Consistency without Inference: Instrumental Variables in Practical Application*

Alwyn Young
London School of Economics
This draft: September 2017

Abstract

I use the bootstrap to study a comprehensive sample of 1400 instrumental variables regressions in 32 papers published in the journals of the American Economic Association. IV estimates are more often found to be falsely significant and more sensitive to outliers than OLS, while having a higher mean squared error around the IV population moment. There is little evidence that OLS estimates are substantively biased, while IV instruments often appear to be irrelevant. In addition, I find that established weak instrument pre-tests are largely uninformative and weak instrument robust methods generally perform no better or substantially worse than 2SLS.

---

*I am grateful to Ruoqi Zhou for excellent research assistance.

I: Introduction

The economics profession is in the midst of a “credibility revolution” (Angrist and Pischke 2010) in which careful research design has become firmly established as a necessary characteristic of applied work. A key element in this revolution has been the use of instruments to identify causal effects free of the potential biases carried by endogenous ordinary least squares regressors. The growing emphasis on research design has not gone hand in hand, however, with equal demands on the quality of inference. Despite the widespread use of Eicker (1963)-Hinkley (1977)-White (1980) robust and clustered covariance estimates, in finite samples heteroskedastic and correlated errors still produce test statistics whose distribution is typically much more dispersed than believed. This complicates inference in both ordinary and two stage least squares, but is compounded in the latter, where first stage joint test statistics are used to buttress the credibility of second stage results. In this paper I show that two stage least squares (hereafter, 2SLS or IV) methods produce estimates that, in practice, rarely identify parameters of interest more accurately or substantively differently than is achieved by biased ordinary least squares (OLS).

I use the bootstrap to study the distribution of test statistics for a comprehensive sample of 1533 instrumented coefficients in 1400 2SLS regressions in 32 papers published in the journals of the American Economic Association. I maintain, throughout, the exact specification used by authors and their identifying assumption that the excluded variables are orthogonal to the second stage residuals. When I bootstrap, I draw samples in a fashion consistent with the error dependence within groups of observations and independence across observations implied by authors’ standard error calculations. Thus, this paper is not about point estimates or the validity of fundamental assumptions, but rather concerns itself with the quality of inference within the framework established by authors themselves. The bootstrap shows that conventional tests have rejection rates much greater than nominal size, i.e. systematically understate confidence intervals, and that these distortions grow in joint tests. Bootstrapping the bootstrap, however, I also find that the bootstrap itself continues to understate confidence intervals. Thus, the results reported below are likely to be generous.

I find that, depending upon the bootstrap method used, 2SLS point estimates are falsely declared significant between ⅓ and ½ of the time, while their bootstrapped 99 percent confidence interval includes the OLS point estimate between 92 and 94 percent of the time and the entirety of the bootstrapped OLS 99 percent confidence interval between 75 and 83 percent of the time. The extraordinary sampling variability of IV estimates is reflected in their sensitivity to outliers. With the removal of only or two clusters or observations 45 and 63 percent, respectively, of reported .01 significant 2SLS results can be rendered insignificant at the same level. I find that only 8 to 14 percent of regressions can reject the null that the OLS estimates are in fact unbiased at the .01 level. This is important because the ln mean squared error of 2SLS around its own population moment is on average 4.77 greater than the ln mean squared error of OLS around its population moment, so if OLS is unbiased the use of 2SLS is, from a quadratic loss point of view, a regrettable choice. Surprisingly, I find that the ln mean squared error of 2SLS around its population moment is on average 1.52 greater than that of OLS around the same moment, i.e. in applied work biased OLS is on average more accurate in estimating the IV population moment than 2SLS itself! Moreover, the bias of 2SLS methods is greater than the bias of OLS (from the 2SLS moment) in about 1/6 of coefficients. I find that the null that all first stage coefficients are zero can only be rejected at the .01 level between 52 to 70 percent of the time, i.e. in ⅓ to ½ of published regressions one cannot reject the null that the instruments are totally irrelevant and the observed correlation between the endogenous variables and the excluded instruments, despite the exogeneity of the latter in the population, is due to a wholly undesirable finite sample correlation between the instruments and the endogenous errors. Only one in ten to twelve instrumented coefficients resides in a regression that rejects the instrument irrelevance and the no-OLS bias nulls at the .01 level. Only 5 to 6 percent of instrumented coefficients meet these standards of credibility while producing a confidence interval that does not contain the OLS point estimate.

Weak instruments can, in principle, play a role in many of the maladies described above. A weak first stage relationship results in variation of predicted values well below that of the original regressors, producing less efficient estimates than OLS. Weak instruments also produce biased coefficients with highly non-normal distributions whose tail variation may be much greater than believed, generating empirical rejection rates under the null much greater than nominal size. I find that conventional first stage F statistics have a much wider sampling distribution than recognized, with those based on the default covariance estimate showing an average rejection probability of .284 at the .01 level when the null of zero effects is true, which remains as high as .160 when clustered/robust covariance estimates are used. In light of this, I construct bootstrap equivalent F-statistics by inverting the bootstrapped p-value and find that instruments are much weaker than believed, with only 30 to 40 percent of regressions showing a bootstrapped F greater than 10. However, outside of the very weakest of instruments, with Fs less than 1, no F-statistic of any form is associated with second stage size biases. While weak instruments distort the distribution of second stage test statistic distributions, beyond the weakest cases these effects are completely dominated by the biases created by the departure of heteroskedastic, correlated and non-normal errors from the iid normal ideal, making inference with both 2SLS and OLS equally inaccurate, despite the use of cluster/robust covariance estimates. The simplest evidence of this is the fact that the average excess size of 2SLS and OLS versions of the same regressions are roughly the same. Non-iid errors confound the measurement of instrument strength, but also make it less relevant, as instrument strength is not the main source of size biases.

There is a growing professional tendency to use weak instrument pre-tests based upon conventional first stage F-statistics to establish the credibility of 2SLS results. I apply the weak instrument pre-tests of Stock and Yogo (2005) and show that, outside of the most extreme cases where the conventional first stage F is less than 1, they provide little protection against excess size as coefficients which pass the tests have rejection rates under the null which are no better than those which do not. The bounds on both size and bias that are supposed to be assured by the tests are grossly violated by regressions which pass. Given this, and the fact that I find that conventional Fs greater than 10 have a .089 to .213 probability of arising when the instruments are in fact completely irrelevant, the increasing use of these pre-tests to legitimize results is unfortunate.

The finding that first stage relations are much weaker than believed might lead to an increased reliance on weak instrument robust methods. My results indicate this would be ill-advised. I examine the performance relative to 2SLS of three weak instrument robust inference and estimation methods. The Anderson-Rubin (1949) method tests second stage significance by projecting the dependent variable directly on the excluded instruments and has been endorsed by a number of econometricians as a solution to the problem of inference with weak instruments (e.g. Dufour 2003, Baum, Schaffer and Stillman 2007, Chernozhukov and Hansen 2008). It performs very poorly, as the projection onto a broader space magnifies size distortions in over-identified equations, while in exactly identified equations it has few advantages since, as already noted, with non-ideal errors outside of the most pathologically weak first stage relations OLS and 2SLS have similar coverage bias. The limited information maximum likelihood (LIML) method has been found, in analysis with iid disturbances, to have better median bias and coverage bias, especially with weak instruments, than conventional 2SLS

(Anderson, Kunitomo & Sawa 1982, Anderson 1983, Staiger and Stock 1997, and Stock and Yogo 2005). In practical application, I find that none of these properties holds, as LIML performs much worse than 2SLS on bias and mean squared error (particularly when instruments are weak), with no advantages in size. Fuller (1977) designed his k method as an adjustment that creates finite moments for the LIML estimator, and, in iid settings, Rothenberg (1984) showed that to a second-order approximation Fuller's method is the unbiased k-class estimator with minimum mean squared error, while Stock and Yogo (2005) concluded that bias in Fuller's k is more robust to weak instruments than 2SLS. I find that Fuller's method has lower mean squared error and bias than 2SLS, but, again outside of extreme cases with conventional Fs less than 1, this is unrelated to any measure of the strength of instruments. In sum, in practical application most of the predictions of iid based theory are found to be untrue.

The concern with the quality of inference in 2SLS raised in this paper is not new. Sargan, in his seminal 1958 paper, raised the issue of efficiency and the possibility of choosing the biased but more accurate OLS estimator, leading later scholars to explore relative efficiency in Monte Carlo settings (e.g. Summers 1965, Feldstein 1974). The current professional emphasis on first stage F-statistics as pre-tests originates in Nelson and Startz (1990a, b), who used examples to show that size distortions can be substantial when the strength of the first stage relationship is weak, and Bound, Jaeger and Baker (1995), who emphasized problems of bias and inconsistency with weak instruments. These papers spurred path-breaking research, such as Staiger and Stock (1997) and Stock and Yogo's (2005) elegant derivation and analysis of weak instrument asymptotic distributions, renewed interest in older weak instrument robust methods, and motivated the use of such techniques in critiques of selected papers (e.g. Albouy 2012, Bazzi and Clemens 2013). The contribution of this paper within this literature is two-fold: first, in showing that departures from the iid ideal that motivates most theoretical work raises important questions, both regarding the measurement of instrument strength and the practical usefulness of iid based results; and, second, in highlighting the importance of redirecting some attention away from uninformative first-stage pre-tests to older concerns regarding mean squared error and the relative efficiency of 2SLS and OLS.

The paper proceeds as follows: Section II below begins by describing the rules used to select the sample and its characteristics. I have not allowed myself any discretion in picking papers or regressions and, subject to some basic rules regarding data and code availability and methods applied, have used all papers produced by a search on the AEA website. Section III provides a brief review of notation and the bootstrap methods used in the paper highlighting the reason why the bootstrap itself may produce rejection rates greater than nominal size. Section IV presents the results described above concerning the sample itself, while sections V and VI review the disappointing performance of weak instrument pre-tests and weak instrument robust methods. Section VII concludes.

All of the results of this research are anonymized. Thus, no information can be provided, in the paper, public use files or private conversation, regarding results for particular papers. Methodological issues are more important than individual results and studies of this sort rely upon the openness and cooperation of authors. For the sake of transparency, I provide complete code (in preparation) that shows how each paper was analysed, but the reader eager to know how a particular paper fared will have to execute this code themselves. Public use data files (in preparation) provide the results and principal characteristics of each 2SLS regression in an anonymized fashion, allowing researchers to reproduce the tables in this paper and use the data in further analysis.

II. The Sample

My sample is based upon a search on www.aeaweb.org using the keyword "instrument" covering the American Economic Review and the American

Economic Journals for Applied Economics, Economic Policy, Microeconomics and Macroeconomics which, at the time of its implementation, yielded papers up through the July 2016 issue of the AER. I then dropped papers that:

(a) did not provide public use data files and Stata do-file code[1];
(b) did not include instrumental variables regressions;
(c) used non-linear methods or non-standard covariance estimates;
(d) provided incomplete data or non-reproducible regressions.

Public use data files are necessary to perform any analysis, and I had prior experience with Stata and hence could analyse do-files for this programme at relatively low cost. Stata is by far the most popular programme as, among papers that provide data, only five make use of other software. The keyword search brought up a number of papers that deal with instruments of policy, rather than instrumental variables, and these were naturally dropped from the sample.

Conventional linear two stage least squares with either the default or clustered/robust covariance estimate is the overwhelmingly dominant approach, so, to keep the discussion focused, I dropped four rare deviations. These include two papers that used non-linear IV methods, one paper which clustered on two variables, the union of which encompassed the entire sample making it impossible to implement a pairs bootstrap that respects the cross-correlations the authors believe exist in the data, and another paper which uniquely used auto-correlation consistent standard errors (in only 6 regressions) using a user-written routine that does not provide formulas or references in its documentation. One paper used Fuller's modification of LIML methods. Since this method is considered a weak instrument robust alternative to 2SLS, I keep the paper in the sample, examining its regressions with conventional 2SLS and, along with the rest of the sample, using LIML and Fuller methods. A smattering of GMM methods appear in two

---

[1] Conditional on a Stata do-file, a non-Stata format data file was accepted.

papers whose 2SLS regressions are otherwise included in the sample. Inference in the generalized method of moments raises issues of its own that are best dealt with elsewhere, so I exclude these regressions from the analysis.

Many papers provide partial data, indicating that users should apply to third parties for confidential data necessary to reproduce the analysis. As the potential delay and likelihood of success in such applications is indeterminate, I dropped these papers from my sample. One paper provided completely "broken" code, with key variables missing from the data file, and was dropped from the sample. Outside of this case, however, code in this literature is remarkably accurate and with the exception of two regressions in one paper (which were dropped),[2] I was able to reproduce, within rounding error or miniscule deviations on some coefficients, all of the regressions reported in tables. I took as my sample only IV regressions that appear in tables. Alternative specifications are sometimes discussed in surrounding text, but catching all such references and linking them to the correct code is extremely difficult. By limiting myself to specifications presented in tables, I was able to use coefficients, standard errors and supplementary information like sample sizes and test statistics to identify, interpret and verify the relevant parts of authors' code. Two papers critique the results of other papers. I use these, as they provide data and code for 73 regressions in 9 papers otherwise not in my sample.[3]

As shown in Table I, my final sample consists of 32 papers, 16 appearing in the American Economic Review and 16 in other AEA journals.[4] Of the 1400 IV regressions in these papers, 1359 involve only one endogenous variable, with 1087

---

[2] I also dropped 2 other regressions that had 10 million observations, as bootstrapping these regressions is beyond the computational resources available to me.

[3] In referring to "papers" below, however, I count all papers reviewed in a single review paper as one paper.

[4] The AEJ: Microeconomics does not appear in the final sample because the six papers that showed up in my search either did not provide a data file or used non-linear methods.

Table I: Characteristics of the Sample

32 papers		1400 2SLS regressions (1533 coefficients)					
	Journal	endogenous regressors & excluded instruments		covariance estimate		distribution	
16	AER	1087	1 & 1	105	default	767	t & F
7	AEJ: A. Econ.	272	1 & >1	1035	clustered	633	N & chi^2
4	AEJ: E. Policy	41	>1 & >1	260	robust		
5	AEJ: Macro						

Notes: AER = American Economic Review; AEJ = American Economic Journal; t, F, N & chi^2 = t, F, standard normal and chi-squared distributions.

of these exactly identified by one excluded instrument. Multiple testing, in the form of multiple coefficients of substantive interest presented within one estimating equation, is extremely rare, with only 41 equations involving more than one instrumented endogenous variable. This contrasts strongly with practice in field and laboratory experiments, where I find in Young (2017) that about half of estimating equations include multiple treatment measures, with an average of 5 treatment measures per equation. When equations are overidentified, the number of instruments can be quite large, with an average of 18 excluded instruments (median of 14) for 1.3 endogenous variables in 300 overidentified equations in 15 papers. Thus, econometric issues concerning the higher dimensionality of instruments are relevant in a substantial subset of equations and papers.

Turning to statistical inference, almost all of the papers in my sample use the Eicker (1963)-Hinkley (1977)-White (1980) robust covariance matrix or its multi-observation cluster extension. One paper uses the default 2SLS covariance estimate, as does a solitary regression in another paper where the old fixed effects command the author chose for that regression did not allow the clustering used in other regressions in the same paper. Different Stata commands make use of different distributions to evaluate the significance of the same 2SLS estimating equation, with the sample roughly equally divided between results evaluated using

Table II: Tests of Normality, Cluster Correlation and Heteroskedasticity
(fraction of regressions rejecting the null at 1x10^{-10} and .01 levels)

	1400 second stage regressions		1533 first stage regressions	
	1x10^{-10}	.01	1x10^{-10}	.01
normality of residuals	.709	.854	.665	.857
no cluster fixed effects	.805	.942	.801	.940
homoskedasticity (Breusch & Pagan 1979)	.780	.873	.730	.863
homoskedasticity (Koenker 1981)	.526	.729	.487	.704
homoskedasticity (Wooldridge 2013)	.573	.756	.530	.728

Notes: As the theory underlying the homoskedasticity tests is developed in an OLS framework, the second stage tests are performed using the OLS version of the authors' estimating equation.

the t and F (with finite sample covariance corrections) and those evaluated using the normal and chi^2. In directly evaluating authors' results, I use the distributions and methods they chose. For more general comparisons, however, I move everything to a consistent basis using, in turn, either the default or clustered/robust[5] covariance estimates and the same t and F distributions (with finite sample covariance corrections) for all 2SLS and OLS results.

Table II shows that non-normality, intra-cluster correlation and heteroskedasticity of the disturbances are important features of the data generating process in my sample. Using Stata's test of normality based upon skewness and kurtosis, I find that about ⅔ of first and second stage regressions reject the null that the residuals are normal at the 1x10^{-10} level and about .85 at the .01 level. In equations which cluster, cluster fixed effects are found to be significant .80 of the time at the 1x10^{-10} level and .94 of the time at the .01 level. In close to ½ of these

---

[5] I use the robust covariance estimate for the paper that used the default covariance estimate throughout, cluster the solitary regression mentioned above, and also cluster three regressions where the author used the robust covariance estimate but otherwise clustered all other regressions in the paper.

regressions the authors' original specification includes cluster fixed effects, but where there is smoke there is likely to be fire, i.e. it is unlikely that the cluster correlation of residuals is limited to a simple mean effect; a view apparently shared by the authors, as they cluster standard errors despite including cluster fixed effects. Tests of homoskedasticity involving the regression of squared residuals on the authors' right-hand size variables and cluster fixed effects (where authors cluster), using the test statistics and distributions suggested by Breusch and Pagan (1979), Koenker (1981) and Wooldridge (2013), reject the null between .487 to .780 of the time at the $1x10^{-10}$ level and between .704 and .873 of the time at the .01 level. These tests are based upon the assumption that the disturbances are iid normal or at least iid, and hence the distribution theory underlying each test is often undermined by the results of the others,[6] so it would be incorrect to conclude that they show that residuals are simultaneously non-normal, correlated and heteroskedastic. They do indicate, however, significant departures, on at least some dimension, from the iid ideal. This is borne out by the poor predictive power of iid based theory in this sample, as shown later.

III. Notation and Bootstrap Methods

(a) Notation

I follow fairly standard notation. With lower case bold letters indicating vectors and upper case bold letters matrices, the data generating process is taken as given by:

$$(1)\quad \mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\delta} + \mathbf{u}$$

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{X}\boldsymbol{\Delta} + \mathbf{V}$$

---

[6] The Breusch & Pagan test assumes the residuals are iid normal, while the other two tests of homoskedasticity assume they are iid. The test of normality assumes the residuals are iid. The test for cluster fixed effects uses the default covariance matrix, as the number of cluster fixed effects equals the maximum possible rank of the cluster covariance matrix, and hence implicitly assumes that with these fixed effects the residuals are iid. I should note that in implementing the normality and heteroskedasticity tests on residuals, where authors weight I use the weights to remove the known heteroskedasticity in the residuals before running the tests.

where **y** is the n x 1 vector of second stage outcomes, **Y** the n x $k_{\mathbf{Y}}$ matrix of endogenous regressors, **X** the n x $k_{\mathbf{X}}$ matrix of included exogenous regressors, **Z** the n x $k_{\mathbf{Z}}$ matrix of excluded exogenous regressors (instruments), **u** the n x 1 vector of second stage disturbances, and **V** the n x $k_{\mathbf{Y}}$ matrix of first stage disturbances. The remaining (Greek) letters are parameters, with **β** representing the parameters of interest. The nuisance variables **X** and their associated parameters are of no substantive interest, so I use $^{\sim}$ to denote the residuals from the projection on **X** and characterize everything in terms of these residuals. For example, with $^{\wedge}$ denoting estimated and predicted values, the coefficient estimates for OLS and 2SLS are given by:

$$(2)\quad \hat{\boldsymbol{\beta}}_{ols} = (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}'\tilde{\mathbf{y}}, \quad \hat{\boldsymbol{\beta}}_{2sls} = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\tilde{\mathbf{Y}}}'\tilde{\mathbf{y}}, \quad \text{where } \hat{\tilde{\mathbf{Y}}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\tilde{\mathbf{Y}}.$$

(b) The Bootstrap

Conventional econometrics uses assumptions and asymptotic theorems to infer the distribution of a statistic f calculated from a sample with empirical distribution F_1 drawn from an infinite parent population with distribution F_0, which can be described as $f(F_1|F_0)$. In contrast, the bootstrap estimates the distribution of $f(F_1|F_0)$ by drawing random samples F_2 from the population distribution F_1 and observing the distribution of $f(F_2|F_1)$ (Hall 1992). If f is a smooth function of the sample, then asymptotically the bootstrapped distribution converges to the true distribution (Lehmann and Romano 2005), as, intuitively, the outcomes observed when sampling F_2 from an infinite sample F_1 approach those arrived at from sampling F_1 from the actual population F_0.

In bootstrapping the distribution of a test statistic one derives a bootstrap estimate of its p-value and simultaneously learns about the size distortions of the conventional test. To be concrete, let μ_0 denote a moment of the parent population that is of interest. In conventional statistics we observe the concomitant moment μ_1 in F_1, construct a test statistic measuring its distance from a hypothesized value of μ_0, say 0, and use theory to evaluate its likelihood of arising in sampling. By drawing samples F_2 and testing $\mu_2 = \mu_1$, i.e. centring the test statistic around the known population moment of F_1, we learn something about the distribution of the test statistic for F_1 of $\mu_1 = \mu_0 = 0$ under the null $\mu_0 = 0$ for F_0. This forms the basis of the bootstrap p-value. At the same time, by applying conventional p-value calculations to each test of $\mu_2 = \mu_1$ we learn about size, as μ_1 is the true null when sampling F_2 from F_1. It is worth emphasizing that these size calculations have no necessary implications regarding the validity of the null for F_0, although it is true that a finding that the test statistic has unexpectedly large dispersion (producing positive size distortions) will typically result in a bootstrap upward adjustment of the conventional p-value for tests of the null for F_0.

The bootstrap can be iterated to identify its own size distortions. Consider drawing a sample F_2 from F_1, calculating the test statistic for the test $\mu_2 = \mu_1$, and then drawing bootstrap samples F_3 from F_2 and using the distribution of the test statistic for $\mu_3 = \mu_2$ to evaluate the p-value of the earlier test statistic. Performing this many times reveals the size distortions of the bootstrap, since, in each case, one is using the bootstrap to evaluate a null that is known to be true. In principle, knowledge of these size distortions could be used to adjust the original bootstrap p-value for the test of $\mu_1 = \mu_0$ (Hall 1986, Hall & Martin 1988). However, given the large number of regressions studied in this paper, I find this to be beyond my computing resources, as it requires the calculation of 1000s of bootstraps for each of the original 1000s of bootstraps used to evaluate the paper's test statistics. I do, however, draw 10 iterated bootstraps for each test, allowing me to calculate the average bootstrap size distortion across all the tests I examine, although not providing enough information to adjust any given one.

I make use of two bootstrap test statistics. The first is the bootstrap-c, which uses the bootstrap distribution of coefficients to calculate their covariance matrix and Wald statistics, computing the probability:

$$\text{(3) } (\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1)'\mathbf{V}(\boldsymbol{\beta}_2)^{-1}(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1) > (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}(\boldsymbol{\beta}_2)^{-1}(\boldsymbol{\beta}_1 - \mathbf{0})$$

where $\boldsymbol{\beta}_1$ is the vector of coefficients estimated using the sample F_1, $\boldsymbol{\beta}_2^i$ is the vector of coefficients estimated in the i^{th} draw of sample F_2 from F_1, $\mathbf{V}(\boldsymbol{\beta}_2)$ is the covariance matrix of $\boldsymbol{\beta}_2$ calculated across all draws, and **0** is the null hypothesis being tested in the original population. In the case of an individual coefficient, the common variance in the denominator on both sides can be cancelled and the method reduces to calculating the probability:

$$\text{(4) } (\beta_2^i - \beta_1)^2 > (\beta_1)^2$$

If the distribution of coefficients is unbiased and normal, this amounts to calculating their variance. Any estimate of variance is, however, subject to sampling variation.[7] Since the distributions used to evaluate test statistics are convex around critical values, this variation tends to produce size greater than nominal value, even in the bootstrap. Iterating the bootstrap by bootstrapping the bootstrap, as described earlier, amounts to estimating the variance of the estimate of the variance, which can be used to adjust the bootstrap critical values. Further iterations could, in turn, estimate the variance of this estimate, and so forth. At each stage, one can take a step up what Mosteller and Tukey (1977) describe as the "misty staircase" of statistical inference, deriving an estimate of the variance of the previous estimate of variance.

The second bootstrap measure I use is the bootstrap-t, which uses the iteration by iteration covariance estimate to calculate the Wald statistic, computing the probability:

$$\text{(5) } (\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1)'\mathbf{V}(\boldsymbol{\beta}_2^i)^{-1}(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1) > (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}(\boldsymbol{\beta}_1)^{-1}(\boldsymbol{\beta}_1 - \mathbf{0})$$

---

[7] This is not a matter of sampling variation arising from using a finite number of bootstrap draws from F_1 to estimate the variance of the coefficients. Rather, just as for any conventional estimate of variance, the bootstrap estimate itself is a function of the sample F_1 drawn from F_0.

where $\mathbf{V}(\boldsymbol{\beta}_2^i)$ is the conventional covariance estimate of $\boldsymbol{\beta}_2^i$ calculated in the ith draw and $\mathbf{V}(\boldsymbol{\beta}_1)$ is the conventional covariance estimate for the original sample. In the case of an individual coefficient, this amounts to estimating the distribution of squared t-statistics calculating the probability:

$$(6)\ \left[\frac{\beta_2^i - \beta_1}{\hat{\sigma}(\beta_2^i)}\right]^2 > \left[\frac{\beta_1}{\hat{\sigma}(\beta_1)}\right]^2$$

where $\hat{\sigma}$ denotes the estimated standard error of the coefficient. If the coefficients and standard errors followed normally based distributions, this would amount to calculating the degrees of freedom of the t-distribution, which identifies the variance of the variance. Thus, in principle the bootstrap-t can place one further up the misty staircase, attaining higher accuracy without the computational cost of iterating the bootstrap. This relies, however, upon the conventional estimate of variance being roughly accurate. If the conventional variance estimate is poor, the bootstrap-t can prove less accurate than the bootstrap-c, as it simply adds noise to the estimated distribution of coefficients (compare equations (4) and (6)), providing not the desired estimate of the variance of the variance of coefficients, but simply a noisier estimate of the variance. I find that in the case of 2SLS estimates, where the conventional estimate of variance is very inaccurate, the bootstrap-t has less accurate size than the bootstrap-c, while in the case of OLS estimates, where conventional variance estimates are closer to actual coefficient variation, it performs as well and often much better than the bootstrap-c. Results with both measures are reported in the tables below.

IV: Consistency without Inference

Table III reports the statistical significance of the coefficients of instrumented right-hand side variables using conventional and bootstrap techniques. As shown, using authors' covariance calculation methods and chosen distribution (normal or t), .322 of instrumented coefficients are statistically

Table III: Coefficient Significance: Rejection Rates and Size Distortions
using Conventional Methods and the Bootstrap
(1533 coefficients in 1400 regressions)

	two stage least squares				ordinary least squares			
	rejection rates		average size		rejection rates		average size	
	.01	.05	.01	.05	.01	.05	.01	.05
authors' methods	.322	.502	.052	.109				
clustered/robust	.297	.478	.046	.100	.509	.603	.049	.102
default	.393	.525	.095	.165	.579	.691	.127	.204
bootstrap - t	.213	.374	.028	.071	.431	.545	.023	.072
bootstrap - c	.151	.277	.019	.054	.449	.569	.024	.071

Notes: .01/.05 = level of the test; rejection rates = fraction of coefficients rejecting the null of 0; average size = based upon the bootstrap, average rejection rate of the null when true. Bootstrap-t implemented using the clustered/robust variance estimate.

significant at the .01 level and .502 at the .05 level. As authors use diverse methods, the remainder of the table, and (unless otherwise noted) all further analysis below, evaluates results using consistent formats and distributions. In the second row I use the robust or clustered covariance matrix for each equation, and in the third I use the default covariance estimate throughout. Both are evaluated using the t-distribution with the same degrees of freedom and finite sample covariance adjustments as OLS, to facilitate comparisons with that method. As expected, use of the t-distribution lowers significance rates slightly relative to those found with the normal distribution used by authors in almost half of the regressions in the first row. Also as expected, the default covariance estimate produces somewhat higher rejection rates, with the difference concentrated in papers which cluster to correct for the well-known bias brought about by the correlation between errors within clusters and instrumented "treatment" which does not vary within clusters (Kloek 1981, Moulton 1986).[8]

---

[8] In regressions which cluster across multiple observations, moving from the default to the clustered covariance estimate lowers the fraction of .01 significant results from .47 to .33, while in

The changes wrought by the use of different conventional distributions or covariance estimates are trivial relative to those found by applying the bootstrap. As shown in Table III, when the distribution of t-statistics (bootstrap-t) is used to evaluate significance, significance rates fall by about ⅓ at each level, while when the distribution of coefficients (bootstrap-c) is used, significance rates fall by ½. The changes in p-values are substantial, as shown in Figure I, which plots the bootstrap p-values against the conventional clustered/robust p-values of the second row of Table III. Among 2SLS coefficients which are found to be .01 significant using authors' methods, but not so using the bootstrap-t, the average p-value rises from .004 to .040, with ¼ of these showing bootstrap p-values in excess of .052, while among those for which the bootstrap-c reverses significance, the average p-value rises from .003 to .075, with ¼ of these showing bootstrap p-values in excess of .108. [9] Using authors' methods all of the 32 papers in my sample have at least one .05 significant instrumented coefficient, and all but 4 have at least one .01 significant coefficient. Using the bootstrap-t and -c, 3 and 6 papers, respectively, have no .05 significant coefficients and a total of 10 and 14, in each case, have no .01 significant coefficients whatsoever.

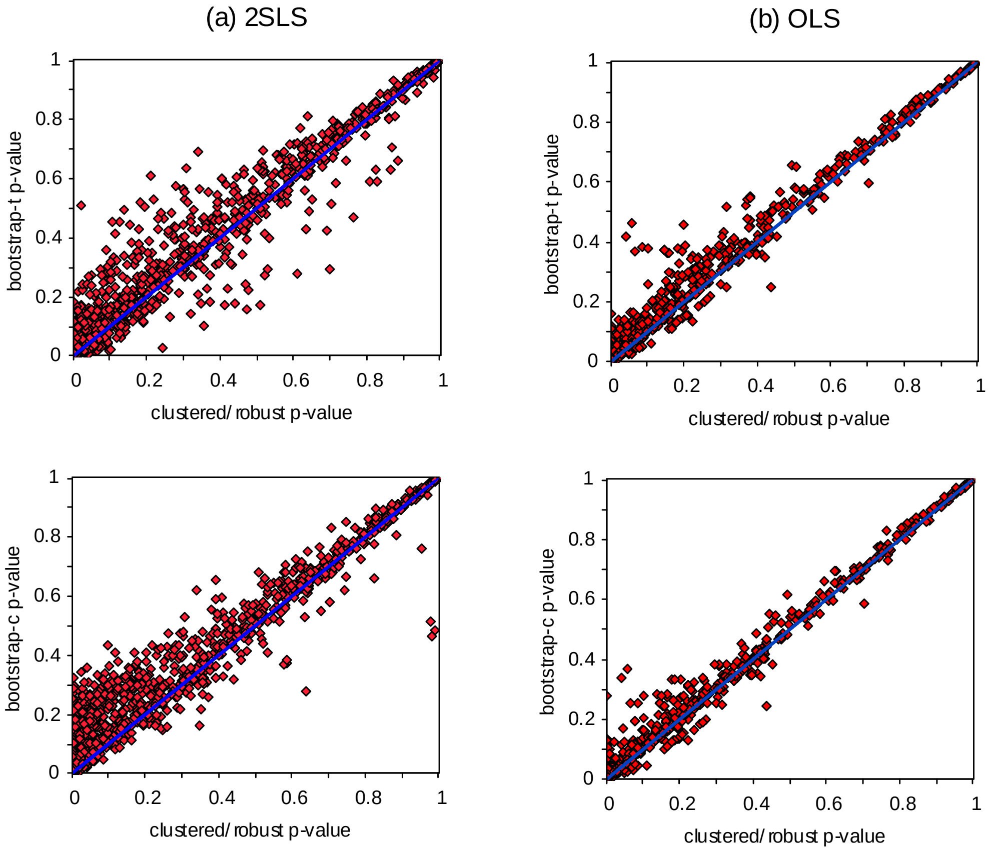
Table III also reports the size of conventional and bootstrap methods estimated by bootstrap sampling the data and testing the null of whether the

---

regressions which use the single-observation robust version the same adjustment actually increases the fraction of significant results from .20 to .23. Although a failure to cluster at the treatment level is not uncommon in randomized experiments (see Young 2017), I find no such cases in my IV sample. Every regression in which instrumented treatment is applied to groups of observations clusters at that or a higher level of aggregation.

[9] I recognize that in a frequentist world, a p-value of .011 is no more significant at the .01 level than a p-value of .11, so all that matters is the frequency of 0/1 significance reported in Table III, not the magnitude of the changes in p-values. However, based upon the comments of seminar participants, most economists appear to operate in a quasi-Bayesian world in which the actual p-value matters (as it affects the posterior probability of the null). I should note that coefficients found to be significant using the bootstrap but not so using conventional methods are very rare. For example, of the 327 and 232 coefficients found to be significant at the .01 level using the bootstrap-t and bootstrap-c, respectively, only 12 and 2 (in turn) are not significant using authors' methods.

Figure I



estimated coefficients equal their known population moments. As shown, conventional 2SLS methods have substantial coverage bias, with clustered/robust methods rejecting the null when true .046 of the time at the .01 level and .100 of the time at the .05 level, while estimates using the default covariance matrix do even worse with, for example, an average rejection rate of .095 at the .01 level. The bootstraps, however, also have empirical size greater than nominal value, with the bootstrap-t rejecting the true null .028 and .071 of the time at the .01 and .05 levels, respectively, and the bootstrap-c doing better, with average rejection rates of .019 and .054 at the two levels. The weaker performance of the bootstrap-t reflects the inaccuracy of the conventional 2SLS variance estimates where, I find, the ln clustered/robust standard error has only a .856 correlation with the ln of the bootstrapped estimate of the coefficient standard error. In contrast, the comparable correlation in the case of OLS is .996, i.e. OLS standard error estimates convey more information, which is why the size of the bootstrap-t is at least as accurate, and often better, than that of the bootstrap-c in the OLS settings examined further on. Regardless, the implication of the excess size of both bootstrap methods, here and later, is that the significance rates reported using bootstrap techniques are likely to be generous.

Table III also compares the results of 2SLS methods with OLS versions of the same equations. While the estimated size of conventional and bootstrap methods in 2SLS and OLS is comparable, conventional OLS results are much more robust to the introduction of the bootstrap, with the number of .01 and .05 significant results falling by only about 15 and 10 percent, respectively. Significant OLS results have, to begin with, lower p-values with, for example, an average p-value of .0008 among .01 significant results as compared to the .0022 achieved by similar 2SLS results. Moreover, those p-value changes which do occur are much more dramatic in the case of 2SLS, as shown in Figure I above. For example, the 5th and 95th percentiles of the difference between the 2SLS bootstrap-t and clustered/robust conventional p-values are -.046 and .154, respectively, while the same percentiles for the difference between the OLS bootstrap-t and conventional p-values are -.008 and .086.

Table IV highlights the extraordinary uncertainty surrounding 2SLS estimates. As shown, the conventional clustered/robust .99 2SLS confidence interval contains the OLS point estimate .866 of the time and the entirety of the OLS confidence interval .675 of the time. Bootstrapped confidence intervals, however, are much wider. In the case of the bootstrap-t, the .99 two-sided confidence interval, arrived at by multiplying the conventional point estimate of the standard error by the bootstrapped estimate of the tail values of the absolute

Table IV: Confidence Intervals and Critical Values (1533 coefficients)

	confidence intervals (CI) and point estimates (β)							
	$\beta_{ols} \in CI_{2sls}$		$CI_{ols} \subset CI_{2sls}$		$\beta_{2sls} \in CI_{ols}$		$CI_{2sls} \subset CI_{ols}$	
	.99	.95	.99	.95	.99	.95	.99	.95
clustered/robust	.866	.723	.675	.542	.325	.267	.003	.003
bootstrap - t	.922	.803	.747	.618	.379	.307	.004	.005
bootstrap - c	.940	.851	.833	.708	.361	.300	.002	.002

	cumulative distribution of t-statistic .01 critical values						
	.01	.10	.25	.50	.75	.90	.99
conventional	2.58	2.58	2.59	2.61	2.68	2.68	2.77
bootstrap-t 2SLS	1.67	2.35	2.66	3.13	3.88	4.84	8.77
bootstrap-t OLS	2.25	2.54	2.68	2.92	3.62	5.22	9.45
bootstrap-c 2SLS	2.19	2.64	2.93	4.52	8.08	30.8	112.8
bootstrap-c OLS	2.41	2.56	2.64	2.82	3.38	5.14	8.20

Notes: .99/.95 = level of the confidence interval. Numbers reported in the top panel are fraction of coefficients meeting the specified criteria. Numbers reported in the bottom panel are the percentiles of the critical values of the absolute value of the t-statistic. For the bootstrap-c a t-statistic equivalent is calculated by dividing the coefficient deviation critical value by the original clustered/robust standard error estimate.

value of the conventional t-statistic, contains the OLS point estimate .922 of the time and the bootstrapped OLS confidence interval .747 of the time. For the bootstrap-c, the two-sided .99 confidence interval, arrived at by calculating the tail values of the absolute value of the coefficient deviations from the parent population moment, contains the OLS point estimate .940 of the time and the entirety of the bootstrapped OLS confidence interval .833 of the time. In contrast, the .99 OLS confidence intervals, whether bootstrapped or conventional, contain the 2SLS point estimate only about ⅓ of the time and the entirety of the 2SLS confidence interval virtually never.

The bottom panel of Table IV reports the cumulative distribution, across the 1533 coefficients, of the two-sided t-statistic .01 critical values of conventional and bootstrap tests. Relative to the conventional distribution, based as it is upon the putative degrees of freedom of the 2SLS and OLS regressions, the bootstrap-t 2SLS and OLS distributions are extraordinarily dispersed, with critical values that are both much smaller and much larger than those assumed by the conventional distribution. Within the bootstrap-t, however, one sees that, with the exception of the extreme ends, bootstrapped 2SLS critical values are not systematically larger than those of bootstrapped OLS. The table also calculates equivalent "t-statistic" critical values for the bootstrap-c by dividing the coefficient deviation critical values by the original sample's conventional estimate of the standard error. For OLS this calculation produces a distribution that is quite similar to that of the bootstrap-t. For 2SLS, however, the bootstrap-c "t-statistic" critical values are systematically larger. Since the difference between the bootstrap-c and bootstrap-t is that the latter divides the sample by sample coefficient estimate (common to both methods) by the sample by sample standard error estimate, this indicates that the standard error estimate is correlated with deviations of coefficients from the population moment. I find that the average correlation between the absolute deviation of the 2SLS coefficient from the population moment and the conventional clustered/robust and default standard error estimates is, extraordinarily, .475 and .525, respectively, while the comparable average correlations for conventional OLS estimates are only .140 and .228 in each case. This is actually a positive feature, in that it limits the frequency with which the extreme coefficient outcomes of 2SLS lead to false conventional rejections, but it also shows that the distributions do not remotely satisfy the assumptions underlying the t-statistic, which is supposedly the ratio of independent random variables. The tails of the actual distribution of 2SLS t-statistics are similar to those of OLS, producing similar size distortions when, as is customary, t-statistics are used to evaluate significance. However, the proportional understatement of conventional confidence intervals is systematically greater in 2SLS (as evidenced by the greater "t-statistic" critical values in the bootstrap-c) indicating a greater downward bias in conventional 2SLS standard error estimates.

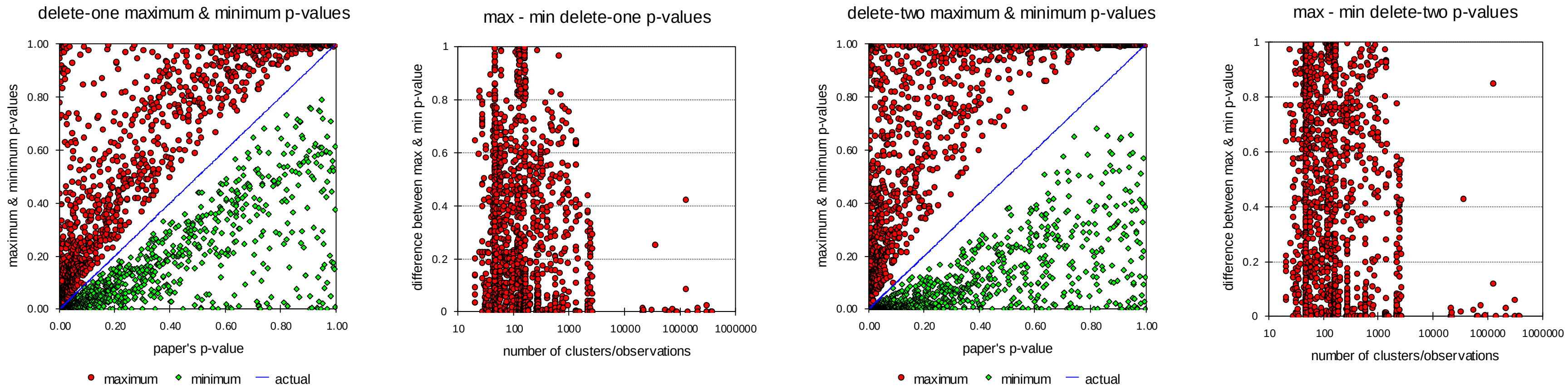
Reported 2SLS results are remarkably dependent upon outliers. Figure II graphs the maximum and minimum coefficient p-values, calculated using authors' methods, found by deleting one cluster or observation in each regression. With the removal of just one cluster or observation, .45 of reported .01 significant 2SLS results can be rendered insignificant at that level, with the average p-value, when such changes occur, rising from .004 to .134. Conversely, .21 of .01 insignificant results can be rendered significant at the same level, with the average p-value falling from .106 to .004. The average gap between the delete-one maximum and minimum p-values is .28, with large differences appearing even in regressions with thousands of clusters or observations. With the deletion of two observations, no less[10] than .63 of .01 significant results can be rendered insignificant (with average p-values rising to .253) and .39 of .01 insignificant results can be made significant, while the average gap between maximum and minimum delete-two p-values is at least .45. In contrast, when OLS versions of the same regressions are examined, insignificant OLS results are found to have a similar sensitivity to outliers, but significant results do not. With the removal of one or two observations, .25 and .39, respectively, of .01 insignificant OLS results can be made significant, but only .15 and .26 of .01 significant OLS results can be made insignificant with the same deletions. In regressions with original p-values greater than .1, the average gap between the maximum and minimum delete-one or -two OLS p-values is .48 and .71, which is similar to the same gaps for 2SLS regressions with original p-values greater than .1 (.50 and .72). In regressions where the original p-value is less than .1, however, the average delete-one/-two

---

[10] "No less" because computation costs prevent me from calculating all possible delete-two combinations. Instead, I delete the cluster/observation with the maximum or minimum delete-one p-value and then calculate the maximum or minimum found by deleting one of the remaining clusters/observations.

Figure II: Sensitivity of P-Values to Outliers (Instrumented Coefficients)

(a) 2SLS



(b) OLS

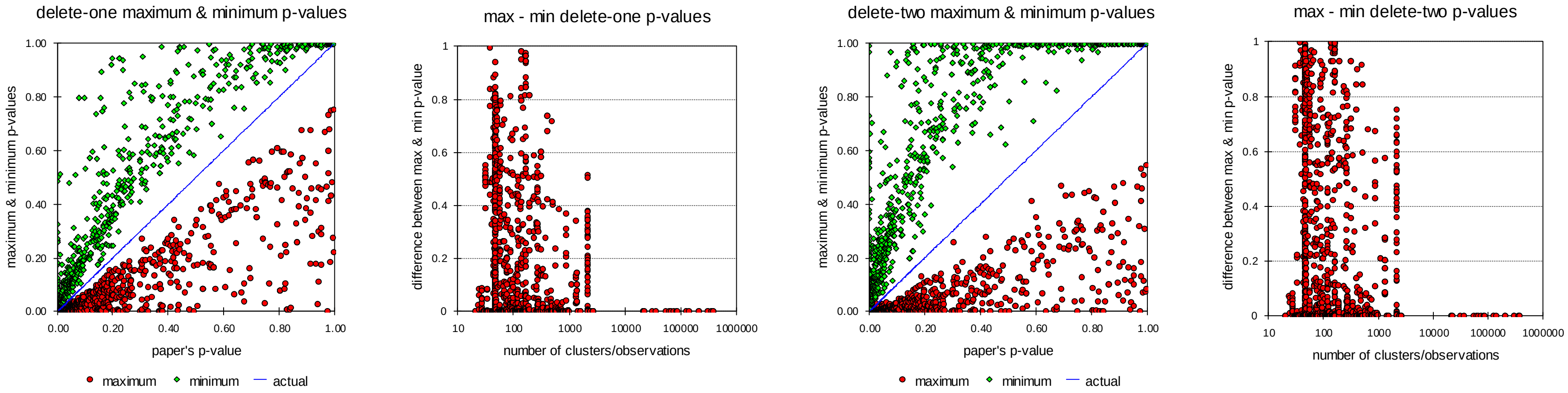


Table V: Testing OLS Bias, Durbin-Wu-Hausman Tests (1400 regressions)

	rejection rates		average size	
	.01	.05	.01	.05
DWH_1	.296	.416	.107	.180
DWH_2	.300	.421	.109	.182
DWH_3	.351	.469	.134	.220
bootstrap – t	.137	.269	.032	.075
bootstrap – c	.084	.190	.021	.061

Notes: Unless otherwise noted, as in Table III above. Conventional test statistics evaluated using the chi^2 distribution. DWH_1-DWH_3 use $\mathbf{V}_1$-$\mathbf{V}_3$, as listed in the text. Bootstrap-t based upon $\mathbf{V}_3$; results using $\mathbf{V}_1$ and $\mathbf{V}_2$ are very similar.

OLS gaps are only .04 and .07, while the same gaps for the corresponding 2SLS regressions are .12 and .25.

The motivation for using 2SLS stems from the fear that the correlation of endogenous regressors with the error term will produce substantially biased and inconsistent estimates of parameters of interest. Table V shows that there is actually limited evidence of this in my sample. I report the Durbin (1954) - Wu (1973) - Hausman (1978) test based upon the Wald statistic formed by the difference between the 2SLS and OLS coefficient estimates. Following Staiger and Stock's (1997) classification, I use three related estimates of the variance of the coefficient difference, namely:

$$(7)\ \ \mathbf{V}_1 = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\sigma}^2_{\text{2sls}} - (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}\hat{\sigma}^2_{\text{ols}};\ \ \mathbf{V}_2 = [(\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1} - (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}]\hat{\sigma}^2_{\text{2sls}};\ \ \mathbf{V}_3 = [(\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1} - (\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}})^{-1}]\hat{\sigma}^2_{\text{ols}}$$

where $\tilde{\mathbf{Y}}$ and $\hat{\tilde{\mathbf{Y}}}$ follow the notation described earlier and $\hat{\sigma}^2_{\text{2sls}}$ and $\hat{\sigma}^2_{\text{ols}}$ denote the 2SLS and OLS estimates of the variance σ^2 of the second-stage disturbances. The different forms of the test arise from the fact that both estimates of σ^2 are consistent under the null. I estimate $\hat{\sigma}^2$ in both cases by dividing the sum of squared residuals by the same finite sample n-$k_{\mathbf{Y}}$-$k_{\mathbf{X}}$ adjustment. This ensures that $\mathbf{V}_1$ is always positive definite and orders the test statistics so that $DWH_1 < DWH_2 < DWH_3$, where DWH_i denotes the Wald statistic calculated with $\mathbf{V}_i$.

As shown in Table V, once again conventional tests show sizeable size distortions, with rejection rates in excess of .10 at the .01 level. Conventional Wald tests reject the null of no bias about ⅓ of the time at the .01 level, but comparable rejection rates for the bootstrap-t and -c are only .137 and .084, respectively, with the bootstrap-c again showing smaller size distortions. Only 13 of the 32 papers in my sample have any regressions in which the zero difference null is rejected by the bootstrap-t at the .01 level, and only 23 have regressions which reject the null at the .05 level. Comparable numbers for the bootstrap-c are 14 and 18 papers, respectively. In the overwhelming majority of regressions reported in published papers, there is actually no compelling evidence that use of OLS methods produces substantively biased estimates. Given the width of 2SLS confidence intervals recorded earlier above, this result is not surprising.

Table VI uses bootstrapped samples from the authors' data sets to estimate the mean squared error (MSE) and bias around values of interest produced by 2SLS and OLS estimation.[11] I consider two scenarios: (1) OLS is inconsistent and the original 2SLS coefficient estimate is the desired population moment; (2) OLS is consistent and either the original OLS or 2SLS coefficient estimate is the desired population moment. As mean squared error and bias varies with units of measurement, I normalize by dividing the estimates for 2SLS by OLS and taking the logarithm, which limits the influence of outliers on the average and ensures that reported results (modulo a sign change) are not sensitive to the choice of denominator. As shown, on average 2SLS's ln mean squared error around its own population moment is 1.52 greater than that of OLS coefficients around the *same* 2SLS population moment, as an average reduction of -1.20 in the 40 percent of cases where 2SLS does better is more than offset by the average 3.27 ln increase in

---

[11] As shown by Kinal (1980), with normal disturbances only the first $k_{\mathbf{Z}} - k_{\mathbf{Y}}$ moments of 2SLS estimates exist. However, the disturbances in my sample are not normal and the dependent variables do not allow for the unbounded outcomes that generate this result. The MSE and bias calculations presented in this paper are for the bounded disturbances actually present in the data.

Table VI: Average Relative Ln Mean Squared Error and Ln Bias
Around 2SLS and OLS Population Moments (2000 bootstrap iterations)

	$\beta_{true} = \beta_{2sls}$				$\beta_{true} = \beta_{2sls}$ or β_{ols}			
	MSE		Bias		MSE		Bias	
	N	mean	N	mean	N	mean	N	mean
all	1533	1.52	1533	-1.75	1533	4.77	1533	2.44
IV < OLS	600	-1.20	1295	-2.24	18	-.259	124	-.796
IV > OLS	933	3.27	238	.941	1515	4.83	1409	2.72

Notes: N = number of coefficients falling into each category; otherwise, numbers reported are the mean ln relative squared deviation around (MSE) or ln absolute deviation from (bias) the 2sls point estimates (left panel) or each method's own point estimate (right panel) of the population moment of interest.

the 60 percent of cases where it does worse. With regards to ln relative bias, 2SLS does better, achieving an average -1.75 ln relative reduction in the absolute deviation from the 2SLS population moment. However, despite the fact that the bias of OLS is motivation for the use of 2SLS methods, in 16 percent of coefficients 2SLS shows a greater bias than OLS around the 2SLS population moment, with an average .941 increase in ln relative bias.

Table V showed that there is generally not much evidence that OLS estimates are, in fact, substantively biased. In consideration of this, the right panel of Table VI calculates the mean squared error and bias of the two methods under the null that OLS is not biased. Since in this case 2SLS remains consistent, albeit inefficient, I give each method the benefit of the doubt and calculate its mean squared error and bias around its own population moment. As shown, in 99 percent of coefficients 2SLS has higher mean squared error, with an average ln ratio of 4.83 in these cases. Similarly, in 92 percent of coefficients 2SLS has a larger bias, with an average ln increase of 2.72. These results highlight the substantial risks involved in using 2SLS and the importance of being very certain that OLS does, in fact, yield intolerably biased estimates.

Table VII: Identification & Strength of the First-Stage (1397 regressions)

	instrument relevance test			
	rejection rates		average size	
	.01	.05	.01	.05
clustered/robust	.905	.951	.160	.217
default	.924	.959	.284	.363
bootstrap - t	.518	.646	.085	.121
bootstrap - c	.704	.891	.096	.135

	first stage F#		
	mean	F > 10	prob (F > 10)
clustered/robust	125	.728	.089
default	474	.790	.213
bootstrap - t	7.4	.338	.063
bootstrap - c	9.0	.397	.071

Notes: .01/.05 = level of the test; (#) for 1359 regressions with one endogenous regressor; F > 10 = fraction of sample with F > 10; prob(F>10) = bootstrapped estimate of probability under the null of zero effects of a first stage F greater than 10.

Table VII asks whether 2SLS equations are even identified by testing the null that all first stage coefficients on the excluded exogenous variables are zero.[12] Using the conventional test with the clustered/robust covariance estimate, .905 of first stage regressions reject the null of a rank zero first stage relation at the .01 level. This share falls to only .518 using the bootstrap-t and .704 using the bootstrap -c. Size distortions are remarkable, with .160 and .284 average rejection rates when the null is true at the .01 level using the clustered/robust and default covariance estimates, respectively. As explored in the on-line appendix, size

---

[12] In the case of the 1359 regressions with one endogenous variable, this is simply the F-test of the significance of the excluded instruments in the first stage regression. In the case of the 41 regressions with more than one endogenous variable, I stack the first stage coefficients and use the covariance matrix for Zellner's (1962) seemingly unrelated regression model with identical regressors (Greene 2012) as the default covariance estimate and White's (1982) sandwich covariance estimator as the clustered/robust covariance estimate, and test joint significance using the chi^2 distribution. I only report statistics for 1397 regressions in the table because for three of the regressions with multiple endogenous variables the total number of coefficients tested is several multiples of the number of observations per equation and the covariance matrix is utterly singular.

distortions increase as more coefficients are tested together and in the 310 equations with more than one excluded instrument examined in the table there are, on average, 25.7 coefficients being simultaneously tested. Size distortions for the bootstrap are also very large, with an average rejection rate at the .01 level of .085 using the bootstrap-t and .096 using the bootstrap-c. Consequently, the bootstrap significance rates reported in the left panel should be considered quite generous.

The strength with which the instrument irrelevance null is rejected, as measured by the size of the first stage F-statistic, is typically used as an indicator of the degree to which the problems of bias and excess size associated with weak (but non-zero) identification are likely to be avoided. The bootstrap indicates, however, that the distribution of these test statistics is much more dispersed than typically recognized, suggesting that they overstate instrument strength. To this end, Table VII calculates an F-equivalent of the bootstrapped p-values by inverting them with the clustered/robust degrees of freedom[13] used to evaluate the paper's conventional F-statistic. While the mean of the conventional F is 474 using the default covariance estimate and 125 using the clustered/robust covariance estimate, it falls to 9.0 and 7.4 when equivalents are calculated using the bootstrap-c and -t, respectively. Based upon conventional methods, about ¾ of regressions have an F greater than 10, which is commonly taken as an indicator of instrument strength, but only about ⅓ to .4 exceed this value when bootstrap p-value equivalents are calculated. The bootstrap distribution of the F-statistics reveals that when the regression is completely unidentified, i.e. all first stage coefficients on excluded exogenous variables are zero, the conventional F statistic with the clustered/robust and default covariance estimates is greater than 10 .089 and .213 of the time, respectively. Thus, allowing that the exclusion restriction holds in the aggregate population, with disturbingly high frequency a report of a strong first

---

[13] As I bootstrap in clusters when the regression clusters.

stage relationship might actually reflect a finite sample correlation of otherwise irrelevant excluded exogenous variables with the residuals, producing biased estimates. Issues associated with the use of the first-stage F statistic as a pre-test are explored more fully in the next section and in the conclusion.

As in the case of 2SLS p-values, the calculated strength of the first stage 2SLS relationship is remarkably dependent upon a few observations, as shown in Figure III below. The first two graphs in each row depict the ratio of the minimum F found by deleting one or two clusters or observations to the actual test statistic in the full sample, while the third and fourth graphs in each row depict the ratio of the actual test statistic to the maximum F found by deleting one or two clusters or observations.[14] The upper and lower panels refer to F-statistics calculated using the default and clustered/robust covariance estimates, respectively. On average, the default F can be reduced to .66 (.51) of its full sample value with the deletion of just one (two) observation(s), while the clustered/robust does slightly better, falling to .72 (.60) of its full sample value. Conversely, the ratio of the actual to delete-one (two) maximum default F averages .84 (.75), while the clustered/robust F is more sensitive in this direction, averaging .65 (.53). The average proportional sensitivity is slightly greater in regressions with original Fs less than ten, but the differences are hardly meaningful.[15] As is to be expected, the largest movements are found in small samples with 100 or less clusters or observations, but proportional changes of .1, .2 or more are disturbingly commonplace in samples with 10s and 100s of thousands of observations, as shown in the figure. These results point to the extraordinary sampling variability of F statistics, explaining the

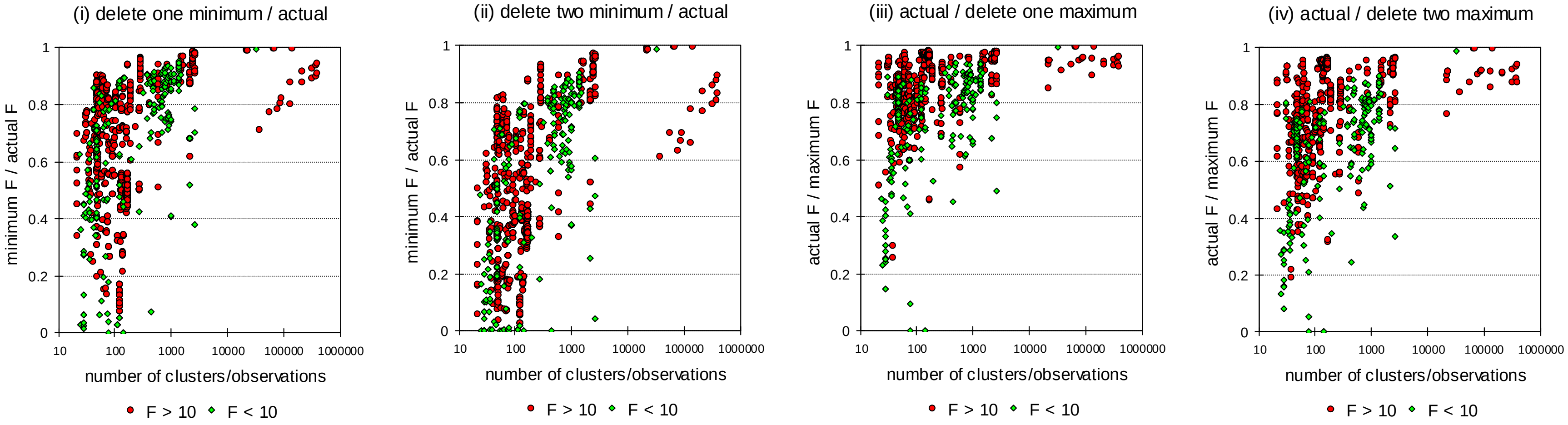
---

[14] The delete-two ratios are upper bounds since, as before, I do not do a full delete-two search but instead simply take the delete-one maximum or minimum and then search across the remaining clusters/observations.

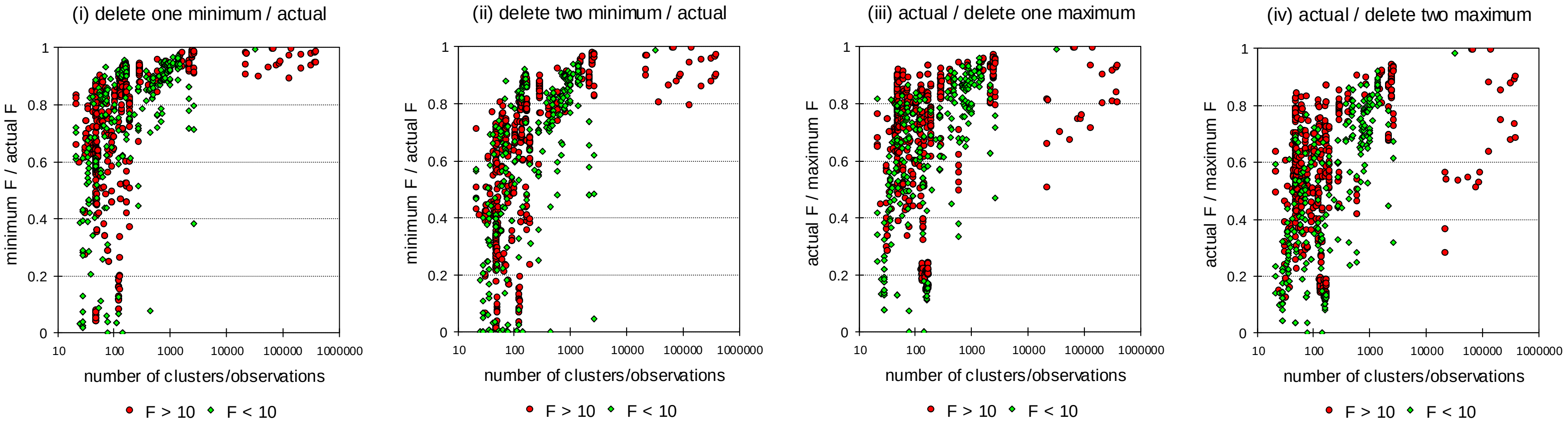
[15] The average ratios for F's greater than 10 (less than 10) moving left to right through the top and then bottom panels are .66 (.64), .52 (.48), .87 (.74), .79 (.63), .72 (.71), .61 (.58), .65 (.64), .and 54 (.50). The only substantial differences are found in the third and fourth panels of the top row, depicting the potential increases in the default F.

Figure III: Proportional Change of First Stage F with Removal of One or Two Clusters or Observations
(1359 regressions with one endogenous regressor)

(a) default covariance estimate



(b) clustered/robust covariance estimate



gross size distortions of conventional methods and the difficulties even the bootstrap finds in providing accurate size, as shown earlier in Table VII.

The sources of delete-one and -two sensitivity are worth exploring. Consider the generic regression on a matrix of regressors **X**. The change in the estimated coefficient for a particular regressor **x** brought about by the deletion of the vector of observations **i** is given by:

$$(8)\ \hat{\beta}_{\sim\mathbf{i}} - \hat{\beta} = -\tilde{\mathbf{x}}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}} / \tilde{\mathbf{x}}'\tilde{\mathbf{x}}$$

where $\tilde{\mathbf{x}}$ is the vector of residuals of **x** projected on the other regressors, $\tilde{\mathbf{x}}_{\mathbf{i}}$ the **i** elements thereof, and $\boldsymbol{\varepsilon}_{\mathbf{i}}$ the vector of residuals for observations **i** calculated using the delete-**i** coefficient estimates. The delete-**i** residuals are related to the estimated residuals through the formula $\boldsymbol{\varepsilon}_{\mathbf{i}} = (\mathbf{I} - \mathbf{H}_{\mathbf{ii}})^{-1}\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}$, where $\mathbf{H}_{\mathbf{ii}}$ denotes the **i** x **i** block of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.[16] The default and clustered/robust covariance estimates are of course given by:

$$(9)\ \text{default:}\ \frac{1}{n-k}\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\tilde{\mathbf{x}}'\tilde{\mathbf{x}}}\ ;\ \text{clustered/robust:}\ \frac{n}{n-k}\frac{\sum_{\mathbf{i}}\tilde{\mathbf{x}}_{\mathbf{i}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\tilde{\mathbf{x}}_{\mathbf{i}}}{(\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^2}$$

Define $\boldsymbol{\varepsilon}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}} / \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$, $\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}} / \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ and $\tilde{\mathbf{x}}_{\mathbf{i}}'\tilde{\mathbf{x}}_{\mathbf{i}} / \tilde{\mathbf{x}}'\tilde{\mathbf{x}}$ as the group **i** shares of squared delete-**i** residuals, squared actual residuals, and coefficient leverage,[17] respectively. Clearly, the standard error estimate and the coefficient estimate relative to the standard error estimate will be more sensitive to the deletion of some observations **i** when these shares are uneven.

Table VIII summarizes the maximum residual and leverage shares found in my sample. In the 1359 first stage regressions with one endogenous variable, the

---

[16] Where there are regressors that are non-zero only in **i**, as in the case of cluster fixed effects, so that **I** - $\mathbf{H}_{\mathbf{ii}}$ is singular, one applies the formula by calculating **H** using the residuals of the projection of the other regressors on these measures (i.e. the partitioned regression version of **H**).

[17] Since "leverage" is typically defined as the diagonal elements of the matrix **H** formed using all regressors, while the measure described above is the equivalent for the partitioned regression on $\tilde{\mathbf{x}}$.

Table VIII: Largest Shares of Squared Residuals & Coefficient Leverage
(1359 regressions with one endogenous variable)

	2SLS first-stage			2SLS second-stage			OLS version		
	one cl/obs	two cl/obs	.05 cl/obs	one cl/obs	two cl/obs	.05 cl/obs	one cl/obs	two cl/obs	.05 cl/obs
$\tilde{\mathbf{x}}_{\mathbf{i}}'\tilde{\mathbf{x}}_{\mathbf{i}} / \tilde{\mathbf{x}}'\tilde{\mathbf{x}}$	.172	.288	.563	.163	.273	.526	.213	.286	.457
$\boldsymbol{\varepsilon}_{\mathbf{i}}'\boldsymbol{\varepsilon}_{\mathbf{i}} / \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$	.155	.244	.440	.162	.242	.437	.160	.239	.434
$\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\hat{\boldsymbol{\varepsilon}}_{\mathbf{i}} / \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$	.135	.226	.424	.148	.223	.412	.152	.229	.422

Note: cl/obs = clusters or observations, depending upon whether the regression is clustered or not. .05 = largest 5 percentiles.

largest one or two clusters or observations account, on average, for .172 and .288 of total leverage, respectively, while the largest delete-**i** (estimated) residual shares are .155 and .244 (.135 and .226). The largest 5 percentiles account for around ½ of total leverage and squared residuals. To put these numbers in perspective, in my study of experimental papers (Young 2017) I find the largest (5th percentile) leverage and residual observation shares of OLS regressions average .05 (.27) and .09 (.34), respectively. With massive outliers, in both residuals and regressors, estimated first stage F statistics are largely determined by a handful of observations and have a volatility and distribution consistent with that fact. Table VIII also reports the concentration of leverage and residuals in the second-stage regression and in its OLS version. For the second-stage regression, I treat the instrumented values of the endogenous variables as unaffected by deletions, and calculate the delete-**i** residuals accordingly, to provide an indication of the volatility of coefficient estimates if the first-stage relation were unchanging.[18] As can be seen, the concentration of leverage and residuals in second stage relations is roughly equal to that found in OLS versions of the regressions (using the uninstrumented endogenous variables), suggesting that they would, on average,

---

[18] The reported estimated residual shares are the shares of the actual 2SLS residuals used in the 2SLS covariance estimate.

Table IX: Consistency without Inference: 2SLS in Practical Application
(1533 coefficients in 1400 2SLS regressions)

	bootstrap-t		bootstrap-c	
	.01	.05	.01	.05
coefficient significant using authors' methods	.322	.502	.322	.502
$\boldsymbol{\beta}_{ols}= \boldsymbol{\beta}_{2sls} < \alpha$ & $\boldsymbol{\Pi} = \mathbf{0} < \alpha$	.101	.189	.083	.176
$\boldsymbol{\beta}_{ols}= \boldsymbol{\beta}_{2sls} < \alpha$, $\boldsymbol{\Pi} = \mathbf{0} < \alpha$, & $CI_{ols} \not\subset CI_{2sls}$	.080	.170	.076	.169
$\boldsymbol{\beta}_{ols}= \boldsymbol{\beta}_{2sls} < \alpha$, $\boldsymbol{\Pi} = \mathbf{0} < \alpha$, & $\beta_{ols} \notin CI_{2sls}$	.048	.132	.056	.136
$\boldsymbol{\beta}_{ols}= \boldsymbol{\beta}_{2sls} < \alpha$, $\boldsymbol{\Pi} = \mathbf{0} < \alpha$, & $\beta_{2sls} = 0 < \alpha$	.059	.136	.057	.142

Notes: Numbers reported are fraction of coefficients meeting the specified criteria. .01/.05 = level of the test (α) and complementary confidence interval. $\boldsymbol{\beta}_{ols}= \boldsymbol{\beta}_{2sls} < \alpha$ = bootstrapped p-value of Durbin-Wu-Hausman test of zero OLS bias less than α; $\boldsymbol{\Pi} = \mathbf{0} < \alpha$ = bootstrapped p-value of instrument irrelevance test less than α; $CI_{ols} \not\subset CI_{2sls}$ or $\beta_{ols} \notin CI_{2sls}$ = bootstrapped OLS confidence interval or OLS point estimate not included in bootstrapped 2SLS confidence interval; $\beta_{2sls} = 0 < \alpha$ = bootstrapped p-value of 2SLS coefficient less than α.

have a similar delete-**i** sensitivity. Unfortunately, the first-stage relation is most certainly not unchanging, and is in fact dependent upon a small set of observations. This imparts an additional, extraordinary, degree of volatility and sensitivity to 2SLS estimates.

Table IX brings the preceding results together. As noted in the top line, using authors' methods, ⅓ and ½ of reported coefficients are significant at the .01 and .05 levels, respectively, leading the reader to conclude that 2SLS methods have revealed something about the world. In the lower lines I consider alternative criteria for evaluating published results. A good starting point seems to be to require that the Durbin-Wu-Hausman test indicate that there is a statistically significant OLS bias, as the relative mean squared error and bias of 2SLS when OLS is unbiased is simply too much to bear, and, moreover, that one can reject the null hypothesis that the model is utterly unidentified with all of the first stage coefficients equal to 0, as in this case "identification" is achieved through an undesirable finite sample correlation between the instruments and the error term. Only .101 and .189 of estimated coefficients are in regressions which meet these criteria at the .01 and .05 levels using the bootstrap-t, while only .083 and .176 of estimated coefficients meet these criteria using the bootstrap-c. I should note that imposing these preliminary requirements at the .01 level largely ensures that 2SLS has a lower mean squared error and bias around the measured 2SLS population moment than OLS, as imposing these additional requirements would only lower the fraction of acceptable coefficients by an additional .003 or .004.

With these basic prerequisites for credibility in place, one might then ask whether 2SLS estimates rule out the OLS results, i.e. accepting that, taking into full account their covariance, the OLS and 2SLS population moments are different, one might still want to know if the OLS estimates are unlikely to be true. The weak form of this demand might be that the 2SLS confidence interval does not encompass the entirety of the OLS confidence interval, while the strong form might be that it does not contain the actual OLS point estimate. At the .01 and .05 levels, only about .080 and .170 of 2SLS results, using either bootstrap measure, meet the weak criterion while satisfying the OLS bias and identification prerequisites. The two bootstrap measures are also in fairly close agreement with regards to the strong criterion, with .048 and .132 of coefficients, at the two significance levels, meeting it using the bootstrap-t and .056 and .136 using the bootstrap-c. Putting aside comparison with OLS, an alternative approach, following the DWH and identification pre-tests, is to ask whether the 2SLS bootstrap p-value rejects the null of zero effects, suggesting that, aside from finding that OLS is biased, we have uncovered a meaningful causal relationship. Here again the two bootstrap measures are in close agreement, with just under .06 and around .140 of coefficients meeting this condition at the .01 and .05 levels, respectively. In sum, while IV estimates may be consistent, in finite samples they allow for little inference. Using either bootstrap measure in only about .05 or .06 of cases are 2SLS coefficient estimates both strongly credible and significantly different from either the OLS point estimate or zero. These results are generous as they are based on bootstrapped tests with positive size biases, particularly in the case of testing instrument relevance.

V. Weak Instruments and Weak Instrument Pre-Tests

A weak first stage relation between the excluded exogenous variables **Z** and the endogenous second stage regressors **Y** is known to create many of the 2SLS ailments noted in the previous section, namely large estimated standard errors relative to OLS, even greater tail variation than estimated (producing size larger than nominal value), and biased point estimates. As shown by Rothenberg (1984),[19] the usual root-N convergence to a distribution can, in the case of 2SLS and iid normal errors, be thought of as a function of the square root of the first stage concentration parameter μ^2 which, in the case of a single endogenous regressor equals $\mathbf{\Pi}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\mathbf{\Pi}/\sigma_v^2$, where, as before, **Π** denotes the first stage coefficients on the excluded instruments **Z** and σ_v^2 is the residual variance of the first-stage equation. The concentration parameter can be thought of as effective sample size, and as it goes to infinity the distribution of the 2SLS estimator converges to the normal distribution with variance equal to the default 2SLS estimate of variance, while the bias of the 2SLS estimator relative to OLS goes to zero. Moreover, for a given sample size, as the concentration parameter increases the efficiency of 2SLS relative to OLS improves, as the variation of predicted values $\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}}$ rises relative to that of the OLS regressors $\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$. With weak instruments, however, the predicted 2SLS variation is small relative to OLS, the distribution of coefficients is grossly non-normal with potentially fat tails, and point estimates are biased in the direction of OLS or, even worse, possibly biased more than OLS if there is any correlation between **Z** and the second stage errors. The sample counterpart of the concentration parameter is the Wald test statistic on the excluded instruments in the first-stage regression, namely $\hat{\mathbf{\Pi}}'\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\hat{\mathbf{\Pi}}/\hat{\sigma}_v^2$. The

---

[19] See also the helpful exposition in Stock, Wright and Yogo (2002).

first stage F-statistic equals this value divided by k_Z and authors increasingly report this to convince readers of the reliability of their results. In this section I show that conventional F-statistics and tests based upon these statistics are a poor predictor of the coverage and coefficient biases found in published results.

Tables X and XI evaluate instrument strength using Stock and Yogo's (2002) weak instrument tests. Staiger and Stock (1997) derived the asymptotic distribution of the 2SLS estimator under the assumption of iid errors and a concentration parameter that asymptotically does not grow with sample size. On the basis of this, Stock and Yogo (2002) developed a remarkable set of weak instrument tests, deriving the critical values of the first stage F-statistic large enough to reject the null that the instruments are sufficiently weak so as to generate a proportional bias relative to OLS greater than some level "b" or a size greater than some level "r" above the nominal level α. In the tables I divide regressions based upon whether or not they reject the weak instrument null (H_0) in favour of the strong instrument alternative (H_1) and report the fraction which, based on bootstrap draws from the paper's data, have size or bias greater than the indicated bound. I also report the maximum fraction of H_1 observations violating the bounds that would be consistent with the Stock & Yogo test having its theoretical nominal size of no greater than .05.[20] With critical values depending upon the number of instruments and endogenous regressors, Stock and Yogo

---

[20] Let N_0 and N_1 denote the known number of regressions classified under H_0 and H_1, respectively, and W_0, W_1, S_0 and S_1 the unknown number of regressions with weak and strong instruments classified under each group, with $W_1 = \alpha(W_0+W_1)$ and $S_1 = p(S_0+S_1)$, where α and p denote size and power and I assume $p > \alpha$. Moreover, let $N_1 > \alpha(N_1+N_0)$, which holds for all cases with $N_1 > 0$ presented below. Solving for W_1/N_1, one finds that it is maximized when $p = 1$ and $\alpha = .05$, with $W_1/N_1 = (1/19)(N_0/N_1)$. The reason why W_1/N_1 is maximized when $p = 1$, i.e. is paradoxically *increasing* in power, is because, holding constant the observed N_1 and N_0 and unknown size, lower power means a greater fraction of the total sample must be strong which in turn means that there are fewer W_1 observations. I should note that given the Stock and Yogo theory, the share of regressions in N_1 with size greater than "r" should actually be less than $(1/19)(N_0/N_1)$ as even weak instruments, depending upon the correlation between the first and second stage error terms, need not have coverage greater than "r".

Table X: Fraction of Regressions with Size Greater than “r” in Specifications that Don’t (H_0) and Do (H_1) Reject the Stock & Yogo Weak Instrument Null

	maximum acceptable size (“r”) for a nominal .05 test											
	.10			.15			.20			.25		
	H_0		H_1	H_0		H_1	H_0		H_1	H_0		H_1
	F < 1	F > 1		F < 1	F > 1		F < 1	F > 1		F < 1	F > 1	
(A) default F used as Stock and Yogo test statistic												
N (number)	17	365	944	17	200	1109	17	152	1157	17	135	1174
default cov	.824	.329	.593	.765	.260	.436	.647	.158	.367	.412	.096	.279
cl/robust cov	.882	.334	.332	.882	.190	.150	.588	.072	.087	.471	.052	.050
maximum			.022			.011			.008			.007
(B) clustered/robust F used as Stock and Yogo test statistic												
N (number)	23	748	555	23	268	1035	23	185	1118	23	161	1142
cl/robust cov	.826	.392	.249	.739	.123	.163	.435	.070	.089	.348	.043	.052
maximum			.074			.015			.010			.009
(C) bootstrap-t equivalent F used as Stock and Yogo test statistic												
N (number)	263	960	103	263	566	497	263	467	596	263	376	687
cl/robust cov	.673	.254	.282	.433	.102	.095	.278	.051	.042	.205	.032	.011
maximum			.625			.088			.064			.049

Notes: N = number of regressions in each category; default and cl/robust cov = using these covariance matrices to calculate t-statistics, the share of regressions with conventional size greater than “r”; maximum = maximum share of the sample that rejects H_0 in favour of H_1 with size greater than “r” consistent with the test having size .05 (see text and accompanying footnote).

provide bias critical values for only 179 of the regressions in my sample, but in the case of size their table of critical values covers 1326 of the 1359 regressions with one endogenous regressor.[21]

---

[21] In the case of multiple endogenous variables, the test statistic is based upon the minimum eigenvalue of the Cragg-Donald (1993) underidentification statistic (the sum of whose eigenvalues quite intuitively equals the default covariance estimate based test statistic used to test the rank zero null for equations with multiple endogenous regressors earlier in Table VII). In numerical analysis of examples, Stock and Yogo find that bias and size are non-increasing in all eigenvalues of the Cragg-Donald statistic, and hence consider the minimum eigenvalue as a conservative worst case scenario. As they provide critical values for less than half of the 41 regressions in my sample with multiple endogenous variables, and as these critical values are based upon a conjecture and not a

Table X begins by using the default covariance estimate to evaluate both the F-statistic and coefficient significance, as this is the measure consistent with Stock and Yogo's iid-based theory. As shown in panel (A), this produces disastrous results. Excluding regressions with an exceptionally weak F less than 1, the fraction of regressions with size greater than "r" at the .05 level is actually always substantially greater in regressions which reject the weak instrument null H_0 in favour of the alternative of strong instruments H_1. Using the clustered/robust covariance estimate to evaluate the significance of regressions, whether with the default F (panel A) or the clustered/robust F (panel B), one finds that the fraction of regressions with rejection rates greater than the maximum desired size "r" is neither systematically higher nor lower in H_1 regressions than it is in H_0 regressions with an F greater than 1.[22] Moreover, the share of regressions with size greater than "r" is grossly inconsistent with the maximum that should arise given the test's putative .05 nominal size. Things go somewhat better in the bias test (Table XI). Regardless of whether one uses the default or clustered/robust F, the fraction of regressions with relative bias greater than "b" falls systematically as one moves from regressions with F less than 1, to those with F greater than 1 that don't reject the weak instrument null H_0, to regressions that reject H_0 in favour of the strong instrument alternative H_1. However, the fraction with a bias greater than "b" in the strong instrument set H_1 is again much too high and inconsistent with the Stock and Yogo test having a nominal size of .05. In contrast, use of the bootstrap-t equivalent F statistic, in the bottom panel of each table, produces size and bias in H_1 regressions that is mostly consistent with the test having a nominal size of .05. Calculation of the bootstrap equivalent F, however, is about as costly

---

result, I keep the analysis as simple as possible by focusing on the simple case of a single endogenous regressor.

[22] The substantial gap between rejection rates for H_0 and H_1 for r = .1 in panel (B), which disappears for r = .15, is due to a large mass with high rejection rates just below the r = .1 cutoff point.

Table XI: Fraction of Regressions with Relative Bias Greater than “b” in Specifications that Don’t (H_0) and Do (H_1) Reject the Weak Instrument Null

	maximum acceptable relative bias “b”											
	.05			.10			.20			.30		
	H_0		H_1	H_0		H_1	H_0		H_1	H_0		H_1
	F < 1	F > 1		F < 1	F > 1		F < 1	F > 1		F < 1	F > 1	
(A) default F used as Stock and Yogo test statistic												
N (number)	2	100	77	2	98	79	2	93	84	2	83	94
share	1.00	.950	.312	1.00	.867	.253	1.00	.667	.202	1.00	.518	.106
maximum			.070			.067			.060			.048
(B) clustered/robust F used as Stock and Yogo test statistic												
N (number)	4	104	71	4	99	76	4	89	86	4	76	99
share	1.00	.913	.310	1.00	.848	.250	1.00	.685	.186	.750	.539	.111
maximum			.080			.071			.057			.043
(C) bootstrap-t equivalent F used as Stock and Yogo test statistic												
N (number)	8	171	0	8	171	0	8	144	27	8	119	52
share	1.00	.661	---	1.00	.579	---	.875	.514	.000	.750	.412	.000
maximum			---			---			.296			.129

Notes: share = share of regressions in each category with bias > “b”; otherwise, as in Table X.

as simply bootstrapping the size and bias of the 2SLS regression.

In the on-line appendix I provide regressions that show that, once Fs less than one are removed from the sample, no F statistic, of any form, is significantly or *even negatively* correlated with size. Once Fs less than one are removed, conventional Fs are not significantly correlated with bias or mean squared error either, although the point estimate of the relationship is at least negative. In the non-iid world of published results, conventional F’s are a very poor measure of the strength of the first-stage relation, as they are far more dispersed than indicated by their putative distribution (Table VII earlier), while size distortions do not differ substantially between 2SLS and OLS frameworks (Table III earlier and XII below), indicating that the finite sample problems of inference based upon clustered/robust covariance matrices, rather than the strength of first stage relations, are the dominant problem. Outside of noting that the F is greater than the absurdly low value of 1, little information (and most certainly no protective bound) is gained in reporting the conventional strength of the first stage relation.

VI: Weak Instrument Robust Inference

The finding that the strength of the first stage relationship may be substantially weaker than indicated by conventional F-statistics and that F-test based pre-tests are largely uninformative might lead practitioners to use well-known "weak-instrument robust" alternatives to 2SLS. Unfortunately, the professional understanding of such alternatives is based upon theory and simulations with iid disturbances. In this section I review three such methods, the Anderson-Rubin (1949) approach, the limited information maximum likelihood (LIML) method, and Fuller's (1977) k modification of LIML, showing that in a world with non-iid disturbances these methods are often substantially inferior to conventional 2SLS.

The Anderson-Rubin significance test makes use of the reduced form for the second stage endogenous variable **y**. Specifically, substituting for the endogenous regressors **Y**, we have:

$$\text{(10)}\quad \mathbf{y} = \mathbf{Y}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\delta} + \mathbf{u} = (\mathbf{Z}\boldsymbol{\Pi} + \mathbf{X}\boldsymbol{\Delta} + \mathbf{v})\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\delta} + \mathbf{u}$$
$$\rightarrow \quad \mathbf{y} = \mathbf{Z}\boldsymbol{\Pi}\boldsymbol{\beta} + \mathbf{X}(\boldsymbol{\Delta}\boldsymbol{\beta} + \boldsymbol{\delta}) + (\mathbf{v}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{Z}\boldsymbol{\beta}_{\mathbf{Z}} + \mathbf{X}\boldsymbol{\beta}_{\mathbf{X}} + \boldsymbol{\varepsilon}$$

If $\boldsymbol{\beta} = \mathbf{0}$ then $\boldsymbol{\beta}_{\mathbf{Z}} = \boldsymbol{\Pi}\boldsymbol{\beta} = \mathbf{0}$, so by running the OLS regression of **y** on the excluded and included exogenous variables and testing the joint significance of the excluded exogenous variables, one can test the null that the coefficients on the instrumented variables are zero.[23] The test is robust to weak instruments as $\boldsymbol{\beta}_{\mathbf{Z}} = \mathbf{0}$ under the

---

[23] One can also test non-zero values $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ by running the equation $\mathbf{y} - \mathbf{Y}\boldsymbol{\beta}_0 = \mathbf{Z}\boldsymbol{\beta}_{\mathbf{Z}} + \mathbf{X}\boldsymbol{\beta}_{\mathbf{X}} + \boldsymbol{\varepsilon}$, as under these circumstances $\boldsymbol{\beta}_{\mathbf{Z}} = \boldsymbol{\Pi}(\boldsymbol{\beta}\text{-}\boldsymbol{\beta}_0)$.

null whatever the value of $\mathbf{\Pi}$ and the strength of the correlation between **Y** and **Z**. As it is an OLS equation, it does not suffer from any of the extra variation brought on by 2SLS estimation with a weak concentration parameter, allowing for more accurate inference. For these reasons, it has been recommended as a solution to the problem of weak instruments by Dufour 2003, Baum, Schaffer and Stillman 2007, and Chernozhukov and Hansen 2008, among others. Its recognized weaknesses include the fact that it does not allow for the testing of individual components of $\boldsymbol{\beta}$ and may have very low power in over-identified equations where the dimensionality of **Z** is much greater than that of **Y**, particularly if some of the excluded instruments are irrelevant (i.e. have first stage coefficients near zero).[24]

Table XII below uses the bootstrap to calculate the size distortions of the Anderson-Rubin approach and compare these to those found using conventional 2SLS, in both cases using the clustered/robust covariance estimate to calculate p-values. As shown, the Anderson-Rubin method actually performs worse than 2SLS. Empirical size is on average slightly larger than 2SLS in exactly identified equations, but much greater in over-identified equations, which have a startling average rejection probability of .285 at the .01 level. Dividing the sample by the default, clustered/robust and bootstrap-t equivalent 2SLS first stage F, I find that only in the case of exactly identified equations with the very weakest of instruments, i.e. with a conventional F less than 1, does the Anderson-Rubin approach provide any improvements over 2SLS. Elsewhere, its performance is systematically worse, particularly in the case of over-identified equations.

The results of Table XII once again show that the inaccuracy of inference in the presence of non-ideal errors is the central problem in both OLS *and* 2SLS, overwhelming and dominating any issues associated with weak instruments in the

---

[24] It is also sensitive to the exclusion restriction, since if **Z** affects **y** other than through **Y**, $\boldsymbol{\beta}_{\mathbf{Z}}$ will be non-zero even when $\boldsymbol{\beta} = \mathbf{0}$. However, I accept, here and throughout, the basic premise that the exclusion restriction applies, as otherwise the entire 2SLS endeavour is ill-conceived.

Table XII: Size Distortions with Anderson-Rubin Weak Instrument Robust Inference

	exactly identified equations					overidentified equations				
		A-Rubin		2SLS			A-Rubin		2SLS	
	N	.01	.05	.01	.05	N	.01	.05	.01	.05
all	1100	.053	.115	.046	.097	297	.285	.373	.055	.123
$F_d < 1$	15	.038	.101	.293	.354	2	.353	.483	.248	.341
$1 < F_d < 10$	133	.048	.116	.037	.082	136	.336	.451	.057	.134
$F_d > 10$	952	.054	.116	.044	.095	159	.240	.305	.051	.112
$F_{cl/r} < 1$	19	.037	.101	.246	.306	4	.235	.362	.138	.222
$1 < F_{cl/r} < 10$	209	.047	.109	.034	.074	138	.331	.445	.056	.132
$F_{cl/r} > 10$	872	.054	.117	.045	.097	155	.244	.309	.052	.113
$F_b < 1$	255	.108	.187	.100	.171	15	.624	.703	.106	.198
$1 < F_b < 10$	379	.039	.099	.028	.066	250	.239	.329	.044	.107
$F_b > 10$	466	.034	.089	.032	.081	32	.482	.558	.120	.213

Notes: N = number of equations in each group; otherwise numbers reported are average size using clustered/robust covariance estimates at the .01 or .05 levels. F_d, $F_{cl/r}$ & F_b = default, clustered/robust and bootstrap-t equivalent 1st stage F statistics.

latter. Generally, needlessly increasing the dimensionality of a test tends to lower its power. In the case of clustered/robust covariance estimates, however, increasing the dimensionality of a test appears to produce the opposite problem, raising the probability of rejecting the null when true. I confirm this in the on-line appendix, where I show that size distortions in joint tests using clustered/robust covariance estimates are systematically increasing in the number of individual coefficient components. As also shown in the appendix, however, while estimates using the default covariance estimate do not appear to have this property, they have greater size distortions overall (as already seen in Table III earlier) and find no advantage in the Anderson-Rubin method outside of conventional first stage Fs less than 1. In a world with correlated and heteroskedastic errors, use of default covariance estimates is fraught with peril, as is use of clustered/robust methods in high-dimensional tests. Clustered/robust covariance estimates do better in testing only one coefficient at a time, as is the case for almost all the exactly identified tests in Table XII, but even here size distortions are large enough that, outside of the very weakest cases, they dominate considerations of instrument strength.

The LIML and Fuller-k estimators are members of the k-class family of estimators (Theil 1953) which form the estimate of the coefficient vector using:

$$(11)\quad \hat{\boldsymbol{\beta}} \;=\; (\tilde{\mathbf{Y}}'(\mathbf{I}-\kappa\mathbf{M}_{\tilde{\mathbf{Z}}})\tilde{\mathbf{Y}})^{-1}\tilde{\mathbf{Y}}'(\mathbf{I}-\kappa\mathbf{M}_{\tilde{\mathbf{Z}}})\mathbf{y}$$

where, as before, ˜ denotes the residuals from the projection on the included instruments **X** and $\mathbf{M}_{\tilde{\mathbf{Z}}} = \mathbf{I} - \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})\tilde{\mathbf{Z}}'$ is the residual-maker from the projection on the excluded instruments $\tilde{\mathbf{Z}}$. OLS and 2SLS correspond to κ equal to 0 and 1, respectively. The LIML estimator sets κ_{LIML} equal to the smallest eigenvalue of $(\tilde{\mathbf{Y}}_1'\mathbf{M}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Y}}_1)^{-½}(\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Y}}_1)(\tilde{\mathbf{Y}}_1'\mathbf{M}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Y}}_1)^{-½}$, where $\tilde{\mathbf{Y}}_1 = [\tilde{\mathbf{y}}, \tilde{\mathbf{Y}}]$,while Fuller's k sets κ_{Fuller} equal to κ_{LIML} – c/N-k_Y-k_X , where c is a pre-determined constant and N-k_Y-k_X is the number of observations minus the number of second stage endogenous and exogenous regressors. It is easily seen that $\kappa_{LIML} \geq 1$ and must equal 1 when the equation is exactly identified,[25] in which case the LIML estimator is the same as 2SLS. Early analysis based upon normal disturbances showed that LIML has less median bias and converges to the normal distribution faster than 2SLS (Anderson, Kunitomo & Sawa 1982, Anderson 1983). In Monte Carlo simulations, Staiger and Stock (1997) found that LIML has much more accurate size than 2SLS, a result later confirmed by Stock and Yogo's (2005) weak instrument asymptotics that concluded that "LIML is far superior to 2SLS when the researcher has weak instruments" with "coverage rates that are quite close to their nominal rates." The LIML estimate, however, is so dispersed that with normal disturbances it has no

---

[25] Provided that the matrices are non-singular, the minimum eigenvalue, by the properties of the Rayleigh quotient, equals the minimum across all **z** ($k_{\mathbf{Y}}$+1 x 1) such that $\mathbf{z}'\mathbf{z}$ = 1 of :

$$\mathbf{z}'\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Y}}_1\mathbf{z} / \mathbf{z}'\tilde{\mathbf{Y}}_1'\mathbf{M}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Y}}_1\mathbf{z} = \mathbf{z}'\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Y}}_1\mathbf{z} / (\mathbf{z}'\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Y}}_1\mathbf{z} - \mathbf{z}'\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})\tilde{\mathbf{Z}}'\tilde{\mathbf{Y}}_1\mathbf{z}) \geq 1$$

When the equation is exactly identified, $\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Z}}$ is $k_{\mathbf{Y}}$+1 x $k_{\mathbf{Y}}$ (i.e. of rank $k_{\mathbf{Y}}$), so there exists a $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{z}'\tilde{\mathbf{Y}}_1'\tilde{\mathbf{Z}} = \mathbf{0}'$, ensuring that the minimum is 1.

Table XIII: Coefficient Size Distortions with LIML and Fuller-k Inference

	overidentified equations					all equations				
		LIML		2SLS			Fuller-k		2SLS	
	N	.01	.05	.01	.05	N	.01	.05	.01	.05
all	385	.065	.109	.049	.117	1524	.047	.097	.046	.100
$F_d < 1$	2	.032	.058	.248	.341	17	.122	.219	.288	.352
$1 < F_d < 10$	136	.087	.130	.057	.134	269	.063	.112	.047	.108
$F_d > 10$	247	.054	.098	.043	.106	1238	.042	.092	.042	.094
$F_{cl/r} < 1$	4	.156	.182	.138	.222	23	.121	.205	.227	.292
$1 < F_{cl/r} < 10$	138	.080	.123	.056	.132	347	.053	.097	.043	.097
$F_{cl/r} > 10$	243	.055	.100	.044	.107	1154	.043	.095	.043	.097
$F_b < 1$	15	.213	.248	.106	.198	270	.094	.166	.100	.172
$1 < F_b < 10$	250	.051	.098	.044	.107	629	.040	.085	.034	.083
$F_b > 10$	120	.077	.116	.054	.127	625	.033	.079	.034	.085

Notes: As in Table XII.

moments (Mariano 1982). Fuller (1977) introduced his approach as a means of guaranteeing that all moments exist. Rothenberg (1984) showed that to a second-order approximation (given iid normal errors) Fuller's k with c set equal to 1 is the unbiased k-class estimator with minimum mean squared error. Stock and Yogo's (2005) weak instrument asymptotics led them to conclude that Fuller's k is "more robust to weak instruments than 2SLS when viewed from the perspective of bias." The one paper in my sample that uses Fuller's k uses this value of c, as do Stock and Yogo (2005), so in the analysis below I set c equal to 1 as well.

Table XIII presents the bootstrap estimated size of LIML and Fuller-k estimators using clustered/robust covariance estimates, contrasting these with 2SLS estimates. In the case of over-identified equations, where LIML results may differ from 2SLS, average size using LIML methods is greater than 2SLS at the .01 level and lower at the .05 level with, outside of better performance in the two regressions with a default F less than 1, no systematic dependence upon the conventional strength of the first-stage relation. Using default covariance

Table XIV: Average Relative (to 2SLS) ln Mean Squared Error and Bias

	LIML (385 coefficients)			Fuller-k (1524 coefficients)		
	MSE	mean bias	median bias	MSE	mean bias	median bias
all regressions	5.35	1.35	.667	-1.11	-.287	.037
$F_d < 1$	12.3	2.55	1.91	-7.45	-2.40	-2.26
$1 < F_d < 10$	7.98	1.94	.992	-1.18	.023	.430
$F_d > 10$	3.84	1.01	.478	-1.01	-.326	-.017
$F_{cl/r} < 1$	13.7	5.22	5.36	-6.58	-1.47	-1.58
$1 < F_{cl/r} < 10$	7.70	1.79	.821	-1.50	-.420	.257
$F_{cl/r} > 10$	3.88	1.03	.502	-.886	-.224	.003
$F_b < 1$	12.2	4.27	3.83	-.498	-.052	-.044
$1 < F_b < 10$	4.00	.824	.176	-1.89	-.530	-.004
$F_b > 10$	7.30	2.07	1.29	-.592	-.145	.113

Notes: Reported values are average ln ratios relative to 2SLS. Numbers used to calculate averages equal those reported in rows of Table XIII. Mean squared error, mean and median bias calculated around each method's own population moment.

estimates to construct p-values, LIML does even worse, with size at the .01 and .05 levels double and 1.5 times, respectively, that of 2SLS (details in the on-line appendix). In sum, the results regarding size based upon iid disturbances and Monte Carlo simulations mentioned above are, in this practical setting, flatly contradicted. Fuller's k modification of LIML does better, producing average results that are quite similar to 2SLS, the only difference being substantially better performance in the few regressions with conventional Fs less than 1.

Table XVI examines the mean squared error and mean and median bias of the different methods around their respective population moments by comparing the distribution of coefficients produced by bootstrap samples from the original data to each method's point estimates for the original data itself.[26] The LIML

---

[26] The three methods are all consistent and hence asymptotically identical, so there is no sense in which one can define a "correct" moment. Consequently, I evaluate each method against its own computation of the desired population moment in the parent data.

estimator performs extraordinarily poorly in almost every respect. Its MSE is greater than that of 2SLS in 99 percent of regressions, with an average ln increase of 5.35, and its mean bias is greater than that of 2SLS ¾ of the time, with an average ln increase of 1.35. Its median bias, which according to the iid normal analysis cited above should be better than 2SLS, is actually .667 worse, as an 1.83 ln increase in the more than half of cases where it does worse more than offsets the -.863 ln reduction in the fewer cases where it does better. The results for Fuller's k are much more encouraging. Fuller's method has lower mean squared error and mean bias than 2SLS about ¾ of the time, achieving overall average ln reductions of -1.11 and -.287 on these measures, while median bias is on par with 2SLS. The LIML estimator does systematically worse in regressions with weaker instruments, while Fullers-k does best in regressions with conventional Fs less than 1, but beyond these has no consistent association with measures of instrument strength.

The results of this section show that established iid based theory is largely misleading: in practical samples weak instrument "robust" methods perform no better, and often much worse, even when first stage Fs are less than 10. Beyond the handful of regressions with the very weakest of instruments, i.e. those with conventional Fs less than 1, there appears to be little to recommend in the Anderson-Rubin approach, particularly in overidentified equations. Similarly, the LIML estimator combines no improvements in size, with grossly increased MSE and mean and median bias. Fuller's method has similar size as 2SLS, but provides substantial improvements in MSE and in bias (at least for Fs less than 1) over 2SLS. This is the only area in which iid based theory, most notably Rothenberg's mean squared error result, works.

Putting aside issues of weak instruments and theory, the results above suggest that use of Fuller's k method can provide substantial advantages in MSE, albeit without improved statistical inference. Repeating Table VI's earlier analysis of MSE around the IV population moment, I find that average ln MSE relative to

OLS across all regressions of 1.52 using 2SLS falls to .27 using Fuller's method. MSE using Fuller's method is worse than OLS only slightly less frequently than 2SLS (.51 vs. .61), but it does better in these circumstances, with the average ln MSE disadvantage of 3.27 using 2SLS falling to 1.55 using Fuller's method. In terms of squared loss around the IV moment, Fuller's k method is still on average less desirable than OLS, but appears to avoid the worst outcomes of 2SLS.

VII. Conclusion

Contemporary IV practice involves the screening of reported results on the basis of the first stage F-statistic, as, beyond argumentation in favour of the exogeneity of instruments, the acceptance of findings rests on evidence of a strong first stage relationship. The results in this paper suggest that this approach is not helpful, and possibly pernicious. Conventional Fs provide none of the bounds on size and bias suggested by asymptotic iid based critical values. Beyond extremely weak cases, with Fs less than 1, conventional first stage F-statistics have no negative relationship whatsoever to size, and no statistically significant relation to bias or mean squared error. In contrast, there is a very substantial probability of a large F arising when there is absolutely no relationship between the excluded instruments and the endogenous second stage variables, with the probability of a clustered/robust or default first stage F greater than 10 in such circumstances exceeding, in my sample, 8 and 20 percent, respectively. In a world in which economists experiment with plausible instruments in the privacy of their offices, publicly reported results could easily be filled with instruments which, while legitimately exogenous in the population, are nevertheless irrelevant or very nearly so, with the strong reported F being the result of an unfortunate finite sample correlation with the endogenous disturbances, producing unpleasantly biased estimates. The widespread and growing use of test statistics with underappreciated fat tails to gain credibility using uninformative critical values is less than ideal.

Economists use 2SLS methods because they wish to gain a more accurate estimate of parameters of interest. In this regard, explicit consideration of the tradeoffs between 2SLS and OLS seems natural. In establishing the conceptual basis for modern 2SLS, Sargan (1958) suggested that, given their inefficiency, 2SLS results only be given consideration if their confidence interval excludes the OLS point estimate. Earlier above, I suggested bootstrapped Durbin-Wu-Hausman and instrument relevance tests as minimal pre-tests, based upon the inefficiency of 2SLS when OLS is unbiased and the dangers of finite sample "identification" when instruments are irrelevant, and then incorporated variants of Sargan's criterion. These approaches, however, throw away information. A more systematic alternative, imaginatively suggested by Feldstein (1974), is to use estimates of mean squared error to form a weighted average of the 2SLS and OLS estimators. The bootstrap, with its estimates of relative bias and mean squared error given the moments found in a paper's sample, can be used to inform this analysis in making inferences about the broader population from which it is drawn. An approach of this sort merits further exploration.

No reader of the instrumental variables papers published in the journals of the American Economic Association can help but be impressed by the ingenuity with which they achieve identification of a wide variety of important effects using thoughtful sources of exogenous variation. The care devoted to research design deserves, however, an equally careful and complementary inference design, one that combines the information in 2SLS and OLS using practical measures of their strengths and weaknesses.

BIBLIOGRAPHY[27]

Anderson, T.W. 1983. "Some Recent Developments of the Distributions of Single-Equation Estimators." In Werner Hildenbrand, ed., Advances in Econometrics, pp. 109-122. Cambridge: Cambridge University Press, 1983.

---

[27] Sources cited in this paper. See on-line appendix for list of papers in the sample.

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.

Anderson, T.W., Naoto Kunitomo and Takamitsu Sawa. 1982. “Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator.” *Econometrica* 50 (4): 1009-1027.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2010. “The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics.” *Journal of Economic Perspectives* 24 (2): 3-30.

Baum, Christopher F., Mark E. Schaffer and Steven Stillman. 2007. “Enhanced routines for instrumental variables/generalized method of moments estimation and testing.” *Stata Journal* 7 (4): 465-506.

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

Cragg, John G. and Stephen G. Donald. 1993. “Testing Identifiability and Specification in Instrumental Variable Models.” *Econometric Theory* 9 (2): 222-240.

Dufour, Jean-Marie. 2003. “Identification, Weak-Instruments, and Statistical Inference in Econometrics.” *Canadian Journal of Economics* 36 (4): 767-808.

Durbin, J. 1954. “Errors in Variables.” *Review of the International Statistical Institute* 22: 23-32.

Greene, William H. 2012. Econometric Analysis, 7th edition. Boston: Pearson Education Ltd.

Hall, Peter. 1986. “On the Bootstrap and Confidence Intervals.” *Annals of Statistics* 14 (4): 1431-1452.

Hall, Peter. 1992. The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag, 1992.

Hall, Peter and Michael A. Martin. 1988. “On Bootstrap Resampling and Iteration.” *Biometrika* 75 (4): 661-671.

Hausman, Jerry A. 1978. “Specification Tests in Econometrics.” *Econometrica* 46 (6): 1251-1271.

Feldstein, Martin. 1974. “Errors in Variables: A Consistent Estimator with Smaller MSE in Finite Samples.” *Journal of the American Statistical Association* 69 (348): 990-996.

Fuller, Wayne A. 1977. “Some Properties of a Modification of the Limited Information Estimator.” *Econometrica* 45 (4): 939-953.

Kinal, Terrence W. 1980. “The Existence of Moments of k-Class Estimators.” *Econometrica* 48 (1): 241-249.

Kloek, Teun. 1981. “OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated.” *Econometrica* 49 (1): 205-207.

Lehmann, E.L and Joseph P. Romano. 2005. Testing Statistical Hypotheses. Third edition. New York: Springer Science + Business Media, 2005.

Mariano, Roberto S. 1982. “Analytical Small-Sample Distribution Theory in Econometrics: The Simultaneous Equations Case.” *International Economic Review* 23 (3): 503-533.

Mosteller, Frederick and John W. Tukey. 1977. Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley Publishing Company.

Moulton, Brent R. 1986. “Random Group Effects and the Precision of Regression Estimates.” *Journal of Econometrics* 32 (3): 385-397.

Nelson, Charles R. and Richard Startz. 1990a. “The Distribution of the Instrumental Variable Estimator and Its t Ratio When the Instrument Is a Poor One.” *Journal of Business* 63 (1): S125–S140.

Nelson, Charles R. and Richard Startz. 1990b. “Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator.” *Econometrica* 58 (4): 967–976.

Rothenberg, Thomas J. 1984. “Approximating the Distribution of Econometric Estimators and Test Statistics.” In Zvi Griliches and Michael E. Intriligator, eds. Handbook of Econometrics, Vol. 2. Amsterdam: North-Holland.

Sargan, J.D. 1958. “The Estimation of Economic Relationships using Instrumental Variables.” *Econometrica* 26 (3): 393-415.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557-586.

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In Andrews, Donald W.K. and James H. Stock, eds, Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. New York: Cambridge University Press.

Stock, James. H., Jonathan H. Wright, and Motohiro Yogo. 2002. "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments." *Journal of Business and Economic Statistics* 20 (4): 518-529.

Summers, Robert. 1965. "A Capital Intensive Approach to the Small Sample Properties of Various Simultaneous Equation Estimators." *Econometrica* 33 (1): 1-41.

Theil, Henri. 1953. "Estimation and Simultaneous Correlation in Complete Equation Systems." The Hague: Central Planning Bureau, memograph. Reprinted in Raj, Baldev and Johan Koerts, eds, Henri Theil's Contributions to Economics and Econometrics, Vol. I. Dordrecht: Kluwer Academic Publishers, 1992.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecificed Models." *Econometrica* 50 (1): 1-25.

Wu, De-Min. 1973. "Alternative Tests of Independence between Stochastic Regressors and Disturbances." *Econometrica* 41 (4): 733-750.

Young, Alwyn. 2017. "Channelling Fisher: Randomization Inference and the Statistical Insignificance of Seemingly Significant Results." Manuscript.

Zellner, Arnold. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association* 57 (298): 348-368.