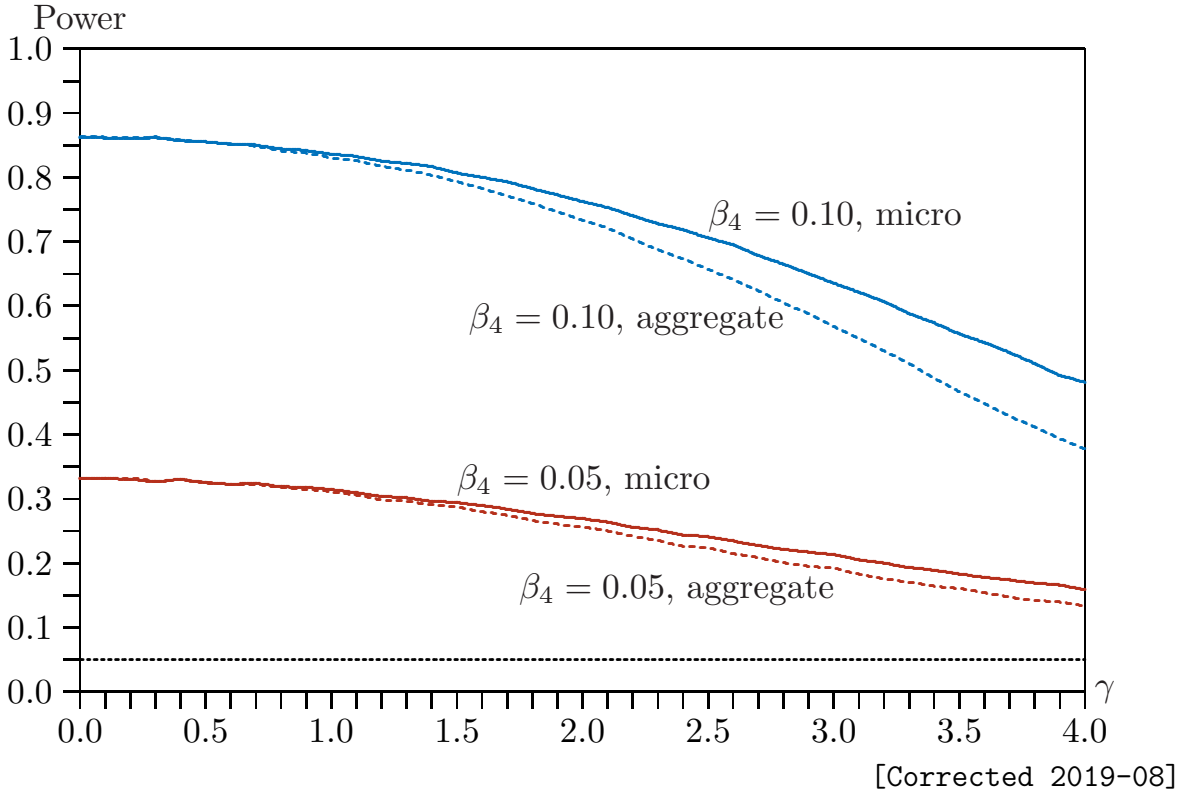
Figure A.11 shows power functions for bootstrap (WCR) tests for three values of γ. Not surprisingly, there is no power loss when $\gamma = 0$. With equal-sized clusters, homoskedasticity, and no within-cluster variation, the aggregate estimates are very similar (but, in this case, not identical) to the micro estimates. As γ increases, however, power declines, and the power loss from using aggregate data becomes steadily greater.

Both of these results should have been expected. Because of the intra-cluster correlation, the OLS estimates are most efficient when cluster sizes are balanced, so power inevitably declines as γ increases. The aggregate estimates give as much weight to cells based on few observations as to cells based on many observations, and this presumably causes them to be less efficient than the micro estimates.

Another way to see how power loss changes as clusters become more unbalanced is to graph power against γ for a given value of β_4. This is done in Figure A.12, which shows the power of WCR bootstrap tests for two values of β_4 (0.05 and 0.10) as functions of γ for $0 \leq \gamma \leq 4$. Power loss is modest for $\gamma \leq 1$ and quite substantial for $\gamma \geq 2$. This is what we would expect to see in view of the results in Figure A.11.

A very different way to investigate power is to modify the placebo laws experiments of Section 7 by adding a small amount to the regressand whenever an observation or cell is treated. In our experiments, we added 0.02, which is equivalent to increasing wages by 2.02%. We report results only for the WCR bootstrap, because it is generally the most

Figure A.12: Bootstrap power as a function of γ for two values of β_4



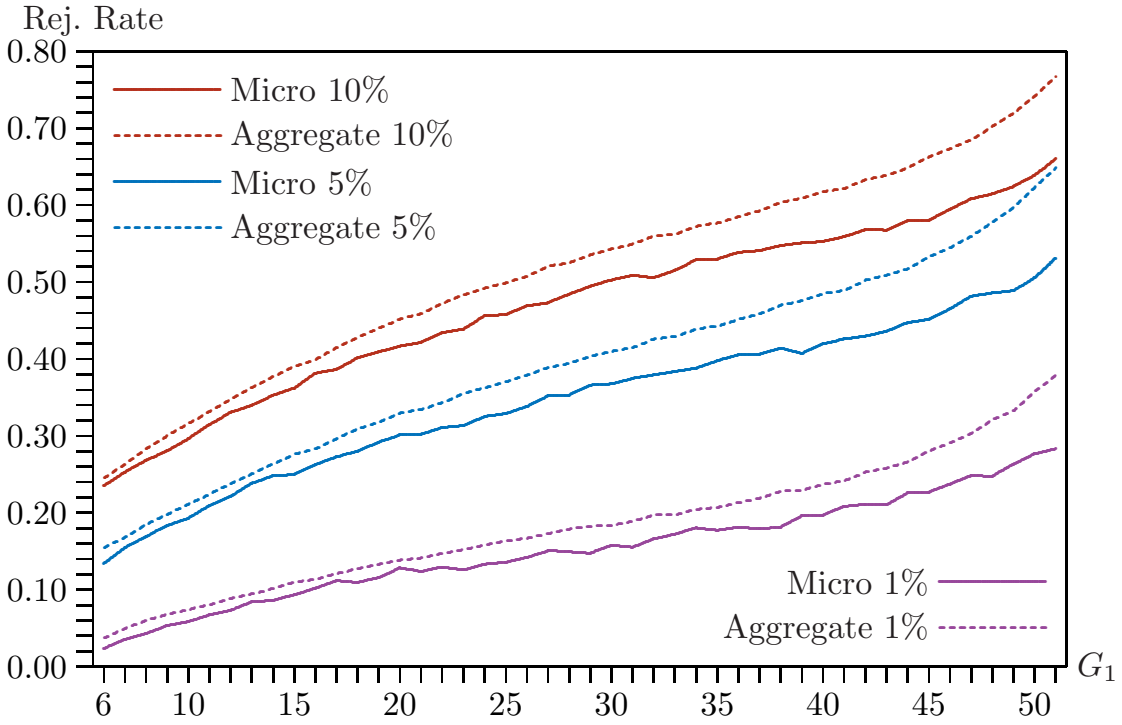
reliable procedure for controlling size. Our experiments used 100,000 replications for the aggregate data and 10,000 for the micro data, with 399 bootstraps in both cases. When we did this for $G_1 \geq 6$ (because the wild cluster bootstrap is not reliable for $G_1 \leq 5$), we obtained the results shown in Figure A.13.

Figure A.13 is surprising. If it is to be believed, aggregation actually increases power, especially when a large number of states is treated. The reason seems to be that, contrary to normal intuition, the micro estimates of β_{treat} are a bit more dispersed than the aggregate estimates. For example, when $G_1 = 20$, the standard errors are 0.01345 and 0.01287, respectively.

Whether the apparent gain in power from aggregation that appears in Figure A.13 should be taken seriously is not clear. Normally, in an experiment designed to measure power, we would hold the DGP constant and draw many realizations of the error terms and, hence, the dependent variable. In the placebo laws experiments, however, we hold most of the data constant and draw many different realizations of the test regressor, arbitrarily adding 0.02 to the regressand whenever the test regressor is equal to 1. Thus the distributions of the estimated coefficients and the distributions of the test statistics are not really the ones that would normally determine test power.

Based on these results, we might be tempted to conclude that, in cases where the test regressor does not vary within cells, aggregation is not likely to lead to very much loss of power. However, the empirical results in Section 8 suggest otherwise. Whether or not power

Figure A.13: Bootstrap power as a function of G_1 for placebo laws



loss due to aggregation is a problem in practice is often very easy to determine. If a test using aggregate data rejects the null hypothesis convincingly, then lack of power is evidently not a problem. Conversely, if a test using micro data that should be reasonably reliable, such as a WCR bootstrap test, rejects the null hypothesis, and a similar test using aggregate data does not, then lack of power is evidently a problem.