Econometrics and Statistics 26 (2023) 52–71

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

Fast cluster bootstrap methods for linear regression models

James G. MacKinnon

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

ARTICLE INFO

Article history:
Received 27 February 2021
Revised 22 November 2021
Accepted 22 November 2021
Available online 27 November 2021

Keywords:
cluster-robust variance estimator
CRVE
wild cluster bootstrap
pairs cluster bootstrap
wild restricted efficient cluster bootstrap
bootstrap Wald test

ABSTRACT

Efficient computational algorithms for bootstrapping linear regression models with clustered data are discussed. For ordinary least squares (OLS) regression, a new algorithm is provided for the pairs cluster bootstrap, along with two algorithms for the wild cluster bootstrap. One of these is a new way to express an existing method. For instrumental variables (IV) regression, an efficient algorithm is provided for the wild restricted efficient cluster (WREC) bootstrap. All computations are based on matrices and vectors that contain sums of squares and cross-products for the observations within each cluster, which have to be computed just once before the bootstrap loop begins. Monte Carlo experiments are used to study the finite-sample properties of bootstrap Wald tests for OLS regression and of WREC bootstrap tests for IV regression.

© 2021 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

1. Introduction

Inference in regression models based on cluster-robust variance estimators can be seriously unreliable when the number of clusters is small, the clusters are not fairly homogeneous both in size and in the features of the regressand and regressors, or the key regressor is a treatment dummy and few clusters are treated. Recent papers that provide theoretical results and/or simulation evidence on which these statements are based include MacKinnon and Webb (2017a, 2017b, 2018), Pustejovsky and Tipton (2018), Djogbenou et al. (2019), and Canay et al. (2021). Unless there are many clusters, and they are all quite homogeneous, it can be dangerous to rely on conventional cluster-robust t-statistics, Wald statistics, and confidence intervals. The tests can over-reject severely, and the confidence intervals may be much too short.

Instead of using procedures based directly on cluster-robust standard errors, it is usually better to base inferences on the restricted wild cluster (WCR) bootstrap, which was proposed in Cameron et al. (2008) and proven to be asymptotically valid in Djogbenou et al. (2019). This method is implemented in the `boottest` package for Stata; see Roodman et al. (2019). The package uses a computational trick that allows it to bootstrap cluster-robust OLS t-statistics and confidence intervals with extraordinary speed when the number of clusters is not too large, even when the sample size is enormous. This paper provides a new algebraic implementation of this procedure, which provides some useful insights but is no faster; see Section 3.2.

The main contribution of the paper is to propose a different computational approach, which applies to the pairs cluster bootstrap (Section 3.1) as well as the wild cluster bootstrap (Section 3.2). In the latter case, it is generally not as fast as the `boottest` approach, but it is easier to understand. Most importantly, the new approach can be used for IV estimation (Section 4), where existing methods can be very slow when the sample size is large.

In addition to being computationally attractive, the methods proposed in this paper are conceptually simple. The key idea is that all computations are based on matrices and vectors that contain sums of products and cross-products over all

E-mail address: mackinno@queensu.ca

https://doi.org/10.1016/j.ecosta.2021.11.009
2452-3062/© 2021 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

the observations within each cluster. Once those vectors and matrices have been computed, they can be used for all the bootstrap samples, and subsequent computational costs are independent of the sample size. This idea can also be used in other contexts. In particular, it makes it easy to compute measures of influence and leverage at the cluster level, a topic that is explored in MacKinnon et al. (2021). This paper, however, focuses exclusively on its application to several bootstrap methods.

Section 2 discusses some key ideas. Then Section 3 shows how to compute pairs cluster bootstrap P values and wild cluster bootstrap P values for both t-tests and Wald tests efficiently. It also discusses wild cluster bootstrap confidence intervals. Section 4 is concerned with IV estimation, where the bootstrap computations are much more complicated. It deals with a bootstrap method called the wild restricted efficient cluster (WREC) bootstrap, an extension of one proposed for models with heteroskedasticity but without clustering in Davidson and MacKinnon (2010), and shows how to compute WREC bootstrap P values efficiently. Section 5 presents simulation results which illustrate the enormous time savings that can be achieved by using these methods. Section 6 presents evidence from Monte Carlo experiments on the finite-sample properties of bootstrap Wald tests for OLS regression and of WREC bootstrap tests for IV regression. Finally, Section 7 concludes.

2. Background and Key Ideas

When the data have been divided into G disjoint clusters and ordered by cluster, the linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ can be written as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{u}_g, \quad g = 1, \ldots, G, \tag{1}$$

where $\boldsymbol{X}_g$ is an $N_g \times k$ matrix of exogenous regressors, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients, $\boldsymbol{y}_g$ is an $N_g \times 1$ vector of observations on the regressand, and $\boldsymbol{u}_g$ is an $N_g \times 1$ vector of disturbances (or error terms). Since the g^{th} cluster is assumed to have N_g observations, the sample size is $N = \sum_{g=1}^{G} N_g$.

The OLS estimator of $\boldsymbol{\beta}$ is, of course, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$. It follows that, when the data-generating process (DGP) is a special case of (1),

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{u}_g = \left(\sum_{g=1}^{G}\boldsymbol{X}_g^\top\boldsymbol{X}_g\right)^{-1}\sum_{g=1}^{G}\boldsymbol{s}_g, \tag{2}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, and $\boldsymbol{s}_g = \boldsymbol{X}_g^\top\boldsymbol{u}_g$ denotes the $k \times 1$ score vector corresponding to the g^{th} cluster. It is assumed that

$$\mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_g^\top) = \boldsymbol{\Sigma}_g, \quad \text{and} \quad \mathrm{E}(\boldsymbol{s}_g\boldsymbol{s}_{g'}^\top) = \boldsymbol{0}, \quad g, g' = 1, \ldots, G, \quad g' \neq g, \tag{3}$$

where the expectations here are conditional on the $\boldsymbol{X}_g$. The matrix $\boldsymbol{\Sigma}_g$, a $k \times k$ symmetric, positive semi-definite matrix, is the (conditional) covariance matrix of the score vector for the g^{th} cluster. It follows from (2) and (3) that the (conditional) covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\boldsymbol{\Sigma}_g\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{4}$$

This well-known result is often stated in a somewhat different way, with the matrix $\boldsymbol{\Sigma}_g$ replaced by $\boldsymbol{X}_g^\top\boldsymbol{\Omega}_g\boldsymbol{X}_g$, where $\boldsymbol{\Omega}_g$ denotes the covariance matrix of $\boldsymbol{u}_g$. However, (4) is a more informative way to write $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, because it makes it clear that the key things to estimate are the $\boldsymbol{\Sigma}_g$, that is, the covariance matrices of the score vectors.

It is natural to estimate the $\boldsymbol{\Sigma}_g$ by using the outer products of the empirical score vectors $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top\hat{\boldsymbol{u}}_g$, in which the disturbance subvectors $\boldsymbol{u}_g$ are replaced by the residual subvectors $\hat{\boldsymbol{u}}_g$. If in addition we multiply by a correction for degrees of freedom, we obtain the most widely-used cluster-robust variance estimator, or CRVE,

$$\mathrm{CV}_1\colon \quad \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{G(N-1)}{(G-1)(N-k)}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\left(\sum_{g=1}^{G}\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top\right)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}, \tag{5}$$

which just depends on the $\boldsymbol{X}^\top\boldsymbol{X}$ matrix and the empirical score vectors. Note that each of the $\hat{\boldsymbol{s}}_g\hat{\boldsymbol{s}}_g^\top$ matrices has rank at most 1, so that (5) has rank at most G (in many cases, it will have rank $G-1$). This suggests, correctly, that asymptotic inference based on CV_1 may be unreliable when G is not large, especially when there is more than one restriction. It is therefore common to use bootstrap methods, as discussed in Section 1.

Remark 1. If interest focuses on a single element of $\boldsymbol{\beta}$, say β_j, then the first and second instances of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ in (5) can be replaced by the j^{th} row and the j^{th} column of that symmetric matrix, respectively.

The basic insight of this paper is that, for the model (1), both the OLS estimates $\hat{\boldsymbol{\beta}}$ and the variance matrix (5) depend on $\boldsymbol{X}$ and $\boldsymbol{y}$ only through the matrices $\boldsymbol{X}_g^\top\boldsymbol{X}_g$ and the vectors $\boldsymbol{X}_g^\top\boldsymbol{y}_g$, for $g = 1, \ldots, G$. These quantities may be thought of as sufficient statistics for the parameter estimates, their estimated covariance matrices, and the bootstrap samples. The first

step in all the proposed procedures is therefore to calculate these sufficient statistics. Doing so requires computation time that is $O(N)$. The sufficient statistics may then be used to compute the OLS estimates

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{X}_g \right)^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^\top \boldsymbol{y}_g \tag{6}$$

and the empirical scores

$$\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top (\boldsymbol{y}_g - \boldsymbol{X}_g \hat{\boldsymbol{\beta}}) = \boldsymbol{X}_g^\top \boldsymbol{y}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g \hat{\boldsymbol{\beta}}, \quad g = 1, \ldots, G. \tag{7}$$

Notice that there is no need to calculate the residual vector $\hat{\boldsymbol{u}}$. Given the empirical scores $\hat{\boldsymbol{s}}_g$ and the matrix $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, it is easy to calculate CV_1 using (5).

Remark 2. In order to obtain OLS estimates that are as numerically accurate as possible, it is often recommended to use a QR decomposition rather than relying on matrix inversion as in (6). However, experience suggests that, unless the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is extremely ill-conditioned, computations based on (6), where the matrix inversion is performed using a Cholesky decomposition, are more than accurate enough.

3. Bootstrap Computations for OLS Estimation

The key feature of efficient computational methods for bootstrapping the linear regression model (1) is that they start by computing the sufficient statistics $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and $\boldsymbol{X}_g^\top \boldsymbol{y}_g$ for every one of the G clusters. These computations, which are evidently $O(k^2 N)$, should take advantage of the fact that the $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ matrices are symmetric. All of the quantities needed for bootstrap inference are subsequently calculated using these sufficient statistics, without the need for any more computations that are $O(N)$. Thus the cost of each bootstrap replication depends on G and k but not on N. If instead we performed a full OLS regression for every bootstrap sample, the cost of each bootstrap replication would be $O(k^2 N)$.

For concreteness, consider testing the hypothesis that $\boldsymbol{a}^\top \boldsymbol{\beta} = 0$, where $\boldsymbol{a}$ is a known k-vector. Then the bootstrap methods for OLS regression all work as follows:

1. Obtain $\hat{\boldsymbol{\beta}}$ using (6) and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ using (5), and use them to compute the test statistic $t_a = \boldsymbol{a}^\top \hat{\boldsymbol{\beta}} / \big(\boldsymbol{a}^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \boldsymbol{a}\big)^{1/2}$. If necessary, re-estimate the model under the null hypothesis to obtain restricted estimates $\tilde{\boldsymbol{\beta}}$; see Section 3.4.
2. Compute bootstrap test statistics t_a^{*b} for B bootstrap samples indexed by b. Just how this can be done efficiently varies from case to case and will be discussed below. Here B should be chosen so that $\alpha(B+1)$ is an integer for any test level α of interest (Racine and MacKinnon, 2007, Section 2). Common values are 999, 9,999 and 99,999.
3. Calculate the bootstrap P value corresponding to the alternative hypothesis of interest. The P value could be left-tailed, right-tailed, equal-tailed, or symmetric (Djogbenou et al., 2019, Section 3.1). The symmetric bootstrap P value is

$$\hat{P}_{\mathrm{S}}^* = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\big(|t_a^{*b}| > |t_a|\big), \tag{8}$$

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

In order to test two or more linear restrictions, say $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, where the matrix $\boldsymbol{R}$ is $r \times k$ and the vector $\boldsymbol{r}$ is $r \times 1$, we can use the Wald statistic

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \big(\boldsymbol{R} \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \boldsymbol{R}^\top\big)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \tag{9}$$

and compute the right-tailed bootstrap P value

$$\hat{P}^*\big(W(\hat{\boldsymbol{\beta}})\big) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\big(W_b^* > W(\hat{\boldsymbol{\beta}})\big), \tag{10}$$

where W_b^* is the Wald statistic for the b$^{\text{th}}$ bootstrap sample; see Section 3.2.

Just how to generate the bootstrap samples varies according to the bootstrap method used. The key is never actually to generate all N observations of a bootstrap dataset. Section 3.1 discusses the pairs cluster bootstrap, and Section 3.2 discusses the wild cluster bootstrap, both restricted and unrestricted. Section 3.3 considers the special case of dummy variables for fixed effects. Section 3.4 explains how to obtain bootstrap confidence intervals by inverting tests based on the restricted wild cluster bootstrap.

3.1. The Pairs Cluster Bootstrap

The pairs cluster bootstrap is very close in spirit to the original resampling bootstrap of Efron (1979). Conceptually, we arrange the data into G pairs, $[\boldsymbol{X}_g, \boldsymbol{y}_g]$, one for each cluster. Then each bootstrap sample consists of G pairs $[\boldsymbol{X}_g^*, \boldsymbol{y}_g^*]$, resampled with replacement from the original G pairs and stacked to form a matrix $\boldsymbol{X}^*$ and a vector $\boldsymbol{y}^*$.

For the model (1), however, there is no need to form $\boldsymbol{X}^*$ and $\boldsymbol{y}^*$. Instead, after we have computed the pairs of matrices and vectors $[\boldsymbol{X}_g^\top \boldsymbol{X}_g, \boldsymbol{X}_g^\top \boldsymbol{y}_g]$ for each cluster, as discussed at the beginning of Section 3, we can resample directly from them. The b$^{\text{th}}$ bootstrap sample then consists of

$$[\boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b}, \boldsymbol{X}_g^{*b\top} \boldsymbol{y}_g^{*b}], \quad g = 1, \ldots, G, \tag{11}$$

where each of the bootstrapped pairs in (11) is chosen with replacement with probability $1/G$ from the original pairs $[\boldsymbol{X}_g^\top \boldsymbol{X}_g, \boldsymbol{X}_g^\top \boldsymbol{y}_g]$. The bootstrap estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{*b} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b} \right)^{-1} \sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{y}_g^{*b}, \tag{12}$$

and the bootstrap empirical scores, which are needed for the CRVE, are

$$\hat{\boldsymbol{s}}_g^{*b} = \boldsymbol{X}_g^{*b\top} \boldsymbol{y}_g^{*b} - \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b} \hat{\boldsymbol{\beta}}^{*b}. \tag{13}$$

Importantly, the $\hat{\boldsymbol{s}}_g^{*b}$ can be calculated without using the residuals $\hat{\boldsymbol{u}}^{*b}$, which never need to be computed. Since the calculations in (12) and (13) are both $O(k^2 G)$, they should not be demanding unless k is large and/or G is extremely large.

The CRVE for $\hat{\boldsymbol{\beta}}^{*b}$ has exactly the same form as (5):

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{*b}) = \frac{G(N-1)}{(G-1)(N-k)} \left(\sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b} \right)^{-1} \left(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g^{*b} (\hat{\boldsymbol{s}}_g^{*b})^\top \right) \left(\sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b} \right)^{-1}, \tag{14}$$

and the bootstrap test statistic for $\boldsymbol{a}^\top \boldsymbol{\beta} = 0$ is given by

$$t_a^{*b} = \frac{\boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^{*b} - \hat{\boldsymbol{\beta}})}{\left(\boldsymbol{a}^\top \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{*b}) \boldsymbol{a} \right)^{1/2}}. \tag{15}$$

It is usually not expensive to compute (15). The outer factors in (14) were already computed for (12). Forming the middle factor is $O(k^2 G)$, since it just involves summing G outer products. Moreover, in many cases, the denominator of t_a^{*b} can be computed without calculating the entire covariance matrix; see Remark 1. However, the pairs cluster bootstrap is considerably more expensive than the wild cluster bootstrap, to be discussed in Section 3.2. Because the matrix of sums of squares and cross-products of the regressors, that is, $\sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b}$, is different for every bootstrap sample, this matrix needs to be created and inverted B times.

In certain cases, it may be impossible to calculate (15) for every bootstrap sample. Suppose, for example, that one of the regressors equals 0 for all observations except some (or all) of the ones in clusters 1 and 2. Then, if the b$^{\text{th}}$ bootstrap sample happens to omit both clusters 1 and 2, the matrix $\sum_{g=1}^{G} \boldsymbol{X}_g^{*b\top} \boldsymbol{X}_g^{*b}$ will be singular, making it impossible to compute $\hat{\boldsymbol{\beta}}^{*b}$. This sort of problem can arise when inference concerns treatment at the cluster level, and few clusters are treated; see MacKinnon and Webb (2017b, 2018). If only a few bootstrap samples are affected, it may be reasonable just to discard them. But if more than a handful of bootstrap samples have to be discarded, it would probably be unwise to base inference on the pairs cluster bootstrap.

Of course, the fact that it is now possible to calculate pairs cluster bootstrap P values fairly inexpensively does not mean that this is a good procedure to use. Because the null hypothesis is not imposed on the bootstrap samples, and because the number of observations varies across bootstrap samples unless all clusters are the same size, it seems likely that the pairs cluster bootstrap may not always perform particularly well. Indeed, simulation evidence suggests that this is often the case; see Cameron et al. (2008), MacKinnon and Webb (2017b), and Section 6.1 below.

3.2. The Wild Cluster Bootstrap

The wild cluster bootstrap (WCB) was proposed in Cameron et al. (2008), and its asymptotic validity was proved in Djogbenou et al. (2019). In many cases, the restricted version of the WCB seems to provide particularly reliable inferences, although (like other methods) it can fail in certain cases. In particular, it can under-reject severely in models for treatment and difference-in-differences when the number of treated clusters is small (MacKinnon and Webb, 2017a; 2017b; 2018), and it can over-reject seriously when cluster sizes vary a lot.

The key idea of the wild cluster bootstrap is to multiply the entire vector of residuals for each cluster g by a single auxiliary random variable v_g^*. Unless G is very small (Webb, 2014), the best choice for the distribution of v_g^* seems to be the Rademacher distribution, which takes the values 1 and -1 with equal probability (Davidson and Flachaire, 2008; Djogbenou et al., 2019). There are two variants of the WCB. One of them uses unrestricted parameter estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{u}}$; it is called the unrestricted wild cluster (WCU) bootstrap. The other estimates the model subject to the restrictions and

uses restricted estimates $\tilde{\boldsymbol{\beta}}$ and restricted residuals $\tilde{\boldsymbol{u}}$ in the bootstrap DGP; it is called the restricted wild cluster (WCR) bootstrap.

The empirical score vectors needed for the WCU bootstrap have already been computed in (7). For the WCR bootstrap, it is necessary to estimate the model subject to the restrictions and then compute the restricted empirical score vectors $\tilde{\boldsymbol{s}}_g$ as $\boldsymbol{X}_g^\top \boldsymbol{y}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g \tilde{\boldsymbol{\beta}}$. This can often be done quite efficiently by reusing calculations for the unrestricted model. Suppose, for example, that

$$\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{u},$$

where the restriction is that $\beta_2 = 0$. Then

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\ 0 \end{bmatrix},$$

where $\boldsymbol{X}_1^\top \boldsymbol{X}_1$ is the upper left $(k-1) \times (k-1)$ block of the matrix $\boldsymbol{X}^\top \boldsymbol{X}$, and $\boldsymbol{X}_1^\top \boldsymbol{y}$ contains the first $k-1$ elements of the vector $\boldsymbol{X}^\top \boldsymbol{y}$.

The wild cluster bootstrap uses the same regressors for every bootstrap sample, and the bootstrap disturbances only affect the estimates through the scores. Thus generating the b$^{\text{th}}$ bootstrap sample simply requires us to compute the bootstrap score vectors

$$\boldsymbol{s}_g^{*b} = v_g^{*b} \ddot{\boldsymbol{s}}_g, \quad g = 1, \ldots, G, \tag{16}$$

where $\ddot{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \boldsymbol{y}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g \ddot{\boldsymbol{\beta}}$, and $\ddot{\boldsymbol{\beta}}$ denotes either $\hat{\boldsymbol{\beta}}$ or $\tilde{\boldsymbol{\beta}}$. Notice that the vectors $\boldsymbol{s}_g^{*b}$ in (16) are the bootstrap analogs of the score vectors $\boldsymbol{s}_g = \boldsymbol{X}_g^\top \boldsymbol{u}_g$, not the bootstrap analogs of the empirical score vectors $\hat{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \hat{\boldsymbol{u}}_g$.

Instead of using (16), the conventional procedure for the WCB would generate the vectors $\boldsymbol{u}_g^{*b}$ of bootstrap disturbances as

$$\boldsymbol{u}_g^{*b} = v_g^{*b} \ddot{\boldsymbol{u}}_g, \quad g = 1, \ldots, G, \tag{17}$$

and then use them to compute first the score vectors $\boldsymbol{s}_g^{*b}$ and then the vector $\hat{\boldsymbol{\beta}}^* - \ddot{\boldsymbol{\beta}}$; see (18) below. Eqs (17) alone involve calculations that are $O(N)$, and obtaining the score vectors requires further calculations that are $O(kN)$. Thus, by using (16), computations that are $O\big((k+1)N\big)$ have been replaced by ones that are $O(kG)$.

Eq. (16) provides an alternative motivation for the wild cluster bootstrap. The usual one is that multiplying all the residuals in each cluster g by the same random variable, as in (17), preserves the covariance matrix of the residuals. This is true. But it is clear from (16) that doing so also preserves the covariance matrix of the scores. Since it is the bootstrap scores that determine $\hat{\boldsymbol{\beta}}^{*b}$, and the empirical bootstrap scores that determine its covariance matrix, this is actually the critical feature of the WCB.

Using the bootstrap scores from (16), we readily obtain

$$\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sum_{g=1}^{G} \boldsymbol{s}_g^{*b}, \quad \text{and} \tag{18}$$

$$\hat{\boldsymbol{s}}_g^{*b} = \boldsymbol{s}_g^{*b} - \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}}), \quad g = 1, \ldots, G, \tag{19}$$

where the $\hat{\boldsymbol{s}}_g^{*b}$ are the empirical bootstrap scores, that is, the analogs of the $\hat{\boldsymbol{s}}_g$. Observe that (18) is the bootstrap analog of Eq. (2). Moreover, equation (7) implies that $\hat{\boldsymbol{s}}_g = \boldsymbol{s}_g - \boldsymbol{X}_g^\top \boldsymbol{X}_g (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, and (19) is the bootstrap analog of this equation. Using (18) and (19), it is straightforward to compute the wild bootstrap version of (5) and the corresponding bootstrap t-statistic or bootstrap Wald statistic.

This involves considerably more work than necessary, however, because we do not actually need the bootstrap parameter estimates $\hat{\boldsymbol{\beta}}^{*b}$. For the numerator of the bootstrap t-statistic, we can avoid evaluating (18) by using

$$\boldsymbol{a}^\top (\hat{\boldsymbol{\beta}}^{*b} - \ddot{\boldsymbol{\beta}}) = \sum_{g=1}^{G} \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \ddot{\boldsymbol{s}}_g v_g^{*b} = \sum_{g=1}^{G} c_g v_g{}^{*b} = \boldsymbol{c}^\top \boldsymbol{v}_b^*. \tag{20}$$

Here the vector $\boldsymbol{c}$ has typical element $c_g \equiv \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \ddot{\boldsymbol{s}}_g$, which can be calculated prior to the bootstrap loop, and the vector $\boldsymbol{v}_b^*$ has typical element v_g^{*b}.

For the denominator of the bootstrap t-statistic, we can save time by forming the matrices $\boldsymbol{A}_g = \boldsymbol{X}_g^\top \boldsymbol{X}_g (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ before the bootstrap loop begins. Then (19) becomes

$$\hat{\boldsymbol{s}}_g^{*b} = \boldsymbol{s}_g^{*b} - \boldsymbol{A}_g \sum_{h=1}^{G} \boldsymbol{s}_h^{*b}, \quad g = 1, \ldots, G, \tag{21}$$

and the square of the denominator is

$$\boldsymbol{a}^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{*b})\boldsymbol{a} = \frac{G(N-1)}{(G-1)(N-k)} \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \Big(\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g^{*b} (\hat{\boldsymbol{s}}_g^{*b})^\top \Big) (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{a}. \tag{22}$$

Computing (22) is quite inexpensive, but the cost can be reduced further by using a trick proposed in Roodman et al. (2019). The discussion there is for Wald tests and involves some rather complicated matrix algebra. The following discussion is much simpler because it deals only with t-tests.

Before bootstrapping begins, the $G \times G$ matrix $\boldsymbol{H}$ with typical element

$$H_{gh} \equiv \boldsymbol{a}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}_g^\top \boldsymbol{X}_g (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \ddot{\boldsymbol{s}}_h \tag{23}$$

is computed. Then, for each bootstrap replication, the H_{gh} and the c_g are used to calculate

$$\boldsymbol{z}_g^{*b} = v_g^{*b} c_g - \sum_{h=1}^{G} v_h^{*b} H_{gh}, \quad g = 1, \ldots, G. \tag{24}$$

Notice that the empirical bootstrap scores $\hat{\boldsymbol{s}}_g^{*b}$ no longer appear explicitly here. Using (24), we obtain (22) as

$$\boldsymbol{a}^\top \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{*b})\boldsymbol{a} = \frac{G(N-1)}{(G-1)(N-k)} \sum_{g=1}^{G} (\boldsymbol{z}_g^{*b})^2. \tag{25}$$

Dividing $\boldsymbol{c}^\top \boldsymbol{v}_b^*$ from (20) by the square root of (25) yields the wild bootstrap t-statistic t_a^{*b}. The method based on (20), (23), (24), and (25) will be referred to as the boottest method, since this is essentially what the `boottest` package does.

The calculations for the boottest method are remarkably cheap. Once the preliminary work has been done, the effort required for each bootstrap sample is $O(G^2)$ and does not depend on either N or k. In contrast, forming the empirical bootstrap scores in (19) or (21) in order to compute (22) requires computations that are $O(k^2 G)$. Thus we should expect the boottest method to be cheaper than the more straightforward one based on (20) and (22), except perhaps when G is extremely large; see Section 5.1.

Asymptotic cluster-robust Wald tests tend to over-reject more and more severely as the number of restrictions increases (Pustejovsky and Tipton, 2018). Thus it is particularly important to bootstrap these tests; see Section 6.1. Consider the r linear restrictions $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, where $\boldsymbol{R}$ and $\boldsymbol{r}$ are $r \times k$ and $r \times 1$, respectively. The Wald test statistic is $W(\hat{\boldsymbol{\beta}})$ given in (9). If the bootstrap DGP imposes the restrictions $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, as the WCR bootstrap does, then the bootstrap analog of (9) is

$$W_b^* = (\boldsymbol{R}\hat{\boldsymbol{\beta}}^{*b} - \boldsymbol{r})^\top \big(\boldsymbol{R} \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{*b}) \boldsymbol{R}^\top \big)^{-1} (\boldsymbol{R}\hat{\boldsymbol{\beta}}^{*b} - \boldsymbol{r}). \tag{26}$$

Because the pairs cluster and WCU bootstraps do not impose the restrictions, however, expression (26) is not valid for them. Both instances of the vector $\boldsymbol{R}\hat{\boldsymbol{\beta}}^{*b} - \boldsymbol{r}$ need to be replaced by the vector $\boldsymbol{R}(\hat{\boldsymbol{\beta}}^{*b} - \hat{\boldsymbol{\beta}})$.

Given $\hat{\boldsymbol{\beta}}^{*b}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{*b})$, expression (26) for the WCR bootstrap is very easily calculated, but computing time can be further reduced by not explicitly calculating either of them. If instead we compute the $r \times k$ matrix

$$\boldsymbol{B} = \boldsymbol{R}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tag{27}$$

before the bootstrap loop begins, then the WCR bootstrap Wald statistic is simply

$$W_b^* = \frac{(G-1)(N-k)}{G(N-1)} \left(\boldsymbol{B} \sum_{g=1}^{G} \boldsymbol{s}_g^{*b} \right)^\top \left(\boldsymbol{B} \sum_{g=1}^{G} \hat{\boldsymbol{s}}_g^{*b} (\hat{\boldsymbol{s}}_g^{*b})^\top \boldsymbol{B}^\top \right)^{-1} \left(\boldsymbol{B} \sum_{g=1}^{G} \boldsymbol{s}_g^{*b} \right). \tag{28}$$

The equality of (28) and (26) follows from (18), (27), and the fact that $\boldsymbol{r} - \boldsymbol{R}\tilde{\boldsymbol{\beta}} = \boldsymbol{0}$, because $\tilde{\boldsymbol{\beta}}$ satisfies the restrictions. Notice that the covariance matrix inverted in (28) depends on the estimated bootstrap scores, while the vectors that measure how far $\boldsymbol{R}\hat{\boldsymbol{\beta}}^{*b}$ is from $\boldsymbol{r}$ only depend on the actual bootstrap scores generated by (16).

Remark 3. If we are performing a bootstrap test without the corresponding asymptotic test, the leading scalar factor in (28) can be dropped, as long as the corresponding one in (5) is also dropped when computing (9).

The boottest method has been extended to Wald statistics, but the algebra is complicated; see Roodman et al. (2019, Section 5). In most cases, the boottest method should be cheaper than the one based on (28), because it avoids explicitly calculating the estimated bootstrap scores. However, (28) is much easier to program, and it might even be computationally attractive when both G and r are large, because the boottest method requires $r(r+1)/2$ calculations similar to (24) when there are r restrictions.

3.3. Fixed Effects and Bootstrap Computations

Many linear regression models where the data appear to be clustered involve fixed effects, often a large number of them. A great deal of computer time can often be saved by partialing out the fixed effects before running the regression. Using the wild cluster bootstrap with data where fixed effects have been partialed out works as expected, but care must be taken when using the pairs cluster bootstrap.

A linear regression model with fixed effects can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{\eta} + \boldsymbol{u}, \tag{29}$$

where $\boldsymbol{D}$ contains N rows of observations on 0-1 dummy variables. The OLS estimate $\hat{\boldsymbol{\beta}}$ from (29) can be obtained by projecting $\boldsymbol{y}$ and $\boldsymbol{X}$ off the columns of $\boldsymbol{D}$ and then regressing $\boldsymbol{M_D y}$ on $\boldsymbol{M_D X}$, where $\boldsymbol{M_D} = \mathbf{I} - \boldsymbol{D}(\boldsymbol{D}^\top\boldsymbol{D})^{-1}\boldsymbol{D}^\top$ is the matrix that takes residuals from a regression on $\boldsymbol{D}$. When there is just one set of fixed effects, each row of $\boldsymbol{D}$ contains just one "1," so that premultiplying by $\boldsymbol{M_D}$ simply takes deviations from the means over every set of observations to which a fixed effect applies. When there are two or more sets of fixed effects, what it does is more complicated.

For the wild cluster bootstrap, bootstrapping the model (29) by treating $\boldsymbol{M_D y}$ and $\boldsymbol{M_D X}$ as if they were the original data works as expected. We can condition on the matrix $\boldsymbol{X}^\top\boldsymbol{M_D X}$ because it is the same for all bootstrap samples. In addition, since the bootstrap test statistics depend on $\boldsymbol{y}^{*b}$ only through the matrix $\boldsymbol{X}^\top\boldsymbol{M_D y}^{*b}$, it is not necessary to generate the $\boldsymbol{y}^{*b}$ and partial out the fixed effects. We can simply generate the bootstrap scores using (16) as usual.

For the pairs cluster bootstrap, however, resampling from the pairs

$$\left[(\boldsymbol{X}^\top\boldsymbol{M_D X})_g, \quad (\boldsymbol{X}^\top\boldsymbol{M_D y})_g\right] \tag{30}$$

to obtain bootstrap quantities $\boldsymbol{X}_g^{*b\top}\boldsymbol{X}_g^{*b}$ and $\boldsymbol{X}_g^{*b\top}\boldsymbol{y}_g^{*b}$, where the fixed effects have already been partialed out, and then using (12) and (13) to obtain $\hat{\boldsymbol{\beta}}^{*b}$ and $\hat{\boldsymbol{s}}_g^{*b}$, respectively, does not yield the same bootstrap estimates as resampling from the triples

$$\left[\boldsymbol{X}_g, \quad \boldsymbol{D}_g, \quad \boldsymbol{y}_g\right] \tag{31}$$

to obtain $\boldsymbol{X}^{*b}$, $\boldsymbol{D}^{*b}$, and $\boldsymbol{y}^{*b}$, and then regressing $\boldsymbol{y}^{*b}$ on $\boldsymbol{X}^{*b}$ and $\boldsymbol{D}^{*b}$, because $\boldsymbol{D}^{*b}$ is different for every bootstrap sample. However, both resampling procedures should be asymptotically valid, and it is not clear which one is likely to work better in finite samples.

There is one important special case in which both resampling procedures yield identical results. Suppose there is a fixed effect for each cluster and no other fixed effects. In that case, for any vector $\boldsymbol{x}$ with components $\boldsymbol{x}_g$, $(\boldsymbol{M_D x})_g = \boldsymbol{x}_g - \bar{x}_g\boldsymbol{\iota}_g$ for $g = 1, \ldots, G$. Here $\bar{x}_g$ is the sample mean of the elements of $\boldsymbol{x}_g$, and $\boldsymbol{\iota}$ is an N_g-vector of 1s. Thus $(\boldsymbol{M_D x})_g$ is just the vector of deviations of the elements of $\boldsymbol{x}_g$ from their group mean. Because resampling by cluster does not change the group means of the $\boldsymbol{X}_g$ or the $\boldsymbol{y}_g$, resampling from (30) yields the same results as resampling from (31) in this special case.

3.4. Bootstrap Confidence Intervals

Inverting a WCR bootstrap test is often a good way to obtain a confidence interval for one of the coefficients in (1); see MacKinnon (2015). The `boottest` package already calculates these WCR bootstrap intervals, but obtaining them is a bit tricky. Since the discussion of confidence intervals in Roodman et al. (2019) is brief and does not focus on computation, it seems worthwhile to discuss the main issues. For concreteness, the model (1) can be rewritten as

$$\boldsymbol{y}_g = \boldsymbol{X}_{1g}\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_{2g} + \boldsymbol{u}_g, \quad g = 1, \ldots, G, \tag{32}$$

where the objective is to form a confidence interval for the parameter β_2. In (32), the vector $\boldsymbol{x}_{2g}$ contains the observations for cluster g on the regressor associated with β_2, and the $N_g \times (k-1)$ matrix $\boldsymbol{X}_{1g}$ contains those observations for the other regressors.

A bootstrap confidence interval for β_2 can always be obtained by inverting two bootstrap tests, one that β_2 equals the lower limit of the interval, and one that it equals the upper limit. We could use a different one-sided bootstrap P value for each end of the interval, but this is equivalent to using the equal-tail P value

$$\hat{P}^*_{\text{ET}}(\beta_{20}) = \frac{2}{B}\min\left(\sum_{b=1}^{B}\mathbb{I}\big(t_2^{*b} > t_2\big), \sum_{b=1}^{B}\mathbb{I}\big(t_2^{*b} \leq t_2\big)\right), \tag{33}$$

where

$$t_2 = \frac{\hat{\beta}_2 - \beta_{20}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_2)}} \quad \text{and} \quad t_2^{*b} = \frac{\hat{\beta}_2^{*b} - \beta_{20}}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_2^{*b})}}, \quad b = 1, \ldots, B. \tag{34}$$

Here β_{20} denotes one of the limits of the confidence interval, say $\beta_{2\text{l}}$ for the lower limit and $\beta_{2\text{u}}$ for the upper one. Neither $\hat{\beta}_2$ nor its standard error depends on β_{20}, but for the WCR bootstrap both $\hat{\beta}_2^{*b}$ and its standard error do so. Thus, in order

to obtain a WCR bootstrap confidence interval, we need to use an iterative procedure to find the lower and upper limits of the interval. In the case of a lower limit, we need to find a value β_{2l}, almost certainly less than $\hat{\beta}_2$, for which $\hat{P}^*_{\text{ET}}(\beta_{2l})$ is greater than α for $\beta_2 > \beta_{2l}$ and less than α for $\beta_2 < \beta_{2l}$. In the case of an upper limit, we need to find a value β_{2u}, almost certainly greater than $\hat{\beta}_2$, for which the two inequalities are reversed.

Finding a WCR bootstrap confidence interval requires evaluating $\hat{P}^*_{\text{ET}}(\beta_{20})$ quite a few times using the same set of realizations of the auxiliary random variable. Unless BG is enormous, it makes sense to generate all of the v_g^{*b} before the procedure begins and store them. Before we can generate the bootstrap samples for any value β_{20}, we need to obtain the c_g implicitly defined in (20) and the H_{gh} defined in (23), with $\ddot{\boldsymbol{s}}_g = \tilde{\boldsymbol{s}}_g$, the score vector that corresponds to β_{20} and the restricted estimates $\tilde{\boldsymbol{\beta}}_1$ conditional on $\beta_2 = \beta_{20}$. In this case, $\boldsymbol{a}$ is a vector with the first k_1 elements equal to 0 and the k^{th} element equal to 1, so that $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{a}$ is just the last column of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, which may be denoted by $\boldsymbol{e}$. Both the c_g and the H_{gh} depend on β_{20}. We observe that

$$c_g(\beta_{20}) = \boldsymbol{y}_g^\top\boldsymbol{X}_g\boldsymbol{e} - \boldsymbol{y}^\top\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_{1g}^\top\boldsymbol{X}_g\boldsymbol{e} - \beta_{20}\big(\boldsymbol{x}_{2g}^\top\boldsymbol{X}_g\boldsymbol{e} - \boldsymbol{x}_2^\top\boldsymbol{X}_1(\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\boldsymbol{X}_{1g}^\top\boldsymbol{X}_g\boldsymbol{e}\big). \tag{35}$$

The first two terms in (35) are scalars that just need to be computed once. The final term is β_{20} times a scalar that also just needs to be computed once. Everything here depends solely on the matrices $\boldsymbol{X}_g^\top\boldsymbol{X}_g$ and the vectors $\boldsymbol{X}_g^\top\boldsymbol{y}_g$, or components of them.

The H_{gh} defined in (23) can be computed in essentially the same way as (35), but with $\boldsymbol{e}$, which appears four times there, replaced by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}_g^\top\boldsymbol{X}_g\boldsymbol{e}$. Thus updating the two expressions in (23) for a new value of β_{20} requires very little computational effort. For each bootstrap sample, we can then use (20) to obtain $\hat{\beta}_2^{*b}$, and we can use (24) and the square root of (25) to obtain the bootstrap t-statistic for testing $\beta_2 = \beta_{20}$.

The limits of the confidence interval can be found in many ways. However, because $\hat{P}^*_{\text{ET}}(\beta_{20})$ is not a continuous function of its argument, methods that depend on derivatives cannot be used, except perhaps to obtain initial approximations. A method that is easy to implement and reliable, although not particularly fast, is bisection. For concreteness, consider the lower limit. First, we need to find two values of β_2, say β_{2a} and β_{2b}, such that

$$f(\beta_{2a}) \equiv \hat{P}^*_{\text{ET}}(\beta_{2a}) - \alpha < 0 \quad \text{and} \quad f(\beta_{2b}) \equiv \hat{P}^*_{\text{ET}}(\beta_{2b}) - \alpha > 0. \tag{36}$$

In many cases, if we start at the lower limit of the asymptotic interval and choose β_{2a} to be, say, half a standard error lower and β_{2b} to be half a standard error larger, conditions (36) will both be satisfied. If the first is not satisfied, we must reduce β_{2a}; if the second is not satisfied, we must increase β_{2b}. The next step is to set $\beta_{2c} = (\beta_{2a} + \beta_{2b})/2$ and find $f(\beta_{2c})$. If $f(\beta_{2c}) < 0$, we replace β_{2a} by β_{2c}. If instead $f(\beta_{2c}) > 0$, we replace β_{2b} by β_{2c}. In either case, the interval between β_{2a} and β_{2b} is now half as long as it was before. Once again, we set $\beta_{2c} = (\beta_{2a} + \beta_{2b})/2$, find $f(\beta_{2c})$, and proceed as before.

Of course, we need a stopping rule. The obvious one is to stop when the distance between β_{2a} and β_{2b} is sufficiently small, although this will depend on how the parameter β_2 is scaled. Using (33) and (34) repeatedly for both limits within the bisection procedure, we eventually obtain the interval $\big[\beta_{2l},\ \beta_{2u}\big]$, where $P^*_{\text{ET}}(\beta_{2l}) \cong \alpha$ and $P^*_{\text{ET}}(\beta_{2u}) \cong \alpha$. The quality of these approximations will increase with B, which should be chosen so that $\alpha(B+1)/2$ is an integer because of the factor of 2 in (33).

A very different approach to solving for the limits of a bootstrap confidence interval that depends on restricted estimates was proposed in Hansen (1999). That approach, which has a nice graphical interpretation, could undoubtedly be adapted to the present case. For sufficiently large values of B, it will yield essentially the same results.

Inverting a WCR bootstrap test to obtain a confidence interval should rarely require much computing time, but it evidently involves a good deal of rather tricky programming. Instead of inverting a WCR bootstrap test, we could use either the pairs cluster bootstrap or the WCU bootstrap to obtain a studentized bootstrap confidence interval. Because no iteration is needed, this would be much simpler than the procedure just described. Let $c^*_{1-\alpha/2}$ and $c^*_{\alpha/2}$ denote, respectively, the $1-\alpha/2$ quantile and the $\alpha/2$ quantile of the t_2^{*b} obtained using either of these unrestricted bootstrap methods. For example, if $B = 999$ and $\alpha = 0.05$, then $c^*_{\alpha/2}$ would be number 25 and $c^*_{1-\alpha/2}$ would be number 975 in the list of the t_2^{*b} sorted from smallest to largest. The studentized bootstrap interval is then simply

$$\big[\hat{\beta}_2 - s(\hat{\beta}_2)c^*_{1-\alpha/2},\quad \hat{\beta}_2 - s(\hat{\beta}_2)c^*_{\alpha/2}\big], \tag{37}$$

where $s(\hat{\beta}_2)$ is the cluster-robust standard error of $\hat{\beta}_2$.

In many cases, the interval (37) will work quite well. However, simulation results in MacKinnon (2015) suggest that studentized bootstrap intervals based on the WCU bootstrap generally under-cover. That is, the proportion of samples for which the interval includes the true value β_{20} is less than $1-\alpha$. Sometimes this under-coverage is very modest, but in other cases it can be substantial. In the experiments reported in MacKinnon (2015), confidence intervals obtaining by inverting WCR bootstrap tests under-cover less than intervals based on (37). When the latter under-cover severely, the former often over-cover.

The finite-sample performance of bootstrap confidence intervals is closely related to the finite-sample performance of bootstrap tests. When WCU and WCR bootstrap tests perform almost equally well, we would expect studentized bootstrap intervals based on the former to perform just about as well as bootstrap intervals obtained by inverting the latter. In the experiments of Section 6.1, where the number of regressors is relatively large, both WCU and WCR bootstrap tests over-reject

noticeably. Since the former always over-reject more than the latter, it must be the case that studentized WCU bootstrap intervals would under-cover more severely than WCR bootstrap intervals in these cases.

4. Bootstrap Computations for IV Estimation

Bootstrap methods for instrumental variables (IV) estimation with clustered disturbances are more complicated than ones for OLS estimation, and their properties in finite samples are much less well understood. In the OLS case, the role of the bootstrap is simply to correct for the (sometimes) poor performance of cluster-robust covariance matrix estimators. In the IV case, the bootstrap still has to do that, but it also has to correct for the discrepancies between the actual and asymptotic distributions of the parameter(s) of interest. These discrepancies, which can include large biases, depend in complicated ways on the number of instruments, the features of the instruments, especially how weak they are, the correlation(s) between the disturbances in different equations, and potentially on many other things.

There is an enormous literature on IV inference in finite samples, much of it focused on weak instruments in models with independent, homoskedastic disturbances. Nelson and Startz (1990), Staiger and Stock (1997), Kleibergen (2002), Chao and Swanson (2005), and Andrews et al. (2006) are influential papers, and Andrews et al. (2019) is a valuable recent survey. Most of this literature does not focus on bootstrap methods, for two reasons. The first is that many bootstrap methods (notably the pairs bootstrap and the standard versions of the residual bootstrap and the wild bootstrap) seem to perform poorly. The second is that even the best bootstrap methods (see Section 4.1) can be shown to fail when the instruments are very numerous and/or extremely weak. However, these are extreme cases, in which the sample probably contains so little information about the parameters of interest that not much can be learned anyway. The theoretical focus on such cases has, in my view, tended to obscure the fact that some bootstrap methods apparently work quite well for models and datasets that do contain a reasonable amount of information, but for which conventional inference based on heteroskedasticity-robust or cluster-robust covariance matrix estimators can be highly unreliable.

For simplicity, and because it is by far the most commonly encountered case in empirical work, consider a model with just one endogenous explanatory variable. There are two equations in total:

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}_1, \text{ and} \tag{38}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2 = \boldsymbol{Z}\boldsymbol{\pi}_1 + \boldsymbol{W}_2\boldsymbol{\pi}_2 + \boldsymbol{u}_2. \tag{39}$$

Here (38) is a structural equation for $\boldsymbol{y}_1$, and (39) is a reduced-form equation for $\boldsymbol{y}_2$. The coefficient of interest is β. The instrument matrix $\boldsymbol{W}$ is $N \times l$. The $N \times k$ matrix of exogenous regressors $\boldsymbol{Z}$ is part of it, so that there are $l - k - 1$ over-identifying restrictions. When the two endogenous variables are in fact determined simultaneously, the i$^{\text{th}}$ elements of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are almost certainly correlated. In consequence, OLS estimation of (38) yields an inconsistent estimate of β, and it is natural to use IV estimation instead.

The key features of bootstrap methods that yield reliable inferences for β, at least in cases that are not too extreme, is that they are based on restricted estimates and that they use efficient estimates of the reduced-form Eq. (39) to generate the bootstrap samples. The first bootstrap method to employ such a method was the restricted efficient (RE) bootstrap proposed in Davidson and MacKinnon (2008). This method is a variant of the residual bootstrap, because it is based on the assumption that the disturbances are independent and homoskedastic. Davidson and MacKinnon (2010) later relaxed the homoskedasticity assumption and proposed the wild restricted efficient (WRE) bootstrap. A straightforward extension of this method allows for clustered disturbances. The wild restricted efficient cluster (WREC) bootstrap uses one value of v^* per cluster instead of one per observation. This method is one of several that were studied in Finlay and Magnusson (2019). It is the only one that will be discussed here.

The key computational idea first discussed in Section 2 and applied to the WCR bootstrap in Section 3.2 can be applied to the WREC bootstrap, as well as to other bootstrap methods for IV regression. The computations are more complicated than the ones for the WCR bootstrap and will inevitably be more expensive, but they share the property that all operations which are $O(N)$ only need to be performed once, before bootstrapping actually begins. For large sample sizes, this can greatly reduce computational costs. The version of `boottest` discussed in Roodman et al. (2019) implemented the WREC bootstrap in a way that was quite slow. Since Version 3.1.0, however, `boottest` has used a much faster algorithm, which was partly inspired by an early draft of this paper. This algorithm still seems to be somewhat slower than the method proposed in Section 4.1, but that may simply be because it is programmed in Mata rather than Fortran.

The IV estimates of β and $\boldsymbol{\gamma}$ can be obtained by regressing $\boldsymbol{y}_1$ on $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2$ and $\boldsymbol{Z}$, where $\boldsymbol{P}_{\boldsymbol{W}}$ is the projection matrix $\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ that yields fitted values from a regression on all the instruments. This gives the usual expression for the IV estimates:

$$\begin{bmatrix} \hat{\beta} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2 & \boldsymbol{Z}^\top \boldsymbol{y}_2 \\ \boldsymbol{y}_2^\top \boldsymbol{Z} & \boldsymbol{Z}^\top \boldsymbol{Z} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2 \\ \boldsymbol{y}_1^\top \boldsymbol{Z} \end{bmatrix}. \tag{40}$$

The Frisch-Waugh-Lovell (or FWL) theorem implies that $\hat{\beta}$ in (40) is identical to the OLS estimate of β from the univariate regression

$$(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}_1 = \beta(\mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2 + \text{residuals} = \beta(\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{y}_2 + \text{residuals}, \tag{41}$$

where $\boldsymbol{P_Z}$ is the projection matrix $\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$, and the second equality uses the fact that $\boldsymbol{P_W Z} = \boldsymbol{Z}$. Using (41), it is then easy to show that

$$\hat{\beta} = \frac{\boldsymbol{y}_2^\top(\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_1}{\boldsymbol{y}_2^\top(\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_2} = \frac{\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_1 - \boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_1}{\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_2 - \boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_2}. \tag{42}$$

The rightmost expression for $\hat{\beta}$ here is the one on which all our computations are based.

When there are G clusters, the four scalars that appear in the rightmost expression in (42) can be written as

$$\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_j = \Big(\sum_{g=1}^{G}\boldsymbol{y}_{2g}^\top\boldsymbol{W}_g\Big)(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\Big(\sum_{g=1}^{G}\boldsymbol{W}_g{}^\top\boldsymbol{y}_{jg}\Big), \ \ j = 1, 2, \ \text{and} \tag{43}$$

$$\boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_j = \Big(\sum_{g=1}^{G}\boldsymbol{y}_{2g}^\top\boldsymbol{Z}_g\Big)(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\Big(\sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{y}_{jg}\Big), \ \ j = 1, 2, \tag{44}$$

where $\boldsymbol{y}_{1g}$, $\boldsymbol{y}_{2g}$, $\boldsymbol{W}_g$, and $\boldsymbol{Z}_g$ are matrices containing the rows of $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\boldsymbol{W}$, and $\boldsymbol{Z}$ that belong to cluster g. Moreover,

$$\boldsymbol{W}^\top\boldsymbol{W} = \sum_{g=1}^{G}\boldsymbol{W}_g{}^\top\boldsymbol{W}_g, \ \ \text{and} \ \ \boldsymbol{Z}^\top\boldsymbol{Z} = \sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{Z}_g. \tag{45}$$

Thus $\hat{\beta}$ depends on the sample only through the various within-cluster cross-products that appear in (43), (44), and (45). These cross-products can be computed quite economically by using the facts that

$$\boldsymbol{W}_g{}^\top\boldsymbol{W}_g = \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{Z}_g & \boldsymbol{Z}_g^\top\boldsymbol{W}_{2g} \\ \boldsymbol{W}_{2g}^\top\boldsymbol{Z}_g & \boldsymbol{W}_{2g}^\top\boldsymbol{W}_{2g} \end{bmatrix} \ \text{and} \ \boldsymbol{W}_g{}^\top\boldsymbol{y}_{jg} = \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{y}_{jg} \\ \boldsymbol{W}_{2g}^\top\boldsymbol{y}_{jg} \end{bmatrix}, \ j = 1, 2. \tag{46}$$

Since it is not clear what the degrees-of-freedom factor should be for a CRVE based on IV estimates, because the IV residuals are not necessarily too small, I follow Stata and omit this factor. The bread in the CRVE sandwich is $\big(\boldsymbol{y}_2^\top(\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_2\big)^{-1}$, and the filling is

$$\hat{\omega}^2 = \sum_{g=1}^{G}\big(\boldsymbol{y}_2^\top(\boldsymbol{P_W} - \boldsymbol{P_Z})\big)_g\hat{\boldsymbol{u}}_{1g}\hat{\boldsymbol{u}}_{1g}^\top\big((\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_2\big)_g, \tag{47}$$

where $\big((\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_2\big)_g$ is a matrix containing the rows of $(\boldsymbol{P_W} - \boldsymbol{P_Z})\boldsymbol{y}_2$ that correspond to group g, and $\big(\boldsymbol{y}_2^\top(\boldsymbol{P_W} - \boldsymbol{P_Z})\big)_g$ is the transpose of that matrix. Expression (47) can be rewritten as

$$\hat{\omega}^2 = \sum_{g=1}^{G}\big(\boldsymbol{y}_2^\top\boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}_g{}^\top\hat{\boldsymbol{u}}_{1g} - \boldsymbol{y}_2^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}_g^\top\hat{\boldsymbol{u}}_{1g}\big)^2. \tag{48}$$

We can then use the facts that

$$\boldsymbol{W}_g{}^\top\hat{\boldsymbol{u}}_{1g} = \boldsymbol{W}_g{}^\top\boldsymbol{y}_{1g} - \hat{\beta}\boldsymbol{W}_g{}^\top\boldsymbol{y}_{2g} - \boldsymbol{W}_g{}^\top\boldsymbol{Z}_g\hat{\boldsymbol{\gamma}} \ \ \text{and} \ \ \boldsymbol{Z}_g^\top\hat{\boldsymbol{u}}_{1g} = \boldsymbol{Z}_g^\top\boldsymbol{y}_{1g} - \hat{\beta}\boldsymbol{Z}_g^\top\boldsymbol{y}_{2g} - \boldsymbol{Z}_g^\top\boldsymbol{Z}_g\hat{\boldsymbol{\gamma}} \tag{49}$$

and substitute these into (48). Of course, (49) requires $\hat{\boldsymbol{\gamma}}$ as well as $\hat{\beta}$. Because we know $\hat{\beta}$, we can just regress $\boldsymbol{y}_1 - \hat{\beta}\boldsymbol{P_W}\boldsymbol{y}_2$ on $\boldsymbol{Z}$. This yields

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_1 - \hat{\beta}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{y}_2 \\ &= (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\Big(\textstyle\sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{y}_{1g} - \hat{\beta}\sum_{g=1}^{G}\boldsymbol{Z}_g^\top\boldsymbol{y}_{2g}\Big). \end{aligned} \tag{50}$$

Everything in (50) depends on within-cluster cross-products that have already been calculated; $\boldsymbol{Z}^\top\boldsymbol{Z}$ is given in (45), and both $\boldsymbol{Z}^\top\boldsymbol{y}_1$ and $\boldsymbol{Z}^\top\boldsymbol{y}_2$ were used in (44). The projection matrix $\boldsymbol{P_W}$ does not appear because $\boldsymbol{P_W Z} = \boldsymbol{Z}$.

From (42) and (48), we see that the t-statistic for $\beta = \beta_0$ is

$$t_{\beta_0} = \frac{1}{\hat{\omega}}\big(\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_1 - \boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_1 - \beta_0(\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_2 - \boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_2)\big), \tag{51}$$

The factor of $\boldsymbol{y}_2^\top\boldsymbol{P_W}\boldsymbol{y}_2 - \boldsymbol{y}_2^\top\boldsymbol{P_Z}\boldsymbol{y}_2$ in the second term here is both the inverse of the bread in the CRVE sandwich and the denominator of $\hat{\beta}$.

Of course, if we simply wanted to compute t_{β_0}, it would not make sense to employ the long and rather complicated sequence of computations described in (42) through (51). It would be easier and not much more expensive just to use a standard program for IV estimation with cluster-robust standard errors. But using the former makes sense when we want to bootstrap t_{β_0}, because the bootstrap t-statistics will depend on the sample only through G sets of vectors and matrices that are computed before the bootstrap loop begins.

4.1. The WREC Bootstrap

As noted earlier, the WREC bootstrap is a slightly modified version of the WRE bootstrap proposed in Davidson and MacKinnon (2010). The bootstrap DGP for the latter uses a single auxiliary random variable, say v_i^*, for the i$^{\text{th}}$ observation for both equations. This preserves the correlations between the structural and reduced-form residuals. The WREC bootstrap needs to preserve not only those correlations but also all the within-cluster correlations of both sets of residuals. It therefore uses the same auxiliary random variable, say v_g^*, for every observation in the g$^{\text{th}}$ cluster for both equations. The fact that the WREC bootstrap uses only G realizations of the auxiliary random variable is what makes it possible to compute the bootstrap t-statistics without performing any operations that are $O(N)$.

If we actually generated bootstrap data for every observation, the WREC bootstrap DGP for cluster g could be written as

$$\boldsymbol{y}_{1g}^* = \beta_0 \boldsymbol{y}_{2g}^* + \boldsymbol{Z}_g \tilde{\boldsymbol{\gamma}} + v_g^* m_1 \tilde{\boldsymbol{u}}_{1g}, \tag{52}$$

$$\boldsymbol{y}_{2g}^* = \boldsymbol{W}_g \tilde{\boldsymbol{\pi}} + v_g^* m_2 \tilde{\boldsymbol{u}}_{2g}, \tag{53}$$

where Eq. (53) is implicitly evaluated before Eq. (52), because $\boldsymbol{y}_{1g}^*$ depends on $\boldsymbol{y}_{2g}^*$. Here $\tilde{\boldsymbol{\gamma}}$ is obtained by regressing $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$ on $\boldsymbol{Z}$, which should be inexpensive because we have already computed $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$, and $\boldsymbol{Z}^\top \boldsymbol{y}_1 = \sum_{g=1}^G \boldsymbol{Z}_g^\top \boldsymbol{y}_{1g}$. The other things that appear in (52), in addition to $\boldsymbol{y}_{2g}^*$, are $m_1 = (N/(N-k))^{1/2}$, a degrees-of-freedom correction, the random variate v_g^*, which (usually) follows the Rademacher distribution, and $\tilde{\boldsymbol{u}}_{1g} = \boldsymbol{y}_{1g} - \beta_0 \boldsymbol{y}_{2g} - \boldsymbol{Z}_g \tilde{\boldsymbol{\gamma}}$. The vector $\tilde{\boldsymbol{u}}_{2g}$ is $\boldsymbol{y}_{2g} - \boldsymbol{W}_g \tilde{\boldsymbol{\pi}}$, where $\tilde{\boldsymbol{\pi}}$ will be defined shortly, and $m_2 = (N/(N-l))^{1/2}$ is another degrees-of-freedom correction. Of course, we never actually calculate the $\tilde{\boldsymbol{u}}_{1g}$, the $\tilde{\boldsymbol{u}}_{2g}$, the $\boldsymbol{y}_{1g}^*$, or the $\boldsymbol{y}_{2g}^*$, because doing so would involve computations that are $O(N)$.

The estimate $\tilde{\boldsymbol{\pi}}$ that appears in (53) is obtained by running the regression

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \rho \tilde{\boldsymbol{u}}_1 + \boldsymbol{\epsilon}, \tag{54}$$

where $\boldsymbol{u}_2$ is divided into a vector that is correlated with $\boldsymbol{u}_1$ and a vector $\boldsymbol{\epsilon}$ that is uncorrelated with it. Unless ρ, the correlation between the structural and reduced-form disturbances, is zero (or at least close to zero), $\tilde{\boldsymbol{\pi}}$ is more efficient than the usual estimate $\hat{\boldsymbol{\pi}}$ obtained by regressing $\boldsymbol{y}_2$ on $\boldsymbol{W}$ alone; see Kleibergen (2002). It is asymptotically equivalent to what we would obtain by estimating Eqs (38) and (39) jointly using full-information maximum likelihood. Because (39) is an unrestricted reduced form, those estimates are numerically identical to restricted LIML estimates of (38); indeed, that is how `boottest` currently estimates $\boldsymbol{\pi}$. Obtaining the restricted LIML estimates requires a fair amount of algebra (Roodman et al., 2019, Appendix B), which could be rewritten in terms of the cluster cross-product matrices used in this section.

It is not difficult to show that

$$\tilde{\boldsymbol{\pi}} = (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \big(\boldsymbol{W}^\top \boldsymbol{y}_2 - \tilde{\rho} \, (\boldsymbol{W}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_1 - \beta_0 \boldsymbol{W}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2) \big), \tag{55}$$

where $\boldsymbol{M}_{\boldsymbol{Z}} = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{Z}}$. Here

$$\tilde{\rho} = \frac{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 - \beta_0 \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_1 - 2\beta_0 \boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2 + \beta_0^2 \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{W}} \boldsymbol{y}_2} \tag{56}$$

is the estimated coefficient on $\tilde{\boldsymbol{u}}_1$ in regression (54), and $\boldsymbol{M}_{\boldsymbol{W}} = \mathbf{I} - \boldsymbol{P}_{\boldsymbol{W}}$. The second term in the second factor of (55) would vanish if $\tilde{\rho} = 0$, and $\tilde{\boldsymbol{\pi}}$ would then equal $\hat{\boldsymbol{\pi}}$, the usual vector of reduced-form estimates. Everything that appears on the right-hand sides of (55) and (56) can be expressed in terms of quantities that were either computed above or can readily be computed just once, notably $\boldsymbol{y}_1^\top \boldsymbol{y}_1$, $\boldsymbol{y}_1^\top \boldsymbol{y}_2$, and $\boldsymbol{y}_2^\top \boldsymbol{y}_2$.

The numerator of the bootstrap t-statistic depends solely on β_0 and the bootstrap analogs of (43) and (44). The latter depend in turn on either $(\boldsymbol{W}^\top \boldsymbol{W})^{-1}$ or $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$ and on one or two of the vectors

$$\boldsymbol{y}_2^{*\top} \boldsymbol{W} = \sum_{g=1}^G \boldsymbol{y}_{2g}^{*\top} \boldsymbol{W}_g = \sum_{g=1}^G \big((1 - v_g^* m_2) \tilde{\boldsymbol{\pi}}^\top \boldsymbol{W}_g{}^\top \boldsymbol{W}_g + v_g^* m_2 \boldsymbol{y}_2^\top \boldsymbol{W}_g \big), \tag{57}$$

$$\boldsymbol{y}_1^{*\top} \boldsymbol{W} = \sum_{g=1}^G \boldsymbol{y}_{1g}^{*\top} \boldsymbol{W}_g = \beta_0 \boldsymbol{y}_2^{*\top} \boldsymbol{W} + \sum_{g=1}^G \big(v_g^* m_1 (\boldsymbol{y}_{1g}^\top \boldsymbol{W}_g - \beta_0 \boldsymbol{y}_{2g}^\top \boldsymbol{W}_g - \tilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_g^\top \boldsymbol{W}_g) \big), \tag{58}$$

and/or the vectors $\boldsymbol{y}_2^{*\top} \boldsymbol{Z}$ and $\boldsymbol{y}_1^{*\top} \boldsymbol{Z}$, which are respectively equal to the first k elements of $\boldsymbol{y}_2^{*\top} \boldsymbol{W}$ and the first k elements of $\boldsymbol{y}_1^{*\top} \boldsymbol{W}$.

Eq. (58) is missing the term $\sum_{g=1}^G \tilde{\boldsymbol{\gamma}}^\top \boldsymbol{Z}_g^\top \boldsymbol{W}_g$, which would seem to arise from the second term on the right-hand side of (52). There is no reason to bother with that term; because $\boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{Z} - \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{Z}$, it vanishes when we compute $\hat{\beta}^*$. Notice that most of the computations in (57) and (58) can be done before the bootstrap loop begins. The only things that differ across bootstrap samples are the auxiliary random variables v_g^*, for $g = 1, \ldots, G$.

From (51), we see that the bootstrap t-statistic is

$$t_{\beta_0}^* = \frac{1}{\hat{\omega}^*} \big(\boldsymbol{y}_2^{*\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_1^* - \boldsymbol{y}_2^{*\top} \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_1^* - \beta_0 (\boldsymbol{y}_2^{*\top} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2^* - \boldsymbol{y}_2^{*\top} \boldsymbol{P}_{\boldsymbol{Z}} \boldsymbol{y}_2^*) \big). \tag{59}$$

The numerator can readily be calculated using $(\boldsymbol{W}^\top\boldsymbol{W})^{-1}$, $(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$ and (57)–(58). The denominator is just the square root of (48) with the $\boldsymbol{y}_g$ and the $\hat{\boldsymbol{u}}_{1g}$ replaced by their analogs from the bootstrap samples:

$$\hat{\omega}^* = \left(\sum_{g=1}^{G} \left(\boldsymbol{y}_2^{*\top} \boldsymbol{W} (\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}_g^\top \hat{\boldsymbol{u}}_{1g}^* - \boldsymbol{y}_2^{*\top} \boldsymbol{Z} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}_g^\top \hat{\boldsymbol{u}}_{1g}^* \right)^2 \right)^{1/2}. \tag{60}$$

We already have expressions for $\boldsymbol{y}_2^{*\top}\boldsymbol{W}$ and $\boldsymbol{y}_2^{*\top}\boldsymbol{Z}$ in (57). We just need ones for $\boldsymbol{W}_g^\top\hat{\boldsymbol{u}}_{1g}^*$ and $\boldsymbol{Z}_g^\top\hat{\boldsymbol{u}}_{1g}^*$, where $\hat{\boldsymbol{u}}_{1g}^*$ is the vector of bootstrap IV residuals for the g$^{\text{th}}$ cluster. Clearly

$$\boldsymbol{W}_g^\top \hat{\boldsymbol{u}}_{1g}^* = \boldsymbol{W}_g^\top \boldsymbol{y}_1^* - \hat{\beta}^* \boldsymbol{W}_g^\top \boldsymbol{y}_2^* - \boldsymbol{W}_g^\top \boldsymbol{Z}_g \hat{\boldsymbol{\gamma}}^*, \tag{61}$$

and $\boldsymbol{Z}_g^\top\hat{\boldsymbol{u}}_{1g}^*$ contains the first k elements of $\boldsymbol{W}_g^\top\hat{\boldsymbol{u}}_{1g}^*$. We have already computed $\boldsymbol{W}_g^\top\boldsymbol{y}_1^*$, $\boldsymbol{W}_g^\top\boldsymbol{y}_2^*$, and $\boldsymbol{W}_g^\top\boldsymbol{Z}_g$, which appear in (61). We have also come very close to calculating $\hat{\beta}^*$, which, for completeness, is

$$\hat{\beta}^* = \frac{\boldsymbol{y}_2^{*\top}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_1^* - \boldsymbol{y}_2^{*\top}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}_1^*}{\boldsymbol{y}_2^{*\top}\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2^* - \boldsymbol{y}_2^{*\top}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}_2^*};$$

compare (42). The only thing in (61) that we have not yet computed is $\hat{\boldsymbol{\gamma}}^*$. As with (50), we just need to regress $\boldsymbol{y}_1^* - \hat{\beta}^*\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2^*$ on $\boldsymbol{Z}$. This yields

$$\hat{\boldsymbol{\gamma}}^* = (\boldsymbol{Z}^\top\boldsymbol{Z})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{Z}_g^\top \boldsymbol{y}_{1g}^* - \hat{\beta}^* \sum_{g=1}^{G} \boldsymbol{Z}_g^\top \boldsymbol{y}_{2g}^* \right), \tag{62}$$

where the matrix and all the vectors in (62) have been computed previously.

We finally have everything needed to compute $t_{\beta_0}^*$ defined in (59), and thus everything needed for the WREC bootstrap. All operations that are $O(N)$ can be done before the bootstrap loop begins, and so can many other calculations. Thus, as is documented in Section 5.2, the cost of obtaining B bootstrap t-statistics is generally far less than the cost of obtaining an actual t-statistic B times.

Because $\hat{\beta}$ is usually biased, often severely so, it makes sense to use an equal-tailed bootstrap P value like (33). In this case, it is given by

$$\hat{P}_{\text{ET}}^*(\beta_0) = \frac{2}{B} \min \left(\sum_{b=1}^{B} \mathbb{I}\big(t_{\beta_0}^{*b} > t_{\beta_0}\big), \sum_{b=1}^{B} \mathbb{I}\big(t_{\beta_0}^{*b} \leq t_{\beta_0}\big) \right), \tag{63}$$

where t_{β_0} is the t-statistic for $\beta = \beta_0$ given in (51) and $t_{\beta_0}^{*b}$ is the bootstrap t-statistic for the b$^{\text{th}}$ bootstrap sample given in (59).

5. Computing Costs

The algorithms proposed in Sections 3 and 4, along with the one for the WCR bootstrap used in `boottest` (Roodman et al., 2019), can evidently be much faster than more straightforward bootstrap methods that generate full bootstrap samples. In realistic cases, costs can be reduced by several orders of magnitude, because all calculations that are $O(N)$ are done just once rather than for every bootstrap sample. This means that bootstrap P values and bootstrap confidence intervals based on the WCR bootstrap, the WREC bootstrap, and even the pairs cluster bootstrap can often be computed routinely, even for extremely large samples. In this section, some timing evidence is presented to support these claims.

The first set of simulation experiments, in Section 5.1, investigates the computational methods proposed in Section 3 for the pairs cluster bootstrap and the WCR bootstrap applied to linear regression models estimated by OLS. The second set, in Section 5.2, studies the cost of computing the WREC bootstrap for linear regression models estimated by IV using the methods proposed in Section 4.1.

5.1. The Pairs Cluster and WCR Bootstraps

From the discussion in Section 3, it is clear that computing time depends on N, G, B, k, and r. These are therefore the quantities that are varied in the simulations. Specifically, N equals 10,000 or 1,000,000, G equals 20 or 40, and B equals 999 or 9,999. The number of regressors k is either 10 or 20, and the number of restrictions r is either 1 or 5.

Table 1 shows computing times in seconds for the pairs cluster bootstrap and the WCR bootstrap programmed in two different ways. For comparison, a benchmark number is also reported. It is simply $B+1$ times the computing time for a single test statistic, because an inefficient bootstrap method typically involves computing the same test statistic that many times. This may be an over-estimate, especially in the case of the wild cluster bootstrap, because it ignores the possibility of using various computational tricks, such as re-using the $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ matrix. On the other hand, it also fails to account for the cost of generating each bootstrap sample. In all cases, the cost of generating the data (which is often larger than the cost of bootstrapping) is excluded from the times reported.

Table 1
Computing Times in Seconds for OLS Bootstrap Methods

N	10,000			1,000,000		
G, k	20,10	40,10	40,20	20,10	40,10	40,20
$r = 1$						
Bench, 999	0.1378	0.1459	0.4656	21.89	20.05	63.65
Pairs, 999	0.0038	0.0068	0.0213	0.0253	0.0264	0.0880
WCR(s), 999	0.0016	0.0029	0.0080	0.0239	0.0222	0.0723
WCR(b), 999	0.0006	0.0011	0.0015	0.0229	0.0210	0.0684
Bench, 9,999	1.378	1.459	4.656	218.9	200.5	636.5
Pairs, 9,999	0.0370	0.0666	0.2102	0.0595	0.0858	0.2848
WCR(s), 9,999	0.0146	0.0280	0.0749	0.0366	0.0474	0.1380
WCR(b), 9,999	0.0045	0.0098	0.0102	0.0260	0.0291	0.0745
$r = 5$						
Bench, 999	0.1397	0.1461	0.4703	20.71	19.02	64.81
Pairs, 999	0.0049	0.0079	0.0246	0.0274	0.0277	0.0887
WCR(s), 999	0.0027	0.0040	0.0111	0.0251	0.0232	0.0755
Bench, 9,999	1.397	1.461	4.703	207.1	190.2	648.1
Pairs, 9,999	0.0480	0.0783	0.2423	0.0695	0.0969	0.3096
WCR(s), 9,999	0.0256	0.0391	0.1066	0.0470	0.0583	0.1699

Notes: All computations were performed in Fortran using one core of an Intel i9-10850K processor running at 3.6 GHz. For accuracy, they were repeated at least 1000 times. For $r = 1$, WCR(s) computes the bootstrap test statistics as Wald statistics using (28), while WCR(b) computes them using the boottest approach described in Eq. (20) and Eqs (23) through (25).

Times for the WCU bootstrap, not shown, are similar to those for the WCR bootstrap, but somewhat smaller (especially for $N = 10^6$) because the restricted estimates do not need to be computed. The pairs cluster bootstrap is more expensive than the wild cluster bootstrap because the former needs to construct the $\boldsymbol{X}^{*b\top}\boldsymbol{X}^{*b}$ matrix for each bootstrap sample, while the latter uses the same $\boldsymbol{X}^\top\boldsymbol{X}$ matrix for all of them. However, the new method of computing the pairs cluster bootstrap proposed in Section 3.1 is still enormously faster than the benchmark.

For $r = 1$, there are two variants of the WCR bootstrap. The simpler one, denoted "WCR(s)", is based on (28). The other one, denoted "WCR(b)", is the boottest method described in Section 3.2. Although both methods are much faster than the benchmark, the boottest method can be quite a bit faster than the simpler one. This is particularly true when N is small, k is large, and B is large. In theory, the marginal cost of generating an additional bootstrap test statistic using WCR(b) is independent of N and k. This seems to be more or less true in Table 1, allowing for some inaccuracy in the reported times. Unless G is very large, in which case there is usually no need to bootstrap, this marginal cost tends to be remarkably small.

For $r = 5$, only one set of WCR results is shown, because the boottest method for computing Wald statistics has not yet been programmed in Fortran. But even the relatively slow method based on (28), which is very easy to program, is enormously faster than the benchmark. However, it is sometimes not much faster than the new version of the pairs bootstrap proposed in Section 3.1.

Except for the benchmark numbers, all of the times in Table 1 are less than one second, and most of them are less than 1/10 of a second. So all the methods proposed here seem to be more than fast enough for practical use with large samples. The benchmark times always greatly exceed those for the WCR(s) and WCR(b) bootstraps. When $N = 10^6$ and $B = 9{,}999$, they do so by factors of several thousand. When $N = 10^6$, the initial computations evidently account for a substantial proportion of the total time. Increasing B from 999 to 9,999 generally has a fairly modest effect. For WCR(b), as noted above, the effect is always very modest indeed and depends solely on G.

Increasing G from 20 to 40 and increasing r from 1 to 5 both have fairly small effects on computing times for the bootstrap, especially when $N = 10^6$. The only parameter that has a large effect, but not for WCR(b), is k. This happens because forming the $\boldsymbol{X}_g^\top\boldsymbol{X}_g$ matrices requires $O(k^2N)$ operations, and creating the "filling" in the CRVE (5) requires $O(k^2G)$. The former only has to be done once for any bootstrap test, but the latter has to be done $B+1$ times for the WCR(s) and pairs cluster bootstraps. For the pairs cluster bootstrap, it is also necessary to compute and invert the matrix $\sum_{g=1}^G \boldsymbol{X}_g^{*b\top}\boldsymbol{X}_g^{*b}$ for each bootstrap sample. Thus the only circumstance in which the new procedures risk becoming computationally challenging is when there are many regressors. In such cases, reducing the number of coefficients that have to be estimated by partialing out fixed effects and perhaps other dummy variables might reduce computing time substantially; see Section 3.3. Of course, this is never a problem for the WCR(b) procedure used by `boottest`, which becomes relatively less costly as the number of regressors increases.

5.2. The WREC Bootstrap

Table 2 shows computing times in seconds for the WREC bootstrap and for a benchmark, which is simply $B+1$ times the cost of running a single IV regression and computing a single cluster-robust t-statistic. This benchmark is only a crude ap-

Table 2
Computing Times in Seconds for the WREC Bootstrap

N	10,000			1,000,000		
k, l	2,20	12,20	12,40	2,20	12,20	12,40
$G = 20$						
Bench, 999	0.6861	1.0366	2.5369	152.08	232.46	580.06
WREC, 999	0.0062	0.0083	0.0268	0.0776	0.0788	0.2782
Bench, 9,999	6.861	10.366	25.369	1520.8	2324.6	5800.6
WREC, 9,999	0.0574	0.0822	0.2453	0.1260	0.1458	0.4902
$G = 40$						
Bench, 999	0.6803	1.0378	2.5175	158.18	242.04	586.23
WREC, 999	0.0132	0.0189	0.0718	0.0843	0.0871	0.2995
Bench, 9,999	6.803	10.378	25.175	1581.8	2420.4	5860.2
WREC, 9,999	0.1304	0.1778	0.5646	0.1902	0.2303	0.7971

Notes: All computations were performed in Fortran using one core of an Intel i9-10850K processor running at 3.6 GHz. For accuracy, they were repeated at least 1000 times.

proximation. It does not allow for some fairly obvious optimizations, but it also ignores the cost of generating the bootstrap samples, which can be substantial if they are generated in a naive way.

As implemented using the method proposed in Section 4.1, the WREC bootstrap is always extremely fast. It never takes as long as one second, even with 10^6 observations, 40 instruments, and 9,999 bootstrap samples. Its relative performance is best when $N = 10^6$, $G = 20$, $l = 20$, and $k = 12$. In this case, it is roughly 16,000 times as fast as the benchmark. Inevitably, computing time goes up with G, k, and (especially) l, the number of instruments. However, it appears that these numbers would all need to be very large indeed for computation to be even slightly challenging.

For $N = 10^4$, the cost of increasing B from 999 to 9,999 is always somewhat less than a factor of 10. In contrast, for $N = 10^6$, it is always much less than a factor of 10. In the latter case, computation time is evidently dominated by the cost of forming the within-cluster sums of cross-products that appear in (43) through (45), which only needs to be done once.

6. Monte Carlo Experiments

The fast bootstrap algorithms developed in Sections 3 and 4 make it feasible to undertake experiments that would otherwise have been computationally challenging. This section takes advantage of that fact to investigate the finite-sample properties of these bootstrap methods. All experiments use 400,000 replications, ensuring that experimental randomness is negligible. Section 6.1 deals with Wald tests in regression models estimated by OLS, and Section 6.2 deals with t-tests of the coefficient on the single endogenous regressor in linear regression models estimated by IV.

6.1. Wald Tests in Finite Samples

There has been a good deal of work on the finite-sample properties of various methods, notably the WCR bootstrap, for cluster-robust inference about a single coefficient in the linear regression model (1); see, among others, Cameron et al. (2008), Imbens and Kolesár (2016), and MacKinnon and Webb (2017a, 2017b, 2018). However, with the exception of Pustejovsky and Tipton (2018), which does not study bootstrap tests, there has apparently been no work on the finite-sample properties of cluster-robust Wald tests for more than one restriction. The Monte Carlo experiments of this subsection therefore focus on that case.

The experiments deal with tests of $r \geq 1$ restrictions based on the Wald statistic $W(\hat{\boldsymbol{\beta}})$ given in (9). Four tests are considered. The only one that does not involve bootstrapping rejects the null hypothesis at level α whenever $W(\hat{\boldsymbol{\beta}})/r$ exceeds the $1-\alpha$ quantile of the $F(r, G-1)$ distribution. Since the results of Pustejovsky and Tipton (2018) suggest that the approximation on which this test is based is generally not a good one, especially for larger values of r, this procedure is not expected to perform well.

The other three tests are based on the pairs cluster bootstrap, discussed in Section 3.1, and two variants of the wild cluster bootstrap, discussed in Section 3.2. For both the pairs cluster and WCU bootstraps, the statistic for the b$^{\text{th}}$ bootstrap sample is

$$W_b^* = (\hat{\boldsymbol{\beta}}^{*b} - \hat{\boldsymbol{\beta}})^\top \boldsymbol{R}^\top \big(\boldsymbol{R}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}^{*b})\boldsymbol{R}^\top\big)^{-1}\boldsymbol{R}(\hat{\boldsymbol{\beta}}^{*b} - \hat{\boldsymbol{\beta}}),$$

because the bootstrap samples do not impose the restriction that $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$. In contrast, for the WCR bootstrap, the bootstrap test statistic is given by (26).

The data are generated from the model (1) with $k = 20$. The number of regressors is quite large because up to 10 restrictions are to be tested. With the exception of the constant term, both the regressors and the disturbances are normally distributed and follow independent random-effects models parametrized so that the intra-cluster correlation of the disturbances is 0.1 and that of the regressors is 0.5. For reasons of space, the values of these parameters were not varied, although

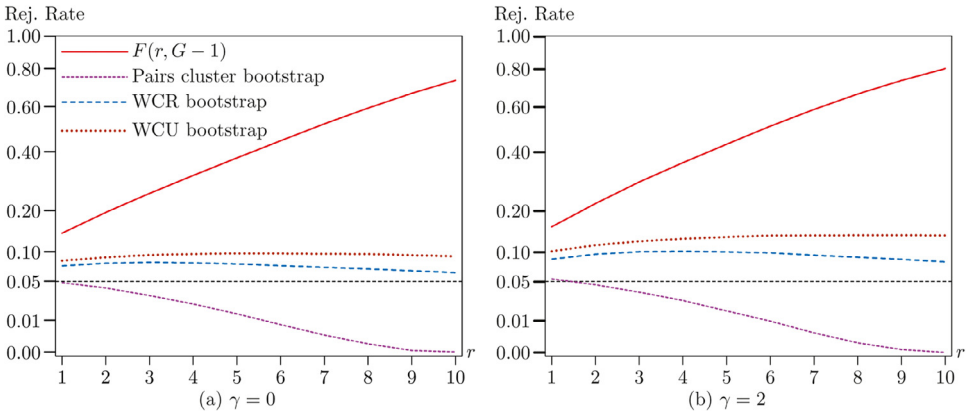

Fig. 1. Rejection Rates for Wald Tests with 20 Clusters. **Notes:** The vertical axis shows estimated rejection rates, after a square root transformation, for tests at the.05 level. These are based on 400,000 replications with $G = 20$, $N = 2000$, $k = 20$, and $B = 399$. The number of restrictions, r, varies from 1 to 10. Clusters are equal-sized in panel (a) and vary from 32 to 229 in panel (b).

they necessarily affect the results. Figures C.1 and C.5 of Djogbenou et al. (2019) suggest that rejection rates for bootstrap tests are quite insensitive to the correlation of the disturbances once it exceeds about 0.05, and only moderately sensitive to the correlation of the regressors once it exceeds about 0.40.

In the experiments, there are $N = 100G$ observations, which are divided among the G clusters using the formula

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{64}$$

where $[x]$ means the integer part of x. The value of N_G is then set to $N - \sum_{g=1}^{G_1} N_g$. The key parameter here is γ, which determines how uneven the cluster sizes are. When $\gamma = 0$ and N/G is an integer, Eq. (64) implies that $N_g = N/G$ for all g. As γ increases, however, cluster sizes vary more and more.

In the first two sets of experiments, $G = 20$. Fig. 1 shows rejection rates (or frequencies) for all four tests as a function of r, the number of restrictions. In panel (a), $\gamma = 0$, so that all clusters have 100 observations. In panel (b), $\gamma = 2$, so that cluster sizes vary from 32 to 229. The vertical axis has been subjected to a square root transformation in order to show results for all four procedures legibly on the same axes. In all experiments, the number of bootstrap replications B is equal to 399. In practice, it would be better to use a larger number, such as 999 or 9,999, and it would not be terribly expensive to do so. However, because randomness in the bootstrap samples averages out across replications, the reported rejection rates would have been essentially unchanged if a larger value of B had been used.

In both panels, the F test over-rejects severely. The extent to which it does so increases with r and is always greater in panel (b) than in panel (a). Although neither of these results is surprising, the figure makes it clear that asymptotic Wald tests can be greatly over-sized, especially when r is not small. If one of these tests fails to reject a null hypothesis, then we can be very confident that evidence against that hypothesis is weak. However, if one of them apparently rejects a null hypothesis, we should seek confirmation from other procedures before drawing any conclusions.

The pairs cluster bootstrap works very well for $r = 1$. In fact, it is by far the most reliable procedure in this case. However, in both panels, it under-rejects more and more severely as r increases. For $r \geq 6$, it rejects less than 1% of the time at the.05 level. For $r = 10$, it either never rejects or rejects just once in 400,000 replications. This property of the pairs cluster bootstrap does not seem to have been noticed before. The problem arises because, as r increases, the bootstrap test statistics become larger and more dispersed to a much greater extent than do the actual test statistics. In consequence, the.95 quantile of the empirical distribution of the W_b^* for each sample tends to become larger and larger relative to the value of $W(\hat{\boldsymbol{\beta}})$. Thus using that quantile as a critical value leads to increasingly severe under-rejection as r increases. It would be interesting to investigate why the bootstrap distribution behaves in this way for the pairs cluster bootstrap, but this would require theoretical work well beyond the scope of this paper.

The two wild cluster bootstrap tests do not work particularly well, but their performance does not deteriorate as r increases. In fact, it eventually improves modestly once r becomes large enough. As expected, the WCU bootstrap always rejects more often than the WCR bootstrap. Even when $r = 1$, both procedures over-reject much more often than they have done in previous Monte Carlo experiments (MacKinnon and Webb, 2017a; 2018). The reason for this seems to be that there are 20 regressors here (including a constant term), instead of just a few, as in previous work. Djogbenou et al. (2019, Appendix C.2) studies the consequences of adding additional regressors and finds that doing so increases rejection rates for both the WCR and WCU bootstraps, but rejection rates are considerably smaller than is seen in Fig. 1. This is almost certainly because the present experiments involve both fewer clusters and more regressors than the ones in that paper.

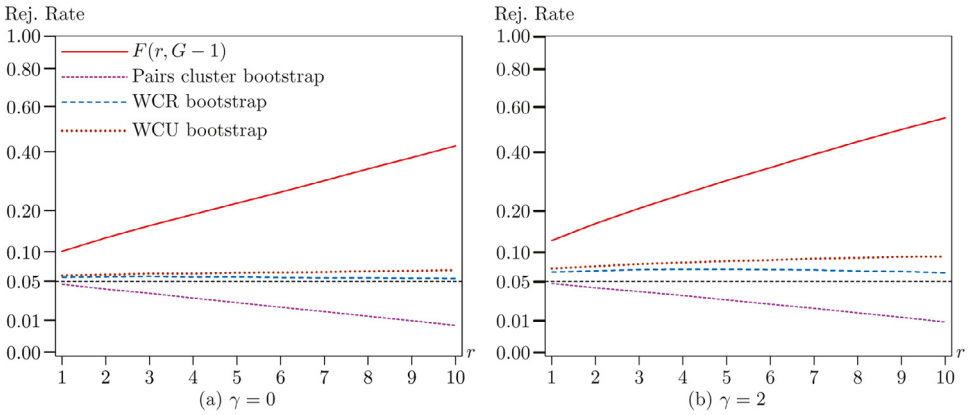

Fig. 2. Rejection Rates for Wald Tests with 40 Clusters. **Notes:** The vertical axis shows estimated rejection rates, after a square root transformation, for tests at the.05 level. These are based on 400,000 replications with $G = 40$, $N = 4000$, $k = 20$, and $B = 399$. The number of restrictions, r, varies from 1 to 10. Clusters are equal-sized in panel (a) and vary from 32 to 246 in panel (b).

Fig. 2 is similar to Fig. 1, except that $G = 40$ and $N = 4000$. In panel (a), all clusters are the same size. In panel (b), they vary from 32 to 246. With the exception of the pairs cluster bootstrap when $r = 1$, all the tests perform much better when $G = 40$ than when $G = 20$. In particular, the WCR bootstrap never rejects more than 5.8% of the time in panel (a) and 6.9% of the time in panel (b). Since the comparable numbers for $G = 20$ were 8.1% and 10.2%, doubling the number of clusters has evidently made WCR-based inference much more reliable. The pairs cluster bootstrap is still the best procedure for $r = 1$, but it continues to under-reject more and more severely as r becomes larger, albeit to a much lesser extent than it did for $G = 20$.

6.2. Bootstrap Tests for IV Regression

There is an enormous literature on the finite-sample properties of IV estimates, especially when the instruments are weak; see Andrews et al. (2019) for a recent survey. However, with the exception of Finlay and Magnusson (2019) and Wang and Zhang (2021), there has been little work on cluster-robust IV bootstrap methods. The experiments in this subsection attempt to remedy this omission, at least to a modest extent. They focus on WREC bootstrap tests for the parameter β in the linear model (38)–(39).

Studying the WREC bootstrap is much more complicated than studying bootstrap methods for OLS regression with clustering, because IV regressions with clustering involve two quite different types of finite-sample size distortion. As in the OLS case, one of these arises because cluster-robust standard errors may be inaccurate (usually too small). The other arises whether or not there is clustering, because the distribution of the IV estimator $\hat{\beta}$ given in (42), and therefore of t-statistics based on it, can differ greatly from its asymptotic distribution. In particular, $\hat{\beta}$ is often severely biased, especially when the number of over-identifying restrictions is large and the instruments are weak.

In the first set of experiments, there are 2000 observations and 20 clusters, each with 100 observations. None of the instruments, regressors, or disturbances actually displays intra-cluster correlation; that will be added below. Thus these experiments should yield results similar in most respects to those for the WRE bootstrap (Davidson and MacKinnon, 2010). Studying the WREC bootstrap is much cheaper than studying the WRE bootstrap, because the computational tricks of Section 4.1 cannot be used with the latter. However, the cluster-robust t-statistics are likely to be somewhat larger, on average, than heteroskedasticity-robust ones, and somewhat more variable across samples, especially when the number of clusters is small. If the WREC bootstrap works well, it should compensate for both of these phenomena at the cost of some power loss.

A key parameter for finite-sample inference about β in the model (38)–(39) is the concentration parameter C, which measures the strength of the instruments. When the disturbances are independent and identically distributed, as they are in the first set of experiments,

$$C = \frac{\mathrm{E}\big(\boldsymbol{\pi}_2^\top \boldsymbol{W}_2^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W}_2 \boldsymbol{\pi}_2\big)}{\sigma_2^2}, \tag{65}$$

where σ_2 is the standard deviation of the elements of $\boldsymbol{u}_2$. Thus C measures the explanatory power in the reduced-form regression (39) of the instruments that are not also regressors in the structural Eq. (38), relative to the variance of the disturbances in (39). In theoretical analyses, it is common to assume that $\boldsymbol{\pi}_2 = O(N^{-1/2})$, so that C remains constant as N increases. In practice, however, C is unknown and roughly proportional to the sample size. Since C is the noncentrality parameter of the F statistic for $\boldsymbol{\pi}_2 = \mathbf{0}$ in regression (39), it can be estimated (inconsistently) by running that regression and

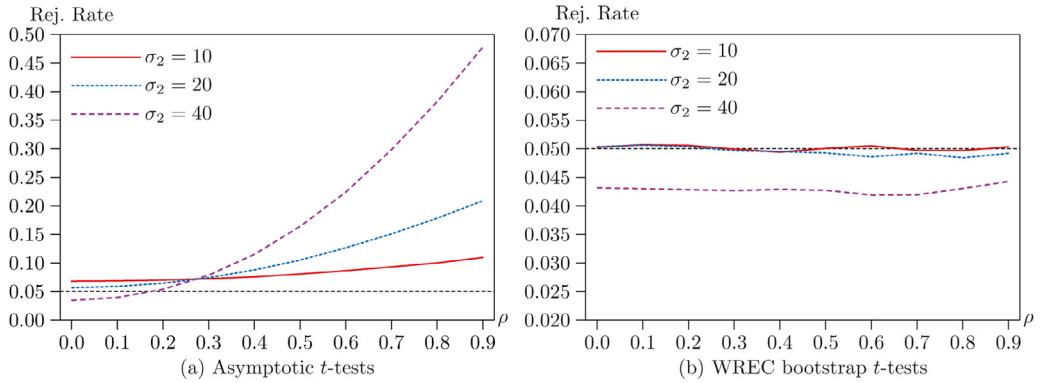

Fig. 3. Rejection Rates for IV Regression t-Tests. **Notes:** The vertical axis shows estimated rejection rates for tests at the.05 level for three values of σ_2. These are based on 400,000 replications with $G = 20$, $N = 2000$, $k = 5$, $l = 10$, and $B = 399$. Clusters are equal-sized, and there is no intra-cluster correlation. The correlation between the structural and reduced-form disturbances, ρ, is shown on the horizontal axis.

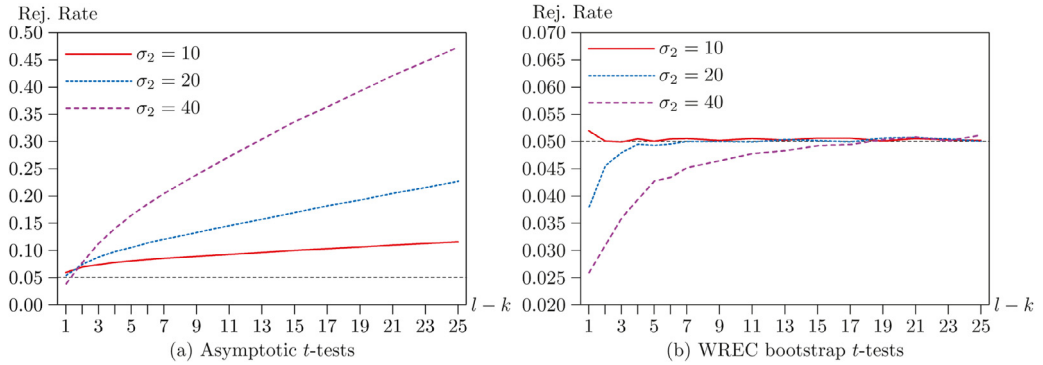

Fig. 4. Rejection Rates for IV Regression t-Tests. **Notes:** The vertical axis shows estimated rejection rates for tests at the.05 level for three values of σ_2. These are based on 400,000 replications with $G = 20$, $N = 2000$, $k = 5$, $\rho = 0.5$, and $B = 399$. Clusters are equal-sized, and there is no intra-cluster correlation. The difference between the number of instruments l and the number of exogenous regressors k is shown on the horizontal axis.

computing the F statistic. Stock and Yogo (2005) shows that this F statistic needs to be much larger than its usual critical value if t-tests and confidence intervals for β are to be reliable.

In the first set of experiments, all elements of $\boldsymbol{\pi}$ equal 1, and σ_2 takes on the values 10, 20, or 40, implying that $C/(l-k)$ equals 20, 5, or 1.25. It seems reasonable to argue, based on the tables in Stock and Yogo (2005), that $C/(l-k) = 20$ corresponds to moderately weak instruments, $C/(l-k) = 5$ corresponds to very weak instruments, and $C/(l-k) = 1.25$ corresponds to extremely weak instruments. In none of these cases would we expect the distribution of the t-statistic (51) to be well approximated by the $N(0,1)$ distribution.

Panel (a) of Fig. 3 shows rejection rates for asymptotic t-tests at the.05 level based on the standard normal critical value of 1.96 as a function of ρ, the correlation between the disturbances of the structural Eq. (38) and the reduced-form Eq. (39). The values of l and k are 10 and 5, respectively, so that there are 4 over-identifying restrictions. It is evident that the tests over-reject severely for all values of σ_2 when ρ is large. Perhaps surprisingly, however, they actually under-reject for $\sigma_2 = 40$ when ρ is small.

Panel (b) of Fig. 3 shows rejection rates for equal-tailed WREC bootstrap tests, defined in (63), for the same experiments. Note the greatly enlarged scale of the vertical axis! For the smallest value of σ_2, the WREC bootstrap tests reject almost exactly 5% of the time for all values of ρ. For the second-smallest value, this is still true, although there is very slight under-rejection for the larger values of ρ. Only for the largest value of σ_2, which corresponds to extremely weak instruments, do the WREC bootstrap tests not perform almost perfectly. They reject between 4.2% and 4.4% of the time. In contrast, the asymptotic tests reject between 3.4% and 47.8% of the time, depending on the value of ρ.

Fig. 4 is similar to Fig. 3, but now ρ is fixed at 0.5 and $l-k$ varies between 1 and 25, taking on the values $1, 2, \ldots, 7$ and $9, 11, \ldots, 25$. In panel (a), we see once again that the asymptotic tests usually over-reject severely, especially when the instruments are weak. However, all tests perform reasonably well when $l-k = 1$, so that there are no over-identifying restrictions.

In panel (b), where once again the vertical axis is on a greatly enlarged scale, we see that the WREC bootstrap tests always perform well for the smallest value of σ_2, but they under-reject in the left of the figure for the two larger values.

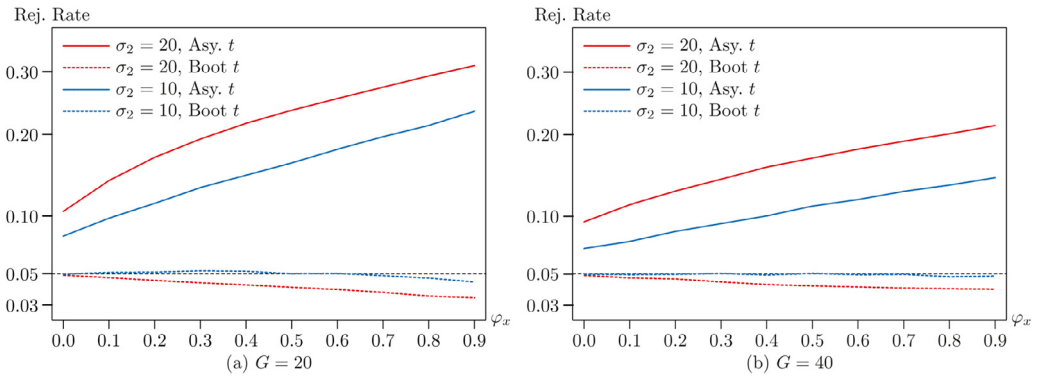

Fig. 5. Rejection Rates for IV Regression t-Tests with Intra-Cluster Correlation. **Notes:** The vertical axis shows estimated rejection rates for tests at the.05 level. These are based on 400,000 replications with $G = 20$ or $G = 40$, $N = 2000$, equal-sized clusters, $k = 5$, $l = 10$, $\rho = 0.5$, $\varphi_u = 0.1$, and $B = 399$. The intra-cluster correlation φ_x for the instruments and exogenous regressors is shown on the horizontal axis.

For $\sigma_2 = 20$, this under-rejection is not severe and is not really evident for $l - k \geq 5$, but for $\sigma_2 = 40$ it is quite severe and is evident even for fairly large values of $l - k$.

The tests shown in Fig. 3 and 4 are two-tailed. Results for one-tailed tests, especially asymptotic ones, are rather different. What usually happens, because $\hat{\beta}$ is biased upwards when $\rho > 0$, is that right-tail tests reject more often than two-tailed tests, and left-tail tests reject less often. For example, in the most extreme case of Fig. 3, where $\rho = 0.9$ and $\sigma_2 = 40$, the right-tail asymptotic test rejects 54.8% of the time, and the left-tail one never rejects. In contrast, the right-tail bootstrap test rejects 5.06% of the time, and the left-tail one rejects 4.23%.

In the next set of experiments, both the disturbances and the instruments are correlated within clusters. The instruments (which include the exogenous regressors), are generated using independent normal random-effects models parametrized so that their intra-cluster correlation is φ_x. Generating the disturbances is somewhat more complicated, because there needs to be both intra-cluster correlation and correlation between $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. This is done by using a normal random-effects model to generate three random N-vectors, $\boldsymbol{\nu}$, $\boldsymbol{\epsilon}_1$, and $\boldsymbol{\epsilon}_2$, all of them with intra-cluster correlation φ_u. Then $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are generated by

$$\boldsymbol{u}_j = \rho^{1/2}\boldsymbol{\nu} + (1-\rho)^{1/2}\boldsymbol{\epsilon}_j, \quad j = 1, 2. \tag{66}$$

The correlation between $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ is evidently ρ. Because both $\boldsymbol{\nu}$ and the $\boldsymbol{\epsilon}_j$ have intra-cluster correlation φ_u, so do the $\boldsymbol{u}_j$. The intra-cluster correlations of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ then arise endogenously from those of $\boldsymbol{Z}$, $\boldsymbol{W}$, $\boldsymbol{u}_1$, and $\boldsymbol{u}_2$.

The results of four sets of experiments are shown in Fig. 5. In all cases, the disturbances are generated using (66), with correlation coefficient $\varphi_u = 0.1$. The intra-cluster correlation φ_x of the instruments and exogenous regressors varies from 0.0 to 0.9 by increments of 0.1. There are two values of σ_2, namely, 10 and 20. No experiments were performed for $\sigma_2 = 40$, because the instruments, already extremely weak when $\varphi_x = 0$, would have been ridiculously weak for larger values of φ_x. Note that the instruments effectively become weaker as the intra-cluster correlations increase. Neither the usual concentration parameter (65) nor the usual F-statistic for $\boldsymbol{\pi}_2 = \boldsymbol{0}$ is valid in this case; see Olea and Pflueger (2013). When $\varphi_x = 0.0$, the rejection rates in both panels of Fig. 5 are essentially the same as the corresponding ones in Fig. 3. However, the rejection rates for asymptotic tests increase sharply with φ_x, especially when $\sigma_2 = 20$. In contrast, the rejection rates for bootstrap tests decline gradually with φ_x when $\sigma_2 = 20$ and hardly change when $\sigma_2 = 10$.

In panel (b) of Fig. 5, there are 40 clusters instead of 20. As expected, all the tests now perform better, and over-rejection by the asymptotic tests increases more slowly with φ_x. Notably, the WREC bootstrap tests perform more or less perfectly when $\sigma_u = 10$. Even when $\sigma_u = 20$, the worst they do is to reject about 4% of the time, when the corresponding asymptotic tests are rejecting 21.4% of the time.

The results in Fig. 5 remind us of a very important feature of cluster-robust inference, which has been known at least since Moulton (1986). What matters for inference are the intra-cluster correlations of the scores; see Section 2. When either the disturbances or the regressors display no intra-cluster correlation, then neither do the scores, and cluster-robust inference is asymptotically equivalent to heteroskedasticity-robust inference.

In another set of experiments, not reported here, the value of φ_x was set to 0.5, and the value of φ_u was varied. As expected, results for $\varphi_u = 0$ were almost identical to the ones for $\varphi_x = 0$ in Fig. 5. Then, as φ_u became larger, the rejection rates for asymptotic tests increased, and the ones for bootstrap tests decreased.

Results for the final set of experiments are shown in Fig. 6. In this case, what varies is γ, the parameter that determines how much cluster sizes differ; see Eq. (64). When $\gamma = 0$, all clusters have either 100 observations (in panel (a), where $G = 20$) or 50 observations (in panel (b), where $G = 40$). At the other extreme, when $\gamma = 4$, cluster sizes vary from 8 to 378 in panel (a) and from 3 to 215 in panel (b). In both panels, the asymptotic tests reject more frequently as cluster sizes become more variable. In panel (a), this is also true for the bootstrap tests, which initially under-reject. In panel (b), however,

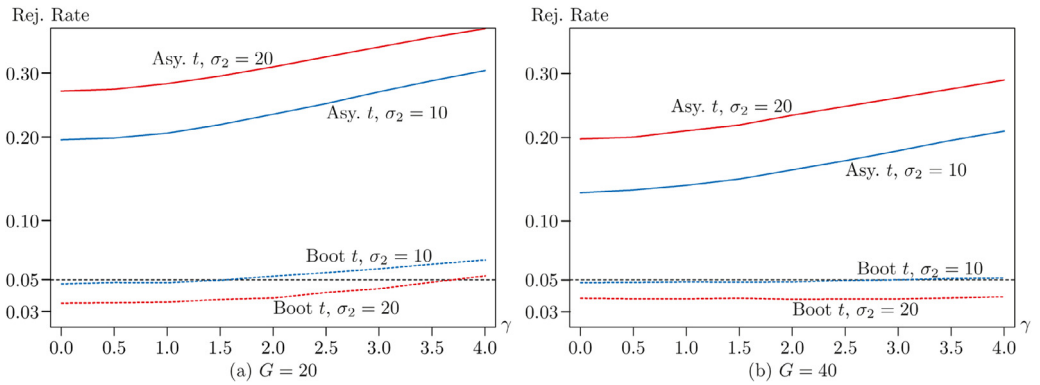

Fig. 6. Rejection Rates for IV Regression t-Tests with Intra-Cluster Correlation. **Notes:** The vertical axis shows estimated rejection rates for tests at the.05 level. These are based on 400,000 replications with $G = 20$ or $G = 40$, $N = 2000$, $k = 5$, $l = 10$, $\rho = 0.5$, $\varphi_x = 0.5$, $\varphi_u = 0.2$, and $B = 399$. The value of γ, which determines how much cluster sizes vary via equation (64), is shown on the horizontal axis.

the bootstrap tests for $\sigma_2 = 10$ perform almost perfectly for all values of γ, and the ones for $\sigma_2 = 20$ always under-reject moderately. Rejection rates for the latter vary between 3.7% and 3.9%.

Because the simultaneous equations model (38)–(39) has quite a few free parameters, the regressors and instruments can in principle be generated in a great many ways. In addition, the number of observations, the number of clusters, and the distribution of the cluster sizes can vary enormously. Thus the experiments reported in this section are far from definitive. It would be easy enough to generate datasets for which asymptotic inference would work almost perfectly. It would also be easy to generate datasets for which asymptotic inference was extremely inaccurate and even the WREC bootstrap worked poorly, perhaps because the instruments were extremely weak, the intra-cluster correlations were high, the number of clusters was small, and/or cluster sizes were highly variable. Nevertheless, it seems reasonable to conclude that tests based on the WREC bootstrap will generally yield more reliable inferences than ones based on IV t-statistics and asymptotic theory.

7. Conclusion

This paper proposes new computational methods for bootstrap tests in linear regression models with clustered disturbances. It also provides a new algebraic formulation for the existing method discussed in Roodman et al. (2019) and implemented in the Stata package `boottest`. For large samples, these methods can be several orders of magnitude faster than conventional ones. The key idea is to pre-compute a set of vectors and matrices of sums of squares and cross-products for each of the G clusters, which serve as sufficient statistics. The actual test statistic and all the bootstrap test statistics then depend on the sample only through these sufficient statistics. The simulations in Section 5 often show enormous reductions in computer time, especially for large sample sizes.

The new methods are particularly advantageous when applied to the pairs cluster bootstrap for models estimated by OLS (Section 3.1) and to the WREC bootstrap for models estimated by instrumental variables (Section 4.1). In the case of the WCR bootstrap for models estimated by OLS, however, they are not quite as efficient as the extraordinarily rapid method already used by `boottest`, which is explained in a new way in Section 3.2.

The Monte Carlo experiments in Section 6 contain a number of new findings. The pairs cluster bootstrap under-rejects severely when used with Wald tests of several restrictions in models estimated by OLS regression. The WCR bootstrap over-rejects more substantially than in previous studies, almost certainly because the number of random regressors is much larger. The WREC bootstrap for IV regression often performs remarkably well in cases where conventional t-tests over-reject greatly. In many cases (but not all) where it does not perform perfectly, the WREC bootstrap tends to under-reject. Increasing the amount of intra-cluster correlation for either the instruments or the disturbances effectively makes the instruments weaker, thus causing both t-tests and bootstrap tests to become less reliable.

Acknowledgements

I am grateful to Colin Cameron, Morten Nielsen, David Roodman, Takuya Ura, and Matt Webb for discussion and comments. I am also grateful to an Associate Editor and four referees for a number of valuable suggestions. Versions of the paper were presented at the University of California Davis and at the Canadian Econometric Study Group in Vancouver. This research was supported, in part, by the Social Sciences and Humanities Research Council of Canada (SSHRC grants 435-2016-0871 and 435-2021-0396).

References

Andrews, D.W.K., Moreira, M.J., Stock, J.H., 2006. Optimal two-sided invariant similar tests for instrumental variables regression. Econometrica 74 (3), 715–752.

Andrews, I., Stock, J.H., Sun, L., 2019. Weak instruments in instrumental variables regression: Theory and practice. Annual Review of Economics 11 (1), 727–753.
Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics 90 (3), 414–427.
Canay, I.A., Santos, A., Shaikh, A., 2021. The wild bootstrap with a 'small' number of 'large' clusters. Review of Economics and Statistics 103 (2), 346–363.
Chao, J.C., Swanson, N.R., 2005. Consistent estimation with a large number of weak instruments. Econometrica 73 (5), 1673–1692.
Davidson, R., Flachaire, E., 2008. The wild bootstrap, tamed at last. Journal of Econometrics 146 (1), 162–169.
Davidson, R., MacKinnon, J.G., 2008. Bootstrap inference in a linear equation estimated by instrumental variables. Econometrics Journal 11 (3), 443–477.
Davidson, R., MacKinnon, J.G., 2010. Wild bootstrap tests for IV regression. Journal of Business & Economic Statistics 28 (1), 128–144.
Djogbenou, A.A., MacKinnon, J.G., Nielsen, M.Ø., 2019. Asymptotic theory and wild bootstrap inference with clustered errors. Journal of Econometrics 212 (2), 393–412.
Efron, B., 1979. Bootstrapping methods: Another look at the jackknife. Annals of Statistics 7 (1), 1–26.
Finlay, K., Magnusson, L.M., 2019. Two applications of wild bootstrap methods to improve inference in cluster-IV models. Journal of Applied Econometrics 34 (6), 911–933.
Hansen, B.E., 1999. The grid bootstrap and the autoregressive model. Review of Economics and Statistics 81 (4), 594–607.
Imbens, G.W., Kolesár, M., 2016. Robust standard errors in small samples: Some practical advice. Review of Economics and Statistics 98 (4), 701–712.
Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70 (5), 1781–1803.
MacKinnon, J.G., 2015. Wild cluster bootstrap confidence intervals. L'Actualité économique 91 (1), 11–33.
MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2021. The summclust package: Leverage and influence in clustered regression models. QED Working Paper. Queen's University.
MacKinnon, J.G., Webb, M.D., 2007a. Wild bootstrap inference for wildly different cluster sizes. Journal of Applied Econometrics 32 (2), 233–254.
MacKinnon, J.G., Webb, M.D., 2017b. Pitfalls when estimating treatment effects using clustered data. The Political Methodologist 24 (2), 20–31.
MacKinnon, J.G., Webb, M.D., 2018. The wild bootstrap for few (treated) clusters. Econometrics Journal 21 (2), 114–135.
Moulton, B.R., 1986. Random group effects and the precision of regression estimates. Journal of Econometrics 32 (3), 385–397.
Nelson, C.R., Startz, R., 1990. Some further results on the exact small sample properties of the instrumental variables estimator. Econometrica 58 (4), 967–976.
Olea, J.L.M., Pflueger, C., 2013. A robust test for weak instruments. Journal of Business and Economic Statistics 31 (3), 358–369.
Pustejovsky, J.E., Tipton, E., 2018. Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. Journal of Business & Economic Statistics 36 (4), 672–683.
Racine, J.S., MacKinnon, J.G., 2007. Simulation-based tests that can use any number of simulations. Communications in Statistics–Simulation and Computation 36 (2), 357–365.
Roodman, D., MacKinnon, J.G., Nielsen, M.Ø., Webb, M.D., 2019. Fast and wild: Bootstrap inference in Stata using boottest. Stata Journal 19 (1), 4–60.
Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.
Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, D.W.K., Stock, J.H. (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, pp. 80–108.
Wang, W., Zhang, Y., 2021. Wild bootstrap for instrumental variables regressions with weak and few clusters. ArXiv e-prints.
Webb, M.D., 2014. Reworking wild bootstrap based inference for clustered errors. QED Working Paper. Queen's University.