CEP 16-13

The Subcluster Wild Bootstrap for Few (Treated) Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

September 2016

CARLETON ECONOMIC PAPERS

Carleton
UNIVERSITY

Department of Economics

1125 Colonel By Drive
Ottawa, Ontario, Canada
K1S 5B6

The Subcluster 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

September 2, 2016

Abstract

Inference based on cluster-robust standard errors is known to fail when the number of clusters is small, and the wild cluster bootstrap fails dramatically when the number of treated clusters is very small. We propose a family of new procedures called the subcluster wild bootstrap. In the case of pure treatment models, where all the observations in each cluster are either treated or not, the new procedures can work astonishingly well. The key requirement is that the sizes of the treated and untreated clusters should be very similar. Unfortunately, the analog 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

[*] We are grateful to participants at the 2016 University of Calgary Empirical Microeconomics Workshop. This research was supported, in part, by grants 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 of linear regression models are correlated within clusters but uncorrelated between them. Inference is then based on a cluster-robust covariance matrix, or CRVE. However, t tests based on cluster-robust standard errors tend to overreject severely when the number of clusters is small. The wild cluster bootstrap proposed by Cameron, Gelbach and Miller (2008) often leads to much more reliable inferences, but, as MacKinnon and Webb (2016a) showed, 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 usual restricted wild cluster bootstrap underreject severely, and tests based on the unrestricted wild cluster bootstrap overreject severely.

An alternative approach for the case of few treated clusters, based on randomization inference, was suggested by Conley and Taber (2011). MacKinnon and Webb (2016b) studied that procedure and proposed an improved one which uses t statistics rather than coefficient estimates and sometimes works well. However, randomization inference fails in many cases. In particular, it fails when cluster sizes vary and there are few treated clusters, it fails when there is heteroskedasticity of unknown form across clusters and there are few treated clusters, and it simply cannot be used when the number of clusters is very small.[1]

In this paper, we propose a family of new 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.[2] In many cases, this will simply be the ordinary wild bootstrap DGP (Wu, 1986; Liu, 1988), which does not cluster at all, but 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.

In Section 2, we study a simple theoretical model for which all the observations in each cluster are either treated or not and explain why t tests and wild cluster bootstrap tests fail when the number of treated clusters is small. In Section 3, we then analyze 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 a certain condition is satisfied. The condition requires that all clusters be the same size, but it allows for heteroskedasticity across clusters. We also explain why such a condition will rarely be satisfied for difference-in-difference (DiD) regressions. Finally, we extend the analysis to the subcluster wild bootstrap.

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 astonishingly well in cases where the wild cluster bootstrap performs very badly either because either the number of clusters is small or the number of treated clusters is very small, perhaps made worse by heteroskedasticity. Bootstrap tests based on the ordinary wild bootstrap often yield surprisingly reliable inferences even when there is just one treated

[1] Ferman and Pinto (2015) proposes a procedure to handle aggregate data with heteroskedasticity, and MacKinnon and Webb (2016b) suggests a method that combines randomization inference and the bootstrap which can be used when the number of clusters is small.

[2] 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).

cluster. Additional simulation experiments confirm all the principal theoretical predictions of Section 2.

In Section 5, we discuss an empirical example for which the subcluster wild bootstrap (the ordinary one in this case) yields sensible results even though there are just eight clusters. Section 6 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, in which all observations in the first G_1 clusters are treated and no observations in the remaining $G_0 = G - G_1$ clusters 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,

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

In this case, $\boldsymbol{X}_g$ 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.[3]

2.1 Why CRVE Inference Can Fail

It is shown in Section 6 of MacKinnon and Webb (2016a) that, under the null hypothesis, the cluster-robust t statistic for $\beta_2 = 0$ in equation (1) can be written as

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

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

[3] Expression (3), which is sometimes called CV_1, is not the only CRVE. Alternatives are discussed in Bell and McCaffrey (2002), Imbens and Kolesar (2016), MacKinnon (2015), Pustejovsky and Tipton (2016), and Young (2016), among others. We focus on CV_1 because it is by far the most commonly used and because it is the easiest to analyze and to compute.

the factor c for simplicity. Since it multiplies both the actual and bootstrap t statistics, it cannot affect bootstrap P values.

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 in expression (6) are evidently supposed to estimate the variances of the corresponding terms in expression (5). However, as was shown in MacKinnon and Webb (2016a), expression (6) is a very poor estimator when either G_1 or G_0 is small.[4]

For simplicity, and because this is the worst 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 is extraordinarily large, $\bar{d}$ 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 are much less variable than those for the other clusters, most of the variance of expression (8) will come from the first term. However, we can see from equation (7) that the variance of that term is incorrectly estimated to be zero.

This argument explains why tests based on the cluster-robust t statistic (4) almost always overreject extremely severely when $G_1 = 1$. The denominator of the test statistic grossly underestimates the variance of the numerator. It is shown in MacKinnon and Webb (2016a) that 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 of the disturbances.

[4] 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.

2.2 The Wild Cluster Bootstrap and Why It Can Fail

The wild cluster bootstrap was suggested by Cameron, Gelbach and Miller (2008) as a way to improve the finite-sample properties of cluster-robust t tests. In the case of the dummy variable 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 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 $G \leq 12$, it is better to use a distribution with more than two mass points; see Webb (2014).

To perform a bootstrap test, the bootstrap DGP (9) is used to generate B bootstrap samples indexed by b, each of which is then 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.[5]

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 (2016a) 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 subcluster wild bootstrap procedure that we introduce in the next section, we now explain why this happens.

The bootstrap t statistic analogous to t_2 is

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

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 $G_1 = 1$. Ignoring the factor c, the numerator of the right-hand side of equation (11) becomes

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

[5] 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.

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.

For the Rademacher distribution, the vectors of bootstrap disturbances for $g = 1$ can have just two values, namely, $\tilde{\boldsymbol{\epsilon}}_1$ and $-\tilde{\boldsymbol{\epsilon}}_1$. When $G_1 = 1$, the distribution of the bootstrap t 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 Figure 4 in MacKinnon and Webb (2016a). 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 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; see Figure 7 in MacKinnon and Webb (2016a).

3 The Subcluster Wild Bootstrap

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 not sufficiently large.

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 DGP is simply the ordinary wild bootstrap DGP; see Davidson and Flachaire (2008). Of course, standard errors are still computed using a CRVE like (3).[6]

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, even when $G_1 = 1$. This is a very important special case.

In Section 3.4, we discuss other variants of the subcluster wild bootstrap in which there are fewer subclusters than observations, so that each subcluster contains more than one observation. However, we focus on the ordinary wild bootstrap because it is the easiest one to describe and implement, and because it is almost certainly the one that should be used in practice most of the time.

[6] Using the same t statistic for the original sample and the bootstrap samples is imperative.

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. Consider once again the special case in which $G_1 = 1$. Provided N_1 is not too small, 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). Instead of there being just two possible vectors of bootstrap disturbances $\boldsymbol{\epsilon}_1^{*b}$ for cluster 1, there are now 2^{N_1} possible vectors.

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 valid inferences. However, it 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 second case is when the covariance matrix of $\boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g$ satisfies certain conditions. Of course, the first case is of no practical interest, since both cluster-robust t tests and the wild cluster bootstrap work extremely well. But the ordinary wild bootstrap (and, more generally, the subcluster wild bootstrap) can be extremely valuable in the second case.

The ordinary wild bootstrap works in the first 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 belong 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 not close to 0 or 1. Therefore, 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. Of course, this argument is not intended to be fully rigorous. Providing a formal treatment of what happens as $G \to \infty$ is tricky, because we need to specify what happens to the N_g as G increases; see Carter, Schnepel and Steigerwald (2015). Since what happens in this case is of no practical interest, we make no attempt to provide such a treatment.

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 (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. This is true even when there is an arbitrary pattern of heteroskedasticity at the cluster level.

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

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

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

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

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

Now assume that all clusters are the same size and that $\boldsymbol{\Omega}_g = \lambda_g\bar{\boldsymbol{\Omega}}$ for all g, with $\lambda_g > 0$. According to this assumption, the covariance matrices for all clusters are proportional, with factors of proportionality λ_g that may differ across clusters. This implies that $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g) = \lambda_g\boldsymbol{\iota}_g'\bar{\boldsymbol{\Omega}}\boldsymbol{\iota}_g \equiv \lambda_g\omega^2$ for all g. The key requirement here is that all the scalars $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ must be proportional to λ_g. Thus we are allowing there to be an arbitrary pattern of heteroskedasticity at the cluster level.

From (15) and the definition of ω^2, we may conclude that, in this special case, the variance of $1/c$ times 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; indeed, we saw in the previous section that $\theta_1 = 0$ when $G_1 = 1$. These two factors will almost always be different, because they depend on the numbers and sizes of the treated and untreated clusters.

Provided that the N_g are not too small, so that the OLS residuals mimic the disturbances sufficiently well, the square of the denominator of (15) must 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

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

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

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 that N is large enough for the residuals to be good estimators of the disturbances, it is approximately $\lambda_g N_g$ times 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.

Of course, the relations (19) and (22) hold only as approximations. They might be poor ones if N were small, because both the residuals and the bootstrap disturbances might provide poor approximations to the true disturbances in such a case. Therefore, both the numerators and the denominators of $\mathrm{Var}(t_2)$ and $\mathrm{Var}(t_2^*)$ might differ substantially from each other and from the expressions that appear in (19) and (22).

The argument above does not claim that t_2 and t_2^* actually follow the same distribution in finite samples. It merely suggests that they have approximately the same variance. For simplicity, we have treated the denominators of t_2 and t_2^* as constants when they are in fact random variables, but this should not be a bad approximation when 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.

More importantly, we have 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 largely 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 (2016a). However, if these factors were substantially different between the actual and bootstrap test statistics, then it would no longer be the case that $\mathrm{Var}(t_2) \cong \mathrm{Var}(t_2^*)$. This is most likely to happen when the sample size is small, because the residuals may then be poor estimators of the disturbances.

3.3 Differing Cluster Sizes and Difference in Differences

The key results (19) and (22) depend critically on the assumption that all clusters are the same size. Without that assumption, 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, 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 (which must surely be positive when there is clustering, at least on average) is proportional to N_g^2.

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 $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ increases faster than N_g, while $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ 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 $\mathrm{Var}(t_2)$ and $\mathrm{Var}(t_2^*)$. 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 $\mathrm{Var}(t_2^*) > \mathrm{Var}(t_2)$. 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 $\mathrm{Var}(t_2) \cong \mathrm{Var}(t_2^*) \cong 1$.

It might seem tempting to create a sample in which every cluster is the same size 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. The problem is that, when we take averages over different numbers of observations, we implicitly create intra-cluster covariance matrices that depend on those numbers. As a result, the condition that all the covariance matrices, and hence all the scalars $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$, must be identical up to a factor λ_g will be violated.

The result that $\mathrm{Var}(t_2) \cong \mathrm{Var}(t_2^*)$ 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. This means that expression (5) for the numerator of the t statistic has to be replaced by

$$(1-\bar{d})\sum_{g=1}^{G_1}\boldsymbol{d}_g'\boldsymbol{\epsilon}_g + \bar{d}\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g - \boldsymbol{d}_g)'\boldsymbol{\epsilon}_g + \bar{d}\sum_{g=G_1+1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g. \tag{23}$$

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 (23) 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 when the number of treated observations in each treated cluster is small relative to the number of untreated observations, and/or relative to the number of observations in each untreated cluster. They also suggest that it is likely to overreject when the number of treated observations in each treated cluster is relatively large. The former situation is likely to be more common than the latter, 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 must be treated. In most cases, we would not expect both these conditions to be satisfied.

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 $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*)$ to $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g)$ 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 location 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

$$y_{itg} = \beta_1 + \beta_2 d_{itg} + \epsilon_{itg}, \tag{24}$$

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:

$$y^{*b}_{itg} = \tilde{\beta}_1 + \tilde{\epsilon}_{itg} v^{*b}_{tg}. \tag{25}$$

This is a variant of the wild cluster bootstrap, since the auxiliary random variable v^{*b}_{tg} 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^{*b}_{g}.

For the DGP (25), 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 $\mathrm{Var}(\boldsymbol{\iota}_g' \boldsymbol{\epsilon}^*_g)$ to provide a better approximation to $\mathrm{Var}(\boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g)$ than would be the case for the ordinary wild bootstrap. In consequence, we would expect $\mathrm{Var}(t^*_2)$ to be closer to $\mathrm{Var}(t_2)$ and bootstrap tests to perform better when cluster sizes vary.

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 $1/S$.[7] 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 statistic will be correlated with the actual t statistic. With the unrestricted subcluster wild bootstrap, in the same cases, the variance of the bootstrap t 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 and Section 6 of MacKinnon and Webb (2016a). 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.

We tentatively conclude 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 will probably only be beneficial if the correlations within subclusters are a lot higher than the correlations between them.

[7] Moreover, with unbalanced subclusters, this fraction would be further reduced.

4 Simulation Experiments

In this section, we report some of the results from 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, uncorrelated across clusters, and equicorrelated within clusters with correlation coefficient ρ. The bootstrap methods use $B = 399$ bootstrap samples.[8] There are always 400,000 replications. Using such a large number is essential in order to distinguish between experimental noise and small but systematic failures of exactness for various bootstrap tests.

Figure 1: Rejection frequencies for existing tests, $G = 14$, $N/G = 200$

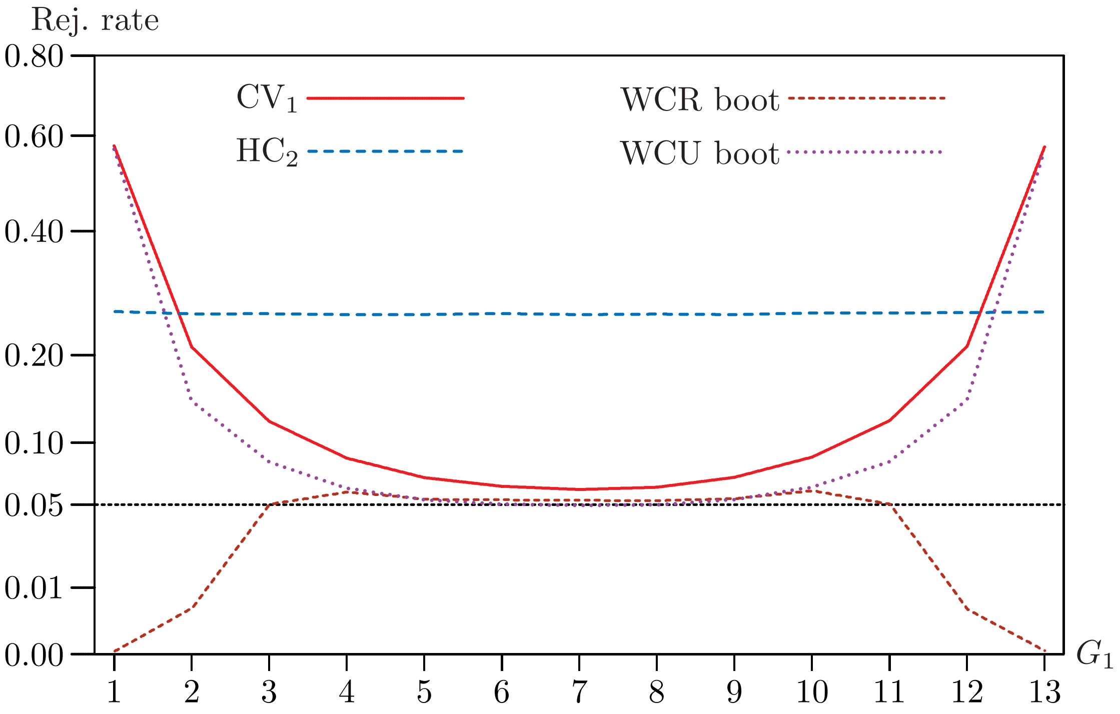


Figure 1 shows rejection frequencies at the .05 level for four conventional tests with $G = 14$, $N/G = 200$, and $\rho = 0.10$. The horizontal axis shows the number of treated clusters, G_1, which varies from 1 to 13. The vertical axis has been subjected to a square root transformation in order to present both large and small rejection frequencies in the same figure.

[8] In empirical analysis, it is desirable to use a larger value for B, but 399 seems to work well in simulation experiments, where any randomness in the bootstrap P values tends to average out across replications.

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 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.

As the analysis of Subsection 2.1 suggests, using t statistics based on the CV_1 covariance matrix (3), combined with the $t(13)$ distribution, leads to extremely 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 $G/2$.

The two wild cluster bootstrap methods perform exactly as the analysis of MacKinnon and Webb (2016a), reviewed in Subsection 2.2, suggests. 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 $G_1 = 11$,[9] 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 $6 \leq G_1 \leq 8$.

Figure 1 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 $6 \leq G_1 \leq 34$.

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 are close, they both seem to be at least fairly reliable. Of course, the P values being similar certainly does not guarantee that they are entirely reliable; consider the cases of $G_1 = 4$ and $G_1 = 10$ in Figure 1.

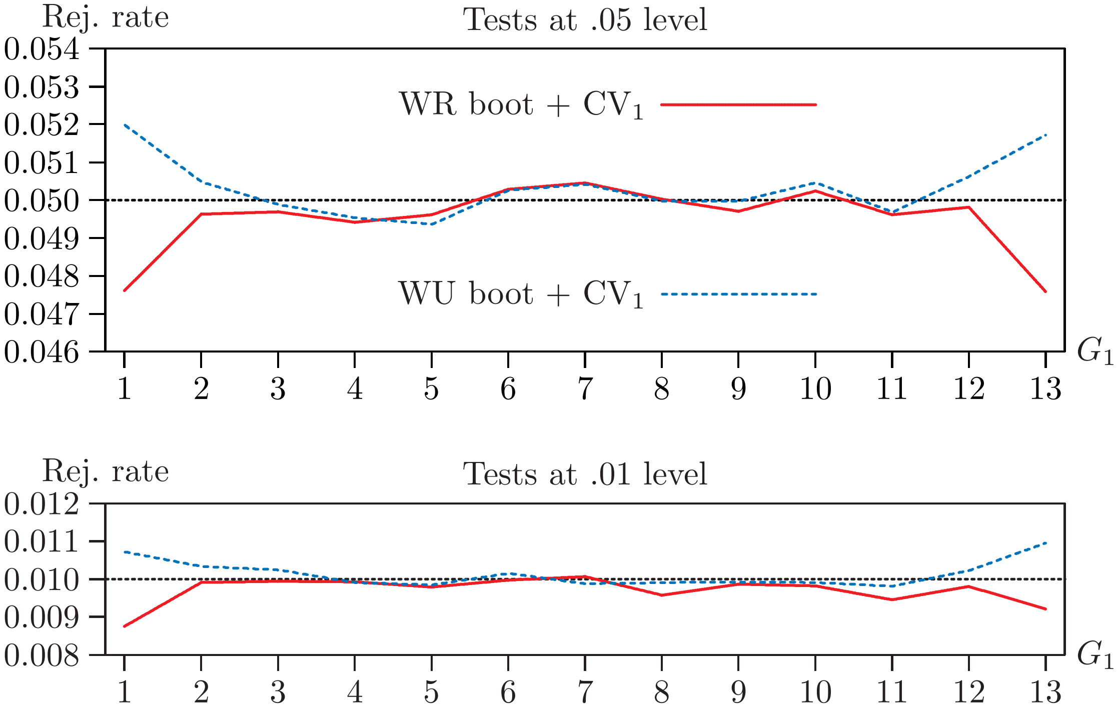
Figure 2 shows rejection frequencies for ordinary wild bootstrap tests at the .05 and .01 levels for the same thirteen experiments. These tests combine the wild bootstrap, either restricted (WR) or unrestricted (WU), with the CV_1 covariance matrix. They perform extraordinarily well. The only deviations between rejection frequencies and nominal levels that are clearly not due to experimental noise are for $G_1 = 1$ and $G_1 = 13$. These are cases where wild cluster bootstrap tests fail dramatically; see Figure 1. The very minor deviations visible in Figure 2 are extraordinarily trivial by comparison.

Figure 3 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 of subclusters S, which varies from 1 (the wild cluster bootstrap) to 256 (the wild bootstrap) by factors of 2. The vertical axis shows rejection frequencies for $G_1 = 1$ and $G_1 = 2$ for restricted and unrestricted bootstrap tests. Note that, as in Figure 1, the vertical axis has been subjected to a square root transformation.

The results in Figure 3 are dramatic. 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 monotonically. For $S = 256$ (the ordinary wild bootstrap), they perform very well for $G_1 = 1$ and almost perfectly for $G_1 = 2$. In additional

[9] This is a coincidence that would not have occurred if G had been larger or smaller.

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



experiments, not reported, we varied the value of ρ. 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 true for all values of S, but it is particularly true for small values.

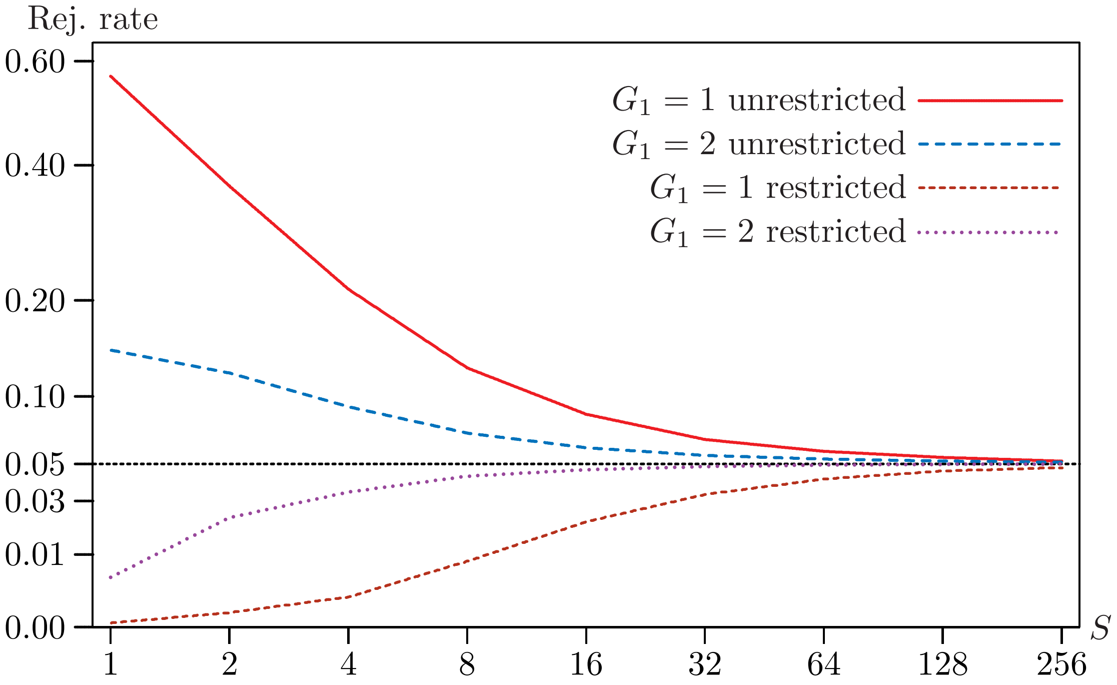
The results in Figure 3, together with additional ones for larger values of G_1 and other values of ρ, suggest that it is better to use the ordinary wild bootstrap than to subcluster at any level. We suspect that this will generally be the case unless correlations within subclusters are substantially larger than correlations across them.

The next set of experiments is designed to investigate the effects of the number of clusters and their size. Figure 4 reports the results of several experiments for $G_1 = 1$, which is always the worst case. There are three panels, for $N/G = 20$, $N/G = 100$, and $N/G = 500$. The vertical axis shows rejection frequencies at the .05 level. The horizontal axis shows G, which varies from 3 to 15.[10] Remarkably, both variants of the ordinary wild bootstrap perform almost perfectly when $G = 3$. 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 only slightly as N/G increases, but the latter improves very substantially.

We showed in Subsection 3.2 that the ordinary wild bootstrap is approximately invariant

[10] $G = 3$ is the smallest value for which it is possible to compute the CV$_1$ standard error of $\hat{\beta}_2$. When $G = 1$, all observations are either treated or not treated, and so β_2 cannot be identified. When $G = 2$ and $G_1 = 1$, both terms in expression (6) equal 0 by an extension of the argument that led to equation (7), causing the CV$_1$ standard error to be zero.

Figure 3: Rejection frequencies as level of subclustering changes, $G = 14$, $N/G = 256$



to heteroskedasticity at the cluster level. To investigate this 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:

$$\sigma_g = \exp\left(\frac{\delta(g-1)}{G-1}\right). \tag{26}$$

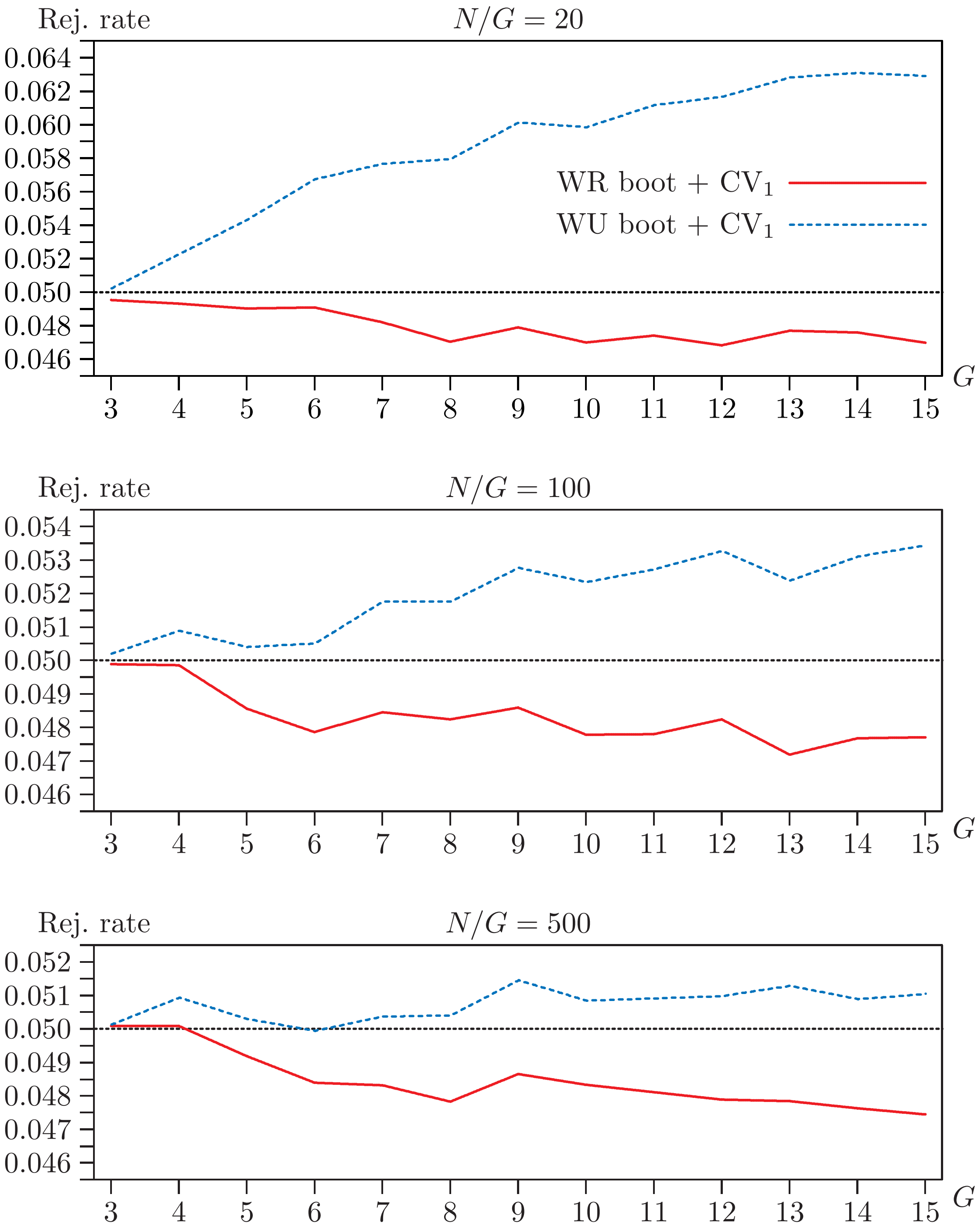
According to equation (26), σ_g equals 1 when $\delta = 0$ and is increasing in δ. In the experiments, δ varies between -2 and 2, so that σ_g varies between 0.135 and 7.39. The treated clusters are always the ones with the highest indices. Therefore, $\exp(\delta)$ can be thought of as the ratio of the 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 5 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 almost perfectly. It is extremely difficult to distinguish their rejection frequencies from each other or from the horizontal line at .05. In contrast, the performance of the two wild cluster bootstrap tests is very sensitive to δ. Remarkably, when $G_1 = 4$ and $\delta > 0.4$, 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 on inference

Figure 4: Effect of G on rejection frequencies for two tests when $G_1 = 1$

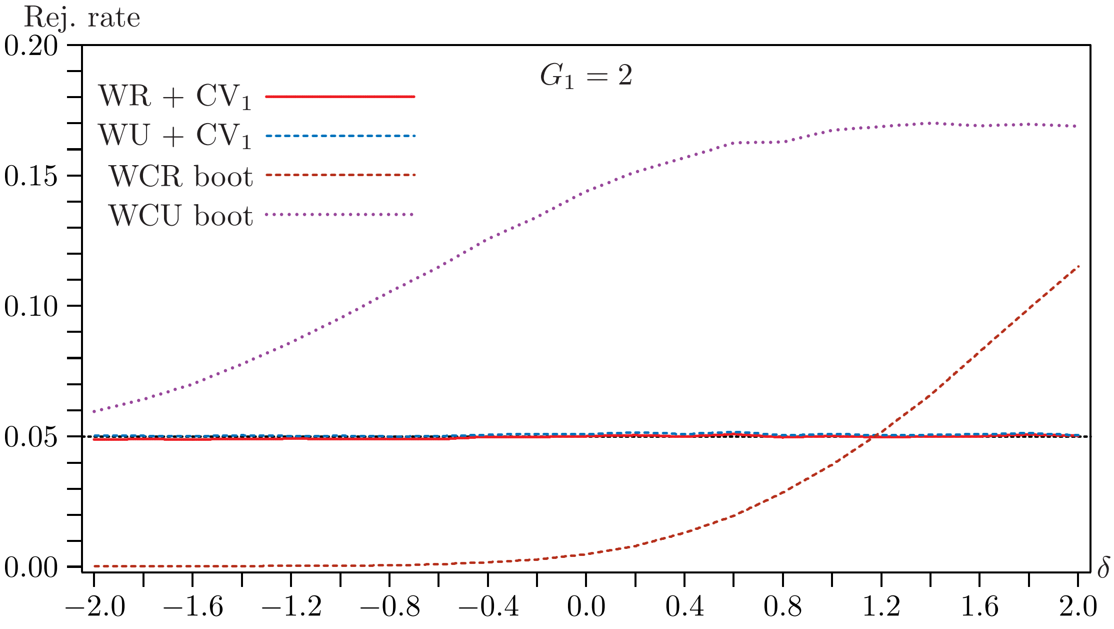


$N/G = 20$

$N/G = 100$

$N/G = 500$

Figure 5: Effects of heteroskedasticity on rejection frequencies



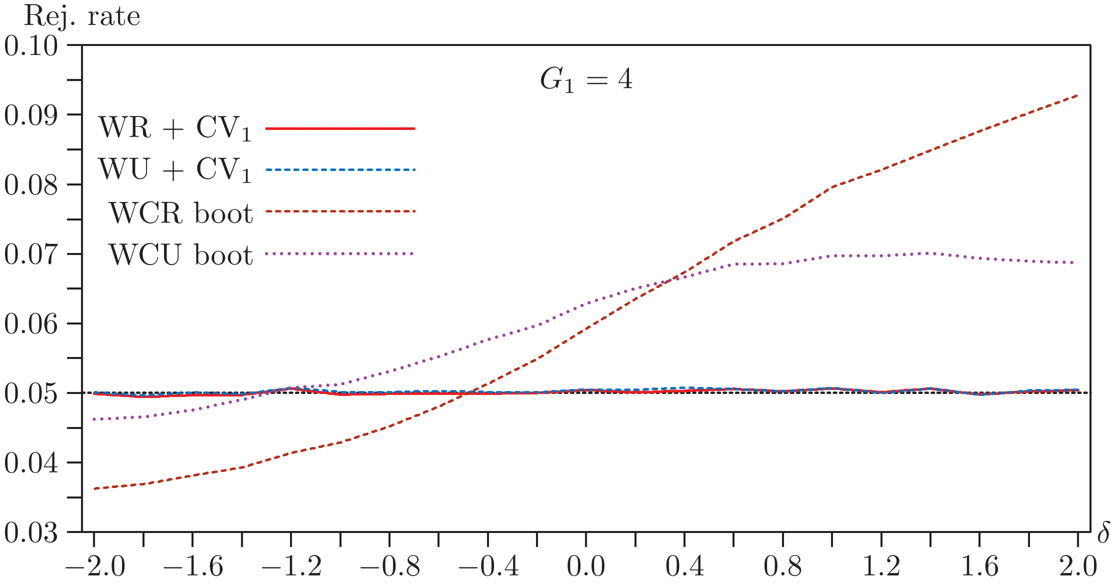
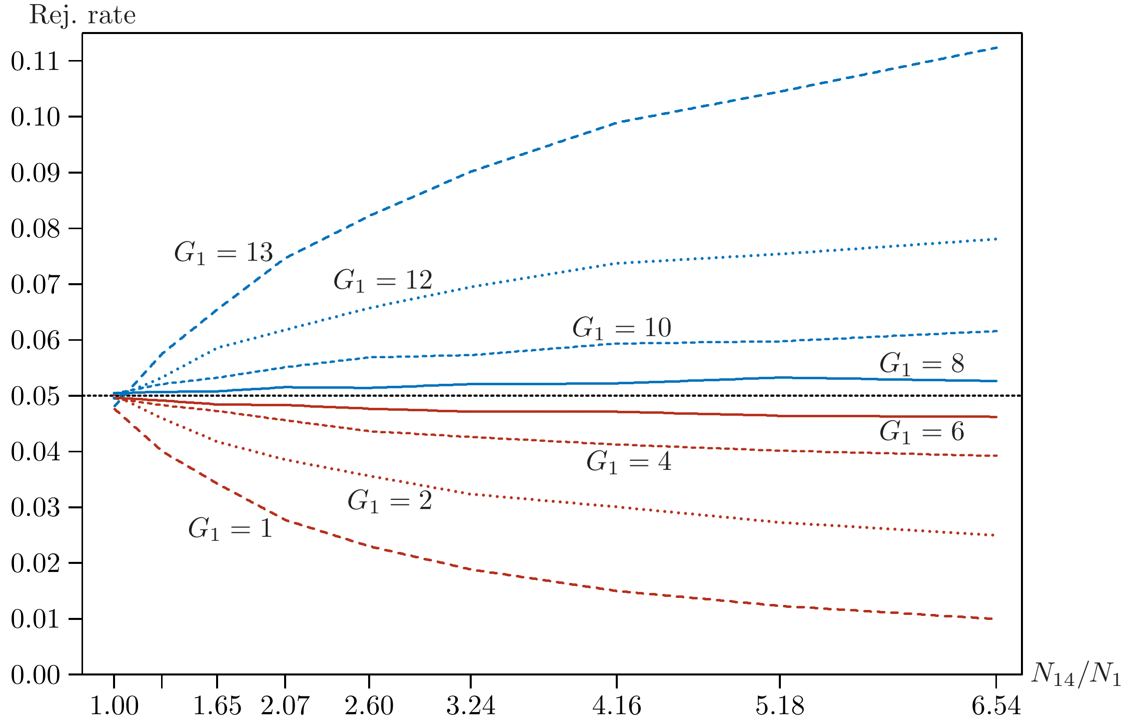


Figure 6: Effect of varying cluster sizes on rejection frequencies for WR + CV_1



using the wild cluster bootstrap. In particular, all of the simulations in MacKinnon and Webb (2016a) assume that the disturbances are homoskedastic. Figure 5 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.[11]

The next set of experiments deals with varying cluster sizes. The theory of Subsection 3.3 suggests that the ordinary wild bootstrap may not perform well when either G_1 or G_0 is small and cluster sizes vary. In these experiments, $G = 14$, $\rho = 0.10$, and the average value of N/G is 200. The actual cluster sizes depend on a 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; with clusters sorted from smallest to largest, $N_1 = 67$ and $N_{14} = 438$. For details, see Section A.6 of the appendix to MacKinnon and Webb (2016a).

Figure 6 plots rejection frequencies at the .05 level for the restricted (WR) variant of the wild bootstrap when clusters are treated from smallest to largest. Results for the WU variant, not shown in the figure, are very similar. 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 underrejection

[11] Similar experiments in MacKinnon and Webb (2016b) show that randomization inference procedures also perform poorly with cluster-specific heteroskedasticity and few treated clusters.

for $G_1 < 7$ and overrejection for $G_1 > 7$, because treating, say, the 12 smallest clusters is equivalent to treating the 2 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. However, it underrejects fairly severely for $G_1 = 1$, and it overrejects fairly severely for $G_1 = 13$ when they vary by as little as a factor of two. 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.[12]

The situation depicted in Figure 6 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, we would expect there 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.

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 $G = 12$, $N = 2400$, $\rho = 0.10$, and $\psi = 0.05, 0.10, \ldots, 1.00$.

Figure 7 reports rejection frequencies at the .05 level as functions of ψ for four tests. The top panel shows results for $G_1 = 1$, and the bottom panel shows results for $G_1 = 2$. As expected, the two wild cluster bootstrap tests perform very badly when $G_1 = 1$. The restricted variant (WCR) almost never rejects, and the unrestricted one (WCU) rejects more than half the time and is therefore not shown on the figure. 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.

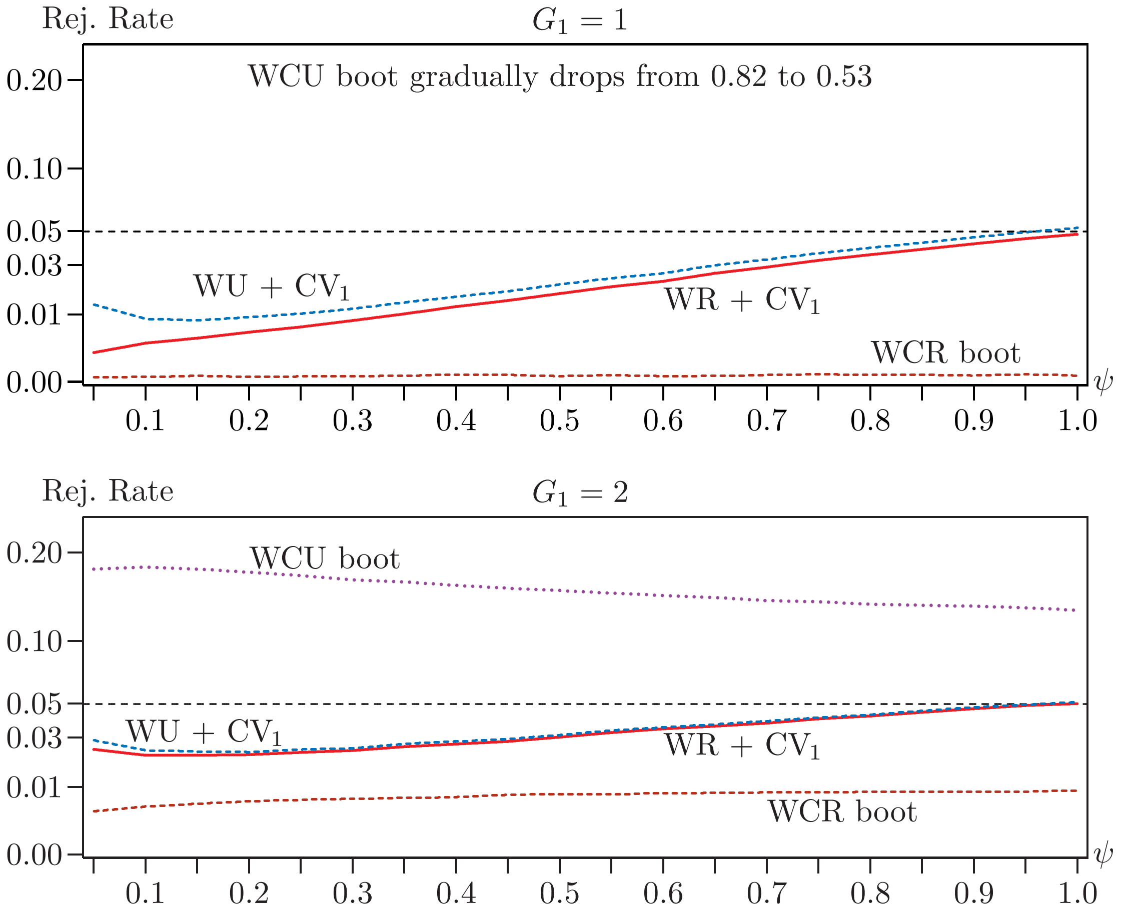
All four tests perform much better when $G_1 = 2$, but WCR still underrejects severely, and WCU still overrejects severely. The two wild bootstrap tests still underreject, but not nearly as much as when $G_1 = 1$. They are fairly reliable for $\psi \geq 0.75$, always rejecting between 4% and 5% of the 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. Therefore, in order for there to be any reason to use a CRVE, we would need to use a 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

[12] Randomization inference also performs poorly when cluster sizes vary and G_1 is small; see MacKinnon and Webb (2016b).

Figure 7: Effects of fraction of treated observations in treated clusters



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 7 are probably much worse than we would see in practice with 12 clusters of 200 observations each. Of course, with clusters that were substantially larger or variable in size, we might well see even worse results.

5 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. Additionally, we restrict attention to the change in math scores between 1995 and 1996, as this coefficient is reported to be quite statistically significant; see Table 5, column 4 of the original paper. The experimental design allows for a very simple identification strategy:

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

Table 1: Effects of Teacher Training on Math Score Difference

	full sample	drop 48	drop 40
coef.	-0.866	-0.778	-0.903
std. error	0.195	0.206	0.205
t stat (P value)	-4.45 (0.003)	-3.78 (0.009)	-4.41 (0.005)
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
N	255	207	215
G	8	7	7
G_1	3	2	2

Notes: The outcome variable is the difference between 1995 and 1996 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 one treated student, the effective values of G and G_1 are probably smaller by 1 than the reported values.

Here diff_{is} is the difference in math scores for student i in school s between 1995 and 1996, and treated_{is} is an indicator for whether a student was in a class taught by a treated teacher. The standard errors are clustered by school.

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 there is one school with 53 untreated students and just one treated student.

Initially, we repeat the analysis of Angrist and Lavy (2001). We also calculate 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 by Webb (2014). The ordinary wild bootstrap DGPs use the Rademacher distribution.

Our results for the full sample are found in column 1 of Table 1. Our coefficient estimate is identical to the one reported in the paper, but our standard error estimate differs slightly. The CRVE P value, which is based on the $t(7)$ distribution, suggests that the treatment has a negative impact which is statistically significant at well below the 1% level. However, all four bootstrap procedures agree that it is significant only at the 5% level.

It may seem surprising that all four bootstrap procedures agree in this case. The reason is that, because G is so small, the two wild cluster bootstrap procedures actually work quite well despite G_1 being small. The equivalent of Figure 1 for $G = 7$ shows very good performance by the WCR bootstrap when $G_1 = 2$. 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 also not surprising that the ordinary wild bootstrap works well.

In order to make inference more difficult, we 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 one student. 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 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 are so far apart that they convey little information. The latter also do not yield unambiguous results, but they are very much closer, and for column 2 they yield the same inferences. 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 of a treatment effect.

6 Conclusion and Recommendations

Although the wild cluster bootstrap works well much of the time, MacKinnon and Webb (2016a) have shown that it often fails when the number of treated clusters is small, whether or not the total number of clusters is small. What very often happens is that the restricted wild cluster bootstrap P value is quite large, and the unrestricted wild cluster bootstrap P value is very much smaller. In such cases, neither of them can be trusted.

We have proposed a family of new bootstrap procedures, called the subcluster wild bootstrap, that often works much better than the wild cluster bootstrap when there are few treated clusters. In principle, the subcluster wild bootstrap can be implemented in a variety of ways. In most cases, however, it seems that the best approach is simply to combine the ordinary wild bootstrap with cluster-robust standard errors.

We showed in Section 3.2 that the ordinary wild bootstrap can be expected to work very well, even with as few as one treated cluster, 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. Finally, the number of observations per cluster must be sufficiently large. Simulation results in Section 4 confirm these predictions.

The conditions in the previous paragraph are quite stringent. Happily, the subcluster wild bootstrap often works reasonably well even when they are violated, provided the violations are not too extreme. With just a few treated clusters, it is very likely to underreject (overreject) when the treated clusters are smaller (larger) than average. It also tends to underreject for difference-in-difference 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.

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, but they are almost certainly not severely misleading. However, WCR and WCU will very often yield different inferences when the number of treated clusters is very 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 5. 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 Section 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 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

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

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

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

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2015) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' Technical Report, University of California, Santa Barbara

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

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

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

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

Imbens, Guido W., and Michal Kolesar (2016) 'Robust standard errors in small samples: Some practical advice.' *Review of Economics and Statistics* 98, to appear

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

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 (2016a) 'Wild bootstrap inference for wildly different cluster sizes.' *Journal of Applied Econometrics* 31, to appear

MacKinnon, James G., and Matthew D. Webb (2016b) 'Randomization inference for difference-in-differences with few treated clusters.' Working Paper 1355, Queen's University, Department of Economics

Pustejovsky, James E., and Elizabeth Tipton (2016) 'Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models.' Technical Report, University of Texas at Austin

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

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