Webb, Matthew D.

Working Paper

Reworking Wild Bootstrap Based Inference for Clustered Errors

Queen's Economics Department Working Paper, No. 1315

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

Suggested Citation: Webb, Matthew D. (2013) : Reworking Wild Bootstrap Based Inference for Clustered Errors, Queen's Economics Department Working Paper, No. 1315, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:
https://hdl.handle.net/10419/97480

Queen's Economics Department Working Paper No. 1315

Reworking Wild Bootstrap Based Inference for Clustered Errors

Matthew D. Webb
University of Calgary

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

12-2013

Reworking Wild Bootstrap Based Inference for Clustered Errors

Matthew D. Webb*

September 25, 2013

Abstract

Many empirical projects are well suited to incorporating a linear difference-in-differences research design. While estimation is straightforward, reliable inference can be a challenge. Past research has not only demonstrated that estimated standard errors are biased dramatically downwards in models possessing a group clustered design, but has also suggested a number of bootstrap-based improvements to the inference procedure. In this paper, I first demonstrate using Monte Carlo experiments, that these bootstrap-based procedures and traditional cluster-robust standard errors perform poorly in situations with fewer than eleven clusters - a setting faced in many empirical applications. With few clusters, the wild cluster bootstrap-t procedure results in p-values that are not point identified. I subsequently introduce two easy-to-implement alternative procedures that involve the wild bootstrap. Further Monte Carlo simulations provide evidence that the use of a 6-point distribution with the wild bootstrap can improve the reliability of inference.

Keywords: CRVE, grouped data, clustered data, panel data, cluster wild bootstrap

1 Introduction

Difference-in-differences (DiD) estimators have a great deal of appeal, as there are often policy changes that affect a subset of the population. The presence of two groups allows us to make inferences about the causal effects of a policy change. The appropriateness of DiD estimators depends on a few critical assumptions being satisfied, beyond having a treatment and a control group for the estimates. The first assumption is that there is common support amongst the two groups. Common support requires that the composition of both groups—in terms of both observable and unobservable characteristics—be similar.

*Department of Economics, University of Calgary, Calgary, Alberta, Canada, T2N 1N4. Email: mwebb@ucalgary.ca. I thank my supervisors, James MacKinnon and Steven Lehrer for their continuous support. I am grateful to Michele Campolieti, Marco Cozzi, and Allan Gregory for thoughtful evaluations on a prior draft. I would also like to thank Russell Davidson, Emmanuel Flachaire, Maximilien Kaffo Melou, and Arthur Sweetman for helpful comments and suggestions. I am grateful to participants at the 47th Canadian Economics Association Conference, the 8th CIREQ Ph.D. Student Conference, the 29th Canadian Econometric Study Group Annual Meeting, and seminar participants at the University of Calgary.

The second assumption of common or parallel trends requires that each of the groups had a similar trend in the dependent variable before the policy change.[1] The common trend assumption implies that in the absence of the change, both the treatment and control groups would have followed along the same trends as before the change. Any change that is observed in the treatment group, differenced by the change in the control group, is then attributed to the policy change. For more details on program evaluation see DiNardo and Lee (2011).[2]

Beyond worrying about the identifying assumptions for DiD, recent research has asked the question: how reliable are the inferences made with DiD? In answering this question the literature has shown that a fundamental problem with difference-in-differences arises from the use of data with clustered errors. Since ignoring clustered errors leads to very unreliable inference, corrections for clustered errors have become commonplace in empirical work.[3] Although asymptotic corrections work well in many cases, recent studies have suggested that estimating cluster-robust variance-estimator (CRVE) standard errors leads to biased inference when the number of clusters is small.

Donald and Lang (2007) first showed the unreliability of DiD estimators in the case when there are few groups and when some variables are fixed within groups. The authors also used Monte Carlo work that estimated rejection frequencies are too high. The authors propose a two-step method to estimate the significance of DiD coefficients. Wooldridge (2003) proposes an alternative two step method. Bertrand, Duflo and Mullainathan (2004) (BDM) show that in Monte Carlo simulations to show that DiD coefficients are estimated to be significant at the 5% level, 45% of the time. They suggest that the over-rejection is largely driven by the serial correlation in their data. The authors propose a number of methods to correct this problem, including a block bootstrap procedure. Conley and Taber (2011) argue that point estimates within DiD frameworks are not consistent because variables for policy interventions are often invariant over time for a given group. They propose a method of inference which relies on information contained in the control groups. Abadie, Diamond and Hainmueller (2010) propose a similar procedure which involves the construction of synthetic cohorts. Finally, Cameron, Gelbach and Miller (2008) (CGM) propose a wild bootstrap-based procedure extending the work of BDM.

Empirical researchers, following CGM, have frequently used wild cluster bootstrap-t generated p-values for improved inference.[4] However, this paper demonstrates that when the number of clusters is quite small the procedure for inference is noisy and imprecise as estimated 'p-values' are intervals rather point estimates. As a result the standard 2-point wild cluster bootstrap is not appropriate when there are few clusters, with the appropriateness decreasing as the number of clusters decreases. There are many real world problems

[1] Abadie (2005) and Athey and Imbens (2006) relax this assumption.

[2] Arguably, the most well-known application of DiD estimators is Card and Krueger (1994), which examined the impact of increasing the minimum wage on employment in the fast food industry. Other well-cited DiD applications have involved analyzing changes in tax laws on health insurance Gruber and Poterba (1994) and changes to the Earned Income Tax Credit on labour supply Eissa and Liebman (1996). Overviews of difference-in-differences estimators are provided in Meyer (1995) and Angrist and Pischke (2008).

[3] A seminal paper on estimating clustered errors, Rogers (1994), has over 1700 citations according to Google Scholar as of June 2013.

[4] As of June 2013 this article has been cited over 417 times according to Google Scholar.

where data sets contain few clusters. For example, policy analysts in Canada often exploit variation across ten provinces, while policy analysts in Australia often examine eight states. Alternatively, clustering is often accounted for in the time dimension, and it is common in panel data to have few time periods. This is particularly true in finance, following the suggestion of Thompson (2011) that when working with a data set in which the number of firms greatly exceeds the number of time periods, clustering by year will eliminate much of the bias.

This paper proposes two procedures when the sample is collected from a small number of clusters, considering both enumerating the bootstrap samples and new bootstrap weight distributions. Enumeration involves systematically calculating all of the possible bootstrap samples, and their associated t-statistics. The enumeration procedure has the benefit of being invariant to resampling variance, but it is limited in the precision of the calculated p-values when the number of clusters is quite small. Expanding the 2-point wild cluster bootstrap to either a 4-point or a 6-point distribution allows for an approximate test for significance. The 4-point and 6-point wild cluster bootstraps have resampling variability, but more precise p-values can be determined. The proposed distributions appear to work well even in the case of five clusters, when the conventional 2-point wild cluster bootstrap is most inappropriate.

The organization of this paper is as follows: Section 2 provides a background on the challenge of clustered errors in empirical research and current strategies to deal with them. Specifically, the limitations of the 2-point wild cluster bootstrap are identified and examined. Alternative bootstrap methods to account for the small cluster problem are discussed in section 3, with an enumeration technique and the aforementioned new bootstrap weight distributions considered. Section 4 discusses the design and results of Monte Carlo simulations. The results expose the limitations of existing techniques when properly calculated, and favor a new 6-point distribution. Section 5 concludes.

2 Background on Methods to Deal With Within Cluster Correlation

Consider a standard two-period linear difference-in-differences model such as:

$$Y_{igt} = \beta_0 + \beta_1 * treat_g + \beta_2 * post_t + \beta_3 * treat_g * post_t + X_{igt}\gamma + u_{igt}. \tag{1}$$

Here Y_{igt} is an observation for person i in group g and time t, $treat_g$ is a dummy variable for whether the observation is in the treatment group, and $post_t$ is a dummy variable for whether the observation is in time period after the treatment occurred. Neither group was treated in the pre-period. The $treat_g * post_t$ variable is an interaction of the two indicator variables. It is an indicator for those individuals in the treated group in the treated period. The coefficient on this term can be interpreted as the difference-in-differences estimate, which can be viewed as a causal parameter. X_{igt} is a vector of other independent variables. It is quite easy to extend this setup to multiple periods.

Models like (1) are quite common in empirical work, though many papers have shown a problem with using conventional OLS or heteroskedasticity-consistent standard errors for inference when data are of a grouped or clustered nature. A data set can be considered

clustered when there is an underlying natural grouping of the observations. Sometimes these groups are based on methodology, as in data on many students in several classrooms within a particular school. More often the grouping is geographic as there are data on many individuals residing within a given state. The problem is most severe when estimating the impact of a common group variable, such as a treatment variable, on individual level outcomes. The first paper that identified this problem is Kloek (1981), though the problem was popularized by Moulton (1990) and Rogers (1994). The problem was considered in the DiD context by Bertrand, Duflo and Mullainathan (2004) as well as Donald and Lang (2007). For a detailed survey of the issues related to clustered data see Cameron and Miller (2010). Recent work involving rescaling either standard errors or covariance matrices has been done by Imbens and Kolesar (2012) and Brewer, Crossley and Joyce (2013).[5]

The estimates of the β coefficients are unaffected by the clustered nature of the data and can be obtained using the OLS estimator. The issue with clustered data is that the estimated error terms, $\hat{u}_{igt}$, can no longer be assumed to be i.i.d. Although the errors are independently distributed across clusters, the errors are correlated within clusters. Expressed formally, clustered data results in $E[u_g] = 0$, $E[u_g u_g^{'}] = \Sigma_g$, $E[u_g u_h^{'}] = 0$ for cluster $h \neq g$. Given that the i.i.d. assumption is violated, the standard OLS variance matrix is an inappropriate estimate of the variance. The Cluster Robust Variance Estimator (CRVE) was developed by Liang and Zeger (1986) in response to the need to correct for within cluster correlation. The standard Cluster Robust Variance Estimate (CRVE) is given by:

$$\hat{V_{CR}[\hat{\beta}]} = (X^{'}X)^{-1} \left(\sum_{g=1}^{G} X_g \hat{u}_g \hat{u}_g^{'} X_g^{'} \right) (X^{'}X)^{-1}. \tag{2}$$

The CRVE estimate takes a familiar sandwich form, though the $\hat{u}_g$ terms can be non-standard and need to be estimated from the data. In the simplest case, the OLS residuals are used with $\hat{u}_g = y_g - X_g\hat{\beta}$. In other cases, the expression in equation (2), $\sum_{g=1}^{G} X_g \hat{u}_g \hat{u}_g^{'} X_g^{'}$, is replaced by $\sum_{g=1}^{G} \tilde{U}_g \tilde{U}_g^{'}$. Software packages have different routines for estimating $\tilde{U}_g$. For example, Stata uses:

$$\tilde{U}_g = \sum_{i=1}^{N_g} \hat{u}_{ig} \begin{pmatrix} 1 \\ X_g \end{pmatrix},$$

where $\hat{u}_{ig}$ is the OLS residual for individual i in group g. CRVE controls for both error heteroskedasticity and quite general correlation and heteroskedasticity within clusters.

General results in White (1984) on covariance matrix estimation imply the consistency of this estimator based on three assumptions:

A1. The number of clusters, G, goes to infinity.

A2. The degree of within-cluster correlation is constant across clusters.

A3. Each cluster contains an equal number of observations.

[5]Another paper that deals with some of the issues of clustered data is Ibragimov and Muller (2010). However, their paper proposes an estimation technique which requires estimating a t-statistic for each cluster separately. While this technique works well in many situations, it does not work at all when there is a binary independent variable which is invariant within a cluster. As this is often the case with DiD estimates, no Monte Carlo simulations testing their technique will be performed in this paper.

Several authors have previously studied the finite sample properties of the estimator when A1 is not satisfied. Carter, Schnepel and Steigerwald (2012) relax both assumptions A2 and A3 and derive a new asymptotic distribution for the test statistic. MacKinnon and Webb (2013) study the finite sample properties when A3 is violated. Simulation results from Bertrand, Duflo and Mullainathan (BDM), Cameron, Gelbach and Miller (CGM), and those presented in this paper, show that the rejection rates based on OLS standard errors are almost an order of magnitude greater than those based on CRVE. In my own simulations with 30 clusters (discussed at length later in the paper and shown in table 2), the estimated 5% rejection rate for OLS is 49.9% whereas it is only 8% with CRVE standard errors. In simulations with 5-clusters, the rejection rate is 47% for OLS and 21% for CRVE.[6]

2.1 Should we use the Wild Bootstrap to Conduct Inference in DiD Models?

Although CRVE is a substantial improvement over OLS in the presence of grouped data, it is not without its weaknesses. If one uses CRVE with $\tilde{U}_g = \hat{u}_g$ it is biased, as $E[\tilde{U}_g\tilde{U}_g^{'}] \neq \Sigma_g = E[u_g u_g^{'}]$. The bias depends on the form of Σ_g but will usually be downward, which results in coefficients being estimated as significant too often.

After presenting Monte Carlo evidence of the over-rejection problem when using standard CRVE techniques, BDM propose a bootstrap procedure as a means of improving the size of the tests. In particular BDM suggest block bootstrapping, where they resample blocks of all observations from a given state.[7] Cameron, Gelbach and Miller (2008) perform additional Monte Carlo experiments and find that when the number of clusters is small, e.g. fewer than 30, the rejection rate of the block bootstrap method proposed by BDM is too large.[8] CGM investigate several alternative bootstrap methods for improved inference and argue that the 'Wild Cluster bootstrap-t' method is preferred.

The wild cluster bootstrap is preferable to the block bootstrap in several ways. When using the wild cluster bootstrap, each bootstrap sample has the same number of observations, equal to the original sample size, while the block bootstrap generates samples of unequal size. Additionally, every observation in the data set is in every bootstrap sample. This is an important characteristic when identification may be coming from only a few observations, such as when a certain policy is operating in a particular state for only a few years, as pointed out by Conley and Taber (2011). Finally, the wild bootstrap preserves the structure of the error correlation within clusters. The structure is preserved as every residual within a cluster is multiplied by the same weight. CGM present Monte Carlo simulations as evidence that the wild cluster bootstrap-t technique allows for valid inference with as few as five clusters. As this paper will discuss, there is a problem with the rejection rates they calculate to justify that claim. This problem is a result of an insufficient number of unique bootstrap samples.

Imagine we are interested in calculating a wild cluster bootstrap-t p-value for β_3 in equation (1). We can construct the p-value by first estimating the t-statistic, $\hat{t}$, in the

[6] On a historical note, it was Moulton (1990) that pointed out just how large the rejection rates were for OLS, and BDM who pointed out that CRVE is still a great distance from the desired size of 5% when there are few clusters.

[7] The bootstrap samples are generated by sampling the clusters with replacement.

[8] The over rejection is a result of the BDM technique using OLS standard errors in generating bootstrap t-statistics, rather than CRVE standard errors, and a result of less-than-desirable features of the block bootstrap.

original sample using cluster-robust standard errors. We then re-estimate the equation by imposing the null hypothesis, to obtain the restricted estimates $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{u}_{igt}$, etc. Then B iterations, or bootstraps, are performed. In each iteration a bootstrap sample is generated from the bootstrap data generating process

$$y^*_{igt} = \tilde{\beta}_0 + \tilde{\beta}_1 * treat_g + \tilde{\beta}_2 * post_t + \tilde{\beta}_3 * treat_g * post_t + X_{igt}\tilde{\gamma} + \tilde{u}_{igt}v_g, \tag{3}$$

where the i^{th} residual in time t from group g, $\tilde{u}_{igt}$, is transformed by the bootstrap weight v_g.[9] The difference between the wild cluster bootstrap and the conventional wild bootstrap is that under the former the same v_g is applied to all observations within the same cluster, while the conventional wild bootstrap applies a v_{igt} to each observation. The bootstrap weight can take many forms, as will be discussed later. In each iteration, a bootstrap t-statistic, t^*_j is generated using cluster-robust standard errors. After B iterations the bootstrap p-value is then calculated by:

$$\hat{p}^*(\hat{t}) = 2 \min \left(\frac{1}{B}\sum_{j=1}^{B} I(t^*_j \leq \hat{t}), \frac{1}{B}\sum_{j=1}^{B} I(t^*_j > \hat{t}) \right), \tag{4}$$

where $I(.)$ is the standard indicator function.[10]

This procedure is based on the assumption that a given number of bootstrap samples, B, are taken from an extremely large pool of potential bootstrap samples. This means that a set of bootstrap samples are drawn that will contain very few, if any, repeated samples. Suppose we are concerned about the significance of our estimated $\hat{\beta}$ by examining our t-statistic, $\hat{t}$, and we have generated a vector of 999 bootstrap t-statistics, $t^* = t^*_1, ..., t^*_{999}$. If we observe that our estimated t-statistic falls between the 90th and 91st bootstrap t-statistic, then we would say that the p-value associated with this t-statistic is 0.180.[11]

With few clusters the number of unique potential bootstrap samples is rather small. The bootstrap samples depend on the choice of a bootstrap weight distribution. In the literature two well-known distributions are the Rademacher and the Mammen distributions, both of which contain only two points. With these distributions, v_g from equation (3) is set to one of two values with a given probability, p. The Rademacher distribution is defined as:

$$v_g = \pm 1 \text{ with probability } 0.5. \tag{5}$$

The Mammen distribution is defined as:

$$v_g = -\frac{\sqrt{5}-1}{2} \text{ w.p. } p = \frac{\sqrt{5}+1}{2\sqrt{5}} \text{ and } v_g = \frac{\sqrt{5}+1}{2} \text{ w.p. } 1-p.$$

Accordingly, there are only 2^G possible bootstrap samples, where G is the number of groups. When G = 5 there are only 32 possible bootstrap samples. Cameron, Gelbach and Miller

[9] In general the bootstrap DGP should impose the null hypothesis, which in this case would mean setting $\tilde{\beta}_3 = 0$

[10] These p-values are equal tail p-values, while the enumeration p-values are symmetric p-values calculated by $\frac{1}{B}\sum_{j=1}^{B} I(|t^*_j| >= |\hat{t}|)$.

[11] Recall that the equal tail p-value is the result of a two-sided test for t-statistics, so $\hat{p}^* = 2\min\left(\frac{90}{999}, 1 - \frac{90}{999}\right)$.

(CGM) recommend using the wild cluster bootstrap-t technique with Rademacher weights. Thus the 32 bootstrap samples yield 32 distinct t-statistics. However, the set of unique absolute value t-statistics is only 2^{G-1} or 16 in the five cluster case. A proof of this result is provided in the Appendix. When $G = 5$ and $B = 399$, by sampling with replacement you are choosing 399 elements from a set of 32. This is not a problem when G is large as you will obtain a vector of mostly unique t-statistics, but when G is small it is quite problematic.

The CGM procedure inaccurately treats the B t-statistics as B unique values; however, in the small cluster case, the majority of bootstrap t-statistics are not unique. Having many repeated t-statistics leaves open the possibility that $\hat{t}$ will be found multiple times in this vector.[12] We should thus regard unique t-statistics as a signal as to the significance of $\hat{\beta}$, and repeated t-statistics as noise which interferes with our ability to make inferences about $\hat{\beta}$. When 2^G is small we cannot perform conventional inference. This problem comes as a result of the inability to point-identify where $\hat{t}$ falls within the sorted vector of bootstrap t-statistics. If $\hat{t}$ is found multiple times within the vector, then the 'p-value' would not be a point but would instead be an interval from the first occurrence of $\hat{t}$ to the last occurrence of $\hat{t}$. Returning to the above example, if we have $B = 999$ and 31 of those bootstrap samples result in $t^* = \hat{t}$ then $\hat{t}$ would appear in the sorted vector 31 times. For example, if $\hat{t} = t^*_{70}, ..., t^*_{100}$, then the 'p-value' would be the interval from 0.140-0.200. Figure 1 plots the observed spread between the first occurrence p-value and the last occurrence p-value across clusters from Monte Carlo simulations using the Rademacher distribution with 999 bootstraps. The figure shows that the p-values occupy a wide interval when the number of clusters is small. This wide interval makes it quite difficult to assess significance at conventional levels. It is not until there are more than eleven clusters that these intervals are quite small.

Calculating additional bootstrap t-statistics will not shrink the width of these intervals. Calculating an infinite number of bootstraps will lead to results equivalent to those of the enumeration p-values, discussed in section 3. One way to generate more unique t-statistics is to increase the number of clusters, though in empirical work the number of clusters will be determined by the data.

When using the Rademacher distribution, one of the possible bootstrap t-statistics, t^*_j is the original estimate of the t-statistic, $\hat{t}$.[13] When 2^G is small, this will be observed, almost surely. As there are only 2^G possible t-statistics, estimating the p-value depends on identifying where $\hat{t}$ lies in the vector of sorted t-statistics. The 2^G possible t-statistics map into a small number of p-values. If inference is done correctly, we should only observe that small number of p-values across Monte Carlo simulations. CGM instead chose to estimate the p-value as being the center of this range. Figure 2 shows a histogram of 50,000 p-values based on the CGM method for 2-point wild cluster bootstrap-t. If the CGM procedure worked appropriately, only 16 unique p-values would be observed.[14] However, the histogram is quite smooth and shows that numerous p-values were calculated. These

[12] There will also likely be repetitions of other t-statistics in the vector of bootstrap t-statistics.

[13] This occurs when $v_g = 1, \forall g$.

[14] The fact there are 16 p-values is a result of $\hat{t} = t^*_2$ and $\hat{t} = t^*_{30}$ resulting in the same p-value in a two tail test.

additional p-values are a result of the noise from their estimates.[15] The noise leads to improper inference and makes the 2-point wild bootstrap inappropriate in cases with few clusters.

3 Alternative Bootstrap Methods

The first technique considered in this paper for improving inference is that of enumeration. The above mentioned issues are a result of using a 2-point distribution in general, and the Rademacher distribution in particular. Inference using a 2-point distribution will be limited in the few cluster case on account of the interval identified p-values. However, if one is convinced that the Rademacher distribution is ideal, then enumeration is the correct way to conduct analysis.[16] The procedure for estimating a p-value is quite similar to the wild cluster bootstrap procedure discussed above. The main difference is that with the wild bootstrap v_g is picked at random from the distribution, while with enumeration v_g is selected methodically. Given the small number of possible bootstrap samples when the number of clusters is small, it is feasible to calculate all possible t-statistics. When all possible test statistics are calculated, it is referred to as full enumeration; when a subset of these test statistics is calculated, it is referred to as partial enumeration.

Under full enumeration, the resulting p-values do not depend on resampling variation. In conventional bootstrap procedures the results will depend in part on the set of samples drawn, and thus are subject to resampling variation. This is not the case with full enumeration as all samples have been drawn. When the number of clusters is large, it is infeasible to calculate all possible t-statistics; however, partial enumeration is possible and will result in a sample of bootstrap t-statistics without any repetitions. The main benefit of enumeration is that you get a sample of t-statistics with no repetition, though there can be a small benefit in terms of computing time.

The resulting p-value of this procedure is different than a conventional p-value. For instance when $G = 5$ there are only 2^{G-1} unique t-statistics in absolute value. If $|\hat{t}| = |t_2^*|$, the p-value is equal to $\frac{2}{16}$, which tells us something about the statistical significance of $\hat{\beta}$. We have to be careful not to think about this p-value as 0.125, as doing so can lead one to incorrectly infer that the observed p-value is drawn from a continuous distribution. In this case the p-value is $\frac{2}{16}$, but it could have alternatively been $\frac{1}{16}$ or $\frac{3}{16}$, and is drawn from a discrete distribution with the p-value $\in \{\frac{1}{16}, ..., \frac{16}{16}\}$. The issue here is that conventional significance levels that applied econometricians work with are not as meaningful. The p-value of $\frac{2}{16}$ spans the space from $0.0625 - 0.125$ and so straddles the 10% level.[17]

Although enumeration has much to its credit, its advantages are largely confined to cases with small G. After G is sufficiently large, say 12, the computational limitations of full enumeration necessitate partial enumeration, which is very similar to the wild cluster

[15] The noise comes in part from repeated $\hat{t}$, but also from repeated t_1^*, t_2^*,etc. The number of $\hat{t}$ repetitions changes from one replication to another, thus the variability of p-values in their technique.

[16] This procedure was alluded to in Efron's seminal bootstrap paper in 1979 and mentioned in Davidson and Flachaire (2008) specifically in the context of the (non-cluster) wild bootstrap.

[17] Perhaps we are best to remain agnostic about whether this observed p-value is significant at the 10% level, thus the recommendation of reporting enumerated p-values as fractions as opposed to decimals to highlight the distinction.

bootstrap. Figure 3 shows the histogram of Monte Carlo p-values using the enumeration method for the 5-cluster case. This figure is comparable to figure 2 though here it is easy to see that only the p-values associated with the 16 unique bootstrap t-statistics have been calculated. Using this technique results in inference being based on the data and the properties of the bootstrap weighting distribution, and not on resampling noise.

3.1 Adding Points to the Bootstrap Weight Distributions

Enumeration will generate unique t-statistics, and thus is more precise than the conventional wild bootstrap procedure. However, the limitation of having only 2^{G-1} t-statistics from which to conduct inference leaves much to be desired. It is possible to find variations to the wild bootstrap technique which expand the number of points in the weight distribution, v_g in equation (3), used to generate bootstrap samples. Following Davidson and Flachaire (2008), who show that the Rademacher distributions has better finite sample properties than the Mammen distribution, I look for variations of the Rademacher distribution.

The first four moments of the 'ideal' distribution would be 0,1,1,1. It is however not possible to satisfy all of these moment conditions.[18] The Rademacher and Mammen distributions differ in the moment conditions that they satisfy. Both distributions have a mean of zero and a variance of one. The Mammen distribution has a third moment equal to one, and a fourth moment of two. The Rademacher distribution has a third moment of zero and a fourth moment of one. The candidate distribution will expand the Rademacher distribution, imposing symmetry. Like the Rademacher, the candidate distributions ignore the third moment. The candidate 4-point distribution I consider is:

$$v_g = -\sqrt{\frac{3}{2}}, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{3}{2}} \qquad w.p. \qquad \frac{1}{4}. \tag{6}$$

In addition to imposing symmetry the 6-point distribution will impose a restriction that two of the points are 1 and -1. The imposition of symmetry means that the third moment will be 0. The candidate distribution will then have 6-points of the form $-A, -1, -B, B, 1, A$ each selected with equal probability. The first four moments of this symmetric 6-point distribution would have to be 0,1,0,1 to match the Rademacher moments. This also is impossible. Any symmetric equal probability distribution will automatically satisfy the first and third moment restrictions. It is then a matter of trying to satisfy the second and fourth moment conditions. Rearranging these moment conditions results in the following equation: $A^2 + B^2 + 1^2 = A^4 + B^4 + 1^4$. This is only satisfied when A and B are $0, 1,$ or -1, which does not result in a 6-point distribution. Thus it is not possible to have a distribution of the form $-A, -1, -B, B, 1, A$ with the first four moments of $0, 1, 0, 1$. The candidate 6-point distribution I consider is:

$$v_g = -\sqrt{\frac{3}{2}}, -\sqrt{\frac{2}{2}}, -\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{2}{2}}, \sqrt{\frac{3}{2}} \qquad w.p. \qquad \frac{1}{6}. \tag{7}$$

The fourth moments of these distributions are $\frac{5}{4}$ for the 4-point, and $\frac{7}{6}$ for the 6-point. There exists the temptation to add additional points to the distribution to increase the

[18] I thank Professor James MacKinnon and Professor Russell Davidson for bringing this to my attention.

potential number of bootstrap samples. There are two concerns about doing so. The first is that the ideal weights will be distinct from one another, as the weights 0.99 and 1.01 will result in very similar bootstrap samples and very similar bootstrap t-statistics, t_j^*. The second is that given the desire to have a distribution with distinct weights, mean zero, and variance one, the inclusion of additional weights will often increase the fourth moment. As a limiting case I consider using the normal distribution to generate weights for the bootstrap sample where $v_g \sim N(0,1)$. Drawing from the normal would allow for infinite possible bootstrap samples. Mammen (1993) considered the distribution $v_g = u_i/\sqrt{2} + \frac{1}{2}(w_i^2 - 1)$, where u_i and w_i are draws from the normal distribution.[19]

The main benefit of adding additional points to the bootstrap weight distribution is that the number of potential bootstrap samples increases exponentially. For instance, with the 2-point distribution the number of bootstrap samples is 2^G, but with the 4-point distribution it increases to 4^G, and to 6^G for the 6-point distribution. So in the case of five clusters, the number of potential bootstrap samples increases from 32 to 1024 to 7776. It should be noted that the unique number of absolute value t-statistics is less than 4^G or 6^G. The proposed distributions are also symmetric, and so have the same feature as the Rademacher distribution. As a result, the number of unique absolute value t-statistics is $\frac{4^G}{2}$ for the 4-point and $\frac{6^G}{2}$ for the 6-point. The large number of possible bootstrap samples should give us confidence that inference conducted using the 6-point distribution is based primarily on the estimated t-statistics, and not on noise introduced from resampling. Figure 4 shows a histogram of 50,000 p-values based on the 6-point wild bootstrap method. In contrast to figure 2, the smoothness seen in this figure comes from the great number of unique and correctly calculated p-values.

4 Monte Carlo Evidence

4.1 Description of Simulations

To enhance the comparability of the simulations, I follow the simulation procedure in section IV.A of Cameron, Gelbach and Miller (2008). Data are generated using

$$\begin{gathered} y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \\ \text{or} \\ y_{ig} = \beta_0 + \beta_1 (z_g + z_{ig}) + (\epsilon_g + \epsilon_{ig}). \end{gathered} \tag{8}$$

With $z_g, z_{ig}, \epsilon_g, \epsilon_{ig}$ each an independent draw from $N(0,1)$. We can think of z_g as a group specific component of x_{ig} and ϵ_g as the group level error. The presence of ϵ_g introduces correlation amongst the error terms. Alternatively, z_{ig} is the idiosyncratic component of x_{ig}, while ϵ_{ig} is the idiosyncratic component of the error term.

The number of observations per group, N_g, is set to 30 for all simulations. I perform 50,000 replications, and each of the bootstrap exercises uses 399 bootstraps. In generating

[19] Mammen (1993) also considered another more complicated distribution. These two distributions are ignored in this paper since simulation results in MacKinnon (2012) show them to be inferior to the Normal distribution.

y_{ig}, I set $\beta_1 = 1$ and test the hypothesis that $\beta_1 = 1$.[20] Following common practice, the null hypothesis is imposed in the bootstrap replications. The rejection rates are estimated across replications as

$$\hat{\alpha} = \frac{1}{R}\sum_{j=1}^{R} I(p_j^* \leq 0.05),$$

where R is the number of replications, and p_j^* is the bootstrap p-value from the j^{th} replication. This $\hat{\alpha}$ is then compared to the true size of the test which is given by $\alpha = 0.05$.

In total eight different rejection rates are compared, across a variety of asymptotic and bootstrap methodologies. A description of the simulations can be found in table 1. Designs 1-3 use asymptotic procedures for generating p-values, while designs 4-8 use bootstrap procedures. Design 1 uses t-statistics which are calculated using OLS standard errors and are assumed to follow a normal distribution. As the OLS standard errors ignore the clustered nature of the data these rejection rates should be rather high, as was pointed out by Moulton (1990). Design 2 uses the CRVE standard errors of equation (2), and the t-statistics are assumed to be distributed normally. Design 3 uses the same t-statistics as in design 2, but the distribution of the t-statistics is assumed to follow a t-distribution with G-1 degrees of freedom, where G is the number of groups.[21]

Designs 4-8 employ the wild cluster bootstrap-t procedure as discussed above, but differ in which bootstrap weight distribution is used. Design 4 generates p-values using the wild cluster bootstrap with v_g drawn from the 2-point Rademacher distribution described above in equation (5), this is the test that was recommended by CGM.[22] Design 5 generates p-values with v_g drawn from $N(0,1)$. Design 6 uses v_g drawn from the 4-point distribution that was proposed above in equation (6). Design 7 uses v_g drawn from the 6-point distribution that was proposed above in equation (7). Finally, design 8 generates p-values by enumerating the Rademacher wild bootstrap t-statistics. When $G \leq 11$ full enumeration is used and all t-statistics are calculated. When $G > 11$ partial enumeration is used and a unique set of t-statistics is calculated. The results of the Monte Carlo experiments are discussed below.

4.2 Simulation Results

Table 2 replicates Cameron, Gelbach and Miller (2008) by performing tests 1-5.[23] The table shows the severe problem of ignoring the clustered nature of the data, as the test using

[20] The code I used for performing the bootstrap simulations is based off the code provided by Douglas Miller, which can be found at: `http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/bs_example.do` I thank the author for making his code publicly available.

[21] This distribution is both recommended by Donald and Lang (2007) and is the default Stata uses with the cluster command. The asymptotic justification for this distribution is provided in Bester, Conley and Hansen (2011).

[22] The values reported are slightly different than the values reported by CGM. They are different because different random numbers were used, but more importantly because CGM use the average value at which $\hat{t}$ matches the bootstrap t-statistics, while I use the max value at which this occurs. The difference is negligible when 2^G is large, but significant when 2^G is small, see figure 1. The difference is largely irrelevant as neither rejection rate is correct in the small G case.

[23] In all of the tables, the simulation standard error is not shown to save space. The standard error is $s_{\hat{\alpha}} = \sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{R-1}}$ with R being the number of replications, and $\hat{\alpha}$ being the estimated rejection rate. Given the observed rejection rates the largest standard error is 0.0022 and the smallest is 0.0008.

OLS standard errors gives rejection rates of close to 50%. Clustering the standard errors and performing inference-based tests 2 and 3 works much better. Assuming that the t-statistics are normally distributed is rather problematic when there are very few clusters. The assumption that the t-statistics follow a t-distribution with $G-1$ degrees of freedom goes a long way in correcting the size of the test, but the rejection rate is still too large when G is very small. The rejection rates for the wild cluster bootstrap-t with Rademacher weights look deceivingly nice. As explained above the results for $G=5$ and $G=10$ should not be trusted as they are based on a very noisy vector of t-statistics, but the results for $G \geq 15$ do not suffer from this problem. A histogram of the 5-cluster wild bootstrap p-values can be seen in figure 2. The under-rejection in the table is evidenced by the under-concentration of p-values in the far left of the figure.

Table 3 shows the results of simulations in which the number of clusters is small.[24] The wild bootstrap with Normal weights does fairly well, though it is outperformed in most cases by the wild bootstrap with either the 4-point or 6-point distribution.[25] Both the 4-point and 6-point distribution work well, though in all cases the 6-point distribution outperforms the 4-point distribution. Note that when $G=5$ the rejection rate is 0.070, which is still noticeably above 0.05, but better than the $T(G-1)$ rate of 0.100. This over-rejection can also be seen in figure 4, as evidenced by the over-concentration of p-values in the left tail of the histogram.

As mentioned previously, the enumerated p-values are not point identified and are instead identified by an interval. Two rejection frequencies are calculated for these p-values, one using the lower bound, and one using the upper bound. The wide differences in these two rejection frequencies are to be expected, as was shown in figure 1. In the 5-cluster case the upper bound never rejects at the 5% level as $\frac{1}{16}$ is above that threshold. The lower bound rejects far too often. This is particularly interesting since both of these rejection frequencies come from the same estimated t-statistic. The upper bound rejection frequency and lower bound rejection frequencies converge as G increases. For the 5-cluster case, a histogram of the 16 t-statistics is shown in figure 3. The over-concentration of p-values in the left of the figure corresponds with the result in the table that the enumeration technique rejects too often when using the lower bound of the interval. Even with 10 clusters the enumeration rejection frequencies are higher than those from the 6-point distribution.

Table 4 shows the results when the number of clusters ranges from 5 to 30. The partial enumeration method which is used for $G \geq 15$ works well in the case of G=15 and G=20. The results are not presented for $G \geq 25$, since they should be equivalent to the results from using the 2-point distribution. The 4-point and 6-point distributions work equally well, though in a few cases the 6-point distribution outperforms the 4-point distribution. In general, the various bootstrap methods work better than the analytic $T(G-1)$ method, test 3. While the 6-point distribution dominates the 2-point distribution in most cases, the 2-point distribution does slightly better when G is equal to 15 or 20, though it slightly under-rejects compared to the 6-point bootstrap when G is larger. Given the small range in which the 6-point distribution is inferior to the 2-point distribution, and the problems

[24] The results for the wild cluster bootstrap with Rademacher weights are not presented in this table since the p-values are not correctly calculated in the 5-10 cluster range.

[25] The normal distribution is not ideal since the fourth moment is too large; see MacKinnon (2012).

with the 2-point distribution in the case of few clusters, the 6-point distribution is generally preferable.

5 Conclusion

While difference-in-differences estimators are widely used to evaluate policy changes, care must be taken in performing inference. This is particularly true when individual-level data are used, and when the data are grouped or clustered. I evaluate the performance of several inference procedures in Monte Carlo simulations and confirm the findings of Cameron, Gelbach and Miller (2008). In this paper I show that substantial improvements can be made to the inference procedure when the researcher faces few clusters. The few cluster concern is quite common in practice, as many data sets have fewer than eleven clusters. The issue with the conventional wild bootstrap procedure is that it uses a 2-point weight distribution. The small number of weights leads to p-values not being point identified when there are few clusters. In cases with 5-clusters these intervals have a width of 0.0625 and make conventional inference quite difficult. Using a 6-point weight distribution solves this problem and works equally well when there are more clusters.

References

Abadie, Alberto (2005) 'Semiparametric difference-in-differences estimators.' *The Review of Economic Studies* 72(1), 1–19

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) 'Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program.' *Journal of the American Statistical Association* 105(490), 493–505

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

Athey, Susan, and Guido W. Imbens (2006) 'Identification and inference in nonlinear difference-in-differences models.' *Econometrica* 74(2), 431–497

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

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011) 'Inference with dependent data using cluster covariance estimators.' *Journal of Econometrics* 165(2), 137–151

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

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

Cameron, A.C., and D.L. Miller (2010) 'Robust inference with clustered data.' Technical Report, UC Davis Department of Economics, February

Card, David, and Alan B Krueger (1994) 'Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania.' *American Economic Review* 84(4), 772–93

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2012) 'Cluster robust inference for heterogeneous cluster samples.' 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

DiNardo, John Enrico, and David S. Lee (2011) 'Program evaluation and research designs.' 1 ed., vol. 4A (Elsevier) chapter 05, pp. 463–536

Donald, Stephen G, and Kevin Lang (2007) 'Inference with difference-in-differences and other panel data.' *The Review of Economics and Statistics* 89(2), 221–233

Efron, B. (1979) 'Bootstrap methods: Another look at the jackknife.' *The Annals of Statistics* 7(1), pp. 1–26

Eissa, Nada, and Jeffrey B Liebman (1996) 'Labor supply response to the earned income tax credit.' *The Quarterly Journal of Economics* 111(2), 605–37

Gruber, Jonathan, and James Poterba (1994) 'Tax incentives and the decision to purchase health insurance: Evidence from the self-employed.' *The Quarterly Journal of Economics* 109(3), 701–733

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

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

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

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

MacKinnon, James G. (2012) 'Inference based on the wild bootstrap.' Seminar presentation given to Carleton University in September 2012.

MacKinnon, James G., and Matthew D. Webb (2013) 'Wild bootstrap inference for wildly different cluster sizes.' Working Papers 1314, Queen's University, Department of Economics, August

Mammen, Enno (1993) 'Bootstrap and wild bootstrap for high dimensional linear models.' *The Annals of Statistics* 21(1), pp. 255–285

Meyer, Bruce D (1995) 'Natural and quasi-experiments in economics.' *Journal of Business & Economic Statistics* 13(2), 151–61

Moulton, Brent R. (1990) 'An illustration of a pitfall in estimating the effects of aggregate variables on micro units.' *Review of Economics & Statistics* 72(2), 334

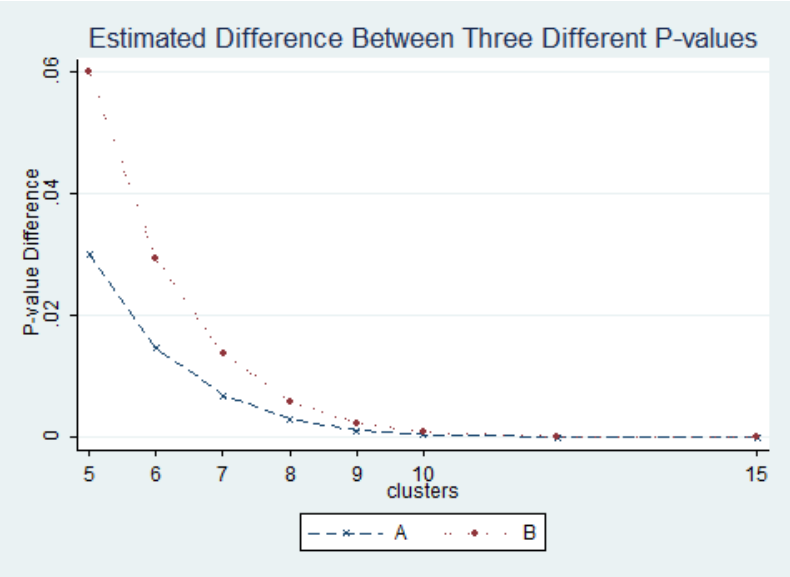
Rogers, William (1994) 'Regression standard errors in clustered samples.' *Stata Technical Bulletin*

Thompson, Samuel B. (2011) 'Simple formulas for standard errors that cluster by both firm and time.' *Journal of Financial Economics* 99(1), 1–10

White, Halbert (1984) *Asymptotic theory for econometricians* (Academic Press)

Wooldridge, Jeffrey M. (2003) 'Cluster-sample methods in applied econometrics.' *American Economic Review* 93(2), 133–138

Figure 1: Estimated Differences From Three Different P-values



Notes: A is the difference between the max p-value and the mean p-value. B is the difference between the max p-value and the min p-value.

Figure 2: Histogram of 50,000 Monte Carlo P-values: Rademacher Distribution

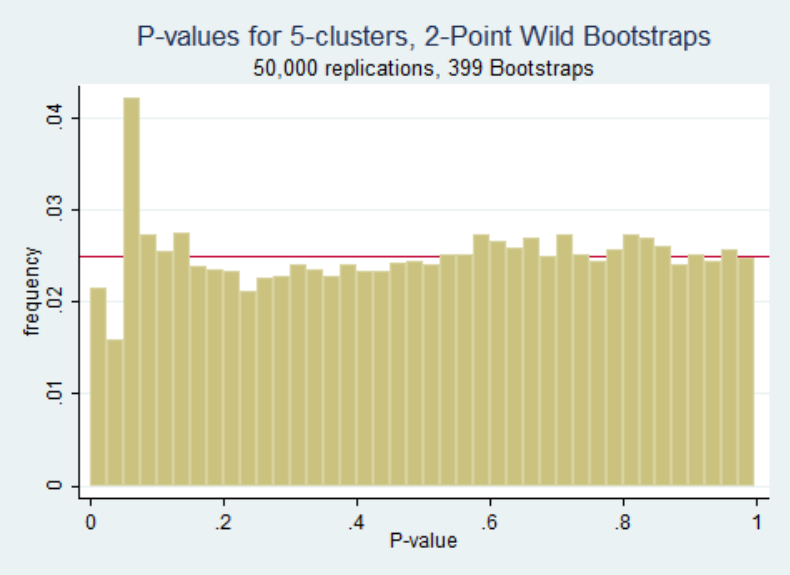


Figure 3: Histogram of 50,000 Monte Carlo P-values: Enumerated Wild Bootstrap

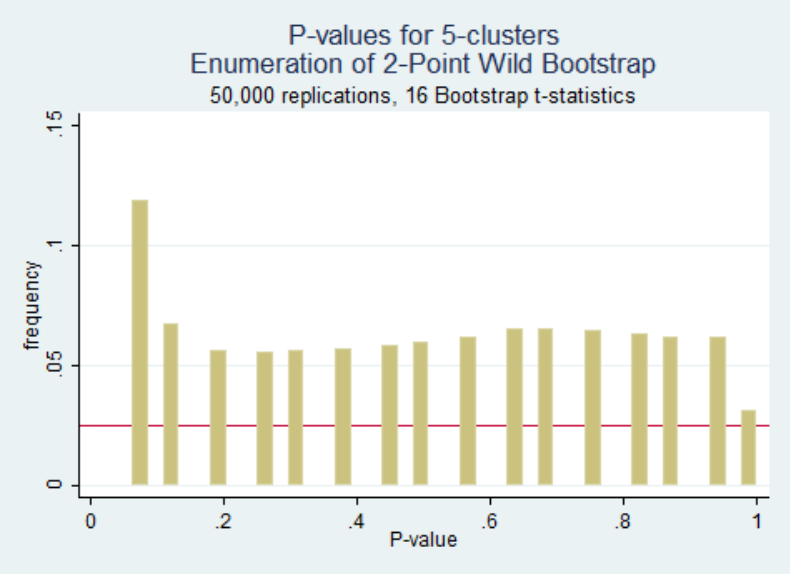


Figure 4: Histogram of 50,000 Monte Carlo P-values: 6-point Distribution

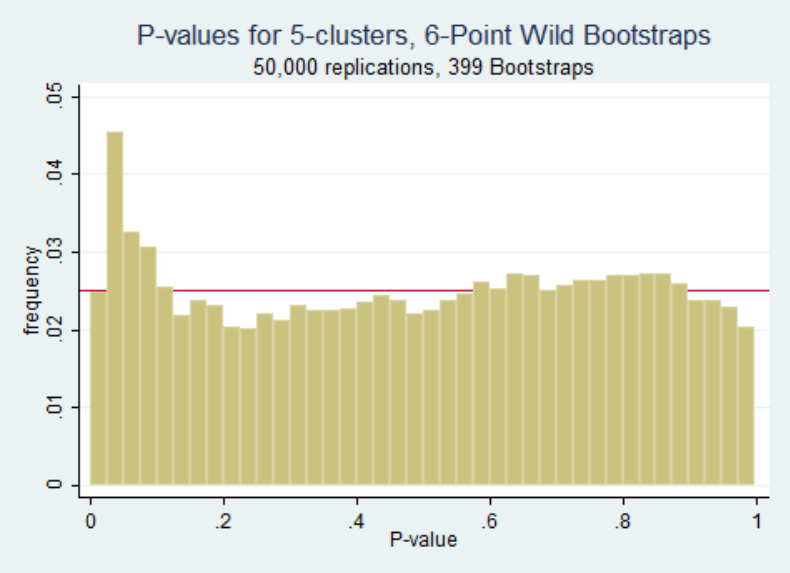


Table 1: Design of Monte Carlo Simulations

Design #	Description	Standard Error	t-statistic distributed as	Bootstrap Weights
1	OLS	OLS	N(0,1)	-
2	Cluster $\sim N$	CRVE	N(0,1)	-
3	Cluster $\sim T$	CRVE	T(G-1)	-
4	Wild Cluster - Rademacher	CRVE	-	2-point - rademacher
5	Wild Cluster - Normal	CRVE	-	$\sim N(0,1)$
6	Wild Cluster - 4-point	CRVE	-	4-point equation (6)
7	Wild Cluster - 6-point	CRVE	-	6-point equation (7)
8	Enumeration - Rademacher	CRVE	-	2-point - rademacher

Table 2: Results from Monte Carlo Study with Different Numbers of Clusters: Replicating Results in Cameron, Gelbach, and Miller (2008)

		Number of Groups (G)					
		5	10	15	20	25	30
1	OLS $\sim N(0,1)$	0.468	0.486	0.493	0.494	0.489	0.499
2	CRVE $\sim N(0,1)$	0.211	0.133	0.108	0.094	0.084	0.080
3	CRVE $\sim T(G-1)$	0.100	0.090	0.081	0.075	0.070	0.069
4	Wild 2pt BS	*0.037	*0.054	0.050	0.050	0.047	0.048

Notes: Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). * - estimate is not accurately calculated. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00084 and the largest standard error is .00224.

Table 3: Results from Monte Carlo Study with Different Numbers of Clusters: Small Number of Clusters Simulation

		Number of Groups (G)					
		5	6	7	8	9	10
3	CRVE $\sim T(G-1)$	0.100	0.100	0.094	0.096	0.088	0.090
5	Wild $N(0,1)$ BS	0.072	0.070	0.072	0.072	0.071	0.069
6	Wild 4pt BS	0.070	0.069	0.064	0.062	0.059	0.057
7	**Wild 6pt BS**	**0.070**	**0.067**	**0.063**	**0.061**	**0.057**	**0.056**
8	Enum. Lower Bound	0.118	0.095	0.084	0.068	0.062	0.060
8	Enum. Upper Bound	0.000	0.059	0.067	0.061	0.058	0.058

Notes: Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). Preferred procedure is presented in **bold**. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00000 and the largest standard error is .00144.

Table 4: Results from Monte Carlo Study with Different Numbers of Clusters: Larger Number of Clusters Simulation

		Number of Groups (G)					
		5	10	15	20	25	30
1	CRVE $\sim T(G-1)$	0.100	0.090	0.081	0.075	0.070	0.069
4	Wild 2pt BS	*0.037	*0.054	0.050	0.050	0.047	0.048
5	Wild $N(0,1)$ BS	0.072	0.069	0.065	0.063	0.059	0.059
6	Wild 4pt BS	0.070	0.057	0.054	0.052	0.048	0.049
7	**Wild 6pt BS**	**0.070**	**0.056**	**0.052**	**0.052**	**0.049**	**0.049**
8	Enum. Lower Bound	0.118	0.060				
8	Enum. 2pt BS			0.052	0.051		
8	Enum. Upper Bound	0.000	0.058				

Notes: Rejection frequencies estimated with 50,000 replications and 399 bootstraps (BS). * - estimate is not accurately calculated. Preferred procedure is presented in **bold**. Simulation standard errors have been omitted from this table. The smallest standard error in this table is .00084 and the largest standard error is .00144.

Appendix - Proof of 2^{G-1} Unique Absolute Value t-statistics

Recall that a bootstrap sample is generated by:

$$y_i^* = X\tilde{\beta} + \tilde{u}_i^*, \tag{9}$$

where $\tilde{u}_i^*$ is the Hadamard product $\tilde{u} \circ v_i$, and v_i is the vector of draws of the bootstrap weights. The Rademacher weights are -1 and $+1$, so every possible v_i is equal to $-1 \circ v_j$ for some $i \neq j$. These two bootstrap weight draws will generate the following bootstrap samples: $y_i^* = X\tilde{\beta} + \tilde{u} \circ v_i$ and $y_j^* = X\tilde{\beta} + \tilde{u} \circ v_j$. Since $v_j = -1 \circ v_i$ we can rewrite y_j^* as $y_j^* = X\tilde{\beta} - \tilde{u} \circ v_i$.

We then test the null hypothesis β_i^* and β_j^* are $= \beta_o$ and calculate a t-statistic of the form:

$$\frac{(X'X)^{-1}X'y_i - \beta_o}{\left(\frac{u_i'u_i}{X'X(n-k)}\right)^{1/2}}. \tag{10}$$

The denominator in equation (10) is constant for either i or j, as X and $n-k$ are invariant and $u_i'u_i = u_j'u_j$ because $u_j = -1 \circ u_i$.

Let us consider the numerator in equation (10), where we have an expression in terms of β_i^* and β_o. If we start with the expression:

$$(X'X)^{-1}X'y_i - \beta_o,$$

using the identity that $y_i = X\tilde{\beta} + \tilde{u} \circ v_i$ we get the following,

$$(X'X)^{-1}X'(X\tilde{\beta} + \tilde{u} \circ v_i) - \beta_o.$$

With a little algebra we get:

$$\begin{aligned}
&(X'X)^{-1}X'(X\tilde{\beta} + \tilde{u} \circ v_i) - \beta_o \\
=&(X'X)^{-1}X'X\tilde{\beta} + (X'X)^{-1}X'(\tilde{u} \circ v_i) - \beta_o \\
=&\tilde{\beta} - \beta_o + (X'X)^{-1}X'(\tilde{u} \circ v_i).
\end{aligned}$$

Because the bootstrap samples impose the null hypothesis, $\tilde{\beta} = \beta_o$. The numerator then simplifies to:

$$(X'X)^{-1}X'(\tilde{u} \circ v_i).$$

Because $v_i = -1 \circ v_j$, the numerator for the t-statistic of β_j^* will be the negative of the numerator for the t-statistic of β_i^*. Because the denominators are also the same, the t-statistics are equal in absolute value. If we reverse the sign on the weight vector v_i we reverse the sign of the t-statistic, but preserve the magnitude. Thus the 2^G unique bootstrap samples will only result in 2^{G-1} unique t-statistics in absolute value.