Bootstrap Methods for Inference with Cluster-Sample IV Models

Keith Finlay and Leandro M. Magnusson*

January 31, 2009

Preliminary and incomplete. Please do not cite. Comments welcome.

Abstract

Microeconomic data often have within-cluster dependence. This dependence affects standard error estimation and inference in regression models, including the instrumental variables model. Standard corrections assume that the number of clusters is large, but when this is not the case, Wald tests can either over-reject or under-reject and weak instrument robust tests can over-reject. We examine the use of bootstrap methods to construct appropriate critical values for these tests when the number of clusters is small. We find that a variant of the Wild bootstrap performs well and reduces absolute size bias significantly, even with a small number of clusters. We also provide guidance in the choice among possible weak instrument robust tests when data have cluster dependence. These results should extend to fixed effect panel data models.

Keywords: Two-stage least squares, instrumental variables, hypothesis testing, weak instruments, clustered errors

JEL codes: C12, C15, C31

*Assistant Professors, Department of Economics, 206 Tilton Hall, Tulane University, New Orleans, LA 70118; kfinlay@tulane.edu and lmagnuss@tulane.edu.

1 Introduction

Microeconometric data often have a group structure. When regression errors are correlated within these groups or clusters, it is well-known that standard error estimates can be biased and hypothesis testing can be misleading. The common solution to this problem is to use cluster-robust standard error estimation methods. The asymptotic performance of these standard error estimators requires a large number of clusters. When the number of clusters is small, tests can be oversized even when cluster-robust methods are used. Research on the implications of cluster data in estimation has focused on ordinary least squares, but the problem is just as important in instrumental variables models.

In the linear IV model, we show that cluster-robust standard error adaptations fail when instruments are weak and when the number of clusters is small. In addition, we show that weak instrument robust tests also perform poorly when the number of clusters is small. We propose bootstrap techniques for weak-instrument robust tests that perform well even under weak instruments and as few as 5 clusters.

Our Monte Carlo simulations provide strong evidence of the benefit of bootstrap techniques in the IV model and with weak instrument robust tests. When we falsely assume that errors are iid, we find rejection rates as high as 0.50 with Wald or weak instrument robust tests. Cluster-robust versions of these tests can reduce the rejection rates to 0.10 to 0.30, but never as low as the nominal size of 0.05. Using our wild cluster constrained residual bootstrap, we get rejection rates that are very close to the 0.05. As is well know, bootstrap techniques cannot improve performance of the Wald test when instruments are weak. Using our bootstrap methods with the weak instrument robust tests provide a comprehensive and practical alternative for testing in the linear IV when data have cluster dependence.

The paper proceeds as follows. First, demonstrate the inference problems with a small number of clusters for standard Wald tests and weak instrument robust tests. Then, we describe our bootstrap techniques and the results of Monte Carlo experiments using them.

2 Cluster-robust inference

We consider a limited information cluster model with G clusters, index by g, and with N_g observations per clusters, indexed by i, represented by the following system:

$$y_{ig} = x_{ig}\beta + w_{ig}\gamma + u_{ig} \quad (1)$$

$$x_{ig} = z_{ig}\pi_z + w_{ig}\pi_w + v_{ig}. \quad (2)$$

where x_{ig} is a $1\times k_x$ vector of endogenous variables, w_{ig} is a $1\times k_w$ row vector of included instruments, z_{ig} is a $1 \times k_z$ vector of excluded instruments, u_{ig} and v_{ig} are residuals. We can rewrite the model as:

$$\mathbf{y}_g = \mathbf{x}_g\beta + \mathbf{x}_g\gamma + \mathbf{u}_g \quad (3)$$

$$\mathbf{x}_g = \mathbf{z}_g\pi_z + \mathbf{w}_g\pi_w + \mathbf{v}_g, \quad (4)$$

or simply

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{W}\gamma + \mathbf{U} \quad (5)$$

$$\mathbf{X} = \mathbf{Z}\pi_z + \mathbf{W}\pi_w + \mathbf{V}. \quad (6)$$

The errors $\mathbf{U}$ and vec($\mathbf{V}$) are independent across clusters, but have some dependence within clusters.

The two-stage least squares estimator is:

$$\hat{\beta}_{\text{2SLS}} = [\bar{\mathbf{X}}'\bar{\mathbf{X}}]^{-1}[\bar{\mathbf{X}}'\bar{\mathbf{y}}], \quad (7)$$

where $\bar{\mathbf{X}} = \mathbf{P}_{\mathbf{M_W Z}}\mathbf{X}$, $\mathbf{P}_{\mathbf{M_W Z}} = \mathbf{M_W Z}(\mathbf{Z}'\mathbf{M_W Z})^{-1}\mathbf{Z}'\mathbf{M_W}$, and $\mathbf{M_W} = \mathbf{I} - \mathbf{P_W}$ with $\mathbf{P_W} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$.

When the errors are assumed to be independent identically distributed, the variance of $\hat{\beta}_{\text{2SLS}}$ is $\sigma_u^2\left[\bar{\mathbf{X}}'\bar{\mathbf{X}}\right]^{-1}$. However, in the presence of intra-cluster dependence, even when this dependence is negligible, the two-stage least squares estimator of the variance of $\hat{\beta}_{\text{2SLS}}$ can be severely downward biased.

To see this point, let us assume that the number of observations per cluster is the same and equal to M, and the residual $\mathbf{u}_g$ can be decompose into individuals and cluster specific shocks, i.e., $\mathbf{u}_g = c_g\boldsymbol{\iota} + \boldsymbol{\varepsilon}_g$, where c_g is a intra-cluster specific effect with $E(c_g^2) = \sigma_c^2$ for all m, $\boldsymbol{\varepsilon}_g = (\varepsilon_{1,g}, \ldots, \varepsilon_{M,g})$ is the vector individual effects with $E(\varepsilon_{ig}^2) = \sigma_\varepsilon^2$ and $E(\varepsilon_{i,g}\varepsilon_{j,g}) = 0$ for $i \neq j$, and $\boldsymbol{\iota}$ is a vector of ones. Let us also assume that $\bar{\boldsymbol{X}}_g$ does not present within cluster variation. We can use the same arguments as in Moulton (1990) to show that ratio between the two-stage least squares variance and the iid asymptotic variance is $1 + (M-1)\rho_{c,\varepsilon}$, where $\rho_{c,\varepsilon} = \frac{\sigma_c^2}{\sigma_c^2+\sigma_\varepsilon^2}$ is the intra-cluster correlation. In this example, the usual iid two-stage least square variance under estimate the true variance by $(M-1)\rho_{c,\varepsilon}$.

The most commonly used variance estimator of $\hat{\beta}_{\text{2SLS}}$ is an adaptation of the Huber-White heteroskedasticity-robust sandwich estimator (White 1980), which does not impose any structure on the variance of the unobserved component:

$$\mathbf{V}[\hat{\beta}_{\text{2SLS}}] \quad = \quad \left[\bar{\mathbf{X}}'\bar{\mathbf{X}}\right]^{-1}\left[\sum_g \bar{\mathbf{X}}_g'\boldsymbol{\Omega}_g\bar{\mathbf{X}}_g\right]\left[\bar{\mathbf{X}}'\bar{\mathbf{X}}\right]^{-1}, \tag{8}$$

where $\boldsymbol{\Omega}_g = E(\boldsymbol{u}_g\boldsymbol{u}_g')$ is the variance-covariance matrix of the errors. Although the sandwich estimator is not the most efficient one, it does not suffer from the underestimation and is general enough to accommodate different residual structures. However, when the number of clusters is small, the asymptotic results from cluster-robust variance estimators are a poor approximation of the true distributions of the test statistics. We will return to this point later.

The crucial assumption behind the two-stage least squares estimator is that the instruments $\boldsymbol{Z}$ should be correlated with the explanatory endogenous variables. There exists a vast literature on weak instruments robust tests. In the context of linear instrumental variable model, we have the pioneering AR-test of Anderson and Rubin (1949), the Kleibergen (2002, 2007) score, or LM, test and its orthogonal complement, the J-test, and Moreira's (2003) conditional likelihood-ratio (CLR) test.

The standard derivation of the AR, LM, and CLR tests uses the optimal covariance matrix estimator (see Kleibergen (2007)), which is different from the non-optimal sandwich cluster matrix in equation (8). Davidson and MacKinnon (2006), Chernozhukov and Hansen (2008), and Finlay and Magnusson (2009) extend the prior weak instruments tests are also robust to arbitrary

heteroskedasticity or cluster dependence.

We can derive these robust versions of the AR, LM, and CLR tests by starting with the unrestricted reduced-form model that follows from equations (3) and (4):

$$\begin{cases} \mathbf{y}_g = \mathbf{Z}_g\delta_z + \mathbf{W}_g\delta_w + \mathbf{e}_g \\ \mathbf{X}_g = \mathbf{Z}_g\pi_z + \mathbf{W}_g\pi_w + \mathbf{v}_g \end{cases}, \tag{9}$$

where $\mathbf{e}_g = \mathbf{v}_g\beta + \mathbf{u}_g$ and $\delta_z = \pi_z\beta$ represents the restrictions imposed by the structural model over the reduced-form parameters. Tests on the structural parameter β can be conducted indirectly by testing if $\delta_z - \pi_z\beta = 0$ under the null hypothesis. In this representation, the identification of β requires that $\|\pi_z\| \neq 0$. As π_z approaches zero, the instruments become weaker.

As in the linear IV model, $\mathbf{W}$ can be projected out by $\mathbf{M_W}$ without affecting the inference. For simplicity, we keep the same notation for the regressors in the following discussion, although they are projected out of the space spanned by the columns of $\mathbf{W}$. We also omit the subscript z in δ_z and π_z. Let $(\hat{\delta}, \hat{\pi})$ be the OLS estimators of the parameters derived from the stacked model

$$\begin{bmatrix} \mathbf{y} \\ \text{vec}(\mathbf{X}) \end{bmatrix} = \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & I_{k_x} \otimes \mathbf{Z} \end{bmatrix} \begin{bmatrix} \delta \\ \text{vec}(\pi) \end{bmatrix} + \begin{bmatrix} \mathbf{e} \\ \text{vec}(\mathbf{v}) \end{bmatrix}, \tag{10}$$

and $\hat{\Lambda}$ be the heteroskedasticity and cluster-autocorrelation consistent cluster variance estimator: $\hat{\Lambda} = \mathbf{B}^{-1}\mathbf{A}\mathbf{B}^{-1}$, where $\mathbf{B} = I \otimes (\mathbf{Z}'\mathbf{Z})$, $\mathbf{A} = \sum_g (I \otimes \mathbf{Z}_g')\hat{\mathbf{\Upsilon}}_g(I \otimes \mathbf{Z}_g)$,

$$\hat{\mathbf{\Upsilon}}_g = \begin{bmatrix} \hat{\mathbf{e}}_g\hat{\mathbf{e}}_g' & \hat{\mathbf{e}}_g\text{vec}(\hat{\mathbf{v}}_g)' \\ \text{vec}(\hat{\mathbf{v}}_g)\hat{\mathbf{e}}_g' & \text{vec}(\hat{\mathbf{v}}_g)\text{vec}(\hat{\mathbf{v}}_g)' \end{bmatrix}, \tag{11}$$

and $(\hat{\mathbf{e}}_g, \hat{\mathbf{v}}_g)$ are the OLS residuals. The consistency properties of the OLS estimators do not depend on the value of the structural parameter β. The $(k_x+1)k_z \times (k_x+1)k_z$ variance $\hat{\Lambda}$ can be partitioned accordingly as

$$\hat{\Lambda} = \begin{bmatrix} \hat{\Lambda}_{\delta\delta} & \hat{\Lambda}_{\delta\pi} \\ \hat{\Lambda}_{\pi\delta} & \hat{\Lambda}_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \hat{\Lambda}_{\delta\delta} & \hat{\Lambda}_{\delta\pi_1} & \cdots & \hat{\Lambda}_{\delta\pi_{k_x}} \\ \hat{\Lambda}_{\pi_1\delta} & \hat{\Lambda}_{\pi_1\pi_1} & \cdots & \hat{\Lambda}_{\pi_1\pi_{k_x}} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Lambda}_{\pi_{k_x}\delta} & \hat{\Lambda}_{\pi_{k_x}\pi_2} & \cdots & \hat{\Lambda}_{\pi_{k_x}\pi_{k_x}} \end{bmatrix}.$$

Define the $k_z \times k_z k_x$ matrix $\hat{\Lambda}_{\pi_i \pi} = \left[\hat{\Lambda}_{\pi_i \pi_1} \quad \dots \quad \hat{\Lambda}_{\pi_i \pi_{k_x}} \right]$ for $i = 1, \dots, k_x$,[1] and $\hat{\Lambda}_{\pi_i \pi, \beta} = \hat{\Lambda}_{\pi_i \pi}(\beta \otimes I_{k_z})$. Let us introduce four more statistics, before introducing the weak instruments robust tests:

$$\begin{aligned}
\hat{\Psi}_\beta &= \hat{\Lambda}_{\delta\delta} - (I_{k_z} \otimes \beta')\hat{\Lambda}_{\delta\pi} - \hat{\Lambda}_{\pi\delta}(\beta \otimes I_{k_z}) + (I_{k_z} \otimes \beta')\hat{\Lambda}_{\pi\pi}(\beta \otimes I_{k_z}) \\
\hat{\bar{\pi}}_\beta &= \hat{\pi} - \left[\hat{\Lambda}_{\pi_1, \delta} - \hat{\Lambda}_{\pi_1 \pi, \beta} \hat{\Psi}_\beta^{-1} (\hat{\delta} - \hat{\pi}\beta), \dots, \hat{\Lambda}_{\pi_{k_x} \delta} - \hat{\Lambda}_{\pi_{k_x} \pi, \beta} \hat{\Psi}_\beta^{-1} (\hat{\delta} - \hat{\pi}\beta) \right] \\
\hat{\Xi}_\beta &= \hat{\Lambda}_{\pi\pi} - \left[\hat{\Lambda}_{\pi\delta} - \hat{\Lambda}_{\pi\pi}(\beta \otimes I_{k_z}) \right] \hat{\Psi}_\beta^{-1} \left[\hat{\Lambda}_{\delta\pi} - (\beta' \otimes I_{k_z}) \hat{\Lambda}_{\pi\pi} \right] \\
\mathrm{rk}(\beta) &= \left\{ \hat{\bar{\pi}}_\beta' \hat{\Xi}_\beta^{-1} \hat{\bar{\pi}}_\beta \right\}.
\end{aligned}$$

The first statistic is an estimate of the asymptotic covariance matrix of $\sqrt{n}(\hat{\delta} - \hat{\pi}\beta)$. The second statistic is an estimator of π which is independent of $(\hat{\pi}_z - \hat{\Pi}_z \beta)$.[2] The third statistic is an estimator of the covariance of $\mathrm{vec}(\hat{\bar{\pi}}_\beta)$, and $\mathrm{rk}(\beta)$ is a test for the rank of $\hat{\bar{\pi}}_\beta$.

We define the weak instruments robust tests for the cluster model as follows:

Definition 1. *(Weak instruments robust tests for the cluster model) Under the null hypothesis $H_0 : \beta = \beta_0$, the AR, LM, J, and CLR are:*

$$AR(\beta_0) = \left(\hat{\delta} - \hat{\pi}\beta_0 \right)' \hat{\Psi}_{\beta_0}^{-1} \left(\hat{\delta} - \hat{\pi}\beta_0 \right) \tag{12}$$

$$LM(\beta_0) = \left[\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \left(\hat{\delta} - \hat{\pi}\beta_0 \right) \right]' \hat{\mathrm{P}}_{\beta_0} \left[\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \left(\hat{\delta} - \hat{\pi}\beta_0 \right) \right] \tag{13}$$

$$J(\beta_0) = \left[\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \left(\hat{\delta} - \hat{\pi}\beta_0 \right) \right]' \hat{\mathrm{M}}_{\beta_0} \left[\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \left(\hat{\delta} - \hat{\pi}\beta_0 \right) \right] \tag{14}$$

$$\begin{aligned}
CLR(\beta_0) = \frac{1}{2} \Big\{ & AR(\beta_0) - \mathrm{rk}(\beta_0) + \\
& \sqrt{[AR(\beta_0) + \mathrm{rk}(\beta_0)]^2 - 4J(\beta_0)\mathrm{rk}(\beta_0)} \Big\}
\end{aligned} \tag{15}$$

where:

$$\hat{\mathrm{P}}_{\beta_0} = \left(\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \hat{\bar{\pi}}_{\beta_0} \right) \left(\hat{\bar{\pi}}_{\beta_0}' \hat{\Psi}_{\beta_0}^{-1} \hat{\bar{\pi}}_{\beta_0} \right)^{-1} \left(\hat{\Psi}_{\beta_0}^{-\frac{1}{2}} \hat{\bar{\pi}}_{\beta_0} \right)'$$

$$\hat{\mathrm{M}}_{\beta_0} = I_{k_z} - \hat{\mathrm{P}}_{\beta_0}$$

[1] The matrix $\hat{\Lambda}_{\pi_i}$ is the i^{th} block row of matrix $\hat{\Lambda}_{\pi\pi}$.

[2] More details are discussed in (Magnusson 2008).

Asymptotically we have:

$$\begin{aligned} AR(\beta_0) &\xrightarrow{d} \chi^2(k_z) \\ LM(\beta_0) &\xrightarrow{d} \chi^2(k_x) \\ J(\beta_0) &\xrightarrow{d} \chi^2(k_z - k_x) \end{aligned}$$

where the value inside the parentheses indicates the χ^2 distribution degrees of freedom. The CLR-test converges to a nonpivotal distribution. However, its critical values, for a given value of $rk(\beta)$, can be simulated from independent chi-squared distributions. The convergence of the tests are independent of whether the instruments are weak.

The AR tests simultaneously the value of the structural parameter and the overidentification restriction. It can be decomposed into two orthogonal two statistics, namely the LM and J tests. Under the null hypothesis, the LM statistic tests the value of the structural parameter given that the overidentification condition holds, while the J statistic tests the overidentification restriction given that the value of β_0. It is well-known that the LM-test suffers from a spurious decline of power at some regions of the parameter space. In those regions, the J-test approximates the AR-test which always has discriminatory power.

The asymptotic distribution of the CLR is not pivotal and depends on $\text{rk}(\beta_0)$. The critical values of this test are calculated by simulating independent values of $\chi^2(1)$ and $\chi^2(k_z - 1)$ for a given value of $\text{rk}(\beta_0)$. This approach is not satisfactory because accuracy demands a large number of simulations which can be computationally intensive. For linear IV models under homoskedasticity, Andrews, Moreira, and Stock (2007) provide a formula for computing the p-value function of the CLR-test. Although this is not the correct p-value function when homoskedasticity is violated, our simulations indicate that it provides a good approximation.

The cluster-robust versions of the weak instrument robust tests can reduce size bias provided that the number of clusters is large. In much applied microeconometrics, there are data with intra-cluster dependence in which the number of clusters are small and the asymptotic results are a poor approximation of the true distributions of the test statistics. For example, many papers in labor economics use research designs that rely on policy changes at the state level, in which the

number of clusters is 50. Tests that use cluster-robust variance estimators still overreject with as many as 50 clusters. We now present discuss the potential for the bootstrap to provide improved test performance even with as few as 5 clusters. The weak instrument robust tests provide valid inference even when instruments are weak, and so the bootstrap techniques also provide valid inference in this case.

3 Cluster bootstrap methods

The bootstrap provides one set of possible solutions to improve inference when cluster-dependent data are used. Resampling is usually done at the same level as the clustering. In this section, we describe some applications to single-equation models and, then, instrumental variables models. We then describe the variant of the bootstrap that we use.

Recent work has highlighted the use of the bootstrap to improve inference when there is intra-cluster dependence. Bertrand, Duflo, and Mullainathan (2004) examine the bootstrap, but their bootstrap sampling techniques do not fully account for the clustered error structure of the data generating process. Recently, Cameron, Gelbach, and Miller (2008) show that a variant of the wild bootstrap with cluster-based sampling performs well in a variety of cases, and bootstrap techniques dominate any cluster-robust standard error estimation technique. But these papers only examine use in OLS.

In general, this literature has not examined the cluster-robust inference in the two-stage least squares model.

Bootstrap methods only offer asymptotic improvements when used with pivotal statistics. The AR, LM, and Wald tests all have χ^2 pivotal distributions. These tests behave like symmetrical, two-tailed tests, and under H_0 the difference between their nominal size and true rejection rates are $O(n^{-1})$ with asymptotic critical values but $O(n^{-2})$ with bootstrap critical values (Horowitz 2001). The CLR test, however, is pivotal only after conditioning on the rk statistic. For linear IV models under homoskedasticity, Andrews et al. (2007) provide a formula for computing the p-value function of the CLR-test. Although this is not the correct p-value function when homoskedasticity is violated, our simulations indicate that it provides a good approximation (Finlay and Magnusson 2009).

Moreira, Porter, and Suarez (2004, forthcoming) develop the asymptotic theory for using the bootstrap with IV models. Using Edgeworth expansions, they show that the bootstrap can improve inference even with weak instruments. Davidson and MacKinnon (2006) develop bootstrap techniques for IV models using weak instrument robust tests, but they assume homoskedasticity.

Although there is some evidence that the bootstrap can improve inference using the Wald test in an IV model when there is strong identification, this is not a solution under weak identification. In fact, in small samples the numerical accuracy of the bootstrapped Wald test may be even worse than that of first-order asymptotic approximations when instruments are weak.

Gelbach, Klick, and Stratmann (2007) implement a variant of the wild cluster bootstrap of Cameron et al. (2008) in an instrumental variables setting. They examine its performance in Monte Carlo simulations and find that it performs well. But they do not consider tests that are robust to weak instruments.

3.1 Wild cluster constrained residual bootstrap

We now discuss our bootstrap algorithm, which combines features from the algorithms of Cameron et al. (2008), Gelbach et al. (2007), and Davidson and MacKinnon (2006). Recall the model in equations (3) and (4). Let $\hat{T}$ be the test statistic generated from the original sample using this model. The wild cluster constrained residual bootstrap works in the following way:

1. Under the null hypothesis $H_0 : \beta = \beta_0$, run the restricted OLS regression $\mathbf{y} - \mathbf{X}\beta_0 = \mathbf{W}\gamma + \mathbf{u}$. Define the constrained estimated residuals by $\hat{\mathbf{u}}^R$ and the constrained estimate $\hat{\gamma}(\beta_0)$.

2. Sample $(\eta_1, \ldots, \eta_G)$ where $\eta_g = \kappa_g \xi_g - E(\kappa_g)E(\xi_g)$,

$$\kappa_g \sim N\left(\frac{\sqrt{\frac{17}{6}} + \sqrt{\frac{1}{6}}}{2}, \frac{1}{2}\right), \text{ and} \tag{16}$$

$$\xi_g \sim N\left(\frac{\sqrt{\frac{17}{6}} - \sqrt{\frac{1}{6}}}{2}, \frac{1}{2}\right). \tag{17}$$

3. Define the bootstrap residuals as $\{\hat{\mathbf{u}}_1^*, \ldots, \hat{\mathbf{u}}_G^*\} = \{\eta_1 \hat{\mathbf{u}}_1^R, \ldots, \eta_G \hat{\mathbf{u}}_G^R\}$ and define the bootstrap sample $\{(\hat{\mathbf{y}}_1^*, \mathbf{X}_1, \mathbf{W}_1), \ldots, (\hat{\mathbf{y}}_G^*, \mathbf{X}_G, \mathbf{W}_G)\}$, where $\hat{\mathbf{y}}_g^* = \mathbf{X}_g\beta_0 + \mathbf{W}_g\hat{\gamma}(\beta_0) + \hat{\mathbf{u}}_g^*$.

4. Estimate the IV model using the bootstrap sample. Compute the bootstrap statistic T_b^* using the bootstrap sample.

5. Repeat steps (2), (3), and (4) B times. Compute the simulated p-value of the original statistic $\hat{T}$:

$$\hat{p}(\beta_0) = 1 - \frac{1}{B}\sum_{b=1}^{B} I(T_b^* < \hat{T}) \tag{18}$$

   Reject or do not reject the test accordingly.

Notes of the bootstrap algorithm

Residual weights. Weights used for the wild bootstrap must have three properties to ensure that the resampled residuals have the same first three moments as the sample residuals: $E[\eta_i] = 0$, $E[\eta_i^2] = 1$, and $E[\eta_i^3] = 1$. Many residual weights have been proposed for the wild bootstrap that satisfy these properties. The most commonly used weights come from discrete distributions. These include the Mammen (1993) weights, defined as

$$\eta_i = \begin{cases} (1-\sqrt{5})/2 & \text{with probability } \frac{1+\sqrt{5}}{2\sqrt{5}} \\ 1-(1-\sqrt{5})/2 & \text{with probability } 1-\frac{1+\sqrt{5}}{2\sqrt{5}} \end{cases},$$

and the Rademacher weights, defined as

$$\eta_i = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}.$$

Liu (1988) proposes a few continuous distributions from which to draw the residual weights. Since we are resampling entire clusters of residuals, we find that the continuous weights performed better because they provide more replication variation. The weights presented in the algorithm above come from Liu (1988), so we call these the Liu normal weights, which are also used in Gelbach et al. (2007). She also proposes a set of weights defined as $\eta_g = \zeta_g - E(\zeta_g)$, where ζ_g is a gamma random variable with shape parameter 4 and scale parameter $\frac{1}{2}$. We call these the Liu gamma weights, and also present results using these weights.

Resampling first-stage residuals. In Davidson and MacKinnon's (2006) bootstrap algorithm, residuals are not just resampling from the second-stage of the IV model, but also from the first-stage. In our simulations (not shown), we found that this step offers no reduction in the size bias, so we present the simpler algorithm.[3]

Bootstrap replications. In our experiments, we use 399 bootstrap replications. In repeated Monte Carlo experiments, the sampling error from a small number of bootstrap replications should cancel out. In actual practice, at least 999 replications should be used.

4 Monte Carlo simulations

We evaluate the performance of the wild cluster bootstrap using Monte Carlo simulations. We experimented with a variety of data generating processes. Our data generation process has a structure that resembles the random effects model. This dgp takes the form:

$$y_{ig} = \beta_0 + \beta_1 x_{ig} + u_{ig} \tag{19}$$

$$= \beta_0 + \beta_1 x_{ig} + (\epsilon_g + \epsilon_{ig}), \tag{20}$$

and

$$x_{ig} = \alpha_0 + \alpha_1 z_{1,ig} + \alpha_2 z_{2,ig} + e_{ig} \tag{21}$$

$$= \alpha_0 + \alpha_1 (a_g + a_{ig}) + \alpha_2 b_{ig} + (\eta_g + \eta_{ig}), \tag{22}$$

where ϵ_g, a_g, a_{ig}, b_{ig}, and η_g are each drawn independently from a standard normal distribution, and ϵ_{ig} and η_{ig} are drawn from a bivariate normal distribution $\mathcal{N}\left[\binom{0}{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$. We set $\beta_0 = 0$, $\beta_1 = 1$, $\alpha_0 = 0$, and $\alpha_2 = 0$.

We set α_1 according to a pseudo concentration parameter $\mu = \alpha_1^2 \left(\text{Var}(\alpha_1)\right)^{-1}$, where

$$\text{Var}(\alpha_1) = \left(\sum_g Z_g' Z_g\right)^{-1} \left(\sum_g Z_g' \Sigma_g Z_g\right) \left(\sum_g Z_g' Z_g\right)^{-1}, \tag{23}$$

[3] Results are available upon request.

and

$$\begin{aligned}\Sigma_g &= E[e_g e_g'] \\ &= \sigma^2_{\eta_g} \iota_g \iota_g' + \sigma^2_{\eta_{ig}} I_g.\end{aligned}$$

Here, ι_g is a vector of ones, I_g is an $N_g \times N_g$ identity matrix, $E[\eta^2_{ig}]\sigma^2_{\eta_{ig}}$ is the individual-level variance, and $E[\eta^2_g] = \sigma^2_{\eta_g}$ is the group-level variance. Using these cluster-robust equations, we experiment with three values for the concentration parameter: $\mu = 1$ for weak identification, $\mu = 10$ for good identification, and $\mu = 50$ for strong identification.

Table 1 shows the results using the group-level random errors dgp using 5 cluster each with 30 observations. For each test, we show the rejection rate if we (1) assume that the errors are iid, (2) use the cluster-robust variance estimator, (3) use the wild cluster bootstrap-t with Liu normal weights, and (4) use the wild cluster bootstrap-t with Liu gamma weights. With 5 clusters, all tests perform poorly when we assume that the errors are independently distributed or if we use the cluster-robust versions of the tests. Assuming iid errors, the weak instrument robust tests have rejections rates between 0.452 and 0.510. Using the cluster-robust variance estimator version of the tests, the rejection rate is reduced to between 0.387 and 0.432. This is minor improvement, and not sufficient for valid inference. The next two rows show the rejection rates when using the two bootstrap methods. For the CLR and LM tests, the rejection rates are close to the nominal size of 0.05 (between 0.046 and 0.059).

With the Wald test, there is much more variation in the rejection rates, and this variation is clearly a result of the level of identification. When the instruments are weak (the concentration parameter $\mu = 1$), the rejection rates are approximately 0.30 when we assume the errors are iid, 0.20 when we use the cluster-robust Wald test, and 0.01 using the wild cluster bootstrap. When the instruments provide identification (the concentration parameter $\mu = 50$), the rejection rates are approximately 0.50 when we assume the errors are iid, 0.30 when we use the cluster-robust Wald test, and 0.01 using the wild cluster bootstrap.

Table 2 shows the results using the group-level random errors dgp using 10 cluster each with 30 observations. With 10 clusters, all tests perform poorly when we assume that the errors are independently distributed or if we use the cluster-robust versions of the tests. Assuming iid errors, the weak instrument robust tests have rejections rates between 0.462 and 0.525. Using the cluster-

robust variance estimator version of the tests, the rejection rate is reduced to between 0.177 and 0.208. While this is an improvement, it is not sufficient for valid inference. The next two rows show the rejection rates when using the two bootstrap methods. For the CLR and LM tests, the rejection rates are close to the nominal size of 0.05 (between 0.038 and 0.060).

With the Wald test, there is much more variation in the rejection rates, and this variation is clearly a result of the level of identification. When the instruments are weak (the concentration parameter $\mu = 1$), the rejection rates are approximately 0.30 when we assume the errors are iid, 0.04 when we use the cluster-robust Wald test, and 0.01 using the wild cluster bootstrap. When the instruments provide identification (the concentration parameter $\mu = 50$), the rejection rates are approximately 0.50 when we assume the errors are iid, 0.15 when we use the cluster-robust Wald test, and 0.02 using the wild cluster bootstrap.

5 Conclusion

In further work, we will examine the performance of the wild cluster bootstrap when different data generating processes are used. We will also describe the power properties of the bootstrapped tests.

References

Anderson, T. W. and Herman Rubin. 1949. Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics* 20(1): 46–63.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2007. Performance of Conditional Wald Tests in IV Regression with Weak Instruments. *Journal of Econometrics* 139(1): 116–32.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics* 119(1): 249–75.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics* 90(3): 414–27.

Chernozhukov, Victor and Christian Hansen. 2008. The Reduced Form: A Simple Approach to Inference with Weak Instruments. *Economics Letters* 100(1): 68–71.

Davidson, Russell and James MacKinnon. 2006. Bootstrap Inference in a Linear Equation Estimated by Instrumental Variables. Queen's Economics Department Working Paper 1024.

Finlay, Keith and Leandro M. Magnusson. 2009. Implementing Weak Instrument Robust Tests for a General Class of Instrumental Variables Models. Tulane University Working Paper 0901.

Gelbach, Jonah B., Jonathan Klick, and Thomas Stratmann. 2007. Cheap Donuts and Expensive Broccoli: The Effect of Relative Prices on Obesity. Working paper.

Horowitz, Joel L. 2001. The Bootstrap. In J.J. Heckman and E.E. Leamer, editors, *Handbook of Econometrics*, volume 5, chapter 52, pp. 3159–228. Elsevier.

Kleibergen, Frank. 2002. Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression. *Econometrica* 70(5): 1781–1803.

———. 2007. Generalizing Weak Instrument Robust IV Statistics Towards Multiple Parameters, Unrestricted Covariance Matrices and Identification Statistics. *Journal of Econometrics* 139(1): 181–216.

Liu, Regina Y. 1988. Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics* 16(4): 1696–1708.

Magnusson, Leandro M. 2008. Inference in Limited Dependent Variable Models Robust to Weak Identification. Tulane Economics Working Paper 0801. URL: `http://ideas.repec.org/p/tul/wpaper/0801.html`.

Mammen, Enno. 1993. Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *The Annals of Statistics* 21(1): 255–85.

Moreira, Marcelo J. 2003. A Conditional Likelihood Ratio Test for Structural Models. *Econometrica* 71(4): 1027–48.

Moreira, Marcelo J., Jack R. Porter, and Gustavo A. Suarez. 2004. Bootstrap and Higher-Order Expansion Validity When Instruments May Be Weak. National Bureau of Economic Research Technical Working Paper 302.

———. forthcoming. Bootstrap Validity for the Score Test When Instruments May Be Weak. *Journal of Econometrics* .

Moulton, Brent. 1990. An Illustration of the Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *Review of Economics and Statistics* 72(2): 334–38.

White, Halbert. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817–38.

Table 1: Size (in percent) for testing $H_0 : \beta = 0$ at the 5% significance level, group-level random errors, 5 clusters with 30 observations each

	μ	1	1	1	10	10	10	50	50	50
Test	ρ	0.1	0.5	0.8	0.1	0.5	0.8	0.1	0.5	0.8
Method		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CLR										
1 Assume iid		47.9	48.0	47.6	50.5	50.5	50.7	50.7	50.7	51.0
		(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		40.3	40.3	40.0	41.0	41.4	41.7	43.0	43.2	42.6
		(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
3 Wild cluster bootstrap-t		5.2	5.3	5.3	5.3	5.4	5.4	5.1	5.1	5.4
(Liu normal weights)		(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
4 Wild cluster bootstrap-t		5.2	5.5	5.6	5.2	5.3	5.6	5.4	5.7	5.9
(Liu gamma weights)		(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.8)
LM										
1 Assume iid		45.3	45.2	45.9	50.4	50.4	50.5	50.7	50.7	51.0
		(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		38.8	38.7	38.9	40.9	41.1	41.9	42.8	43.1	42.6
		(1.5)	(1.5)	(1.5)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
3 Wild cluster bootstrap-t		4.8	4.6	5.0	5.2	4.9	5.2	5.1	5.1	5.3
(Liu normal weights)		(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
4 Wild cluster bootstrap-t		4.9	4.9	5.0	5.1	5.2	5.2	5.4	5.6	5.9
(Liu gamma weights)		(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.8)
Wald										
1 Assume iid		28.9	30.4	31.8	48.0	47.8	47.2	50.2	50.0	49.9
		(1.4)	(1.5)	(1.5)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		19.0	19.8	20.7	27.9	28.0	27.4	30.3	30.4	30.4
		(1.2)	(1.3)	(1.3)	(1.4)	(1.4)	(1.4)	(1.5)	(1.5)	(1.5)
3 Wild cluster bootstrap-t		0.7	0.5	0.7	0.9	0.8	0.8	0.9	0.9	0.8
(Liu normal weights)		(0.3)	(0.2)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)
4 Wild cluster bootstrap-t		0.8	0.6	0.9	1.1	1.1	0.8	1.3	1.0	1.1
(Liu gamma weights)		(0.3)	(0.2)	(0.3)	(0.3)	(0.3)	(0.3)	(0.4)	(0.3)	(0.3)

Note: Authors' calculation from 999 Monte Carlo simulations. For bootstrap methods, each experiment consists of 399 bootstrap replications. Simulation standard errors in parentheses.

Table 2: Size (in percent) for testing $H_0 : \beta = 0$ at the 5% significance level, group-level random errors, 10 clusters with 30 observations each

	μ	1	1	1	10	10	10	50	50	50
Test	ρ	0.1	0.5	0.8	0.1	0.5	0.8	0.1	0.5	0.8
Method		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CLR										
1 Assume iid		49.6	49.5	49.6	51.8	52.1	52.1	52.5	52.4	52.4
		(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		19.8	19.5	20.0	20.5	20.4	20.8	20.8	20.5	20.5
		(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)
3 Wild cluster bootstrap-t		5.7	5.4	5.4	4.8	4.9	5.1	4.8	4.8	4.7
(Liu normal weights)		(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
4 Wild cluster bootstrap-t		5.9	6.0	6.0	5.4	5.3	5.5	5.1	5.0	5.1
(Liu gamma weights)		(0.8)	(0.8)	(0.8)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
LM										
1 Assume iid		46.9	46.9	47.7	51.8	51.9	52.3	52.5	52.4	52.4
		(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		18.0	17.7	18.9	19.8	19.5	19.6	20.4	20.3	20.3
		(1.2)	(1.2)	(1.2)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)	(1.3)
3 Wild cluster bootstrap-t		3.8	3.8	4.1	4.3	4.5	4.6	4.7	4.7	4.6
(Liu normal weights)		(0.6)	(0.6)	(0.6)	(0.6)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
4 Wild cluster bootstrap-t		4.2	4.0	4.5	4.7	4.9	5.0	5.0	4.9	4.8
(Liu gamma weights)		(0.6)	(0.6)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)	(0.7)
Wald										
1 Assume iid		29.1	31.1	33.4	49.7	49.3	49.7	52.2	52.4	52.4
		(1.4)	(1.5)	(1.5)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)	(1.6)
2 Cluster-robust		4.4	4.4	4.4	12.9	12.3	12.6	15.8	15.5	15.4
		(0.7)	(0.7)	(0.7)	(1.1)	(1.0)	(1.1)	(1.2)	(1.1)	(1.1)
3 Wild cluster bootstrap-t		0.4	0.4	0.6	0.9	0.9	0.7	2.4	2.4	2.2
(Liu normal weights)		(0.2)	(0.2)	(0.2)	(0.3)	(0.3)	(0.3)	(0.5)	(0.5)	(0.5)
4 Wild cluster bootstrap-t		0.1	0.3	0.6	0.8	0.9	1.0	2.3	2.4	2.4
(Liu gamma weights)		(0.1)	(0.2)	(0.2)	(0.3)	(0.3)	(0.3)	(0.5)	(0.5)	(0.5)

Note: Authors' calculation from 999 Monte Carlo simulations. For bootstrap methods, each experiment consists of 399 bootstrap replications. Simulation standard errors in parentheses.