Pitfalls when Estimating Treatment Effects Using Clustered Data*

James G. MacKinnon
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

September 14, 2017

Abstract

Inference for estimates of treatment effects with clustered data requires great care when treatment is assigned at the group level. This is true for both pure treatment models and difference-in-differences regressions. Even when the number of clusters is quite large, cluster-robust standard errors can be much too small if the number of treated (or control) clusters is small. Standard errors also tend to be too small when cluster sizes vary a lot, resulting in too many false positives. Bootstrap methods generally perform better than t tests, but they can also yield very misleading inferences in some cases.

Keywords: CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, difference-in-differences, DiD regression

*We are grateful to Justin Esarey for several very helpful suggestions and to Joshua Roxborough for valuable research assistance. This research was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. Some of the computations were performed at the Centre for Advanced Computing at Queen's University.

1 Introduction

There is a large and rapidly growing literature on inference with clustered data, that is, data where the disturbances (error terms) are correlated within clusters. The clusters might be associated with, for example, jurisdictions, schools, hospitals, industries, or time periods. Cameron and Miller (2015) provides a very good survey. However, the literature is growing rapidly. More recent papers include Imbens and Kolesár (2016), MacKinnon and Webb (2017b), Carter, Schnepel, and Steigerwald (2017), Pustejovsky and Tipton (2017), Djogbenou, MacKinnon, and Nielsen (2017), MacKinnon, Nielsen, and Webb (2017), and Esarey and Menger (2017). Most of these papers are written by and for economists, but the Esarey-Menger paper is specifically aimed at researchers in political science.

Since the theoretical justification for cluster-robust standard errors is asymptotic, it is evident that we need sufficiently large samples if we are to make valid inferences. It has long been recognized that what matters is not the number of observations but rather the number of clusters. Based on the limited simulation evidence available at the time it was written, Angrist and Pischke (2008, Chapter 8) suggested, not entirely seriously, that it is safe to use cluster-robust standard errors whenever there are at least 42 clusters. But the evidence on which they based this suggestion was for the best-case scenario of equal-size clusters and continuous regressors.

More recent work suggests that, by itself, the number of clusters does not tell us whether inference is likely to be reliable. MacKinnon and Webb (2017b) shows by simulation that using cluster-robust t statistics can be quite unreliable when there are either 50 or 100 clusters that are proportional to the sizes of U.S. states (with each state appearing twice in the latter case). In general, it seems to be the case that, as cluster sizes become more unequal, inference becomes less reliable.

Perhaps more surprisingly, MacKinnon and Webb (2017b) also shows that, when the regressor of interest is a treatment dummy, cluster-robust standard errors can be very much too small whenever the number of treated clusters is small. This result is obtained by theory and confirmed by simulation. It holds even when the total number of clusters is very large. Previous simulation evidence on this point may be found in Bell and McCaffrey (2002) and Conley and Taber (2011). Since many applied studies in economics and political science involve either pure treatment models or difference-in-differences (DiD) models, this finding has serious implications for applied work in both fields.

Bootstrap methods typically perform better than t tests, but they can also yield very misleading inferences in some cases. In particular, what would otherwise be the best variant of the wild bootstrap can underreject extremely severely when the number of treated clusters is very small. Other bootstrap methods can overreject extremely severely in that case.

In Section 2, we briefly review the key ideas of cluster-robust covariance matrices and standard errors. In Section 3, we then explain why inference based on these standard errors can fail when there are few treated clusters. In Section 4, we discuss bootstrap methods for cluster-robust inference. In Section 5, we report (graphically) the results of several simulation experiments which illustrate just how severely both conventional and bootstrap methods can overreject or underreject when there are few treated clusters. In Section 6, the implications of these results are illustrated using an empirical example from

Burden, Canon, Mayer, and Moynihan (2017). Finally, Section 7 concludes and provides some recommendations for empirical work.

2 Cluster-Robust Standard Errors

We are concerned with the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \mathrm{E}(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega}, \tag{1}$$

where $\boldsymbol{y}$ and $\boldsymbol{u}$ are $N \times 1$ vectors of observations and disturbances, $\boldsymbol{X}$ is an $N \times k$ matrix of covariates, $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector, and the $N \times N$ covariance matrix $\boldsymbol{\Omega}$ is

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Omega}_2 & \dots & \mathbf{O} \\ \vdots & \vdots & & \vdots \\ \mathbf{O} & \mathbf{O} & \dots & \boldsymbol{\Omega}_G \end{bmatrix}. \tag{2}$$

There are G clusters, indexed by g, with N_g observations in the g^{th} cluster. For notational convenience, the observations are assumed to be ordered by cluster, although this is not necessary in practice. The $N_g \times N_g$ matrix $\boldsymbol{\Omega}_g$ is positive definite. It is the covariance matrix for the observations belonging to the g^{th} cluster. Thus $\boldsymbol{\Omega}$ is block-diagonal, with G diagonal blocks that correspond to the G clusters.

The true covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ for the model (1) is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{\Omega}_g\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3}$$

where $\boldsymbol{X}_g$ denotes the N_g rows of $\boldsymbol{X}$ that belong to the g^{th} cluster. The most widely-used cluster-robust variance estimator, or CRVE, is

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\hat{\boldsymbol{u}}_g\hat{\boldsymbol{u}}_g'\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{4}$$

The first factor here is asymptotically negligible, but it always makes the CRVE larger when G and N are finite.[1] Covariance matrix estimators like equation (4) are often referred to as "sandwich estimators" because there are two identical pieces of "bread" on the outside and a "filling" in the middle. The filling in the sandwich on the right-hand side of equation (4) is evidently intended to estimate the corresponding factor in equation (3). In both cases, the filling involves a sum of G $k \times k$ matrices. In the case of (4), each of these matrices has rank one. Therefore, the matrix (4) can have rank at most G.

The CRVE (4) is often called CV_1, because it is the analog of the heteroskedasticity-robust covariance matrix estimator HC_1; see MacKinnon (2012). A more complicated CRVE, often called CV_2, which is the analog of the HC_2 estimator studied in MacKinnon and White (1985), was proposed in Bell and McCaffrey (2002) and has recently been advocated by Imbens and Kolesár (2016); see also Pustejovsky and Tipton (2017). CV_2 generally

[1] This particular factor is used by Stata and seems to have been introduced by them.

has better finite-sample properties than CV_1. It does not solve the problems associated with few treated clusters, but it can make them less severe. Unfortunately, CV_2 is considerably more expensive to compute than CV_1 when the clusters are large.[2] Although we do not discuss CV_2 further, it should certainly be considered for samples of moderate size.

The most common way to make inferences about any element of $\boldsymbol{\beta}$, say β_k, is to divide the OLS estimate $\hat{\beta}_k$ by the square root of the k^{th} diagonal element of the CRVE (4) and compare the resulting t statistic to the $t(G-1)$ distribution. This procedure, which can be much more conservative than using the standard normal distribution when G is small, was suggested in Bester, Conley, and Hansen (2011). There are also several (moderately complicated) procedures for calculating the degrees of freedom based on $\boldsymbol{X}$ and the assumed cluster structure; see Bell and McCaffrey (2002), Imbens and Kolesár (2016), Young (2016), and Carter, Schnepel, and Steigerwald (2017). These typically yield non-integer degrees of freedom that are even smaller than $G-1$.

3 Inference with Few Treated Clusters

The fundamental problem with CRVE-based inference when there are few treated clusters is that, in such cases, the residuals typically provide very poor estimates of the disturbances for the treated clusters. Following MacKinnon and Webb (2017b, Section 6), we consider the pure treatment model

$$y_{gi} = \beta_1 + \beta_2 d_{gi} + u_{gi}, \quad i = 1 \ldots, N_g, \quad g = 1, \ldots, G, \tag{5}$$

where d_{gi} equals 1 for the first G_1 clusters and 0 for the remaining $G_0 = G - G_1$ clusters. In this model, every observation in the g^{th} cluster is either treated ($d_{gi} = 1$) or not treated ($d_{gi} = 0$). The analysis would be more complicated if we included additional regressors, or allowed only some observations within treated clusters to be treated, but it would not change in any fundamental way.

From expression (4), it is not hard to show that, if we omit the initial scalar factor, the CRVE for $\hat{\beta}_2$, the OLS estimator of β_2 in equation (5), is equal to

$$\frac{\sum_{g=1}^{G} (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)' \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g' (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)}{\left((\boldsymbol{d} - \bar{d}\boldsymbol{\iota})' (\boldsymbol{d} - \bar{d}\boldsymbol{\iota}) \right)^2}. \tag{6}$$

Here $\boldsymbol{d}$ denotes the vector with typical element d_{gi}, $\boldsymbol{d}_g$ is the subvector corresponding to cluster g, $\bar{d}$ is the mean of the d_{gi} (that is, the proportion of the observations that are treated), and $\boldsymbol{\iota}$ and $\boldsymbol{\iota}_g$ are vectors of 1s, of lengths N and N_g, respectively. The numerator of (6) is essentially the filling in the CRVE sandwich, and the denominator is the square of the bread. Only the numerator depends on the residuals.

Expression (6) would provide a good estimate of $\text{Var}(\hat{\beta}_2)$ if its numerator provided a good estimate of the filling in the sandwich (3) specialized to the case of β_2 in the treatment model (5). In scalar notation, this numerator can be written as

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

[2] In fact, with current computer hardware and software, it seems to become difficult to compute CV_2 once any of the N_g exceeds 5000 or so; see MacKinnon and Webb (2017a).

Expression (7) is supposed to estimate the quantity

$$(1-\bar{d})^2 \sum_{g=1}^{G_1} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \omega_{ij}^g + \bar{d}^2 \sum_{g=G_1+1}^{G} \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \omega_{ij}^g, \tag{8}$$

where ω_{ij}^g is the ij^{th} element of $\boldsymbol{\Omega}_g$. But it does a terrible job of doing so when the number of treated clusters is small.

The problem is that the residuals must sum to zero over all the treated observations. Consider first the extreme case in which only the observations in cluster 1 are treated. In that case, $\sum_{i=1}^{N_1} \hat{u}_{1i} = 0$. This implies that the first term in expression (7) equals 0. The corresponding term on the right-hand side of equation (8) is certainly not 0. In fact, unless the proportion of treated observations is very large (which seems improbable when only one cluster is treated), the first term on the right-hand side of equation (8) will typically be larger than the second term, because $(1-\bar{d})^2$ will typically be much larger than $\bar{d}^2$.

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 (7) will always underestimate the corresponding triple summation in (8). As MacKinnon and Webb (2017b, Appendix A.3) shows, the underestimation should go away quite rapidly as G_1 increases. However, it is difficult to say just how large G_1 needs to be, even for the very simple model (5), because how well (7) estimates (8) depends on the sizes of the treated and untreated clusters and on the $\boldsymbol{\Omega}_g$ matrices. For more general models, it will also depend on the $\boldsymbol{X}_g$ matrices.

When the number of treated clusters is very small, it is quite possible for the CRVE (6) to underestimate the true variance of $\hat{\beta}_2$ by a factor of 25 or more, which implies that cluster-robust t statistics may be too large by a factor of five or more. This inevitably leads to confidence intervals that are much too narrow and tests that overreject very severely.

4 Bootstrap Methods

One widely-used way to obtain more reliable inferences than simply comparing a cluster-robust t statistic with the $t(G-1)$ distribution is to use a bootstrap test. Several different bootstrap tests are available, and they can produce very different results. As we will see in Section 5, bootstrap tests often yield considerably more reliable inferences than cluster-robust t tests, but they do not always work well. For a general introduction to bootstrap hypothesis testing, see Davidson and MacKinnon (2006).

When we perform a bootstrap test, we have to generate a large number of bootstrap samples. The alternative bootstrap methods that we consider differ only in how these bootstrap samples are generated. One particularly important issue concerns whether or not the bootstrap data-generating process, or DGP, imposes the null hypothesis. For concreteness, suppose we wish to test the hypothesis that β_k, the last element of $\boldsymbol{\beta}$, is zero. Then the bootstrap DGP could use either $\hat{\boldsymbol{\beta}}$, the unrestricted vector of OLS estimates, or $\tilde{\boldsymbol{\beta}}$, the vector with 0 as the last element and all other elements equal to the restricted OLS estimates.

For example, in the case of the pure treatment model (5), if the null hypothesis is that $\beta_2 = 0$, then $\tilde{\beta}_1 = \bar{y}$, the sample mean, and $\tilde{\beta}_2 = 0$. The estimate of β_1 under the null is just the sample mean in this simple case because, when $\beta_2 = 0$, equation (5) only contains

a constant term. More generally, we would need to re-estimate the model subject to the restriction that $\beta_k = 0$.

It is good to choose B, the number of bootstrap samples, so that $\alpha(B+1)$ is an integer when α is the level of the test. Thus good values of B include 999 and 9999. However, if bootstrapping is very expensive, it may be possible to get away with a much smaller value of B by using a sequential procedure; see Davidson and MacKinnon (2000).

Assuming that B is fixed, the algorithm for computing a bootstrap P value for the hypothesis that $\beta_k = 0$ works as follows:

1. Estimate model (1) by OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to obtain $\hat{\boldsymbol{\beta}}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. Use these quantities to compute the cluster-robust t statistic $t_k = \hat{\beta}_k/\text{s.e.}(\hat{\beta}_k)$, where s.e.$(\hat{\beta}_k)$ denotes the square root of the k^{th} diagonal element of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$.

2. Generate bootstrap samples $\boldsymbol{y}^{*b}$ for $b = 1, \ldots, B$, and re-estimate the model (1) to obtain $\hat{\boldsymbol{\beta}}^{*b}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{*b})$. The bootstrap samples may be generated in several different ways; see below.

3. For each bootstrap sample, compute the bootstrap t statistic, t_k^{*b}, as either

$$t_{kr}^{*b} = \frac{\hat{\beta}_k^{*b}}{\text{s.e.}(\hat{\beta}_k^{*b})} \quad \text{or} \quad t_{ku}^{*b} = \frac{\hat{\beta}_k^{*b} - \hat{\beta}_k}{\text{s.e.}(\hat{\beta}_k^{*b})}.$$

   If the bootstrap data generating process (DGP) imposes the null hypothesis, use t_{kr}^{*b} (the restricted version). If the bootstrap DGP does not impose the null hypothesis, use t_{ku}^{*b} (the unrestricted version).

4. Compute either the symmetric bootstrap P value

$$\hat{P}_{\text{S}}^{*} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\Big(|t_k^{*b}| > |t_k|\Big) \tag{9}$$

   or the equal-tail bootstrap P value

$$\hat{P}_{\text{ET}}^{*} = \frac{1}{B} \min\left(\sum_{b=1}^{B} \mathbb{I}(t_k^{*b} < t_k), \sum_{b=1}^{B} \mathbb{I}(t_k^{*b} \geq t_k) \right), \tag{10}$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function, which equals 1 when its argument is true and 0 otherwise.

For many pure treatment and DiD models, we would expect t_k to be approximately symmetric around zero under the null hypothesis, so that equations (9) and (10) should yield very similar P values. Of course, this would generally not be true for dynamic models or models estimated by instrumental variables, since in both cases $\hat{\beta}_k$ might well be biased. In such cases, it would make sense to use $\hat{P}_{\text{ET}}^{*}$ rather than $\hat{P}_{\text{S}}^{*}$.

A number of different bootstrap methods can be used to generate the bootstrap samples, and, as we show in the next section, the finite-sample properties of these methods can differ

enormously. The best-known method is probably the wild cluster bootstrap proposed in Cameron, Gelbach, and Miller (2008). For the restricted version of this procedure, we first re-estimate the model subject to the restriction(s) to be tested, so as to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$. The bootstrap DGP is then

$$\boldsymbol{y}_g^{*b} = \boldsymbol{X}_g\tilde{\boldsymbol{\beta}} + v_g^{*b}\tilde{\boldsymbol{u}}_g, \quad g = 1, \dots, G, \tag{11}$$

where v_g^{*b} is a random variate with mean 0 and variance 1. A good choice in most cases is the Rademacher distribution, which takes the values 1 and -1 with equal probability; see Davidson and Flachaire (2008). Notice that, for each bootstrap sample, there is only one realization of this random variable for each cluster. In consequence, the bootstrap disturbances, $v_g^{*b}\tilde{\boldsymbol{u}}_g$, are independent across clusters but are supposed to mimic the covariance matrices of the residuals within each cluster.[3]

Combining the bootstrap DGP (11) with the algorithm given above yields a restricted wild cluster, or WCR, P value. If we replaced $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{u}}_g$ in (11) by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}_g$, respectively, we would obtain an unrestricted wild cluster, or WCU, P value. Note that, in the latter case, we must use t_{ku}^{*b} rather than t_{kr}^{*b} for the bootstrap test statistics. Otherwise, the test would have no useful power. That is, the power of the test would be very similar to its size.

Recently, MacKinnon and Webb (2017a) suggested that it may be desirable to use the ordinary wild bootstrap instead of the wild cluster bootstrap for the model (5) when the number of treated clusters is small. The only difference between the restricted wild bootstrap (WR) and the restricted wild cluster bootstrap (WCR) is that the random variate v_g^{b*} in equation (11) is replaced by N_g random variates v_{gi}^{b*}, $i = 1, \dots, N_g$. Since this eliminates the intra-cluster correlation that the wild cluster bootstrap is trying to capture, it may seem inappropriate. However, MacKinnon and Webb (2017a) shows that it can work very well in some cases, and Djogbenou, MacKinnon, and Nielsen (2017) proves that it is asymptotically valid. Of course, there is also an unrestricted wild bootstrap procedure (WU) which differs from WR in exactly the same way as WCU differs from WCR.

One important limitation of the wild and wild cluster bootstraps is that they can only be used with regression models. The models do not have to be linear, like (1), but they must have the form

$$y_{gi} = f(\boldsymbol{X}_{gi}, \boldsymbol{\beta}) + u_{gi}, \tag{12}$$

where $f(\boldsymbol{X}_{gi}, \boldsymbol{\beta})$ is a possibly nonlinear regression function that depends on a parameter vector $\boldsymbol{\beta}$ and a row vector of regressors $\boldsymbol{X}_{gi}$.

A very different way to generate bootstrap samples is to use the pairs cluster bootstrap, which was proposed (under a different name) in Bertrand, Duflo, and Mullainathan (2004). The idea of the pairs cluster bootstrap is to resample the entire $[\boldsymbol{y} \;\; \boldsymbol{X}]$ matrix by cluster. Thus each bootstrap sample consists of G submatrices chosen at random, with replacement, from the G submatrices $[\boldsymbol{y}_g \;\; \boldsymbol{X}_g]$ and stacked to form a matrix $[\boldsymbol{y}^{*b} \;\; \boldsymbol{X}^{*b}]$.

The pairs cluster bootstrap has one major advantage. It can be used for any sort of model in which the disturbances may be clustered, not just regression models. In particular, it can

[3] When the number of clusters is small, the fact that a Rademacher random variate can take on only two values means that the number of possible bootstrap samples, which is 2^G, is rather small. For $G \leq 12$, it may be better to use another discrete distribution which can take on six values; see Webb (2014).

be used for binary response models such as the logit and probit models. However, it also has two serious disadvantages. The first is that, unless $N_g = N/G$ for all clusters, the bootstrap samples will almost never be the same size as the original sample. Some bootstrap samples will happen to contain a relatively large proportion of big clusters, and others will happen to contain a relatively large proportion of small clusters. When the N_g vary a lot, so will the sizes of the bootstrap samples. This will make it difficult for the distribution of the t_k^{*b} to mimic the distribution of t_k.

The second disadvantage of the pairs cluster bootstrap applies specifically to models with few treated clusters. We know from the results in MacKinnon and Webb (2017b) (and will also see in the next section) that the number of treated clusters, G_1, has an enormous impact on the rejection frequency of a cluster-robust t test. But G_1 necessarily varies when we use the pairs cluster bootstrap. Some bootstrap samples will contain more treated clusters than the actual sample, and some will contain fewer. Indeed, when G_1 is small, some bootstrap samples may well contain no treated clusters at all. This will happen for about 36.8% of them when $G_1 = 1$.[4] When it does happen, $\hat{\boldsymbol{\beta}}^{*b}$ cannot be computed, and the bootstrap sample has to be thrown out.

Another class of simulation-based procedures that has been proposed for making inferences about treatment effects at the cluster level is randomization inference, or RI. It was first suggested in this context by Conley and Taber (2011). Alternative RI procedures have been investigated by Canay, Romano, and Shaikh (2017), Ferman and Pinto (2015), and MacKinnon and Webb (2016). To keep this paper focused and reasonable in length, we do not consider any of these procedures in our simulations.

5 Simulation Experiments

In this section, we study the performance of cluster-robust t tests and five different bootstrap tests using simulation experiments. We study both the pure treatment model (5) and a DiD regression model. Our focus is on the number of treated clusters, G_1, holding N and G fixed. We report all our results graphically, with 5% rejection frequencies[5] on the vertical axis and G_1 on the horizontal axis.

The first set of experiments is for the pure treatment model (5). The value of G is either 12 or 24, and we set $N = 100G$. We would have obtained extremely similar results for most methods if N had been 5, 10, or even 100 times larger. The disturbances, u_{gi}, are normally distributed and generated by a random effects model with intra-cluster correlation coefficient $\rho_g = 0.05$. We do not study the role of ρ_g in detail because previous research (along with some experiments that we do not report) has shown that it typically has a very modest effect for the WCR and WCU bootstraps. For obvious reasons, it does have a somewhat larger effect for the WR and WU bootstraps, however.

[4] The probability that any given cluster will not appear in a particular bootstrap sample is $\big((G-1)/G\big)^G$. This converges to $\exp(-1) = 1/(2.71828) = 0.36788$ as $G \to \infty$.

[5] A 5% rejection frequency is the proportion of the P values over the 400,000 replications that are below 0.05. If the null is true and the procedure works perfectly, then the 5% rejection frequency should be 0.05, plus or minus a little bit of experimental error. When the true rejection frequency is 0.05, the standard error of the estimate is 0.00034. Thus we would expect to see numbers between 0.04932 and 0.05068 about 95% of the time for procedures that work perfectly.

Results in MacKinnon and Webb (2017b) show that the amount of variation in cluster sizes can be very important. In order to allow for possibly unbalanced cluster sizes, N_g is determined by a parameter γ, as follows:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{13}$$

where $[\cdot]$ denotes the integer part of its argument, and $N_G = N - \sum_{j=1}^{G-1} N_g$. Every N_g is equal to $N/G = 100$ when $\gamma = 0$. As γ increases or decreases, cluster sizes become increasingly unbalanced. For $\gamma > 0$, the N_g increase as g increases, and for $\gamma < 0$, they decrease.

All experiments have 400,000 replications. We use such a large number because rejection frequencies often differ very little among alternative methods, and it would otherwise be difficult to distinguish between systematic differences and experimental errors. All bootstrap methods use $B = 399$.

Panels a) and b) of Figure 1 show results for experiments with $G = 12$. In panel a), $\gamma = 0$, so that $N_g = 100$ for all g. In panel b), $\gamma = 2$, so that N_g increases from 34 to 217 as g increases from 1 to 12. The clusters are always treated in increasing order. Thus, if $G_1 = 2$ and $\gamma > 0$, only the smallest and second-smallest clusters are treated. This is probably not very realistic, but it illustrates the importance of differing cluster sizes. In practice, even if cluster sizes varied a lot, it would probably not often be the case that only the very smallest (or very largest) clusters were treated. Thus the results in panel b) are probably for an extreme case.

All the results in panel a) of Figure 1 are symmetric around $G_1 = 6$. This reflects the fact that, for constant cluster sizes, there is no real difference between G_1 and G_0. It is evident that the cluster-robust t statistic always overrejects, and it does so very severely for $G_1 \leq 2$ and $G_1 \geq 10$. The pairs cluster and WCU bootstraps also overreject very severely for $G_1 = 1$ and $G_1 = 11$, but they improve more rapidly as G_1 becomes less extreme, and the latter actually performs very well for $4 \leq G_1 \leq 8$. In contrast, the pairs bootstrap underrejects quite severely for those cases.

The WCR bootstrap underrejects extremely severely for $G_1 = 1$ and $G_1 = 11$ (the actual rejection rates are about 1 in 10,000) and very severely for $G_1 = 2$ and $G_1 = 10$. However, it works quite well for $3 \leq G_1 \leq 9$. The performance of the cluster-robust t statistic and the WCR and WCU bootstraps here are precisely what the theoretical results in MacKinnon and Webb (2017b, Section 6) predict.

To our knowledge, the performance of the pairs cluster bootstrap for treatment models has not been studied. However, it could have been predicted from the results of Davidson and MacKinnon (1999). That paper studies the size distortion of parametric bootstrap tests, and the value of G_1 here plays essentially the same role as the parameters in that paper. The paper shows that whether a bootstrap test overrejects or underrejects is determined by how much the distributions of the bootstrap test statistics depend on the parameter estimates and on how biased and noisy those estimates are.

In the case of Figure 1, many of the errors in rejection frequency for the pairs cluster bootstrap arise from the fact that the bootstrap samples contain different groups. Imagine a sample with four groups denoted A, B, C, and D, of which only A is treated, so that $G_1 = 1$. If we call the number of treated clusters in a bootstrap sample G_1^*, then the

bootstrap sample $\{A, B, A, C\}$ would have $G_1^* = 2$, whereas the sample $\{B, D, B, C\}$ would have $G_1^* = 0$. When $G_1^* = 0$, we obviously cannot estimate the model, so any samples that do not contain A would have to be discarded. Thus all the bootstrap samples that we can actually use have $G_1^* \geq 1$ (and also $G_0^* \leq 3$). On average, the values of G_1^* must be greater than G_1 whenever $G_1 = 1$.

The same thing happens in our experiments. When G_1 is very small or very large, many of the bootstrap samples have values of G_1^* that are less extreme than G_1 itself. This means that the actual t statistic t_k is being compared with a mixture of distributions, many of which involve less extreme values of G_1^*. In contrast, when G_1 takes on values between 4 and 8, the bootstrap t statistics with which t_k is being compared are often associated with substantially more extreme values of G_1^*. In the former case, we get severe overrejection, and in the latter case quite severe underrejection.

The WR and WU bootstraps work extraordinarily well in panel a) of Figure 1. They perform well even for extreme values of G_1, and they differ noticeably from each other only for $G_1 = 1$ and $G_1 = 11$. Note that they would have differed more for $\rho_g > 0.05$ and even less for $\rho_g < 0.05$. This is precisely what the results of MacKinnon and Webb (2017a) predict. Unfortunately, those results require that all clusters be the same size and have the same patterns of intra-cluster correlation, which is the case in panel a) of the figure.

In panel b) of Figure 1, the variation in cluster sizes causes a number of asymmetries. This is evident for all the methods, but particularly for WCR, WR, and WU. In the case of WCR, there is now underrejection for $G_1 = 3$ as well as $G_1 \leq 2$, and there is quite substantial overrejection for $G_1 = 9$ and $G_1 = 10$. The two wild bootstrap methods, WR and WU, both underreject for small values of G_1 and overreject for large values, as predicted in MacKinnon and Webb (2017a). However, they still perform better than WCR and WCU for the two largest and three smallest values of G_1.

Panels c) and d) of Figure 1 are similar to panels a) and b), respectively, except that $G = 24$ and G_1 runs from 1 to 23. Many of the results do not change much when we double the number of clusters. However, it is noteworthy that rejection rates for the t statistic, the WCU bootstrap, and the pairs cluster bootstrap are actually worse for $G_1 = 1$ and $G_0 = 1$ than they were before. On the other hand, all methods now work better for intermediate values of G_1, with WCR and WCU performing very well for $8 \leq G_1 \leq 16$.

The pairs cluster bootstrap underrejects for all but the three most extreme values of G_1 and G_0, but it does so much less severely than in the top two panels. The latter could have been predicted from the analysis above. The range of values of G_1 for which the rejection rates of the t statistic do not change very much is now much wider. This means that these rates do not vary as much across the G_1^* for the various bootstrap samples as they did when $G = 12$, so that the differences between the actual G_1 for t_k and the various G_1^* for the corresponding t_k^{b*} do not matter as much.

Overall, when we compare the results in the bottom two panels with those in the top two, it appears that increasing the number of clusters improves the performance of all methods, but only if G_1/G is not too large or small. This is precisely what would be expected from the asymptotic theory in Djogbenou, MacKinnon, and Nielsen (2017), where regularity conditions effectively require the number of treated clusters to rise with (but not necessarily as fast as) the sample size.

The second set of experiments is for a DiD regression model. The model that we estimate can be written as

$$y_{gi} = \sum_{t=1}^{T} \delta_t D_{gi}^t + \beta_2 d_{gi} + u_{gi}, \quad i = 1 \ldots, N_g, \quad g = 1, \ldots, G. \tag{14}$$

The constant term in (5) has been replaced by T dummy variables D_{gi}^t, each of which takes the value 1 for observations associated with year t and 0 otherwise. The treatment variable d_{gi} now takes the value 1 only for treated clusters during the years in which they are treated.

In practice, investigators would often add $G - 1$ dummy variables for all but one of the clusters to regression (14). We did not do so because it would have made the experiments more expensive, and the cluster dummies would have simply offset the cluster-specific shocks in the random effects specification of the u_{gi}.[6] We choose which clusters are to be treated and then allow our program to decide at random when treatment is to commence. We set $T = 20$ and allow treatment to start as early as year 6 and as late as year 16.

The big difference between the treatment model (5) and the DiD model (14) is that, for the latter, there is no symmetry between G_1 and G_0. In fact, it is entirely possible to have $G_1 = G$, so that $G_0 = 0$. In that case, identification comes from the fact that all of the treated clusters contain some untreated observations. We therefore expect results for small values of G_1 to resemble the ones in Figure 1 but results for large values to be very different, and that is indeed what we see in Figure 2.

The top two panels of Figure 2 are comparable to the same two panels of Figure 1. Notice that G_1 runs from 1 to 12 instead of from 1 to 11. In both cases, results for small values of G_1 are quite similar across the two figures. There is severe overrejection for the t statistic, the WCU bootstrap, and the pairs bootstrap, together with severe underrejection for the WCR bootstrap. However, the results for other values of G_1 are not very similar across the two figures. In panel a), all the wild bootstrap methods work fairly well for intermediate and large values of G_1, and even the pairs cluster bootstrap performs much better than it did before. Perhaps surprisingly, the results for $G_1 = 12$ are noticeably better than for $G_1 = 11$. This probably happens because $G_0 = 1$ when $G_1 = 11$, and having just one non-treated cluster tends to cause problems. In contrast, when $G_1 = 12$, there are no longer any non-treated clusters, just 12 clusters with various proportions of treated observations.

Panel c) of Figure 2 deals with the case of $G = 12$ and $\gamma = -2$. The variation in cluster sizes here is very similar to the variation in panel b), but now the largest clusters are treated first. There are some very substantial differences between panels b) and c). In the latter, WR and WU always overreject, especially for small values of G_1, and WCR overrejects for $G_1 \geq 2$. Perhaps surprisingly, the pairs bootstrap is actually the best method for $G_1 \geq 5$.

Finally, panel d) of Figure 2 deals with the case of $G = 24$ and $\gamma = 0$. All the bootstrap methods work quite well for $G_1 \geq 4$, and WCR works extremely well for $G_1 \geq 7$. It is difficult to see in the figure, but WCU and the two ordinary wild bootstrap methods

[6] This does not mean that adding group-specific fixed effects solves the problem of intracluster correlation; it would merely do so for our specific DGP. In the "placebo law" experiments of Bertrand, Duflo, and Mullainathan (2004) and MacKinnon and Webb (2017b), which use real data from the Current Population Survey, standard errors that are robust to heteroskedasticity but not to intracluster correlation yield extremely severe errors of inference despite the presence of group-specific fixed effects.

overreject slightly for the largest values of G_1. The latter work quite well even for small values of G_1, but this undoubtedly reflects the rather simple experimental design.

We remark that all the results for DiD models would have been different if the "years" in which treatment begins had been different. They might have been quite different if we had conditioned on a particular set of years for treatment rather than choosing them at random. With our procedure, results tend to average out across replications.

6 Empirical Example

In this section, we consider an empirical example from Burden, Canon, Mayer, and Moynihan (2017). It uses county-level data to analyze the effects of certain state-level voting laws, particularly early voting, on Democratic vote share. In a portion of the analysis, the authors attempt to estimate the following non-panel DiD model to analyze the effect of early voting laws:

$$\text{demdiff}_{cs} = \beta_0 + \beta_1 \text{EV}_{cs} + \beta_2 \texttt{felon}_{cs} + \beta_3 \texttt{id}_{cs} + \boldsymbol{X}_{cs}\boldsymbol{\beta}_4 + \epsilon_{cs}. \tag{15}$$

Here demdiff_{cs} represents the difference in Democratic vote share between the 2008 and 2012 elections, for county c in state s. β_1 is the coefficient of interest. EV takes on the values -1, 0, or 1, if a state either repealed, did not change, or adopted early voting laws, respectively. Between 2008 and 2012, California and Maryland adopted, while New Jersey repealed, their early voting laws. This specification assumes a symmetric treatment effect for repealing or adopting; we will later re-estimate this model using asymmetric treatment effects. Additionally, `felon` and `id` are binary variables representing changes in felon disenfranchisement and ID requirement laws, and $\boldsymbol{X}$ represents various county-level demographic covariates. All estimates are weighted by county population. The U.S. has a total of 3,144 counties and county equivalents. The dataset has $N = 3,112$ usable observations collected from every state including D.C. but excluding Alaska, so that $G = 50$. Cluster sizes range from 1 county in D.C. to 254 counties in Texas.

We attempt to reproduce Column 4 from Table 7 of Burden et al. (2017), which contains their difference-in-differences analysis comparing the 2008 and 2012 elections. While attempting to reproduce the results, two problems were found in the replication files provided. Firstly, three control counties were mistakenly coded as treated: one in Alaska, one in Colorado, and one in DC. Additionally, California, which adopted early voting, was mistakenly coded as a control. We estimate the model using the data as coded in the 'miscoded' column, and then re-estimate it using the proper coding in the 'corrected' column.

Secondly, standard errors were clustered by county Federal Information Processing Standards (fips). County fips codes are unique within but not across states. The prevailing commonality amongst these counties is their position alphabetically within state. Clustering by fips assumes that there is potential correlation across the alphabetically first, second, third, etc. clusters across states, but no correlation within states. It appears that the authors intended to cluster at the individual county level; however, since observations are at the county level, each cluster would only contain one observation in that case. Clustering by county would then result in heteroskedasticity-robust rather than cluster-robust standard errors. Since earlier results in Burden et al. (2017) were clustered by state, and laws change at the state level, it seems sensible to cluster by state here as well.

Table 1: Effects of Early Voting Laws on Democratic Vote Shares

	Symmetric Effect		Asymmetric Effect	
	Miscoded	Corrected	EV Adopted	EV Repealed
$\hat{\beta}_1$	−1.458	−0.254	0.700	2.795
Robust	0.000	0.401	0.041	0.000
CRVE (Ctyfips)	0.000	0.325	0.055	0.000
CRVE (State)	0.088	0.722	0.144	0.000
WCU	0.500	0.799	0.208	0.000
WCR	0.740	0.880	0.300	0.423
Treated States	-	3	2	1

Empirical example from Burden, Canon, Mayer, and Moynihan (2017).
Treated states in 'Miscoded' column omitted due to coding issues in treatment assignment.
All entries except the estimates of β_1 are P values.
Estimates are weighted by county population.
WCU and WCR P values are clustered at the state level and based on 99,999 bootstraps.

Table 1 demonstrates the limitations of inference when few groups are treated. The first two columns seek to estimate the model used in Burden et al. (2017). The other two columns show this model without the symmetry assumption, where adopting states (CA, MD) and repealing states (NJ) are considered separately. The top row of Table 1 reports the estimate of β_1 or its asymmetric equivalent.[7]

The second panel reports three asymptotic P values based on robust, CRVE-fips, and CRVE-state standard errors. The bottom panel reports two bootstrap P values (WCR and WCU), both of which are calculated using $B = 99,999$ and are clustered by state.

In column 1, we replicate the analysis of Burden et al. (2017). The Robust and CRVE-fips P values are highly significant. However, clustering by state greatly reduces the evidence against the null, and the bootstrap P values are highly insignificant. In column 2, where treatment status has been corrected, all of the P values are insignificant.

When we estimate the asymmetric effects, we can really see the problems of few treated clusters, especially in evaluating the effects of a repeal of early voting where $G_1 = 1$. Here we observe that all of the asymptotic and WCU P values are highly significant, whereas the

[7]For the asymmetric equivalent, rather than β_1 being the coefficient when EV_{cs} is equal to -1 or $+1$, we in effect estimate two coefficients. The first, β_1^a, is for states that adopted early voting, so that $\text{EV}_{cs} = -1$, and the second, β_1^r, is for states that repealed early voting, so that $\text{EV}_{cs} = 1$. For the asymmetric estimates, we have two binary variables, one for repeal, which is equal to 1 for New Jersey and equal to 0 for all other states, and one for adopt, which is equal to 1 for California and Maryland and equal to 0 for all other states. This explains why the coefficient is negative in the "corrected" column, even though both coefficients are positive in the asymmetric estimates. New Jersey saw its Democratic vote share increase from 2008 to 2012, despite eliminating early voting. The coding of $\text{EV}_{cs} = -1$ essentially forces the coefficient to be negative to pick up the large increase that occurred in New Jersey, despite the fact that Maryland and California also had (difference-in-differences) increases after adopting early voting.

WCR P value is highly insignificant. This is exactly what the theory predicts and what the simulation results in Figure 2 show.

While the example above highlights the difficulties of statistical inference when there are few treated groups, we do not regard these findings as challenging the conclusions in Burden et al. (2017). In fact, when we evaluate the asymmetric estimates, we see that the states which adopted early voting had a statistically insignificant increase in the Democratic vote share, while the state that repealed early voting had a statistically ambiguous *increase* in that share. Taken together, these results present evidence against the conventional wisdom that early voting laws favor Democrats.

7 Conclusions

When a regression model is used to estimate treatment effects, cluster-robust standard errors can be extremely misleading when cluster sizes vary a lot and/or when the number of treated clusters (G_1) is small. This is true for both pure treatment models and difference-in-differences (DiD) models. One simple way to see whether there is a problem is to calculate bootstrap P values using two variants of the wild cluster bootstrap. If the WCR (restricted) and WCU (unrestricted) bootstraps yield very different inferences, then there is definitely a problem. Normally, in such cases, WCU will reject and WCR will fail to reject. Other methods may or may not yield reliable results.

Unfortunately, even when WCR and WCU roughly agree, there may be a problem; consider the results for $G_1 = 10$ in panel b) of Figure 1. However, based on the simulation results here and in MacKinnon and Webb (2017b,a), agreement between WCR and WCU seems to rule out really severe errors of inference. These are often associated with very small values of G_1, where WCR and WCU tend to disagree sharply.

It is also worth trying other bootstrap methods. When cluster sizes are similar, the ordinary wild bootstrap can work surprisingly well, even when G_1 is very small, but the pairs cluster bootstrap typically overrejects in that case. The latter is rarely the procedure of choice for regression models, but it can be useful for nonlinear models such as logit and probit. However, it tends to overreject severely when G_1 is small, and it can underreject severely when the number of clusters, G, is small and G_1/G is not close to 0 or 1.

There is an important difference between pure treatment models and DiD models. For the former, small numbers of untreated clusters are just as bad as small numbers of treated ones. For the latter, reasonably reliable results can be obtained even when all clusters are treated, provided treatment starts at different times for different clusters.

References

Angrist, J. D. and J.-S. Pischke (2008, December). *Mostly Harmless Econometrics: An Empiricist's Companion* (First ed.). Princeton: Princeton University Press.

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

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*, 249–275.

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

Burden, B. C., D. T. Canon, K. R. Mayer, and D. P. Moynihan (2017). The complicated partisan effects of state election laws. *Political Research Quarterly 70*, 564–576.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008, 05). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–427.

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

Canay, I. A., J. P. Romano, and A. M. Shaikh (2017). Randomization tests under an approximate symmetry assumption. *Econometrica 85*, 1013–1030.

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2017). Asymptotic behavior of a t test robust to cluster heterogeneity. *Review of Economics and Statistics 99*, to appear.

Conley, T. G. and C. R. Taber (2011, February). Inference with "Difference in Differences" with a small number of policy changes. *Review of Economics and Statistics 93*, 113–125.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*, 162–169.

Davidson, R. and J. G. MacKinnon (1999). The size distortion of bootstrap tests. *Econometric Theory 15*, 361–376.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews 19*, 55–68.

Davidson, R. and J. G. MacKinnon (2006). Bootstrap methods in econometrics. In T. C. Mills and K. D. Patterson (Eds.), *Palgrave Handbooks of Econometrics: Volume 1 Econometric Theory*, pp. 812–838. Palgrave.

Djogbenou, A., J. G. MacKinnon, and M. O. Nielsen (2017). Bootstrap inference with clustered errors. QED working paper, Queen's University.

Esarey, J. and A. Menger (2017). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods 5*, to appear.

Ferman, B. and C. Pinto (2015). Inference in differences-in-differences with few treated groups and heteroskedasticity. Technical report, Sao Paulo School of Economics.

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

MacKinnon, J. G. (2012). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer.

MacKinnon, J. G., M. O. Nielsen, and M. D. Webb (2017). Bootstrap and asymptotic inference with multiway clustering. QED Working Paper 1386, Queen's University.

MacKinnon, J. G. and M. D. Webb (2016). Randomization inference for difference-in-differences with few treated clusters. QED Working Paper 1355, Queen's University.

MacKinnon, J. G. and M. D. Webb (2017a). The subcluster wild bootstrap for few (treated) clusters. QED Working Paper 1364, Queen's University.

MacKinnon, J. G. and M. D. Webb (2017b). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*, 233–254.

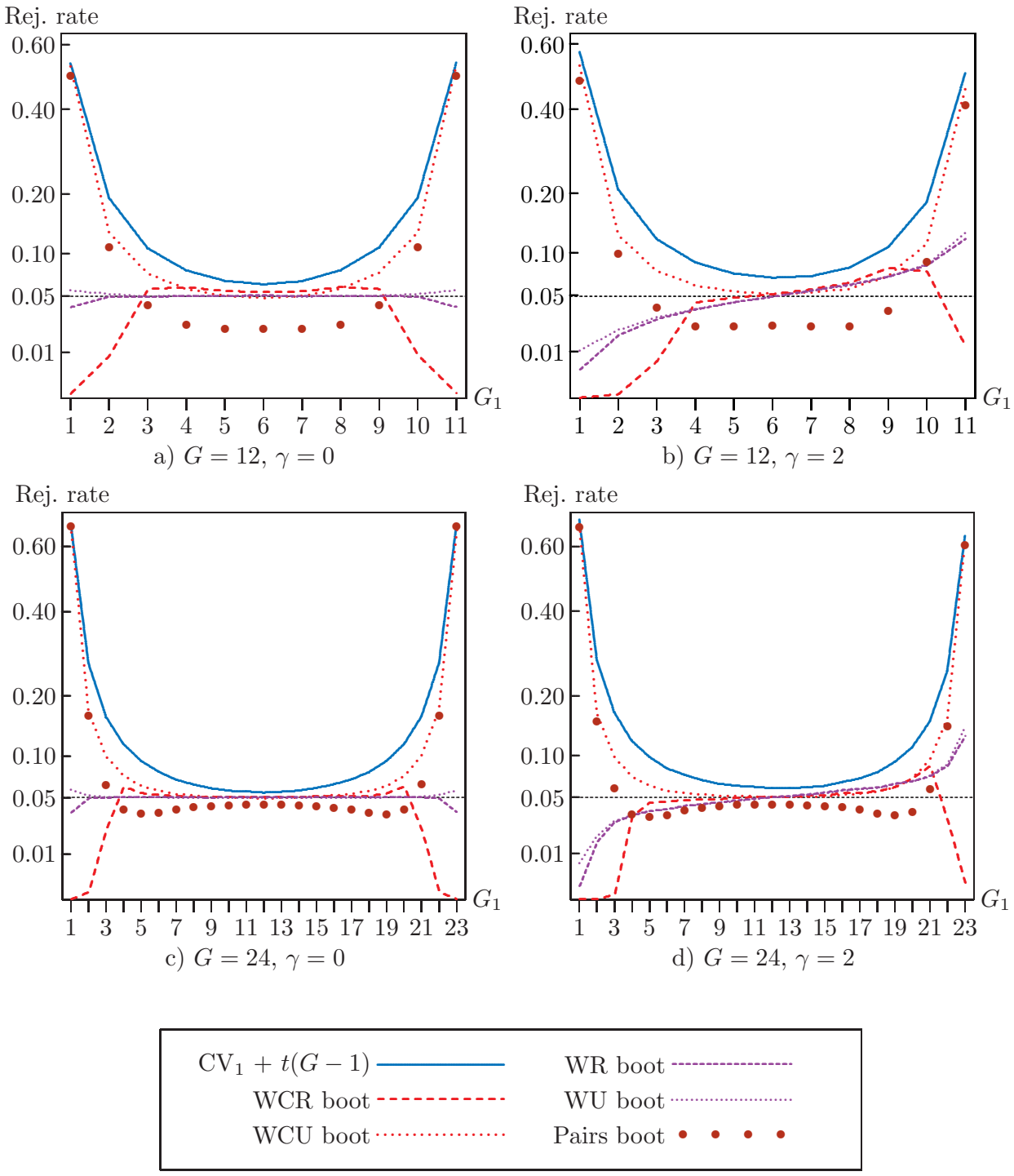
MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–325.

Pustejovsky, J. E. and E. Tipton (2017). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business and Economic Statistics 35*, to appear.

Webb, M. D. (2014, August). Reworking wild bootstrap based inference for clustered errors. QED Working Paper 1315, Queen's University.

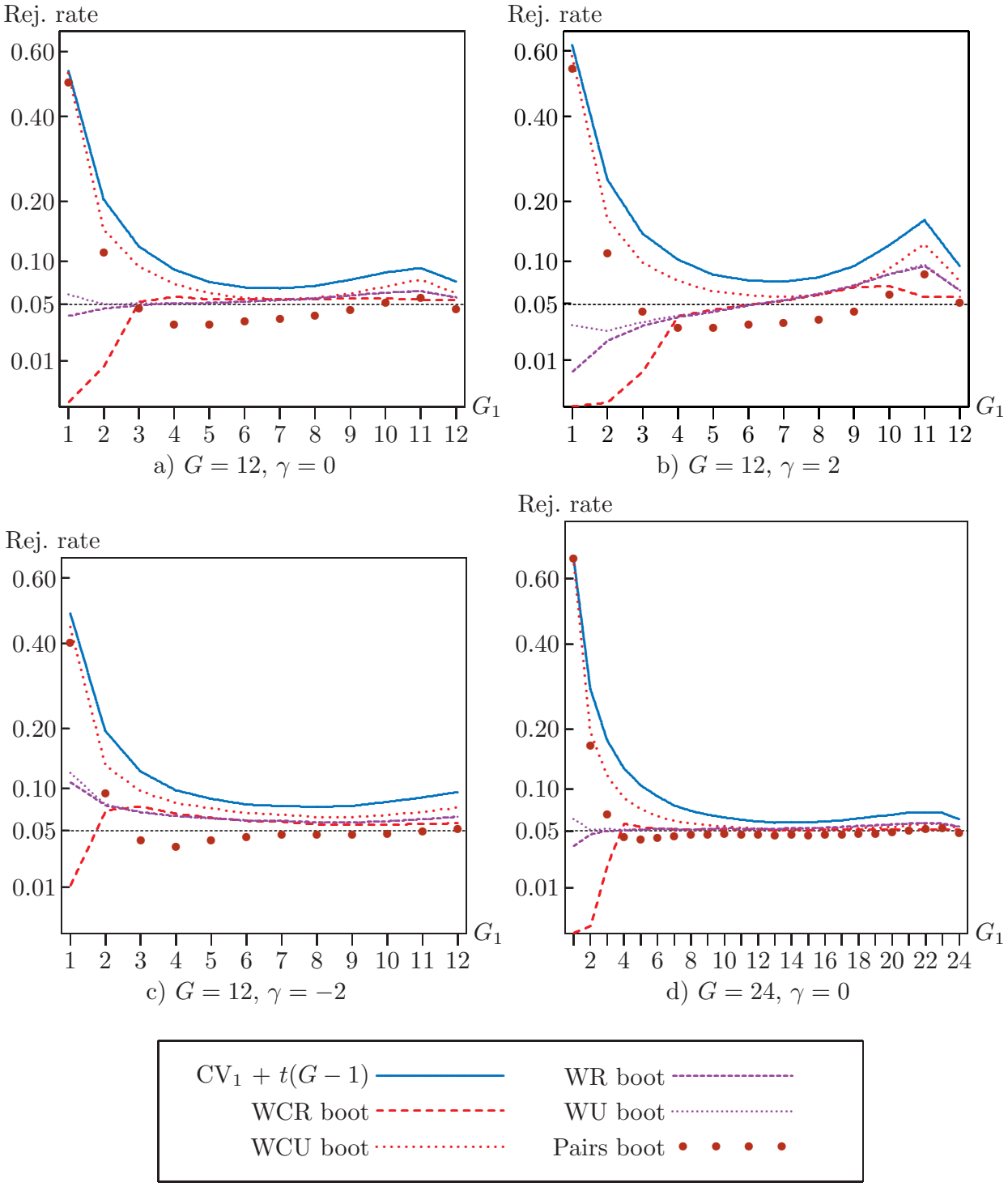
Young, A. (2016). Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections. Technical report, London School of Economics.

Figure 1: Rejection Frequencies for Pure Treatment Model



In all panels, the number of treated clusters (G_1) is on the horizontal axis.

Figure 2: Rejection Frequencies for Difference-in-Differences Model



In all panels, the number of treated clusters (G_1) is on the horizontal axis.