Queen's Economics Department Working Paper No. 1364

The Wild Bootstrap for Few (Treated) Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

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

11-2017

The Wild Bootstrap for Few (Treated) Clusters*

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

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

This paper will appear in the *Econometrics Journal* in 2018.

Abstract

Inference based on cluster-robust standard errors in linear regression models, using either the Student's t distribution or the wild cluster bootstrap, is known to fail when the number of treated clusters is very small. We propose a family of new procedures called the subcluster wild bootstrap, which includes the ordinary wild bootstrap as a limiting case. In the case of pure treatment models, where all observations within clusters are either treated or not, the latter procedure can work remarkably well. The key requirement is that all cluster sizes, regardless of treatment, should be similar. Unfortunately, the analogue of this requirement is not likely to hold for difference-in-differences regressions. Our theoretical results are supported by extensive simulations and an empirical example.

Keywords: CRVE, grouped data, clustered data, wild bootstrap, wild cluster bootstrap, subclustering, treatment model, difference-in-differences, robust inference

*An earlier version of this paper was entitled "The subcluster wild bootstrap for few (treated) clusters." The authors are grateful to participants at the 2016 University of Calgary Empirical Microeconomics Workshop, McMaster University, the 2016 Canadian Econometric Study Group, the 2016 Atlantic Canada Economics Association Meeting, the 2016 Southern Economics Association Conference, New York Camp Econometrics XII, Society for Labor Economics 2017 meeting, International Association for Applied Econometrics 2017 conference, and 2017 European Meeting of the Econometric Society for helpful comments, as well as to Phanindra Goyari, Andreas Hagemann, Doug Steigerwald, Brennan Thompson, an editor, and a referee. This research was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. Much of the computation was performed at the Centre for Advanced Computing of Queen's University.

1 Introduction

It is common in many areas of economics to assume that the disturbances (error terms) of regression models are correlated within clusters but uncorrelated between them. Inference is then based on a cluster-robust variance estimator, or CRVE. However, t tests based on cluster-robust standard errors tend to overreject severely when the number of clusters is small. How many clusters are required to avoid serious overrejection depends on several things, including how the observations are distributed among clusters and, for the important special case of binary regressors that do not vary within clusters, how many clusters are "treated"; see MacKinnon and Webb (2017b).

The wild cluster bootstrap (WCB) of Cameron, Gelbach and Miller (2008) often leads to much more reliable inferences, but, as MacKinnon and Webb (2017b) shows, this procedure can also fail dramatically. When the regressor of interest is a dummy variable that is nonzero for only a few clusters, tests based on the restricted WCB can underreject severely, and tests based on the unrestricted WCB can overreject severely.

In this paper, we investigate a family of procedures that we call the subcluster wild bootstrap. The key idea is to employ a wild bootstrap data generating process (DGP) which clusters at a finer level than the covariance matrix.[1] In many cases, this will simply be the ordinary wild bootstrap DGP of Wu (1986) and Liu (1988), which does not cluster at all. However, it could also be, for example, a DGP that clusters by state-year pair when the covariance matrix clusters by state. Thus the subcluster wild bootstrap DGP deliberately fails to match a key feature of the (unknown) true DGP. This is done in order to reduce the dependence of the bootstrap DGP on the actual sample.

In Section 2, we study a simple theoretical model for which all the observations in each cluster are either treated or not, and we explain why t tests and wild cluster bootstrap tests fail when the number of treated clusters is small. In Section 3, we then analyse the performance of the ordinary wild bootstrap for this pure treatment model. We show that, even when the number of clusters is very small, the procedure can be expected to work well if certain conditions are satisfied. The key condition is that all clusters should be (approximately) the same size. We then explain why such a condition will rarely be satisfied for difference-in-differences (DiD) regressions. Finally, we extend the analysis to the case of genuine subclusters.

In Section 4, we report the results of a large number of simulation experiments. We show that the ordinary wild bootstrap, combined with CRVE standard errors, often works very well in cases where the wild cluster bootstrap performs very badly either because the number of clusters is small or the number of treated clusters is very small, occasionally made worse by heteroskedasticity. Bootstrap tests based on the ordinary wild bootstrap often yield surprisingly reliable inferences even when there are just two treated clusters, and sometimes when there is just one.

Combining the wild bootstrap with a popular CRVE is not the only way to obtain improved finite-sample inferences in linear regression models with clustered disturbances. In Section 5 and the appendix, we discuss several alternative methods that involve using

[1] We assume that the covariance matrix is clustered at the coarsest possible level, in terms of nested clusters, which is usually the appropriate thing to do; see Cameron and Miller (2015).

a different CRVE and/or t tests with a calculated, and usually non-integer, number of degrees of freedom.

A completely different approach for the case of few treated clusters was suggested in Conley and Taber (2011). It is based on randomization inference. MacKinnon and Webb (2018a) studied that procedure and proposed an improved one which uses t statistics rather than coefficient estimates and sometimes works well. However, randomization inference with few treated clusters fails when cluster sizes vary or there is heteroskedasticity of unknown form across clusters, and it cannot be used when the number of clusters is very small.[2] We therefore do not consider randomization inference in this paper.

In Section 6, we discuss an empirical example for which the ordinary wild bootstrap yields sensible results even though there are just eight clusters. We also present results for several alternative procedures.

In the appendix, we provide a more complete treatment than the one in Section 5 of some alternative procedures for cluster-robust inference that are not based on the bootstrap. We also report the results of some additional simulation experiments which investigate the performance of those procedures when there are few treated clusters.

Section 7 concludes and provides recommendations for applied work.

2 A Pure Treatment Model

In general, we are concerned with linear regression models in which there are N observations divided among G clusters, with N_g observations in the g^{th} cluster. However, we focus on the special case of a pure treatment model, for which in the first G_1 clusters all observations are treated and in the remaining $G_0 = G - G_1$ clusters no observations are treated. This model can be written as

$$y_{ig} = \beta_1 + \beta_2 d_{ig} + \epsilon_{ig}, \tag{1}$$

where y_{ig} denotes the i^{th} observation on the dependent variable within cluster g, and d_{ig} equals 1 for the first G_1 clusters and 0 for the remaining $G_0 = G - G_1$ clusters. As usual in the literature on cluster-robust inference, we assume that

$$\mathrm{E}(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_g') = \boldsymbol{\Omega}_g \quad \text{and} \quad \mathrm{E}(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_h') = \mathbf{0} \ \text{ for } g \neq h, \tag{2}$$

where the $\boldsymbol{\epsilon}_g$ are vectors with typical elements ϵ_{ig}, and the $\boldsymbol{\Omega}_g$ are $N_g \times N_g$ positive definite covariance matrices. The model (1) is estimated by OLS, and standard errors are based on the cluster-robust variance estimator, or CRVE,

$$\frac{G(N-1)}{(G-1)(N-k)} (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{3}$$

[2] Ferman and Pinto (2015) proposes a procedure to handle aggregate data with heteroskedasticity, and MacKinnon and Webb (2018b) suggests a method that combines randomization inference and the bootstrap which can be used when the number of clusters is small. Ibragimov and Müller (2016) proposes an alternative procedure which requires at least two treated clusters.

In this case, where $k = 2$, $\boldsymbol{X}_g$ has typical row $[1 \;\; d_{ig}]$, $\hat{\boldsymbol{\epsilon}}_g$ is the N_g-vector of OLS residuals for cluster g, and $\boldsymbol{X}$ is the $N \times 2$ matrix formed by stacking the $\boldsymbol{X}_g$ matrices vertically.

Expression (3) is often called CV_1. There is more than one way to make inferences based on it. The most popular way is to compare a t statistic based on the square root of the appropriate diagonal element with the $t(G-1)$ distribution; see Bester, Conley and Hansen (2011). There are also other covariance matrix estimators, and any of the estimators can be combined with more sophisticated procedures to determine the degrees of freedom; see Section 5 and the appendix.

2.1 Why CRVE Inference Can Fail

It is shown in MacKinnon and Webb (2017b, Section 6), that the cluster-robust t statistic for $\beta_2 = 0$ in equation (1) can be written under the null hypothesis as

$$t_2 = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{\Big(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\Big)^{1/2}}, \tag{4}$$

where the N-vectors $\boldsymbol{d}$, $\boldsymbol{\iota}$, and $\boldsymbol{\epsilon}$ have typical elements d_{ig}, 1, and ϵ_{ig}, respectively, $\boldsymbol{\iota}_g$ is an N_g-vector of 1s, $\boldsymbol{d}_g$ is the subvector of $\boldsymbol{d}$ corresponding to cluster g, and $\bar{d}$ is the fraction of treated observations. The scalar c is the square root of $\Big((G-1)(N-2)\Big)/\Big(G(N-1)\Big)$, the inverse of the degrees-of-freedom correction in expression (3). In what follows, we omit the factor c, since it does not affect any of the arguments.

With c omitted, the numerator of the t statistic (4) can be written as

$$(1-\bar{d})\sum_{g=1}^{G_1}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g - \bar{d}\sum_{g=G_1+1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g. \tag{5}$$

The first term is the contribution of the treated clusters, and the second term is the contribution of the untreated ones. Similarly, the summation inside the square root in the denominator can be written as

$$(1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2. \tag{6}$$

The first and second terms here are evidently supposed to estimate the variances of the corresponding terms in expression (5). However, MacKinnon and Webb (2017b) showed that expression (6) is a very poor estimator when either G_1 or G_0 is small.

To see why this is the case, suppose that $G_1 = 1$. Then expression (6) reduces to

$$(1-\bar{d})^2(\boldsymbol{\iota}_1'\hat{\boldsymbol{\epsilon}}_1)^2 + \bar{d}^2\sum_{g=2}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 = \bar{d}^2\sum_{g=2}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2, \tag{7}$$

where the first term is zero because the residual subvector $\hat{\boldsymbol{\epsilon}}_1$ must be orthogonal to the treatment dummy $\boldsymbol{d}$. It is obvious from equation (7) that expression (6) provides a

dreadful estimator of the variance of

$$(1-\bar{d})\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1 - \bar{d}\sum_{g=2}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g, \tag{8}$$

which is what expression (5) reduces to when $G_1 = 1$. Unless cluster 1 contains a substantial fraction of the population, $\bar{d}$ will be much less than one half, and $(1-\bar{d})^2$ will therefore be very much larger than $\bar{d}^2$. Thus, unless the disturbances for the first cluster (the elements of $\boldsymbol{\epsilon}_1$) are much less variable than those for the other clusters, most of the variance of expression (8) will come from the first term. However, from equation (7) it is evident that the variance of that term is incorrectly estimated to be zero.

Note that, for the pure treatment model (1), small values of G_0 have the same consequences as small values of G_1. In contrast, for DiD models, only small values of G_1 cause problems. It is not difficult to make inferences from such models even when $G_0 = 0$, provided treatment starts at different times for different clusters.

This argument explains why tests based on the cluster-robust t statistic (4) using conventional critical values almost always overreject very severely when $G_1 = 1$ or $G_0 = 1$. The denominator of (4) grossly underestimates the variance of the numerator. As MacKinnon and Webb (2017b) shows, this underestimation, and the resulting overrejection, become much less severe as G_1 increases. Just how rapidly this happens depends on the sizes of the treated and untreated clusters and on the covariance matrices $\boldsymbol{\Omega}_g$ of the disturbances within each cluster.

2.2 The Wild Cluster Bootstrap and Why It Can Fail

Suppose there are B bootstrap samples indexed by b. In the case of regression (1), the restricted wild cluster bootstrap DGP for bootstrap sample b is

$$y_{ig}^{*b} = \tilde{\beta}_1 + \tilde{\epsilon}_{ig}v_g^{*b}, \tag{9}$$

where $\tilde{\beta}_1$ is the restricted OLS estimate of β_1, which in this case is just the sample mean of the dependent variable, $\tilde{\epsilon}_{ig}$ is the restricted residual for observation i in cluster g, and v_g^{*b} is a random variable that typically follows the Rademacher distribution and takes the values 1 and -1 with equal probability. Other auxiliary distributions can also be used, but the Rademacher distribution seems to work best in most cases; see Davidson and Flachaire (2008) and MacKinnon (2015). However, when $G \leq 11$, it is better to use a distribution with more than two mass points; see Webb (2014).

To perform a bootstrap test, each of the B bootstrap samples generated by the bootstrap DGP (9) is used to compute a bootstrap test statistic t_2^{*b}; see below. The symmetric bootstrap P value is then calculated as

$$\hat{p}^* = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\Big(|t_2^{*b}| > |t_2|\Big), \tag{10}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. It would of course be valid to use an equal-tail P value instead of (10), and the latter would surely be preferable if the distribution of

the t_2^{*b} were not symmetric around the origin.

In most cases, the wild cluster bootstrap works well. Even when G is quite small (say, between 15 and 20), simulation results in MacKinnon and Webb (2017b) and MacKinnon (2015) suggest that rejection frequencies tend to be very close to nominal levels, provided that cluster sizes do not vary extremely and the number of treated clusters is not too small. However, the restricted wild cluster bootstrap tends to underreject very severely when G_1 is small. When $G_1 = 1$, it typically never rejects at any conventional level. In order to motivate the wild bootstrap procedures that we introduce in the next section, we now explain why this happens.

The bootstrap t statistic analogous to t_2 is

$$t_2^{*b} = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}_b^*}{\left(\sum_{g=1}^G (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g^{*b}\hat{\boldsymbol{\epsilon}}_g^{*b\prime}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\right)^{1/2}}, \tag{11}$$

where $\boldsymbol{\epsilon}_b^*$ is an N-vector formed by stacking the vectors of bootstrap disturbances $\boldsymbol{\epsilon}_g^{*b}$ with typical elements $\tilde{\epsilon}_{ig}v_g^{*b}$, and $\hat{\boldsymbol{\epsilon}}_g^{*b}$ is the vector of OLS residuals for cluster g and bootstrap sample b; compare equation (4).

Now consider the extreme case in which $G_1 = 1$. The numerator of the right-hand side of equation (11) becomes

$$(1 - \bar{d})\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1^{*b} - \bar{d}\sum_{g=2}^G \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^{*b}; \tag{12}$$

this is the bootstrap analog of expression (8). Because $\bar{d} = N_1/N$, the first term in expression (12) must be the dominant one unless N_1 is extraordinarily large or the variance of the disturbances in the first cluster is extraordinarily small. In expression (12) and henceforth, we omit the factor c. Because it multiplies both the actual and bootstrap t statistics, it cannot affect bootstrap P values.

For the Rademacher distribution, the bootstrap disturbance vectors $\boldsymbol{\epsilon}_1^{*b}$ can have just two values, namely, $\tilde{\boldsymbol{\epsilon}}_1$ and $-\tilde{\boldsymbol{\epsilon}}_1$. When $G_1 = 1$, the distribution of the bootstrap statistics t_2^{*b} is then bimodal, with half the realizations in the neighborhood of t_2 and the other half in the neighborhood of $-t_2$; see MacKinnon and Webb (2017b, Figure 4). The wild cluster bootstrap fails for $G_1 = 1$ because the absolute value of the bootstrap test statistic is highly correlated with the absolute value of the actual test statistic. This makes it very difficult to obtain a bootstrap P value below any specified small level and leads to severe underrejection. However, the problem rapidly becomes less severe as G_1 increases.

It might seem that this problem could be solved by using unrestricted instead of restricted residuals in the bootstrap DGP (9). However, this creates a new problem, which is just as severe. When unrestricted residuals are used with $G_1 = 1$, the first term in expression (12) always equals zero, just like the first term on the left-hand side of equation (7), because the unrestricted residuals sum to zero for the single treated cluster. As a consequence, the bootstrap t statistics have far less variance than the actual t statistics, and the bootstrap test overrejects very severely. Again, the problem rapidly becomes less severe as G_1 increases.

Figure 1: Rejection frequencies for several tests, $G = 14$, $N/G = 200$, $\rho = 0.1$

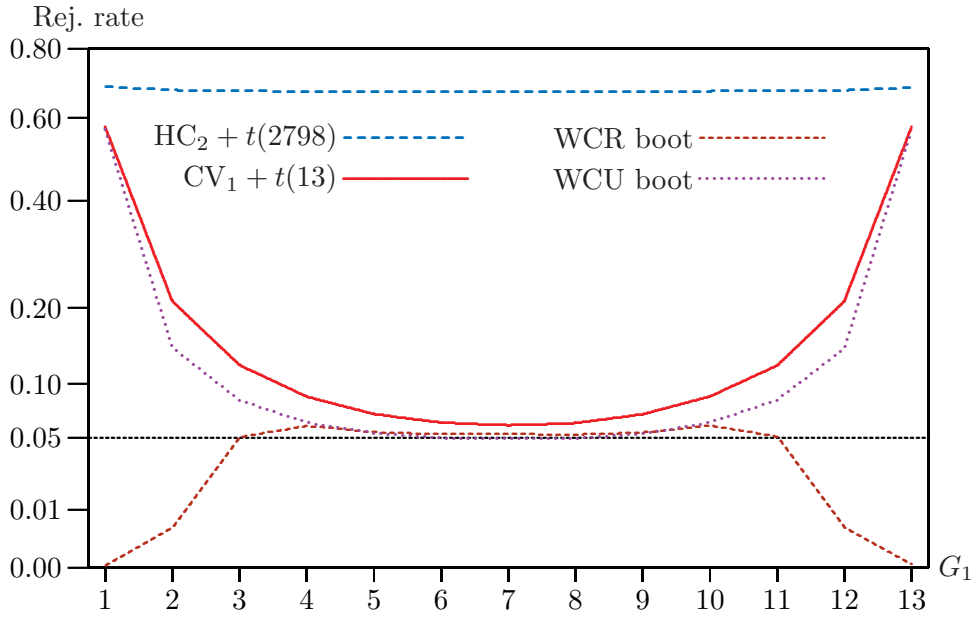


Figure 1 illustrates the poor performance of the procedures discussed so far when the number of treated clusters is small. It shows rejection frequencies at the .05 level for four tests with $G = 14$, $N/G = 200$ for all g, and $\rho = 0.10$. The horizontal axis shows the number of treated clusters, G_1, which varies from 1 to 13. The vertical axis has been subjected to a square root transformation in order to present both large and small rejection frequencies in the same figure. The rejection frequencies are based on 400,000 replications. For details of the experiments, see Section 4.

Simply using t statistics based on heteroskedasticity-robust standard errors—specifically, the HC_2 variant proposed in MacKinnon and White (1985)—combined with the $t(2798)$ distribution results in very severe overrejection for all values of G_1. This overrejection would have been even more severe if either N/G or ρ had been larger.

Using t statistics based on the CV_1 covariance matrix (3), combined with the $t(13)$ distribution, leads to severe overrejection when $G_1 = 1$ and $G_1 = 13$, but the overrejection is much less severe for values of G_1 that are not too far from $G/2$. This is exactly what the arguments of Subsection 2.1 suggest.

The two wild cluster bootstrap methods perform exactly as the analysis of MacKinnon and Webb (2017b) predicts. The restricted wild cluster bootstrap (WCR) almost never rejects for $G_1 = 1$ and $G_1 = 13$, underrejects severely for $G_1 = 2$ and $G_1 = 12$, performs almost perfectly for $G_1 = 3$ and $G_1 = 11$ (a coincidence that would not have occurred if G had been larger or smaller), and overrejects modestly for other values of G_1. In contrast, the unrestricted wild cluster bootstrap (WCU) overrejects very severely for $G_1 = 1$ and $G_1 = 13$, but it improves rapidly as G_1 becomes less extreme and performs extremely well for $6 \leq G_1 \leq 8$.

A very different bootstrap procedure is the pairs cluster bootstrap, in which the boot-

strap samples are obtained by resampling the matrices $[\boldsymbol{y}_g \quad \boldsymbol{X}_g]$ with replacement for $g = 1, \ldots, G$. This procedure has at least one major drawback: G_1 varies across the bootstrap samples and may well equal 0 for many of them. Because this procedure tends to overreject very severely when G_1 is small, we do not study it further; see MacKinnon and Webb (2017a).

3 The Wild and Subcluster Wild Bootstraps

The wild cluster bootstrap fails when $G_1 = 1$ because the same value of the auxiliary random variable v_g^{*b} multiplies every residual for cluster g. Thus the vector of bootstrap disturbances for the treated cluster is always proportional to the vector of residuals. This is an essential feature of the wild cluster bootstrap, because it allows the bootstrap samples to mimic the (unknown) covariance structure of the $\boldsymbol{\epsilon}_g$. But it leads to highly unreliable inferences when either G_1 or (in the pure treatment case) G_0 is small.

The idea of the subcluster wild bootstrap is to break up the vector of residuals within each cluster into mutually exclusive subvectors and multiply each subvector by an auxiliary random variable. In the simplest case, each subvector has just one element, and the subcluster wild bootstrap corresponds to the ordinary wild bootstrap. Of course, standard errors are still computed using a CRVE like (3); using the same form of t statistic for the original sample and the bootstrap samples is imperative.

Even though the wild bootstrap fails to capture some important features of the true DGP, it yields asymptotically valid inferences when both G_1 and G_0 are large, and it often yields greatly improved inferences when one or both of them is small. Most importantly, it yields (approximately) valid inferences for the pure treatment model (1) whenever all clusters are the same size and the amount of intra-cluster correlation is not too large, even when $G_1 = 1$. This is a very important special case.

In Section 3.4, we discuss variants of the subcluster wild bootstrap in which there are fewer subclusters than observations, so that each subcluster contains more than one observation. However, in the next three subsections, we focus on the ordinary wild bootstrap. It is the easiest one to describe and implement, and, in the case of cross-sectional data, it seems to be the one that should be used in practice most of the time.

3.1 The Ordinary Wild Bootstrap

The restricted wild bootstrap DGP analogous to equation (9) is

$$y_{ig}^{*b} = \tilde{\beta}_1 + \tilde{\epsilon}_{ig} v_{ig}^{*b}. \tag{13}$$

The only difference between equations (9) and (13) is that, for the former, the auxiliary random variable takes the same value for every observation in cluster g, and, for the latter, it takes an independent value for every observation. Instead of just two possible vectors of bootstrap disturbances $\boldsymbol{\epsilon}_g^{*b}$ for cluster g, there are now 2^{N_g} possible vectors.

Consider once again the special case in which $G_1 = 1$. Provided N_1 is not too small and the amount of intra-cluster correlation is not too large, the DGP (13) solves the problem of the absolute value of the numerator of the bootstrap test statistic being highly correlated with the absolute value of the numerator of the actual test statistic; see expression (12). Of course, solving this problem comes at a cost: The bootstrap disturbances no longer

mimic the covariance structure of the $\boldsymbol{\epsilon}_g$. Thus it may seem that using the bootstrap DGP (13) cannot possibly yield (approximately) valid inferences. However, it actually does so in at least two important cases.

The first case is when G tends to infinity and the limit of $\phi \equiv G_1/G$ is strictly between 0 and 1. The ordinary wild bootstrap works in this case because, whenever we bootstrap an asymptotically pivotal test statistic, the asymptotic validity of bootstrap tests does not require the bootstrap DGP to mimic the true, unknown DGP. It merely requires that the bootstrap DGP belongs to the family of DGPs for which the test statistic is asymptotically pivotal. Two papers in which this point has been explicitly recognized are Davidson and MacKinnon (2010) and Gonçalves and Vogelsang (2011).

Consider the t statistic (4) and its bootstrap analog (11). Under the wild bootstrap DGP (13), the numerators of (4) and (11) do not have the same distributions. But, in both cases, the denominator correctly estimates the standard deviation of the numerator when G is large and ϕ is bounded away from 0 and 1. Therefore, assuming that we can invoke a central limit theorem, both test statistics are approximately distributed as standard normal for large G, so that computing a bootstrap P value for (4) using the empirical distribution of B realizations of (11) is asymptotically valid. A formal proof of the asymptotic validity of the ordinary wild bootstrap for linear regression models with clustered disturbances is given in Djogbenou, MacKinnon and Nielsen (2018), which was written after this paper.

The second case, which we discuss in detail in Subsection 3.2, is when cluster sizes are equal and the covariance matrices $\boldsymbol{\Omega}_g$ for every g are the same up to a scalar factor λ_g. This implies that the patterns of intra-cluster correlation must be the same for all clusters, but there can be heteroskedasticity across them.

The wild bootstrap DGP (13) imposes the null hypothesis. We could instead use the unrestricted wild bootstrap DGP

$$y_{ig}^{*b} = \hat{\beta}_1 + \hat{\beta}_2 d_{ig} + \hat{\epsilon}_{ig} v_{ig}^{*b}, \tag{14}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are unrestricted OLS estimates, and the $\hat{\epsilon}_{ig}$ are unrestricted residuals. If the restricted wild bootstrap works well, then so should the unrestricted one, provided the bootstrap t statistic is redefined so that it is testing the hypothesis $\beta_2 = \hat{\beta}_2$ instead of the hypothesis $\beta_2 = 0$. Using (14) instead of (13) will inevitably affect the finite-sample properties of bootstrap tests, often making P values smaller, but it makes it much easier to compute confidence intervals. In the simulation experiments of Section 4, we study both the restricted and unrestricted wild and wild cluster bootstraps.

3.2 Equal Cluster Sizes

Our most important, and most surprising, result is that the ordinary wild bootstrap can yield approximately valid inferences even when G_1 is very small, provided all cluster sizes are the same, so that $N_g = N/G$. It is also essential that there not be too much intra-cluster correlation, especially when $G_1 = 1$, and that the covariance matrices $\boldsymbol{\Omega}_g$ satisfy a certain condition. The result is true even when there is an arbitrary pattern of heteroskedasticity at the cluster level.

Whenever we make approximations in this section, they are not asymptotic approx-

imations in the usual sense. The problem is that, when any of G, G_1, or G_0 is fixed as $N \to \infty$, the OLS estimator $\hat{\boldsymbol{\beta}} \equiv [\hat{\beta}_1 \quad \hat{\beta}_2]'$ in the model (1) is not consistent, at least not without very unrealistic assumptions about the intra-cluster correlations; see Carter, Schnepel and Steigerwald (2017). This inconsistency is implied by the results on regression with common shocks in Andrews (2005).

In the cases that interest us, where G and G_1 are fixed, the vector $\hat{\boldsymbol{\beta}}$ is asymptotically equal to $\boldsymbol{\beta}_0$ plus a random term, so that neither consistency nor asymptotic normality holds. However, when N, and hence the N_g, is large and the amount of intra-cluster correlation is not too large, we may reasonably expect this random term to be very small. The experiments in Section 4 confirm both the accuracy of this conjecture and the dependence of the quality of the approximation on N_g and the amount of intra-cluster correlation. We use the symbol "$\cong$" to denote approximations that should generally be accurate when these conditions hold.

From expressions (5) and (6), the actual t statistic under the null hypothesis is

$$t_2 = \frac{(1-\bar{d})\sum_{g=1}^{G_1} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g + \bar{d}\sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g}{\Big((1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2\Big)^{1/2}}. \tag{15}$$

Now consider the bootstrap t statistic based on the ordinary wild bootstrap DGP (13). Omitting the b superscripts for clarity, it is

$$t_2^* = \frac{(1-\bar{d})\sum_{g=1}^{G_1} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^* + \bar{d}\sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*}{\Big((1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g^*)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g^*)^2\Big)^{1/2}}. \tag{16}$$

The bootstrap t statistic (16) evidently has the same form as the t statistic (15), but with bootstrap disturbances replacing actual disturbances and bootstrap residuals replacing actual residuals in the numerator and denominator, respectively.

We now make the following key assumptions:

1. G, G_1, and N are fixed, with $N_g = N/G$ for $g = 1, \ldots, G$.
2. $\boldsymbol{\Omega}_g = \lambda_g\bar{\boldsymbol{\Omega}}$ for all g, for some positive definite matrix $\bar{\boldsymbol{\Omega}}$, with $\lambda_1 = 1$ and $\lambda_g > 0$.
3. The average intra-cluster correlation, say ρ, is small if G_1 is small.

Assumption 1, that the cluster sizes are equal, can always be verified, because they can be observed. In practice, the N_g only need to be approximately equal. Assumption 3 will be discussed below. Assumption 2 is important. It says that the covariance matrices for all clusters are proportional, with factors of proportionality λ_g that may differ. It follows that $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g) = \lambda_g\boldsymbol{\iota}_g'\bar{\boldsymbol{\Omega}}\boldsymbol{\iota}_g \equiv \lambda_g\omega^2$ for all g. Thus we are allowing there to be an arbitrary pattern of cross-cluster heteroskedasticity, but the same pattern of within-cluster correlation and heteroskedasticity for all clusters. The condition that $\lambda_1 = 1$ is just an arbitrary normalization.

From (15) and the definition of ω^2, we may conclude that, in this special case, the

variance of the numerator of t_2 is simply

$$(1-\bar{d})^2 \sum_{g=1}^{G_1} \lambda_g \omega^2 + \bar{d}^2 \sum_{g=G_1+1}^{G} \lambda_g \omega^2. \tag{17}$$

The variance of t_2 itself depends on how well the denominator of (15) estimates expression (17). This denominator involves two terms. The first involves a summation over G_1 random scalars $(\boldsymbol{\iota}_g' \hat{\boldsymbol{\epsilon}}_g)^2$ that estimates the first term in (17), and the second involves a summation over G_0 random scalars that estimates the second term.

Now define θ_1 as $1/(\lambda_g \omega^2)$ times the expectation of a typical element $(\boldsymbol{\iota}_g' \hat{\boldsymbol{\epsilon}}_g)^2$ in the first summation, and θ_0 as $1/\lambda_g \omega^2$ times the expectation of the same typical element in the second summation. In most cases, the factors θ_1 and θ_0 will be less than one, sometimes much less when G_1 or G_0 is very small; indeed, we saw in the previous section that $\theta_1 = 0$ when $G_1 = 1$. This point is discussed further at the end of this subsection. These two factors will almost always be different, because they depend on the numbers and sizes of the treated and untreated clusters.

We now assume that $\hat{\boldsymbol{\beta}} \cong \boldsymbol{\beta}_0$, which implies that $\hat{\epsilon}_{ig} \cong \epsilon_{ig}$ for all observations. For the approximation to be good, N should not be too small, and Assumption 3 must hold. If there were a substantial amount of intra-cluster correlation and G_1 were small, then $\hat{\boldsymbol{\beta}}$ might depend excessively on the the common component(s) of the disturbances for the treated cluster(s). When the approximation is a good one, the square of the denominator of (15) will be approximately equal to

$$(1-\bar{d})^2 \theta_1 \sum_{g=1}^{G_1} \lambda_g \omega^2 + \bar{d}^2 \theta_0 \sum_{g=G_1+1}^{G} \lambda_g \omega^2. \tag{18}$$

Thus, from (17) and (18), we conclude that

$$\text{Var}(t_2) \cong \frac{(1-\bar{d})^2 \sum_{g=1}^{G_1} \lambda_g + \bar{d}^2 \sum_{g=G_1+1}^{G} \lambda_g}{(1-\bar{d})^2 \theta_1 \sum_{g=1}^{G_1} \lambda_g + \bar{d}^2 \theta_0 \sum_{g=G_1+1}^{G} \lambda_g}. \tag{19}$$

Notice that ω^2 does not appear in this expression.

We now turn our attention to the bootstrap t statistic t_2^*. Because the ordinary wild bootstrap does not preserve intra-cluster correlations, the variance of $\boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g^*$ is not $\lambda_g \omega^2$. Instead, assuming again that $\hat{\epsilon}_{ig} \cong \epsilon_{ig}$ for all observations, it is approximately $\lambda_g N_g$ times σ^2, the average diagonal element of $\bar{\boldsymbol{\Omega}}$. Thus the variance of the numerator of t_2^* is approximately

$$(1-\bar{d})^2 \sum_{g=1}^{G_1} \lambda_g N_g \sigma^2 + \bar{d}^2 \sum_{g=G_1+1}^{G} \lambda_g N_g \sigma^2. \tag{20}$$

By essentially the same argument that led to expression (18), the square of the denominator of t_2^* must be approximately equal to

$$(1-\bar{d})^2 \theta_1 \sum_{g=1}^{G_1} \lambda_g N_g \sigma^2 + \bar{d}^2 \theta_0 \sum_{g=G_1+1}^{G} \lambda_g N_g \sigma^2. \tag{21}$$

Therefore, using (20) and (21), we conclude that

$$\mathrm{Var}(t_2^*) \cong \frac{(1-\bar{d})^2 \sum_{g=1}^{G_1} \lambda_g + \bar{d}^2 \sum_{g=G_1+1}^{G} \lambda_g}{(1-\bar{d})^2 \theta_1 \sum_{g=1}^{G_1} \lambda_g + \bar{d}^2 \theta_0 \sum_{g=G_1+1}^{G} \lambda_g}, \tag{22}$$

which is just expression (19). The factors of $N_g\sigma^2$ have cancelled out in the same way that the factors of ω^2 did previously. The same factors of λ_g appear in both (19) and (22) because the wild bootstrap preserves the heteroskedasticity of the original disturbances.

Our key result, from (19) and (22), is that

$$\mathrm{Var}(t_2^*) \cong \mathrm{Var}(t_2). \tag{23}$$

This result depends critically on Assumptions 1, 2, and 3. If N_g differed across clusters, in violation of Assumption 1, then the values of $N_g\sigma^2$ would differ across clusters. So would the values of $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$, which would now equal $\lambda_g\omega_g^2$ instead of $\lambda_g\omega^2$, for appropriately defined scalars ω_g^2. Without Assumption 1, we could not have made Assumption 2. If only the latter assumption were violated, it would again be the case that $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g) = \lambda_g\omega_g^2$ instead of $\lambda_g\omega^2$. Then the ratio of $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ to $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ would not be the same for all g, which is essential for the result (23) to hold. Assumptions 1 and 2 are not actually necessary. In principle, both the N_g and the $\boldsymbol{\Omega}_g$ could vary across clusters in such a way that the ratio of $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ to $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ is constant. Larger clusters would need to have less intra-cluster correlation than smaller ones.

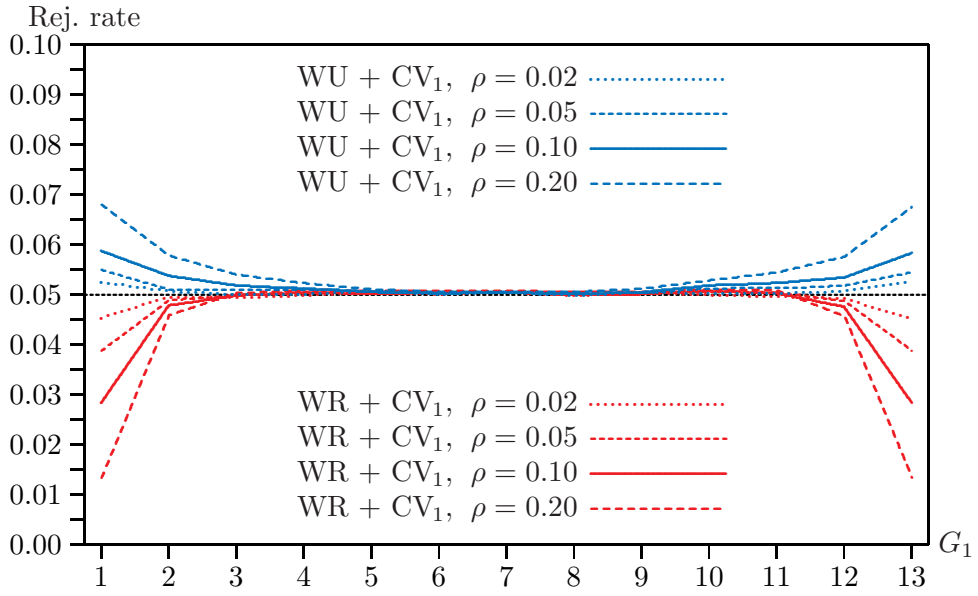
Assumption 3 is not stated precisely, because it seems to be impossible to do so. Just how much intra-cluster correlation is allowable necessarily depends on G, G_1, the sizes of both treated and untreated clusters and the patterns of intra-cluster correlation within them, the error in rejection frequency that is tolerable, and so on. When Assumption 3 is seriously violated, the wild bootstrap will fail in almost the same way as the wild cluster bootstrap fails. Suppose that $G_1 = 1$, which is by far the worst case. If the disturbances for cluster 1 happen to be unusually large in absolute value, so will be t_2, and so will be the absolute values of the restricted residuals $\tilde{\epsilon}_{i1}$. If the ϵ_{i1} are correlated, then the $|\tilde{\epsilon}_{i1}|$ will tend to be large when $\hat{\beta}_2$ is large. This will cause exactly the same sort of failure as occurs for the restricted wild cluster bootstrap; see the discussion around equation (12). We expect the restricted wild bootstrap to underreject in this case.

A similar argument applies to the unrestricted wild bootstrap. When the ϵ_{i1} are correlated, the $|\hat{\epsilon}_{i1}|$ will tend to be too small, causing the variance of the $\hat{\beta}_2^*$ to be too small. This will cause exactly the same sort of failure as occurs for the unrestricted wild cluster bootstrap; see the last paragraph of Subsection 2.2. We expect the unrestricted wild bootstrap to overreject in this case.

Our simulation results (see Section 4) suggest that the failure of Assumption 3 can cause serious errors of inference when $G_1 = 1$, but not when $G_1 \geq 2$, unless the amount of intra-cluster correlation is very large. Because the signs of the distortions caused by its failure are known, we can be confident that Assumption 3 is not seriously violated whenever the bootstrap P values for the restricted and unrestricted wild bootstraps are similar, with the former larger than the latter.

The argument that led to (23) does not imply that t_2 and t_2^* actually follow the same

Figure 2: Rejection frequencies for ordinary wild bootstrap tests, $G = 14$, $N/G = 200$



distribution under the null hypothesis. It merely suggests that they have approximately the same variance. When G is fixed, neither t_2 nor t_2^* will be asymptotically $\mathrm{N}(0,1)$ under the null. However, since the numerators of both test statistics are weighted sums of disturbances that have mean zero —compare (4) and (11)—it seems plausible that they will both be *approximately* normally distributed when N is large.

In order to obtain an asymptotic normality result, it is essential that G should tend to infinity as N tends to infinity, although perhaps at a slower rate; see Djogbenou, MacKinnon and Nielsen (2018). To see the problem, consider a random-effects model in which the disturbance ϵ_{ig} is equal to a cluster-level random effect v_i plus an individual random effect u_{ig}. When the number of clusters G is fixed, there are only G realizations of the v_i. Each of them must have a non-negligible effect on the OLS estimates. Therefore, the distribution of those estimates, and of t statistics based on them, must depend on the distribution of the v_i. Only by letting $G \to \infty$ could we invoke a central limit theorem in order to make the dependence on that distribution vanish asymptotically.

In the analysis that led to (23), we treated the denominators of t_2 and t_2^* as constants when they are in fact random variables. This should be a good approximation when Assumption 3 holds and N is reasonably large. Moreover, if those random variables have similar distributions for the actual and bootstrap samples, that should help to make the distribution of t_2^* mimic the distribution of t_2.

We also assumed that the factors θ_1 and θ_0, which determine how badly the two terms in the denominators of (15) and (16) underestimate the quantities they are trying to estimate, are the same for t_2 and t_2^*. It makes sense that these factors should be approximately the same, because the underestimation arises from the orthogonality between the OLS residuals and the treatment dummy, which is present for both the actual residuals and the bootstrap ones. The orthogonality causes the variances of sums of residuals to

be smaller than the variances of the corresponding sums of disturbances in a manner that depends on G_1, G_0, and the number of elements in each of the sums; see Section A.3 of the appendix to MacKinnon and Webb (2017b). If these factors were substantially different between the actual and bootstrap test statistics, then the approximation (23) would no longer hold. This is most likely to happen when Assumption 3 fails or N is small, because the residuals, which are used to construct the $\boldsymbol{\epsilon}_g^*$, might then be poor estimators of the disturbances.

Figure 2 shows rejection frequencies for the tests proposed in this section for the same cases as Figure 1, but for four values of ρ. These tests combine the ordinary wild bootstrap, either restricted (WR) or unrestricted (WU), with the CV_1 covariance matrix. We use the w2 wild bootstrap—see Davidson and Flachaire (2008) and MacKinnon (2013)—in which the i^{th} residual is divided by the square root of the i^{th} diagonal element of either the projection matrix $\boldsymbol{M_X} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, or its restricted version, as appropriate, before being multiplied by the auxiliary random variable. This procedure is analogous to using HC_2 standard errors.

The new tests perform extraordinarily well, with two exceptions. They do not perform well when $G_1 = 1$ or $G_1 = 13$ and $\rho > 0.02$, or when $G_1 = 2$ or $G_1 = 12$ and $\rho = 0.20$. These are cases where Assumption 3 is seriously violated and wild cluster bootstrap tests fail dramatically; see Figure 1. Even when $G_1 = 1$ and $G_1 = 13$, the new tests perform quite well for $\rho = 0.02$ and, arguably, for $\rho = 0.05$. They always perform very much better than the wild cluster bootstrap.

3.3 Differing Cluster Sizes and Difference in Differences

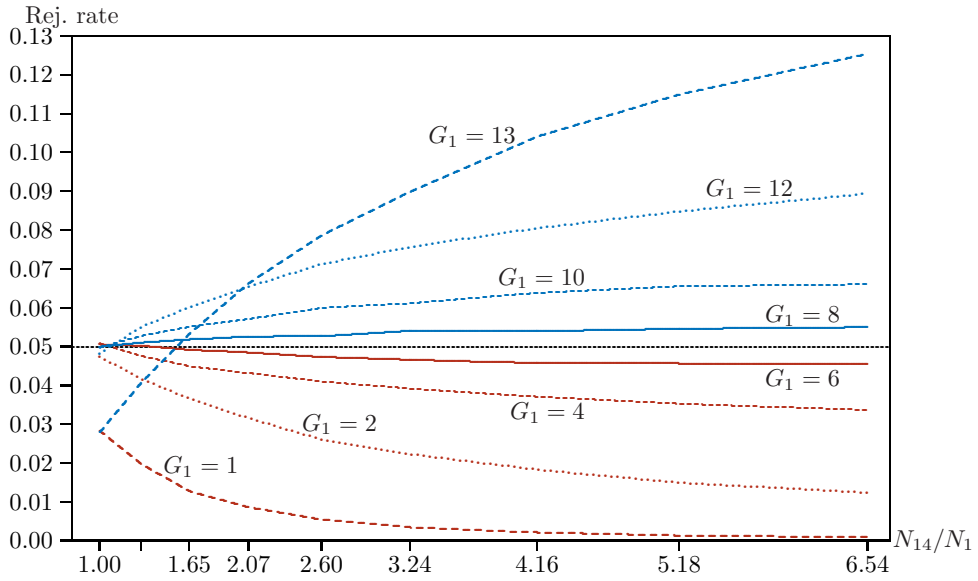
The key result (23) depends critically on Assumption 1. Without it, the ratio of $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ to $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ would not be the same for all g, and t_2^* would not have approximately the same variance as t_2 when G_1 or G_0 is small. The ratio would evidently be larger for large clusters than for small ones, because the number of off-diagonal terms is proportional to N_g^2, and these terms must surely be positive, at least on average.

Suppose that, instead of being the same size, the treated clusters were all smaller than the untreated ones. This would make the variance of the first term in the numerator of t_2 smaller relative to the variance of the second term, and likewise for the first and second terms in the numerator of t_2^*; see equations (15) and (16). However, the effect would be stronger for t_2 than for t_2^*, because $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ increases faster than N_g, while $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ is proportional to N_g. Since $1 - \bar{d} >> \bar{d}$ unless a large proportion of the clusters is being treated, it is primarily the first terms that determine $\text{Var}(t_2)$ and $\text{Var}(t_2^*)$. Moreover, it is the first terms that the corresponding terms in the denominators of t_2 and t_2^* underestimate (often severely) when G_1 or G_0 is small.

We conclude that, when G_1 is small (at any rate, not too much larger than $G/2$), and the treated clusters are smaller than the untreated ones, it must be the case that $\text{Var}(t_2^*) > \text{Var}(t_2)$. This will lead the ordinary wild bootstrap test to underreject. By a similar argument, the test will overreject whenever the treated clusters are larger than the untreated ones. Of course, this is only a problem when at least one of G_1 and G_0 is small. For G_1 and G_0 sufficiently large, the denominators of t_2 and t_2^* correctly estimate the variances of the numerators, and so $\text{Var}(t_2) \cong \text{Var}(t_2^*) \cong 1$.

To investigate the effect of varying cluster sizes, we allow the N_g to depend on a

Figure 3: Effects of varying cluster sizes on rejection frequencies for WR + CV_1



parameter γ that varies between 0 and 2. When $\gamma = 0$, all clusters are the same size. When $\gamma = 2$, the largest cluster is about 6.5 times as large as the smallest one. For details, see Section 4. In the experiments, $G = 14$, $\rho = 0.10$, and the average value of N/G is 200.

Figure 3 plots rejection frequencies at the .05 level for the restricted (WR) variant of the wild bootstrap when clusters are treated from smallest to largest. Instead of γ, which is hard to interpret, the horizontal axis shows the ratio of the largest to the smallest cluster size. There are eight curves, which correspond to $G_1 = 1, 2, 4, 6, 8, 10, 12, 13$. We expect to see increasing underrejection for $G_1 < 7$ as cluster sizes become more variable, and increasing overrejection for $G_1 > 7$, because treating the G_1 smallest clusters is equivalent to treating the $G - G_1$ largest clusters.

The ordinary wild bootstrap performs just as the theory of Subsection 3.3 predicts. It works quite well for $4 \leq G_1 \leq 10$ even when cluster sizes vary by a factor of more than six. Because $\rho = 0.10$, it underrejects fairly severely for both $G_1 = 1$ and $G_1 = 13$ when all clusters are the same size. It then underrejects more and more severely for $G_1 = 1$, and it overrejects more and more severely for $G_1 = 13$, as cluster sizes become more variable. Performance for $G_1 = 2$ and $G_1 = 12$ is much better than for $G_1 = 1$ and $G_1 = 13$ but still not very good when cluster sizes vary by a factor of three or more. Results for the WU variant, not shown in the figure, are quite similar except for $G_1 = 1$ and $G_1 = 13$, where there is overrejection instead of underrejection when all clusters are the same size.

The situation depicted in Figure 3 is a rather extreme one. In practice, it should be rare for only the largest or the smallest clusters to be treated. Thus, for $G_1 \geq 2$, we would generally expect to see better performance than is shown in the figure. Moreover, since the investigator knows the cluster sizes, he or she will know whether the wild bootstrap is likely to overreject or underreject. For example, if the treated clusters are, on average, smaller than the untreated ones, there is likely to be underrejection. In that case, a

significant bootstrap P value would provide strong evidence against the null hypothesis, but an insignificant one might be misleading.

We could create a sample with equal-sized clusters by taking averages of individual observations. For example, if every observation is associated with a jurisdiction and a time period, we could create a balanced panel by averaging over all the observations associated with each jurisdiction and time period. Unfortunately, this will probably not yield good results if the sample is not balanced originally. When we take averages over different numbers of observations, we implicitly create intra-cluster covariance matrices that depend on those numbers. As a result, Assumption 2 will be violated.

The result that $\text{Var}(t_2) \cong \text{Var}(t_2^*)$ when cluster sizes are equal applies only to pure treatment models like (1). In the case of difference-in-differences regressions, only some of the observations in the treated clusters are actually treated. Untreated observations may belong either to the control clusters in any period or to the treated clusters in the pre-treatment period. This means that expression (5) for the numerator of the t statistic has to be replaced by

$$(1 - \bar{d}) \sum_{g=1}^{G_1} \boldsymbol{d}_g' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=1}^{G_1} (\boldsymbol{\iota}_g - \boldsymbol{d}_g)' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g. \tag{24}$$

Recall that the $\boldsymbol{d}_g$ are N_g-vectors equal to 1 for treated observations and 0 for untreated ones. The numerator of the t statistic now has three terms instead of two. The first term corresponds to the treated observations in the treated clusters, the second corresponds to the untreated observations in the treated clusters, and the third corresponds to the untreated clusters. The first two terms are not independent, because they both depend on the same set of treated clusters.

It is clear from expression (24) that the analysis which led to the approximations (19) and (22) does not apply to the DiD case. The previous arguments about what happens when cluster sizes differ suggest that the subcluster bootstrap is likely to underreject (overreject) when the number of treated observations in each treated cluster is small (large) relative to the number of untreated observations, and/or relative to the number of observations in each untreated cluster. Underrejection will probably be more common than overrejection, however, because the number of treated observations per treated cluster can only be relatively large if two conditions are satisfied: The treated clusters must be relatively large, and a substantial fraction of the observations in them must be treated. In most cases, we would not expect both these conditions to be satisfied. Section 4 provides some evidence on how well the wild and wild cluster bootstraps perform in the DiD case.

3.4 Using Actual Subclusters

Up to this point, we have only discussed the wild cluster bootstrap and the ordinary wild bootstrap. In general, the subcluster wild bootstrap is a sequence of procedures with the former as one limiting case and the latter as the other. In between, there could potentially be a large number of bootstrap DGPs that involve some degree of clustering, but at a finer level than the covariance matrix estimator.

Recall from Subsection 3.3 that the ordinary wild bootstrap fails when cluster sizes vary and at least one of G_1 and G_0 is small, so that the denominators of the actual and

bootstrap t statistics do a poor job of estimating the variance of the numerators. The fundamental reason for this failure is that the ratio of $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ to $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ varies across clusters. This happens because, with the ordinary wild bootstrap, the elements of $\boldsymbol{\epsilon}_g^*$ are uncorrelated, while those of $\boldsymbol{\epsilon}_g$ are not.

Suppose the observations within each cluster fall naturally into subclusters. For example, with panel data, every observation will be associated with a time period as well as a jurisdiction. With survey data, every observation might be associated with a city or a county within a larger region. In such a case, equation (1) can be rewritten as

$$y_{itg} = \beta_1 + \beta_2 d_{itg} + \epsilon_{itg}, \tag{25}$$

where g indexes jurisdictions or regions, the level at which the covariance matrix is clustered, t indexes time periods or locations, and i indexes individual observations. In this case, there is a natural subcluster wild bootstrap DGP:

$$y_{itg}^{*b} = \tilde{\beta}_1 + \tilde{\epsilon}_{itg} v_{tg}^{*b}. \tag{26}$$

This is a variant of the wild cluster bootstrap, since the auxiliary random variable v_{tg}^{*b} is the same for all i within each tg pair. But it is not the usual wild cluster bootstrap, for which the auxiliary random variable would be v_g^{*b}.

For the DGP (26), the bootstrap disturbances will be correlated within subclusters but uncorrelated across them. If the correlations between ϵ_{itg} and ϵ_{jtg} are substantially larger than the correlations between ϵ_{itg} and ϵ_{jsg}, for $i \neq j$ and $s \neq t$, then much of the intra-cluster correlation is really intra-subcluster correlation. In this case, we would expect $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ to provide a better approximation to $\text{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ than would be the case for the ordinary wild bootstrap. In consequence, we would expect $\text{Var}(t_2^*)$ to be closer to $\text{Var}(t_2)$ and bootstrap tests to perform better when cluster sizes vary.

There is a contrary argument, however. Suppose that each cluster contains M observations that can be evenly divided into S equal-sized subclusters. Therefore, the total number of unique off-diagonal elements is $M(M-1)/2$, and the number of those that are contained within the S diagonal blocks is $M(M/S-1)/2$. The ratio of these numbers is $(M-1)/(M/S-1)$, which is always greater than S. Therefore, using S subclusters will capture a fraction of the intra-cluster correlations that is less than $1/S$. With unbalanced subclusters, this fraction would be further reduced. We conclude that, unless the intra-subcluster correlations are large relative to the remaining intra-cluster correlations, the potential gain from using actual subclusters instead of the ordinary wild bootstrap is likely to be modest.

Moreover, there is a cost to subclustering at anything but the individual level. With the restricted subcluster wild bootstrap, when the number of treated or untreated subclusters is small, the bootstrap t statistics will be correlated with the actual t statistic. With the unrestricted subcluster wild bootstrap, in the same cases, the variance of the bootstrap t statistics will be too small. These are precisely the reasons why the two variants of the wild cluster bootstrap fail when G_1 or G_0 is too small; see Subsection 2.2 above. The whole point of the subcluster wild bootstrap is to avoid this type of failure, but we are very likely to encounter it if we subcluster at too coarse a level.

Our tentative conclusion is that subclustering at a very fine level should yield results similar to those from using the ordinary wild bootstrap DGP, and subclustering at a very coarse level is likely to yield unreliable results unless G_1 and G_0 are both fairly large (in which case subclustering may not be necessary at all). Subclustering at an intermediate level may be beneficial if the correlations within subclusters are a lot higher than the correlations between them.

Subclustering at an intermediate level may also perform well when cluster sizes vary. Suppose, for example, that WR overrejects (as it is likely to do when the treated clusters are large) and WCR underrejects (as it is likely to do whenever G_1 is small). Then there may well be intermediate levels of subclustering for which the restricted subcluster wild bootstrap outperforms both of them. We consider this case, and also one in which the treated clusters are small, in Section 4.

4 Simulation Experiments

We perform a very extensive set of simulation experiments, mainly for the pure treatment model (1) with G small and G_1 often very small. The primary objective is to see whether combining the ordinary wild bootstrap DGP with CV_1 standard errors works as well the analysis of Subsection 3.2, which necessarily involves some approximations, suggests that it should. Secondary objectives are to study subclustering and to investigate situations in which the theory of Subsection 3.3 suggests that the ordinary wild bootstrap should not work well.

In all experiments, the disturbances are normally distributed, equicorrelated within clusters with correlation coefficient ρ, and uncorrelated across clusters. In most of them, there are 400,000 replications, and the bootstrap methods use $B = 399$ bootstrap samples.[3] Using such a large number of replications is essential in order to distinguish between experimental noise and small but systematic failures of exactness for the bootstrap tests.

Figure 1, in Subsection 2.2, shows rejection frequencies at the .05 level for four existing tests with $G = 14$, $N/G = 200$ for all g, and $\rho = 0.10$. This figure would have looked more or less the same for any moderate value of G. As G increases, the range of extreme values of G_1 for which the WCR bootstrap severely underrejects and the WCU bootstrap severely overrejects gradually becomes a little wider, but the range of moderate values for which both bootstrap tests perform well becomes larger relative to G. When $G = 40$, for example, both wild cluster bootstrap tests perform extremely well for $6 \leq G_1 \leq 34$. With the exception of the t test based on HC_2 standard errors, all of these tests appear to be almost invariant to the value of ρ.

Numerous experiments suggest that, whenever the WCR and WCU P values differ substantially, at least one of them must be seriously misleading. Thus it is often easy to tell when G_1 is too small. In contrast, when the two P values do not differ much and lead to the same conclusion, they both seem to be at least fairly reliable. Of course, the P values being similar does not guarantee that they are entirely reliable; consider the cases of $G_1 = 4$ and $G_1 = 10$ in Figure 1, where both methods overreject slightly.

[3] In empirical analysis, it is desirable to use a larger value for B, but 399 seems to work well in simulation experiments, where randomness in the bootstrap P values tends to average out across replications. We use 999 instead of 399 for the experiments that involve power, because there is power loss of order $1/B$.

Figure 4: Rejection frequencies for $G = 7$, $N/G = 200$

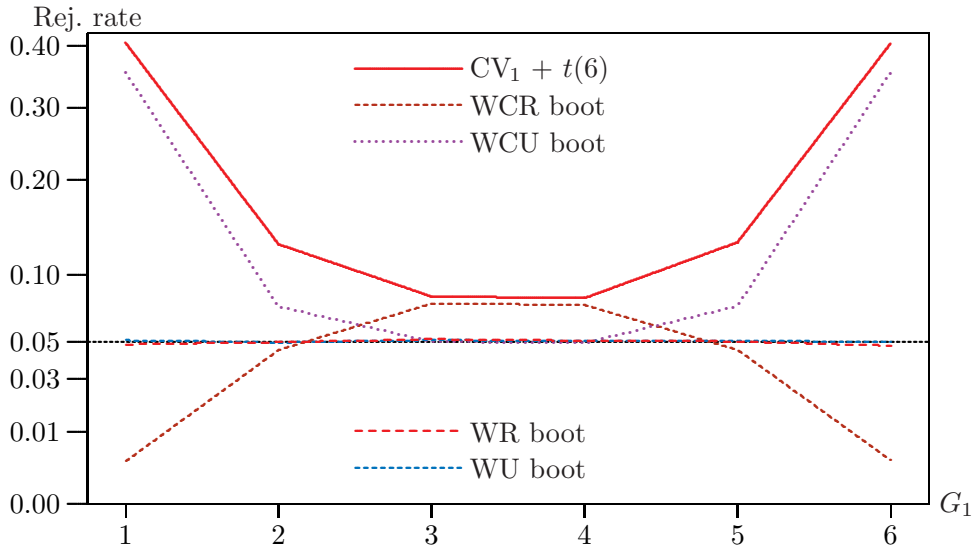


Figure 2, in Subsection 3.2, shows rejection frequencies at the .05 level for ordinary wild bootstrap tests, both WR and WU, for the same cases as Figure 1 and also for other values of ρ. Instead of the w2 wild bootstrap, we could have used the slightly simpler w0 procedure, which employs the restricted or unrestricted residuals without any rescaling. For the model (1), the two WR procedures are numerically identical, because the only regressor in the restricted model is a constant. Thus the rescaling involves the same factor for every observation, which for computing bootstrap t statistics is equivalent to not rescaling at all. However, the two WU procedures differ. Limited evidence suggests that the w2 procedure we use rejects slightly less often than the w0 procedure.

Figure 3, in Subsection 3.3, shows what happens when cluster sizes differ in a particular way. In those experiments, N_g is determined by a parameter γ, as follows:

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

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 $N/G = 200$ when $\gamma = 0$. As γ increases, cluster sizes become increasingly unbalanced. For the most extreme case in the figure, where $\gamma = 2$ and the clusters are sorted from smallest to largest, $N_1 = 67$ and $N_{14} = 438$.

Because the empirical example of Section 6 effectively involves 7 clusters, we performed a set of experiments similar to the ones in Figures 1 and 2, but with $G = 7$. These used only 100,000 replications. Results are shown in Figure 4. The ordinary wild bootstrap procedures continue to perform extraordinarily well, although WR underrejects very slightly for $G_1 = 1$ and $G_1 = 6$. WCR works remarkably well for $G_1 = 2$ and $G_1 = 5$, and WCU works almost perfectly for $G_1 = 3$ and $G_1 = 4$.

Figure 5 investigates the consequences of using genuine subclusters. In these experiments, $G = 14$, $\rho = 0.10$, and $N_g = 256$ for all g. The horizontal axis shows the number

Figure 5: Rejection frequencies as level of subclustering changes, $G = 14$, $N/G = 256$

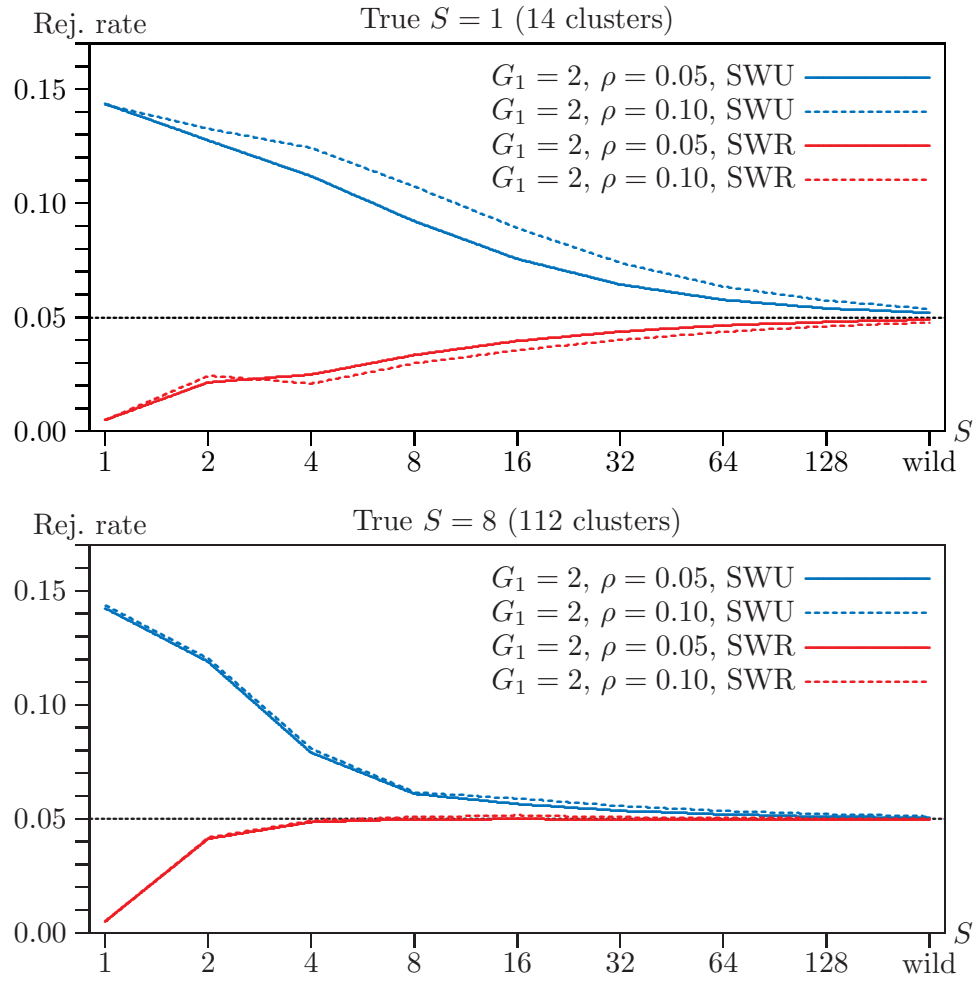
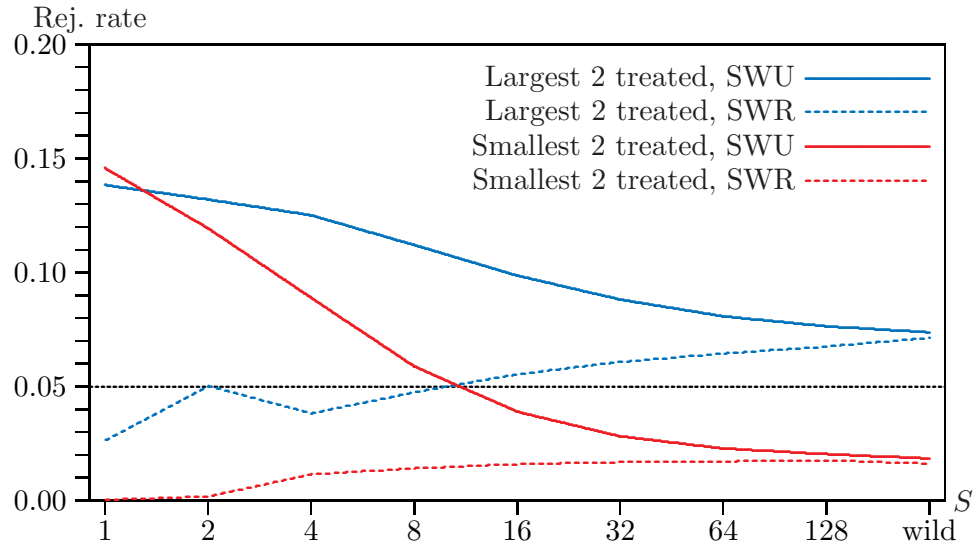


Figure 6: Rejection frequencies for subclustering when cluster sizes vary, $G = 16$



of subclusters S per cluster, which varies from 1 (the wild cluster bootstrap) to 256 (the wild bootstrap) by factors of 2. The vertical axis shows rejection frequencies for $G_1 = 2$ for restricted and unrestricted bootstrap tests for two values of ρ. The CRVE is always calculated using just 14 clusters, but the subcluster wild bootstrap (denoted SWR or SWU in the figure) uses between 14 ($S = 1$) and 3584 ($S = 256$) clusters.

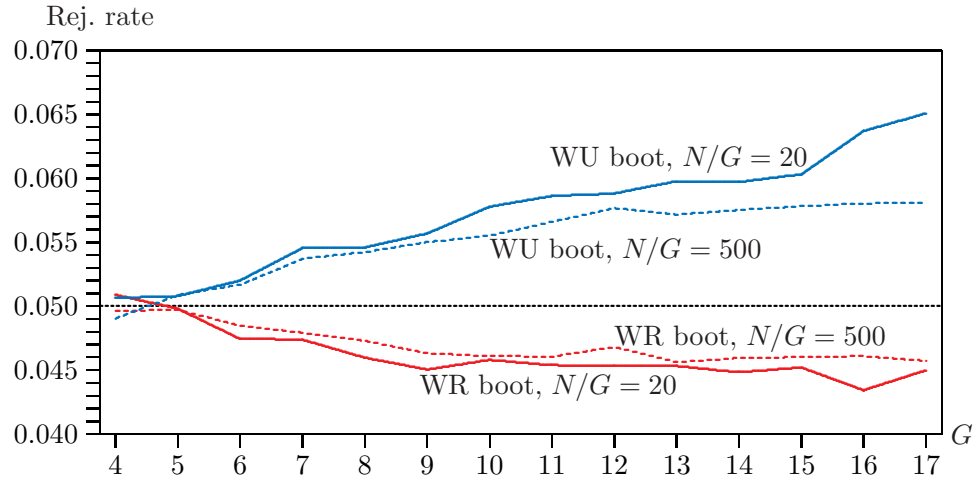
In the top panel of the figure, there are actually 14 clusters. Thus the "correct" value of S is 1. However, the unrestricted wild cluster bootstrap overrejects very severely, and the restricted one underrejects very severely. As the level of subclustering becomes finer, both procedures improve. For $S = 256$ (the ordinary wild bootstrap), they perform almost perfectly when $\rho = 0.05$ and extremely well when $\rho = 0.10$. The restricted versions always underreject less severely than the unrestricted ones overreject.

In the bottom panel of the figure, there are actually 112 clusters. Thus the "correct" value of S is 8, but the investigator mistakenly allows for much larger clusters than actually exist. One can think of this as analogous to clustering at the state level when in reality there is only clustering at the county level and the counties are nested within states. In this case, the restricted wild subcluster bootstraps work extremely well for $S \geq 4$. The unrestricted ones do not work quite as well, but for $S \geq 8$ they certainly perform very much better than they do in the top panel.

In Figure 5, results are shown for two values of ρ. We obtained results for larger and smaller values as well. As expected, the performance of WCR and WCU is almost invariant to ρ, but the performance of all the subclustering procedures deteriorates as ρ increases. This is particularly true for small values of S.

The ordinary wild bootstrap works particularly well in Figure 5 because all of the assumptions of Subsection 3.2 are satisfied. We relax assumption 1, that cluster sizes are equal, in Figure 6. There are now 16 clusters, four each with 128, 256, 384, and 512 observations, for a total of $N = 5120$. When cluster sizes vary, it matters which clusters are treated. The figure shows two extreme cases, in which either two of the largest clusters

Figure 7: Effect of G on rejection frequencies when $G_1 = 2$ and $\rho = 0.2$



($N_g = 512$) or two of the smallest clusters ($N_g = 128$) are treated. The horizontal axis again shows S, the number of subclusters per cluster. This varies from 1 (the wild cluster bootstrap) to 128 by factors of 2, and then finally to the wild bootstrap.

When the two largest clusters are treated, both WR and WU overreject, as the results of Subsection 3.3 predict. Since WCR underrejects (although nothing like as severely as it does for equal cluster sizes), there are several values of S for which SWR performs better than either WR or WCR. In fact, for $S = 8$, the rejection frequency is 0.0472. Similarly, when the two smallest clusters are treated, both WR and WU underreject, again as predicted. Since WCU overrejects, as usual, there are several values of S for which SWU performs better than either WR or WCR. The best case is again $S = 8$, for which the rejection frequency is 0.0587.

The results in Figure 5, together with additional ones for larger values of G_1 and other values of ρ, suggest that, when cluster sizes are equal, it is better to use the ordinary wild bootstrap than to subcluster at any level. However, Figure 6 shows that this is not so when cluster sizes vary. There must be many cases that look like the ones in the figure, in which WR and WU both either overreject or underreject, WCR underrejects, and WCU overrejects. In such cases, it seems likely that there will be subclustering schemes for which either SWR or SWU (but not both) outperform WR and WU.

There may also be cases in which Assumption 2 is violated and subclustering works well because the correlations within subclusters are substantially larger than the correlations across them. Recall that, in our experiments, the correlations both within and across subclusters within the same cluster are all the same.

The next set of experiments is designed to investigate the effects of the number of clusters and their (common) size. Figure 7 reports the results of several experiments for $G_1 = 2$ with balanced clusters. The value of ρ is 0.20, which in our view corresponds to the worst realistic case. The vertical axis shows rejection frequencies at the .05 level. The horizontal axis shows G, which varies from 4 to 17, because $G = 4$ is the smallest value for which $G_0 > 1$ when $G_1 = 2$. Both variants of the ordinary wild bootstrap perform almost perfectly when $G = 4$. As G increases, their performance gradually deteriorates, but it

is still generally quite good. As in Figure 2, the restricted variant underrejects, and the unrestricted variant overrejects. The former improves modestly as N/G increases, and the latter improves quite substantially.

We also obtained results with $\rho = 0.10$. As expected, both methods worked better than they do in Figure 7. The worst rejection rates for WR were 0.0459 for $N/G = 20$ and 0.0472 for $N/G = 500$. The worst rejection rates for WU were 0.0608 for $N/G = 20$ and 0.0538 for $N/G = 500$. The worst cases occurred when G was 16 or 17.

We showed in Subsection 3.2 that the ordinary wild bootstrap is approximately invariant to heteroskedasticity at the cluster level. To investigate this potentially important result, we perform a number of experiments in which the standard deviation σ_g for cluster g depends on a parameter δ, as follows:

$$\sigma_g = \exp\left(\frac{\delta(g-1)}{G-1}\right). \tag{27}$$

According to equation (27), σ_g equals 1 when $\delta = 0$ and is increasing in δ. In the experiments, δ varies between -2 and 2, so that σ_g varies between 0.135 and 7.39. The treated clusters are always the ones with the highest indices. Therefore, $\exp(\delta)$ can be thought of as the ratio of the highest standard deviation for a treated cluster to the lowest standard deviation for an untreated cluster.

All the experiments have 400,000 replications, with $G = 14$, $N = 2800$, $\rho = 0.10$, and $B = 399$. They are therefore comparable to the experiments of Figures 1 and 2. We performed five sets of experiments, for $G_1 = 1, 2, \ldots, 5$. However, for reasons of space, we only report results for $G_1 = 2$ and $G_1 = 4$.

Figure 8 plots rejection frequencies at the .05 level against δ for all four tests. As the theory of Subsection 3.2 predicts, the two ordinary wild bootstrap tests work very well, especially when $G_1 = 4$. In contrast, the performance of the two wild cluster bootstrap tests is very sensitive to δ. Remarkably, when $G_1 = 4$ and $\delta \geq 0.4$, the WCR bootstrap rejects more often than WCU. Given the slopes of the two curves near $\delta = 2$, it seems very likely that this would also be the case for $G_1 = 2$ when δ is large enough.

There has been very little investigation of the effects of heteroskedasticity across clusters on inference using the wild cluster bootstrap. In particular, all of the simulations in MacKinnon and Webb (2017b) assume that the disturbances are homoskedastic. Figure 8 suggests that the wild cluster bootstrap can perform much worse under heteroskedasticity than under homoskedasticity. Since that is not the case for the ordinary wild bootstrap, it may be attractive to use the latter when there is cluster-specific heteroskedasticity even when G_1 is not particularly small.[4]

All of the experimental results so far are for the pure treatment case, in which every observation in the treated clusters is treated. In Subsection 3.3, we showed that the key results (19) and (22) do not apply to DiD regression models. To investigate the performance of the ordinary wild bootstrap for these models, we performed another set of experiments in which only a fraction ψ of the observations in the treated clusters is treated. The experiments have $G = 20$, $N = 4000$, $\rho = 0.10$, and $\psi = 0.05, 0.10, \ldots, 1.00$.

[4]Experiments in MacKinnon and Webb (2018a) show that randomization inference procedures also perform poorly with cluster-specific heteroskedasticity and few treated clusters.

Figure 8: Effects of heteroskedasticity on rejection frequencies

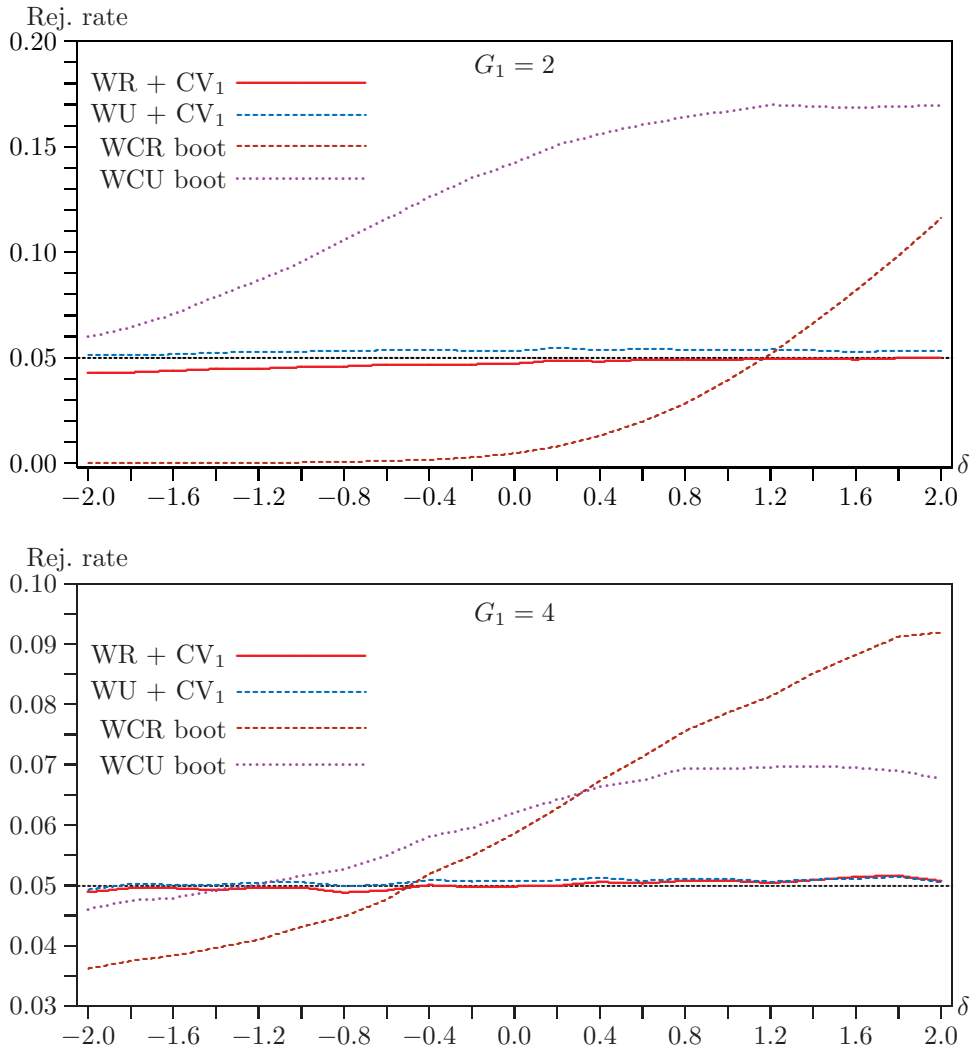
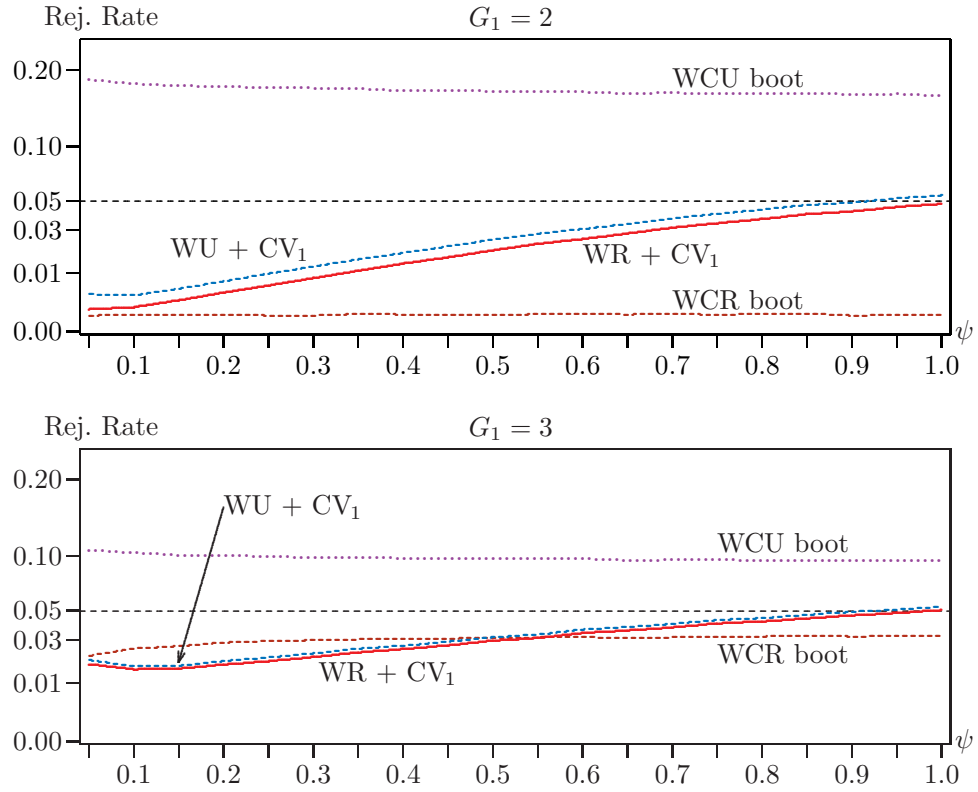


Figure 9: Effects of fraction of treated observations in treated clusters



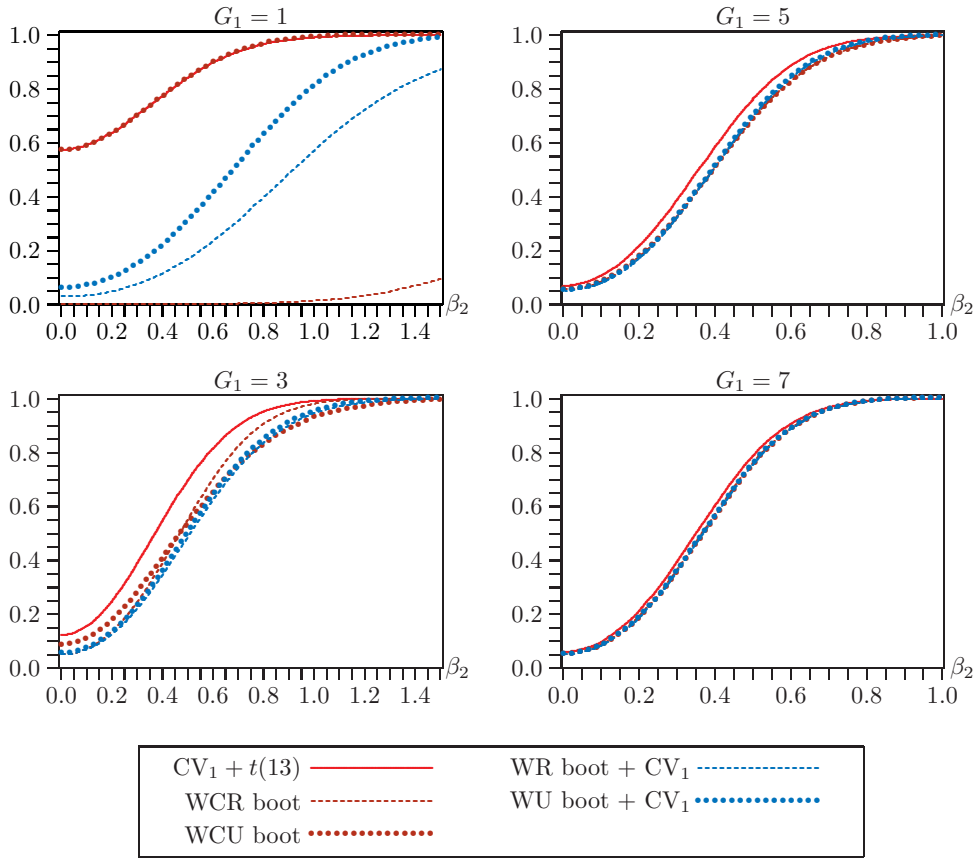
These experiments use $G = 20$ rather than $G = 14$ because the wild cluster bootstrap happens to perform very well when $G = 14$ and $G_1 = 3$, even though it usually performs poorly with such a small number of treated clusters.

Figure 9 reports rejection frequencies at the .05 level as functions of ψ for four tests. The top panel shows results for $G_1 = 2$, and the bottom panel shows results for $G_1 = 3$. As expected, the two wild cluster bootstrap tests perform badly when $G_1 = 2$. The restricted variant (WCR) very rarely rejects, and the unrestricted one (WCU) always rejects more than 16% of the time. In contrast, the two wild bootstrap tests perform about the same, with the unrestricted variant (WU) always rejecting a bit more often than the restricted one (WR). Both tests underreject severely when ψ is small, but the extent of the underrejection diminishes steadily as ψ increases. When $\psi = 1$, WU actually overrejects very slightly.

All four tests perform much better when $G_1 = 3$, but WCR still underrejects quite noticeably, and WCU still overrejects fairly severely. The two wild bootstrap tests still underreject (except for WU when $\psi \geq 0.95$) but not nearly as much as when $G_1 = 2$. They are fairly reliable for $\psi \geq 0.75$, always rejecting at least 4% of the time.

An actual DiD model with treatments starting at different times would normally include a full set of time and cluster dummy variables. We did not use such a model here, partly for reasons of computational cost, but more importantly because the dummies would eliminate any intra-cluster correlations in the disturbances of the DGP. There-

Figure 10: Power of Various Procedures



fore, in order for there to be any reason to use a CRVE, we would need to use a more complicated DGP that creates intra-cluster correlations which dummies cannot eliminate.

It seems very unlikely that the amount of intra-cluster correlation left after regressing on a full set of dummy variables would be anything like as large as 0.10 on average. Thus the results for the wild bootstrap tests in Figure 9 are probably quite a bit worse than we would see in practice with 20 clusters of 200 observations each. Of course, with clusters that were substantially larger or variable in size, we might well see even worse results.

The final simulation experiments concern power. Figure 10 presents results from four sets of experiments for equation (1) with 100,000 replications and $\rho = 0.10$. In all cases, $G = 14$, $N = 2800$, $N_g = 200$, and $B = 999$. The four panels show results for $G_1 = 1$, 3, 5, and 7. In the experiments, β_2 varies from 0.00 to 1.50 for $G_1 = 1$ and $G_1 = 3$, and from 0.00 to 1.00 for $G_1 = 5$ and $G_1 = 7$. Nominal 5% rejection frequencies are calculated for tests based on CV$_1$ and the $t(13)$ distribution, as well as for the WCR, WCU, WR, and WU bootstrap tests.

In the top left panel of Figure 10, where $G_1 = 1$, we see that the WCR bootstrap is severely lacking in power, while CV$_1$ t tests and the WCU bootstrap apparently have a great deal of power. The theoretical results in Section 6 of MacKinnon and Webb (2017b) suggest that, under the null when $G_1 = 1$, WCR should underreject severely and CV$_1$

and WCU should be similar to each other and overreject severely. The same relationships among these three tests evidently hold for all values of β_2. The only useful tests here are WU and WR, which have rather different power functions. The former appears to have considerably more power, but it also rejects quite a bit more often under the null.

As G_1 increases, the power functions for the five tests converge. There are still noticeable differences when $G_1 = 3$, however. In this case (see the lower left-hand panel), the power functions for the bootstrap tests cross. WCU overrejects under the null and consequently has higher power than the other tests for small values of β_2, but WCR is the most powerful bootstrap test for large values. When $G_1 = 7$ (see the lower right-hand panel), there are almost no observable differences among the four bootstrap procedures. CV_1 seems to have a little more power, but it is also slightly oversized. For $\beta_2 = 0$, it rejects 5.97% of the time, compared with between 4.89% (WCU) and 5.25% (WCR) for the bootstrap procedures.

Based on these experiments, it appears that size and power generally go hand in hand. Tests that have approximately the correct size generally have very similar power functions. Tests that overreject under the null have relatively high power, and tests that underreject under the null have relatively low power.

5 Alternative Procedures

The bootstrap is not the only way to obtain inferences that are more accurate than using CV_1 standard errors. The most widely-used approach, which is due to Bell and McCaffrey (2002), is to replace CV_1 by the alternative estimator

$$\text{CV}_2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{M}_{gg}^{-1/2}\boldsymbol{X}_g\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{28}$$

where $\boldsymbol{M}_{gg}^{-1/2}$ is the inverse symmetric square root of the g^{th} diagonal block of the $N \times N$ projection matrix $\boldsymbol{M}_{\boldsymbol{X}} = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. This block is the $N_g \times N_g$ symmetric matrix $\boldsymbol{M}_{gg} = \mathbf{I}_{N_g} - \boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g'$. Thus CV_2 omits the scalar factor in CV_1 and replaces the residual subvectors $\hat{\boldsymbol{\epsilon}}_g$ by rescaled subvectors $\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{\epsilon}}_g$.

The CV_2 estimator generalizes the HC_2 heteroskedasticity-consistent covariance matrix estimator discussed in MacKinnon and White (1985), and the former reduces to the latter when all the N_g are equal to 1. Both these estimators are intended to correct the downward bias of the OLS residuals. The fact that the CV_2 estimator would be unbiased if the $\boldsymbol{\Omega}$ matrix were proportional to an $N \times N$ identity matrix suggests that it may be attractive; see Young (2016).

Methods that employ bias-reduced standard errors generally also adjust the degrees of freedom used to compute P values or critical values with the t distribution. The first implementation of such a procedure is in Bell and McCaffrey (2002). More recently, Imbens and Kolesár (2016) suggests a slightly modified version of the Bell-McCaffrey procedure, and Young (2016) proposes a way to reduce the bias of the diagonal elements of CV_1 and compute critical values. All these procedures are discussed, and their performance studied by simulation, in the appendix. We mention them here because we implement them in the empirical example of Section 6.

6 Empirical Example

Angrist and Lavy (2001) studies the impact of teacher training on student outcomes using a matched comparisons design in Jerusalem schools. The paper tests whether students who were taught by teachers that received additional training increased their test scores by more than students taught by teachers with no additional training. The analysis is done separately for students in religious and secular schools. We focus our attention on 255 students taught in eight religious schools. With one exception, each school was either treated or not treated. The eight schools had 54, 48, 41, 40, 28, 24, 19, and 1 students, respectively. Although the example nominally has $G = 8$ and $G_1 = 3$, it effectively has $G = 7$ and $G_1 = 2$, because there is one untreated school with just one student and one school with 52 untreated students and just two treated students.[5]

We restrict attention to the change in math scores between 1994 and 1995, as this coefficient is puzzling but reported to be quite statistically significant; see column 4 of Table 5 in the original paper. The experimental design allows for a very simple identification strategy:

$$\text{diff}_{is} = \beta_0 + \beta_1 \text{treated}_{is} + \epsilon_{is}.$$

Here diff_{is} is the difference in math scores for student i in school s between 1994 and 1995, and treated_{is} is an indicator for whether a student was in a school taught by a treated teacher. For the religious schools, both 1994 and 1995 are pre-treatment years, so that a test of $\beta_1 = 0$ can be regarded as a test for common trends. The standard errors are clustered by school.

In column 1 of Table 1, we repeat the analysis of Angrist and Lavy (2001) and add numerous additional results. Our coefficient estimate is essentially the same as the one reported in the paper, but our standard error estimate is somewhat smaller.[6] The CRVE P value, based on the $t(7)$ distribution, suggests that the treatment has a negative impact that is statistically significant at well below the 1% level.

In column 1 of the second block of results, we report four bootstrap P values, using wild cluster and wild bootstraps, both restricted and unrestricted. All bootstrap P values use $B = 99,999$ replications. Because $G = 8$, the wild cluster bootstrap DGPs use the six-point distribution proposed in Webb (2014). The ordinary wild bootstrap DGPs use the Rademacher distribution. All four bootstrap procedures agree that the coefficient is significant only at the 5% level.

It may seem surprising that all four bootstrap procedures agree in this case. Since the two treated schools are only a little larger than the average size of $254/7 = 36.3$ (ignoring the school with just one student), it is not surprising that the ordinary wild bootstrap works well. The two wild cluster bootstrap procedures actually work well despite the fact that G_1 is very small because G is extremely small. Figure 4 in Section 4 shows rejection frequencies for $G = 7$ and $G_1 = 1, 2, \ldots, 6$ with equal-sized clusters, and the WCB procedures (especially WCR) work reasonably well when $G_1 = 2$ and $G_1 = 5$.

[5] Example code for estimating WCR and WR P values can be found at http://qed.econ.queensu.ca/pub/faculty/mackinnon/wild-few/

[6] Our coefficient estimate is actually -0.866476, which we report as -0.866. Angrist and Lavy (2001) reports a value of -0.867, which is what would have been obtained if the original estimate were first rounded to 4 and then to 3 digits.

Table 1: Effects of Teacher Training on Math Score Difference

	full sample	drop 48	drop 40
coefficient	−0.866	−0.778	−0.903
CV_1 std. error	0.195	0.206	0.205
t stat. (P value)	−4.45 (.003)	−3.78 (.009)	−4.41 (.005)
WCR P value	0.031	0.411	0.322
WCU P value	0.024	0.053	0.033
WR P value	0.020	0.247	0.109
WU P value	0.014	0.152	0.039
CV_1^{br} std. error	0.233	0.397	0.387
CV_2 std. error	0.207	0.279	0.285
df_Y	3.055	3.499	3.765
df_{BM}	2.366	1.534	1.642
df_{IK}	2.030	2.792	3.102
$\hat{\rho}$	0.081	0.100	0.111
CV_2 + $t(7)$ P value	0.004	0.032	0.019
CV_1^{br} + df_Y P value	0.033	0.132	0.084
CV_2 + df_{BM} P value	0.039	0.144	0.111
CV_2 + df_{IK} P value	0.051	0.075	0.048
N	255	207	215
G	8	7	7
G_1	3	2	2

Notes:

The outcome variable is the difference between 1994 and 1995 math test scores. All bootstrap P values use $B = 99{,}999$.

Because there is one school with just one student, and one otherwise untreated school with just two treated students, the effective values of G and G_1 are probably smaller by 1 than the reported values.

CV_1^{br} is the bias-reduced standard error proposed in Young (2016). df_Y, df_{BM}, and df_{IK} are, respectively, the degrees of freedom obtained by the methods of Young (2016), Bell and McCaffrey (2002), and Imbens and Kolesár (2016); see the appendix for details.

The next block in the table reports two alternative standard errors, both of which are somewhat larger than the usual CV_1 standard error. The following block reports the degrees of freedom calculated by three different methods, which are described in the appendix. These are much smaller than $G - 1 = 7$. The penultimate block reports four P values based on the alternative standard errors and various degrees of freedom. At 0.004, the P value based on the CV_2 standard error and the $t(7)$ distribution is not much larger than the one based on the CV_1 standard error and $t(7)$, but the others are a good deal larger. The procedure of Imbens and Kolesár (2016) actually yields a P value that is slightly greater than 0.05.

In order to make inference more difficult, we next drop either the school with 48 treated students or the school with 40 treated students from the sample; see columns 2 and 3 of Table 1. After dropping either of these schools, we are left with two treated schools, one of which only has two treated students. When we do this, neither the coefficient nor the standard error changes much. Both alternate samples yield CRVE P values, based on the $t(6)$ distribution, that are significant at the 1% level.

It seems very strange that dropping roughly half the treated students apparently has very little effect on the significance of the estimated coefficient. In fact, it does have a substantial effect, which is masked by the unreliability of cluster-robust standard errors when G_1 is very small. This is clear from the bootstrap P values. In all cases, the P values based on restricted estimates are much larger than the ones based on unrestricted estimates. None of the former suggest that the null hypothesis should be rejected.

The difference between the P values based on restricted and unrestricted estimates is much more pronounced for the wild cluster bootstrap (WCR and WCU) than for the wild bootstrap (WR and WU). The former is precisely what the theory reviewed in Subsection 2.2 implies, so that the WCR and WCU P values evidently convey very little information. The WR and WU and P values also do not yield unambiguous results, but they are very much closer, and for column 2 they yield the same inferences.[7] Moreover, there are two reasons to suspect that the WU P value of 0.039 in column 3 is too small: The treated school in that case is relatively large, and the WR P value is quite a bit larger than the WU one. Thus, if the results in column 3 were the only ones we had, it would be reasonable to conclude that there is insufficient evidence against the null hypothesis.

For the full sample, there is not much conflict among the four bootstrap P values and the three P values that use bias-reduced standard errors together with calculated degrees of freedom. Every procedure rejects the null or comes very close to doing so. This contrasts with the very small P values obtained using either CV_1 or CV_2 standard errors together with the $t(7)$ distribution. There is more conflict for the two subsamples, which is not at all surprising, because for both of them almost all the treated observations belong to a single cluster. Using a number of different procedures has revealed how fragile the results for each of the subsamples is.

7 Conclusion and Recommendations

Although the wild cluster bootstrap works well much of the time, MacKinnon and Webb (2017b) has shown that it often fails when the number of treated clusters is small, whether or not the total number of clusters is small; see Subsection 2.2. What very often happens in these cases is that the restricted wild cluster bootstrap P value is quite large, and the unrestricted wild cluster bootstrap P value is very much smaller. When that happens, neither of them can be trusted.

We have proposed a family of new bootstrap procedures, called the subcluster wild bootstrap, which includes the ordinary wild bootstrap as a limiting case. These procedures often work much better than the wild cluster bootstrap when there are few treated clusters.

[7] The differences between WR and WU are roughly what we would expect given the level of intra-cluster correlation. Our estimates of ρ are 0.0808 for the full sample, 0.0997 for the sample of column 2, and 0.1114 for the sample of column 3.

In principle, the subcluster wild bootstrap can be implemented in a variety of ways. However, it seems that the best approach is usually just to combine the ordinary wild bootstrap with cluster-robust standard errors.

We showed in Subsection 3.2 that, for a pure treatment model, the ordinary wild bootstrap can be expected to work very well under certain conditions. Firstly, clusters must be either treated or untreated. That is, if any observation in a cluster is treated, then every observation must be treated. Secondly, every cluster must have the same number of observations and the same covariance matrix up to a scalar factor which may be different for every cluster. Thirdly, the number of observations per cluster must be sufficiently large. Finally, if there is just one treated (or untreated) cluster, the intra-cluster correlations must be small, and if there are just two treated (or untreated) clusters, they must not be very large. When the last of these conditions is violated, the unrestricted (WU) P value will almost certainly be smaller than the restricted (WR) P value, so that it is easy to tell when there is a problem.

The conditions discussed in the previous paragraph are quite stringent. With just a few treated clusters, it is very likely that the ordinary wild bootstrap will underreject (overreject) when the treated clusters are smaller (larger) than average. It is also likely to underreject for difference-in-differences regression models with few treated clusters, unless the treated clusters are relatively large and have a large proportion of treated observations. In that case, it may overreject.

We have obtained a large number of simulation results. These results strongly confirm the theoretical results of Section 3 which predict when the ordinary wild bootstrap will or will not perform well. One unexpected result is that the wild cluster bootstrap, unlike the ordinary wild bootstrap, is very sensitive to heteroskedasticity across clusters when the number of treated clusters is small. This is a disturbing feature of the WCB that does not seem to have been observed previously.

Of course, bootstrap-based procedures are not the only ones that may be able to provide reasonably reliable inferences when there are few treated clusters. In the appendix, we discuss several recently proposed procedures which employ less-biased cluster-robust standard errors and calculate the appropriate degrees of freedom for each test. Procedures of this type can work very well in many cases, but none of them appears to dominate either the wild cluster bootstrap or the ordinary wild bootstrap across a wide range of cases. Moreover, some of these procedures can be computationally burdensome or even infeasible for sample sizes that are not large by current standards. In contrast, the wild and wild cluster bootstraps are perfectly feasible for samples with millions of observations in total and hundreds of thousands per cluster.

When the restricted (WCR) and unrestricted (WCU) variants of the wild cluster bootstrap yield similar inferences, there is no real need to employ any other procedure. The results may not be entirely reliable, especially if the number of treated clusters is small. However, unless the sample is dominated by one or two very large clusters, as in some of the experiments in Djogbenou, MacKinnon and Nielsen (2018), it seems to be very uncommon for both of them to be severely misleading in the same direction.

In practice, WCR and WCU will very often yield different inferences when the number of treated clusters is small. Typically, the latter will reject the null and the former will not. When that happens, we evidently cannot rely on the wild cluster bootstrap. In such

cases, the ordinary (or subcluster) wild bootstrap can often allow us to make reasonable, albeit imperfect, inferences, as in the empirical example of Section 6. Moreover, the wild bootstrap will probably outperform the wild cluster bootstrap when there is a substantial amount of cluster-specific heteroskedasticity, unless the numbers of treated and untreated clusters are so large that both procedures work very well.

In principle, for the ordinary wild bootstrap to provide valid inferences, we need the conditions of Subsection 3.2 to be satisfied. In practice, however, we are likely to obtain reasonably reliable inferences when the number of treated clusters is not too small (2 is a lot better than 1), when the treated and untreated clusters are approximately the same size, and when the sample size is not too small (50 observations per cluster is a lot better than 10 when there are not many clusters). It can also be useful as a conservative procedure even in the case of DiD models, where it will often tend to underreject. However, like the wild cluster bootstrap, the procedure should never be relied upon if the restricted and unrestricted wild bootstrap P values are not quite similar.

References

Andrews, Donald W. K. (2005) 'Cross-section regression with common shocks.' *Econometrica* 73(5), 1551–1585

Angrist, Joshua D, and Victor Lavy (2001) 'Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools.' *Journal of Labor Economics* 19(2), 343–69

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

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

Cameron, A. Colin, and Douglas L. Miller (2015) 'A practitioner's guide to cluster robust inference.' *Journal of Human Resources* 50(2), 317–372

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

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

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

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

Davidson, Russell, and James G. MacKinnon (2010) 'Wild bootstrap tests for IV regression.' *Journal of Business and Economic Statistics* 28(1), 128–144

Djogbenou, Antoine, James G. MacKinnon, and Morten Ø. Nielsen (2018) 'Asymptotic theory and wild bootstrap inference with clustered errors.' Working Paper 1399, Queen's University, Department of Economics

Ferman, Bruno, and Christine Pinto (2015) 'Inference in differences-in-differences with few treated groups and heteroskedasticity.' Technical Report, Sao Paulo School of Economics

Gonçalves, Silvia, and Timothy J. Vogelsang (2011) 'Block bootstrap HAC robust tests: The sophistication of the naive bootstrap.' *Econometric Theory* 27(4), 745–791

Ibragimov, Rustam, and Ulrich K. Müller (2016) 'Inference with few heterogeneous clusters.' *Review of Economics and Statistics* 98(1), 83–96

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

Liu, Regina Y. (1988) 'Bootstrap procedures under some non-I.I.D. models.' *Annals of Statistics* 16(4), 1696–1708

MacKinnon, James G. (2013) 'Thirty years of heteroskedasticity-robust inference.' In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis,* ed. Xaiohong Chen and Norman R. Swanson (Springer) pp. 437–461

MacKinnon, James G. (2015) 'Wild cluster bootstrap confidence intervals.' *L'Actualité Economique* 91(1-2), 11–33

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

MacKinnon, James G., and Matthew D. Webb (2017a) 'Pitfalls when estimating treatment effects with clustered data.' *The Political Methodologist* 24(2), 20–31

MacKinnon, James G., and Matthew D. Webb (2017b) 'Wild bootstrap inference for wildly different cluster sizes.' *Journal of Applied Econometrics* 32(2), 233–254

MacKinnon, James G., and Matthew D. Webb (2018a) 'Randomization inference for difference-in-differences with few treated clusters.' Working Paper 1355, Queen's University, Department of Economics

MacKinnon, James G., and Matthew D. Webb (2018b) 'Wild bootstrap randomization inference for few treated clusters.' Working Paper 1404, Queen's University, Department of Economics

Pustejovsky, James E., and Elizabeth 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

Satterthwaite, F. (1946) 'An approximate distribution of estimates of variance components.' *Biometrics Bulletin* 2(6), 110–114

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

Wu, C. F. J. (1986) 'Jackknife, bootstrap and other resampling methods in regression analysis.' *Annals of Statistics* 14(4), 1261–1295

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

Appendix: More about Alternative Procedures

As we discussed in Section 5, an alternative to the widely-used CV_1 covariance matrix is the CV_2 matrix defined in equation (28). This can be combined with a procedure to calculate the degrees of freedom for a t test. The first implementation of such a procedure is in Bell and McCaffrey (2002), the paper that proposed CV_2. The Bell-McCaffrey degrees-of-freedom parameter, $\mathrm{df}_{\mathrm{BM}}$, for testing a hypothesis about β_j is based on results in Satterthwaite (1946) and is computed as follows:

1. Calculate the $N_g \times 1$ vectors $\boldsymbol{z}_j^g$, for $g = 1, \ldots, G$, as the j^{th} columns of the $N_g \times k$ matrices $\boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

2. Create the $N \times G$ matrix $\boldsymbol{Z}$, where the g^{th} column of $\boldsymbol{Z}$ is the column vector

$$\boldsymbol{Z}_g = \boldsymbol{M}_g' \boldsymbol{M}_{gg}^{-1/2} \boldsymbol{z}_j^g. \tag{29}$$

   Here $\boldsymbol{M}_g$ is the $N_g \times N$ matrix containing the rows of $\boldsymbol{M}_{\boldsymbol{X}}$ that correspond to cluster g.

3. Form the matrix $\boldsymbol{Z}'\boldsymbol{Z}$ and find its eigenvalues, λ_1 to λ_G. Then compute

$$\mathrm{df}_{\mathrm{BM}} = \frac{\left(\sum_{g=1}^G \lambda_g\right)^2}{\sum_{g=1}^G \lambda_g^2}. \tag{30}$$

   For a pure treatment model like (1), $\mathrm{df}_{\mathrm{BM}}$ can be much smaller than $G-1$, especially when G_1 or G_0 is small.

Since it depends on the vectors $\boldsymbol{z}_j^g$, $\mathrm{df}_{\mathrm{BM}}$ is designed only for testing hypotheses about β_j. It would need to be recomputed to test a hypothesis about any other coefficient.

It is worth mentioning that CV_2 runs into difficulty when the regression includes cluster-level fixed effects, because the $\boldsymbol{M}_{gg}$ matrices are singular. For such models, it is necessary to partial out the fixed effects. That is, we need to remove the cluster means from the regressand and all the other regressors before running the regression. For a detailed treatment of fixed effects in this context, see Pustejovsky and Tipton (2017).

Imbens and Kolesár (2016) suggests a slightly modified version of the Bell-McCaffrey procedure. The latter implicitly assumes that the $\boldsymbol{\Omega}$ matrix is proportional to an identity matrix. Instead, Imbens and Kolesár (2016) assumes that $\boldsymbol{\Omega}$ is generated by a random-effects model. Thus each of the $\boldsymbol{\Omega}_g$ has diagonal elements equal to 1 and off-diagonal elements equal to ρ. The residuals are used to obtain $\hat{\rho}$, an estimate of ρ.[8] Substituting $\hat{\rho}$ into the $\boldsymbol{\Omega}_g$ blocks of $\boldsymbol{\Omega}$ yields an estimated covariance matrix $\hat{\boldsymbol{\Omega}}$. The Imbens-Kolesár degrees-of-freedom parameter, $\mathrm{df}_{\mathrm{IK}}$, is then calculated using (30), except that the eigenvalues are those of the matrix $\boldsymbol{Z}'\hat{\boldsymbol{\Omega}}\boldsymbol{Z}$ instead of the matrix $\boldsymbol{Z}'\boldsymbol{Z}$. In the pure treatment model with equal-sized clusters, $\mathrm{df}_{\mathrm{IK}}$ is numerically equal to $\mathrm{df}_{\mathrm{BM}}$.

Although using CV_2 is conceptually the easiest way to reduce the bias of the CRVE, it can be computationally challenging; see below. As an alternative, Young (2016) has recently proposed a way to reduce the bias of the diagonal elements of CV_1. As with $\mathrm{df}_{\mathrm{BM}}$, there is a different calculation for each coefficient β_j.[9]

1. Calculate the vector $\boldsymbol{z}_j$ as the j^{th} column of the matrix $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. This vector can be created by stacking the subvectors $\boldsymbol{z}_j^g$ that were formed in step 1 of the $\mathrm{df}_{\mathrm{BM}}$ calculation above. Then calculate the scalar $\Psi = \boldsymbol{z}_j'\boldsymbol{z}_j$.

2. Define the $G \times k$ matrix $\boldsymbol{D}$ with g^{th} row $\boldsymbol{D}_g = (\boldsymbol{z}_j^g)'\boldsymbol{X}_g$. The bias factor for coefficient j is then
$$\mathrm{BF}_j = \left(\frac{\Psi - \mathrm{Tr}\Big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}'\boldsymbol{D}\Big)}{\Psi}\right)\left(\frac{G(N-1)}{(G-1)(N-k)}\right). \tag{31}$$

3. Calculate the bias-reduced CV_1 standard error of β_j as
$$\sqrt{(\mathrm{CV}_1)_{jj}/\mathrm{BF}_j}\,. \tag{32}$$
The second factor in expression (31) is the first factor in equation (3), which defines CV_1. It is simply there so that these factors cancel out in expression (32).

For convenience, we refer to Young's bias-reduced estimator as $\mathrm{CV}_1^{\mathrm{br}}$, even though it does not actually yield an estimator of the covariance matrix. Expression (32) just provides a standard error for $\hat{\beta}_j$ that is larger than the square root of $(\mathrm{CV}_1)_{jj}$, because the first factor in (31) is always less than unity, since the trace is necessarily positive.

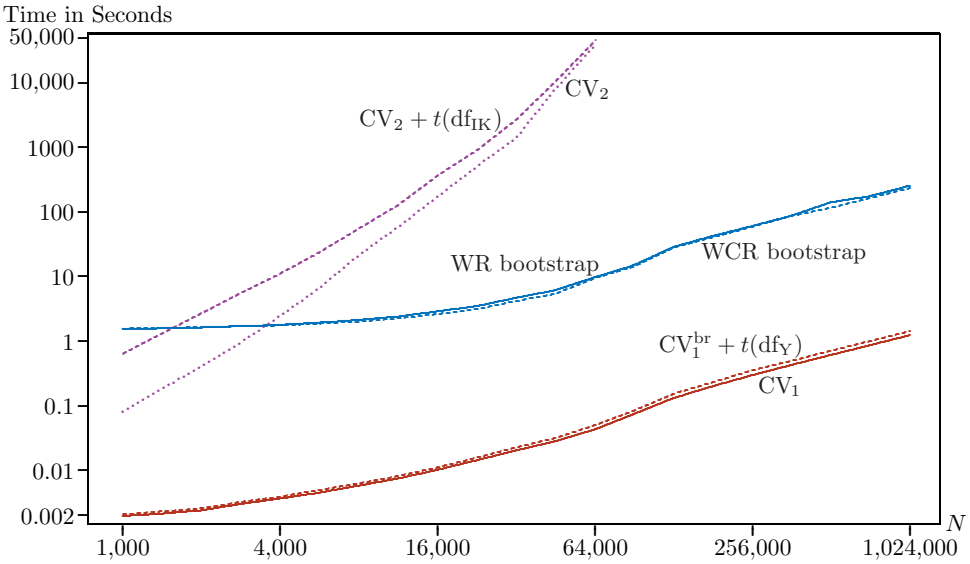
Young (2016) also proposes a degrees-of-freedom parameter for the CV_1 and $\mathrm{CV}_1^{\mathrm{br}}$ estimators. It is computed as follows:

1. Calculate $\Psi_g = (\boldsymbol{z}_j^g)'\boldsymbol{z}_j^g$ for all g. Then form the G-vector $\boldsymbol{\Psi}$ with typical element Ψ_g. Note that the scalar Ψ defined above is equal to $\sum_{g=1}^{G}\Psi_g$.

[8] This can be done in more than one way. The procedure that we use gives equal weight to all off-diagonal elements of every $\boldsymbol{\Omega}_g$ matrix. Also, we normalize the diagonal elements of the $\boldsymbol{\Omega}_g$ to be unity, which Imbens and Kolesár (2016) does not do. This normalization is valid because the ratios of the eigenvalues, which are what matters for (30), are invariant to any rescaling of the $\hat{\boldsymbol{\Omega}}$ matrix.

[9] We are grateful to Alwyn Young for providing a Stata program, which clearly explains how his bias reduction procedure works.

Figure 11: Times for several tests with $G = 20$, $k = 50$



2. Compute $B = \sum_{g=1}^{G} \Psi_g^2$ and the matrix $\boldsymbol{D}$ defined just before (31) above.

3. Young's degrees-of-freedom parameter is

$$\mathrm{df}_{\mathrm{Y}} = \frac{\Big(\Psi - \mathrm{Tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}'\boldsymbol{D})\Big)^2}{B - 2\mathrm{Tr}\Big((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{D}\circ\boldsymbol{\Psi})'\boldsymbol{D}\Big) + \mathrm{Tr}\Big((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}'\boldsymbol{D}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}'\boldsymbol{D}\Big)}, \quad (33)$$

   where "$\circ$" denotes the direct or Hadamard product.

In our experience, even though it is calculated very differently, df_{Y} tends to be quite similar to $\mathrm{df}_{\mathrm{BM}}$.

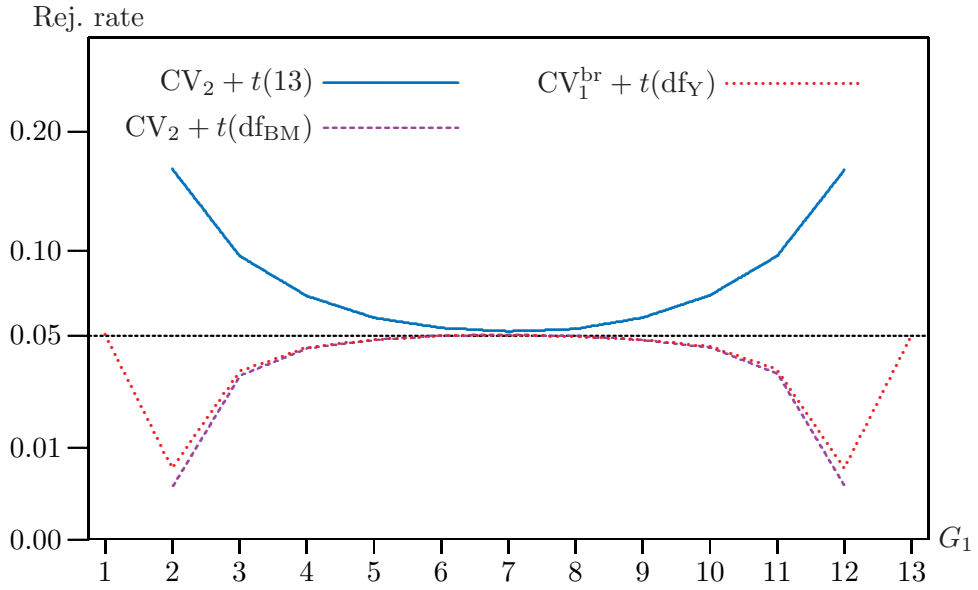
The big advantage of Young's procedures is that they are much less computationally demanding than the ones based on CV_2. Although calculating CV_2 is quite easy when N and all of the N_g are of moderate size, doing so becomes difficult, or even impossible, when any of the N_g is very large. The problem is that, for every cluster, it is necessary to form, store, invert, and take the symmetric square root of an $N_g \times N_g$ matrix. This requires a great deal of memory, and, for the models we estimate, it seems to become numerically unstable when any of the N_g is greater than about 2000.

Figure 11 plots computer time (on a Xeon server with 64 GB of memory, but using only one core) against sample size for a model with $G = 20$ and $k = 50$. The program used was written in Fortran 95 and compiled with the Intel Fortran Compiler. Both axes are logarithmic. The values of N are 1000, 1400, 2000, and so on up to 1,024,000. The two bootstrap methods use $B = 999$ bootstrap samples.

Using CV_1 standard errors is always the least expensive method. Perhaps surprisingly, the additional cost to calculate Young's bias factor and degrees of freedom is very modest. Even for $N = 1{,}024{,}000$, it only increases CPU time by about 15%.

Both bootstrap methods (WCR and WR) are much more expensive than CV_1 and Young's variant of it, but their cost increases quite slowly at first because many of the

Figure 12: Rejection frequencies for BR-DF tests, $G = 14$, $N/G = 200$, $\rho = 0.10$



calculations are reused across bootstrap samples. Even for large samples, their cost is roughly proportional to N. Thus the bootstrap methods are entirely feasible for very large sample sizes. WR is a little more expensive than WCR, especially for large sample sizes, because it needs BN random numbers instead of BG.

In contrast, simply computing CV_2 is more expensive than either bootstrap method for $N \geq 4000$, and calculating $\mathrm{df_{IK}}$ adds substantially to the cost. We ran out of memory attempting to compute CV_2 for $N > 64{,}000$, and the cost for $N = 64{,}000$ was already about 4000 times the cost of either bootstrap method. An applied researcher working on a laptop with non-optimized code would probably run out of memory much sooner or have to endure much longer computing times.

The computational cost and memory requirements of methods based on CV_2 increase rapidly with N. Therefore, even with substantial increases in the performance of standard computers over time, and assuming that problems of numerical instability can be solved, it will probably be many years before CV_2 can routinely be calculated for sample sizes as large as, say, 200,000.

It should be noted that, in computing $\mathrm{df_{IK}}$, we did not calculate or store the $N \times N$ matrix $\hat{\mathbf{\Omega}}$. Instead, we used the fact that its diagonal elements are 1, its within-cluster off-diagonal elements are $\hat{\rho}$, and all other elements are 0 so as to compute $\boldsymbol{Z}'\hat{\mathbf{\Omega}}\boldsymbol{Z}$ in the most efficient fashion that we could devise.

We perform a few additional simulation experiments for the procedures just discussed. We refer to these collectively as BR-DF procedures, since they all involve bias reduction, and most also involve calculating the number of degrees of freedom.

Figure 12 is comparable to Figures 1 and 2. It shows rejection frequencies for three tests for a pure treatment model with $G = 14$, $N_g = 200$ for all g, $\rho = 0.10$, and 13 values of G_1. Using CV_2 instead of CV_1, again in conjunction with critical values from the $t(13)$

distribution, results in somewhat better performance, although there is still overrejection. The improvement is not nearly as dramatic as the scale of the vertical axis suggests, because results for $G_1 = 1$ and $G_1 = 13$ are not included. They cannot be computed due to the singularity of some of the $\boldsymbol{M}_{gg}$ matrices.

When df_{BM} is used with CV_2, there is severe underrejection rather than severe overrejection for the more extreme values of G_1. However, this procedure works very well for $5 \leq G_1 \leq 9$. For this experiment, df_{IK} and df_{BM} are always numerically identical, so these results apply to both procedures.

When df_{Y} is used with CV_1^{br}, the rejection frequencies for $2 \leq G_1 \leq 12$ are very similar to the ones for df_{BM} with CV_2. However, the test works perfectly for $G_1 = 1$ and $G_1 = 13$. This remarkable result occurs because, for those two very special cases, $\text{df}_{\text{Y}} = 12$, and the bias factor, which is very small, is apparently just the right magnitude to correct the extreme bias in the standard error. When $G_1 = 2$ instead of 1, $\text{df}_{\text{Y}} = 1.69$, and the bias factor is very much larger.

Figure 13 is comparable to Figure 8. It shows that, like the WCR and WCU bootstraps, but unlike WR and WU, all of the BR-DF tests are quite sensitive to heteroskedasticity. In all cases, rejection frequencies increase with the amount of heteroskedasticity. Because cluster sizes are equal, df_{BM} and df_{IK} are once again numerically equal, so the figure does not show results for the latter.

Figure 14 deals with the same case as Figure 3. Cluster sizes vary, and the horizontal axis shows the ratio of the largest to the smallest value of N_g. For readability, only results for $G_1 = 3$ and $G_1 = 11$ are shown. It is evident that all the procedures, including the WR bootstrap, are sensitive to cluster sizes. Using CV_2 with df_{IK} seems to be a bit less sensitive than the other methods, but the WR bootstrap is the only procedure that performs well for equal cluster sizes. Results for CV_2 with df_{BM} are not shown. They typically lie between the ones for CV_2 with df_{IK} and the ones for CV_1^{br} with df_{Y}.

Figure 13: Effects of heteroskedasticity on rejection frequencies for BR-DF tests

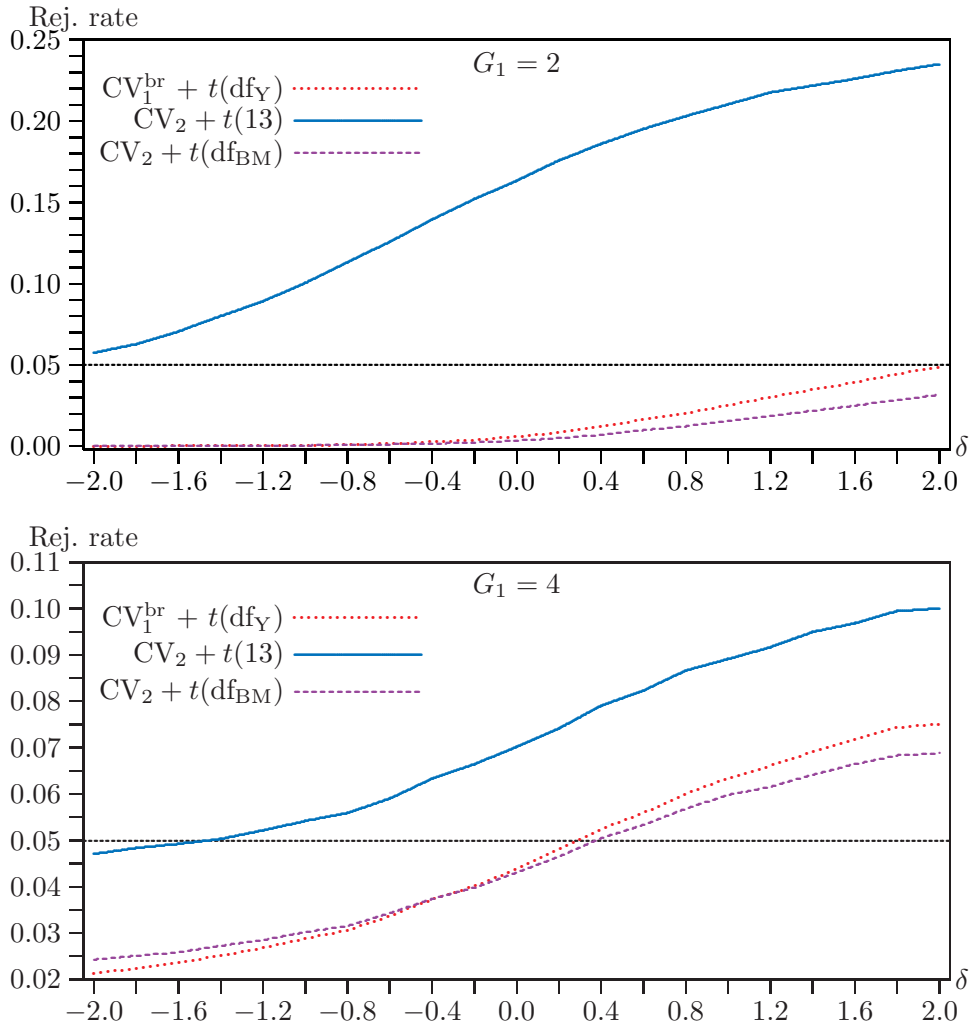


Figure 14: Effects of varying cluster sizes on rejection frequencies for several tests