Queen's Economics Department Working Paper No. 1404

Wild Bootstrap Randomization Inference for Few Treated Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

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

6-2018

Wild Bootstrap Randomization Inference for Few Treated Clusters[*]

James G. MacKinnon
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

August 15, 2018

Abstract

When there are few treated clusters in a pure treatment or difference-in-differences setting, t tests based on a cluster-robust variance estimator (CRVE) can severely over-reject. Although procedures based on the wild cluster bootstrap often work well when the number of treated clusters is not too small, they can either over-reject or under-reject seriously when it is. In a previous paper, we showed that procedures based on randomization inference (RI) can work well in such cases. However, RI can be impractical when the number of possible randomizations is small. We propose a bootstrap-based alternative to randomization inference, which mitigates the discrete nature of RI P values in the few-clusters case. We also compare it to two other procedures. None of them works perfectly when the number of clusters is very small, but they can work surprisingly well.

Keywords: CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, difference-in-differences, DiD, randomization inference, kernel-smoothed P value

[*] The WBRI procedure discussed in this paper was originally proposed in a working paper circulated as "Randomization Inference for Difference-in-Differences with Few Treated Clusters." However, a revised version of that paper no longer discusses the WBRI procedure. We are grateful to Jeffrey Wooldridge, seminar participants at the Complex Survey Data conference on October 19-20, 2017 and at New York Camp Econometrics XIII on April 6-8, 2018, and two anonymous referees for helpful comments. This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. Joshua Roxborough and Oladapo Odumosu provided excellent research assistance.

1 Introduction

During the past decade or two, it has become common for empirical work in many areas of economics to involve models where the error terms are allowed to be correlated within clusters. Much of this work employs difference-in-differences (DiD) estimators, where the dataset has both a time and a cross-section dimension, and clustering is typically at the cross-section level (say, by state or province). Cameron and Miller (2015) provides a recent and comprehensive survey of econometric methods for cluster-robust inference.

Despite considerable progress in the development of suitable econometric methods over the past decade, it can still be a challenge to make reliable inferences. Doing so is particularly challenging in the DiD context when there are very few treated clusters. Past research, including Conley and Taber (2011), has shown that inference based on cluster-robust test statistics can greatly over-reject in this case. MacKinnon and Webb (2017b) explains why this happens and why the wild cluster bootstrap of Cameron, Gelbach and Miller (2008) does not solve the problem; for a less technical discussion, see also MacKinnon and Webb (2017a). When there are very few treated clusters, the restricted wild cluster bootstrap often severely under-rejects, and the unrestricted wild cluster bootstrap often severely over-rejects.

One potentially attractive way to obtain tests with accurate size when there are few treated clusters is to use randomization inference (RI). This approach involves comparing estimates based on the clusters that were actually treated with estimates based on control clusters that were not treated. Several authors have recently investigated this approach; see Conley and Taber (2011), Canay, Romano and Shaikh (2017), Ferman and Pinto (2017), and MacKinnon and Webb (2018a).

Randomization inference procedures necessarily rely on strong assumptions about how similar the control clusters are to the treated clusters. MacKinnon and Webb (2018a) shows that, for RI procedures which use coefficient estimates, like the one of Conley and Taber (2011), these assumptions almost always fail to hold when the treated clusters have either more or fewer observations than the control clusters. As a consequence, the procedure may over-reject or under-reject quite noticeably when the treated clusters are substantially smaller or larger than the controls. MacKinnon and Webb (2018a) suggests that more reliable inferences can often be obtained by basing randomization inference on t statistics rather than coefficient estimates. However, such procedures can involve noticeable power loss relative to ones based on coefficient estimates.

In Section 2, we briefly discuss conventional asymptotic procedures for inference with clustered errors. In Subsection 2.1, we then explain how the wild cluster bootstrap works. In Section 3, we introduce randomization inference and discuss two variants of it, one based on coefficient estimates which is essentially what was proposed in Conley and Taber (2011), and one based on t statistics proposed in MacKinnon and Webb (2018a).

All RI procedures encounter a serious practical problem when the number of controls is small. Since there are not many ways to compare the treated clusters with the control clusters, the RI P value can take on only a small number of values in such cases. We discuss this problem in Subsection 3.1. Section 4 then introduces a modified RI procedure, which we call "wild bootstrap randomization inference," or WBRI, that combines RI with the wild cluster bootstrap. There are two variants, one based on t statistics and one based on coefficient estimates. The WBRI procedure is the main contribution of the paper.

In Section 5, we briefly discuss two alternative procedures. One of them is to use P values obtained by kernel smoothing; see Racine and MacKinnon (2007b). The second, which makes much stronger assumptions about the error terms, is to estimate the model at the cluster level, with just one observation per cluster; see Donald and Lang (2007). We do not discuss the "synthetic controls" method of Abadie, Diamond and Hainmueller (2010), because it is, in our view, fundamentally different from WBRI and the other procedures that we consider. It involves "matching" the treated clusters with untreated ones according to their characteristics.

In Section 6, we show that both WBRI and the other procedures we discuss can substantially improve inference in cases where the only problem is an insufficient number of control clusters. All these methods can work surprisingly well even when the number of treated clusters is very small.

Finally, in Section 7, we present an empirical example from Decarolis (2014). This example involves just one treated cluster. Section 8 concludes.

2 Cluster-Robust Inference

A linear regression model with clustered errors may be written as

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

where each of the G clusters, indexed by g, has N_g observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ have $N = \sum_{g=1}^{G} N_g$ rows, $\boldsymbol{X}$ has k columns, and the parameter vector $\boldsymbol{\beta}$ has k rows. Each subvector $\boldsymbol{\epsilon}_g$ is assumed to have covariance matrix $\boldsymbol{\Omega}_g$ and to be uncorrelated with every other subvector. The covariance matrix $\boldsymbol{\Omega}$ of the entire error vector is block diagonal with diagonal blocks the $\boldsymbol{\Omega}_g$. OLS estimation of equation (1) yields estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$.

Because the elements of the $\boldsymbol{\epsilon}_g$ are in general neither independent nor identically distributed, both classical OLS and heteroskedasticity-robust standard errors for $\hat{\boldsymbol{\beta}}$ are invalid. As a result, conventional inference can be severely unreliable. The true covariance matrix for the model (1) is

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Omega}_g \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2}$$

This can be estimated by using a cluster-robust variance estimator, or CRVE. The most popular CRVE is:

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

where $\hat{\boldsymbol{\epsilon}}_g$ is the subvector of $\hat{\boldsymbol{\epsilon}}$ that corresponds to cluster g. This is the estimator that is used when the `cluster` command is invoked in Stata.[1] Consistent with the results of Bester, Conley and Hansen (2011), it is common to assume that the t statistics follow a $t(G-1)$ distribution; this is what Stata does by default.

It is not obvious that using t statistics based on the CRVE (3) is valid asymptotically. The proof requires technical assumptions about the distributions of the errors and the regressors and how the number of clusters and their sizes change as the sample size tends to infinity; see Djogbenou, MacKinnon and Nielsen (2018). Nevertheless, test statistics based on (3) seem to yield reliable inferences when the number of clusters is large and there is not too much heterogeneity across clusters. In particular, the number of observations per cluster must not vary too much; see Carter, Schnepel and Steigerwald (2017) and MacKinnon and Webb (2017b). However, t statistics based on (3) tend to over-reject severely when the parameter of interest is the coefficient on a treatment dummy and there are very few treated clusters; see Conley and Taber (2011) and MacKinnon and Webb (2017b). Rejection frequencies can be over 75% when all the treated observations belong to the same cluster.

In this paper, we are primarily concerned with the difference-in-differences (DiD) model, which is often appropriate for studies that use individual data in which there is variation in treatment across both clusters (or groups) and time periods. We can write such a model as

$$\begin{aligned} y_{igt} &= \beta_1 + \beta_2 \mathrm{GT}_g + \beta_3 \mathrm{PT}_t + \beta_4 \mathrm{TREAT}_{gt} + \epsilon_{igt}, \\ & i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \quad t = 1, \ldots, T, \end{aligned} \tag{4}$$

where i indexes individuals, g indexes groups, and t indexes time periods. Here GT_g is a "group treated" dummy that equals 1 if group g is treated in any time period, PT_t is a "period treated" dummy that equals 1 if any group is treated in time period t, and TREAT_{gt} is a dummy that equals 1 if an observation is actually treated.

The coefficient of most interest, on which we focus in this paper, is β_4, which measures the effect on treated groups in periods when there is treatment. In many cases, of course, regression (4) would also contain additional regressors, such as group and/or time dummies, which might make it necessary to drop GT_g, PT_t, or both. Following the literature, we divide the G groups into G_0 control groups, for which $\mathrm{GT}_g = 0$, and G_1 treated groups, for which $\mathrm{GT}_g = 1$, so that $G = G_0 + G_1$.

We are concerned with the case in which G_1 is small. In this case, as previously noted, CRVE-based inference fails. It also fails when G_0 is small in a pure treatment model where every cluster is either entirely treated or entirely not treated. However, in a DiD model where treatment only takes place in some time periods, it is possible for CRVE-based inference to perform well even when $G_0 = 0$; see MacKinnon and Webb (2017a, b). In the remainder of the paper, since we are focusing on the DiD case, we assume that only G_1 may be small.

The reason for the failure of CRVE-based inference when G_1 is small is explained in detail in MacKinnon and Webb (2017b, Section 6). Essentially, the problem is that the least squares residuals must be orthogonal to the treatment dummy variable. This implies that they sum to zero over all the treated observations. When those treated observations are spread over many clusters, there is no problem. But when they are concentrated in just a

[1] One of the earliest CRVEs was suggested in Liang and Zeger (1986). Alternatives to (3) have been proposed in Bell and McCaffrey (2002) and Imbens and Kolesár (2016), among others.

few clusters, some of the terms that are summed in the middle matrix of (3) severely underestimate the corresponding quantities in the matrices $\boldsymbol{X}'\boldsymbol{\Omega}_g\boldsymbol{X}$.[2] This causes the standard error of $\hat{\beta}_4$ to be seriously underestimated.

2.1 The Wild Cluster Bootstrap

The wild cluster bootstrap (WCB) was proposed in Cameron, Gelbach and Miller (2008) as a method for reliable inference in cases with a small number of clusters, and its asymptotic validity is proved in Djogbenou, MacKinnon and Nielsen (2018). A different, but less effective, bootstrap procedure for cluster-robust inference, often referred to as the "pairs cluster bootstrap," was previously suggested in Bertrand, Duflo and Mullainathan (2004); see MacKinnon and Webb (2017a). The WCB was studied extensively in MacKinnon and Webb (2017b) for the cases of unbalanced clusters and/or few treated clusters. Because we will be proposing a new procedure that is closely related to the wild cluster bootstrap in Section 4, we review how the latter works.

Without loss of generality, we consider how to test the hypothesis that β_4, the DiD coefficient in equation (4), is zero. Then the (restricted) wild cluster bootstrap works as follows:

1. Estimate equation (4) by OLS.

2. Calculate $\hat{t}_4$, the t statistic for $\beta_4 = 0$, using the square root of the 4th diagonal element of (3) as a cluster-robust standard error.

3. Re-estimate the model (4) subject to the restriction that $\beta_4 = 0$, so as to obtain restricted residuals $\tilde{\boldsymbol{\epsilon}}$ and restricted estimates $\tilde{\boldsymbol{\beta}}$.

4. For each of B bootstrap replications, indexed by b, generate a new set of bootstrap dependent variables y_{ig}^{*b} using the bootstrap DGP

$$y_{igt} = \tilde{\beta}_1 + \tilde{\beta}_2 \mathrm{GT}_g + \tilde{\beta}_3 \mathrm{PT}_t + \tilde{\epsilon}_{igt} v_g^{*b}, \qquad (5)$$
$$i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \quad t = 1, \ldots, T,$$

   Here y_{igt}^{*b} is an element of the vector $\boldsymbol{y}^{*b}$ of observations on the bootstrap dependent variable, $\mathrm{GT}_g, \mathrm{PT}_t$ are the corresponding row of $\boldsymbol{X}$, and v_g^{*b} is an auxiliary random variable that follows the Rademacher distribution; see Davidson and Flachaire (2008). It takes the values 1 and -1 with equal probability.[3]

5. For each bootstrap replication, estimate regression (4) using $\boldsymbol{y}^{*b}$ as the regressand. Calculate t_4^{*b}, the bootstrap t statistic for $\beta_4 = 0$, using the square root of the 4th diagonal element of (3), with bootstrap residuals replacing the OLS residuals, as the standard error.

[2] Of course, even when G_1 is not small, the matrices $N_g^{-1}\boldsymbol{X}_g'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'\boldsymbol{X}_g$ in (3) do not estimate the corresponding matrices $N_g^{-1}\boldsymbol{X}'\boldsymbol{\Omega}_g\boldsymbol{X}$ in (2) consistently, because the former matrices necessarily have rank 1. But the summation in the middle of expression (3), appropriately normalized, does consistently estimate the matrix $\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}$, appropriately normalized. See Djogbenou, MacKinnon and Nielsen (2018) for details.

[3] Because v_g^{*b} takes the same value for all observations within each group, we would not want to use the Rademacher distribution if G were smaller than about 12; see Webb (2014), which proposes an alternative for such cases.

6. Calculate the symmetric bootstrap P value as

$$\hat{p}_{\mathrm{s}}^{*} = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\big(|t_4^{*b}| > |t_4|\big), \tag{6}$$

   where $\mathbb{I}(\cdot)$ denotes the indicator function. Equation (6) assumes that, under the null hypothesis, the distribution of t_4 is symmetric around zero. Alternatively, one can use a slightly more complicated formula to calculate an equal-tail bootstrap P value.

The procedure just described is known as the restricted wild cluster, or WCR, bootstrap, because the bootstrap DGP (5) uses restricted parameter estimates and restricted residuals.[4] We could instead use unrestricted estimates and residuals in step 4 and calculate bootstrap t statistics for the hypothesis that $\beta_4 = \hat{\beta}_4$ in step 5. This yields the unrestricted wild cluster, or WCU, bootstrap.

MacKinnon and Webb (2017b) explains why the wild cluster bootstrap fails when the number of treated clusters is small. The WCR bootstrap, which imposes the null hypothesis, leads to severe under-rejection. In contrast, the WCU bootstrap, which does not impose the null hypothesis, leads to severe over-rejection. When just one cluster is treated, it over-rejects at almost the same rate as using CRVE t statistics with the $t(G-1)$ distribution.

The poor performance of WCR and WCU when there are few treated clusters is a consequence of the fact that the bootstrap DGP attempts to replicate the within-cluster correlations of the errors using residuals that have very odd properties. MacKinnon and Webb (2018b) therefore suggests using the ordinary wild bootstrap instead, and Djogbenou, MacKinnon and Nielsen (2018) proves that combining the ordinary wild bootstrap for the model (1) with the CRVE (3) leads to asymptotically valid inference. When clusters are sufficiently homogeneous, this procedure can work well even when the number of treated clusters is small.

3 Randomization Inference

Randomization inference, first proposed in Fisher (1935), is a procedure for performing exact tests in the context of experiments. The idea is to compare an observed test statistic $\hat{\tau}$ with an empirical distribution of test statistics τ_j^* for $j = 1, \ldots, S$ generated by re-randomizing the assignment of treatment across experimental units. To compute each of the τ_j^*, we use the actual outcomes while pretending that certain non-treated experimental units were treated. If $\hat{\tau}$ is in the tails of the empirical distribution of the τ_j^*, then this is evidence against the null hypothesis of no treatment effect.

Randomization tests are valid only when the distribution of the test statistic is invariant to the realization of the re-randomizations across permutations of assigned treatments; see Lehmann and Romano (2008) and Imbens and Rubin (2015). Whether this key assumption is true in the context of policy changes such as those typically studied in the DiD literature is debatable. Any endogeneity in the way policies are implemented over jurisdictions and time would presumably cast doubt on the assumption.

[4] For more details on how to implement the wild cluster bootstrap in Stata at minimal computational cost, see Roodman, MacKinnon, Nielsen and Webb (2018).

When treatment is randomly assigned at the individual level, the invariance of the distribution of the test statistic to re-randomization follows naturally. However, if treatment assignment is instead at the group level, as is always the case for DiD models like (4), then the extent of unbalancedness can determine how close the distribution is to being invariant.

It is obvious that the proportion of treated observations matters for $\hat{\beta}_4$ in (4) and its cluster-robust standard error. Let $\bar{d} = \big(\sum_{g=1}^{G_1} N_g\big)/N$ denote this proportion. When clusters are balanced, the value of $\bar{d}$ will be constant across re-randomizations. However, when clusters are unbalanced, $\bar{d}$ may vary considerably across re-randomizations. This implies that the distributions of $\hat{\beta}_4$ may also vary substantially. Randomization inference may not work well in such cases.

MacKinnon and Webb (2018a) studies two types of RI procedure. One uses $\hat{\beta}_4$ in (4) as $\hat{\tau}$, and the other uses the cluster-robust t statistic that corresponds to $\hat{\beta}_4$. The former procedure, which we refer to as RI-β, is quite similar to a procedure proposed in Conley and Taber (2011). It is only valid, even in large samples, if re-randomizing does not change the distribution of the $\hat{\beta}^*_{4j}$. The latter procedure, which we refer to as RI-t, is evidently valid in large samples whenever the cluster-robust t statistics follow an asymptotic distribution that is invariant to $\bar{d}$ and to any other features of the individual clusters. However, as MacKinnon and Webb (2018a) shows, it is generally not valid in finite samples when $\bar{d}$ varies across re-randomizations, especially when G_1 is small. Nevertheless, the RI-t procedure typically works better than the RI-β procedure, especially when G_1 is not too small.

When there is just one treated group, it is natural to compare $\hat{\tau}$ to the empirical distribution of G_0 different τ^*_j statistics. However, when there are two or more treated groups and G_0 is not quite small, the number of potential τ^*_j to compare with can be very large. In such cases, we may pick S of them at random. To avoid ties, we never include the actual $\hat{\tau}$ among the τ^*_j. Some RI procedures do in fact include $\hat{\tau}$, however. Provided S is large, this is inconsequential.

The randomization inference procedures discussed in MacKinnon and Webb (2018a) for the model (4) work as follows. Here $\hat{\tau}$ denotes either $\hat{\beta}_4$ or its cluster-robust t statistic, and τ^*_j denotes the corresponding quantity for the j^{th} re-randomization.

1. Estimate the regression model and calculate $\hat{\tau}$.

2. Generate a number of τ^*_j statistics, S, to compare $\hat{\tau}$ with.

   - When $G_1 = 1$, assign a group from the G_0 control groups as the "treated" group g^* for each repetition, re-estimate the model using the observations from all G groups, and calculate a new statistic, τ^*_j, indicating randomized treatment. Repeat this process for all G_0 control groups. Thus the empirical distribution of the τ^*_j will have G_0 elements.

   - When $G_1 > 1$, sequentially treat every set of G_1 groups except the set actually treated, re-estimate equation (4), and calculate a new τ^*_j. There are potentially ${}_G\mathrm{C}_{G_1} - 1$ sets of groups to compare with, where ${}_n\mathrm{C}_k$ denotes "n choose k." When this number is not too large, obtain all of the τ^*_j by enumeration. When it exceeds B (picked on the basis of computational cost), choose the comparators randomly, without replacement, from the set of potential comparators. Thus the empirical distribution will have $S = \min({}_G\mathrm{C}_{G_1} - 1, B)$ elements.

3. Sort the vector of τ_j^* statistics.

4. Determine the location of $\hat{\tau}$ within the sorted vector of the τ_j^*, and compute a P value. This may be done in more than one way, as we discuss in the next subsection.

In the above procedures, we need to assign a starting period for "treatment" in each re-randomization if we are dealing with a DiD model like (4). The method used in the simulation experiments in MacKinnon and Webb (2018a) and in Subsection 6 below is to make the treatment period(s) the same for each re-randomization as for the actual sample. Thus if, for example, $G_1 = 1$ and treatment began in 1978, the single "treated" group in all re-randomizations would start treatment in 1978. If $G_1 = 2$ and treatment began in 1978 and 1982, then, for each re-randomization, one group would begin treatment in 1978 and the other in 1982. In our simulations, we ordered both the actually treated groups and the controls by size. Thus if, for example, treatment began in 1978 for group 3 and in 1982 for group 11, and $N_3 > N_{11}$, then treatment would begin in 1978 for the larger control group and in 1982 for the smaller one. We also experimented with assigning treatment years at random and found that doing so made very little difference.

3.1 The Problem of Interval P Values

The most natural way to calculate an RI P value is probably to use the equivalent of equation (6). As before, S denotes the number of repetitions, which would be G_0 when $G_1 = 1$ and the minimum of ${}_G\mathrm{C}_{G_1} - 1$ and B when $G_1 > 1$, where B is a user-specified target number of replications. Then the analog of equation (6) is

$$\hat{p}_1^* = \frac{1}{S} \sum_{j=1}^{S} \mathbb{I}\big(|\tau_j^*| > |\hat{\tau}|\big). \tag{7}$$

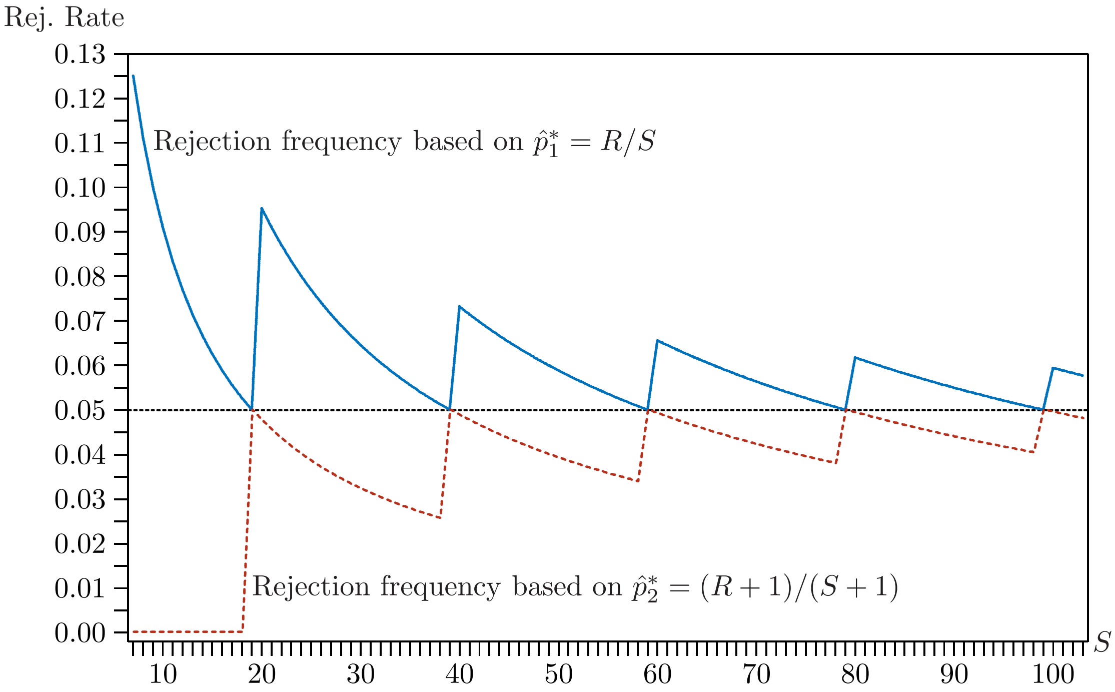
This makes sense if we are testing the null hypothesis that $\beta_4 = 0$ and expect the τ_j^* to be symmetrically distributed around zero. If we were instead testing the one-sided null hypothesis that $\beta_4 \leq 0$, we would want to remove the absolute value signs.

Equation (7) is not the only way to compute an RI P value for a point null hypothesis, however. A widely-used alternative is

$$\hat{p}_2^* = \frac{1}{S+1} \left(1 + \sum_{j=1}^{S} \mathbb{I}\big(|\tau_j^*| > |\hat{\tau}|\big)\right). \tag{8}$$

It is easy to see that the difference between $\hat{p}_1^*$ and $\hat{p}_2^*$ is $O(1/S)$, so that they tend to the same value as $S \to \infty$. There is evidently no problem if S is large, but the two P values can yield quite different inferences when S is small. The analogous issue should rarely arise for bootstrap tests, because the investigator can almost always choose B (the number of bootstrap samples, which plays the same role as S here) in such a way that equations (7)

Figure 1: Rejection Frequencies and Number of Simulations



and (8) yield the same inferences. This will happen whenever $\alpha(B+1)$ is an integer, where α is the level of the test. That is why it is common to see $B = 99$, $B = 999$, and so on.

With randomization inference, however, we generally cannot choose S such that $\alpha(S+1)$ is an integer. In every other case, we could in principle use any P value between $\hat{p}_1^*$ and $\hat{p}_2^*$. Thus P values based on a finite number of simulations are generally interval-identified rather than point-identified, where the interval is $[\hat{p}_1^*, \hat{p}_2^*]$; see Webb (2014).

For small values of S, the conflict between inferences based on $\hat{p}_1^*$ and $\hat{p}_2^*$ can be substantial. Figure 1 shows analytical rejection frequencies for tests at the .05 level based on equations (7) and (8), respectively. The tests would reject exactly 5% of the time if S were infinite, but the figure is drawn for values of S between 7 and 103. In the figure, R denotes the number of times that t is more extreme than t_j^*, so that $\hat{p}_1^* = R/S$ and $\hat{p}_2^* = (R+1)/(S+1)$. It is evident that $\hat{p}_1^*$ always rejects more often than $\hat{p}_2^*$, except when (for tests at the .05 level) $S = 19$, 39, 59, and so on. Even for fairly large values of S, the difference between the two rejection frequencies can be substantial.

Suppose the data come from Canada, which has just ten provinces. If one province is treated, then $G_1 = 1$, $G_0 = 9$, and the P value can lie in only one of nine intervals: 0 to 1/10, 1/9 to 2/10, 2/9 to 3/10, and so on. Even if $R = 0$, it would never be reasonable to reject at the .01 or .05 levels.

One way to eliminate the interval and obtain a single P value is to use a random number generator. Such a procedure is described, in the context of the bootstrap, in Racine and MacKinnon (2007b). The idea is simply to replace the 1 after the large left parenthesis in (8) with a draw from the U[0, 1] distribution. Similar procedures have been used for many years in the RI literature; see Young (2015). However, these procedures have the unfortunate property that the outcome of the test depends on the realization of a single random variable drawn by the investigator. The gap between $\hat{p}_1$ and $\hat{p}_2$ still remains. We have simply chosen a number between the two by, in effect, flipping a coin. This means that two different researchers using the same dataset will randomly obtain different P values.

4 Wild Bootstrap Randomization Inference

In this section, we suggest a novel way to overcome the problem of interval P values. We propose a procedure that we refer to as wild bootstrap randomization inference, or WBRI. The WBRI procedure essentially combines the wild cluster bootstrap of Subsection 2.1 with either the RI-t or RI-β procedures of Section 3. We focus on RI-t, because, at least under the null hypothesis, it seems to be better to use t statistics rather than coefficients for randomization inference.

The key idea of the WBRI procedure is to augment the small number (S) of test statistics obtained by randomization with a much larger number generated by a restricted wild cluster bootstrap DGP like (5). All the bootstrap samples are generated in the same way. However, they are used to test $S+1$ different null hypotheses, corresponding to the actual treatment and the S re-randomized ones.

Why should this procedure work? Under the (fairly strict) conditions for randomization inference to be valid (Imbens and Rubin 2015), the RI-t procedure would work perfectly if it were not for the interval P value problem. Provided the clusters are reasonably homogeneous and S is not too small, it generally seems to work very well; see MacKinnon and Webb (2018a). The idea of WBRI is to keep the good properties of RI-t for large S even when S is not large by generating a large number of bootstrap statistics that resemble the t_j^* obtained by re-randomization.

Of course, we could obtain as many bootstrap statistics t_b^* as we desire simply by using the wild cluster bootstrap. But, when G_1 is small, the $|t_b^*|$ tend to be positively correlated with $|t|$. This is the reason for the failure of the WCR bootstrap with few treated clusters; see MacKinnon and Webb (2017b). When $G_1 = 1$, the correlation tends to be very high, and this often leads to extreme under-rejection.

With the WBRI procedure, the bootstrap statistics $|t_{bj}^*|$ that correspond to the j^{th} re-randomization will undoubtedly be correlated with $|t_j|$ when G_1 is small. But only the ones that correspond to the actual null hypothesis should be strongly correlated with $|t|$. Thus WBRI should not encounter anything like the sort of extreme failure that WCR routinely does when G_1 is small. Of course, we do not expect that WBRI will ever work perfectly, especially when the number of clusters is very small. But it seems plausible that it should yield P values which are reasonably accurate and much more precise than the interval $[\hat{p}_1, \hat{p}_2]$. We provide evidence on this point in Section 6.

We make no effort to prove the asymptotic validity of WBRI, because any proof would require that G tend to infinity; see Djogbenou, MacKinnon and Nielsen (2018). But when G is large, $S = G-1$ for $G_1 = 1$ and $S >> G$ for $G_1 > 1$, so that there would be no problem of interval P values and no reason to employ WBRI.

Formally, the WBRI procedure for generating the t_b^* and t_{bj}^* statistics is as follows:

1. Estimate equation (4) by OLS and calculate t for the coefficient of interest using CRVE

standard errors.

2. Obtain S test statistics t_j^* by re-randomization, as in Section 3.

3. Estimate a restricted version of equation (4) with $\beta_4 = 0$, and retain the restricted estimates $\tilde{\boldsymbol{\beta}}$ and residuals $\tilde{\boldsymbol{\epsilon}}$.

4. For the original test statistic and each of the S possible re-randomizations, indexed by $j = 0, \ldots, S$, construct B bootstrap samples indexed by b, say $\boldsymbol{y}_{jb}^*$, using the restricted wild cluster bootstrap procedure discussed in Subsection 2.1. For each bootstrap sample, estimate equation (4) using $\boldsymbol{y}_{jb}^*$ and calculate a bootstrap t statistic t_{jb}^* based on CRVE standard errors.[5]

5. Use one of equations (7) or (8) to calculate a P value for t based on the $(B{+}1)(S{+}1){-}1$ bootstrap and randomized test statistics.

Since every possible set of G_1 clusters is "treated" in the bootstrap samples, the number of bootstrap test statistics is $B \times {}_G\mathrm{C}_{G_1} = B(S{+}1)$. In addition, there are ${}_G\mathrm{C}_{G_1}{-}1 = S$ statistics based on the original sample. Thus the total number of test statistics is $B(S + 1) + S = (B + 1)(S + 1) - 1$. We suggest choosing B so that this number is at least 1000.

The number of possible bootstrap DGPs is only 2^G if one uses the Rademacher distribution. Therefore, when G is small, it is better to use an alternative bootstrap distribution such as the 6-point distribution suggested in Webb (2014). In the case of the latter, the number of possible bootstrap DGPs is 6^G.

In general, it makes sense to use the WBRI procedure only when the RI-t procedure does not provide enough t_j^* for the interval P value problem to be negligible. As a rule of thumb, we suggest using WBRI when $G_1 = 1$ and $G < 300$, or $G_1 = 2$ and $G < 30$, or $G_1 = 3$ and $G < 15$. Code for this procedure is available from the authors.[6]

5 Alternative Procedures

In this section, we briefly discuss two very different procedures that can be used instead of WBRI. Their performance will be compared with that of the latter in the simulation experiments of Section 6.

Racine and MacKinnon (2007a) suggested a way to solve the interval P value problem in the context of bootstrap tests. For those tests, the problem only arises if computation is so expensive that making $\alpha(B + 1)$ an integer for all test levels α of interest is infeasible. But since the problem arises quite frequently for randomization tests, their procedure may be useful in this context.

Recall the example of Canadian provinces given in Section 3.1. Suppose the treated province has a more extreme outcome than any of the others, so that $R = 0$. In the strict context of randomization inference, all we can say is that the P value is between 0, according

[5] Note that, in this procedure, B denotes the number of bootstrap samples per re-randomization. The total number of bootstrap samples is $B(S{+}1)$. It might seem tempting to use the same B bootstrap samples for every re-randomization. However, this would create dependence among the S different test statistics that depend on each bootstrap sample. This sort of dependence should be avoided.

[6] Code for the WBRI procedure can be found at https://sites.google.com/site/matthewdwebb/code

to equation (7), and 0.10, according to equation (8). In saying this, however, we have made no use of the actual values of t and the t^*_j. Only the location of $|t|$ in the sorted list affects either P value. If the outcome for the treated province differed a lot from the outcomes for the other nine provinces, that is, if $|t|$ were much larger than any of the $|t^*_j|$, then the evidence against the null hypothesis would seem to be quite strong. On the other hand, if $|t|$ were just slightly larger than the largest of the $|t^*_j|$, the evidence against the null would seem to be rather weak. But neither of the RI P values takes this into account.

The procedure of Racine and MacKinnon (2007a) does take the values of the actual and re-randomized test statistics into account. It is based on the smoothed P value

$$\hat{p}_h = 1 - \hat{F}_h(t) = 1 - \frac{1}{S}\sum_{j=1}^{S} K(t^*_j, t, h), \tag{9}$$

where $\hat{F}_h(t)$ is a kernel-smoothed CDF of the t^*_j evaluated at the actual test statistic t. When t is much more extreme than any of the t^*_j, it will surely lie in the far tail of the CDF, and $\hat{p}_h$ will be very small. On the other hand, when t is near one or more of the t^*_j, $\hat{p}_h$ is unlikely to be very small.

This procedure requires the choice of a kernel function $K(\cdot)$ and a bandwidth h. Because $\hat{p}_h$ is an estimated probability rather than an estimated density, $K(\cdot)$ must be a cumulative kernel. A natural choice is the cumulative standard normal CDF. The choice of h is more difficult, and it matters a lot when S is small. Based largely on simulation evidence, Racine and MacKinnon (2007a) suggested choosing $h = scS^{-4/9}$, where s is the standard deviation of the t^*_j, and the values of c are 2.418, 1.575, and 1.3167 for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$, respectively.[7] Thus the bandwidth h should be larger the more variable are the t^*_j and the smaller is the level of the test. The latter makes sense, because values of t^*_j will be scarcer near more extreme values of t.

The kernel smoothing procedure of Racine and MacKinnon (2007a) can evidently be used with coefficients as well as t statistics, and we consider both methods in the next section. Note that we estimate s from the t^*_j but then apply the procedure to $|t|$ and the $|t^*_j|$, whereas Racine and MacKinnon (2007a) considered upper-tail tests. Despite this difference, their smoothing procedure generally performed best for tests at the .05 level when using the value of c recommended for that level, namely, 1.575. All of the results reported in Section 6 use that value.

A radically different approach, which was studied in Donald and Lang (2007), is to collapse the original, individual data to the cluster level. Instead of N observations, the regression uses just G of them. Precisely how this works depends on the model. Consider the simple case in which

$$\boldsymbol{y}_g = \gamma\boldsymbol{\iota}_{N_g} + \beta\boldsymbol{x}_g + \boldsymbol{u}_g, \quad g = 1, \ldots, G, \tag{10}$$

where each of the subscripted vectors corresponds to a single cluster and has N_g observations, and the vector $\boldsymbol{\iota}_{N_g}$ contains N_g ones. If we take the averages of each of the vectors here, we

[7] There is a typo on page 5955 of Racine and MacKinnon (2007a) which causes the optimal values of $\alpha = .01$ and $\alpha = .10$ to be reversed. That this is incorrect can be seen from Figure 6 of the paper.

obtain $\bar{y}_g = \boldsymbol{\iota}'_{N_g}\boldsymbol{y}_g/N_g$, $\bar{x}_g = \boldsymbol{\iota}'_{N_g}\boldsymbol{x}_g/N_g$, and $\bar{u}_g = \boldsymbol{\iota}'_{N_g}\boldsymbol{u}_g/N_g$. This allows us to write

$$\bar{\boldsymbol{y}} = \gamma\boldsymbol{\iota}_G + \beta\bar{\boldsymbol{x}} + \bar{\boldsymbol{u}}, \tag{11}$$

where the G-vectors $\bar{\boldsymbol{y}}$, $\bar{\boldsymbol{x}}$, and $\bar{\boldsymbol{u}}$ have typical elements $\bar{y}_g$, $\bar{x}_g$, and $\bar{u}_g$, respectively. Since all the variables in regression (11) are cluster means, we refer to it as a "cluster-means regression," or CMR.

Donald and Lang (2007) argues that the ordinary t statistic for $\beta = 0$ in the cluster-means regression (11) will be approximately distributed as $t(G-2)$ if two restrictive but not unreasonable assumptions are satisfied. The first is that all clusters are the same size, so that $N_g = m$ for all g, with all of the $\boldsymbol{u}_g$ having the same covariance matrices. The second is either that the original error terms are normally distributed or that m is sufficient large so that a central limit theorem applies to the elements of $\bar{\boldsymbol{u}}$.

The advantage of collapsing individual data to the cluster level, as in (11), is that we no longer have to estimate a CRVE. Because of the first assumption, we do not even have to use heteroskedasticity-robust standard errors. This allows us to make inferences about β when just one cluster is treated. In that case, only one element of $\bar{\boldsymbol{x}}$ is non-zero, but we can still make valid inferences because all G observations are used to estimate the variance of the error terms.

6 Simulation Experiments

In this section, we report the results of some simulation experiments designed to assess the performance of WBRI and the procedures discussed in Section 5. The model is very simple. It is essentially equation (4), but without any group dummies. This model can also be thought of as equation (10) with time dummies instead of the constant term. To make inference a bit more difficult, the error terms follow a lognormal distribution. The group dummies are omitted because the error terms have constant intra-cluster correlations of 0.05 (prior to being exponentiated), and group dummies would soak up all of this correlation.

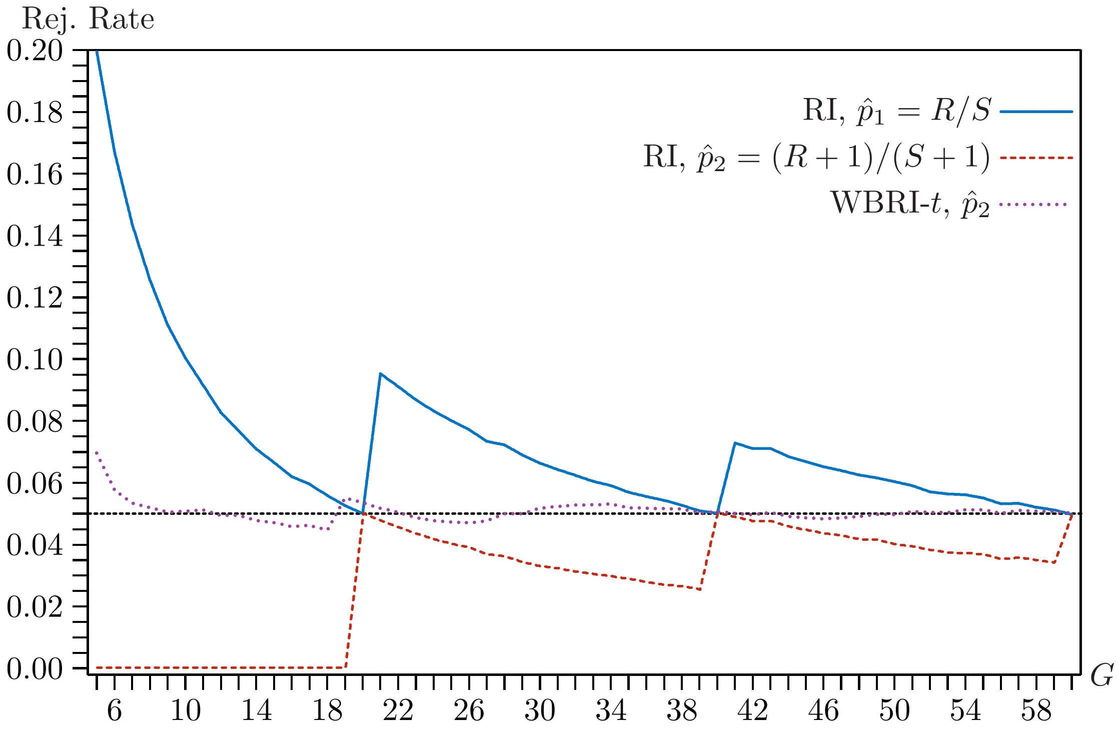
In the experiments that we report, there are G clusters, each with 100 observations divided evenly among 10 time periods. When a cluster is treated, treatment is always for 5 of the 10 periods. Because all clusters are the same size, and the number of treated observations per treated cluster is always the same, randomization inference would work perfectly if it were not for the interval P value problem. If we relaxed either of these assumptions, of course, it would not work perfectly, even when G is large; see MacKinnon and Webb (2018a).

Figure 2 shows rejection frequencies for tests at the .05 level for three procedures (RI-t using $\hat{p}_1$, RI-t using $\hat{p}_2$, and WBRI-t using $\hat{p}_2$) for 56 different experiments, each with 400,000 replications.[8] The number of clusters varies from 5 to 60, and only one cluster is treated. For any value of G, this is the case for which the interval P value problem is most severe, because $S = G - 1$ is small unless there are many clusters. The number of bootstraps per randomization is always chosen so that $(B + 1)(S + 1) \geq 1000$.

One striking feature of Figure 2 is that rejection frequencies for the two RI procedures are almost exactly what theory predicts; see Figure 1. When $G = 20$, 40, and 60, the two

[8] WBRI would have rejected slightly more often if we had used $\hat{p}_1$ instead of $\hat{p}_2$; the difference in rejection frequencies was almost always less than 0.001.

Figure 2: WBRI and RI Rejection Frequencies



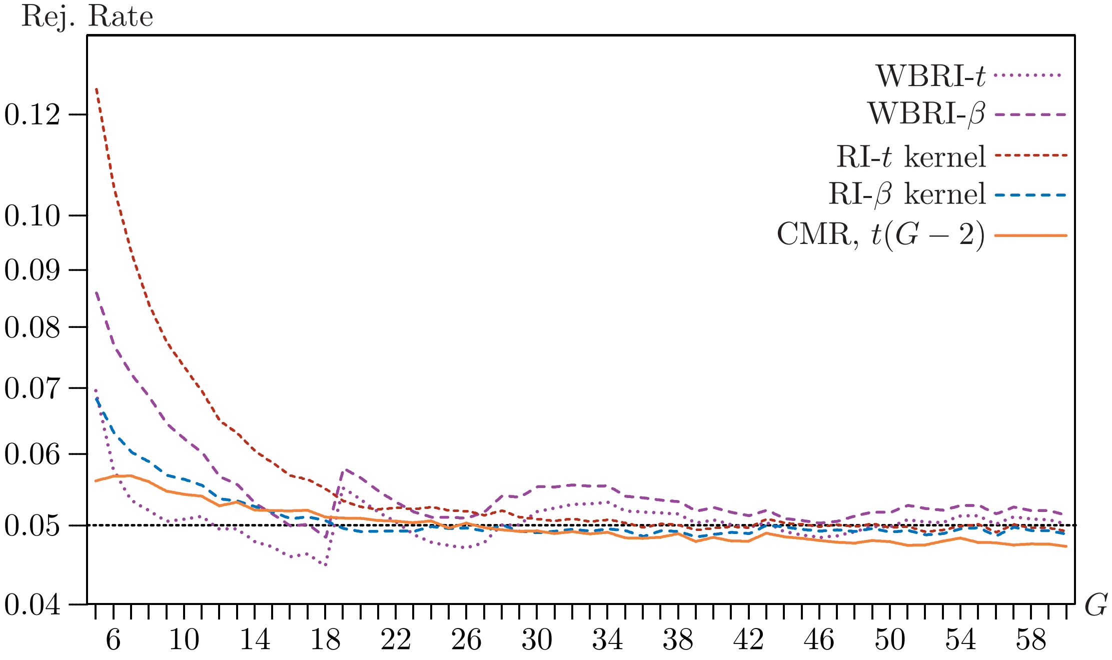
RI P values yield precisely the same outcomes, as they must. In every other case, however, $\hat{p}_1^* = R/S$ rejects more often than $\hat{p}_2^* = (R+1)/(S+1)$.

In Figure 2, the WBRI rejection frequencies are almost always between the two RI rejection frequencies (although this is not true for $G = 19$ and $G = 20$), and they are always quite close to 5% except when G is very small. This is what we would like to see. However, it must be remembered that the figure deals with a very special case in which all clusters are the same size and the error terms are homoskedastic. The WBRI procedure cannot be expected to work any better than the RI-t procedure when the treated clusters are smaller or larger than the untreated clusters, or when their error terms have different variances.

In the experiments of Figure 2, we used the Rademacher distribution for $G \geq 19$ and the 6-point distribution for $G \leq 18$. This accounts for the sharp jump between 18 and 19. Rejection frequencies for small values of G would have been much larger if we had used Rademacher, while those for large values of G would have been noticeably smaller if we had used 6-point. It is not clear why WBRI tends to under-reject for tests with $G_1 = 1$ (but not for tests with $G_1 = 2$; see below) when the 6-point distribution is used. As MacKinnon and Webb (2017b, 2018b) showed, OLS residuals have strange properties when just one cluster is treated. We speculate that these cause the choice of the wild bootstrap auxiliary distribution to be unusually important in this case.

Figure 3 shows rejection frequencies for tests at the .05 level for five procedures. For readability, the vertical axis has been subjected to a square root transformation, and the conventional RI procedures have been dropped. The results for WBRI-t are the same ones

Figure 3: WBRI, Smoothed RI, and CMR Rejection Frequencies, $G_1 = 1$



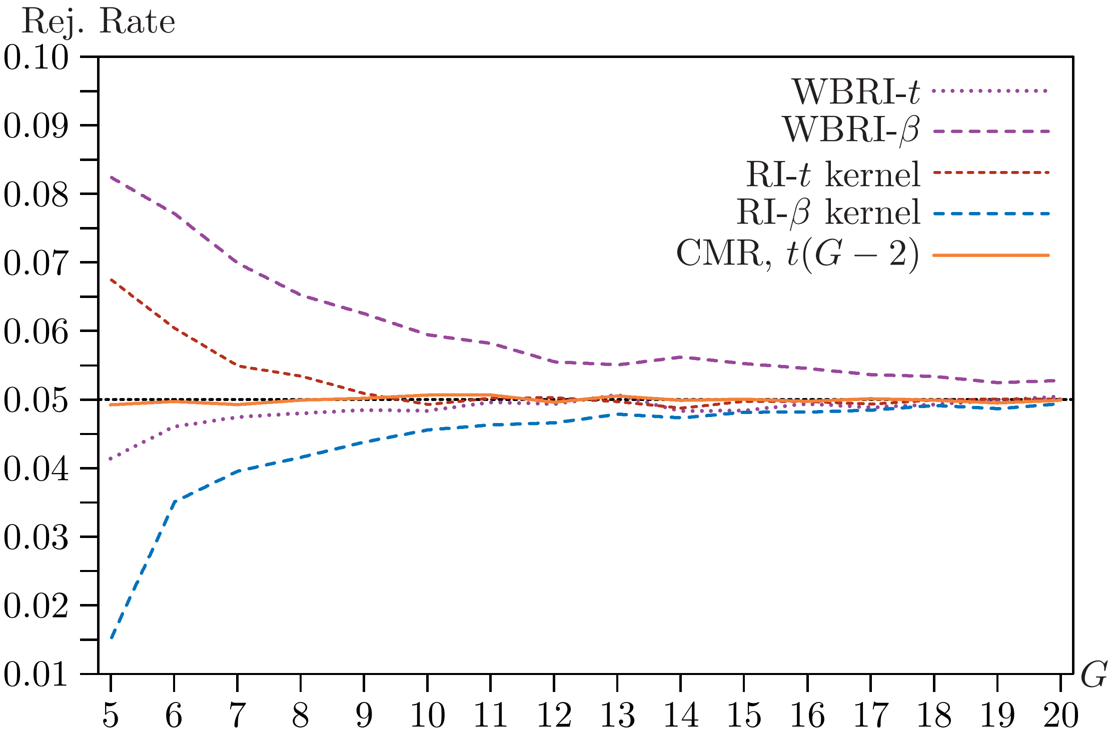
shown in Figure 2. Results for WBRI-β are also shown, and it is evident that WBRI-β always rejects more often than WBRI-t. The difference is quite substantial for small values of G, and WBRI-t is clearly preferred.

In Figure 3, the two tests based on kernel-smoothed P values work remarkably well for larger values of G (say $G > 25$), but they over-reject quite severely for really small values. The over-rejection is more severe for RI-t than for RI-β. All reported results are for $c = 1.575$, the value suggested in Racine and MacKinnon (2007a) for tests at the .05 level. When the larger value of 2.418 (suggested for tests at the .01 level) was used instead, all rejection frequencies were noticeably lower.

The test based on the CMR (11) and the $t(G-2)$ distribution works remarkably well even for very small values of G. It over-rejects slightly when G is small and under-rejects slightly when G is large. It would have performed even better if the errors had been normally rather than lognormally distributed. Since this test is very easy to perform (it requires neither randomization nor the bootstrap), one might well feel, on the basis of these results, that there is no point worrying about the more complicated procedures based on individual data. However, this test does have one serious limitation. As we will see below, it can be seriously lacking in power.

All the experiments reported so far have just one treated group. This is generally the most difficult case. In Figure 4, we show results for several tests with $G_1 = 2$. For these tests, the values of G vary from 5 to 20, and the values of S consequently vary from 9 to 189. The CMR works extremely well for all values of G. WBRI-t (using the 6-point distribution for $G \leq 13$ and the Rademacher distribution for $G \geq 14$) under-rejects slightly for very small values of G but works very well whenever $G \geq 11$. Smoothed RI-t over-rejects for

Figure 4: WBRI, Smoothed RI, and CMR Rejection Frequencies, $G_1 = 2$



very small values of G but works very well for $G \geq 9$. However, the two procedures that are based on coefficients instead of t statistics do not work particularly well.

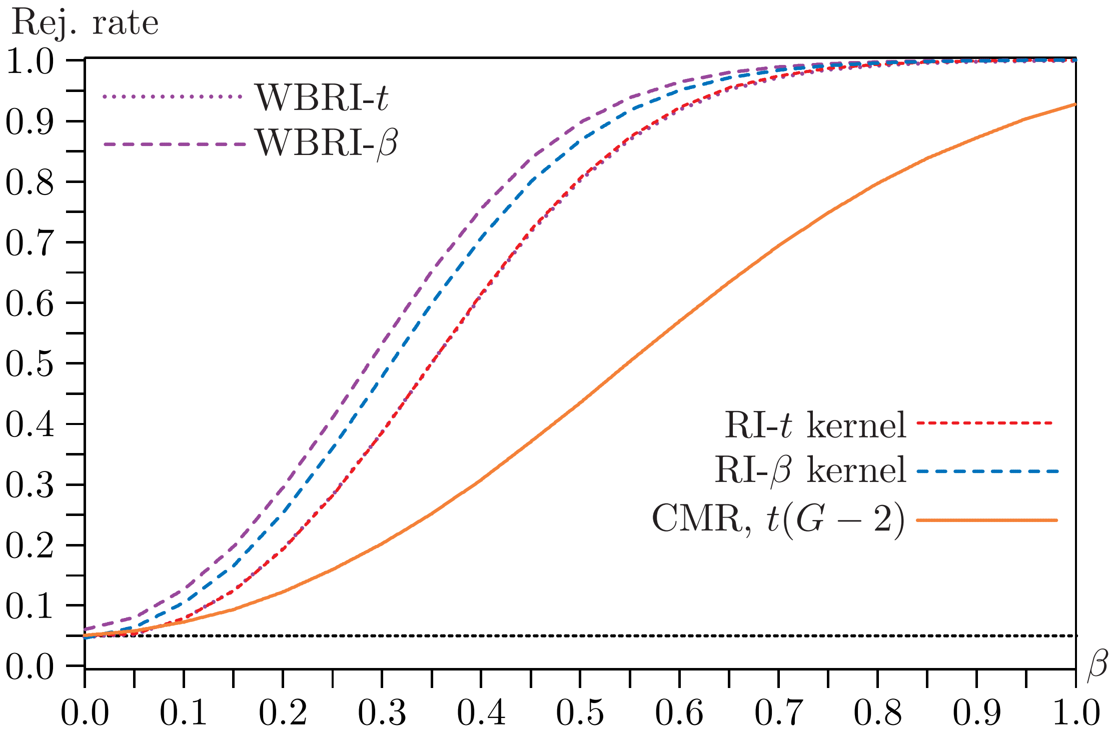
The results presented so far may seem to suggest that the cluster-means regression is the most reliable, as well as the easiest, way to make inferences. However, this approach has one serious shortcoming. When the value of the treatment variable is not constant within groups, aggregation to the group level can seriously reduce power.

Figure 5 presents results for a case with $G = 10$, $G_1 = 2$, and $S = 44$, where the value of β varies from 0 (the null hypothesis) to 1. The most striking result is that tests based on the CMR (11) are much less powerful than the other tests. As noted earlier, in all of our experiments there are 10 "years," only 5 of which are treated. Every cluster has 100 observations, 10 for each "year." Therefore, the regressor $\bar{x}_g$ either takes the value 0 (when cluster g is not treated) or the value 0.5 (when half the observations in cluster g are treated). Not surprisingly, this results in very substantial power loss.[9] Of course, if all the observations in every treated cluster were treated, this power loss would not occur. Additional experiments suggest that, when all "years" are treated, tests based on regression (11) have excellent power.

Some of the other results in Figure 5 are also interesting. The two procedures based on t statistics, WBRI-t and smoothed RI-t, have power functions that are essentially identical. In contrast, the two procedures based on coefficients are noticeably more powerful than the

[9] We note that Donald and Lang (2007) did *not* suggest using equation (11) for DiD models in the way that we have used it here.

Figure 5: Power for Several Procedures



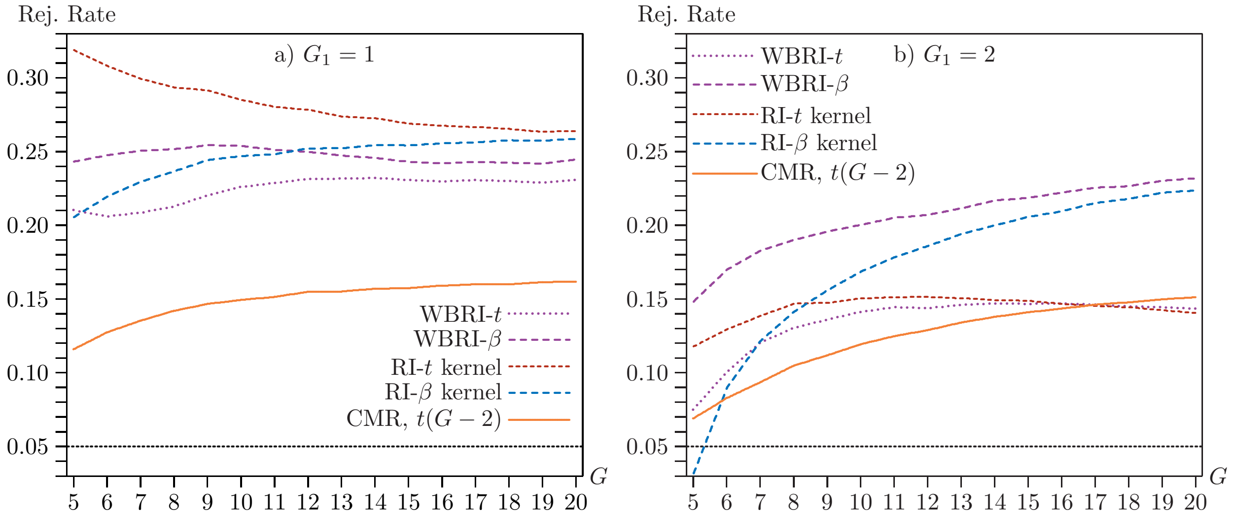
ones based on t statistics. This is consistent with results for RI tests in MacKinnon and Webb (2018a), and it makes sense, because the tests based on coefficients do not have to estimate standard errors. The somewhat higher power of WBRI-β relative to smoothed RI-β can probably be attributed to its somewhat larger size (it rejects 5.92% of the time at the 5% level, versus 4.49%).

It is important to remember that all the procedures we have discussed are very sensitive to the assumption that the clusters are homogeneous. When that assumption is violated, no randomization inference procedure can be expected to perform well, even when G and G_1 are large. Since MacKinnon and Webb (2018a) documents the mediocre performance of RI tests for a number of cases where cluster sizes vary, there is no need to perform similar experiments here. In general, RI tests tend to over-reject when the treated clusters are relatively small and under-reject when they are relatively large.

In Figure 6, we investigate the effects of a particular type of heteroskedasticity which was not studied in MacKinnon and Webb (2018a). Instead of the error terms being homoskedastic, their standard deviation is twice as large for treated observations as for untreated ones. Whether this is a realistic specification is debatable, although it does not seem unreasonable that some treatments could affect the second moment of the outcome as well as the first.

In both panels of Figure 6, G varies between 5 and 20, as in Figures 3 and 4. In the left panel, $G_1 = 1$, and in the right panel, $G_1 = 2$. It is evident that no method yields reliable inferences. The results for $G_1 = 2$ are generally better than for $G_1 = 1$, but they are far from satisfactory. Moreover, the performance of the cluster-means regression and of the two

Figure 6: Rejection Frequencies with Heteroskedasticity



methods based on RI-β actually deteriorates as G increases when $G_1 = 2$.

7 Empirical Example

In this section, we consider an empirical example from Decarolis (2014). Part of the analysis deals with how the introduction of first price auctions (FPA) in Italy affected winning discounts in public works procurement. From January 2000 to June 2006, the use of average bid auctions (ABA) was required for all contracts with reserve prices below €5 million. However, after a case of collusion in ABAs was discovered, the Municipality of Turin and the County of Turin switched from ABAs to FPAs in early 2003. The central government mounted a legal challenge against these reforms that essentially prevented all other public administrations (PA) from making a similar switch.

The timing and exclusivity of the switch in Turin is exploited to estimate a regression analogous to difference-in-differences. Each of the two treated PAs (the county and the municipality) is considered separately in the following model:

$$\text{W.Discount}_{ist} = a_s + b_t + cX_{ist} + \beta\text{FPA}_{st} + \epsilon_{ist}. \tag{12}$$

The outcome of interest, W.Discount_{ist}, is the winning discount offered in auction i of PA s in year t. FPA is a binary indicator equal to 1 for an FPA and 0 otherwise. The coefficient of interest, β, is the effect of using an FPA on the winning discount conditional on fixed effects for PA (a_s), time (b_t), and other covariates (X_{ist}). Analysis is restricted to public works auctions with reserve prices between €300,000 and €5 million, consisting of simple work types such as roadwork construction and repair jobs.

Table 7 presents our results. We first recreate the first two columns of Table 5 in Decarolis (2014). That paper implements a matching strategy, based on similarities in total number of auctions held in each PA during the sample period, to define control groups from other jurisdictions for each of the two treated PAs. This results in 14 control groups for the

Municipality of Turin and 17 control groups for the County of Turin. Thus, $G = 15$ for the Municipality of Turin, and $G = 18$ for the County of Turin, with $G_1 = 1$ in both cases. In the municipality regression, Turin is the largest cluster with 200 observations out of 1,262; the smallest cluster has 28. In the county regression, Turin is again the largest cluster with 147 observations out of 1,355; the smallest cluster has 27. Results in MacKinnon and Webb (2018a) suggest that the RI tests should be conservative when the largest clusters are treated, as is the case in both samples here.

The model above is used to estimate 95% confidence intervals for β under two specifications. Both specifications control for year, PA, a municipality dummy, type of public work dummies, and reserve price. The first specification, which we call Model 1 and is called "W. Discount (1)" in the paper, controls for fiscal efficiency, the ratio of total yearly realized revenue to estimated revenue of the PA. The second specification, which we call Model 2, and is called "W. Discount (2)" in the paper, controls for time trends and PA-specific time trends, but not fiscal efficiency. For each panel, the first and second rows provide estimates when standard errors are clustered at the PA-Year and PA levels, while the third row uses the method of constructing confidence intervals proposed in Conley and Taber (2011).[10]

In addition to reproducing the original results, we compute RI-β and RI-t P values using both formulae, as well as smoothed P values and both types of WBRI P value using the same two samples and two models. We do this clustering only by PA. As expected, the RI-β P values are identical to the RI-t P values for Model 1, because there is only one treated cluster; see MacKinnon and Webb (2018a) for details.[11] The four RI P value intervals for Model 1 contain .05, while the four RI P value intervals for Model 2 contain .10. In the former case, this makes it impossible to tell whether we should reject or not reject at the .05 level. In the latter case, we evidently cannot reject at the .05 level, but it is impossible to tell whether we should reject or not reject at the .10 level.

The WBRI-t P values shown in the table are obtained with $B = 700$ for Panel A and $B = 600$ for Panel B. This means that there are $701 \times {}_{15}C_1 - 1 = 10{,}514$ and $601 \times {}_{18}C_1 - 1 = 10{,}817$ bootstrap/randomization t statistics, respectively. Under Model 1, we find WBRI-t P values that are very close to $\hat{p}_1^*$ and highly significant. Under Model 2, we again find that the WBRI-t P value is very close to $\hat{p}_1^*$ for the municipality sample, but below $\hat{p}_1^*$ for the county sample. Except for Model 2 using the county sample, the smoothed RI-t P values are very similar to the WBRI-t ones. The WBRI-β P values are in general similar to both the WBRI-t values and the smoothed RI-β P values. Interestingly, for Model 2 using both samples, the WBRI-β P values are below $\hat{p}_1^*$.

We also consider an aggregation procedure, which we call cluster-means regression, or CMR, that is similar to one suggested in Donald and Lang (2007). This procedure yields sensible results for Model 1 for both samples. However, for Model 2, it yields much larger P values than any of the other procedures. This is probably a consequence of the fact that Model 2 contains both a DiD term for just one cluster in addition to a time trend for only that cluster, which does not fit easily into the aggregation framework of equation (11).

[10] Following the original paper, confidence intervals for the CT procedure are rounded to the nearest integer values.

[11] We should not expect them to be the same for Model 2, however, because there are two variables that need to be randomized, the DiD variable and the trend-treatment variable.

The evidence against the null hypothesis is probably even stronger than these results suggest. In MacKinnon and Webb (2018a), we showed that RI procedures tend to under-reject when the treated clusters are unusually large. Since the only treated cluster is either the Municipality or the County of Turin, and each of those is the largest cluster in its sample, we would expect all forms of RI P value to be biased upwards. Thus the fact that the WBRI-t test rejects at the .001 level for Model 1 for both datasets and at either the .05 or .10 level for Model 2 suggests that there is quite strong evidence against the null hypothesis.

8 Conclusion

We introduce a bootstrap-based modification of randomization inference which can solve the problem of interval P values when there are few possible randomizations, a problem that often arises when there are very few treated groups. This procedure, which we call WBRI for "wild bootstrap randomization inference," is easiest to understand as a modified version of the wild cluster bootstrap. Like the WCB, it generates a large number of bootstrap samples and uses them to compute bootstrap test statistics. However, unlike the WCB, only some of the bootstrap test statistics are testing the actual null hypothesis. Most of them are testing fictional null hypotheses obtained by re-randomizing the treatment.

The WBRI procedure can be used to generate as many bootstrap test statistics as desired by making B large enough. Thus it can solve the problem of interval P values. However, it shares some of the properties of RI procedures, which perform conventional randomization inference based on either coefficients or cluster-robust t statistics; see MacKinnon and Webb (2018a). In particular, like RI-β and RI-t, WBRI-β and WBRI-t can be expected to over-reject (or under-reject) when the treated clusters are smaller (or larger) than the control clusters. This tendency is greater for WBRI-β than for WBRI-t. Thus we cannot expect WBRI procedures to yield reliable inferences in every case.

We also consider two other procedures. One of them applies the kernel-smoothed P value approach of Racine and MacKinnon (2007a) to randomization inference. This method seems to perform very similarly to WBRI in many cases. The other, based on Donald and Lang (2007), aggregates individual data to the cluster level and uses the t distribution with degrees of freedom equal to the number of clusters minus 2. This cluster-means regression approach can work remarkably well in some cases, but it can be seriously lacking in power when not all observations within treated clusters are treated.

References

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

Bell, Robert M., and Daniel F. McCaffrey (2002) 'Bias reduction in standard errors for linear regression with multi-stage samples.' *Survey Methodology* 28(2), 169–181

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

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

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

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

Canay, Ivan A., Joseph P. Romano, and Azeem M. Shaikh (2017) 'Randomization tests under an approximate symmetry assumption.' *Econometrica* 85(3), 1013–1030

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2017) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' *Review of Economics and Statistics* 99(4), 698–709

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

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

Decarolis, Francesco (2014) 'Awarding Price, Contract Performance, and Bids Screening: Evidence from Procurement Auctions.' *American Economic Journal: Applied Economics* 6(1), 108–132

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

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

Ferman, Bruno, and Christine Pinto (2017) 'Inference in differences-in-differences with few treated groups and heteroskedasticity.' Technical Report, Sao Paulo School of Economics

Fisher, R.A. (1935) *The Design of Experiments* (Edinburgh: Oliver and Boyd)

Imbens, Guido W., and Donald B. Rubin (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press)

Imbens, Guido W., and Michal Kolesár (2016) 'Robust standard errors in small samples: Some practical advice.' *Review of Economics and Statistics* 98(4), 701–712

Lehmann, E. L., and Joseph P. Romano (2008) *Testing Statistical Hypotheses* (New York: Springer)

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

MacKinnon, James G., and Matthew D. Webb (2017a) 'Pitfalls when estimating treatment effects using clustered data.' *The Political Methodologist* 24(2), 20–31

MacKinnon, James G., and Matthew D. Webb (2017b) 'Wild bootstrap inference for wildly different cluster sizes.' *Journal of Applied Econometrics* 32(2), 233–254

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

MacKinnon, James G., and Matthew D. Webb (2018b) 'The wild bootstrap for few (treated) clusters.' *Econometrics Journal* 21(2), 114–135

Racine, Jeffrey S., and James G. MacKinnon (2007a) 'Inference via kernel smoothing of bootstrap P values.' *Computational Statistics & Data Analysis* 51(12), 5949–5957

Racine, Jeffrey S., and James G. MacKinnon (2007b) 'Simulation-based tests that can use any number of simulations.' *Communications in Statistics: Simulation and Computation* 36(2), 357–365

Roodman, David, James G. MacKinnon, Morten Ø. Nielsen, and Matthew D. Webb (2018) 'Fast and wild: Bootstrap inference in stata using boottest.' Working Paper 14xx, Queen's University, Department of Economics

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

Young, Alwyn (2015) 'Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.' Technical Report, London School of Economics

Table 1: 95% Confidence Intervals and P values for FPA coefficient

	Model 1	Model 2
Panel A: Municipality of Turin		
$\hat{\beta}$	12.18	6.14
t statistic	14.86	7.82
PA-Year Clustering (CI)	(9.54, 14.81)	(3.55, 8.72)
PA Clustering (CI)	(10.42, 13.94)	(4.45, 7.82)
CMR P value	0.0203	0.6698
Conley-Taber (CI)	(10, 16)	(5, 8)
RI-β P values	(0.000, 0.063)	(0.133, 0.188)
Smoothed RI-β P value	0.0000	0.0885
RI-t P values	(0.000, 0.063)	(0.067, 0.125)
Smoothed RI-t P value	0.0000	0.0716
WBRI-t P value	0.0000	0.0799
WBRI-β P value	0.0000	0.0595
N	1,262	1,262
G	15	15
Panel B: County of Turin		
$\hat{\beta}$	8.71	5.69
t statistic	19.22	8.34
PA-Year Clustering (CI)	(6.55, 10.85)	(3.19, 8.18)
PA Clustering (CI)	(7.75, 9.66)	(4.25, 7.12)
CMR P value	0.0041	0.4684
Conley-Taber (CI)	(7, 14)	(4, 8)
RI-β P values	(0.000, 0.056)	(0.118, 0.187)
Smoothed RI-β P value	0.0004	0.1046
RI-t P values	(0.000, 0.056)	(0.058, 0.111)
Smoothed RI-t P value	0.0000	0.0570
WBRI-t P value	0.0014	0.0181
WBRI-β P value	0.0000	0.0446
N	1,355	1,355
G	18	18
Regressors		
Fiscal Efficiency	Yes	No
PA Specific Time Trends	No	Yes

Notes: Entries of the form (0.000, 0.067) represent the P value pairs $(\hat{p}_1^*,\ \hat{p}_2^*)$. WBRI P values are obtained with $B = 700$ for Panel A and $B = 600$ for Panel B, ensuring that $B{\times}_G C_1 > 10,000$ for both panels.