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

11-2014

Reworking Wild Bootstrap Based Inference for Clustered Errors

Matthew D. Webb*

November 13, 2014

Abstract

Many empirical projects involve estimation with clustered data. While estimation is straightforward, reliable inference can be challenging. Past research has suggested a number of bootstrap procedures when there are few clusters. I demonstrate, using Monte Carlo experiments, that these bootstrap procedures perform poorly with fewer than eleven clusters. With few clusters, the wild cluster bootstrap results in p-values that are not point identified. I suggest two alternative wild bootstrap procedures. Monte Carlo simulations provide evidence that a 6-point bootstrap weight distribution improves the reliability of inference. A brief empirical example concerning education tax credits highlights the implications of these findings.

JEL:C15, C21, C23
Keywords:CRVE, grouped data, clustered data, panel data, wild cluster bootstrap

*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 continued support. I am grateful to Michele Campolieti, Marco Cozzi, Allan Gregory, and anonymous referees for thoughtful evaluations on a prior draft. I would also like to thank Russell Davidson, Jonah Gelbach, 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, Université du Québec à Montréal, Wilfrid Laurier University, and Ryerson University.

1 Introduction

Research often involves controlling for dependence within clusters. Clusters can be regarded as a natural grouping of observations. Common examples of clusters are students within classrooms, firms within industries, and individuals within states. When the data are clustered OLS or 'robust' means of inference are quite unreliable. This problem occurs whenever there is strong correlation of independent variables or error terms within a cluster or group. It is most severe whenever an independent variable is invariant within a cluster, as discussed in Moulton (1990). A very thorough survey is provided by Cameron and Miller (2014).

Issues of within cluster dependence have been of concern to applied researchers for quite sometime, and packages such as Stata's 'cluster' command are now commonplace within statistical analysis packages.[1] These packages implement Cluster Robust Variance Estimator (CRVE) routines and often work very well. However, problems can occur when the data under analysis fail to meet the asymptotic requirements of the CRVE. This frequently occurs when there are a small number of clusters in the dataset, a result known since Bertrand, Duflo and Mullainathan (2004) (BDM) and Donald and Lang (2007). With few clusters, the CRVE can result in p-values that are, on average, too small resulting in type I errors occurring too frequently.

A common correction for the small clusters problem is to use a wild cluster bootstrap, due to Cameron, Gelbach and Miller (2008) (CGM).[2] This technique works very well with an intermediate number of clusters; however, this paper demonstrates that with few clusters the procedure results in p-values that are not point identified. The appropriateness of the conventional wild cluster bootstrap increases with the number of clusters. Yet, there are many real world problems where data sets contain few clusters. For example, policy analysts in Australia and Canada often exploit variation across eight or ten regions, while others exploit variation between and within regions in the United States. Alternatively, following Thompson (2011) clustering is often accounted for in the time dimension, and many panel data sets have few time periods.

This paper suggests two procedures when working with few clusters, considering

[1] Rogers (1994) implemented cluster robust inference within Stata and has over 1980 citations according to Google Scholar as of September, 2014.

[2] Cameron, Gelbach and Miller (2008) has over 670 citations according to Google Scholar as of September, 2014.

both enumerating the bootstrap samples and alternative bootstrap weight distributions. Enumeration involves systematically calculating all of the possible bootstrap samples, and their associated t-statistics. Expanding the 2-point wild cluster bootstrap to a 6-point distribution appears to perform well, even in settings with five clusters.

Section 2 of this paper discusses the limitations of the 2-point wild cluster bootstrap. Alternative bootstrap methods to account for the few clusters problem are considered in section 3. Section 4 addresses the design and results of Monte Carlo simulations which expose the limitations of existing techniques when properly calculated, and favor a new 6-point distribution. A brief empirical example applies these procedures to an analysis of education related tax credits in section 5 and section 6 concludes.

2 Background on Methods to Deal With Within Cluster Correlation

A data set can be considered clustered when there is a natural grouping of the observations. A common correction for clustered errors is to estimate standard errors using a Cluster Robust Variance Estimator (CRVE).[3] 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.

Several authors have previously studied the performance of the CRVE when G is small. Simulation results from Bertrand, Duflo and Mullainathan (2004) and others show that with 6 clusters CRVE rejection rates can be several times the desired size.[4]

BDM propose a block bootstrap procedure as a means of improving test sizes, and

[3]Kloek (1981) identified the problem of constant regressors within grouped data, though it was popularized by Liang and Zeger (1986), Moulton (1990), and Rogers (1994). The problem was considered in the Difference-in-Differences context by Bertrand, Duflo and Mullainathan (2004) and Donald and Lang (2007). Recent work has been done by Ibragimov and Muller (2010), Imbens and Kolesar (2012) and Brewer, Crossley and Joyce (2013). For a detailed survey on cluster robust inference see Cameron and Miller (2014).

[4]Carter, Schnepel and Steigerwald (2013) relax assumptions A2 and A3 and derive a new asymptotic distribution for the test statistic. Imbens and Kolesar (2012) also deal with violations of A3. MacKinnon and Webb (2014) study the finite sample properties when A3 is violated.

CGM suggest that with fewer than 30 clusters, the block bootstrap rejection rate is too large. CGM propose the wild cluster bootstrap, a variant of the wild bootstrap due to Wu (1986).[5] The wild cluster bootstrap has many desirable features: each and every observation in the original dataset is in each bootstrap sample exactly once, and the structure of the error correlation within clusters is preserved. The procedure for the wild cluster bootstrap is as follows. First consider the OLS regression model:

$$Y_{ig} = \beta_0 + \beta_1 X_{ig} + u_{ig}. \tag{1}$$

Imagine we are interested in calculating a wild cluster bootstrap-t p-value for β_1 in equation (1). We can construct the p-value by first estimating the t-statistic, $\hat{t}$, in the original sample using cluster-robust standard errors. We then re-estimate the equation by imposing the null hypothesis, to obtain the restricted estimates $\tilde{\beta}_0$, $\tilde{\beta}_1$, $\tilde{u}_{ig}$. Then B iterations, or bootstraps, are performed. In each iteration a bootstrap sample is generated from the bootstrap data generating process

$$y^*_{ig} = \tilde{\beta}_0 + \tilde{\beta}_1 X_{ig} + \tilde{u}_{ig} v_g, \tag{2}$$

where the i^{th} residual in group g, $\tilde{u}_{ig}$, is multiplied by the bootstrap weight v_g. In general the bootstrap DGP should impose the null hypothesis, which in this case would mean setting $\tilde{\beta}_1 = 0$.

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_{ig} to each observation. The bootstrap weight can take many forms. 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:

$$\hat{p}^*(\hat{t}) = 2 \min \left(\frac{1}{B} \sum_{j=1}^{B} I\left(t^*_j \leq \hat{t}\right), \frac{1}{B} \sum_{j=1}^{B} I\left(t^*_j > \hat{t}\right) \right), \tag{3}$$

where $I(.)$ is the standard indicator function.[6]

[5] MacKinnon and Webb (2014) propose using the wild cluster bootstrap for clusters of unequal size. Hagemann (2014) proposes a wild cluster bootstrap for quantile regression.

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

Simplifying, the DGP generates unique bootstrap samples solely as a function of the transformed residuals. This procedure is based on the assumption that B bootstrap samples, are drawn from an extremely large pool of potential bootstrap samples. Inference on $\hat{\beta}$ depends on where $\hat{t}$ falls in the sorted vector of bootstrap t-statistics, $t^* = t_1^*, ..., t_B^*$. If our estimated t-statistic falls between the 90th and 91st bootstrap t-statistic, then the p-value of this t-statistic is 0.180. When there is a large number of potential samples, the generated set of bootstrap samples will contain few, if any, repeated samples. Accordingly, the location of $\hat{t}$ can be precisely identified, and the resulting p-value is point identified.

CGM present evidence that the wild cluster bootstrap-t method is preferable to several alternative bootstrap methods, and allows for reliable inference with as few as five clusters. However, when the number of clusters is small, so is the number of unique bootstrap samples for the method advocated by CGM. As I now show, this makes reliable inference difficult. With few clusters, the number of unique potential bootstrap samples is rather small, as bootstrap samples depend on the choice of a bootstrap weight distribution. Two well-known distributions are the Rademacher and the Mammen, both of which contain only two points. With these distributions, v_g from equation (2) is set to one of two values with a given probability, p.[7]

Cameron, Gelbach and Miller (2008) recommend the Rademacher weights, as do Davidson, Monticini and Peel (2007) and Davidson and Flachaire (2008). Accordingly, there are only 2^G possible bootstrap samples, where G is the number of groups. The number of unique absolute value t-statistics is only 2^{G-1}, see Appendix A for a proof of this result. When G is large, this is not a problem as the vector will contain mostly unique t-statistics. When G is small problems arise since $\hat{t}$ and the various other unique values of t_j^* will be found multiple times in the vector. For example, when G $= 5$ there are only 32 unique bootstrap samples; if $B = 999$ one will be drawing 999 samples from a set of only 32 unique samples.

The CGM procedure incorrectly treats the B t-statistics as B unique values. Having many repeated t-statistics leaves open the possibility that $\hat{t} = t^*$ multiple times. When 2^G is small we cannot perform conventional inference. This limitation comes as a result of the inability to point-identify where $\hat{t}$ falls within the sorted vector of bootstrap t-statistics. When using the Rademacher distribution, one of the possible bootstrap t-statistics, t_j^*, is equal to the t-statistic, $\hat{t}$. When 2^G is small, this

[7]The Rademacher distribution is defined as: $v_g = \pm 1$ with probability 0.5.

will be observed within the vector of t^*, almost surely. 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 $\hat{t}$. For example, if $B = 999$ and in 31 replications $t^* = \hat{t}$ then $\hat{t}$ would appear in the sorted vector 31 times, such as $\hat{t} = t^*_{70}, ..., t^*_{100}$. In this case, 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 for different numbers of clusters from Monte Carlo simulations. The figure shows that the p-values occupy a wide interval when there are few clusters, with the width shrinking as the number of clusters increases. This wide interval makes it quite difficult to assess significance at conventional levels.

The 2^{G-1} unique t-statistics map into 2^{G-1} unique p-values. With few clusters these p-values will be intervals. An appropriate inference procedure should result in 2^{G-1} unique p-values across Monte Carlo simulations. CGM instead chose to estimate the p-value as being point identified at the midpoints of these intervals. In simulations discussed later in this paper, the CGM procedure with $G = 5$ resulted in 199 p-values across 50,000 replications rather than the 16 unique p-values.

When using a bootstrap method where the empirical distribution of t^* has few elements, one can calculate these elements systematically, or enumerate them, rather than trying to estimate the distribution through resampling.[8] However, inference using only 2^{G-1} t-statistics will be limited. The enumeration procedure for estimating a p-value is quite similar to the wild cluster bootstrap procedure discussed above. While the wild bootstrap picks v_g at random from a distribution, enumeration selects v_g methodically. A benefit when 2^G is small is that it is feasible to calculate all possible t-statistics. After G is sufficiently large, say 12, the computational burdens make full enumeration unattractive. Similarly, with very small values of G one could enumerate all the p-values resulting from the 6-point distribution proposed in section 3. This is considered in the empirical application in section 5.

The p-value from this procedure is different than a conventional p-value. These p-values are drawn from a discrete, as opposed to a continuous, distribution where the p-value belongs to the set $\{1/2^{G-1}, 2/2^{G-1} \ldots, (2^{G-1}-1)/2^{G-1}, 2^{G-1}/2^{G-1}\}$. For example, if $|\hat{t}| = |t^*_2|$, with $G = 5$ the p-value is 2/6, but not 0.125, since it could have alternatively been 1/16 or 3/16. The discrete nature of these p-values makes conven-

[8]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.

tional significance levels less meaningful. The p-value of 2/6 spans from $0.0625-0.125$ and so straddles the 10% level. With enumeration, inference is based on the data and the properties of the bootstrap weight distribution, and not on resampling noise. Enumeration will generate unique t-statistics; however, the limitation of having only 2^{G-1} t-statistics from which to conduct inference leaves much to be desired.

3 Alternative Bootstrap Methods

It is possible to find an alternative bootstrap weight distribution which expands the number of points. I propose a 6-point distribution, which mimics features of the Rademacher distribution.[9] The first four moments of the 'ideal' distribution are 0,1,1,1. It is not possible to satisfy all of these moment conditions.[10] The Rademacher distribution has the first four moments of 0,1,0,1 and the Mammen has the first four moments of 0,1,1,2. The candidate distribution will be symmetric, have the first three moments of 0,1,0, and a fourth moment not much larger than 1.[11] The candidate 6-point distribution I consider is:

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

The fourth moment of this distribution is 7/6.

There exists the temptation to add additional points to the distribution to increase the potential number of bootstrap samples, but there are two concerns about doing so. The weights will ideally be distinct from one another, as weights of 0.99 and 1.01 will result in very similar bootstrap samples and t-statistics. Given this desire, and restrictions on the first two moments, the inclusion of additional points will often increase the fourth moment. As a limiting case, I consider using the Normal distribution where $v_g \sim N(0,1)$. Drawing from the Normal allows for infinite possible

[9] Previous simulations, not included in this paper, have shown the Rademacher distribution to be preferable to the Mammen distribution.

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

[11] It is not possible to match the first four moments of the Rademacher distribution, if we impose a restriction that two of the points are 1 and -1. The candidate distribution will then have 6-points of the form $-A, -1, -B, B, 1, A$ each selected with equal probability. The imposition of symmetry means that the first and third moments will be 0. 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 $0, 1, \text{or} -1$, which does not result in a 6-point distribution.

bootstrap samples.[12]

The main benefit of adding additional points to the bootstrap weight distribution is that the number of potential bootstrap samples increases exponentially. For instance, in moving from a 2-point distribution to a 6-point distribution the number of bootstrap samples increases from 2^G to 6^G, or from 32 to 7776 when $G = 5$. The symmetry of the distribution results in the number of unique absolute value t-statistics being $(6^G)/2$. Monte Carlo simulations using this method with $G = 5$ resulted in 200 unique p-values across 50,000 replications. In contrast to the 2-point results, these p-values are a result of unique t-statistics.

4 Monte Carlo Evidence

4.1 Description of Simulations

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

$$\begin{gathered} y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \\ \text{or} \\ y_{ig} = \beta_0 + \beta_1 (z_g + z_{ig}) + (\epsilon_g + \epsilon_{ig}), \end{gathered} \tag{5}$$

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 y_{ig}, I set $\beta_1 = 1$ and test the hypothesis that $\beta_1 = 1$.[13]

[12] Mammen (1993) considered two continuous distributions that are not considered in this paper since simulation results in MacKinnon (2014) show them to be inferior to the Normal distribution. Liu (1988) proposes two continuous distributions, one using draws from a gamma distribution and the other using a mixture of normals. These were both chosen to satisfy the third moment restriction, thus the fourth moment must ≥ 2, see MacKinnon (2014). For this reason these distributions are not considered in this paper.

[13] I base my simulations off code provided by Douglas Miller, which is available at: `http://www.econ.ucdavis.edu/faculty/dlmiller/statafiles/bs_example.do`

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. $\hat{\alpha}$ is then compared to the true size of the test, $\alpha = 0.05$.

In total, seven different rejection rates are compared across a variety of asymptotic and bootstrap procedures. Table 1 describes the simulations. Designs 1-3 use asymptotic procedures and designs 4-7 use bootstrap procedures. Design 1 uses t-statistics calculated with OLS standard errors, while design 2 uses CRVE standard errors. Both assume the t-statistics are distributed normally. Design 3 uses CRVE standard errors, but the t-statistics are assumed to follow a t-distribution with G-1 degrees of freedom.[14] Design 4 generates p-values using the wild cluster bootstrap with v_g drawn from the 2-point Rademacher distribution, the test recommended by CGM.[15] Design 5 generates p-values with v_g drawn from $N(0,1)$. Design 6 uses v_g drawn from the 6-point distribution proposed in equation (4). Finally, design 7 generates p-values by enumerating the Rademacher wild bootstrap t-statistics, for $G \leq 10$. The results of the Monte Carlo experiments are discussed below.

4.2 Simulation Results

Table 2 presents results for $G = 5$ to $G = 10$ in the top panel and results for $G = 15$ to $G = 30$ in the bottom panel.[16] The panels show the severe problem of ignoring the clustered nature of the data. Tests using OLS standard errors give rejection rates of nearly 50%. Calculating CRVE standard errors and using tests 2 and 3 performs much better. Assuming that the t-statistics are normally distributed results in severe overrejection when there are very few clusters. The assumption that the t-statistics follow a t-distribution with $G-1$ degrees of freedom substantially improves the size of the test, but overrejection still occurs with very few clusters. For $G = 5$ to 10 the rejection rates for the wild cluster bootstrap-t with Rademacher

[14] This distribution is both recommended by Donald and Lang (2007) and Bester, Conley and Hansen (2011), and is the default within the Stata 'cluster' command.

[15] CGM use the average value for which $\hat{t} = t^*$, while I use the max value. The difference is negligible when G is large, but significant when G is small, see figure 1.

[16] The simulation standard errors, which range from 0.0008 to 0.0022, are not shown to save space.

weights appear deceivingly reliable. The empirical application in section 5 shows that the CGM procedure can result in a p-value that is quite different than the underlying enumerated p-value. The results for $G \geq 15$ do not suffer from this problem.

The top panel of Table 2 shows the results of simulations in which the number of clusters is small. The wild bootstrap with Normal weights performs fairly well, though it is outperformed in most cases by the 6-point wild bootstrap. The 6-point distribution works well. Even when $G = 5$, it is only moderately oversized with a rejection rate of 0.070, versus 0.097 for CRVE with the $T(4)$ distribution.

As mentioned previously, the enumerated p-values are interval rather than point identified. 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 illustrated in figure 1. When $G = 5$ the upper bound never rejects at the 5% level since 1/16 is above that threshold. The lower bound rejects far too often. The upper and lower bound rejection frequencies converge as G increases. Even with 10 clusters the lower bound enumeration rejection frequencies are higher than those from the 6-point distribution.

The bottom panel of table 2 shows the results when the number of clusters ranges from 15 to 30. In general, the various bootstrap methods work better than the analytic methods. The wild bootstrap with Normal weights is outperformed by both the 2-point and 6-point wild bootstrap. The wild bootstrap with the 6-point distribution dominates the wild bootstrap with the 2-point distribution in most cases.

5 Empirical Application

This section evaluates the effectiveness of a set of public finance experiments, known as graduate retention programs, to illustrate the practical application of the methods developed. Beginning in approximately 2006, these policies offered generous tax credits to new graduates with post-secondary degrees. The credits were conditional on residency within a given province. This section analyses the impact of these programs within the Atlantic Provinces of Canada.[17]

Although the programs were designed solely to retain graduates, I instead study the effects of the programs on a number of educational outcomes. The availability of such credits could affect the decision to enroll in post-secondary education, the

[17] Four provinces have introduced these programs: Nova Scotia, New Brunswick, Manitoba, and Saskatchewan. A more detailed analysis of these programs can be found in Webb (2013).

decision to drop out of post-secondary education, and the residency decision. The impact of these programs on these decisions is investigated using a linear difference-in-differences estimator within the Atlantic Provinces. These provinces are ideally suited for analysis as there are two treated provinces and two control provinces. The analysis is conducted using public use data from Statistics Canada's Labour Force Survey (LFS).[18]

The analyzed outcomes are University Graduate, College Graduate, University or College Graduate, University or College Dropout, University Student, College Student, and University or College Student.[19] The sample contains observations from individuals living in the Atlantic provinces for the years 2000-2012. For the graduation and dropout outcomes, the sample is restricted to those aged 22-29, while the sample is extended to those 17-29 for the student outcomes. The means of these variables for the various pre and post, treatment and comparison groups can be found in Table (3).

The estimation is conducted using a linear difference-in-differences equation, of the following form:

$$\begin{aligned} Y_{istm} = c + \beta_{GRP} * I[\text{ProvGRP}_{istm} * \text{YearGRP}_{istm}] \\ + \text{PROV}_s + \text{YEAR}_t + \text{MONTH}_m + \text{AGE}_{istm} + \epsilon_{istm}. \end{aligned} \tag{6}$$

The data varies along four dimensions, as the outcome variable Y_{istm} contains an observation for individual i, in province s, in year t, in month m. In the equation there is a set of province dummy variables PROV, a set of year dummy variables YEAR, a set of month dummy variables MONTH, and a set of age dummies AGE. The coefficient of interest is β_{GRP} which will capture the marginal impact for those individuals living in a province offering a credit, in a year in which a retention credit was available. This variable will be equal to one for individuals in Nova Scotia in 2006-2012 and individuals in New Brunswick in 2005-2012, and zero otherwise.

[18]The LFS is a monthly survey of over 50,000 Canadian households, on labour market and economic outcomes. Respondents are surveyed in waves, with each wave lasting for six months. Observations from a given month have responses from individuals in six overlapping waves.

[19]All of these outcomes are binary, and equal to one if the individual has obtained that level of education or is of that educational status. For example, *University Graduate* is set equal to one if the individual has graduated from university, and is set equal to zero otherwise. Similarly, *University Student* is set equal to one for individuals currently enrolled in a university program, and is set equal to zero otherwise. For the analysis, all variables are multiplied by 1000 to make the coefficients more comparable.

The estimates of equation (6) can be found in Table (4). The implication of these estimates depend on what procedure is used for inference. The table presents the estimated coefficients along with both asymptotic and bootstrap based p-values. The asymptotic p-values are based on procedures 1, 2, and 3 from the Monte Carlo simulations. The bootstrap p-values are based on procedures 4, 6, and 7 from the Monte Carlo simulations. Additionally, for both the Rademacher and 6-point distributions, both bootstrap p-values and enumerated p-values are calculated.[20]

If one were to use OLS standard errors, then one would infer that four out of the seven coefficients were statistically significant at the 5% level. Conversely, if one were to use the CRVE standard errors, one would infer that two were statistically significant using a Normal distribution, and only one was significant using a t-distribution. Finally, if one were to rely on the wild cluster bootstrap, only one coefficient is statistically significant at the 10% level.

The table also highlights the difference between the 2-point and the 6-point distributions. The table shows three p-values each for both the Rademacher distribution and the 6-point distribution, namely the bootstrap p-value (with $B = 999$) and the upper and lower bound of the enumerated p-value.[21] The Rademacher p-values show that with only four clusters the width of the intervals for the enumerated p-values are quite large. Conversely, with the 6-point distribution the width of the p-value intervals are quite small. With the 6-point distribution all of the bootstrap p-values are within one or two percentage points of the enumerated intervals. However, with the Rademacher distribution, in some cases the bootstrap p-value can be greater than seven percentage points away from the enumerated intervals. With few clusters the bootstrap approximates the empirical distribution quite well when using the 6-point distribution, but it fails to do so when using the Rademacher distribution.

The p-values for the *University Graduate* coefficient are particularly illuminating. If one were to rely on the Rademacher bootstrap p-value, then one would infer that this coefficient is not statistically significant at the 5% level. Conversely, if one were to rely on the 6-point bootstrap p-value, then one would infer that the coefficient is statistically significant at the 5% level. When using either the Rademacher or the 6-point distribution, both the bootstrap p-value and the enumerated p-value result

[20] As there are only four clusters in this example, it is sensible to enumerate the 6-point distribution.

[21] The number of bootstraps was chosen to be in line with what an empirical researcher might normally choose.

in the same inference. The enumerated Rademacher p-value spans the interval from 0%-12.5% and thus is not significant at the 5% level, while the enumerated 6-point p-value spans the interval from 4.2%-4.3% and thus is significant at the 5% level. The magnitude of the coefficient suggests that the share of university graduates declined by 11.229/1000, while the goal of the programs was to increase the share of graduates.

6 Conclusion

When data are grouped or clustered, reliable inference is a challenge. With a small number of clusters, rejection rates using cluster robust standard errors can be far too large. The wild cluster bootstrap technique works well when there are few clusters. When there are very few clusters, the 2-point Rademacher bootstrap weight distribution results in bootstrap p-values which are not point identified. With five clusters, the p-values are intervals with a width of 0.0625. I propose a new 6-point distribution which allows for improved inference with few clusters. Monte Carlo evidence suggests that this distribution works well with any number of clusters. An empirical application shows that with very few clusters the wild cluster bootstrap using the Rademacher distribution fails to approximate the empirical distribution, but the bootstrap approximates it quite well using the 6-point distribution.

References

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *Quarterly Journal of Economics* 119(1), 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.' *Review of Economics and Statistics* 90(3), 414–427

Cameron, A.C., and D.L. Miller (2014) 'A practitoner's guide to cluster robust inference.' *Journal of Human Resources.* forthcoming

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2013) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' Technical Report, University of California, Santa Barbara

Davidson, James, Andrea Monticini, and David Peel (2007) 'Implementing the wild bootstrap using a two-point distribution.' *Economics Letters* 96(3), 309–315

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

Donald, Stephen G, and Kevin Lang (2007) 'Inference with difference-in-differences and other panel data.' *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), 1–26

Hagemann, Andreas (2014) 'Cluster-robust bootstrap inference in quantile regression models.' *arXiv preprint arXiv:1407.7166*

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

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

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

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

MacKinnon, James G. (2014) 'Wild cluster bootstrap confidence intervals.' Working Paper 1329, Queen's University, Department of Economics

MacKinnon, James G., and Matthew D. Webb (2014) 'Wild bootstrap inference for wildly different cluster sizes.' Working Paper 1314, Queen's University, Department of Economics

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

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), 334–338

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

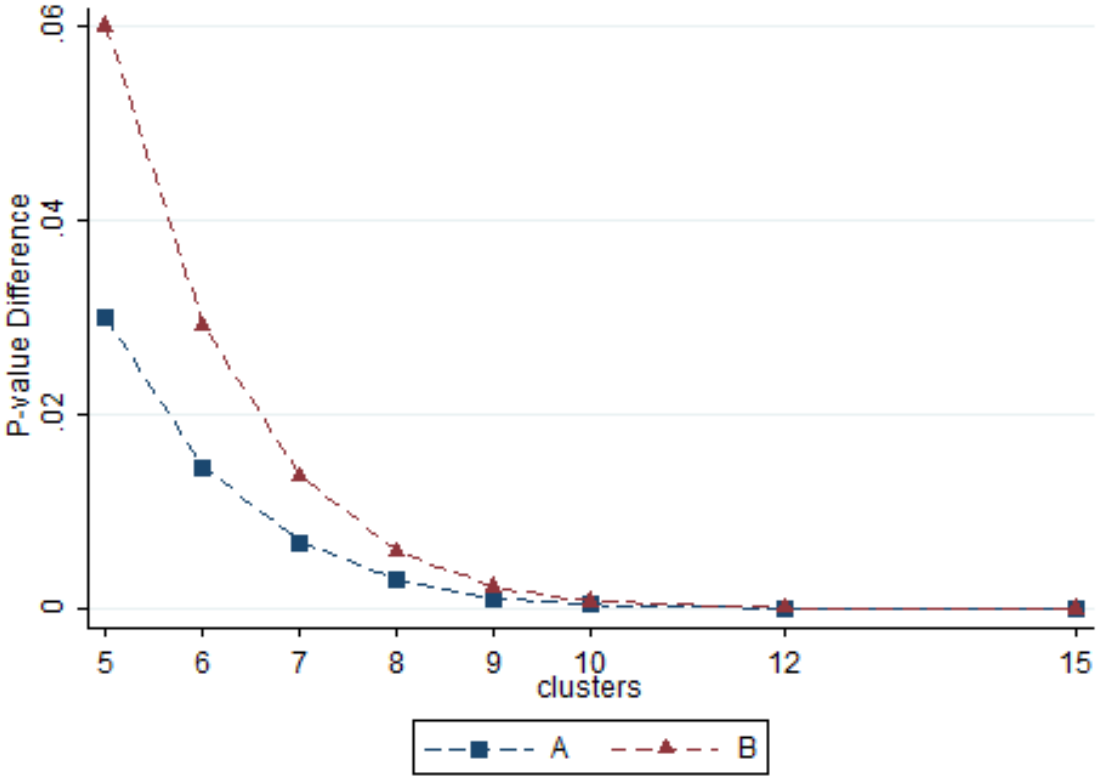
Thompson, Samuel B. (2011) 'Simple formulas for standard errors that cluster by both firm and time.' *Journal of Financial Economics* 99(1), 1–10

Webb, Matthew D. (2013) 'Econometric analysis of labour market interventions.' PhD dissertation, Queen's University

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

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

Figure 1: Estimated Differences From Three Different P-values



Notes: A is the difference between the maximum p-value and the mean p-value. B is the difference between the maximum p-value and the minimum p-value. Differences are calculated with 999 bootstraps using Rademacher weights.

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	-	Rademacher
5	Wild Cluster - Normal	CRVE	-	$\sim N(0,1)$
6	Wild Cluster - 6-point	CRVE	-	6-point equation (4)
7	Enumeration - Rademacher	CRVE	-	Rademacher

Table 2: Results from Monte Carlo Study with Different Numbers of Clusters

		Number of Groups (G)					
		5	6	7	8	9	10
1	OLS $\sim N(0,1)$	0.471	0.478	0.483	0.485	0.488	0.488
2	CRVE $\sim N(0,1)$	0.210	0.185	0.168	0.154	0.143	0.134
3	CRVE $\sim T(G-1)$	0.097	0.098	0.096	0.094	0.092	0.089
4	Wild 2pt Boot	*0.037	*0.053	*0.056	*0.056	*0.055	*0.054
5	Wild $N(0,1)$ Boot	0.072	0.070	0.072	0.072	0.071	0.069
6	Wild 6pt Boot	0.070	0.067	0.063	0.061	0.057	0.056
7	Enum. Lower Bound	0.118	0.095	0.084	0.068	0.062	0.060
7	Enum. Upper Bound	0.000	0.059	0.067	0.061	0.058	0.058
		Number of Groups (G)					
		15	20	25	30		
1	OLS $\sim N(0,1)$	0.489	0.495	0.490	0.496		
2	CRVE $\sim N(0,1)$	0.105	0.093	0.083	0.081		
3	CRVE $\sim t(G-1)$	0.080	0.075	0.069	0.070		
4	Wild 2pt Boot	0.050	0.050	0.047	0.048		
5	Wild $N(0,1)$ Boot	0.065	0.063	0.059	0.059		
6	Wild 6pt Boot	0.052	0.052	0.049	0.049		

Notes: Rejection frequencies estimated with 50,000 replications and 399 bootstraps (Boot). * - estimate is not accurately calculated.

Table 3: Variable Means from Labour Force Survey

	Provinces without GRP		Provinces with GRP		
	Pre	**Post**	**Pre**	**Post**	**Sample**
LFS					
University Graduate	15.9%	20.2%	20.0%	22.7%	22-29
College Graduate	40.7%	39.0%	37.6%	35.0%	22-29
University or College Graduate	57.4%	59.5%	58.2%	58.1%	22-29
University or College Dropout	10.2%	10.0%	9.9%	10.4%	22-29
University Student	14.7%	15.7%	13.5%	14.7%	17-29
College Student	6.9%	7.8%	5.1%	5.6%	17-29
University or College Student	21.6%	23.5%	18.5%	20.3%	17-29

Sample: LFS data from years 2000-2013, ages 17-29, unless otherwise noted.

Table 4: Coefficient Estimates and P-values from Various Procedures

		Asymptotic p-values		
		OLS	**CRVE**	**CRVE**
	coeff	$N(0,1)$	$N(0,1)$	$t(G-1)$
University Graduate	−11.229	0.002	0.000	0.028
College Graduate	−5.494	0.207	0.724	0.747
University or College Graduate	−15.708	0.000	0.231	0.317
University or College Dropout	9.088	0.001	0.034	0.124
University Student	−0.188	0.935	0.973	0.975
College Student	2.998	0.059	0.385	0.449
University or College Student	2.810	0.284	0.596	0.633

		Rademacher p-values		
	coeff	**boot**	**enum–L**	**enum–U**
University Graduate	−11.229	0.059	0.000	0.125
College Graduate	−5.494	0.677	0.625	0.750
University or College Graduate	−15.708	0.574	0.375	0.500
University or College Dropout	9.088	0.441	0.375	0.500
University Student	−0.188	0.936	0.875	1.000
College Student	2.998	0.532	0.375	0.500
University or College Student	2.810	0.828	0.625	0.750

		6-point p-values		
	coeff	**boot**	**enum–L**	**enum–U**
University Graduate	−11.229	0.039	0.042	0.043
College Graduate	−5.494	0.726	0.728	0.730
University or College Graduate	−15.708	0.524	0.523	0.525
University or College Dropout	9.088	0.401	0.421	0.423
University Student	−0.188	0.980	0.971	0.972
College Student	2.998	0.576	0.583	0.585
University or College Student	2.810	0.726	0.719	0.721

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

Recall that a bootstrap sample is generated by:

$$y_i^* = X\tilde{\beta} + \tilde{u}_i^*, \tag{7}$$

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 $\beta_i^*, \beta_j^* = \beta_o$ and calculate t-statistics of the form:

$$\frac{(X'X)^{-1}X'y_i - \beta_o}{\left(\frac{u_i'u_i}{X'X(n-k)}\right)^{1/2}}. \tag{8}$$

The denominator in equation (8) 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 (8), 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:

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

Since $v_i = -1 \circ v_j$, the numerator for the t-statistic of β_j^* will be the negative of the numerator for the t-statistic of β_i^*. The t-statistics are equal in absolute value, because the denominators are also the same. 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.