MacKinnon, James G.; Webb, Matthew D.

**Working Paper**

# Wild Bootstrap Inference for Wildly Different Cluster Sizes

Queen's Economics Department Working Paper, No. 1314

**Provided in Cooperation with:**
Queen’s Economics Department (QED), Queen’s University

Suggested Citation: MacKinnon, James G.; Webb, Matthew D. (2014) : Wild Bootstrap Inference for Wildly Different Cluster Sizes, Queen's Economics Department Working Paper, No. 1314, Queen's Economics Dep., Queen's Univ., Kingston, Ont.

This Version is available at:
http://hdl.handle.net/10419/97471

Queen's Economics Department Working Paper No. 1314

# Wild Bootstrap Inference for Wildly Different Cluster Sizes

James G. MacKinnon
Queen

Matthew D. Webb
University of Calgary

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

5-2014

# Wild Bootstrap Inference for Wildly Different Cluster Sizes *

James G. MacKinnon
jgm@econ.queensu.ca

Matthew D. Webb
mwebb@ucalgary.ca

May 27, 2014

**Abstract**

The cluster robust variance estimator (CRVE) relies on the number of clusters being large. A shorthand 'rule of 42' has emerged among practitioners, but we show that unbalanced clusters invalidate this rule. Monte Carlo evidence suggests that rejection frequencies are higher for datasets with 50 clusters proportional to US state populations rather than 50 balanced clusters. Using critical values based on either the wild cluster bootstrap or the 'effective number' of clusters performs much better. Simulations of placebo laws with dummy variable regressors also favor these alternative procedures, and an empirical example illustrates the consequences of using them.

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

## 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). This is the estimator that is used when the `cluster` command is invoked in Stata.

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

*We are grateful to two 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, and Indiana University for comments on earlier versions. We also thank Arthur Sweetman, Yulia Kotlyarova, and Yuanyuan Wan for helpful suggestions, and Kevin Schnepel and Doug Steigerwald for their computer code. MacKinnon's research was supported, in part, by a grant from the Social Sciences and Humanities Research Council.

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 fails to hold are now well-known; see, among others, Bertrand, Duflo and Mullainathan (2004), Donald and Lang (2007), and Brewer, Crossley and Joyce (2013). 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 (2013). Assumptions A2 and A3 were relaxed by Carter, Schnepel and Steigerwald (2013), which also showed how to calculate the "effective number" of clusters for cases with heterogeneous within-cluster correlation and unequal (unbalanced) cluster sizes. A thorough recent survey of cluster robust inference may be found in Cameron and Miller (2014).

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), have primarily used datasets with equal-sized clusters. Both papers also perform experiments with data from the Current Population Survey (CPS), as discussed in section 8. Most of the simulations in the former paper used aggregate data. The process of aggregation creates an average of the residuals for each state-year pair. This imposes assumption A3, because each state has only one observation per year. Some simulations did involve micro data (unbalanced, clustered by state or state-year pair), but the paper did not calculate rejection rates for 51 states with clustering at the state level. The latter paper conducted some simulations using micro data, but it also did not calculate CRVE rejection rates for 51 states. Table 1 provides a summary of previous experimental results. In the table, "N/A" corresponds to the lack of results for micro data and 51 states. One of the main contributions of this paper is to remedy this omission.

Previous results have led to a 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 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 estimates. For treatment effects, cluster sizes matter, but the proportion of clusters that is treated matters even more. The wild bootstrap works very well in most cases, but all the methods fail badly when that proportion is close to zero or (in some cases) to one. Section 5 uses simulated data, Section 6 explains why the wild bootstrap fails, and Section 7 briefly considers power. Section 8 extends the 'placebo laws' Monte Carlo experiments of Bertrand, Duflo and Mullainathan (2004). Section 9 contains a brief empirical example based on Angrist and Kugler (2008), and Section 10 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. 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

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

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 $\beta_k$, the last coefficient of $\boldsymbol{\beta}$. The procedure for using the wild cluster bootstrap of Cameron, Gelbach and Miller (2008) to test the hypothesis that $\beta_k = 0$ is as follows:

1. Estimate equation (1) by OLS.

2. Calculate $\hat{t}_k$, the $t$ statistic for $\beta_k = 0$, using the square root of the $k^{\text{th}}$ 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 13; see Webb (2013), which proposes an alternative for such cases.

5. For each bootstrap replication, estimate regression (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 $k^{\text{th}}$ diagonal element of (2), with bootstrap residuals replacing the OLS residuals, as the standard error.

6. Calculate the bootstrap $P$ value either as

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

or as

$$\hat{p}_{\text{et}}^{*} = 2\min\left(\frac{1}{B}\sum_{i=1}^{B} I\big(t_k^{*j} \leq \hat{t}_k\big), \frac{1}{B}\sum_{i=1}^{B} I\big(t_k^{*j} > \hat{t}_k\big)\right). \tag{5}$$

Equation (4) yields a symmetric bootstrap $P$ value, and equation (5) yields an equal-tailed bootstrap $P$ value; see MacKinnon (2006). For the experiments of this paper, the two $P$ values were always extremely close, and we report rejection frequencies based on $\hat{p}_{\text{s}}^{*}$. However, for two-tailed test statistics that do not have mean zero, it would be better to use $\hat{p}_{\text{et}}^{*}$.

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 $\epsilon_{ig}$, the bootstrap DGP also preserves these properties. However, it can cause serious problems in certain cases; see Section 6.

The second key feature of this procedure is that the bootstrap DGP (3) uses estimates under the null hypothesis. This improves its finite-sample properties (see Davidson and MacKinnon, 1999), but it means that it cannot be used directly to construct confidence intervals. Studentized bootstrap confidence intervals can easily be constructed when the null hypothesis is not imposed, but obtaining bootstrap intervals as accurate as the tests studied here would require an iterative procedure such as the one discussed in Davidson and MacKinnon (2014, Section 3).

# 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 (2013), 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 $\rho$ that measures within-cluster correlation. CSS focuses on the use of $G^*$ as a diagnostic. However, the paper also suggests, but does not investigate, 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.

The CSS procedure for computing $G^*$ is too detailed to describe here. However, it is important to mention one key step, which involves calculating the matrices

$$\boldsymbol{\gamma}_g \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g\,(\boldsymbol{X}'\boldsymbol{X})^{-1},\ g = 1, \ldots, G, \tag{6}$$

where the $\boldsymbol{\Omega}_g$ are $N_g \times N_g$ matrices with 1 on the principal diagonal and a constant $\rho$ everywhere else. Expression (6) may seem straightforward to calculate, but, when $N_g$ is large, the matrix $\boldsymbol{\Omega}_g$ can easily be so large that doing so is impractical. For example, in the placebo laws experiments of Section 8, the largest value of $N_g$ is 42,625. This implies that $\boldsymbol{\Omega}_g$ has 1,816,890,625 elements, which would require over 14 gigabytes of main memory to store.

Of course, it is possible to calculate the middle matrix in (6) without explicitly creating the matrix $\boldsymbol{\Omega}_g$ by using loops rather than matrix operations. This is feasible in a language like Fortran or C++. The program just involves four loops and one if statement. However, when $N_g$ is very large, even this method of calculating the middle matrix becomes extremely expensive; see Section 8.

CSS suggest setting $\rho = 1$. It seems more natural to estimate $\rho$, 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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{u}, \tag{7}$$

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 $\rho$ is $\hat{\rho} = \hat{\sigma}_\eta^2/(s^2 + \hat{\sigma}_\eta^2)$. Of course, this estimator is based on the assumption that the error terms are equicorrelated within each cluster with the same correlation for every cluster, as the form of the matrices $\Omega_g$ implies.

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 Kolesar (2012). Simulation results in Cameron and Miller (2014) suggest that the Imbens and Kolesar 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 $g = 1, \ldots, G$, and calculating the degrees-of-freedom parameter requires

even more extensive computations. In view of the large sample sizes in all our experiments (extremely large for the placebo laws experiments of Section 8 and the empirical example of Section 9), it was not feasible to study these procedures.[1] Another approach to inference in the model (1) was suggested by Ibragimov and Müller (2010), but it requires that $\beta_2$ be separately identifiable from the data for each cluster, which is not the case in most of our experiments.

# 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 $G^*$.

The simulations are grouped into four sets, two with 50 clusters and two with 100 clusters. For the 50-cluster simulations, one set has clusters of equal size, while the other set has clusters with sizes proportional to the US states without the District of Columbia. The 100-cluster simulations are constructed in a similar fashion, with the first set containing 100 equal-size clusters and the second containing two sets of clusters proportional to US state populations. In a sense, the latter dataset contains two Californias, two Ohios, and so on. These four sets of simulations allow us to test the implications of violating assumption A3, that cluster sizes are equal. Since 50 clusters satisfies the 'rule of 42,' we would expect to see reliable inference in all cases if the rule actually held.

The model is

$$y_{ig} = \beta_1 + \beta_2 X_{ig} + \epsilon_{ig}, \quad i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \tag{8}$$

where there are $G$ clusters, and the $g^{\text{th}}$ cluster has $N_g$ observations. Both the $X_{ig}$ and the $\epsilon_{ig}$ are standard normal and uncorrelated across clusters. The within-cluster correlation is $\rho_x$ for the $X_{ig}$ and $\rho_\epsilon$ for the $\epsilon_{ig}$. We do not allow $\rho_\epsilon$ 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. Within each set of simulations, we construct simulated data in which $\rho_x$ and $\rho_\epsilon$ are varied. All of the simulated datasets have 2000 observations, and each experiment involves 400,000 replications.[2] 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 $\epsilon_{ig}$ and use equation (8) to compute $y_{ig}$, with $\beta_2 = 0$.
3. Estimate equation (8) by OLS.
4. 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.

[1] 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 $B = 399$.

[2] 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 are essentially invariant to $N$.

5. Repeat steps (2), (3), and (4) 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 (7). For reasons of space, however, we report only two sets of results. We omit the results for $t(N-2)$ critical values, which always overrejected more than $t(G-1)$ ones. In most cases, basing $G^*$ on $\rho = \hat{\rho}$ worked better than basing it on $\rho = 0.99$. Therefore, the only results for the effective number of clusters that we report are for $t(G^*-1)$ based on $\hat{\rho}$.

Table 2 presents results from Monte Carlo simulations with 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 $\rho_\epsilon$ and $\rho_x$. When $\rho_x$ is close to 0, rejection rates are close to the desired .05 level. As $\rho_x$ gets closer to 1, however, they increase, aways exceeding 0.065 when $\rho_x = 1$. In general, increasing $\rho_\epsilon$ increases rejection rates slightly, although to a lesser degree than increasing $\rho_x$. The impact is most severe when $\rho_x$ is large but less than 1.

Using critical values from the $t(G^*-1)$ distribution frequently, but not always, results in more accurate inferences. However, there is a tendency to overreject when $\rho_x$ is large and to underreject when $\rho_\epsilon$ 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 $\rho$. In all other cases, setting $\rho = 0.99$ results in (often substantially) lower values of $G^*$ than using $\hat{\rho}$, which tends to cause noticeable underrejection. In the most extreme case ($\rho_\epsilon = 0.8$, $\rho_x = 0.0$), the rejection frequency was 0.0395 when $G^*$ was based on $\rho = 0.99$.

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 2 is 0.04942, and the largest is 0.05081. These numbers are close enough to .05 to be explained by chance. The standard error when the true rejection frequency is .05 is 0.000345, so the largest implied $t$ statistics for the hypothesis that the rejection frequency is 0.05 are less than 2.35.

Since assumption A3 holds, it is perhaps not surprising that the 'rule of 42' holds pretty well in these simulations. Table 3 presents results from a second set of experiments in which that assumption is severely violated. Cluster sizes are now roughly proportional to US state populations; the smallest clusters have just 4 observations, and the largest has 242. Even when $\rho_\epsilon$ and $\rho_x$ 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 $\rho_x$ leads to an increase in rejection rates. So does increasing $\rho_\epsilon$, 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 $\rho_x$ is large and generally becomes worse as $\rho_\epsilon$ 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 reported 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 results are obtained by using the wild bootstrap, although it does not work quite as well as it did with equal-sized clusters. Rejection frequencies range from 0.0497 to 0.0528. There is thus a very modest tendency to overreject in some cases, although this 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. Results for 100 clusters of size 20 are not shown, because they are largely predictable from the ones in Table 2. 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 shows what happens when there are 100 clusters that are roughly proportional to US state populations, with each state appearing twice. The two smallest clusters have just 2 observations, and the two largest have 121. The CRVE rejection frequencies are always closer to .05 than with only 50 clusters, and they exhibit patterns similar to those in Table 3. There is often substantial overrejection with $t(G-1)$ critical values and substantial underrejection with $t(G^* - 1)$ critical values. As before, underrejection is more severe with $G^*(0.99)$ (not reported) instead of $G^*(\hat{\rho})$. The wild bootstrap does not perform flawlessly (it overrejects more often than it underrejects), but since its rejection rate never exceeds 0.0514, it performs extraordinarily well overall.

The results in Tables 2 to 4 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, especially when $G^*$ is small, is apparently to use the wild 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. In order to investigate this type of application, we conducted two sets of experiments. In the first set, 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 these experiments, there is no role for $\rho_x$, and $\rho_\epsilon$ seems to have little effect on rejection frequencies. What appears to matter is $P$.

In Figures 1, 2, and 3, we report results for 50 clusters with 2000 observations, $\rho_\epsilon = 0.50$, and $P$ that varies between 0.02 and 0.98 at intervals of 0.02. 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 tests based on CRVE standard errors and $t(G - 1)$ critical values. There is very severe overrejection when $P$, the proportion of clusters treated, is close to 0 or 1. This result is consistent with Monte Carlo results in Bell and McCaffrey (2002) and Conley and Taber (2011). The latter paper develops procedures for inference when there are just a few treated groups. Another procedure for inference when there is only one treated group has been developed by Abadie, Diamond and Hainmueller (2010)

based on the idea of "synthetic controls."

In Figure 1, overrejection is quite modest when $P$ is far from 0 and 1. With equal-sized clusters, rejection frequencies are very close to 0.05 for $P$ between 0.35 and 0.65. With state-sized clusters, they are somewhat higher, never falling below 0.063. The graph for equal-sized clusters is symmetric around $P = 0.50$, while the one for state-sized clusters is somewhat asymmetric. In the latter case, overrejection is a bit more severe when $P$ is very small (so that only a few small clusters are treated) than when it is very large (so that only a few large clusters are not treated). For clarity, the figure does not show results for state-sized clusters with the largest states treated first, which would be the mirror image of the ones with the smallest states treated first.

Figure 2 shows results for tests based on $t(G^* - 1)$ critical values. $G^*$ is based on $\hat{\rho}$, but the results would have been almost identical for other values of $\rho$ because $G^*$ is invariant to $\rho$ in the equal-sized case and very insensitive to it in the state-sized cases. There is extreme underrejection when $P = 0.02$ and $P = 0.98$, 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 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 $0.06 \leq P \leq 0.94$. The rejection frequencies appear to be symmetric around $P = 0.50$ for equal-sized clusters, but quite asymmetric for state-sized ones.

Figure 3 shows results for wild bootstrap tests based on simulations with 399 bootstraps. In all cases, there is severe underrejection when $P$ is very close to either 0 or 1. In the most extreme cases, there are no rejections at all. For equal-sized clusters, there is modest overrejection when the proportion of treated or untreated clusters is between 0.08 and 0.12, but the wild bootstrap tests work extremely well for $P$ between about 0.14 and 0.86.

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 $P$ is very close to 0 or 1, and they overreject severely when $P$ is between 0.88 and 0.96. When the states are treated from largest to smallest, the opposite problem occurs, with severe overrejection when $P$ is between 0.04 and 0.12, and severe underrejection when $P$ is very close to 0 or 1. The reason why the wild bootstrap fails for extreme values of $P$ will be discussed in Section 6.

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

$$y_{igt} = \beta_1 + \beta_2 \mathrm{GT}_{igt} + \beta_3 \mathrm{PT}_{igt} + \beta_4 \mathrm{GT}_{igt} \mathrm{PT}_{igt} + \epsilon_{igt}, \tag{9}$$

for $i = 1, \ldots, N_g$, $g = 1, \ldots, G$, and $t = 1, \ldots, T$. Here $\mathrm{GT}_{igt}$ is a "group treated" dummy that equals 1 if group $g$ is treated in any time period, and $\mathrm{PT}_{igt}$ is a "period treated" dummy that equals 1 if any group is treated in time period $t$. The coefficient of most interest is $\beta_4$, which shows the effect on treated groups in periods when there is treatment.

Figure 4, which shows results for $t(G - 1)$ critical values, is comparable to Figure 1, except that it is based on equation (9) with either zero or half of the observations in each

cluster treated.[3] Thus $\text{PT}_{igt} = 1$ for half the observations in each cluster, while $\text{GT}_{igt} = 1$ for $P$% of the clusters, with $P$ once again varying between 0.02 and 0.98 at intervals of 0.02. The results for equal-sized clusters are quite similar, but the ones for state-sized clusters are very different. For small values of $P$, there is very severe overrejection when the smallest clusters are treated first. For large values of $P$, 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 (9) is the same when the fraction of observations treated is $P$ as it is when that fraction is $1 - P$. We may conclude from the figure that overrejection tends to be most severe when $\min(P, 1-P)$ is small and the observations that are in the minority are from the smallest clusters.

Figure 5, which shows results for $t(G^* - 1)$ critical values, is comparable to Figure 2. In all cases, $G^*$ was invariant to $\rho$. Except when the proportion of clusters treated is very small or very large, the tests always overreject, but they never do so severely.

Figure 6, which shows results for the wild bootstrap, is comparable to Figure 3, and the results for equal-sized clusters are very similar in the two figures. However, the results for state-sized clusters are much better in Figure 6, closer to the results for equal-sized clusters. When between approximately 12% and 88% of clusters are treated, the wild bootstrap works extremely well. This range of excellent performance is extended somewhat when the observations that are in the minority are from the largest clusters.

It is common to allow for cluster fixed effects in models like equation (9) by dropping the constant term and the $\text{GT}_{ig}$ variable and adding $G$ dummy variables, one for each cluster. We repeated the DiD experiments for this case and obtained results, not reported, that were quite similar to the ones in Figures 4, 5, and 6. Most importantly, the wild bootstrap continued to work extremely well for $0.12 \leq P \leq 0.88$.

# 6 Why the Wild Bootstrap Can Fail

As we have seen, wild bootstrap tests tend to underreject, often very severely, when the proportion of treated clusters, $P$, is close to 0 or 1. In this section, we explain why this happens. For simplicity, consider the dummy variable regression model

$$y_{ig} = \beta_1 + \beta_2 d_{ig} + \epsilon_{ig}, \tag{10}$$

where $d_{ig}$ equals 1 for the first $PG$ clusters and 0 for the remaining $(1-P)G$ clusters. There is no loss of generality in restricting attention to the model (10), since adding additional regressors would not change the analysis in any fundamental way.

When $\beta_2 = 0$, the OLS estimate of $\beta_2$ in regression (10) is

$$\hat{\beta}_2 = \left( \sum_{g=1}^{G} \sum_{i=1}^{N_g} (d_{ig} - P)^2 \right)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (d_{ig} - P)\epsilon_{ig}, \tag{11}$$

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

because the sample mean of the $d_{ig}$ is just $P$. The estimate $\hat{\beta}_2$ depends on all the $\epsilon_{ig}$, but the weight for treated observations is proportional to $1 - P$, and the weight for untreated ones is proportional to $-P$. Therefore, when $P$ is close to 0, the error term for each treated observation receives a great deal more weight than the one for each untreated observation. When $P$ is close to 1, the opposite is true. Thus equation (11) implies that, on average, the $\epsilon_{ig}$ for treated (untreated) observations must be unusually large in absolute value whenever $|\hat{\beta}_2|$ is large and $P$ is close to 0 (close to 1).

In the special case of the dummy variable regression (10), the wild cluster bootstrap DGP (3) is just

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

where $\tilde{\beta}_1$ is the sample mean of the $y_{ig}$, and $\tilde{\epsilon}_{ig} = \epsilon_{ig} - \tilde{\beta}_1$. Consider the extreme case in which $PG = 1$, so that only observations with $g = 1$ are treated. For the Rademacher distribution, each of the bootstrap error terms for $g = 1$ can have just two values, namely, $\tilde{\epsilon}_{i1}$ and $-\tilde{\epsilon}_{i1}$. Therefore, the values of the bootstrap dependent variable for the treated cluster will tend to be far from the mean $\beta_1$ precisely when $|\hat{\beta}_2|$ is large.

This explains why the wild cluster bootstrap underrejects when $PG = 1$. The $t$ statistic $\hat{t}_2$ for $\hat{\beta}_2 = 0$ is likely to be large in absolute value when $|\hat{\beta}_2|$ happens to be large. In this case, the $|y_{i1}^* - \tilde{\beta}_1|$ must necessarily also be large, on average. This causes the distribution of the bootstrap $t$ statistics $t_2^{*j}$ to be bimodal with a large variance. In consequence, when $|\hat{t}_2|$ is large, $|t_2^{*j}|$ tends to be even larger for a substantial number of bootstrap samples, so that it is rare to obtain a bootstrap $P$ value below the critical value of .05.[4]

When $PG = 2$, there are four possible pairs of values for the first two bootstrap error terms. Two of these pairs are $[\tilde{\epsilon}_{i1}, \tilde{\epsilon}_{i2}]$ and $[-\tilde{\epsilon}_{i1}, -\tilde{\epsilon}_{i2}]$, which must yield large values of $|t_2^{*j}|$, while the other two are $[\tilde{\epsilon}_{i1}, -\tilde{\epsilon}_{i2}]$ and $[-\tilde{\epsilon}_{i1}, \tilde{\epsilon}_{i2}]$, which are not assured to do so. Thus the distribution of the $t_2^{*j}$ will still tend to have a large variance when $|\hat{t}_2|$ is large, but not as large as when $PG = 1$.

For the Rademacher distribution, the number of possible sets of bootstrap error terms for the treated observations is $2^{PG}$. As $PG$ increases, the proportion of bootstrap samples for which the bootstrap error terms are equal to plus or minus the restricted residuals for every treated observations rapidly declines. Once $PG$ becomes large enough, the problem goes away.

Although the reason for the failure of the wild cluster bootstrap when $PG$ is large or small is easiest to see in the case of the Rademacher distribution, the problem exists for any valid choice of the distribution of the $v_g^{*j}$, because the wild bootstrap requires that $E(v_g^{*j}) = 1$. When $PG = 1$, the bootstrap errors for the first cluster will tend to be centered at $\tilde{\epsilon}_{i1}$ and $-\tilde{\epsilon}_{i1}$, even if they are not actually equal to those values. The problem remains when, for example, the $v_g^{*j}$ follow the standard normal distribution, and limited experimental evidence suggests that it is no less severe.

The foregoing analysis suggests that the key parameter is $PG$ not $P$. Figure 7 provides support for this conclusion by showing what happens when $G$ is varied for the case of equal-

[4] Of course, this argument implicitly assumes that the disturbances are homoskedastic. If the variance of the $\epsilon_{ig}$ for the only treated cluster were unusually large, the link between large values of $|\hat{\beta}_2|$ and $|\hat{t}_2|$ would break down. In such cases, the wild cluster bootstrap will almost certainly not work well, but it may not always underreject.

sized clusters. In the figure, $G$ takes the values 25, 50, 100, and 200, with $N_g = 40$ in the first three cases and (to save time) $N_g = 20$ in the last one. The left-hand panel, which shows rejection frequencies as a function of $P$, demonstrates that the range of values of $P$ for which the wild bootstrap yields accurate inferences becomes wider as $G$ increases. The right-hand panel, which shows rejection frequencies as a function of $PG$, demonstrates that the relationship between rejection frequencies and the number of treated clusters is essentially invariant to the total number of clusters. These results suggest that, for the DiD case with equal-sized clusters, the cluster wild bootstrap can probably be used safely when the number of treated clusters is between roughly 8 and $G-8$, but it is likely to be seriously misleading when $PG \leq 4$.

# 7 Power of the Tests

Suppose that an asymptotic test overrejects under the null hypothesis and a bootstrap version of it has approximately the correct rejection frequency. Then it is very likely that the asymptotic test will reject more often under every alternative than the bootstrap test. While it might seem that the former test is more powerful, its additional power would actually be spurious, and its apparently higher power would provide no reason to prefer the asymptotic test.

Based on all the results reported so far, it seems likely that wild bootstrap tests and tests which use $t(G^* - 1)$ critical values will have less power than conventional tests which use $t(G-1)$ critical values. Provided the power loss is moderate, however, this should not be a cause for concern. To investigate this issue, we perform a second set of experiments based on the DiD regression model (9). These use 100,000 replications and $B = 999$. The number of bootstrap samples is higher than in the other experiments, because bootstrap tests suffer from power loss that is proportional to $1/B$; see Davidson and MacKinnon (2000).

We actually perform several sets of experiments, but we report results for only one of them, which is typical. Figure 8 shows rejection frequencies for three tests of $\beta_4 = 0$ at the .05 level in equation (9). Cluster sizes are proportional to state sizes, and the smallest 20% of states are treated. As anticipated, the test that uses $t(G-1)$ critical values always rejects more often than the one that uses $t(G^* - 1)$ critical values, which in turn always rejects more often than the wild bootstrap test. In this case, the latter rejects just over 5.1% of the time under the null, so that its performance is almost perfect. Achieving almost perfect size necessarily involves a small loss of power.

# 8 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 US Current Population Survey. The dependent variable is the log of weekly earnings for women aged 25-50 from 1979 to 1999. The objective of the simulations is to show how often random difference-in-differences coefficients are found to be significant if we ignore the intra-cluster correlation in the data. 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 the log of women's earnings is

$$\ln(\text{earnings}) = \beta_1 + \beta_{\text{treat}}\text{TREAT} + \text{YEARS}\,\boldsymbol{\beta}_{\text{years}} + \text{STATES}\,\boldsymbol{\beta}_{\text{states}} + \text{controls} + \epsilon, \quad (12)$$

where YEARS and STATES are full sets of fixed effects, and the controls are a quadratic in age 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 (12), the treatment variable is instead set to 1 randomly, so that $\beta_{\text{treat}}$ should be insignificantly different from zero. This is repeated for many replications with different random values of the treatment variable. If the tests were working properly, we would expect $\beta_{\text{treat}}$ to be significantly different from zero 5% of the time when testing at the 5% 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, 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, the ones in BDM. In that paper, states are always sorted into equal proportions of treatment and control states. Instead, 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 size 547,518. Our simulations are also different in that we use the micro data throughout, whereas the majority of the BDM simulations use data that are aggregated at the state-year level.

We estimate equation (12) for each replication and compute four different tests of the hypothesis that $\beta_{\text{treat}} = 0$. The first employs a $t$ statistic based on the classic heteroskedasticity robust standard error and the $N(0,1)$ distribution. The second employs a $t$ statistic based on cluster robust standard errors and the $t(G-1)$ distribution, as we did 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 $\rho$. The fourth uses a $P$ value computed by the wild cluster bootstrap technique. Results for the first three tests are based on 100,000 replications. However, the ones for the wild bootstrap tests are based on only 10,000 replications because of computational cost.

The largest cluster (California) contains 42,625 observations. This makes it impossible to compute expression (6) using the $N_g \times N_g$ matrix $\boldsymbol{\Omega}_g$ and extremely expensive to compute it using loops in Fortran. A single replication using optimized code takes about 40 minutes. In order to make the experiments feasible, we therefore compute $G^*$ using only 1/100 of each sample. The entire sample is used to estimate $\hat{\rho}$, which is always about 0.031 or 0.032, but 99 out of each 100 observations are discarded before computing $G^*$. 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.

Figure 9 plots rejection frequencies against 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 exceed 0.60 whenever 42 or fewer states are treated, and they exceed 0.50 except when all 51 states are treated. Of course, we would not expect standard errors that are robust to heteroskedasticity but not to clustering to yield valid inferences in this case.

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. However, this procedure is still not very reliable. Rejection frequencies at the .05 level always exceed 0.117, and they exceed 0.20 when five or fewer states are treated. This is in stark contrast with the aggregate results in BDM, which suggest that cluster robust inference is quite reliable when $G = 50$. Table 1 reproduces the rejection frequency of 0.063 for aggregate data that appears in their Table VIII. Based on the results in Figure 9, a conservative estimate of the rejection frequency for micro data that corresponds to the "N/A" in Table 1 is 0.120.

In contrast to the results in Figures 1 and 4, there is quite substantial overrejection in the middle of Figure 9, but it increases much less rapidly as the proportion of states treated becomes large. Because a random number of years for each state is being treated, treating many states is evidently not equivalent to treating few states for these experiments.

Figure 10 plots rejection frequencies against the number of states treated for $t(G^* - 1)$ critical values and for wild cluster bootstrap tests. The latter are based on symmetric bootstrap $P$ values, but results for equal-tail $P$ values would have looked very similar. Both tests underreject severely when $PG$ is very small. For larger values, the wild bootstrap performs extremely well, but using $t(G^* - 1)$ critical values leads to moderate overrejection which increases with the number of treated states.

# 9 Empirical Example

The results in Sections 4, 5, and 8 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 in practice by replicating a few select results from Angrist and Kugler (2008).[5]

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

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

concentrated in the rural areas. The equation can be written as

$$\begin{aligned} y_{ijt} &= \boldsymbol{X}_i'\boldsymbol{\mu} + \beta_j + \delta_t + \alpha_{0,95-97}\, g_{jt,95-97} + \alpha_{0,98-00}\, g_{jt,98-00} \\ &\quad + \alpha_{1,95-97}\, d_{jt,95-97} + \alpha_{1,98-00}\, d_{jt,98-00} + \epsilon_{ijt}. \end{aligned} \tag{13}$$

Here $\boldsymbol{X}_i$ is a vector of control variables, $\beta_j$ is a department dummy, $\delta_t$ is a year dummy, the $\alpha_0$ terms are DiD coefficients for the rural growing areas, and the $\alpha_1$ terms are DiD coefficients for the urban areas. We replicate two estimated equations from the paper's Table 6, namely, column 6 and column 9. The former estimates log hours for men, and it excludes departments that are medium producers of coca from the analysis. The latter estimates log hours for teenage boys, and it includes the medium producers.[6] The sample for men has 181,882 observations, and the sample for teenage boys has 22,141 observations. In 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 bootstrap $P$ values are symmetric and are based on 99,999 bootstrap samples (far more than would normally be needed); equal-tail bootstrap $P$ values, not reported, are almost identical. There are two $t(G^*)$ $P$ values, one based on $\hat{\rho} = 0.025$ and the other based on $\rho = 0.99$, and two corresponding $t(G^*-1)$ $P$ values. 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 (13), 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 $\rho$ is much larger for the boys than for the men. Using $\rho = 0.99$ instead of $\rho = 0.155$ changes the values of $G^*$ quite a bit, but its effect on the $P$ values is fairly modest because none of the $G^*$ values is really small.

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 $2 \times 2$ 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 (2014), 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.

[6] These regressions were chosen in part because the estimated $P$ values implicitly reported for the test regressors are quite low.

# 10 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. With continuous regressors, there can be serious overrejection when clusters are of wildly unequal sizes. With dummy regressors, there can be extremely severe overrejection when the proportion 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. Similar results are found for placebo laws, where there tends to be substantial overrejection in all cases, which becomes very severe when the proportion of treated clusters is small.

These results contrast with 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 critical values based on the the effective number of clusters, as suggested by Carter, Schnepel and Steigerwald (2013), instead of the actual number, often improves matters substantially, although it does not always work as well as one might hope. In contrast, with one notable exception, the wild cluster bootstrap generally yields very reliable inferences for the cases we study, which involve sample sizes of 1000 or more with 50 or more clusters. The exception is that it can underreject very severely when only a few clusters are treated or untreated, for reasons that are explained in Section 6. In order to obtain reliable inference with the wild cluster bootstrap, it appears that there should be at least 7 or 8 treated clusters (and also, in many cases, at least 7 or 8 untreated ones).

# 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 8

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 3, 14, 23

Angrist, Joshua D., and Jorn-Steffen Pischke (2008) *Mostly Harmless Econometrics: An Empiricist's Companion,* 1 ed. (Princeton University Press) 2

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 5, 8

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), pp. 249–275 2, 3, 12, 16, 19

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 4

Brewer, Mike, Thomas F. Crossley, and Robert Joyce (2013) 'Inference with difference-in-differences revisted.' Technical Report, Institute for Fiscal Studies 2

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 2, 3, 16, 19

Cameron, A.C., and D.L. Miller (2014) 'A practitoner's guide to cluster robust inference.' *Journal of Human Resources.* forthcoming 2, 5, 15

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2013) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' Technical Report, University of California, Santa Barbara 2, 5, 16

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 8

Davidson, Russell, and Emmanuel Flachaire (2008) 'The wild bootstrap, tamed at last.' *Journal of Econometrics* 146(1), 162 – 169 4

Davidson, Russell, and James G. MacKinnon (1999) 'The size distortion of bootstrap tests.' *Econometric Theory* 15(3), 361–376 4

Davidson, Russell, and James G. MacKinnon (2000) 'Bootstrap tests: How many bootstraps?' *Econometric Reviews* 19(1), 55–68 12

Davidson, Russell, and James G. MacKinnon (2004) *Econometric Theory and Methods,* 1 ed. (Oxford University Press) 5

Davidson, Russell, and James G. MacKinnon (2014) 'Bootstrap confidence sets with weak instruments.' *Econometric Reviews* 33(6), 651–675 4

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 2, 4, 5

Ibragimov, Rustam, and Ulrich K. Müller (2010) 't-statistic based correlation and heterogeneity robust inference.' *Journal of Business & Economic Statistics* 28(4), 453–468 6

Imbens, Guido W., and Michal Kolesar (2012) 'Robust standard errors in small samples: Some practical advice.' Working Paper 18478, National Bureau of Economic Research, October 5

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

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

MacKinnon, James G. (2006) 'Bootstrap methods in econometrics.' *The Economic Record* 82(s1), S2–S18 4

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 5

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 1

Webb, Matthew D. (2013) 'Reworking wild bootstrap based inference for clustered errors.' Working Papers 1315, Queen's University, Department of Economics, August 2, 4

White, Halbert (1984) *Asymptotic Theory for Econometricians* (Academic Press) 2

# Tables and Figures

Table 1: Summary of Previous Monte Carlo Experiments

| | micro data | | aggregate | |
|---|---|---|---|---|
| | G=10 | G= 51 | G=10 | G=50 |
| OLS | – | 0.675 | 0.443 | 0.490 |
| cluster state-year | *0.444 | 0.440 | – | – |
| cluster state | *0.100 | N/A | 0.080 | 0.063 |

**Notes:** The table shows estimated rejection frequencies for tests at the .05 level. All results are from simulations using data from the Current Population Survey reported in Bertrand, Duflo and Mullainathan (2004) and Cameron, Gelbach and Miller (2008). An asterisk indicates that results are from the latter paper. In all cases, half (or very close to half) of the clusters are treated. Our simulations in section 8 would replace the 'N/A' in the table with a value of 0.120.

Table 2: Rejection Frequencies with 50 Equal-Sized Clusters

| | | $\rho_x$ | | | | | |
|---|---|---|---|---|---|---|---|
| $\rho_\epsilon$ | | **0.0** | **0.2** | **0.4** | **0.6** | **0.8** | **1.0** |
| **0.0** | $\mathbf{t(G-1)}$ | 0.0503 | 0.0503 | 0.0508 | 0.0530 | 0.0572 | 0.0658 |
| | $\mathbf{t(G^*-1)}$ | 0.0500 | 0.0500 | 0.0496 | 0.0498 | 0.0501 | 0.0535 |
| | **bootstrap** | 0.0501 | 0.0498 | 0.0497 | 0.0503 | 0.0505 | 0.0494 |
| **0.1** | $\mathbf{t(G-1)}$ | 0.0502 | 0.0505 | 0.0507 | 0.0535 | 0.0587 | 0.0661 |
| | $\mathbf{t(G^*-1)}$ | 0.0499 | 0.0501 | 0.0494 | 0.0499 | 0.0511 | 0.0541 |
| | **bootstrap** | 0.0496 | 0.0502 | 0.0498 | 0.0500 | 0.0503 | 0.0503 |
| **0.2** | $\mathbf{t(G-1)}$ | 0.0505 | 0.0504 | 0.0519 | 0.0551 | 0.0602 | 0.0663 |
| | $\mathbf{t(G^*-1)}$ | 0.0503 | 0.0498 | 0.0498 | 0.0498 | 0.0518 | 0.0540 |
| | **bootstrap** | 0.0498 | 0.0498 | 0.0498 | 0.0498 | 0.0501 | 0.0502 |
| **0.3** | $\mathbf{t(G-1)}$ | 0.0506 | 0.0505 | 0.0524 | 0.0565 | 0.0619 | 0.0665 |
| | $\mathbf{t(G^*-1)}$ | 0.0502 | 0.0495 | 0.0489 | 0.0498 | 0.0520 | 0.0543 |
| | **bootstrap** | 0.0501 | 0.0495 | 0.0494 | 0.0502 | 0.0501 | 0.0502 |
| **0.4** | $\mathbf{t(G-1)}$ | 0.0506 | 0.0511 | 0.0535 | 0.0573 | 0.0630 | 0.0666 |
| | $\mathbf{t(G^*-1)}$ | 0.0500 | 0.0493 | 0.0487 | 0.0493 | 0.0523 | 0.0542 |
| | **bootstrap** | 0.0499 | 0.0499 | 0.0501 | 0.0501 | 0.0505 | 0.0503 |
| **0.5** | $\mathbf{t(G-1)}$ | 0.0502 | 0.0517 | 0.0538 | 0.0579 | 0.0628 | 0.0661 |
| | $\mathbf{t(G^*-1)}$ | 0.0491 | 0.0487 | 0.0472 | 0.0489 | 0.0518 | 0.0540 |
| | **bootstrap** | 0.0495 | 0.0503 | 0.0501 | 0.0503 | 0.0500 | 0.0502 |
| **0.6** | $\mathbf{t(G-1)}$ | 0.0511 | 0.0520 | 0.0541 | 0.0586 | 0.0630 | 0.0665 |
| | $\mathbf{t(G^*-1)}$ | 0.0495 | 0.0478 | 0.0462 | 0.0483 | 0.0516 | 0.0544 |
| | **bootstrap** | 0.0502 | 0.0505 | 0.0504 | 0.0501 | 0.0497 | 0.0503 |
| **0.7** | $\mathbf{t(G-1)}$ | 0.0507 | 0.0516 | 0.0543 | 0.0593 | 0.0642 | 0.0663 |
| | $\mathbf{t(G^*-1)}$ | 0.0479 | 0.0455 | 0.0452 | 0.0484 | 0.0525 | 0.0541 |
| | **bootstrap** | 0.0498 | 0.0501 | 0.0502 | 0.0504 | 0.0507 | 0.0500 |
| **0.8** | $\mathbf{t(G-1)}$ | 0.0503 | 0.0514 | 0.0552 | 0.0587 | 0.0631 | 0.0664 |
| | $\mathbf{t(G^*-1)}$ | 0.0457 | 0.0439 | 0.0451 | 0.0476 | 0.0514 | 0.0541 |
| | **bootstrap** | 0.0496 | 0.0503 | 0.0508 | 0.0498 | 0.0499 | 0.0501 |
| **0.9** | $\mathbf{t(G-1)}$ | 0.0500 | 0.0508 | 0.0543 | 0.0592 | 0.0633 | 0.0655 |
| | $\mathbf{t(G^*-1)}$ | 0.0430 | 0.0416 | 0.0433 | 0.0480 | 0.0515 | 0.0537 |
| | **bootstrap** | 0.0495 | 0.0497 | 0.0498 | 0.0505 | 0.0498 | 0.0496 |

**Notes:** Rejection frequencies at the .05 level are based on 400,000 replications. There are 50 equal-sized clusters with 2000 observations. The effective number of clusters is $G^*(\hat{\rho})$. Wild bootstrap $P$ values are based on 399 bootstraps using the Rademacher distribution.

Table 3: Rejection Frequencies with 50 State-Sized Clusters

| | | $\rho_x$ | | | | | |
|---|---|---|---|---|---|---|---|
| $\rho_\epsilon$ | | **0.0** | **0.2** | **0.4** | **0.6** | **0.8** | **1.0** |
| **0.0** | $\mathbf{t(G-1)}$ | 0.0594 | 0.0589 | 0.0597 | 0.0628 | 0.0673 | 0.0821 |
| | $\mathbf{t(G^*-1)}$ | 0.0514 | 0.0467 | 0.0368 | 0.0297 | 0.0259 | 0.0255 |
| | **bootstrap** | 0.0504 | 0.0497 | 0.0497 | 0.0508 | 0.0508 | 0.0507 |
| **0.1** | $\mathbf{t(G-1)}$ | 0.0595 | 0.0593 | 0.0625 | 0.0680 | 0.0789 | 0.0938 |
| | $\mathbf{t(G^*-1)}$ | 0.0517 | 0.0466 | 0.0377 | 0.0311 | 0.0289 | 0.0283 |
| | **bootstrap** | 0.0508 | 0.0499 | 0.0505 | 0.0506 | 0.0514 | 0.0513 |
| **0.2** | $\mathbf{t(G-1)}$ | 0.0588 | 0.0614 | 0.0680 | 0.0769 | 0.0889 | 0.1019 |
| | $\mathbf{t(G^*-1)}$ | 0.0508 | 0.0460 | 0.0369 | 0.0319 | 0.0310 | 0.0294 |
| | **bootstrap** | 0.0501 | 0.0506 | 0.0514 | 0.0518 | 0.0522 | 0.0514 |
| **0.3** | $\mathbf{t(G-1)}$ | 0.0588 | 0.0636 | 0.0712 | 0.0810 | 0.0945 | 0.1053 |
| | $\mathbf{t(G^*-1)}$ | 0.0509 | 0.0447 | 0.0339 | 0.0298 | 0.0300 | 0.0297 |
| | **bootstrap** | 0.0502 | 0.0510 | 0.0516 | 0.0515 | 0.0519 | 0.0516 |
| **0.4** | $\mathbf{t(G-1)}$ | 0.0584 | 0.0648 | 0.0737 | 0.0840 | 0.0964 | 0.1060 |
| | $\mathbf{t(G^*-1)}$ | 0.0501 | 0.0409 | 0.0308 | 0.0276 | 0.0290 | 0.0294 |
| | **bootstrap** | 0.0498 | 0.0505 | 0.0519 | 0.0522 | 0.0522 | 0.0517 |
| **0.5** | $\mathbf{t(G-1)}$ | 0.0583 | 0.0666 | 0.0739 | 0.0852 | 0.0982 | 0.1072 |
| | $\mathbf{t(G^*-1)}$ | 0.0495 | 0.0381 | 0.0273 | 0.0265 | 0.0287 | 0.0293 |
| | **bootstrap** | 0.0498 | 0.0516 | 0.0515 | 0.0523 | 0.0522 | 0.0520 |
| **0.6** | $\mathbf{t(G-1)}$ | 0.0585 | 0.0663 | 0.0755 | 0.0864 | 0.0987 | 0.1065 |
| | $\mathbf{t(G^*-1)}$ | 0.0484 | 0.0335 | 0.0251 | 0.0252 | 0.0281 | 0.0289 |
| | **bootstrap** | 0.0502 | 0.0510 | 0.0521 | 0.0525 | 0.0522 | 0.0512 |
| **0.7** | $\mathbf{t(G-1)}$ | 0.0585 | 0.0679 | 0.0756 | 0.0869 | 0.0987 | 0.1065 |
| | $\mathbf{t(G^*-1)}$ | 0.0467 | 0.0304 | 0.0233 | 0.0243 | 0.0270 | 0.0290 |
| | **bootstrap** | 0.0507 | 0.0524 | 0.0522 | 0.0523 | 0.0516 | 0.0515 |
| **0.8** | $\mathbf{t(G-1)}$ | 0.0583 | 0.0678 | 0.0762 | 0.0874 | 0.0994 | 0.1073 |
| | $\mathbf{t(G^*-1)}$ | 0.0436 | 0.0272 | 0.0219 | 0.0237 | 0.0274 | 0.0289 |
| | **bootstrap** | 0.0514 | 0.0524 | 0.0524 | 0.0521 | 0.0520 | 0.0520 |
| **0.9** | $\mathbf{t(G-1)}$ | 0.0566 | 0.0673 | 0.0765 | 0.0884 | 0.0994 | 0.1073 |
| | $\mathbf{t(G^*-1)}$ | 0.0377 | 0.0233 | 0.0210 | 0.0237 | 0.0265 | 0.0288 |
| | **bootstrap** | 0.0514 | 0.0520 | 0.0525 | 0.0528 | 0.0521 | 0.0510 |

**Notes:** Rejection frequencies at the .05 level are based on 400,000 replications. There are 50 clusters proportional to US state populations with 2000 observations. The effective number of clusters is $G^*(\hat{\rho})$. Wild bootstrap $P$ values are based on 399 bootstraps using the Rademacher distribution.

Table 4: Rejection Frequencies with 100 State-Sized Clusters

| | | $\rho_x$ | | | | | |
|---|---|---|---|---|---|---|---|
| $\rho_\epsilon$ | | **0.0** | **0.2** | **0.4** | **0.6** | **0.8** | **1.0** |
| **0.0** | $\mathbf{t(G-1)}$ | 0.0547 | 0.0553 | 0.0557 | 0.0570 | 0.0604 | 0.0685 |
| | $\mathbf{t(G^*-1)}$ | 0.0509 | 0.0484 | 0.0408 | 0.0344 | 0.0314 | 0.0307 |
| | **bootstrap** | 0.0499 | 0.0503 | 0.0503 | 0.0502 | 0.0501 | 0.0499 |
| **0.1** | $\mathbf{t(G-1)}$ | 0.0551 | 0.0551 | 0.0567 | 0.0599 | 0.0658 | 0.0735 |
| | $\mathbf{t(G^*-1)}$ | 0.0510 | 0.0482 | 0.0413 | 0.0359 | 0.0341 | 0.0330 |
| | **bootstrap** | 0.0504 | 0.0499 | 0.0504 | 0.0505 | 0.0512 | 0.0503 |
| **0.2** | $\mathbf{t(G-1)}$ | 0.0552 | 0.0554 | 0.0590 | 0.0639 | 0.0700 | 0.0792 |
| | $\mathbf{t(G^*-1)}$ | 0.0511 | 0.0474 | 0.0413 | 0.0371 | 0.0355 | 0.0350 |
| | **bootstrap** | 0.0503 | 0.0496 | 0.0502 | 0.0510 | 0.0501 | 0.0505 |
| **0.3** | $\mathbf{t(G-1)}$ | 0.0551 | 0.0570 | 0.0610 | 0.0666 | 0.0743 | 0.0824 |
| | $\mathbf{t(G^*-1)}$ | 0.0508 | 0.0476 | 0.0404 | 0.0364 | 0.0365 | 0.0363 |
| | **bootstrap** | 0.0505 | 0.0502 | 0.0506 | 0.0507 | 0.0509 | 0.0510 |
| **0.4** | $\mathbf{t(G-1)}$ | 0.0554 | 0.0570 | 0.0622 | 0.0683 | 0.0753 | 0.0825 |
| | $\mathbf{t(G^*-1)}$ | 0.0510 | 0.0460 | 0.0385 | 0.0355 | 0.0355 | 0.0358 |
| | **bootstrap** | 0.0504 | 0.0497 | 0.0506 | 0.0507 | 0.0500 | 0.0505 |
| **0.5** | $\mathbf{t(G-1)}$ | 0.0550 | 0.0583 | 0.0631 | 0.0695 | 0.0778 | 0.0833 |
| | $\mathbf{t(G^*-1)}$ | 0.0501 | 0.0448 | 0.0362 | 0.0342 | 0.0359 | 0.0359 |
| | **bootstrap** | 0.0498 | 0.0507 | 0.0502 | 0.0504 | 0.0514 | 0.0507 |
| **0.6** | $\mathbf{t(G-1)}$ | 0.0552 | 0.0586 | 0.0644 | 0.0709 | 0.0777 | 0.0836 |
| | $\mathbf{t(G^*-1)}$ | 0.0497 | 0.0423 | 0.0348 | 0.0333 | 0.0348 | 0.0352 |
| | **bootstrap** | 0.0502 | 0.0504 | 0.0513 | 0.0510 | 0.0506 | 0.0504 |
| **0.7** | $\mathbf{t(G-1)}$ | 0.0547 | 0.0597 | 0.0642 | 0.0712 | 0.0783 | 0.0834 |
| | $\mathbf{t(G^*-1)}$ | 0.0479 | 0.0401 | 0.0327 | 0.0322 | 0.0343 | 0.0354 |
| | **bootstrap** | 0.0499 | 0.0506 | 0.0507 | 0.0510 | 0.0509 | 0.0503 |
| **0.8** | $\mathbf{t(G-1)}$ | 0.0547 | 0.0588 | 0.0644 | 0.0711 | 0.0789 | 0.0843 |
| | $\mathbf{t(G^*-1)}$ | 0.0464 | 0.0368 | 0.0310 | 0.0309 | 0.0339 | 0.0357 |
| | **bootstrap** | 0.0502 | 0.0501 | 0.0505 | 0.0505 | 0.0509 | 0.0508 |
| **0.9** | $\mathbf{t(G-1)}$ | 0.0543 | 0.0587 | 0.0641 | 0.0708 | 0.0793 | 0.0852 |
| | $\mathbf{t(G^*-1)}$ | 0.0437 | 0.0334 | 0.0292 | 0.0304 | 0.0335 | 0.0356 |
| | **bootstrap** | 0.0504 | 0.0499 | 0.0501 | 0.0504 | 0.0511 | 0.0510 |

**Notes:** Rejection frequencies at the .05 level are based on 400,000 replications. There are 100 clusters proportional to US state populations with 2000 observations. The effective number of clusters is $G^*(\hat{\rho})$. Wild bootstrap $P$ values are based on 399 bootstraps using the Rademacher distribution.

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

| **Panel A** | | | | |
|---|---|---|---|---|
| **Log Hours – Adult Men – No Medium Producers (32 departments)** | | | | |
| | rural 95-97 | urban 98-00 | rural 95-97 | urban 98-00 |
| coeff. | 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 |
| $G^*_{0.99}$ | 5.540 | 3.082 | 4.318 | 1.424 |
| $P$ values: | | | | |
| $t(G-1)$ | 0.045 | 0.002 | 0.044 | 0.071 |
| $t(G^*_{0.025})$ | 0.079 | 0.035 | 0.096 | 0.240 |
| $t(G^*_{0.99})$ | 0.085 | 0.041 | 0.099 | 0.251 |
| $t(G^*_{0.025}-1)$ | 0.087 | 0.059 | 0.114 | 0.447 |
| $t(G^*_{0.99}-1)$ | 0.096 | 0.073 | 0.118 | 0.497 |
| bootstrap | 0.186 | 0.028 | 0.092 | 0.090 |
| **Panel B** | | | | |
| **Log Hours – Teenage Boys – All Producers (38 departments)** | | | | |
| | rural 95-97 | rural 98-00 | urban 95-97 | urban 98-00 |
| coeff. | 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 |
| $G^*_{0.99}$ | 6.942 | 8.905 | 6.265 | 10.686 |
| $P$ values: | | | | |
| $t(G-1)$ | 0.028 | 0.048 | 0.954 | 0.605 |
| $t(G^*_{0.155})$ | 0.047 | 0.064 | 0.955 | 0.611 |
| $t(G^*_{0.99})$ | 0.057 | 0.072 | 0.955 | 0.613 |
| $t(G^*_{0.155}-1)$ | 0.051 | 0.067 | 0.955 | 0.611 |
| $t(G^*_{0.99}-1)$ | 0.063 | 0.076 | 0.956 | 0.614 |
| bootstrap | 0.030 | 0.050 | 0.958 | 0.628 |
| **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 |

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

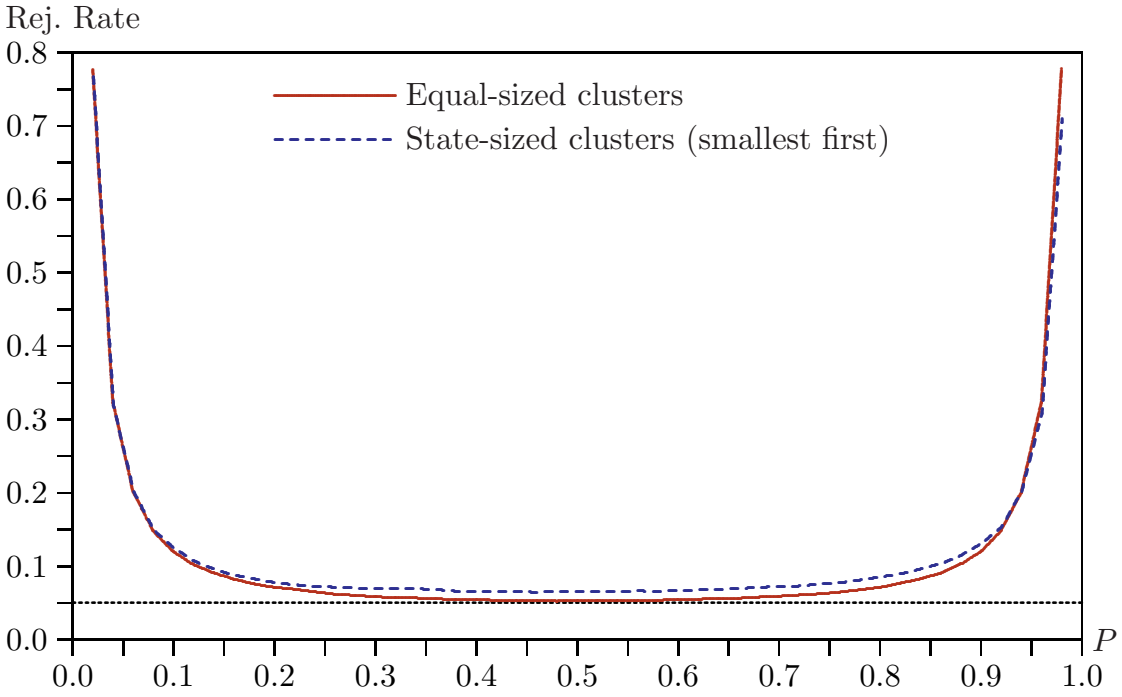


Figure 2: Rejection rates and proportion of clusters treated, $t(G^*-1)$

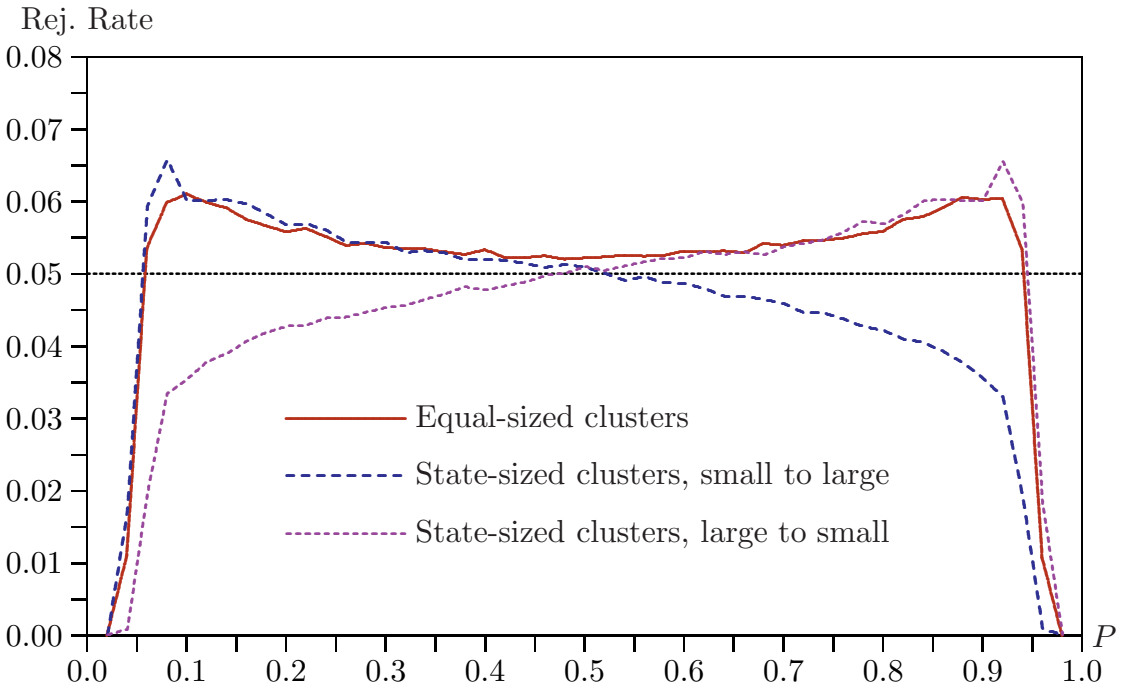

Figure 3: Rejection rates and proportion of clusters treated, wild bootstrap

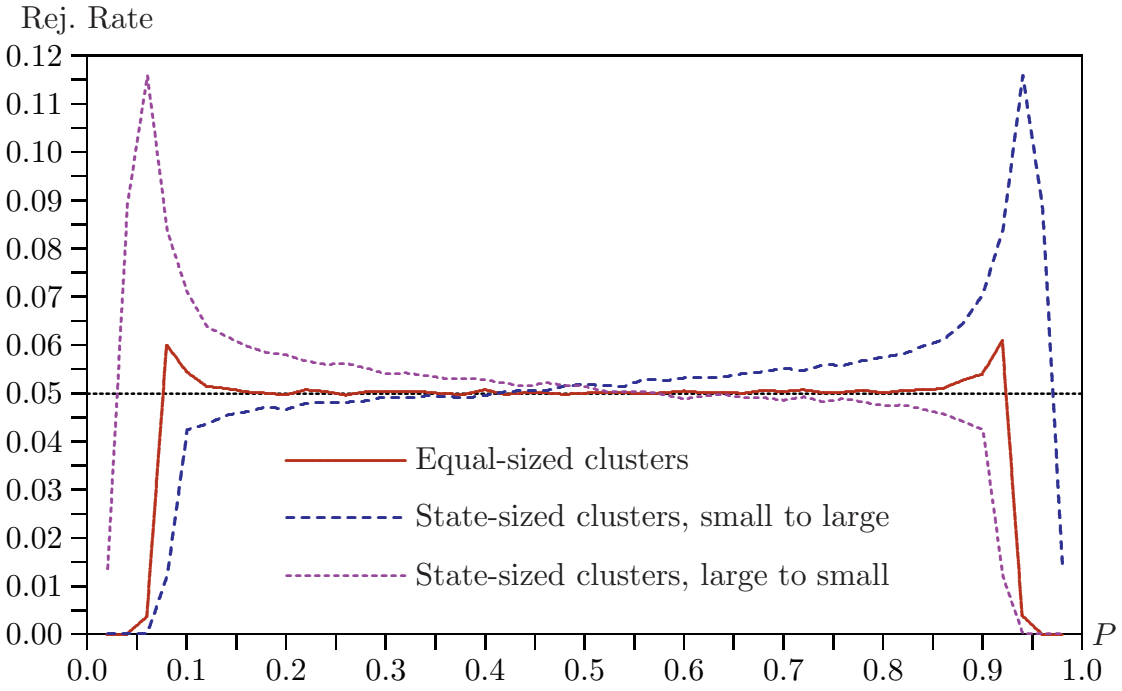


Figure 4: Rejection rates and proportion treated, DiD, $t(G-1)$)

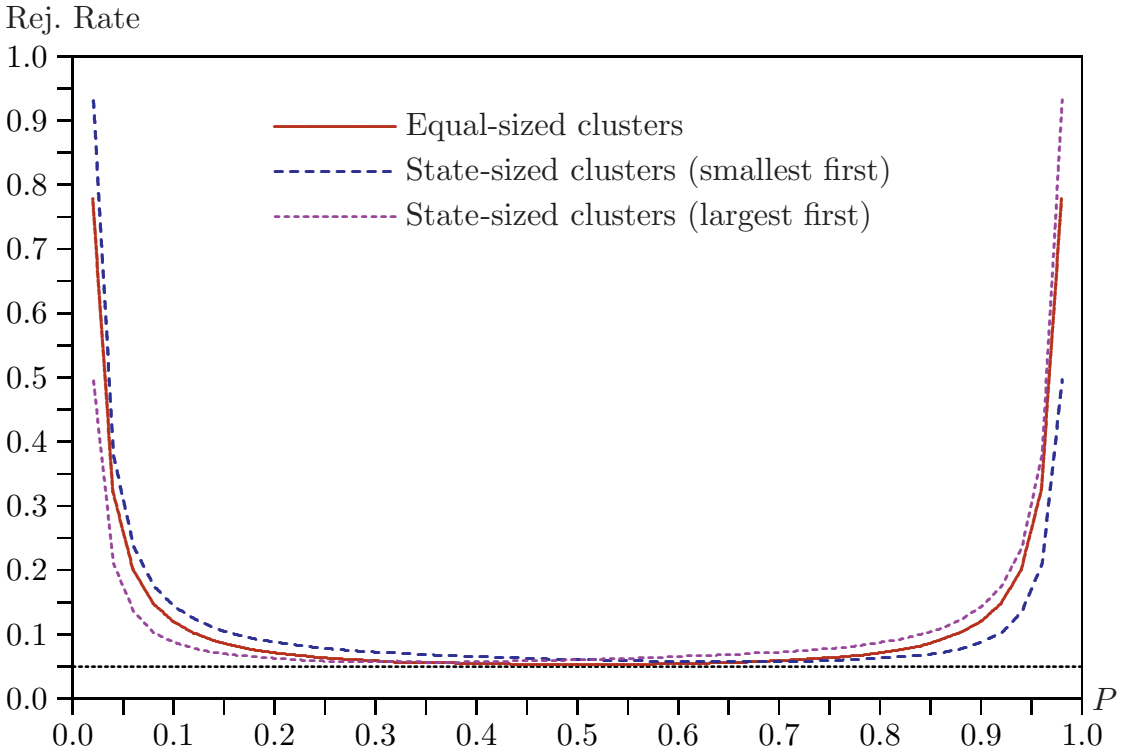

Figure 5: Rejection rates and proportion treated, DiD, $t(G^* - 1)$

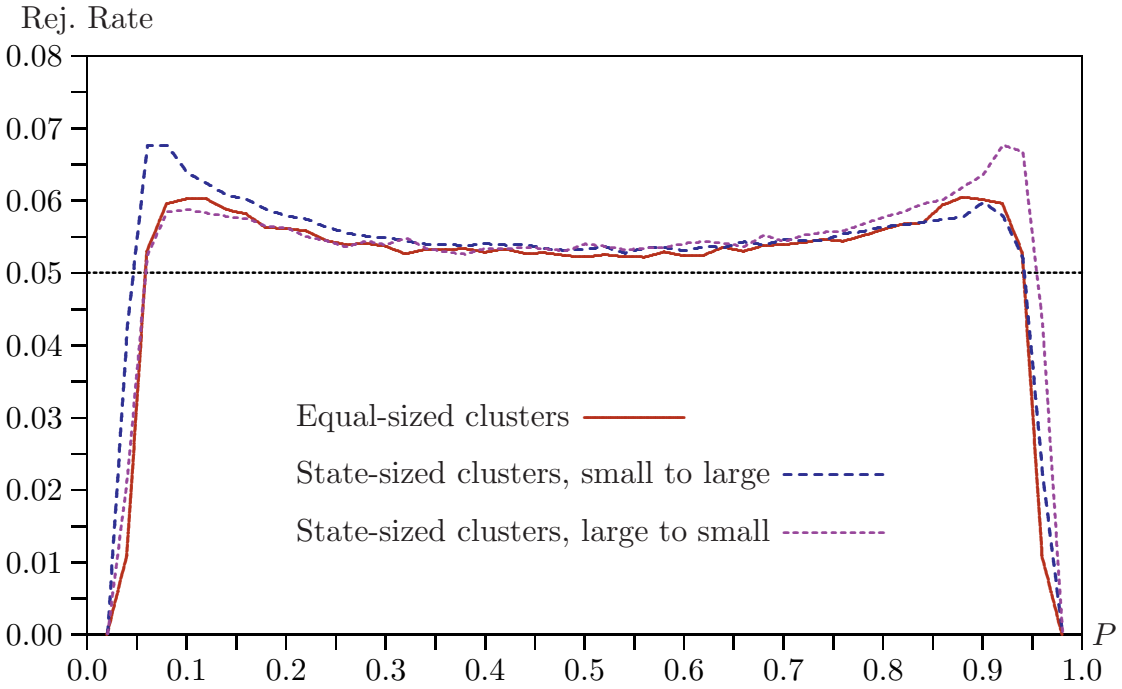


Figure 6: Rejection rates and proportion treated, DiD, wild bootstrap

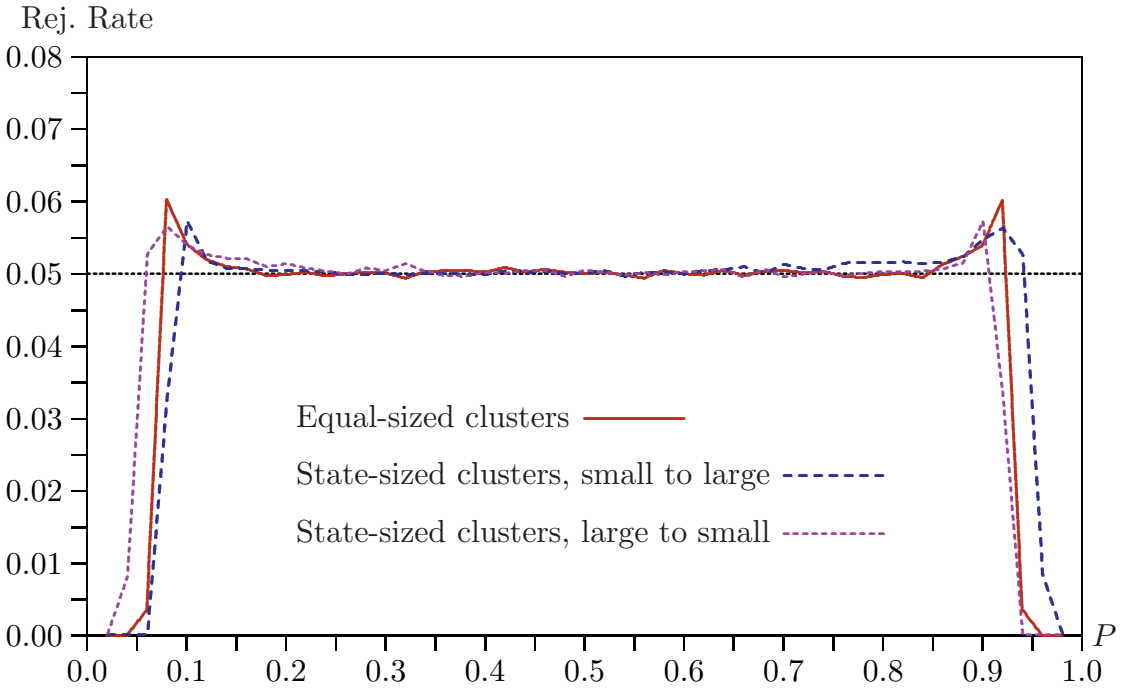

Figure 7: Wild bootstrap rejection rates, DiD

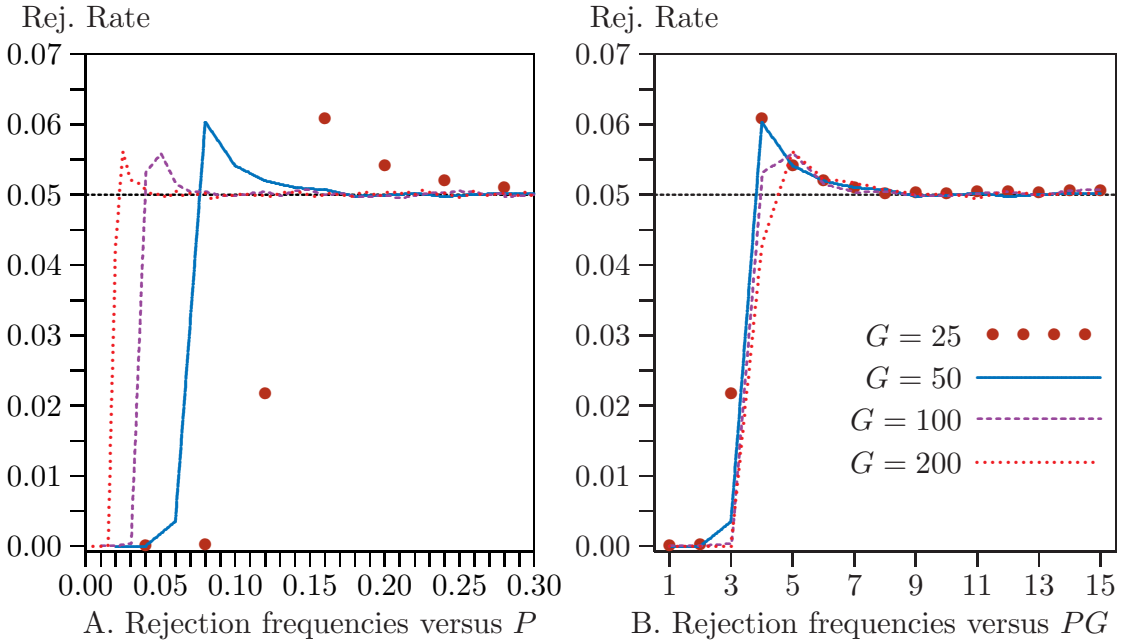


Figure 8: Power of DiD tests, state-sized clusters, $P = 0.2$

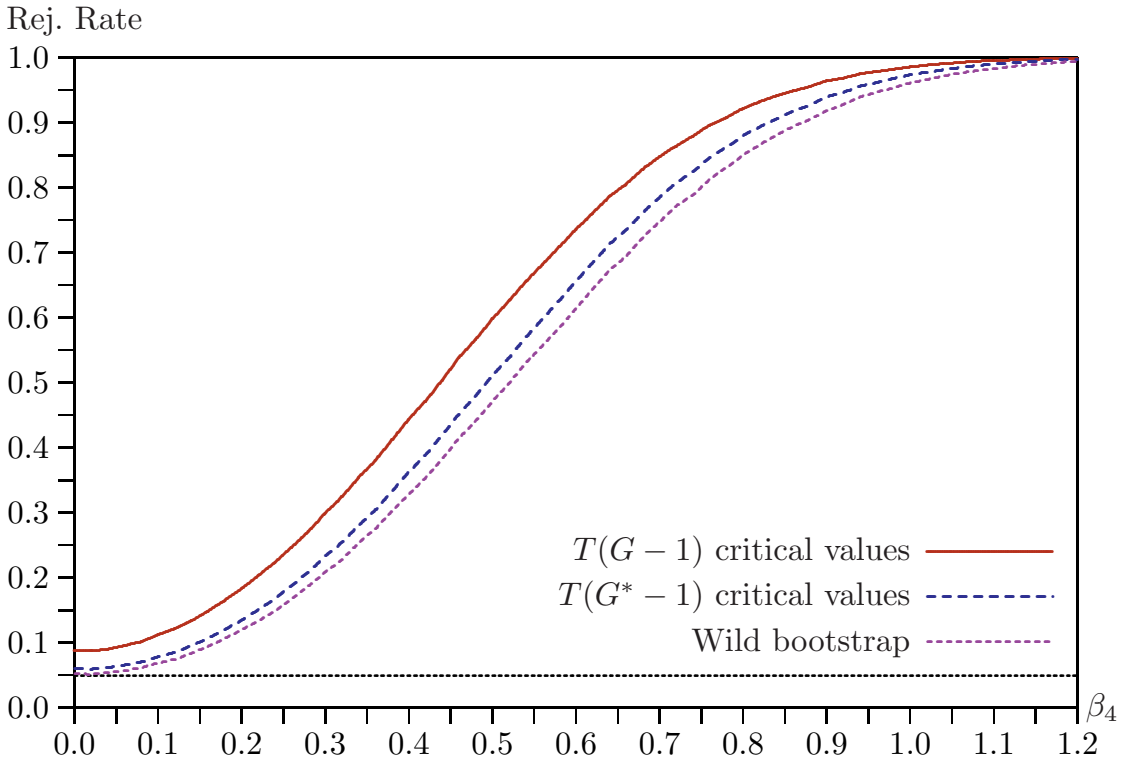

Figure 9: Rejection rates and states treated for placebo laws

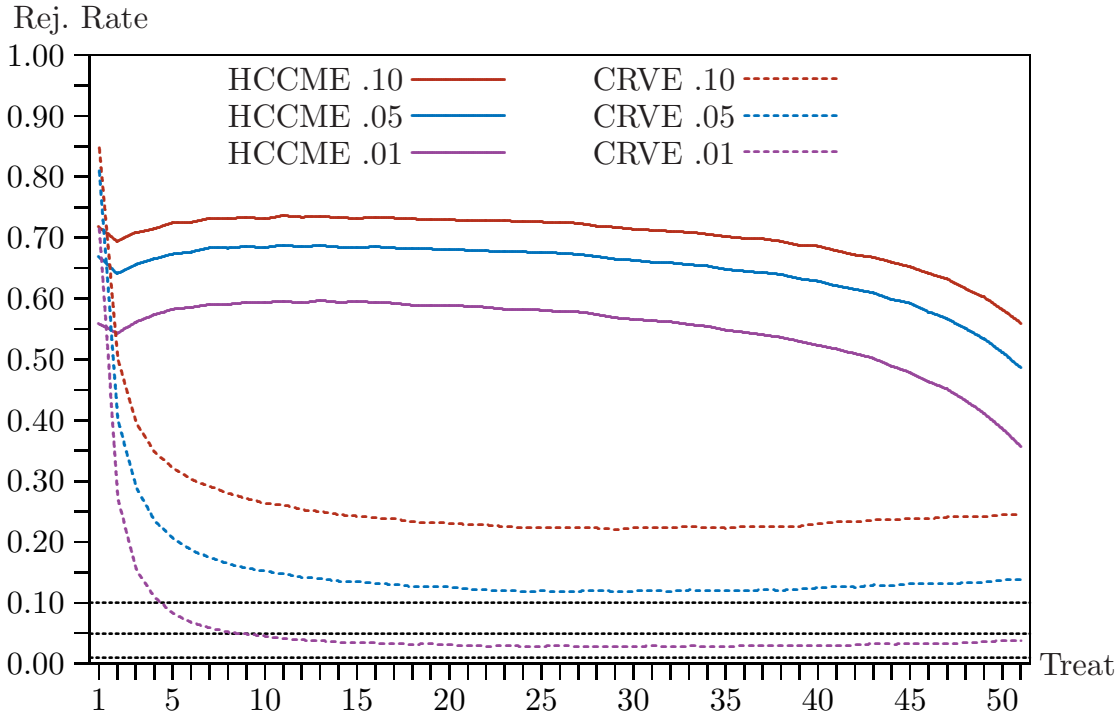


Figure 10: Rejection rates and states treated for placebo laws