Queen's Economics Department Working Paper No. 1329

Wild cluster bootstrap confidence intervals

James G. MacKinnon
Queen's University

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

10-2014

Wild Cluster Bootstrap Confidence Intervals

James G. MacKinnon

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

jgm@econ.queensu.ca
http://www.econ.queensu.ca/faculty/mackinnon/

Abstract

Confidence intervals based on cluster-robust covariance matrices can be constructed in many ways. In addition to conventional intervals obtained by inverting Wald (t) tests, the paper studies intervals obtained by inverting LM tests, studentized bootstrap intervals based on the wild cluster bootstrap, and restricted bootstrap intervals obtained by inverting bootstrap Wald and LM tests. It also studies the choice of an auxiliary distribution for the wild bootstrap, a modified covariance matrix based on transforming the residuals that was proposed some years ago, and new wild bootstrap procedures based on the same idea. Some procedures perform extraordinarily well even when the number of clusters is small.

February 14, 2015

Research for this paper was supported, in part, by a grant from the Social Sciences and Humanities Research Council of Canada. I am grateful to Russell Davidson for valuable insights and to Matthew Webb and a referee for useful comments.

1. Introduction

It is now routine to employ cluster-robust standard errors whenever observations at the individual level are associated with a number of geographical areas and/or with a number of time periods. Each geographical area, or each time period, or perhaps each area-period pair, can be thought of as a cluster. When key regressors are measured at the cluster level, as is often the case when assessing the effects of policy changes, fixed effects cannot be used to account for intra-cluster correlation, because the fixed-effect dummy variables would explain all the variation in the regressor(s) of interest. Instead, it is common to use cluster-robust standard errors, because they allow for heteroskedasticity within and across clusters and also for intra-cluster correlation.

In large datasets, even very small levels of intra-cluster correlation can cause severe errors of inference if standard errors are not robust to clustering. For example, in a "placebo laws" experiment with over 500,000 observations on employment income data, where the average intra-cluster correlation coefficient is roughly 0.032, Bertrand, Duflo, and Mullanaithan (2004) find that using standard errors which are robust to heteroskedasticity but not to clustering yields rejection frequencies for interventions that did not actually take place which exceed 0.67 at the .05 level.

There has been a good deal of recent work on cluster-robust inference; see Cameron and Miller (2015) for a comprehensive survey. Much of this work has focused on testing, including bootstrap testing; see Cameron, Gelbach, and Miller (2008) and MacKinnon and Webb (2014). This paper focuses instead on confidence intervals. The next section discusses the conventional cluster-robust confidence interval, which is implicitly based on inverting a Wald test, and proposes a new interval based on inverting a Lagrange Multiplier, or LM, test. The latter is more computationally intensive than the former, but it should be quite feasible in most cases.

Section 3 then reviews the procedure for constructing a studentized bootstrap interval based on the wild cluster bootstrap. Section 4 proposes two new "restricted bootstrap" intervals, which are based on inverting bootstrap P values for Wald and LM tests, respectively. Unfortunately, these procedures are very computationally intensive. Section 5 describes the design of simulation experiments to compare the performance of the five intervals considered in the previous two sections, and Section 6 presents the experimental results.

The remainder of the paper deals with two different issues. Section 7 discusses the choice of an auxiliary distribution for the wild cluster bootstrap and presents some further experimental results. Section 8 discusses several ways in which cluster-robust confidence intervals can be improved by using transformed residuals in covariance matrices and/or wild bootstrap DGPs and presents additional simulation results. Finally, Section 9 concludes.

2. Cluster-Robust Confidence Intervals

Consider the linear regression model

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \vdots \\ \boldsymbol{u}_G \end{bmatrix}, \tag{1}$$

where the data are divided into G clusters, indexed by g. The g^{th} cluster has N_g observations, and the entire sample has $N = \sum_{g=1}^{G} N_g$ observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{u}$ have N rows, $\boldsymbol{X}$ has K columns, and the parameter vector $\boldsymbol{\beta}$ has K elements. Least squares estimation of equation (1) yields OLS estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{u}}$. The disturbances are assumed to be uncorrelated across clusters but potentially correlated and heteroskedastic within clusters, so that

$$\mathrm{E}(\boldsymbol{u}_g \boldsymbol{u}_g') = \boldsymbol{\Omega}_g, \quad g = 1, \ldots, G,$$

where the $N_g \times N_g$ covariance matrices $\boldsymbol{\Omega}_g$ are unknown. Thus the covariance matrix of $\boldsymbol{u}$ is assumed to be block diagonal.

Following Liang and Zeger (1986), the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be estimated by using a cluster-robust [co]variance matrix, or CRVE. The most widely-used CRVE is

$$\frac{G(N-1)}{(G-1)(N-K)} (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g' \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2}$$

This has the familiar sandwich form, with $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ serving as the bread and a summation of $K \times K$ matrices over all clusters serving as the filling. The matrix $\hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g'$ contains the squares and cross-products of all the residuals for cluster g. It evidently provides an inconsistent estimate of $\boldsymbol{\Omega}_g$. Nevertheless, $1/N$ times the sum of the $\boldsymbol{X}_g' \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g' \boldsymbol{X}_g$ matrices does consistently estimate the filling in the asymptotic covariance matrix, and N times the CRVE consistently estimates the covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. These results require that G tends to infinity with N at a fast enough rate; see Carter, Schnepel, and Steigerwald (2013).

The CRVE (2) resembles a standard heteroskedasticity-consistent covariance matrix, or HCCME. In fact, if $N_g = 1$ for all g and the factor of $G/(G-1)$ is omitted, it reduces to the HC_1 matrix of MacKinnon and White (1985). It will therefore be referred to as the CV_1 matrix. The first factor in (2) is a form of degrees of freedom correction. It is asymptotically negligible, but it always makes CV_1 larger when G and N are finite. When G is small, it can have a non-negligible effect.

When the CV_1 matrix is used to compute t statistics, it is common to base inferences on the $t(G-1)$ distribution; see Donald and Lang (2007) and Bester, Conley, and Hansen

(2011). However, hypothesis tests based on this distribution tend to overreject when G is small, especially when the N_g vary substantially across clusters; see MacKinnon and Webb (2014). This suggests that conventional confidence intervals will tend to undercover.

In order to focus on the confidence interval for one parameter, say the k^{th}, we can partition the parameter vector $\boldsymbol{\beta}$ into a scalar β_k and a $(K-1)$–vector $\boldsymbol{\beta}_1$. The most commonly used $(1-\alpha)\%$ confidence interval for β_k is

$$\big[\hat{\beta}_k - c_{1-\alpha/2}\,\text{se}(\hat{\beta}_k),\ \ \hat{\beta}_k + c_{1-\alpha/2}\,\text{se}(\hat{\beta}_k)\big], \tag{3}$$

where $\text{se}(\hat{\beta}_k)$ is the square root of the k^{th} diagonal element of the CV_1 matrix (2), and $c_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(G-1)$ distribution.

The confidence interval (3) is implicitly obtained by inverting a cluster-robust t test, which is really a Wald test. We could instead invert a Lagrange Multiplier test. In this case, the LM statistic can be computed by using the Frisch-Waugh-Lovell, or FWL, regression

$$\boldsymbol{M}_1(\boldsymbol{y} - \beta_k^0 \boldsymbol{x}_k) = \boldsymbol{M}_1 \boldsymbol{x}_k b_k + \text{residuals}, \tag{4}$$

where β_k^0 is a candidate value of β_k, $\boldsymbol{X} \equiv [\boldsymbol{X}_1 \ \ \boldsymbol{x}_k]$, and $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. The regressand in (4) is the vector of residuals from regressing $\boldsymbol{y} - \beta_k^0\boldsymbol{x}_k$ on $\boldsymbol{X}_1$, and the regressor is the vector of residuals from regressing $\boldsymbol{x}_k$ on $\boldsymbol{X}_1$. It is easy to show that the test statistic $\text{LM}(\beta_k^0)$ can be written as

$$\frac{(G-1)(N-K-1)}{G(N-1)}\big((\boldsymbol{y} - \beta_k^0\boldsymbol{x}_k)'\boldsymbol{M}_1\boldsymbol{x}_k\big)^2 \Big/ \left(\sum_{g=1}^{G}(\boldsymbol{M}_1\boldsymbol{x}_k)_g'\tilde{\boldsymbol{u}}_g\tilde{\boldsymbol{u}}_g'(\boldsymbol{M}_1\boldsymbol{x}_k)_g\right), \tag{5}$$

where $\tilde{\boldsymbol{u}}_g$ denotes the vector of restricted residuals for cluster g, that is, the elements of the vector $\boldsymbol{M}_1(\boldsymbol{y} - \beta_k^0\boldsymbol{x}_k)$ that correspond to cluster g. Similarly, $(\boldsymbol{M}_1\boldsymbol{x}_k)_g$ denotes the rows of the vector $\boldsymbol{M}_1\boldsymbol{x}_k$ that correspond to cluster g. Expression (5) is just the square of the scalar $(\boldsymbol{y} - \beta_k^0\boldsymbol{x}_k)'\boldsymbol{M}_1\boldsymbol{x}_k$ divided by a cluster-robust estimate of its variance.

In order to obtain a confidence interval, we need to invert the test statistic (5). That is, we need to find the set of values of β_k^0 which satisfy the inequality

$$\text{LM}(\beta_k^0) \leq c_{1-\alpha}^F,$$

where $c_{1-\alpha}^F$ denotes the $1-\alpha$ quantile of the $F(1, G-1)$ distribution.[1] This needs to be done numerically. However, because the problem is one-dimensional and $\text{LM}(\beta_k^0)$ is smooth, it should not be very difficult. The resulting interval will have the form $[\beta_k^l, \ \ \beta_k^u]$, where

$$\text{LM}(\beta_k^l) = \text{LM}(\beta_k^u) = c_{1-\alpha}^F. \tag{6}$$

[1] Asymptotically, it would also be valid to use the $1-\alpha$ quantile of the $\chi^2(1)$ distribution.

Unlike the Wald interval (3), the LM interval defined by (6) will generally not be symmetric around $\hat{\beta}_k$

The Wald interval (3) can be expected to provide reliable inferences whenever Wald test statistics based on the CV_1 matrix (2) do so. However, several studies, including Cameron, Gelbach, and Miller (2008) and MacKinnon and Webb (2014), suggest that this will generally not be the case when G is small and/or the N_g vary substantially across clusters. In these cases, the Wald interval is likely to undercover. Whether the LM interval (6) will perform better in such cases is an open question. In linear regression models, LM test statistics are often numerically smaller than corresponding Wald statistics; see Breusch (1979). Even if such an inequality does not hold strictly in this case, it seems very likely that the LM intervals will be longer, and therefore less prone to undercover, than the Wald intervals.

3. The Wild Cluster Bootstrap

The wild bootstrap was proposed in Liu (1988) based on a suggestion in Wu (1986). Key papers include Mammen (1993) and Davidson and Flachaire (2008). An extension to clustered data was suggested in Cameron, Gelbach, and Miller (2008) in the context of hypothesis testing. Simulation evidence in that paper and in MacKinnon and Webb (2014) have shown that the wild cluster bootstrap can provide remarkably accurate inferences in cases where cluster-robust t statistics can overreject severely.

The idea of the wild cluster bootstrap is very simple. For the ordinary wild bootstrap, the residual associated with each observation is multiplied by an auxiliary random variable that has mean 0 and variance 1. For the wild cluster bootstrap, the residuals associated with all the observations in a given cluster are multiplied by the same auxiliary random variable. This ensures that the bootstrap DGP mimics both the intra-cluster correlations and the heteroskedasticity of the residuals.

There are at least two ways in which the wild cluster bootstrap can be used to construct $(1 - \alpha)\%$ confidence intervals. The most widely used and computationally efficient approach is to construct a "studentized bootstrap" interval. This works as follows:

1. Estimate equation (1) by OLS to obtain estimates $\hat{\boldsymbol{\beta}}$, residuals $\hat{\boldsymbol{u}}$, and the cluster-robust standard error $\text{se}(\hat{\beta}_k)$.
2. Calculate $\hat{t}_k = \hat{\beta}_k/\text{se}(\hat{\beta}_k)$, the t statistic for $\beta_k = 0$.
3. For each of B bootstrap replications, indexed by j, generate a new set of bootstrap dependent variables $\boldsymbol{y}_g^{*j}$ using the bootstrap DGP

$$\boldsymbol{y}_g^{*j} = \boldsymbol{X}_g\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{u}}_g v_g^{*j}, \quad g = 1, \ldots, G, \tag{7}$$

   where $\boldsymbol{y}_g^{*j}$ is the vector of observations on the bootstrap dependent variable for cluster g, and v_g^{*j} is a random variable drawn from an auxiliary distribution with mean 0 and variance 1. A good choice for the latter is usually the Rademacher

distribution, which takes the values 1 and -1 with equal probability; see Davidson and Flachaire (2008). Other choices are discussed in Section 7.

4. For each bootstrap replication, estimate regression (1) using $\boldsymbol{y}^{*j}$ as the regressand, and calculate t_k^{*j}, the t statistic for $\beta_k = \hat{\beta}_k$, using the square root of the k^{th} diagonal element of (2), with bootstrap residuals replacing the OLS residuals, as the standard error.

5. Sort the t_k^{*j} from smallest to largest, and denote by $c^*_{\alpha/2}$ and $c^*_{1-\alpha/2}$, respectively, the $(B+1)(\alpha/2)^{\text{th}}$ and $(B+1)(1-\alpha/2)^{\text{th}}$ entries in the sorted list. For these indices to be integers, B must have been chosen so that $(B+1)(\alpha/2)$ is an integer. Natural choices are $B = 999$ and $B = 9{,}999$.

6. Construct the $(1-\alpha)\%$ studentized bootstrap interval as

$$\big[\hat{\beta}_k - \text{se}(\hat{\beta}_k)\, c^*_{1-\alpha/2},\;\; \hat{\beta}_k - \text{se}(\hat{\beta}_k)\, c^*_{\alpha/2}\big]. \tag{8}$$

Studentized bootstrap confidence intervals are widely used. See Davison and Hinkley (1997, Chapter 5) and Davidson and MacKinnon (2004, Chapter 5) for introductory expositions. The key difference between the studentized bootstrap interval (8) and the Wald interval (3) is that the $1-\alpha/2$ quantile of the $t(G-1)$ distribution is replaced by either the $1-\alpha/2$ quantile or the $\alpha/2$ quantile of the bootstrap distribution. Because the interval (8) uses two different quantiles, it will in general be asymmetric.

4. Bootstrap Intervals that Impose the Null

The bootstrap DGP (7) does not impose the null hypothesis. Because doing so makes the estimates more efficient, it is generally a good idea to impose the null whenever possible; see Davidson and MacKinnon (1999). In the context of a confidence interval, however, imposing the null is computationally demanding. There are two null hypotheses that correspond to the two ends of the interval, and neither of them is known initially. Thus an iterative procedure is necessary. However, the computational cost may be worth it, because there are circumstances in which such a "restricted bootstrap Wald interval" can work very much better than a studentized bootstrap interval; for an extreme example, see Davidson and MacKinnon (2014).

The procedure for constructing a restricted bootstrap Wald interval is similar to the one for the LM interval of Section 2 and is not difficult to describe. Step 1 is unchanged from the first step for the studentized bootstrap interval. The procedure for determining the upper limit β_k^u then continues as follows:

2. Pick a candidate upper limit, say $\beta_k^{u\dagger}$. This might be the upper limit of either the Wald interval (3) or the studentized bootstrap interval (8). Then calculate $\hat{t}_k^{u\dagger} = (\hat{\beta}_k - \beta_k^{u\dagger})/\text{se}(\hat{\beta}_k)$, the t statistic for $\beta_k = \beta_k^{u\dagger}$.

3. Calculate the residual vector

$$\tilde{\boldsymbol{u}}^\dagger \equiv \boldsymbol{M}_1(\boldsymbol{y} - \beta_k^{u\dagger}\boldsymbol{x}_k) = \boldsymbol{M}_1\boldsymbol{y} - \beta_k^{u\dagger}\boldsymbol{M}_1\boldsymbol{x}_k. \tag{9}$$

These are the residuals from a regression of $\boldsymbol{y} - \beta_k^{u\dagger}\boldsymbol{x}_k$ on $\boldsymbol{X}_1$.

4. Generate B bootstrap samples using the bootstrap DGP

$$\boldsymbol{y}_g^{*j} = \beta_k^{u\dagger}\boldsymbol{x}_{kg} + \tilde{\boldsymbol{u}}_g^{\dagger} v_g^{*j}, \quad g = 1, \ldots, G, \tag{10}$$

where $\tilde{\boldsymbol{u}}_g^{\dagger}$ is a subvector of $\tilde{\boldsymbol{u}}^{\dagger}$. The right-hand side of equation (10) could also include $\boldsymbol{X}_{1g}\tilde{\boldsymbol{\beta}}_1^{\dagger}$, where $\tilde{\boldsymbol{\beta}}_1^{\dagger}$ denotes the estimates of $\boldsymbol{\beta}_1$ conditional on $\beta_k = \beta_k^{u\dagger}$, but there is no need to include that term, because doing so would not change the bootstrap test statistic.

5. For each bootstrap sample, calculate the bootstrap test statistic t_k^{*j} for the hypothesis that $\beta_k = \beta_k^{u\dagger}$ by regressing $\boldsymbol{y}^{*j} - \beta_k^{u\dagger}\boldsymbol{x}_k$ on $\boldsymbol{X}$, using the CV_1 matrix (2) to calculate the standard error of $\hat{\beta}_k^{*j}$.

6. Calculate the equal-tail bootstrap P value

$$\hat{p}^{\dagger} \equiv \hat{p}^*(\beta_k^{u\dagger}) = 2\min\left(\frac{1}{B}\sum_{j=1}^{B} I\big(t_k^{*j} \leq \hat{t}_k^{u\dagger}\big), \frac{1}{B}\sum_{j=1}^{B} I\big(t_k^{*j} > \hat{t}_k^{u\dagger}\big)\right), \tag{11}$$

where $I(\cdot)$ denotes the indicator function, which equals 1 if its argument is true and 0 otherwise.

7. If $\hat{p}^{\dagger} < \alpha$, the candidate upper limit $\beta_k^{u\dagger}$ must be too large. If $\hat{p}^{\dagger} > \alpha$, it must be too small. Repeating steps 2 through 6 as many times as necessary, search over β_k^u using a root-finding algorithm that does not require derivatives, such as bisection, until it finds a value β_k^{u*} such that $\hat{p}^*(\beta_k^{u*}) = \alpha$. This is the upper limit of the confidence interval.

The procedure for finding the lower limit is almost identical. First, pick a candidate lower limit, say $\beta_k^{l\dagger}$. Then repeat steps 2 through 6 with appropriate modifications. If $\hat{p}^{\dagger} < \alpha$, the candidate lower limit $\beta_k^{l\dagger}$ must be too small. If $\hat{p}^{\dagger} > \alpha$, it must be too large. Use the root-finding algorithm again to find a value β_k^{l*} such that $\hat{p}^*(\beta_k^{l*}) = \alpha$. This is the lower limit of the confidence interval.

As with all simulation-based optimization procedures, it is essential that the same random numbers be used for each set of B bootstrap samples. Otherwise, the root-finding algorithm would probably never converge.

Instead of forming a confidence interval by inverting a bootstrap Wald test, we could form one by inverting a bootstrap test based on the LM statistic (5). The procedure for constructing this "restricted bootstrap LM interval" is very similar to the one for the restricted bootstrap Wald interval. In this case, both the test statistic itself and the bootstrap samples are conditional on the upper and lower limits of the interval. Thus step 1 is omitted. The remainder of the algorithm proceeds as follows:

2. Given a candidate upper limit $\beta_k^{u\dagger}$, use expression (5) to compute the test statistic $\text{LM}(\beta_k^{u\dagger})$. Optionally, convert it into a signed statistic by taking the signed square root of expression (5).

3. Use equation (9) to compute the residual vector $\tilde{\boldsymbol{u}}^\dagger$.
4. Generate B bootstrap samples using equation (10).
5. For each bootstrap sample, calculate the bootstrap test statistic LM_k^{*j}, using the same procedure as in step 2. Optionally, convert it into a signed statistic.
6. Calculate the upper-tail bootstrap P value

$$\hat{p}^\dagger \equiv \hat{p}^*(\beta_k^{u\dagger}) = \frac{1}{B}\sum_{j=1}^{B} I\big(\mathrm{LM}_k^{*j} > \mathrm{LM}(\beta_k^{u\dagger})\big). \tag{12}$$

   If using signed statistics, calculate an equal-tail bootstrap P value, similar to (11), instead of (12).
7. Use a root-finding algorithm to find β_k^{u*}, as before.
8. Repeat steps 2 through 7, with appropriate modifications, to find the lower limit β_k^{l*}, as before.

When $\hat{\beta}_k$ is expected to be unbiased, there is no reason to convert the LM statistic into a signed statistic. However, when $\hat{\beta}_k$ is likely to be biased, as may well be the case when, for example, instrumental variables are being used to correct for endogeneity, doing so can make a substantial difference.

5. Design of the Experiments

The simulation experiments investigate the model

$$y_{gi} = \beta_1 + \beta_2 d_{gi} + \beta_3 D_{gi} + \beta_4 d_{gi} D_{gi} + u_{gi}, \;\; g = 1, \ldots, G, \;\; i = 1, \ldots, N_g, \tag{13}$$

where $d_{gi} = 1$ if any of the observations in cluster g is treated, and $D_{gi} = 1$ if i corresponds to a time period in which there is treatment, which takes place for a constant fraction π of the observations in each treated cluster. Since an observation is actually treated if $d_{gi}D_{gi} = 1$, the coefficient of interest is β_4. The dummy variable d_{gi} is included to account for non-random effects that may characterize treated versus untreated clusters, and the dummy variable D_{gi} is included to account for non-random effects that may characterize the time periods in which treatment occurs.

The model (13) can be thought of as a "difference-in-differences" regression, in which some groups are never treated, so that $d_{gi} = 0$ for all i, and other groups are treated for some but not all time periods (the same ones for each treated cluster). If the data took the form of a balanced panel with one observation for each cluster and time period, then N_g would just be the number of time periods.

For simplicity, consider a balanced panel with two time periods, indexed by 1 and 2, and two groups, indexed by a and b, where group a is never treated and group b is treated in period 2. For group a, equation (13) implies that

$$\mathrm{E}(y_{ai}) = \beta_1 + \beta_3 D_i, \;\; i = 1, 2,$$

since $D_{ai} = D_{bi} \equiv D_i$ for both time periods. For this group, the difference between the conditional means for the two periods is

$$\mathrm{E}(y_{a2}) - \mathrm{E}(y_{a1}) = \beta_3(D_2 - D_1) = \beta_3. \tag{14}$$

For group b, equation (13) implies that

$$\mathrm{E}(y_{bi}) = \beta_1 + \beta_2 + \beta_3 D_i + \beta_4 D_i, \;\; i = 1, 2.$$

For this group, the difference between the conditional means for the two periods is

$$\mathrm{E}(y_{b2}) - \mathrm{E}(y_{b1}) = \beta_3 + \beta_4. \tag{15}$$

The difference between the difference for group b in (15) and the difference for group a in (14) is simply β_4. This explains why the estimate of that parameter in equation (13) can be thought of as a "difference-in-differences" estimate. For a more detailed discussion, see Angrist and Pischke (2009, Chapter 5).

In most of the experiments, $G = 20$ and $N = 1000$. However, the way in which the N observations are allocated to G clusters depends on a parameter $\gamma \geq 0$. Specifically,

$$N_g = \frac{N \exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)}, \;\; \text{for } g = 1, \ldots, G-1,$$

where N_g is truncated to the nearest integer, and $N_G = N - \sum_{g=1}^{G-1} N_g$. When $\gamma = 0$, the clusters are equal-sized, with $N_g = 50$ for all g. As γ increases, the cluster sizes become more and more unequal. The largest value of γ used in the experiments is 4.5, for which the smallest and largest values of N_g are 2 and 216, respectively.

The disturbances u_{gi} are homoskedastic, normally distributed, equicorrelated within clusters with correlation coefficient ρ, and independent across clusters. Although the value of ρ affects the results, it does so to a remarkably limited extent, with almost no observable effect for $0 \leq \rho \leq 0.5$. Since its value does not matter much, ρ is set to 0.2 for all the reported experiments.

One feature of the model (13) is that all the regressors are the same for every replication of a given experiment. This makes it possible to perform many of the computations just once, which greatly reduces the cost of the simulations. The Wald interval is extremely inexpensive to compute. The LM interval is much more expensive, but still quite cheap. The studentized bootstrap interval is somewhat expensive, and the two restricted bootstrap intervals are quite expensive.

It may seem surprising that neither N nor G is varied in the principal experiments. Increasing N would have almost no effect on the results, but it would raise computational costs substantially. Increasing G would make all the intervals perform better, but it

would not affect any conclusions about their relative performance. Some evidence on the latter point is provided in the next section.

Most experiments use 100,000 replications with $B = 999$. For the same reason that the power of bootstrap tests increases with B (see Davidson and MacKinnon, 2000), the length of bootstrap confidence intervals tends to decline (slightly) as B increases. It is therefore desirable not to use too small a value of B in the experiments. With 100,000 replications, the standard error of an estimated coverage level that is truly 0.95 is 0.00069. Because the same simulated data are used for all five intervals, however, the difference between any two estimated coverage levels is actually much smaller than this. In no case do simulation errors lead to results that are at all ambiguous.

6. Performance of Alternative Confidence Intervals

Figure 1 shows the coverage of the five intervals discussed in Sections 2, 3, and 4 at the nominal 0.95 level as functions of P, the fraction of clusters treated. In these experiments, $G = 20$, $N = 1000$, and $\gamma = 3$, which implies that the smallest cluster has 8 observations and the largest has 155. Clusters are always treated from smallest to largest, and the results would undoubtedly be different if any other ordering were used. The fraction of observations within each treated cluster that is treated, π, is 0.4. The actual number of treated clusters is obtained by truncation. For example, since $8 \times 0.4 = 3.2$, three observations are treated when $N_g = 8$.

The Wald interval (3) always undercovers, and it does so quite severely when P is large or small. This is what would be expected based on the results for t tests in MacKinnon and Webb (2014). In contrast, the LM interval defined by equations (6) always overcovers. No results for LM intervals are shown for $P < 0.20$ or $P > 0.85$, because, in those cases, there often appeared to be no finite solution to equations (6).

The studentized bootstrap interval (8) always performs better than the Wald interval. Like the latter, it always undercovers, but it does so to a very limited extent for intermediate values of P. When P is large or small, however, the undercoverage can be quite severe. In the extreme case of $P = 0.05$, in which just one cluster is treated, the Wald and studentized intervals cover the true value just 14.2% and 15.2% of the time, respectively. This case is not shown in order to avoid making the main part of the figure difficult to read.

The two restricted bootstrap intervals behave very similarly. They perform extremely well for moderate values of P, say $0.30 \leq P \leq 0.75$, but they undercover slightly for somewhat more extreme values, and they overcover severely for $P \leq 0.15$ and $P \geq 0.90$. Note that, if G had been larger, the range of excellent performance would have been wider. MacKinnon and Webb (2014) shows that bootstrap tests perform badly only when the number of treated clusters, PG, rather than P itself, is small or large. The results of that paper also apply here, since the restricted bootstrap intervals are obtained by inverting bootstrap tests.

Figure 2 shows coverage as a function of γ, holding P and π constant at 0.3 and 0.4, respectively. As expected, all the intervals perform less well as γ increases and the values of N_g consequently become more dispersed. With increasing γ, the Wald and studentized bootstrap intervals undercover more severely, and the LM interval overcovers more severely. The two restricted bootstrap intervals always undercover slightly, but they perform extremely well for all values of γ.

Figure 3 shows coverage as a function of π, the fraction of observations within treated clusters that is treated. Coverage improves sharply for both the Wald and LM intervals, and to a lesser extent for the studentized bootstrap interval, as π increases from 0.10 to 0.25, but there is little further improvement as π continues to increase.

Figure 4 shows what happens as both G and N increase together. In the figure, G takes the values $15, 18, 21, \ldots, 48$, and $N = 50G$. The value of P is $1/3$, so that the numbers of treated clusters are $4, 5, 6, \ldots, 16$. As we would expect, all the intervals perform better as G increases. In particular, the restricted bootstrap intervals perform almost perfectly for $G \geq 21$, and the studentized bootstrap interval performs very well for $G \geq 39$. In contrast, it appears that G would have to be very large indeed for the Wald and LM intervals to perform really well.

In the experiments reported on so far, the fraction π of observations that is treated is the same for all treated clusters. This assumption substantially simplifies the task of running the experiments and reporting the results, but it is somewhat restrictive. In order to see how much it matters, the experiments reported in Figure 1 were rerun with one major change. Instead of $\pi = 0.4$ for all clusters, odd-numbered clusters now have $\pi = 0.2$, and even-numbered clusters have $\pi = 0.6$. Note that, with this modification, equation (13) is no longer interpretable as a difference-in-differences model. Figure 5 shows the results.

At first glance, Figure 5 looks very much like Figure 1. Upon closer inspection, however, a number of differences become apparent. The studentized bootstrap interval now performs substantially worse, especially when P is small or large. So does the Wald interval, although its performance does not deteriorate as much. The LM interval can no longer be calculated for $P = 0.20$, and it now undercovers for some large values of P. The two restricted bootstrap intervals also perform a bit less well, especially for large values of P. Thus the assumption that the fraction of treated clusters is constant certainly matters. However, there is no indication that the results would change radically if this assumption were relaxed.

The results in Figures 1 through 5 should not have been surprising. Conventional Wald intervals always undercover, and they sometimes do so severely. In contrast, LM intervals usually overcover. Studentized bootstrap intervals always outperform the Wald intervals on which they are based, but their performance can be problematical, especially when G is not large and cluster sizes are quite dispersed. In contrast, the two restricted bootstrap intervals perform extremely well, and almost identically, except when PG, the number of treated clusters, is very small or very large. Thus,

even though these intervals are relatively difficult and expensive to compute, they are probably worth using in many cases.

7. Wild Bootstrap Auxiliary Distributions

In principle, a great many different auxiliary distributions could be used to generate the random variables v_g^* that play a key role in the bootstrap DGPs (7) and (10). These include the asymmetric two-point distribution proposed in Mammen (1993), which is probably the most widely used, and the Rademacher distribution proposed in Davidson and Flachaire (2008), which seems to be a much better choice in most cases.

For the wild bootstrap to work well, the residuals must provide good approximations to the unknown disturbances. Sometimes, the residuals are transformed in order to make the approximations better; see Section 8. Provided the approximations are in fact good, we want the disturbances in the bootstrap DGP to have mean zero and the same higher moments as the (possibly transformed) residuals. For that to be the case up to the fourth moment, the auxiliary distribution must satisfy the conditions

$$\mathrm{E}(v^*) = 0, \;\; \mathrm{E}(v^{*2}) = 1, \;\; \mathrm{E}(v^{*3}) = 1, \;\; \text{and} \;\; \mathrm{E}(v^{*4}) = 1. \tag{16}$$

Unfortunately, it is impossible for any distribution to satisfy these conditions.

To see why not, consider the outer product of the vector $[1 \;\; v^* \;\; v^{*2}]'$ with itself for a random variable v^* with expectation 0 and variance 1. This yields a 3×3 matrix with expectation

$$\mathrm{E}\begin{bmatrix} 1 & v^* & v^{*2} \\ v^* & v^{*2} & v^{*3} \\ v^{*2} & v^{*3} & v^{*4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu_3 \\ 1 & \mu_3 & \mu_4 \end{bmatrix}, \tag{17}$$

where μ_3 and μ_4 denote the third and fourth moments of v^*. The determinant of the matrix on the right-hand side of equation (17) must be nonnegative because the matrix is positive semidefinite. This determinant is $\mu_4 - \mu_3^2 - 1$. Therefore,

$$\mu_4 - \mu_3^2 \geq 1. \tag{18}$$

If $\mu_3 = 1$, equation (18) implies that $\mu_4 \geq 2$. Conversely, if $\mu_4 = 1$, it implies that $\mu_3 = 0$. Thus there exists no distribution of v^* which satisfies conditions (16). This means that there is no ideal auxiliary distribution. We either need to relax the requirement that $\mu_3 = 1$ or allow $\mu_4 \geq 2$.

Since it takes the values 1 and -1 with equal probability, it is easy to see that the Rademacher distribution has $\mu_3 = 0$ and $\mu_4 = 1$. Thus it satisfies three of the four conditions in (16). However, because its third moment is zero, it imposes symmetry on the bootstrap disturbances.

Mammen (1993) suggests the two-point distribution

$$v^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{with probability } (\sqrt{5}+1)/(2\sqrt{5}), \\ (\sqrt{5}+1)/2 & \text{with probability } (\sqrt{5}-1)/(2\sqrt{5}). \end{cases} \tag{19}$$

This distribution satisfies the first three conditions in (16), but it has $\mu_4 = 2$. Thus it violates the last condition, although it does come as close to satisfying it as any distribution with $\mu_3 = 1$ can, because the inequality (18) holds as an equality.

Davidson and Flachaire (2008) provides evidence that the Rademacher distribution is a better choice than Mammen's two-point distribution (19) even when the disturbances are not symmetric. Davidson, Monticini, and Peel (2007) considers a class of two-point distributions of which Rademacher and (19) are special cases. In experiments with disturbances that are heavily skewed and severely heteroskedastic, the Rademacher distribution clearly outperforms Mammen's distribution (19) and all the others considered. Thus it appears that having a fourth moment equal to 1 is more important for an auxiliary distribution than having a third moment equal to 1.

With a two-point distribution, each observation can have only two bootstrap disturbances associated with it. In the cluster case, this means that there are only 2^G possible bootstrap samples. When G is small (say, less than 12) this can cause serious problems, as Webb (2013) points out. That paper therefore suggests an auxiliary distribution with six mass points,

$$-\sqrt{1.5},\ -1,\ -\sqrt{0.5},\ \sqrt{0.5},\ 1,\ \sqrt{1.5},$$

each of which has probability 1/6. It is easy to see that:

$$\mathrm{E}(v^*) = 0,\ \ \mathrm{E}(v^{*2}) = 1,\ \ \mathrm{E}(v^{*3}) = 0,\ \text{ and } \mathrm{E}(v^{*4}) = 7/6.$$

Because 6^G is very much larger than 2^G, the six-point distribution can safely be used even when G is very small. Its only disadvantage, relative to Rademacher, is that the fourth moment is slightly higher than 1.

Of course, it is not essential to limit ourselves to auxiliary distributions with a finite number of mass points. Since the standard normal distribution has mean 0 and variance 1, it may seem to be a natural choice for the distribution of v^*. However, $\mu_3 = 0$ and $\mu_4 = 3$, so that the standard normal violates two of the conditions in (16). It violates the last condition much more severely than the six-point distribution does.

Another continuous distribution with the correct mean and variance is the uniform distribution

$$\mathrm{U}\left[-\sqrt{3}, \sqrt{3}\right],$$

which has $\mu_3 = 0$ and $\mu_4 = 1.8$. It also violates two of the conditions in (16), but it violates the fourth-moment condition less severely than the standard normal does.

In addition to the two-point distribution, Mammen (1993) suggests the continuous distribution

$$v^* = u/\sqrt{2} + \tfrac{1}{2}(w^2 - 1),$$

where u and w are independent standard normal random variables. It can be shown that

$$\mathrm{E}(v^*) = 0, \;\; \mathrm{E}(v^{*2}) = 1, \;\; \mathrm{E}(v^{*3}) = 1, \;\; \text{and} \;\; \mathrm{E}(v^{*4}) = 6.$$

Thus the first three moments satisfy conditions (16), but the fourth moment is very much larger than 1.

None of the simulation evidence in Davidson and Flachaire (2008) and Davidson, Monticini, and Peel (2007) concerns the wild cluster bootstrap. I therefore investigate studentized bootstrap intervals using the six auxiliary distributions discussed above. Because the differences among the auxiliary distributions may be quite small, all experiments use 400,000 replications. This is feasible because studentized bootstrap intervals are much less expensive to compute than restricted bootstrap intervals.

Figure 6 shows the coverage of studentized bootstrap intervals at the nominal 0.95 level as functions of G for $G = 9, 12, 15, \ldots, 30$, with $N = 50G$, $P = 1/3$, and $\pi = 0.4$. It is similar to Figure 4, except that $G = 9$ and $G = 12$ are added, because the failures of the algorithm for the LM interval for those cases are now irrelevant, and values of G greater than 30 are omitted. The results are striking. For every value of G, the Rademacher distribution yields the most accurate coverage, followed closely by Webb's six-point distribution. Surprisingly, this is true even for $G = 9$, where there are only 512 distinct bootstrap samples. The uniform distribution comes next, but after a noticeable gap, followed by the standard normal and Mammen continuous distributions. The worst undercoverage, by a considerable margin, is provided by the Mammen two-point distribution (19), which is probably still the most widely used auxiliary distribution in practice.

Figure 7 shows the coverage of studentized bootstrap intervals at the nominal 0.95 level as functions of P, the proportion of clusters treated, for $G = 16$ and $N = 800$, again with $\pi = 0.4$. The actual number of clusters treated varies from 2 to 14. The ordering of the six auxiliary distributions is precisely the same as in Figure 6. The Rademacher distribution always performs best, followed closely by the six-point distribution, and the Mammen two-point distribution always performs worst.

These results strongly support the use of the Rademacher distribution, even when G is very small, although the six-point distribution also works well and may be safer in that case. Other distributions should not be employed. Using Mammen's classic asymmetric two-point distribution (19) appears to be a particularly bad idea.

8. Modified CRVEs and Bootstrap DGPs

Following MacKinnon and White (1985), it is common to transform residuals prior to using them in the filling of a sandwich HCCME. The most popular such HCCME is probably the HC_2 covariance matrix, which uses the transformation

$$\ddot{u}_i = \left(1 - \boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i'\right)^{-1/2}\hat{u}_i, \quad i = 1, \ldots, N, \tag{20}$$

where $\boldsymbol{X}_i$ denotes the i^{th} row of the regressor matrix $\boldsymbol{X}$. The HC_2 matrix has been studied extensively. Both theoretical and simulation results suggest that it usually yields more accurate inferences than HC_1, in which $\hat{u}_i$ is effectively just multiplied by a degrees-of-freedom correction. For a recent survey on heteroskedasticity-robust inference, see MacKinnon(2012).

Bell and McCaffrey (2002) proposed the cluster-robust analog of HC_2 as an alternative to the widely-used CV_1 matrix (2). It seems logical to refer to their covariance matrix estimator as CV_2. It omits the first factor in (2), which is essentially a degrees-of-freedom correction, and replaces the residual subvectors $\hat{\boldsymbol{u}}_g$ by the subvectors

$$\ddot{\boldsymbol{u}}_g = (\mathbf{I} - \boldsymbol{P}_{gg})^{-1/2}\hat{\boldsymbol{u}}_g, \quad g = 1, \ldots, G, \tag{21}$$

where $(\cdot)^{-1/2}$ denotes the symmetric square root of the inverse of the matrix inside the parentheses,

$$\boldsymbol{P}_{gg} \equiv \boldsymbol{X}_g(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_g',$$

and $\boldsymbol{X}_g$ denotes the $N_g \times K$ submatrix of $\boldsymbol{X}$ corresponding to the observations in cluster g. Thus $\boldsymbol{P}_{gg}$ is the $N_g \times N_g$ block that corresponds to cluster g on the diagonal of the projection matrix $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Equation (21) evidently reduces to equation (20) if all clusters have just one member, so that $G = N$ and $N_g = 1$ for all g.

Imbens and Kolesar (2012) provides some evidence that CV_2 outperforms CV_1. Based on this evidence and the much more extensive body of evidence that HC_2 outperforms HC_1, it might seem logical to use CV_2 all the time. There is a problem, however. The matrices $\mathbf{I} - \boldsymbol{P}_{gg}$ are $N_g \times N_g$. When N_g is large, finding the symmetric square root can be expensive. Indeed, when N_g is very large, simply creating and storing these matrices may be infeasible. This is a very real problem, because empirical work that uses cluster-robust inference often employs very large samples. For example, the largest cluster (for California) in the placebo laws experiments of MacKinnon and Webb (2014), which are based on the experiments of Bertrand, Duflo, and Mullanaithan (2004), has 42,625 observations. The corresponding $\boldsymbol{P}_{gg}$ matrix would take up more than 14 GB of memory.

The basic idea of the CV_2 matrix can be extended in several ways. First, we could define a CV_3 matrix similar to the simplified HC_3 matrix of Davidson and MacKinnon

(1993).[2] This would involve replacing $(\cdot)^{-1/2}$ by $(\cdot)^{-1}$ in equation (21). Although it seems extremely unlikely that CV_3 would outperform CV_2 in general, it might well do so in many cases, just as HC_3 often outperforms HC_2.

A more interesting application of the CV_2 idea is to use transformed residuals in the bootstrap DGP. Davidson and Flachaire (2008) suggest transforming the residuals in wild bootstrap DGPs in ways analogous to the transformations used in the HC_2 and HC_3 covariance matrices. That paper and MacKinnon (2012) find that bootstrap DGPs based on transformed residuals typically yield improved results, even when the covariance matrix does not employ the same transformation.

At least four different wild cluster bootstrap DGPs can be based on these ideas. Two of the four use a transformation like the one used in CV_2, and the other two use a transformation like the one used in CV_3. One of each pair, like (10), imposes the null, and the other, like (7), does not. The former would be appropriate for hypothesis testing and for restricted bootstrap intervals, and the latter would be appropriate for studentized bootstrap intervals. For example, the CV_2-like bootstrap DGP analogous to (7) is

$$\boldsymbol{y}_g^{*j} = \boldsymbol{X}_g\hat{\boldsymbol{\beta}} + \ddot{\boldsymbol{u}}_g v_g^{*j}, \quad g = 1, \dots, G, \tag{22}$$

with the vectors $\ddot{\boldsymbol{u}}_g$ defined by (21). The wild cluster bootstrap DGPs (7) and (22) will be referred to as wc_1 and wc_2, respectively.

Figures 8 and 9 report simulation results for the performance of the CV_1 and CV_2 covariance matrices and the wc_1 and wc_2 bootstrap DGPs. These are similar to Figures 6 and 7, respectively, except that they are based on only 100,000 replications because the CV_2 covariance matrix is much more expensive to compute than CV_1 even when the $(\mathbf{I} - \boldsymbol{P}_{gg})^{-1/2}$ matrices have been precomputed.

It is evident in both figures that Wald confidence intervals based on CV_2 perform substantially better than ones based on CV_1. Thus, if the sample size is small enough to make computation of CV_2 feasible, and the bootstrap is not going to be used, it is apparently a good idea to employ CV_2.

The studentized bootstrap confidence intervals almost always perform better than the Wald intervals. The improvement is always quite substantial in Figure 8, but it is very small for some values of P in Figure 9. Using the wc_2 bootstrap DGP always works better than using the wc_1 bootstrap DGP, except when $P = 0.125$. Interestingly, however, it makes almost no difference whether wc_2 is paired with CV_1 or CV_2. Since the CV_2 matrix is a great deal more expensive to compute than the CV_1 matrix, these

[2] As MacKinnon (2012) explains, this is not quite the same as the jackknife HC_3 matrix originally proposed in MacKinnon and White (1985).

results suggest that combining wc_2 with CV_1 may be the most attractive variant of the studentized bootstrap when the sample size is large but not extremely large.[3]

The CV_1+wc_2 intervals perform quite well except when G is very small and/or P is small or very large. However, if we compare Figure 9 with Figure 1, there appear to be no values of P where they would perform as well as restricted bootstrap intervals based on either LM or Wald tests.

9. Conclusion

Conventional cluster-robust confidence intervals are implicitly obtained by inverting t statistics based on cluster-robust standard errors. The simulation results in Section 6, combined with the ones in MacKinnon and Webb (2014), suggest that these intervals work well when the number of clusters is reasonably large, cluster sizes are not very dispersed, and the fraction of clusters subject to treatment (if the key regressor can be thought of as a treatment dummy) is moderate. However, when any of these conditions fails, they are prone to undercover, sometimes severely.

Various alternative confidence intervals are studied. The ones that usually work best are obtained by inverting bootstrap tests, but the procedure for calculating them can be computationally challenging; see Section 4. Studentized bootstrap intervals work well in many cases and are much less expensive to compute.

The performance of both conventional intervals and studentized bootstrap ones can be improved by using the CV_2 covariance matrix instead of the much more common CV_1 matrix. Unfortunately, the cost of computing CV_2 rises very rapidly with the number of observations in the largest cluster(s). For moderately large samples, it can easily be more expensive than the studentized bootstrap, and, for very large samples, it can be infeasible.

When it is feasible to compute CV_2, it should also be feasible to compute studentized bootstrap intervals that use transformed residuals in the bootstrap DGP together with CV_1 in the test statistics. These seem to work a little better than studentized bootstrap intervals where the bootstrap DGP does not use transformed residuals.

Section 7 studies the choice of an auxiliary distribution for the wild bootstrap. The results provide additional evidence in favor of the two-point Rademacher distribution and against the popular two-point Mammen distribution. They also suggest that it is unwise to use the standard normal or several other continuous distributions.

[3] Just what "a great deal more expensive" means depends on N, G, K, and the N_g. In the experiments, the CV_2+wc_2 intervals are about 4.7 times as expensive to compute as the CV_1+wc_2 intervals. When N is increased from 800 to 1600 or 3200, however, the CV_2+wc_2 intervals become 7.7 and 52 times as expensive, respectively.

References

Angrist, J. D., and J.-S. Pischke, *Mostly Harmless Econometrics*, Princeton, Princeton University Press.

Bell, R. M., and D. F. McCaffrey (2002). "Bias reduction in standard errors for linear regression with multi-stage samples," *Survey Methodology* 28, 169–181.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). "How much should we trust differences-in-differences estimates?" *Quarterly Journal of Economics*, 119, 249–275.

Bester, C. A., T. G. Conley, and C. B. Hansen (2011). "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 165, 137–151.

Breusch, T. S. (1979). "Conflict among criteria for testing hypotheses: Extensions and comments," *Econometrica*, 47, 203–207.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 90, 414–427.

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

Carter, A. V., K. T. Schnepel, and D. G. Steigerwald (2013). "Asymptotic behavior of a t test robust to cluster heterogeneity," Technical report, University of California, Santa Barbara.

Davidson, J., A. Monticini, and D. Peel (2007). "Implementing the wild bootstrap using a two-point distribution," *Economics Letters*, 96, 309–315.

Davidson, R., and E. Flachaire (2008). "The wild bootstrap, tamed at last," *Journal of Econometrics*, 146, 162–169.

Davidson, R., and J. G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, Oxford University Press.

Davidson, R., and J. G. MacKinnon (1999). "The size distortion of bootstrap tests," *Econometric Theory*, 15, 361–376.

Davidson, R., and J. G. MacKinnon (2000). "Bootstrap tests: How many bootstraps?" *Econometric Reviews*, 19, 55–68.

Davidson, R., and J. G. MacKinnon (2004). *Econometric Theory and Methods*, New York, Oxford University Press.

Davidson, R., and J. G. MacKinnon (2014). "Bootstrap confidence sets with weak instruments," *Econometric Reviews*, 33, 651–675.

Davison, A. C., and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*, Cambridge, Cambridge University Press.

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

Imbens, G. W., and M. Kolesar (2012). "Robust standard errors in small samples: Some practical advice," Working Paper 18478, National Bureau of Economic Research.

Liang, K.-Y., and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." *Biometrika* 73, 13–22.

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

MacKinnon, J. G. (2012). "Thirty years of heteroskedasticity-robust inference," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, ed. X. Chen and N. R. Swanson, New York, Springer, 2012, 437–461.

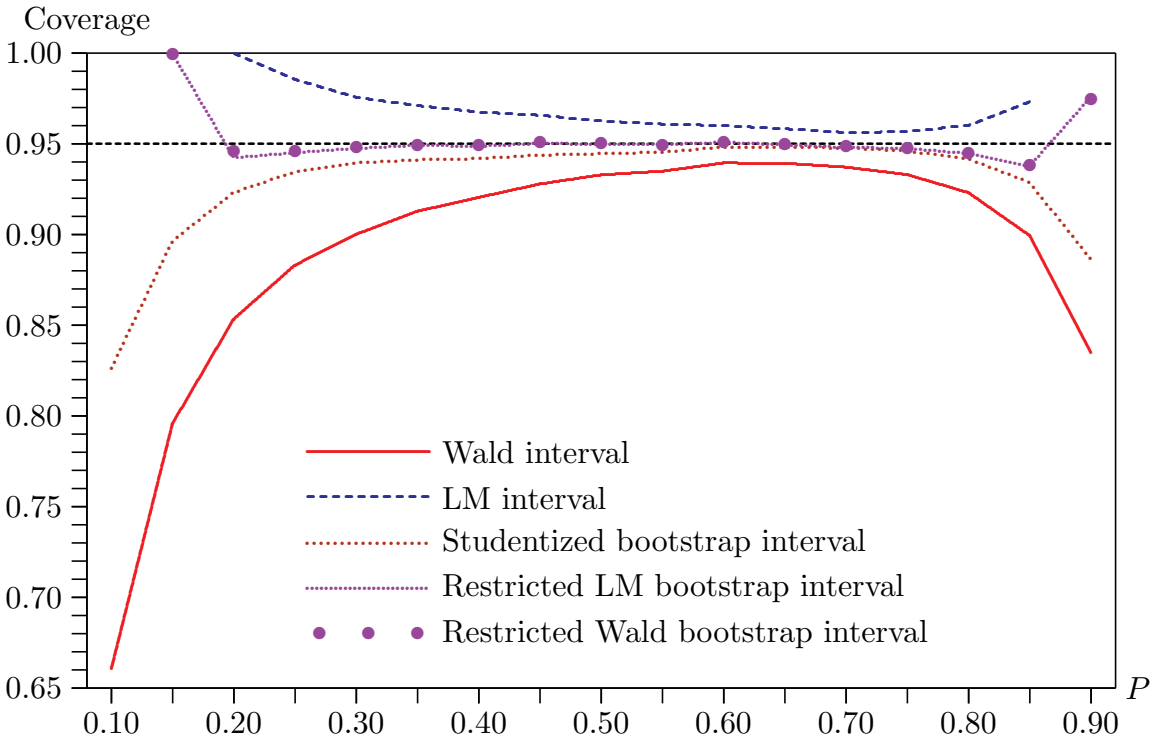
MacKinnon, J. G., and M. D. Webb (2014). "Wild bootstrap inference for wildly different cluster sizes," Working Papers 1314 (revised), Queen's University, Department of Economics.

MacKinnon, J. G., and H. White (1985). "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

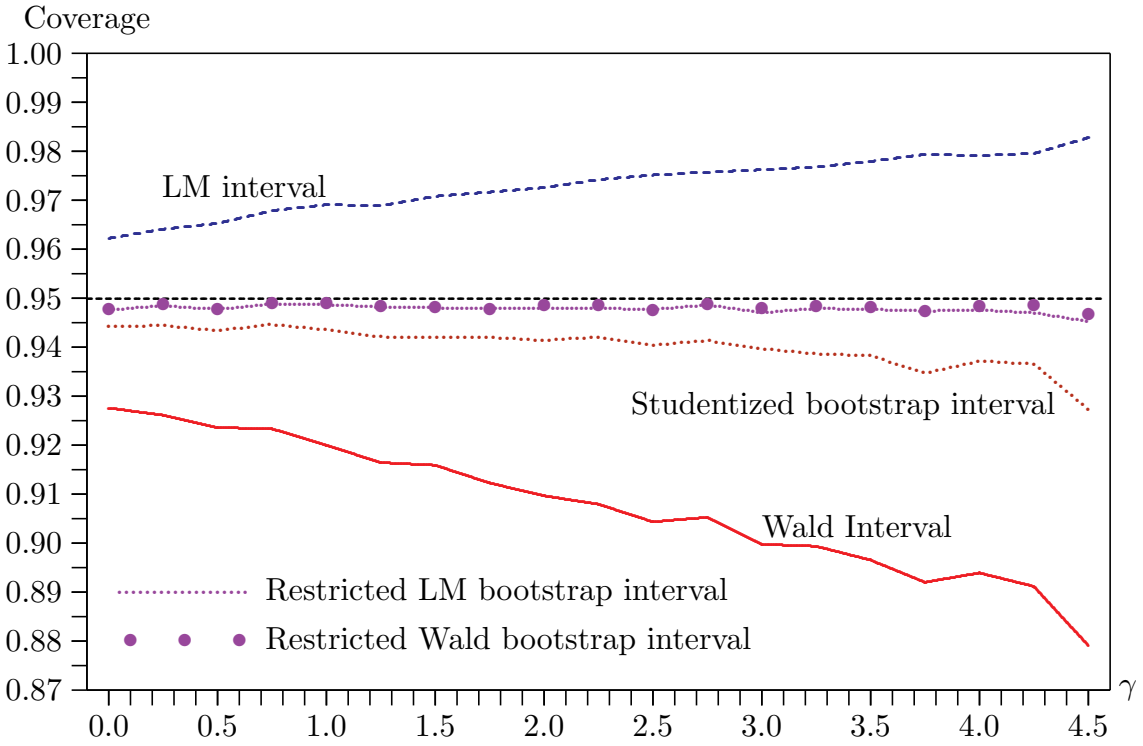
Mammen, E. (1993). "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics* 21, 255–285.

Webb, M. D. (2013). "Reworking wild bootstrap based inference for clustered errors," Working Papers 1315, Queen's University, Department of Economics.

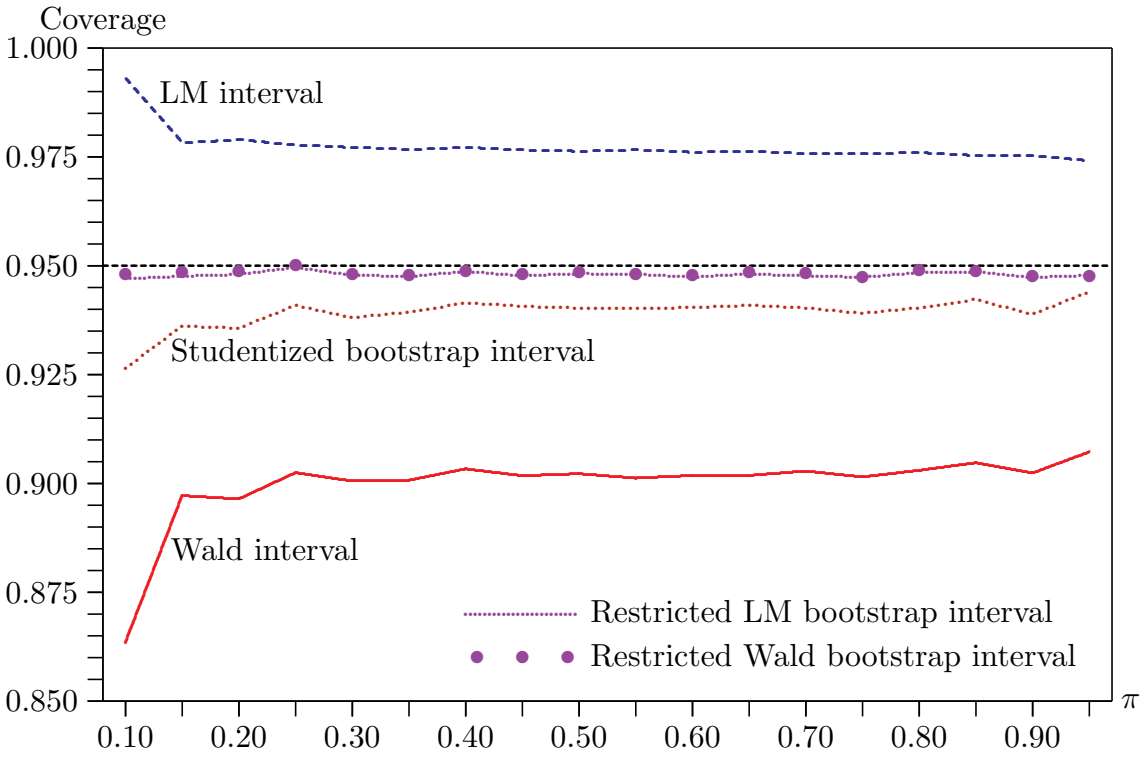
Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis," *Annals of Statistics*, 14, 1261–1295.



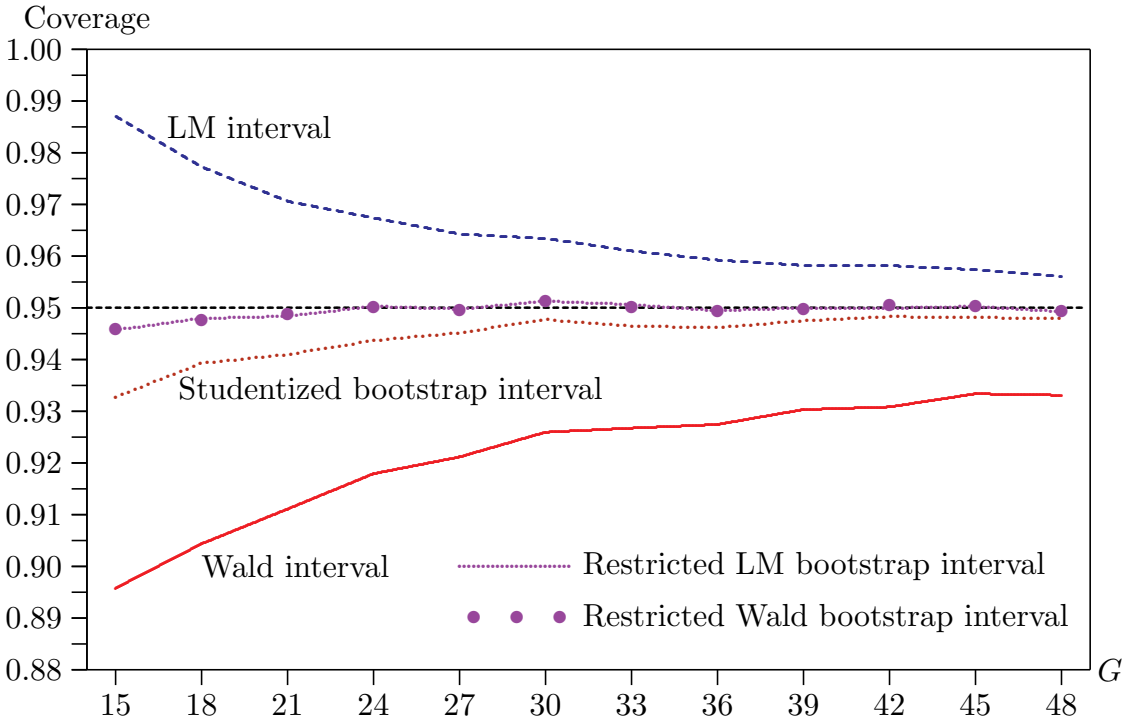
1. Coverage versus proportion of clusters treated for $G = 20$, $\gamma = 3$, $\pi = 0.4$



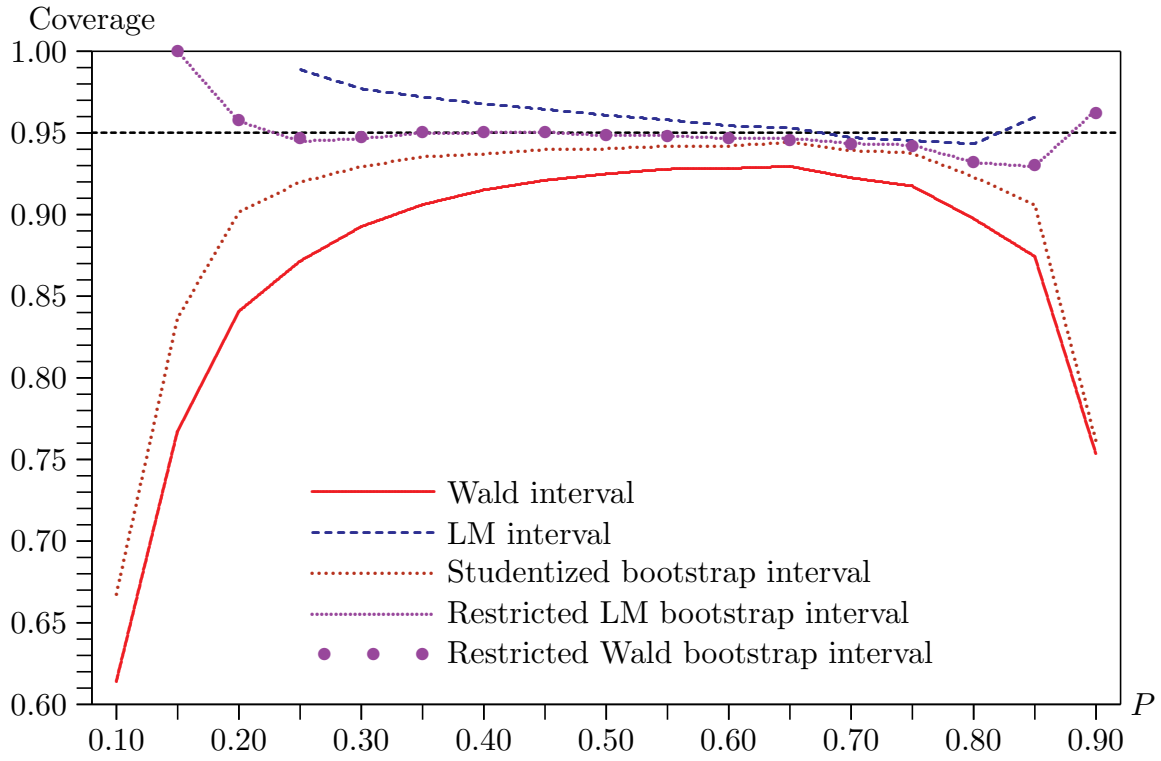
2. Coverage versus γ for $G = 20$, $P = 0.3$, $\pi = 0.4$



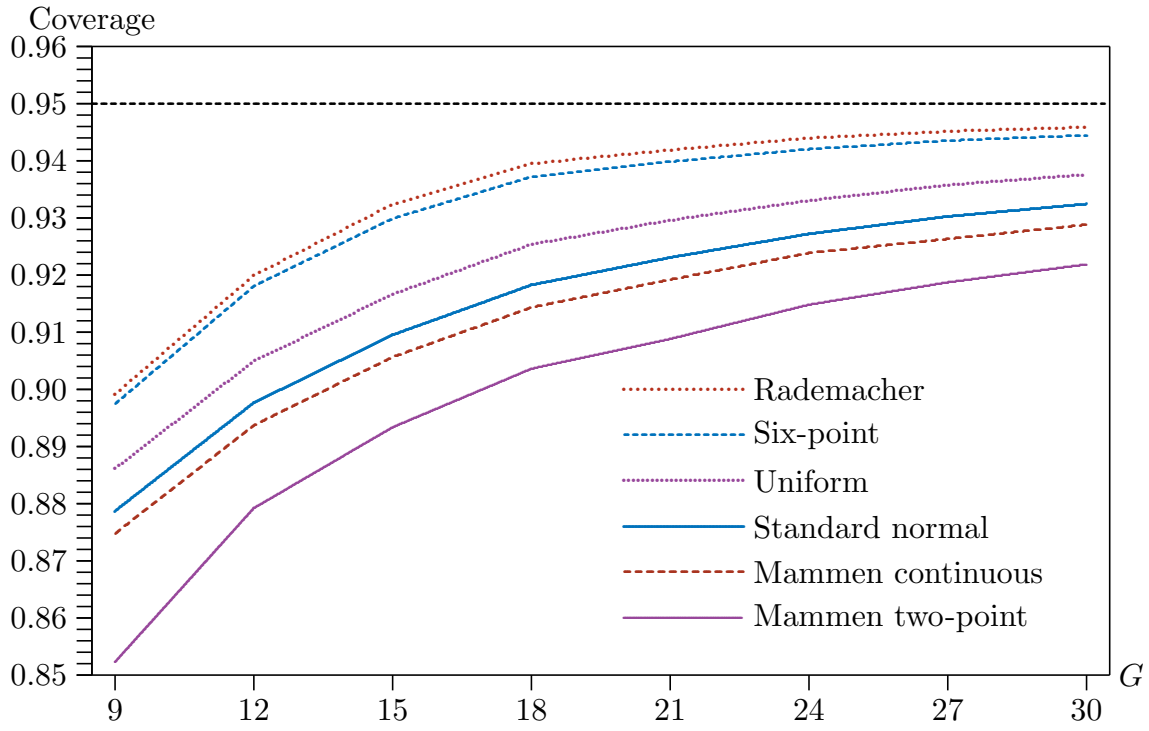
3. Coverage versus fraction of treated clusters treated for $G = 20$, $\gamma = 3$, $P = 0.3$



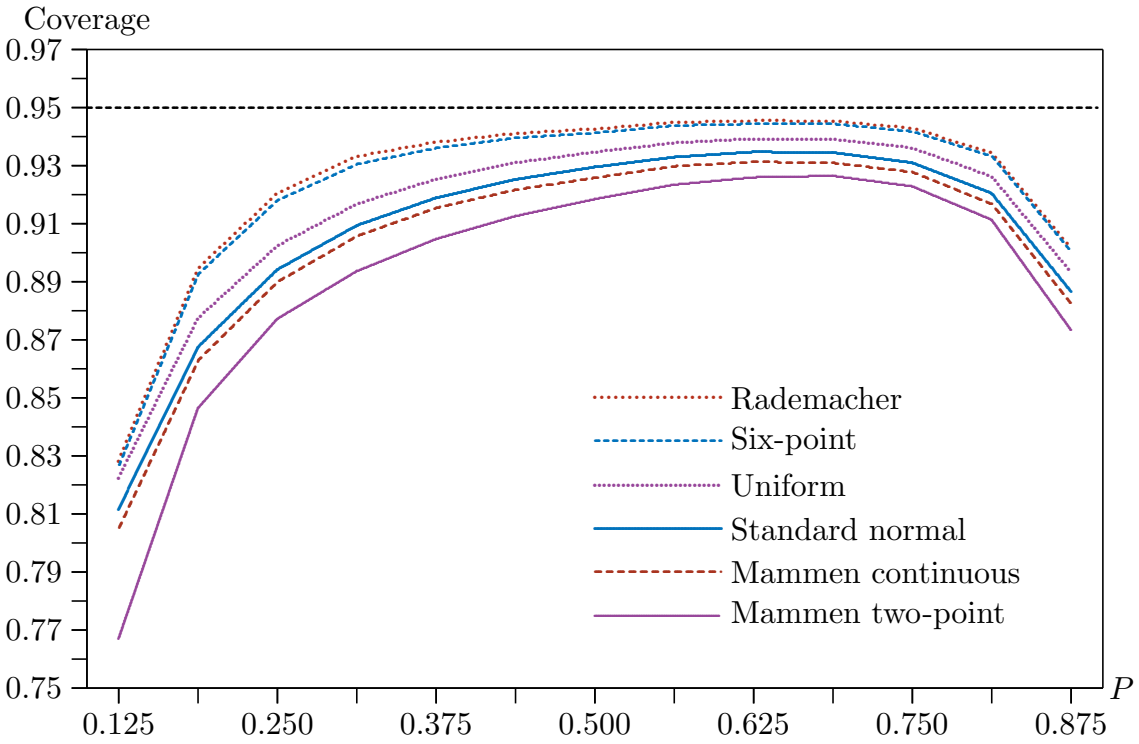
4. Coverage versus G for $\gamma = 3$, $P = 0.333$, $\pi = 0.4$



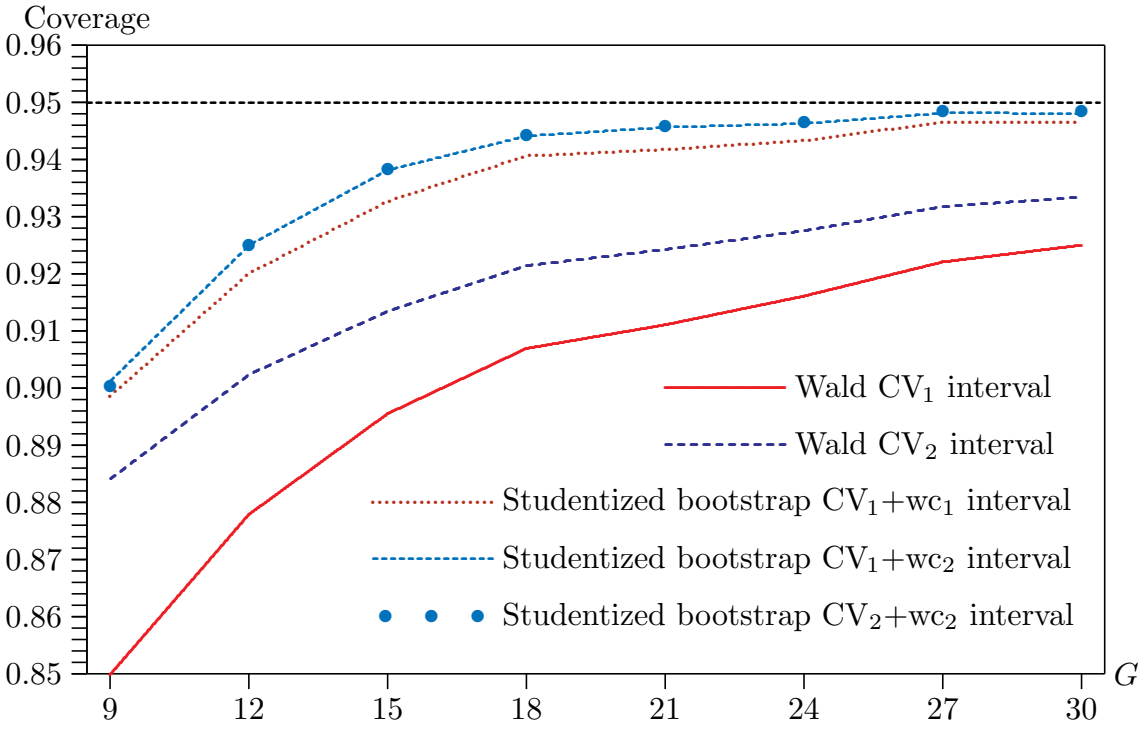
5. Coverage versus proportion of clusters treated for $G = 20$, $\gamma = 3$, $\pi = 0.2$ and 0.6



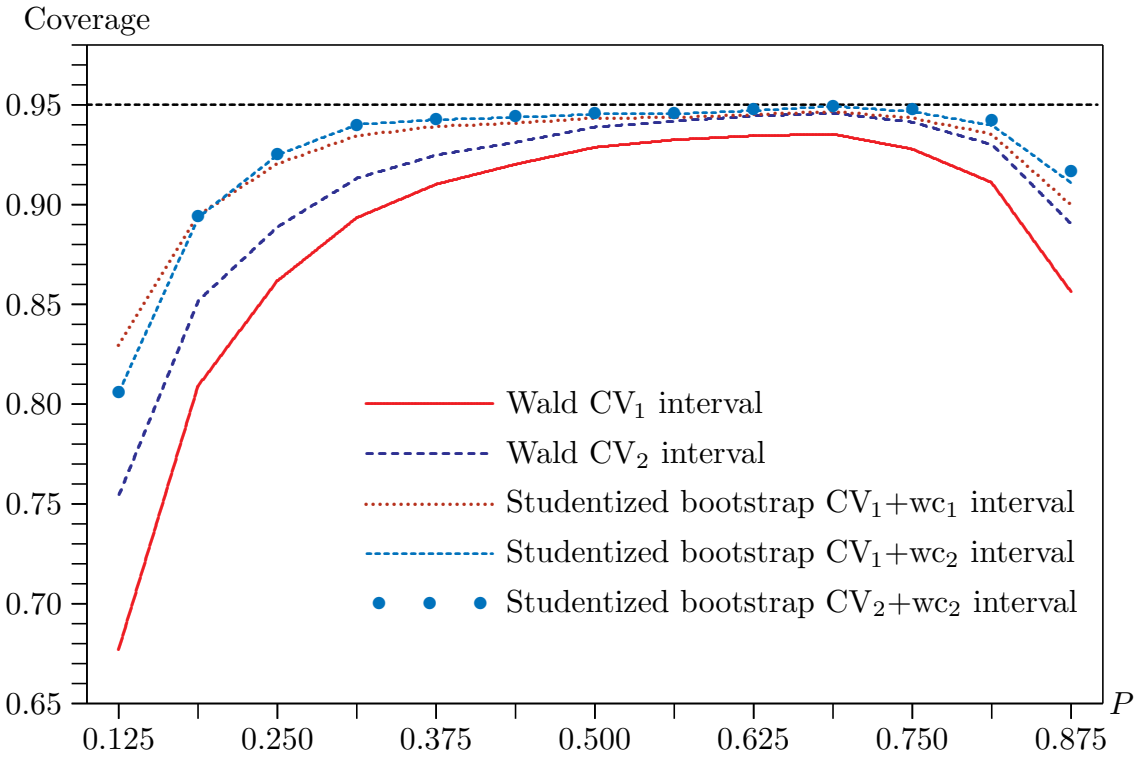
6. Coverage of studentized bootstrap intervals versus G for $\gamma = 3$, $P = 0.333$, $\pi = 0.4$



7. Coverage of studentized bootstrap intervals versus P for $G = 16$, $\gamma = 3$, $\pi = 0.4$



8. Coverage versus G for $N = 50G$, $\gamma = 3$, $P = 0.333$, $\pi = 0.4$



9. Coverage versus P for $G = 16$, $\gamma = 3$, $\pi = 0.4$