The Econometrics Journal

Econometrics Journal (2018), volume **21**, pp. 114–135.
doi: 10.1111/ectj.12107

The wild bootstrap for few (treated) clusters

JAMES G. MACKINNON[†] AND MATTHEW D. WEBB[‡]

[†]*Department of Economics, Queen's University, Kingston, ON, K7L 3N6, Canada*.
E-mail: jgm@econ.queensu.ca

[‡]*Department of Economics, Carleton University, Ottawa, ON, K1S 5B6, Canada*.
E-mail: matt.webb@carleton.ca

First version received: June 2017; final version accepted: October 2017

Summary 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: *Clustered data, Cluster-robust variance estimator, Difference-in-differences, Grouped data, Robust inference, Subclustering, Treatment model, Wild bootstrap, Wild cluster bootstrap.*

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 over-reject severely when the number of clusters is small. The number of clusters required to avoid serious over-rejection 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 et al. (2008) often leads to much more reliable inferences, but, as MacKinnon and Webb (2017b) show, this procedure can also fail dramatically. When the regressor of interest is a dummy variable that is non-zero for only a few clusters, tests based on the restricted WCB can under-reject severely, and tests based on the unrestricted WCB can over-reject 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

[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).

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. Then, in Section 3, we 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.

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 4 and the associated parts of the online Appendix, we briefly discuss several alternative methods that involve using 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 by Conley and Taber (2011), which is based on randomization inference. MacKinnon and Webb (2018a) studied that procedure and proposed an improved one that 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 heteroscedasticity of unknown form across clusters, and it cannot be used when the number of clusters is very small.[2] Therefore, we do not consider randomization inference in this paper.

In Section 5, 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 Section 6, we conclude and provide recommendations for applied work.

In online Appendix A, 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 because the number of treated clusters is very small, occasionally made worse by heteroscedasticity. 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.

In online Appendix B, we discuss some alternative procedures for cluster-robust inference that are not based on the bootstrap. In online Appendix C, we report the results of some additional simulation experiments, which investigate the performance of those procedures when there are few treated clusters.

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

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 gth 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

$$y_{ig} = \beta_1 + \beta_2 d_{ig} + \epsilon_{ig}, \tag{2.1}$$

where y_{ig} denotes the ith 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 is usual in the literature on cluster-robust inference, we assume that

$$E[\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_g'] = \boldsymbol{\Omega}_g \quad \text{and} \quad E[\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}_h'] = \mathbf{O} \quad \text{for } g \neq h, \tag{2.2}$$

where $\boldsymbol{\epsilon}_g$ are vectors with typical elements ϵ_{ig}, and $\boldsymbol{\Omega}_g$ are $N_g \times N_g$ positive definite covariance matrices. The model (2.1) is estimated by ordinary least-squares (OLS), and standard errors are based on the CRVE:

$$\frac{G(N-1)}{(G-1)(N-k)} (\boldsymbol{X}'\boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' \boldsymbol{X}_g \Big) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.3}$$

In this case, where $k = 2$, $\boldsymbol{X}_g$ has a 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 (2.3) 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 et al. (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 4 and online Appendix B.

2.1. The reason 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 (2.1) can be written under the null hypothesis as

$$t_2 = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{\big(\sum_{g=1}^{G} (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g) \big)^{1/2}}, \tag{2.4}$$

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 ones, $\boldsymbol{d}_g$ 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 $((G-1)(N-2))/(G(N-1))$, the inverse of the degrees-of-freedom correction in (2.3). In what follows, we omit the factor c, as it does not affect any of the arguments.

With c omitted, the numerator of the t-statistic (2.4) can be written as

$$(1 - \bar{d}) \sum_{g=1}^{G_1} \boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g. \tag{2.5}$$

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

$$(1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2+\bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2. \tag{2.6}$$

The first and second terms here are evidently supposed to estimate the variances of the corresponding terms in (2.5). However, MacKinnon and Webb (2017b) showed that (2.6) 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, (2.6) reduces to

$$(1-\bar{d})^2(\boldsymbol{\iota}_1'\hat{\boldsymbol{\epsilon}}_1)^2+\bar{d}^2\sum_{g=2}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2=\bar{d}^2\sum_{g=2}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2, \tag{2.7}$$

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 (2.7) that (2.6) provides a dreadful estimator of the variance of

$$(1-\bar{d})\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1-\bar{d}\sum_{g=2}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g, \tag{2.8}$$

which is what (2.5) 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 (2.8) will come from the first term. However, from (2.7) it is evident that the variance of that term is incorrectly estimated to be zero.

Note that, for the pure treatment model (2.1), 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 (2.4) using conventional critical values almost always over-reject very severely when $G_1 = 1$ or $G_0 = 1$. The denominator of (2.4) grossly underestimates the variance of the numerator. As MacKinnon and Webb (2017b) show, this underestimation, and the resulting over-rejection, 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 (2.1), the restricted wild cluster bootstrap DGP for bootstrap sample b is

$$y_{ig}^{*b}=\tilde{\beta}_1+\tilde{\epsilon}_{ig}v_g^{*b}, \tag{2.9}$$

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 (2.9) is used to compute a bootstrap test statistic t_2^{*b} (see below). The symmetric bootstrap P-value is then calculated as

$$\hat{p}^* = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\big(|t_2^{*b}| > |t_2|\big), \tag{2.10}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. It would, of course, be valid to use an equal-tail P-value instead of (2.10), 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 (i.e. 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 under-reject 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

$$t_2^{*b} = \frac{c(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}_b^*}{\big(\sum_{g=1}^{G}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g^{*b}\hat{\boldsymbol{\epsilon}}_g^{*b\prime}(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)\big)^{1/2}}, \tag{2.11}$$

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 (2.4).

Now consider the extreme case in which $G_1 = 1$. The numerator of the right-hand side of (2.11) becomes

$$(1 - \bar{d})\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1^{*b} - \bar{d}\sum_{g=2}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^{*b}; \tag{2.12}$$

this is the bootstrap analogue of (2.8). Because $\bar{d} = N_1/N$, the first term in (2.12) 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 (2.12) 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: $\tilde{\boldsymbol{\epsilon}}_1$ 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 neighbourhood of t_2 and the other half in the neighbourhood 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 thus leads to severe under-rejection. However, the problem rapidly becomes less severe as G_1 increases.

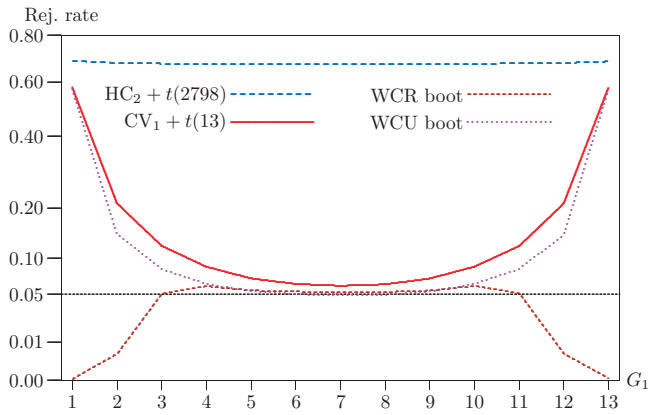

Figure 1. Rejection frequencies for several tests, $G = 14$, $N/G = 200$, $\rho = 0.1$. [Colour figure can be viewed at wileyonlinelibrary.com]

It might seem that this problem could be solved by using unrestricted instead of restricted residuals in the bootstrap DGP (2.9). However, this creates a new problem, which is just as severe. When unrestricted residuals are used with $G_1 = 1$, the first term in (2.12) always equals zero, just like the first term on the left-hand side of (2.7), 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 over-rejects very severely. Again, the problem rapidly becomes less severe as G_1 increases.

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 0.05 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 online Appendix A.

Simply using t-statistics based on heteroscedasticity-robust standard errors – specifically, the HC$_2$ variant proposed in MacKinnon and White (1985) – combined with the $t(2798)$ distribution results in very severe over-rejection for all values of G_1. This over-rejection 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 (2.3), combined with the $t(13)$ distribution, leads to severe over-rejection when $G_1 = 1$ and $G_1 = 13$, but the over-rejection is much less severe for values of G_1 that are not too far from $G/2$. This is exactly what the arguments of Section 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 $G_1 = 13$, under-rejects 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 over-rejects modestly for other values of G_1. In contrast, the unrestricted wild cluster bootstrap (WCU) over-rejects very severely for $G_1 = 1$ and $G_1 = 13$, but it improves rapidly as G_1 becomes less extreme and it performs extremely well for $6 \leq G_1 \leq 8$.

A very different bootstrap procedure is the pairs cluster bootstrap, in which the bootstrap samples are obtained by resampling the matrices $[\boldsymbol{y}_g \;\; \boldsymbol{X}_g]$ 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 over-reject 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 $\boldsymbol{\epsilon}_g$. However, 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 to 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 such as (2.3); it is imperative to use the same form of t-statistic for the original sample and the bootstrap samples.

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 (2.1) 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.

Later, 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. First, however, in the next three subsections, we focus on the ordinary wild bootstrap, which is the easiest one to describe and implement; in the case of cross-sectional data, it seems to be the method that should be used in practice most of the time.

3.1. The ordinary wild bootstrap

The restricted wild bootstrap DGP analogous to equation (2.9) is

$$y_{ig}^{*b} = \tilde{\beta}_1 + \tilde{\epsilon}_{ig} v_{ig}^{*b}. \tag{3.1}$$

The only difference between (2.9) and (3.1) 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 (3.1) 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 (2.12). 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 (3.1) 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 (2.4) and its bootstrap analogue (2.11). Under the wild bootstrap DGP (3.1), the numerators of (2.4) and (2.11) do not have the same distributions. However, 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 (2.4) using the empirical distribution of B realizations of (2.11) 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 et al. (2018), which was written after this paper.

The second case, which we discuss in detail in Section 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 heteroscedasticity across them.

The wild bootstrap DGP (3.1) imposes the null hypothesis. We could instead use the unrestricted wild bootstrap DGP

$$y_{ig}^{*b} = \hat{\beta}_1 + \hat{\beta}_2 d_{ig} + \hat{\epsilon}_{ig} v_{ig}^{*b}, \tag{3.2}$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are unrestricted OLS estimates, and $\hat{\epsilon}_{ig}$ 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 $\beta_2 = 0$. The use of (3.2) instead of (3.1) 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 online Appendix A, 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 is not 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 heteroscedasticity at the cluster level.

Whenever we make approximations in this section, they are not asymptotic approximations 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 $\hat{\boldsymbol{\beta}} \equiv [\hat{\beta}_1 \;\; \hat{\beta}_2]'$ in the model (2.1) is not consistent, at least not without very unrealistic

assumptions about the intra-cluster correlations; see Carter et al. (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 N_g, is large and the amount of intra-cluster correlation is not too large, we can reasonably expect this random term to be very small. The experiments in online Appendix A 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 (2.5) and (2.6), the actual t-statistic under the null hypothesis is

$$t_2 = \frac{(1-\bar{d})\sum_{g=1}^{G_1} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g + \bar{d}\sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g}{\left((1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2\right)^{1/2}}. \tag{3.3}$$

Now, consider the bootstrap t-statistic based on the ordinary wild bootstrap DGP (3.1). Omitting the b superscripts for clarity, it is

$$t_2^* = \frac{(1-\bar{d})\sum_{g=1}^{G_1} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^* + \bar{d}\sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g^*}{\left((1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g^*)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g^*)^2\right)^{1/2}}. \tag{3.4}$$

The bootstrap t-statistic (3.4) evidently has the same form as the t-statistic (3.3), 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.

ASSUMPTION 3.1. *G, G_1, and N are fixed, with $N_g = N/G$ for $g = 1, \ldots, G$.*

ASSUMPTION 3.2. *$\boldsymbol{\Omega}_g = \lambda_g\bar{\boldsymbol{\Omega}}$ for all g, for some positive definite matrix $\bar{\boldsymbol{\Omega}}$, with $\lambda_1 = 1$ and $\lambda_g > 0$.*

ASSUMPTION 3.3. *The average intra-cluster correlation, say ρ, is small if G_1 is small.*

Assumption 3.1, 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 3.3 will be discussed below. Assumption 3.2 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 $\mathrm{Var}(\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g) = \lambda_g\boldsymbol{\iota}_g'\bar{\boldsymbol{\Omega}}\boldsymbol{\iota}_g \equiv \lambda_g\omega^2$ for all g. Thus, we are allowing there to be an arbitrary pattern of cross-cluster heteroscedasticity, but the same pattern of within-cluster correlation and heteroscedasticity for all clusters. The condition that $\lambda_1 = 1$ is just an arbitrary normalization.

From (3.3) and the definition of ω^2, we can conclude that, in this special case, the variance of the numerator of t_2 is simply

$$(1-\bar{d})^2\sum_{g=1}^{G_1}\lambda_g\omega^2 + \bar{d}^2\sum_{g=G_1+1}^{G}\lambda_g\omega^2. \tag{3.5}$$

The variance of t_2 itself depends on how well the denominator of (3.3) estimates (3.5). 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 (3.5), 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 1, 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 3.3 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 (3.3) will be approximately equal to

$$(1-\bar{d})^2\theta_1\sum_{g=1}^{G_1}\lambda_g\omega^2 + \bar{d}^2\theta_0\sum_{g=G_1+1}^{G}\lambda_g\omega^2. \tag{3.6}$$

Thus, from (3.5) and (3.6), we conclude that

$$\mathrm{Var}(t_2) \cong \frac{(1-\bar{d})^2\sum_{g=1}^{G_1}\lambda_g + \bar{d}^2\sum_{g=G_1+1}^{G}\lambda_g}{(1-\bar{d})^2\theta_1\sum_{g=1}^{G_1}\lambda_g + \bar{d}^2\theta_0\sum_{g=G_1+1}^{G}\lambda_g}. \tag{3.7}$$

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

$$(1-\bar{d})^2\sum_{g=1}^{G_1}\lambda_g N_g\sigma^2 + \bar{d}^2\sum_{g=G_1+1}^{G}\lambda_g N_g\sigma^2. \tag{3.8}$$

By essentially the same argument that led to (3.6), the square of the denominator of t_2^* must be approximately equal to

$$(1-\bar{d})^2\theta_1\sum_{g=1}^{G_1}\lambda_g N_g\sigma^2 + \bar{d}^2\theta_0\sum_{g=G_1+1}^{G}\lambda_g N_g\sigma^2. \tag{3.9}$$

Therefore, using (3.8) and (3.9), we conclude that

$$\mathrm{Var}(t_2^*) \cong \frac{(1-\bar{d})^2\sum_{g=1}^{G_1}\lambda_g + \bar{d}^2\sum_{g=G_1+1}^{G}\lambda_g}{(1-\bar{d})^2\theta_1\sum_{g=1}^{G_1}\lambda_g + \bar{d}^2\theta_0\sum_{g=G_1+1}^{G}\lambda_g}, \tag{3.10}$$

which is just (3.7). 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 (3.7) and (3.10) because the wild bootstrap preserves the heteroscedasticity of the original disturbances.

Our key result, from (3.7) and (3.10), is that

$$\mathrm{Var}(t_2^*) \cong \mathrm{Var}(t_2). \tag{3.11}$$

This result depends critically on Assumptions 3.1–3.3. If N_g differed across clusters, in violation of Assumption 3.1, 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 3.1, we could not have made Assumption 3.2. 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 (3.11) to hold. Assumptions 3.1 and 3.2 are not actually necessary. In principle, both N_g and $\boldsymbol{\Omega}_g$ 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 clusters.

Assumption 3.3 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 3.3 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 t_2, and so will the absolute values of the restricted residuals $\tilde{\epsilon}_{i1}$. If the ϵ_{i1} are correlated, then $|\tilde{\epsilon}_{i1}|$ 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 (2.12). We expect the restricted wild bootstrap to under-reject in this case.

A similar argument applies to the unrestricted wild bootstrap. When ϵ_{i1} are correlated, $|\hat{\epsilon}_{i1}|$ will tend to be too small, causing the variance of $\hat{\beta}_2^*$ 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 Section 2.2. We expect the unrestricted wild bootstrap to over-reject in this case.

Our simulation results (see online Appendix A) suggest that the failure of Assumption 3.3 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 3.3 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 (3.11) does not imply that t_2 and t_2^* actually follow the same 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 $N(0, 1)$ under the null. However, as the numerators of both test statistics are weighted sums of disturbances that have mean zero – compare (2.4) and (2.11) – 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 et al. (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 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 (3.11), 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 3.3 holds and N is reasonably large. Moreover, if those random variables have similar distributions for the actual and bootstrap samples, then this should help to make the distribution of t_2^* mimic the distribution of t_2.

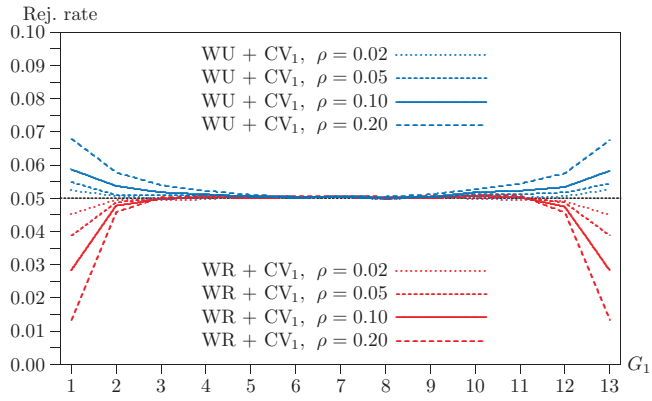

Figure 2. Rejection frequencies for ordinary wild bootstrap tests, $G = 14$, $N/G = 200$. [Colour figure can be viewed at wileyonlinelibrary.com]

We also assumed that the factors θ_1 and θ_0, which determine how badly the two terms in the denominators of (3.3) and (3.4) 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 residuals. 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 (3.11) would no longer hold. This is most likely to happen when Assumption 3.3 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 ith residual is divided by the square root of the ith 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 3.3 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 (3.11) depends critically on Assumption 3.1. 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

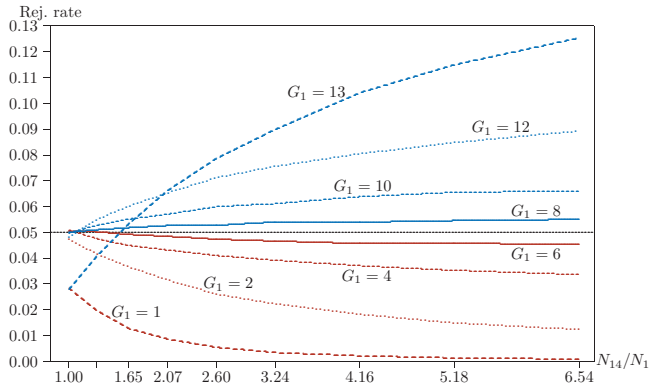

Figure 3. Effects of varying cluster sizes on rejection frequencies for WR + CV_1. [Colour figure can be viewed at wileyonlinelibrary.com]

as t_2 when G_1 or G_0 is small. The ratio would evidently be larger for large clusters than for small clusters, 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 clusters. 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 (3.3) and (3.4). 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. Because $1 - \bar{d} \gg \bar{d}$ 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 clusters, it must be the case that $\text{Var}(t_2^*) > \text{Var}(t_2)$. This will lead the ordinary wild bootstrap test to under-reject. By a similar argument, the test will over-reject whenever the treated clusters are larger than the untreated clusters. 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 N_g to 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 cluster. For details, see online Appendix A. In the experiments, $G = 14$, $\rho = 0.10$, and the average value of N/G is 200.

Figure 3 plots rejection frequencies at the 0.05 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 and 13. We expect to see increasing under-rejection for $G_1 < 7$ as cluster sizes become more variable, and increasing over-rejection 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 Section 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 6. Because $\rho = 0.10$, it under-rejects fairly severely for both $G_1 = 1$ and $G_1 = 13$ when all clusters are the same size. It then under-rejects more and more severely for $G_1 = 1$, and it over-rejects 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 3 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 over-rejection instead of under-rejection 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, because the investigator knows the cluster sizes, he or she will know whether the wild bootstrap is likely to over-reject or under-reject. For example, if the treated clusters are, on average, smaller than the untreated clusters, there is likely to be under-rejection. 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 3.2 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 such as (2.1). In the case of DiD 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 (2.5) for the numerator of the t-statistic has to be replaced by

$$(1 - \bar{d}) \sum_{g=1}^{G_1} \boldsymbol{d}_g' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=1}^{G_1} (\boldsymbol{\iota}_g - \boldsymbol{d}_g)' \boldsymbol{\epsilon}_g - \bar{d} \sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g' \boldsymbol{\epsilon}_g. \tag{3.12}$$

Recall that the $\boldsymbol{d}_g$ are N_g-vectors equal to 1 for treated observations and 0 for untreated observations. 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 (3.12) that the analysis that led to the approximations (3.7) and (3.10) 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 under-reject (over-reject) 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. Under-rejection will probably be more common than over-rejection, 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. Online

Appendix A 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 Section 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, (2.1) can be rewritten as

$$y_{itg} = \beta_1 + \beta_2 d_{itg} + \epsilon_{itg}, \tag{3.13}$$

where g indexes jurisdictions or regions (i.e. 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:

$$y_{itg}^{*b} = \tilde{\beta}_1 + \tilde{\epsilon}_{itg} v_{tg}^{*b}. \tag{3.14}$$

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

For the DGP (3.14), 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 Section 2.2. 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 over-rejects (as it is likely to do when the treated clusters are large) and WCR under-rejects (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 online Appendix A.

4. 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

$$\mathrm{CV}_2 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\Big(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{M}_{gg}^{-1/2}\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{M}_{gg}^{-1/2}\boldsymbol{X}_g\Big)(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.1}$$

where $\boldsymbol{M}_{gg}^{-1/2}$ is the inverse symmetric square root of the gth diagonal block of the $N \times N$ projection matrix $\boldsymbol{M}_X = \mathbf{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. 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 heteroscedasticity-consistent covariance matrix estimator discussed by MacKinnon and White (1985), and the former reduces to the latter when all 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) have suggested a slightly modified version of the Bell–McCaffrey procedure, and Young (2016) has proposed a way to reduce the bias of the diagonal elements of CV_1 and to compute critical values. All these procedures are discussed in online Appendix B, and their performance is studied by simulation in online Appendix C. We mention them here because we implement them in the empirical example of Section 5.

5. EMPIRICAL EXAMPLE

Angrist and Lavy (2001) studied the impact of teacher training on student outcomes using a matched comparisons design in Jerusalem schools. They test whether students who were taught by teachers who 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.[3]

We restrict attention to the change in mathematics 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 mathematics 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 their paper, but our standard error estimate is somewhat smaller.[4] 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 by 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. Because 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 A.4 in online Appendix A 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 $G_1 = 5$.

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

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

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

Table 1. Effects of teacher training on mathematics 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-statistics (P-value)	−4.45	−3.78	−4.41
	(0.003)	(0.009)	(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
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

Note: The outcome variable is the difference between 1994 and 1995 mathematics 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). Here, 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 online Appendix B for details.

of freedom calculated by three different methods, which are described in online Appendix B. 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 alternative 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 suggests 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 Section 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.[5] 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 P-value. Thus, if the results in Column 3 were the only results 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 are for each of the subsamples.

6. CONCLUSION AND RECOMMENDATIONS

Although the wild cluster bootstrap works well much of the time, MacKinnon and Webb (2017b) have shown that it often fails when the number of treated clusters is small, whether or not the total number of clusters is small; see Section 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. 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 Section 3.2 that, for a pure treatment model, the ordinary wild bootstrap can be expected to work very well under certain conditions. First, 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,

[5] 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.

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 under-reject (over-reject) when the treated clusters are smaller (larger) than average. It is also likely to under-reject for DiD 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 over-reject.

We have obtained a large number of simulation results. A few of these were reported above, but most of them are reported in online Appendix A. 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 heteroscedasticity 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 procedures that might be able to provide reasonably reliable inferences when there are few treated clusters. In online Appendices B and C, we discuss several recently proposed procedures that 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 bootstrap and wild cluster bootstrap 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 et al. (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 5. Moreover, the wild bootstrap will probably outperform the wild cluster bootstrap when there is a substantial amount of cluster-specific heteroscedasticity, 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 (two is a lot better than one), 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 under-reject. 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.

ACKNOWLEDGEMENTS

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.

REFERENCES

Andrews, D. W. K. (2005). Cross-section regression with common shocks. *Econometrica 73*, 1551–85.

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

Bell, R. M. and D. F. McCaffrey (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology 28*, 169–81.

Bester, C. A., T. G. Conley and C. B. Hansen (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics 165*, 137–51.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster robust inference. *Journal of Human Resources 50*, 317–72.

Cameron, A. C., J. B. Gelbach and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*, 414–27.

Carter, A. V., K. T. Schnepel and D. G. Steigerwald (2017). Asymptotic behavior of a *t* test robust to cluster heterogeneity. *Review of Economics and Statistics 99*, 698–709.

Conley, T. G. and C. R. Taber (2011). Inference with "Difference in Differences" with a small number of policy changes. *Review of Economics and Statistics 93*, 113–25.

Davidson, R. and E. Flachaire (2008). The wild bootstrap, tamed at last. *Journal of Econometrics 146*, 162–69.

Davidson, R. and J. G. MacKinnon (2010). Wild bootstrap tests for IV regression. *Journal of Business and Economic Statistics 28*, 128–44.

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

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

Gonçalves, S. and T. J. Vogelsang (2011). Block bootstrap HAC robust tests: the sophistication of the naive bootstrap. *Econometric Theory 27*, 745–91.

Ibragimov, R. and U. K. Müller (2016). Inference with few heterogeneous clusters. *Review of Economics and Statistics 98*, 83–96.

Imbens, G. W. and M. Kolesár (2016). Robust standard errors in small samples: some practical advice. *Review of Economics and Statistics 98*, 701–12.

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Annals of Statistics 16*, 1696–708.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, 437–61. Berlin: Springer.

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

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

MacKinnon, J. G. and M. D. Webb (2017b). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics 32*, 233–54.

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

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

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics 29*, 305–25.

Webb, M. 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*, 1261–95.

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

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Online Appendices
Replication files