European Economic Review 147 (2022) 104112

Contents lists available at ScienceDirect

European Economic Review

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

# Consistency without Inference: Instrumental Variables in Practical Application[#]

Alwyn Young [*]

*London School of Economics, U.K*

A B S T R A C T

I use Monte Carlo simulations, the jackknife and multiple forms of the bootstrap to study a comprehensive sample of 1309 instrumental variables regressions in 30 papers published in the journals of the American Economic Association. Monte Carlo simulations based upon published regressions show that non-iid error processes in highly leveraged regressions, both prominent features of published work, adversely affect the size and power of IV tests, while increasing the bias and mean squared error of IV relative to OLS. Weak instrument pre-tests based upon F-statistics are found to be largely uninformative of both size and bias. In published papers IV has little power as, despite producing substantively different estimates, it rarely rejects the OLS point estimate or the null that OLS is unbiased, while the statistical significance of excluded instruments is exaggerated.

## 1. 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) heteroskedasticity robust covariance estimates and their clustered extensions, the implications of non-iid error processes for the quality of inference, and their interaction in this regard with regression and research design, has not received the attention it deserves. Heteroskedastic and correlated errors in highly leveraged regressions produce test statistics whose dispersion is typically much greater than believed, exaggerating the statistical significance of both 1$^{st}$ and 2$^{nd}$ stage tests, while lowering power to detect meaningful alternatives. Furthermore, the bias of 2SLS relative to OLS rises as predicted second stage values are increasingly determined by the realization of a few errors, thereby eliminating much of the benefit of IV. This paper shows that these problems exist in a substantial fraction of published work.

In this paper I use Monte Carlos, the jackknife and multiple forms of the bootstrap to study the distribution of coefficients and test statistics in a comprehensive sample of 1309 2SLS regressions in 30 papers published in the journals of the American Economic Association. Subject to some basic rules regarding methods applied, data and code availability, and computational feasibility, I use all papers produced by a keyword search on the AEA website. I maintain, throughout, the exact specification used by authors and their identifying assumption that the excluded instruments are orthogonal to the second stage residuals. When bootstrapping, jackknifing or generating residuals for Monte Carlos, 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 laid down by

[#] I am grateful to Isaiah Andrews, David Broadstone, Brian Finley and Frank Windmeijer for helpful comments.

[*] Corresponding author.
*E-mail address:* a.young@lse.ac.uk.

https://doi.org/10.1016/j.euroecorev.2022.104112

Available online 10 April 2022

authors themselves.

Monte Carlos, using the regression design and residuals found in my sample, as well as controlled artificial error disturbances with a covariance structure matching that observed in 1st and 2nd stage residuals, show how non-iid errors damage the relative quality of inference using 2SLS. Non-iid errors weaken 1st stage relations, reducing the bias advantage of 2SLS and generating mean squared error that is usually larger than biased OLS. Non-iid errors also increase the probability of spuriously large test statistics when the instruments are irrelevant, particularly in highly leveraged regressions and especially in joint tests of coefficients, i.e. 1st stage F tests. Consequently, while 1st stage relations weaken, 1st stage pre-tests become uninformative, providing little or no protection against 2SLS size distortions or bias. 2SLS standard error estimates become larger and much more volatile, producing null rejection probabilities well in excess of the level of the test, while power falls and 2SLS is increasingly unable to distinguish between a null of zero and the alternative given by the parameter estimates found in published tables.

Monte Carlos show, however, that the bootstrap allows for 2SLS and OLS inference with more accurate size and a much higher ratio of power to size than is achieved using clustered/robust covariance estimates. Thus, while the bootstrap does not undo the increased bias of 2SLS brought on by non-iid errors, it nevertheless allows for improved inference under these circumstances. When published results are examined through the lens of the jackknife and bootstrap, a number of weaknesses are revealed. In published papers, statistical significance rests upon a finding of unusually large t-statistics rather than surprising (under the null) coefficient estimates. First stage relations, when re-examined through the jackknife or bootstrap, are often insignificant, while jackknifed and bootstrapped Hausman (1978) tests find little statistical evidence that OLS is substantively biased, despite large proportional and frequent sign differences between OLS and 2SLS point estimates, as 2SLS estimation is found to be so inaccurate that 2SLS confidence intervals almost always include OLS point estimates. Headline results in the third of my sample with the lowest maximum observational leverage do better on all metrics, but even here at the .01 and .05 levels on average only .23 and .35 of results when bootstrapped or jackknifed reject the null that the instruments are irrelevant and either reject the OLS point estimate or the null that it is unbiased. These results do not validate OLS estimation. Rather, they show that the combination of non-iid errors, highly leveraged regression design, and the intrinsic inefficiency of 2SLS produce results which, while substantively different from OLS, have very little statistical power. 2SLS may be prized for its asymptotic consistency, but in finite samples it often allows for very little inference.

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, 1990b), who used examples to show that size distortions can be substantial when the strength of the first stage relationship is weak, and (Bound et al., 1995), who emphasized problems of bias and inconsistency with weak instruments. These papers spurred Staiger and Stock (1997) and Stock and Yogo's (2005) elegant derivation of weak instrument asymptotic distributions and specific tests to ensure bounds on the size distortions and bias relative to OLS of 2SLS. The theoretical and Monte Carlo work that motivates this literature is largely iid based, a notable exception being Olea & Pflueger (2013), who argue that heteroskedastic error processes weaken 1st stage relations and propose a bias test closely related to the 1st stage clustered/robust F-statistic. This paper supports Olea & Pflueger's insight that non-iid errors effectively weaken 1st stage relations, revives concerns regarding the practical efficiency of 2SLS in the context of leverage, regression design and the power to produce results significantly different from OLS, shows that iid-motivated weak instrument pre-tests perform poorly when misapplied in non-iid settings, and highlights the errors induced by finite sample inference using asymptotically valid clustered/robust covariance estimates in highly leveraged settings, including even the Olea & Pflueger bias test.

The paper proceeds as follows: After a brief review of notation in Section 2, Section 3 describes the rules used to select the sample and its defining characteristics, highlighting the presence of high leverage, sensitivity to outliers and non-iid errors. Section 4 presents Monte Carlos patterned on the regression design and errors found in my sample, showing how non-iid errors worsen inference of all sorts, but especially degrade the ratio of power to size in IV tests while raising the bias relative to OLS of 2SLS estimation. 1st stage pre-tests are found to be largely uninformative, although the Olea & Pflueger bias test does separate low and high bias in over-identified 2SLS regressions with moderate maximal leverage, albeit not with the accuracy suggested by asymptotic results. Section 5 provides a thumbnail review of jackknife and "pairs" and "wild" bootstrap methods. The resampling of the coefficient distribution is found to provide as accurate tail rejection probabilities as the computationally more costly resampling of the t-statistic distribution, particularly in tests of IV coefficients. Section 6 re-examines the 2SLS regressions in my sample using all of the jackknife and bootstrap methods, finding the results mentioned above, while Section 7 concludes with some suggestions for improved practice. An on-line appendix provides alternative versions of tables and comparisons of the Monte Carlo accuracy of different bootstrap methods and outcomes when they are applied to the sample itself.

All of the results of this research are anonymized. Thus, no information is provided, in the paper, public use files or private conversation, regarding results for particular papers. Methodological issues matter more than individual results and studies of this sort rely upon the openness and cooperation of current and future authors. For the sake of transparency, I provide complete code 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.

## 2. Notation and Formulae

It is useful to begin with some notation and basic formulae, to facilitate the discussion which follows. With bold lowercase and uppercase letters indicating vectors and matrices, respectively, the data generating process is taken as given by:

$$\mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \mathbf{u} \text{ and } \mathbf{Y} = \mathbf{Z}\boldsymbol{\pi} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{v}, \tag{1}$$

where **y** is the n x 1 vector of second stage outcomes, **Y** the n x 1 matrix of endogenous regressors, **X** the n x $k_X$ matrix of included exogenous regressors, **Z** the n x $k_Z$ matrix of excluded exogenous regressors (instruments), and **u** and **v** the n x 1 vectors of second and first stage disturbances. The remaining (Greek) letters are parameters, with $\beta$ representing the parameter of interest. Although in principal there might be more than one endogenous right-hand side variable, i.e. **Y** is n x $k_Y$, in practical work this is exceedingly rare (see below) and this paper focuses on the common case where $k_Y$ equals 1.

The nuisance variables **X** and their associated parameters are of no substantive interest, so I use $\tilde{}$ to denote the residuals from the projection on **X** and characterize everything in terms of these residuals. For example, with $\hat{}$ denoting estimated and predicted values, the important first and second stage coefficient estimates are given by:

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

To avoid any confusion, it is also worth spelling out that in referring to "homoskedastic" or "default" covariance estimates below I mean

$$\mathrm{V}(\hat{\boldsymbol{\pi}}) = (\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\hat{\sigma}_v^2 \quad \text{and} \quad \mathrm{V}(\hat{\beta}_{2sls}) = (\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^{-1}\hat{\sigma}_u^2, \tag{3}$$

where $\hat{\sigma}_v^2$ and $\hat{\sigma}_u^2$ equal the sum of the first and second stage squared residuals divided by n minus the k right hand side variables, while in the case of "heteroskedastic" or "clustered/robust" covariance estimates I mean:

$$\mathrm{V}(\hat{\boldsymbol{\pi}}) = c\sum_{\mathbf{i}\in\mathbf{I}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{\mathbf{i}}'\hat{\mathbf{v}}_{\mathbf{i}}\hat{\mathbf{v}}_{\mathbf{i}}'\tilde{\mathbf{Z}}_{\mathbf{i}}(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1} \quad \text{and} \quad \mathrm{V}(\hat{\beta}_{2sls}) = c\sum_{\mathbf{i}\in\mathbf{I}}\hat{\tilde{\mathbf{Y}}}_{\mathbf{i}}'\hat{\mathbf{u}}_{\mathbf{i}}\hat{\mathbf{u}}_{\mathbf{i}}'\hat{\tilde{\mathbf{Y}}}_{\mathbf{i}}\Big/(\hat{\tilde{\mathbf{Y}}}'\hat{\tilde{\mathbf{Y}}})^2, \tag{4}$$

where **i** denotes the group of clustered observations (or individual observations when merely robust) and subscripted **i** the rows of a matrix or vector associated with that group, **I** the set of all such groupings, $\hat{\mathbf{v}}$ and $\hat{\mathbf{u}}$ the first and second stage residuals, *c* a finite sample adjustment (e.g. n/(n-k) in the robust case), and I make use of the fact that the inner-product of **Y** is a scalar.

## 3. The Sample

My sample is based upon a search on www.aeaweb.org using the keyword "instrument" covering the American Economic Review and American Economic Journals which at the time yielded papers up through the July 2016 issue of the AER. I dropped papers that:

a) did not provide public use data files and Stata do-file code;
b) used non-linear methods or non-standard covariance estimates;
c) provided incomplete data or non-reproducible regressions.

I had prior experience with Stata and among papers that provide data only five make use of other software. Conventional linear two stage least squares with either the default or clustered/robust covariance estimate is the overwhelmingly dominant approach in this literature, so I dropped exceedingly rare deviations. This consisted of (only) four papers that used non-linear IV methods, uniquely clustered on two variables or used auto-correlation consistent standard errors, as well as a handful of GMM regressions in two papers whose 2SLS regressions are otherwise included in the sample. There is little to be learnt from a handful of specifications, and clustered/robust linear IV is, virtually without exception, the industry practice.

My search yielded 22 papers that indicated that users should apply to third parties for the confidential data necessary to reproduce the analysis. As the delay and likelihood of success in such applications is indeterminate, I dropped these papers from my sample. Sample sizes in half of these papers are within the mid-range observed in my analysis, as detailed below. I only examined IV regressions that appear in tables, as this allowed me 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. Cleaning of the sample based upon the criteria described above produced 1400 2SLS regressions in 32 papers. Only 41 of these, however, contain more than one endogenous right hand side variable. As 41 regressions are insufficient to draw meaningful conclusions, I further restricted the analysis to regressions with only one endogenous variable. Sample sizes in one paper were in the millions in 90 percent of the IV regressions, with 70 to 250 regressors. I lacked the computer resources to execute the full analysis for this paper and dropped it as well.[1]

The final sample is listed in the on-line appendix and consists of 30 papers, 15 appearing in the AER and 15 in other AEA journals. 27 of these provide JEL codes, and of these all but one reference public, health, labor or development/growth (codes H, I, J and O). Although instrumental variables regressions are central to the argument in all of these papers, with the keyword "instrument" appearing in either the abstract or the title, the actual number of IV regressions varies greatly, as shown in Table 1. While 5 papers

[1] A single run of the IV regressions for this paper requires 2.5 hours of computing time, and executing all of the simulations and analysis for the paper would require roughly 250K such runs, plus additional calculations. Despite the large sample sizes, the regressions in this paper have only about 2000 clusters, putting them in the range observed in the remaining sample. In a similar vein, I dropped two regressions in one paper with more than 10 million observations (but only 166 clusters). As these are not central to the paper and appear as an exploration of "mechanisms", I kept the paper and its other regressions in the sample.

present 98 to 286 IV regressions, 9 have only between 2 to 10.[2] As there is a great deal of similarity within papers in regression design, in presenting averages in tables and text below unless otherwise noted I always take the average *across* papers of the *within* paper average. Consequently, each paper carries an equal weight in determining summary results. Of the 1309 IV regressions in these papers, 1083 are exactly identified by one excluded instrument and 226 (in 12 papers) are over-identified (Table 1). Over-identification magnifies size distortions in first stage tests, as shown below.

Turning to statistical inference, all but one of the papers in my sample use the robust covariance matrix or its multi-observation clustered extension. Sample sizes are generally large, with 7 papers showing an average of 300 to 900 observations per 2SLS regression and another 15 having between 1.4 and 210 thousand. However, the number of statistically independent data groupings, as indicated by authors' clustering decisions, is often much smaller. 14 papers have on average only between 20 and 90 clusters or observations (when not clustered) per 2SLS regression, while another 11 have between 100 and 850. Most tables report the number of observations, but the regression specific number of clusters is only ever given in 6 of the 25 papers which cluster. Although the maximum possible number of clusters can be inferred from the text in another 15 papers, the actual number of clusters often falls far below this limit in specific regressions. The $R^2$ found in regressing the paper average number of clusters on the paper average number of observations is only .04, while the partial $R^2$ from regressing within paper variation in the number of clusters on within paper variation in the number of observations is .02, so reported information on the number of observations provides almost no information on between or within paper variation in the number of independent units used to construct standard error estimates.[3] Future authors might consider reporting the number of clusters in each regression specification.

While the focus of this paper is 2SLS, rather than the substantive results of the sample, at the request of reviewers I separate out headline results in the analysis below. I define a headline 2SLS result as one noted in the abstract, introduction or conclusion and select the estimating equations noted in the text as the "preferred specifications", given precedence by authors based upon the strength of the first stage, sample size, or fewer data caveats, or whose estimates are used in analysis elsewhere in the paper. I rule out results associated with "robustness checks" and "mechanisms", as well as, where numerous effects are mentioned in the introduction and conclusion, those presented in the last page or two of a paper. [4] Altogether I code 61 headline results (listed in the on-line appendix), with 17 papers having one, 10 two or three, and the remaining 3 four to eight (authors often look at multiple outcomes). Results below are given in terms of the cross-paper average of the within paper average for headline results, so that each paper carries equal weight. Headline results tend to be statistically more significant and have stronger first stages. While in the average paper .56 of 2SLS coefficients are significant at the .05 level and the average first stage F is 151, for headline results these figures are .79 and 259, respectively.[5] Despite the larger Fs, headline results share the leverage characteristics of the full sample (see below), and hence their p-values share a similar proportionate sensitivity to alternative inference methods, usefully reinforcing arguments presented in this paper.

The defining characteristic of my sample is the extraordinary sensitivity of reported results to outliers. Fig. 1 graphs the maximum and minimum p-values that can be found by deleting one cluster or observation in each 2SLS regression against the authors' p-value for that instrumented coefficient.[6] With the removal of just one cluster or observation, in the average paper .39 of reported .05 significant 2SLS results can be rendered insignificant at that level, with the average p-value when such changes occur rising from .028 to .158. With the deletion of two observations (panel c), in the average paper no less[7] than .58 of .05 significant IV results can be rendered insignificant. When statistical significance is changed in this manner, .62 of formerly .05 significant results have a delete-two maximum p-value in excess of .10, while their average p-value rises to .252. Conversely, it must be noted that in the average paper .37 and .57 of .05 insignificant IV results can be rendered .05 significant with the removal of one or two clusters or observations, respectively. Headline results are equally sensitive, with .38 of .05 significant results delete-one sensitive and .49 delete-two sensitive, with average p-values in the latter case rising from .022 to .342 with .90 of p-values moving above .1. As panels a and c show, changes can be extraordinary, with p-values moving from close to 0 to near 1.0, and vice-versa. Not surprisingly, the gap between maximum and minimum delete-one and -two IV p-values is decreasing in the number of clusters or observations, as shown in panels b and d of the figure, but very large max-min gaps remain common even with 1000s of clusters and observations.

In my sample the F-statistics authors use to assure readers of the strength of the 1st stage relation are also very sensitive to outliers. Fig. 2 graphs the ratio of the minimum clustered/robust F-statistic found by deleting one or two clusters or observations to the full

[2] These are in the published papers themselves, as I do not code or use results presented in on-line appendices.

[3] Turning to the 22 papers with confidential data mentioned earlier, the average number of observations per regression in these ranges from a minimum of 1500 to a maximum of 1.7 million. Half of these papers have average sample sizes that lie between the minimum and 73rd percentiles of my 30 paper sample. 19 of the 22 papers cluster and information in the text allows the maximum number of clusters to be inferred for 10 of these. It has a min of 70, median of 420 and max of 10K, which lie below the 36th, 84th and 100th percentiles of the 25 papers which cluster in my sample. In sum, sample sizes in such papers are often not extraordinarily large, and for half of these lie within the mid-range of my sample, as noted above.

[4] In the case of 3 papers, the authors critique a standard specification, generally showing how the first stage or 2SLS coefficient can be rendered insignificant with a change of specification. In these cases I use the statistically stronger standard specification as the "headline result".

[5] In some cases authors emphasize results which are not .05 significant either to argue there are "no effects" or because the strong first stage makes the point estimates preferable to those with lower p-values.

[6] I use authors' methods to calculate p-values and where authors cluster, I delete clusters, otherwise I delete individual observations. All averages reported in the paragraph above, as elsewhere in the paper, refer to the average across papers of the within paper average measure.

[7] "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.

**Table 1**
Characteristics of the Sample

| 30 papers | | | | | | 1309 2SLS regressions | | | |
|---|---|---|---|---|---|---|---|---|---|
| # of 2SLS regressions | | average number of | | | | excluded instruments | | covariance estimate | |
| | | observations | | clusters/observations | | | | | |
| 9 | 2-10 | 8 | 40 - 180 | 14 | 20 - 90 | 1083 | 1 | 105 | default |
| 9 | 11-26 | 7 | 300 - 900 | 11 | 100 - 850 | 92 | 2-5 | 992 | clustered |
| 7 | 35-72 | 6 | 1.4K - 2.4K | 5 | 1K - 210K | 134 | 6-60 | 212 | robust |
| 5 | 98-286 | 9 | 8K - 210K | | | | | | |

Notes: K = thousand; M = million; cl/observations = clusters where authors cluster, otherwise observations.

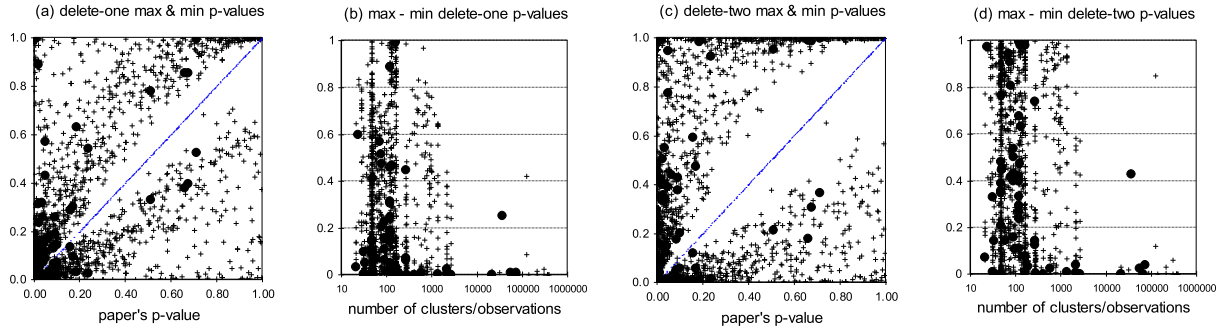


**Fig. 1.** Sensitivity of P-Values to Outliers (Instrumented 2SLS Coefficients), Notes: Solid circles = headline results, plus marks = other results. Panels (a) and (c), above and below 45 degree line are delete-one/two max and min, respectively.

sample F (panels a and b) and the ratio of the full sample F to the maximum delete-one or -two F (panels c and d). With the removal of just one or two observations, the average paper F can be lowered to .71 and .57 of its original value, respectively, or increased to the point that the original value is just .68 or .55, respectively, of the new delete-one or -two F. Headline results again show a similar sensitivity, with the average F falling to .73 and .59 of its original value with the deletion of one or two clusters/observations, and increasing so that the original value is just .69 or .56 of the new delete-one or -two F. As shown in the figure, substantial sensitivity is found in samples with thousands, if not hundreds of thousands, of observations/clusters.

Sample sensitivity of p-values and F-statistics reflects a concentration of "leverage" in a few clusters and observations. Consider the generic OLS regression of a vector **y** 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:

$$\widehat{\beta}_{\sim \mathbf{i}} - \widehat{\beta} = -\widetilde{\mathbf{x}}_{\mathbf{i}}'\breve{\boldsymbol{\varepsilon}}_{\mathbf{i}} \Big/ \widetilde{\mathbf{x}}'\widetilde{\mathbf{x}} \tag{5}$$

where $\widetilde{\mathbf{x}}$ is the vector of residuals of **x** projected on the other regressors, $\widetilde{\mathbf{x}}_{\mathbf{i}}$ the **i** elements thereof, and $\breve{\boldsymbol{\varepsilon}}_{\mathbf{i}}$ the vector of residuals for observations **i** calculated using the delete-**i** coefficient estimates. Clustered/robust covariance estimates are of course given by:

$$c\sum_{\mathbf{i}} \widetilde{\mathbf{x}}_{\mathbf{i}}'\widehat{\boldsymbol{\varepsilon}}_{\mathbf{i}}\widehat{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\widetilde{\mathbf{x}}_{\mathbf{i}} \Big/ (\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}})^2 \tag{6}$$

where the $\widehat{\boldsymbol{\varepsilon}}_{\mathbf{i}}$ are the estimated residuals for observations **i**. Define $\widetilde{\mathbf{x}}_{\mathbf{i}}'\widetilde{\mathbf{x}}_{\mathbf{i}}/\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}}$ as the group **i** share of "coefficient leverage".[8] When leverage is concentrated in a few observations, both OLS coefficient and cl/robust standard error estimates will be heavily influenced by the realizations of the errors for those observations and hence potentially sensitive to their exclusion.[9] These OLS equations obviously have relevance for the IV first stage, but also for the instrumented estimates since (2) and (4) earlier can be re-expressed as functions of OLS coefficients

$$\widehat{\beta}_{2sls} = \frac{\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}'(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})\widehat{\boldsymbol{\beta}}_{\mathbf{yZ}}}{\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}'(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}} \text{ and } \mathrm{V}(\widehat{\beta}_{2sls}) = \frac{c\sum_{\mathbf{i}\in\mathbf{I}}\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}'\widetilde{\mathbf{Z}}_{\mathbf{i}}'\widehat{\mathbf{u}}_{\mathbf{i}}\widehat{\mathbf{u}}_{\mathbf{i}}'\widetilde{\mathbf{Z}}_{\mathbf{i}}\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}}{(\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}'(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}})^2}, \tag{7}$$

where $\widehat{\boldsymbol{\beta}}_{\mathbf{YZ}}$ and $\widehat{\boldsymbol{\beta}}_{\mathbf{yZ}}$ are the OLS first stage and reduced form coefficient estimates derived from regressing the endogenous variable **Y**

[8] So called since leverage is typically defined as the diagonal elements of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ formed using all regressors, while the measure described above is the equivalent for the partitioned regression on $\widetilde{\mathbf{x}}$.

[9] One might usefully contrast the cl/robust covariance estimate in (6) above, which in estimating the covariance between the regressors and residuals uses a leverage share weighted average of the estimated residuals, with the homoskedastic covariance estimate $(N-k)^{-1}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}/(\widetilde{\mathbf{x}}'\widetilde{\mathbf{x}})$, where each residual receives equal weight.

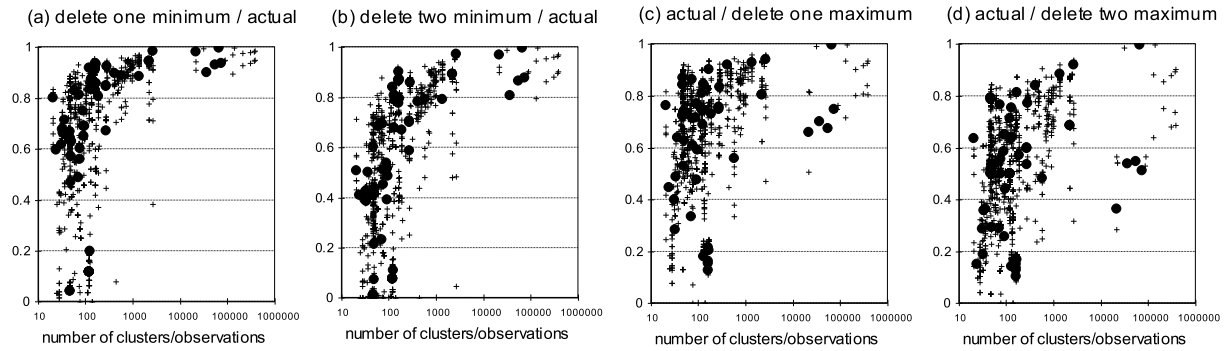


**Fig. 2.** Proportional Change of First Stage F with Removal of One or Two Clusters or Observations, Notes: Solid circles = headline results, plus marks = other results.

and dependent variable **y** on the excluded instruments **Z** (with the included instruments **X** partialed out).

Table 2 summarizes the maximum coefficient leverage shares of the excluded instruments in my sample. In the average paper one cluster or observation on average accounts for .18 of the residual (to other regressors) variation of these instruments and two clusters/observations account for .27. The concentration of leverage is very similar in headline results which as already noted share a similar sensitivity to outliers as the average regression in their paper. Dividing the sample into thirds based upon each paper's average maximum leverage share, in "low" leverage papers this accounts for only .05 of total residual instrument variation, while across "high" leverage papers it averages .33 of instrument variation, reaching an extraordinary high of .70 in the average 2SLS regression of one paper. In the low leverage sample, on average only .16 and .28 of .05 significant results can be rendered insignificant with the deletion of one or two clusters or observations, and the changes in p-values when such significance changes occur are very small, with the average rising to only .09 and .12. In contrast, in the high leverage sample .57 and .84 of .05 significant results are delete-one or -two sensitive, with the average p-value in such circumstances rising to .20 and .34, respectively. A similar association between leverage and the sensitivity of p-values is found in headline results. High maximum leverage is a consequence of the values the instruments take on, and not of conditioning on the included instruments **X**, as leverage shares removing all such covariates other than the constant term from the regression are if anything slightly higher (panel d). For this reason, maximal leverage is quite similar across 2SLS specifications in a given paper, be they headline results or other regressions.

The second defining characteristic of my sample is the deviation of the residuals from the iid normal ideal. As shown in Table 3, using Stata's test of normality based upon skewness and kurtosis, in the average paper more than 80% of the OLS regressions which make up the 2SLS point estimates reject the null that the residuals are normal. In equations which cluster, cluster fixed effects are also found to be significant more than 80% of the time. In close to ½ of these regressions the authors' original specification includes cluster fixed effects, but it is unlikely that the cluster correlation of residuals is limited to a simple mean effect; a view apparently shared by 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 side variables using the test statistics and distributions suggested by Koenker (1981) and Wooldridge (2013) reject the homoskedastic null between $^{2}/_{3}$ and $^{4}/_{5}$ of the time. Headline results share similar residual characteristics to those found elsewhere in the papers. The appeal to "average treatment effects" so often used to motivate the interpretation of

**Table 2**
Coefficient Leverage & Delete-One or -Two Sensitivity

| | all 2SLS results | | | | headline 2SLS results | | | |
|---|---|---|---|---|---|---|---|---|
| | all 30 papers | grouped by leverage | | | all 30 papers | grouped by leverage | | |
| | | 10 low | 10 med | 10 high | | 10 low | 10 med | 10 high |
| (a) maximum shares of instrument leverage ($\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_i/\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}$) | | | | | | | | |
| one cluster/observation | .18 | .05 | .15 | .33 | .17 | .04 | .14 | .33 |
| two clusters/observations | .27 | .08 | .27 | .46 | .26 | .07 | .25 | .46 |
| (b) share of .05 significant results sensitive to deletion of one or two clusters/observations | | | | | | | | |
| one cluster/observation | .39 | .16 | .46 | .57 | .38 | .22 | .42 | .54 |
| two clusters/observations | .58 | .28 | .65 | .84 | .49 | .22 | .42 | .88 |
| (c) max delete-one or -two p-value when .05 significance is delete-one or -two sensitive | | | | | | | | |
| one cluster/observation | .16 | .09 | .17 | .20 | .23 | .08 | .16 | .33 |
| two clusters/observations | .25 | .12 | .29 | .34 | .34 | .14 | .40 | .37 |
| (d) maximum shares of instrument leverage without covariates | | | | | | | | |
| one cluster/observation | .20 | .04 | .18 | .38 | .19 | .04 | .18 | .35 |
| two clusters/observations | .30 | .07 | .30 | .53 | .29 | .07 | .29 | .50 |

Notes: Reported figures are the average across papers of the within paper average measure. Maximum shares refer to the largest share of one or two clusters or observations (when not clustered). Low, med(ium), & high refer to papers grouped by the average maximum leverage of a single cluster/observation in all regressions or headline results. In overidentified equations, leverage shares are the average of those of the **Z** variables.

**Table 3**
Tests of Normality, Cluster Correlation and Heteroskedasticity (average across 30 papers of fraction of regressions rejecting the null)

| | **Y** on **Z, X** (1st stage) | | | | **y** on **Z, X** (reduced form) | | | |
|---|---|---|---|---|---|---|---|---|
| | all results | | headline results | | all results | | headline results | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| normally distributed residuals | .802 | .826 | .767 | .778 | .805 | .878 | .803 | .914 |
| no cluster fixed effects | .839 | .880 | .870 | .870 | .849 | .895 | .792 | .885 |
| homoskedastic (Koenker 1981) | .733 | .802 | .722 | .800 | .641 | .703 | .589 | .649 |
| homoskedastic (Wooldridge 2013) | .736 | .802 | .722 | .800 | .667 | .719 | .622 | .657 |

Notes: .01/.05 = level of the test. Cluster fixed effects only calculated for papers which cluster. Where authors weight I use the weights to remove the known heteroskedasticity in the residuals before running the tests.

coefficients *necessarily* implies heteroskedastic residuals whose variance is correlated with extreme values of the regressors,[10] and public use data provide plenty of evidence that such correlations exist.

Concentrated leverage and heteroskedasticity together undermine the quality of statistical inference in a given sample, the subject and focus of this paper. With concentrated leverage and heteroskedastic errors, coefficients and cl/robust standard errors are heavily determined by the realization of a few residuals, making them unusually volatile and conferring on the regression extremely small sample characteristics. This produces rejection probabilities well above nominal value when standard finite sample N - k or C (# of clusters) - 1 degrees of freedom adjustments for the volatility of variance estimates are used. In the specific context of 2SLS, first stage relations weaken and the bias advantage of 2SLS deteriorates as estimated coefficients are affected by the realization of a few residuals which are correlated with the second stage. Large first stage test statistics become more likely and consequently 1st stage pre-tests become uninformative. Section 4 below uses Monte Carlos to show how the combination of leverage and heteroskedasticity undermine 2SLS, while Section 5 shows that the bootstrap allows for more accurate inference. In the analysis of the sample in Section 6, I find that deviations between bootstrap & jackknife results and those found using conventional techniques are concentrated in papers and regressions with high leverage and evidence of heteroskedasticity, while 2SLS estimates are statistically all but indistinguishable from OLS results in high leverage papers.

## 4. Monte Carlos: 2SLS in IID & Non-IID Settings

This section explores how leverage and clustered heteroskedastic disturbances affect 2SLS using Monte Carlos based on the practical regressions that appear in my sample. I use two sets of Monte Carlos, one based upon artificial errors, and another based upon the actual residuals of my sample. The former allow for a controlled presentation of how leverage and heteroskedasticity interact to undermine 2SLS, while the latter provide a measure of how these forces play out in the sample itself. For all simulations, I estimate the sample's 2SLS system and then use these point estimates as the parameters of a Monte Carlo data generating function:

$$\begin{aligned} &\text{Estimation}: \mathbf{y} = \mathbf{Y}\widehat{\beta}_{iv} + \mathbf{X}\widehat{\boldsymbol{\delta}} + \widehat{\mathbf{u}}, \mathbf{Y} = \mathbf{Z}\widehat{\boldsymbol{\pi}} + \mathbf{X}\widehat{\boldsymbol{\gamma}} + \widehat{\mathbf{v}} \\ &\text{MonteCarlo}: \mathbf{y} = \mathbf{Y}\widehat{\beta}_{iv} + \mathbf{X}\widehat{\boldsymbol{\delta}} + \mathbf{u}, \mathbf{Y} = \mathbf{Z}\widehat{\boldsymbol{\pi}} + \mathbf{X}\widehat{\boldsymbol{\gamma}} + \mathbf{v} \end{aligned} \tag{8}$$

The methods differ in the manner in which the new errors [**u,v**] are created.

In the case of simulations with artificial errors, I calculate the Cholesky decomposition $\mathbf{CC}'$ of the covariance matrix **V** of $[\widehat{\mathbf{u}}, \widehat{\mathbf{v}}]$ and generate $[\mathbf{u},\mathbf{v}] = [\boldsymbol{\varepsilon}_1,\boldsymbol{\varepsilon}_2]\mathbf{C}'$, where $\boldsymbol{\varepsilon}_1$ & $\boldsymbol{\varepsilon}_2$ are independent random variables drawn from standardized distributions (i.e. demeaned and divided by their standard deviation). I use six data generating processes for the observation specific values ($\varepsilon_i$) of $\boldsymbol{\varepsilon}_1$ & $\boldsymbol{\varepsilon}_2$:

9.1. iid standard normal
9.2. heteroskedastic standard normal, where $\varepsilon_i = h_i\eta_i$, $\eta \sim$ iid standard normal
9.3. heteroskedastic clustered standard normal, where $\varepsilon_i = h_i(\eta_i + \eta_c)/2^{1/2}$, $\eta \sim$ iid standard normal
9.4. iid standardized chi$^2$
9.5. heteroskedastic standardized chi$^2$, where $\varepsilon_i = h_i\eta_i$, $\eta \sim$ iid standardized chi$^2$
9.6. heteroskedastic clustered standardized chi$^2$, where $\varepsilon_i = h_i(\eta_i + \eta_c)/2^{1/2}$, $\eta \sim$ iid standardized chi$^2$

To produce heteroskedastic residuals, I use **h** equal to the sample standardized value of the first element **z** in **Z**. As noted earlier, heteroskedastic effects of this kind arise naturally when there is heterogeneity in the effects of **z** on **Y** and **Y** on **y**. In modelling unaccounted for intracluster correlation, there is little point in using simple cluster random effects, as more than half of clustered regressions have cluster fixed effects. Instead, I model the cluster effect as representing iid cluster level draws in the heterogeneity of the impact of **z** on **Y** and **Y** on **y**, with the independent cluster ($\eta_c$) and observation specific ($\eta_i$) components carrying equal weight. Sample standardizing **z** and dividing by $\sqrt{2}$ with clustered errors ensures that the covariance matrix of the disturbances remains unchanged across the six data generating processes. To allow for non-normality, I use standardized chi$^2$ errors, which range from -.7 to infinity and

[10] As a simple example, let $Y_i = (\pi+\pi_i)z_i = \pi z_i + \pi_i z_i = \pi z_i + u_i$, while $y_i = (\beta+\beta_i)Y_i = \beta Y_i + \beta_i(\pi+\pi_i)z_i = \beta Y_i + v_i$, where $\pi_i$ and $\beta_i$ are mean zero random variables that are independent of $z_i$.

are decidedly skewed and non-normal. Results using these errors, however, are similar to those found in the normally based simulations and are consigned to the on-line appendix. When simulations where OLS is unbiased are called for, the off-diagonal elements of **V** are set to 0. Such simulations are noted as having "uncorrelated errors", as opposed to the "correlated errors" of the baseline analysis.

To more closely mimic the actual errors in my sample, I estimate residuals using delete-**i** coefficient estimates

$$\breve{\mathbf{u}}_{\mathbf{i}} = \mathbf{y}_{\mathbf{i}} - \mathbf{Y}_{\mathbf{i}}\widehat{\beta}_{iv\sim\mathbf{i}} - \mathbf{X}_{\mathbf{i}}\widehat{\boldsymbol{\delta}}_{\sim\mathbf{i}} \text{ and } \breve{\mathbf{v}}_{\mathbf{i}} = \mathbf{Y}_{\mathbf{i}} - \mathbf{Z}_{\mathbf{i}}\widehat{\boldsymbol{\pi}}_{\sim\mathbf{i}} - \mathbf{X}_{\mathbf{i}}\widehat{\boldsymbol{\gamma}}_{\sim\mathbf{i}} \tag{10}$$

where $\sim$**i** indicates coefficient estimates excluding cluster **i** (or an individual observation when the regression is not clustered). These "jackknifed" residual pairs $[\breve{\mathbf{u}}, \breve{\mathbf{v}}]$ are then transformed to generate the **[u,v]** added to predicted values based upon the full sample coefficient estimates (as in (8) earlier), creating three different distributions of errors

11.1 iid - jackknifed residual pairs multiplied by a 50/50 iid draw from $\pm 1$ at the observation level and randomly shuffled across observations
11.2 heteroskedastic - jackknifed residual pairs multiplied by a 50/50 iid draw from $\pm 1$ at the observation level, but not shuffled
11.3 heteroskedastic & clustered - jackknifed residual pairs multiplied by a 50/50 iid draw from $\pm 1$ at the cluster level and not shuffled

Where uncorrelated errors are desired, the same procedures are followed, but with the $[\breve{\mathbf{u}}, \breve{\mathbf{v}}]$ residual pairs multiplied by independent $\pm 1$ random variables.

As shown in the on-line appendix, use of 11.1-11.3 to generate Monte Carlo data, when tested on artificial data produced by the data generating processes described in 9.1-9.6, produces results which mimic the patterns produced by these data generating processes. In contrast, applying 11.1-11.3 using the estimated full sample residuals $[\widehat{\mathbf{u}}, \widehat{\mathbf{v}}]$, which are shrunken towards zero in high leverage observations, produces results which bear no resemblance to those produced by 9.1-9.6. That said, it must be borne in mind that jackknifed residuals are not true errors and I find (on-line appendix) that the results produced by such residuals fail to match the full deterioration of outcomes that actually occurs in the most highly leveraged papers using 9.1-9.6.

**(a) Inference and Bias**

Table 4 below begins by reporting null rejection probabilities at the .01 and .05 levels. "$H_0 = \beta_{dgp}$" tests the null that the underlying parameter equals the $\beta$ value used in the data generating process (8), while "$H_0 = 0$" tests the incorrect null that it equals 0. I run 1000 Monte Carlo simulations for each data generating process for each of the 1309 equations and, as usual, report cross paper averages of within paper average rejection rates. Our main interest lies in correlated 1st and 2nd stage errors, but I also report results with uncorrelated errors, where OLS is unbiased and functions properly, to allow a clearer understanding of which features are unique to IV. Results with "actual" errors show less extreme outcomes, particularly in high leverage papers, than those with artificial errors (where all of the heteroskedasticity stems from the heterogeneous effects of the instruments), but the patterns are very much the same.[11] This is repeated in all subsequent tables and not commented on further.

As shown in the table, heteroskedasticity raises the probability of a Type I error in medium and high leverage regressions well above nominal level, while lowering the power to reject the incorrect null of 0 across the board. The increased likelihood of a Type I error emerges from the fact that standard errors become more volatile, producing more dispersed t-statistics, while the degrees of freedom used to evaluate the distribution remain constant.[12] Headline results, despite their stronger first stages, have similar Type I error probabilities. As the OLS results with uncorrelated errors show, the growing probability of a Type I error brought about by the interaction between heteroskedasticity and leverage is not unique to IV, and hence it should not be surprising that IV pre-tests based upon the strength of the 1st stage do not guarantee accurate size, as already seen in the similarity between all and headline results in the table and explored more formally further below. Power to reject a false null falls as the standard error estimate grows in response to the increased volatility of coefficient estimates, and this decline appears to be more severe, both proportionately and in absolute terms, in IV, which has less power to begin with. The bottom right-hand corner of the table shows that with uncorrelated errors IV is an inefficient low-powered substitute for OLS. In contrast, when errors are correlated OLS provides misleadingly precise estimates of biased values, producing gross size distortions. To be sure, these lead to increased power to reject the incorrect null of zero effects, but this is unlikely to be the balance between size and power practitioners are seeking.

Table 5 below reports Monte Carlo estimates of the average truncated ln proportional OLS bias and relative 2SLS to OLS bias and mean squared error. With normal disturbances only the first $k_Z$–$k_Y$ finite sample moments of 2SLS estimates exist (see Kinal 1980 and citations therein). Consequently, in these simulations moments do not exist for most of my sample, which is only exactly identified. However, the moments of the truncated distributions always exist. The table reports moments of estimated coefficients whose absolute

[11] Although, as already noted, jacknifed errors do not reproduce the extreme average outcomes found in high leveraged papers when the underlying data generating process is that of the artificial simulations (on-line appendix).

[12] The average ratio of the 95th percentile of the absolute IV coefficient estimate deviation from $\beta_{dgp}$ divided by the mean of the standard error estimate falls from 2.0 with iid normal errors (1.8 with iid "actual") to 1.2 with heteroskedastic & clustered normal errors (1.6 with "actual"), so standard errors rise more than the dispersion of coefficient estimates. The fraction of t-statistics exceeding the .01 & .05 critical values rises because the volatility of the standard error estimate increases, with the ln of its standard deviation increasing an average of 6.0 (in ln terms!) with normal errors (1.4 with "actual") in the movement from iid to heteroskedastic & clustered errors.

**Table 4**
Average Null Rejection Probabilities at the .01 & .05 Levels (1000 Monte Carlo simulations for each of 1309 equations)

| | $H_0=\beta_{dgp}$ | | | | | | | | $H_0=0$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | all | | low | | medium | | high | | all | | all | | low | | medium | | high | | all | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| (a) correlated errors (all results) | | | | | | | | | | | | | | | | | | | | |
| | 2SLS | | | | | | | | OLS | | | | 2SLS | | | | | | OLS | |
| iid normal | .029 | .077 | .011 | .049 | .036 | .082 | .039 | .101 | .718 | .782 | .461 | .590 | .578 | .694 | .285 | .440 | .519 | .636 | .579 | .682 |
| h normal | .069 | .126 | .012 | .045 | .052 | .106 | .142 | .226 | .528 | .613 | .276 | .375 | .372 | .457 | .132 | .249 | .324 | .418 | .430 | .534 |
| h cl normal | .069 | .123 | .010 | .040 | .054 | .106 | .143 | .224 | .439 | .535 | .182 | .272 | .256 | .333 | .107 | .212 | .183 | .271 | .379 | .488 |
| iid "actual" | .025 | .072 | .014 | .051 | .021 | .064 | .039 | .101 | .693 | .771 | .432 | .553 | .585 | .702 | .231 | .362 | .481 | .594 | .525 | .623 |
| h "actual" | .035 | .085 | .010 | .046 | .042 | .092 | .053 | .117 | .678 | .752 | .409 | .539 | .583 | .708 | .207 | .349 | .436 | .559 | .491 | .593 |
| h cl "actual" | .037 | .085 | .012 | .047 | .044 | .094 | .056 | .115 | .668 | .747 | .297 | .446 | .478 | .634 | .158 | .307 | .255 | .395 | .483 | .589 |
| (b) correlated errors (headline results) | | | | | | | | | | | | | | | | | | | | |
| | 2SLS | | | | | | | | OLS | | | | 2SLS | | | | | | OLS | |
| iid normal | .023 | .072 | .008 | .045 | .023 | .075 | .038 | .097 | .728 | .788 | .566 | .701 | .630 | .754 | .528 | .677 | .541 | .672 | .610 | .699 |
| h normal | .060 | .118 | .009 | .044 | .037 | .096 | .136 | .214 | .520 | .606 | .348 | .455 | .387 | .483 | .277 | .404 | .379 | .478 | .444 | .552 |
| h cl normal | .062 | .117 | .007 | .039 | .044 | .099 | .136 | .213 | .424 | .520 | .229 | .333 | .284 | .372 | .223 | .349 | .179 | .278 | .374 | .495 |
| iid "actual" | .021 | .070 | .007 | .043 | .023 | .076 | .031 | .088 | .698 | .780 | .537 | .661 | .633 | .761 | .458 | .582 | .521 | .641 | .560 | .649 |
| h "actual" | .038 | .089 | .009 | .045 | .057 | .116 | .049 | .106 | .691 | .752 | .518 | .660 | .604 | .764 | .473 | .612 | .479 | .604 | .550 | .630 |
| h cl "actual" | .046 | .092 | .010 | .042 | .062 | .120 | .067 | .114 | .708 | .769 | .404 | .573 | .489 | .689 | .431 | .577 | .293 | .453 | .570 | .644 |
| (c) uncorrelated errors (all results) | | | | | | | | | | | | | | | | | | | | |
| | OLS | | | | | | | | 2SLS | | | | OLS | | | | | | 2SLS | |
| iid normal | .012 | .053 | .010 | .048 | .013 | .056 | .013 | .055 | .018 | .063 | .830 | .874 | .947 | .963 | .740 | .812 | .802 | .848 | .472 | .594 |
| h normal | .068 | .140 | .013 | .055 | .048 | .121 | .143 | .245 | .054 | .107 | .648 | .727 | .850 | .895 | .542 | .646 | .552 | .640 | .295 | .394 |
| h cl normal | .078 | .155 | .017 | .064 | .056 | .132 | .161 | .268 | .053 | .103 | .570 | .665 | .749 | .816 | .500 | .612 | .460 | .567 | .200 | .290 |
| iid "actual" | .017 | .058 | .025 | .066 | .012 | .052 | .013 | .057 | .017 | .058 | .832 | .875 | .949 | .965 | .732 | .805 | .814 | .855 | .449 | .566 |
| h "actual" | .042 | .103 | .031 | .078 | .035 | .096 | .061 | .134 | .028 | .076 | .797 | .851 | .922 | .944 | .684 | .772 | .786 | .836 | .429 | .552 |
| h cl "actual" | .056 | .123 | .027 | .075 | .040 | .102 | .101 | .193 | .026 | .071 | .740 | .815 | .893 | .920 | .638 | .753 | .689 | .772 | .337 | .468 |

Notes: Correlated and uncorrelated refer to the relation between 1st and 2nd stage residuals; h and cl refer to heteroskedastic and clustered data generating processes as described in 9.1-9.3 and 11.1-11.3; low, medium and high leverage divide the sample based upon maximum **Z** leverage (Table 2).

**Table 5**
Ln Truncated OLS Bias & Relative 2SLS to OLS Bias & Mean Squared Error (correlated errors, 1000 Monte Carlo simulations for each of 1309 equations)

| | $\lvert\widehat{\beta}\rvert < 1000*\lvert\beta_{dgp}\rvert$ | | | | | | $\lvert\widehat{\beta}\rvert < 10*\lvert\beta_{dgp}\rvert$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OLS bias all | relative bias all | low | medium | high | relative mse all | OLS bias all | relative bias all | low | medium | high | relative mse all |
| | | | | | | (a) all results | | | | | | |
| iid normal | -0.5 | -3.5 | -4.2 | -2.5 | -3.8 | -0.3 | -0.5 | -3.5 | -4.2 | -2.5 | -3.8 | -0.6 |
| h normal | -0.5 | -2.1 | -2.9 | -1.6 | -1.7 | 1.9 | -0.6 | -2.4 | -3.3 | -1.9 | -2.0 | 0.5 |
| h cl normal | -0.5 | -1.2 | -2.0 | -1.3 | -0.2 | 3.3 | -0.6 | -1.7 | -2.5 | -1.4 | -1.2 | 1.2 |
| iid "actual" | -0.4 | -3.3 | -3.9 | -2.1 | -3.8 | 0.1 | -0.4 | -3.4 | -4.0 | -2.3 | -3.9 | -0.5 |
| h "actual" | -0.4 | -3.0 | -3.8 | -2.0 | -3.1 | 0.4 | -0.5 | -3.0 | -3.9 | -2.1 | -3.1 | -0.3 |
| h cl "actual" | -0.4 | -2.1 | -3.0 | -1.6 | -1.8 | 1.3 | -0.5 | -2.3 | -3.1 | -1.8 | -2.2 | 0.3 |
| | | | | | | (b) headline results | | | | | | |
| iid normal | -0.7 | -3.8 | -4.3 | -3.4 | -3.7 | -0.8 | -0.7 | -3.8 | -4.3 | -3.4 | -3.7 | -0.9 |
| h normal | -0.8 | -2.2 | -3.2 | -1.6 | -1.8 | 1.7 | -0.8 | -2.5 | -3.3 | -2.0 | -2.2 | 0.2 |
| h cl normal | -0.8 | -1.2 | -2.2 | -1.2 | -0.3 | 3.2 | -0.8 | -1.8 | -2.8 | -1.5 | -1.1 | 1.1 |
| iid "actual" | -0.5 | -3.8 | -4.2 | -3.3 | -3.8 | -0.7 | -0.5 | -3.9 | -4.4 | -3.4 | -3.9 | -1.0 |
| h "actual" | -0.6 | -3.4 | -4.0 | -2.7 | -3.6 | -0.4 | -0.6 | -3.5 | -4.0 | -2.7 | -3.6 | -0.7 |
| h cl "actual" | -0.6 | -2.6 | -3.2 | -2.3 | -2.3 | 0.4 | -0.6 | -2.5 | -3.0 | -2.3 | -2.3 | -0.1 |

Notes: Calculated using coefficient estimates whose absolute value is less than 1000 or 10 times the absolute value of the parameter $\beta$ of the data generating process. Low, medium and high refer to papers or headline results by leverage group, as in Table 2. Bias and mse around the parameter $\beta$ of the data generating process. OLS bias = ln(|OLS bias/$\beta$|), relative bias = ln(|IV bias|/|OLS bias|), relative mse = ln(IV mse/OLS mse). Reported figures are the average across papers of the within paper average.

value is less than 1000 or 10 times the absolute value of the parameter $\beta$ of the data generating process. Similar truncation might arise if extreme estimates are dismissed on the grounds of being economically implausible.

As shown in panel (a) of the table, while the bias of OLS does not move meaningfully with the error process, heteroskedastic and clustered errors reduce the bias advantage of 2SLS, especially in high leverage papers.[13] With heteroskedastic errors and high leverage, 1st stage predicted values are heavily influenced by the realization of a few errors that are correlated with 2nd stage disturbances and much of the finite sample bias advantage of 2SLS is lost. Practitioners whose central concern is bias might want to avoid highly leveraged regression specifications. Those who consider second moments will note that, because of its intrinsic inefficiency, the decline in 2SLS' bias advantage eventually leads to a mean squared error that on average is substantially greater than OLS. This problem is ameliorated with greater truncation, but even with truncation to within 10 times the magnitude of the parameter of the data generating process, 2SLS still has higher average mse in 26 and 19 of 30 papers with artificial and "actual" heteroskedastic clustered errors, respectively. Inference using precise but biased OLS estimates seems nonsensical (Table 4), but arguably so is decision-making that does not take into account the volatility and potential bias of IV. Headline results, in panel (b), have somewhat smaller relative bias, but suffer the same deterioration with heteroskedastic errors, resulting in average mse error that with truncation to within 10 times the magnitude of the underlying parameter is still greater than OLS in 23 and 14 of papers with artificial and "actual" heteroskedastic clustered errors, respectively. The formal analysis below shows that stronger first stage Fs improve relative bias, but by no means within the bounds implied by asymptotic theory.

**(b) First stage Pre-Tests and F-tests**

Following the influential work of Nelson and Startz (1990a, 1990b) and (Bound et al., 1995), which identified the problems of size, bias and inconsistency associated with a weak 1st stage relation, all of the papers in my sample try to assure the reader that the relationship between the excluded instruments and right-hand side endogenous variable is strong and results are often singled out based upon the strength of the 1st stage. Twenty-one papers explicitly report 1st stage F statistics in at least some tables, with the remainder using coefficients, standard errors, p-values and graphs to make their case. The reporting of 1st stage F-statistics is, in particular, motivated by Staiger and Stock's (1997) derivation of the weak instrument asymptotic distribution of the 2SLS estimator in an iid world and, on the basis of this, Stock and Yogo's (2005) development of weak instrument pre-tests using the first stage F-statistic to guarantee no more than a .05 probability that 2SLS has size under the null or proportional bias relative to OLS greater than specified levels. In this section I show that in non-iid settings these tests are largely uninformative. Clustered/robust modifications work somewhat better, but only when maximal leverage is low.

[13] Although the effects are not monotonic in the broad categories used in the table, regression analysis (on-line appendix) finds that the increase in relative bias with heteroskedastic errors (normal, chi$^2$ or "actual" and at various levels of truncation) is positively and at the .05 level significantly associated with maximum leverage.

Tables 6 and 7 apply Stock and Yogo's weak instrument pre-tests to each of the 1000 draws for each IV regression from each of the normal and "actual" data generating processes described earlier. 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 of regressions so classified which, based upon the entire Monte Carlo distribution, have rejection probabilities of true nulls or bias greater than the indicated bound.[14] I also report (in parentheses) the maximum fraction of $H_1$ observations violating the bounds that would be consistent with the test having its theoretical nominal size of no greater than .05.[15] With critical values dependent upon the number of instruments and endogenous regressors, Stock and Yogo provide size critical values for 1277 of the 1309 regressions in my sample, but bias critical values for only 134 of the 226 over-identified regressions, where the finite sample first moment can be taken as existing.

Table 6 begins by using the default covariance estimate to evaluate both the F-statistic and coefficient significance when the data generating process is consistent with Stock and Yogo's iid-based theory.[16] In this context, the test performs remarkably well. Only a miniscule share of the regressions which reject the weak instrument null $H_0$ in favour of the strong alternative $H_1$ have Type I error rates greater than the desired bound. Outside of this ideal environment, however, the test rapidly becomes uninformative. When the cl/robust covariance estimate is used to evaluate coefficient significance the test still provides some protection against large size distortions with iid errors, but otherwise the fraction of regressions with Type I error probabilities greater than the specified level in $H_1$ regressions is often greater than that found in $H_0$ and always much larger than the maximum consistent with the test itself having a nominal size of .05. Use of the clustered/robust 1st stage F-statistic as the test-statistic, an ad-hoc adjustment of Stock and Yogo's iid-based theory generally implemented by users,[17] provides no improvement whatsoever. Stock and Yogo's bias test, as shown in Table 7, performs noticeably better, but still quite poorly. In non-iid settings the fraction of regressions with IV to OLS relative bias greater than the specified amount in $H_1$ is always lower than in the $H_0$ sample, but, at levels ranging from $^1/_3$ to .9 with heteroskedastic clustered errors, too high to either be consistent with the test having .05 size or provide much comfort to users. The misapplication of Stock & Yogo's iid based test in non-iid settings does not yield useful results.[18]

Olea and Pflueger (2013), noting that the widespread application of Stock & Yogo's test in non-iid settings is not justified by theory, undertake the challenging task of extending the test to non-iid environments, deriving critical values for the null hypothesis that the IV Nagar bias is smaller than a "worst-case" benchmark. The Nagar bias is that of an approximating distribution based on a third-order Taylor series expansion of the asymptotic distribution, while the worst-case benchmark equals the OLS bias in the case of iid errors. The test statistic is related to the clustered/robust 1st stage F-statistic, but the calculation of sample dependent degrees of freedom for the test is computationally costly and impractical for the many simulations underlying the table which follows. Olea and Pflueger note, however, that conservative degrees of freedom can be estimated using only the eigenvalues of the robust 1st stage F-statistic, and I make use of this approach along with the table of critical values they provide. These conservative degrees of freedom should lower the probability of a Type-I error, i.e. classifying as $H_1$ a regression with a relative bias greater than the desired level, below the .05 size of the test.

Table 8 applies Olea & Pflueger's test to the Monte Carlo sample. As before, I divide regressions by whether or not they reject the weak instrument null and report the fraction of regressions in each group where the relative bias of IV to OLS, as estimated from the Monte Carlo distribution, exceeds the acceptable bound. In fairness, this relative bias is not the object of the test, which concerns asymptotic bias relative to a worst case IV-approximation benchmark, but I would argue it is the object of interest to users, who use 2SLS in order to avoid OLS bias. As shown in the table, for over-identified regressions in low and medium leverage papers bias levels in

[14] That is, each individual data draw is classified into $H_0$ or $H_1$ based upon its 1st stage F statistic, but the size or bias characteristics of a particular regression specification are evaluated using the combined distribution from 1000 draws. I follow Stock & Yogo's theory using the asymptotic $chi^2$ distribution to calculate p-values (Stock & Yogo 2005, pp. 83-84, 88). Results using the t-distribution in the on-line appendix are similar.

[15] 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 in each group, with $W_1 = \alpha(W_0+W_1)$ and $S_0 = (1\text{-}p)(S_0+S_1)$, where $\alpha \leq .05$ and p denote size and power. Then $W_1/N_1 = (\alpha/(1\text{-}\alpha))(N_0\text{-}S_0)/N_1$, which, for given $N_0$ & $N_1$, is maximized when $p = 1$ and $\alpha = .05$, with $W_1/N_1 = (1/19)(N_0/N_1)$. The relative number of regressions in the $N_0$ and $N_1$ groups for each test in the table can be calculated by inverting this equation.

[16] As the number of papers with any results classified in $H_0$ or $H_1$ varies substantially as one moves down the columns or across the rows of the table, here and in Tables 7 & 8 below I depart from the practice of reporting averages across papers of within paper averages, and simply weight each simulation regression equally. These tables only report results for a subset of size and bias bounds. Results for all bounds, leverage groups and including $chi^2$ errors are in the on-line appendix.

[17] Ten of the papers in my sample that report F-statistics make direct reference to the work of Stock and his co-authors. All of these report clustered/robust measures, although two report default F-statistics as well. This ad hoc adjustment may have been motivated by the Stata command ivreg2, which reports the Kleibergen-Paap F (identical to the cl/robust F with one endogenous variable) and compares it to Stock & Yogo's critical values.

[18] Results for the size test broken down by paper leverage (in the on-line appendix) do not find it to be informative in low, medium or high leverage sub-samples. Results for the bias test cannot be meaningfully broken down by leverage group. The 134 regressions for which Stock & Yogo provide bias bounds only cover one high leverage paper and three low leverage papers, and in the latter almost all observations, but for those from one regression, exceed the test bounds.

**Table 6**
Fraction of Regressions with Null Rejection Probabilities Greater than Size Bound in Specifications that Don't ($H_0$) and Do ($H_1$) Reject the Stock & Yogo Weak Instrument Null (1000 simulations for each error process in 1277 IV regressions)

| | size = .10 | | size = .15 | | size = .20 | | size = .25 | |
|---|---|---|---|---|---|---|---|---|
| (A) default IV coefficient covariance estimate, default F used as Stock and Yogo test statistic | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) |
| iid normal | .126 | .000 (.022) | .094 | .000 (.013) | .067 | .000 (.010) | .053 | .000 (.009) |
| iid "actual" | .085 | .003 (.028) | .058 | .002 (.017) | .036 | .002 (.013) | .040 | .002 (.011) |
| (B) cl/robust IV coefficient covariance estimate with | default F used as test statistic | | | | cl/robust F used as test statistic | | | |
| | size = .10 | | size = .25 | | size = .10 | | size = .25 | |
| | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) |
| iid normal | .258 | .267 (.022) | .058 | .009 (.009) | .247 | .270 (.019) | .063 | .009 (.008) |
| h normal | .425 | .268 (.020) | .042 | .061 (.011) | .394 | .247 (.041) | .045 | .062 (.018) |
| h cl normal | .415 | .449 (.019) | .050 | .083 (.011) | .470 | .383 (.101) | .055 | .094 (.037) |
| iid "actual" | .251 | .269 (.028) | .036 | .011 (.011) | .236 | .275 (.022) | .041 | .011 (.010) |
| h "actual" | .254 | .389 (.026) | .091 | .045 (.012) | .244 | .398 (.028) | .098 | .043 (.012) |
| h cl "actual" | .316 | .385 (.026) | .094 | .099 (.012) | .349 | .378 (.060) | .128 | .084 (.025) |

Notes: Regressions for which Stock & Yogo (2005) provide critical values; max = maximum share of the sample in $H_1$ with size greater than bound consistent with the test itself having size .05; IV Type I error rates based upon 1000 Monte Carlo simulations per regression, with coefficient significance evaluated using the normal distribution (following Stock & Yogo). Results using the t-distribution (on line appendix) are very similar.

**Table 7**
Fraction of Regressions with Relative Bias Greater than Bias Bound in Specifications that Don't and Do Reject the Stock & Yogo Weak Instrument Null (1000 simulations for each error process in 134 over-identified IV regressions)

| | default F used as test statistic | | | | cl/robust F used as test statistic | | | |
|---|---|---|---|---|---|---|---|---|
| | bias = .05 | | bias = .30 | | bias = .05 | | bias = .30 | |
| | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) |
| iid normal | .988 | .153 (.162) | .668 | .043 (.068) | .991 | .174 (.155) | .655 | .248 (.030) |
| h normal | .992 | .216 (.137) | .768 | .415 (.025) | .984 | .649 (.032) | .546 | .528 (.006) |
| h cl normal | .995 | .869 (.114) | .833 | .762 (.023) | .972 | .944 (.034) | .970 | .759 (.007) |
| iid "actual" | .971 | .139 (.181) | .705 | .040 (.084) | .989 | .172 (.157) | .725 | .273 (.033) |
| h "actual" | .961 | .116 (.178) | .580 | .176 (.067) | .983 | .105 (.163) | .625 | .181 (.052) |
| h cl "actual" | .966 | .671 (.193) | .589 | .402 (.069) | .966 | .604 (.252) | .615 | .396 (.054) |

Notes: Unless otherwise noted, as in Table 6 above. Bias calculated using the full (not truncated) distribution, as with normal errors the first moment exists when the regression is over-identified.

**Table 8**
Fraction of Regressions with Relative Bias Greater than Bias Bound in Specifications that Don't and Do Reject the Olea & Pflueger Weak Instrument Null (1000 simulations for each error process)

| | bias = .05 | | bias = .10 | | bias = .20 | | bias = $^1/_3$ | |
|---|---|---|---|---|---|---|---|---|
| | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) | $H_0$ | $H_1$ (max) |
| | 174 over-identified IV regressions in 8 low and medium leverage papers | | | | | | | |
| iid normal | .939 | .040 (.249) | .861 | .045 (.226) | .815 | .033 (.196) | .587 | .041 (.146) |
| h normal | .907 | .240 (.391) | .907 | .182 (.266) | .871 | .183 (.204) | .650 | .194 (.177) |
| h cl normal | .938 | .432 (.698) | .894 | .258 (.376) | .880 | .264 (.242) | .767 | .235 (.199) |
| iid "actual" | .930 | .074 (.251) | .876 | .001 (.229) | .799 | .001 (.199) | .648 | .002 (.169) |
| h "actual" | .877 | .093 (.334) | .879 | .044 (.242) | .729 | .043 (.210) | .538 | .062 (.181) |
| h cl "actual" | .899 | .219 (.381) | .886 | .135 (.258) | .736 | .071 (.211) | .543 | .071 (.181) |
| | 52 over-identified regressions in 4 high leverage papers | | | | | | | |
| iid normal | .000 | .197 (.024) | .000 | .118 (.012) | .000 | .000 (.005) | .000 | .000 (.003) |
| h normal | .969 | .206 (.050) | .878 | .207 (.036) | .839 | .191 (.026) | .865 | .219 (.021) |
| h cl normal | .985 | .906 (.908) | .978 | .842 (.376) | .968 | .847 (.198) | .899 | .843 (.147) |
| iid "actual" | .000 | .083 (.023) | .000 | .070 (.011) | .000 | .064 (.006) | .000 | .041 (.004) |
| h "actual" | .485 | .162 (.034) | .197 | .074 (.027) | .000 | .026 (.017) | .000 | .024 (.012) |
| h cl "actual" | .528 | .480 (.246) | .485 | .471 (.111) | .326 | .365 (.049) | .305 | .277 (.037) |

Notes: As in Table 6 above.

**Table 9**
Average Rejection Rates of True Nulls at the .05 Level in 1st Stage Tests (1000 Monte Carlo simulations for each of 1309 equations)

| | default | | | clustered/robust | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | low leverage | | | medium leverage | | | high leverage | | |
| | all | $k_Z > 1$ | | all | $k_Z > 1$ | | all | $k_Z > 1$ | | all | $k_Z > 1$ | |
| | | coef | joint | | coef | joint | | coef | joint | | coef | joint |
| iid normal | .051 | .050 | .050 | .056 | .057 | .061 | .149 | .071 | .235 | .134 | .111 | .355 |
| h normal | .404 | .253 | .463 | .062 | .061 | .070 | .132 | .053 | .149 | .281 | .156 | .481 |
| h cl normal | .595 | .355 | .652 | .066 | .064 | .068 | .133 | .054 | .144 | .308 | .199 | .500 |
| iid "actual" | .054 | .051 | .056 | .056 | .057 | .059 | .132 | .065 | .203 | .124 | .101 | .342 |
| h "actual" | .196 | .138 | .223 | .057 | .066 | .070 | .208 | .084 | .273 | .203 | .136 | .359 |
| h cl "actual" | .372 | .232 | .390 | .061 | .074 | .075 | .211 | .083 | .276 | .226 | .136 | .397 |

Notes: Reported figures are averages of paper average rejection rates; $k_Z > 1$ = average across 3 low, 5 medium and 4 high leverage papers in equations with more than 1 excluded instrument; coef = test of individual coefficients on excluded instruments; joint = joint test of all excluded instruments.

regressions which reject $H_0$ in favour of $H_1$ are generally much lower, although they sometimes exceed the maximum bound consistent with the test having no more than a .05 probability of Type-I error.[19] In highly leveraged regressions, however, the test performs rather poorly, as with heteroskedastic clustered errors bias levels in $H_1$ regressions are always as high as in those which cannot reject the weak instrument null $H_0$.[20]

Table 9 reports Monte Carlo estimates of Type I error probabilities in 1st stage F-tests using default and clustered/robust covariance estimates. As expected, null rejection probabilities with default covariance estimates are close to nominal level with iid disturbances, but explode with non-iid errors. Clustered/robust covariance estimates provide better results, especially in low leverage papers, but rejection rates are very high in medium and high leverage papers, particularly in over-identified equations. Type I errors appear to increase when more than one coefficient is tested, which the table shows by comparing the average rejection probability of coefficient level (t) tests of the excluded instruments in over-identified equations with the much higher rejection rates found in the joint (F) tests of these instruments.[21] In the asymptotic world that forms the foundation of Olea & Pflueger's results, clustered/robust covariance estimates should allow for exact inference. As shown by Table 9, in the finite sample highly-leveraged world of published papers this is very far from the case, and problems of inaccurate inference appear to be compounded in higher dimensional tests, making large clustered/robust 1st stage Fs much more likely than suggested by asymptotic theory. This probably renders the Olea/Pflueger test less informative than it might otherwise be.

## 5. Improved Finite Sample Inference Using the JackKnife and Bootstrap

This section shows that the jackknife and bootstrap provide improved finite sample inference, with rejection probabilities closer to nominal level and greater relative power than found using standard clustered/robust covariance estimates and their associated degrees of freedom. These methods are often evaluated based upon their asymptotic properties, but their usefulness lies in their superior finite sample performance, which is often unrelated to asymptotic results. I begin by describing the methods and then use Monte Carlos to establish their finite sample benefits.

**(a) The JackKnife**

The jackknife covariance estimate based on the full sample ($\widehat{\boldsymbol{\beta}}$) and m delete-**i** ($\widehat{\boldsymbol{\beta}}_{\sim \mathbf{i}}$) coefficient values is given by:

$$\frac{m-1}{m}\sum_{\mathbf{i}}(\widehat{\boldsymbol{\beta}}_{\sim \mathbf{i}}-\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{\sim \mathbf{i}}-\widehat{\boldsymbol{\beta}})' \quad = \quad \frac{m-1}{m}\sum_{\mathbf{i}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{\mathbf{i}}'\breve{\boldsymbol{\varepsilon}}_{\mathbf{i}}\breve{\boldsymbol{\varepsilon}}_{\mathbf{i}}'\mathbf{X}_{\mathbf{i}}(\mathbf{X}'\mathbf{X})^{-1} \tag{12}$$

where the $\breve{\boldsymbol{\varepsilon}}_{\mathbf{i}}$ are the delete-**i** residuals for observations **i** (as in (10) earlier), and where, for expositional purposes, in the second expression I substitute using the formula for the delete-**i** change in coefficient estimates in the OLS regression on variables **X**. The jackknife was shown by Hinkley (1977) to be asymptotically robust to arbitrary heteroskedasticity. For OLS, its use of delete-**i** residuals rather than the estimated residuals is equivalent to the "hc3" finite sample correction of the standard cl/robust covariance estimate (MacKinnon and White 1985).

[19] With chi$^2$ errors (see the on-line appendix) this is actually the case for all bias bounds with non-iid errors.

[20] Olea & Pflueger also provide critical values for exactly identified equations, as the Nagar bias always exists even if the first moment does not. Applying these and comparing relative 2SLS to OLS bias in the truncated distributions in the on-line appendix, I find the test performs worse in this sample. Although bias levels in $H_1$ are generally lower than in the $H_0$ group in low and medium leverage papers, in all leverage groups and for all error processes they are multiples of the limit consistent with the test having a maximum .05 Type-I error rate.

[21] Intuition for this may lie in the fact that the familiar F-statistic actually equals 1/k times the maximum squared t-statistic that can be found by searching over all possible linear combinations **w** of the estimated coefficients, that is $k^{-1}\widehat{\boldsymbol{\beta}}'\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\beta}} = k^{-1}\mathrm{Max}_{\mathbf{w}}(\mathbf{w}'\widehat{\boldsymbol{\beta}})^2/\mathbf{w}'\widehat{\mathbf{V}}\mathbf{w}$. In the test of a single coefficient, the clustered/robust covariance estimate may be biased and have a volatility greater than nominal degrees of freedom, but a joint test involves a search across all possible combinations of this bias and volatility to generate maximal test statistics, producing tail probabilities that are more distorted away from iid-based nominal values than the tests of the individual coefficients.

**(b) The Bootstrap**

I use two forms of the bootstrap, the non-parametric "pairs" resampling of the data and the parametric "wild" bootstrap transformation of residuals. 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 resampling 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 $F_0$.

The resampling bootstrap described above is fully nonparametric, as the only assumption is that the sample can be divided into groups that are independent draws from the population distribution.[22] From a regression perspective, however, the samples are "pairs" of dependent outcomes and regressors and, as such, the estimated distribution of the test statistic is that with both stochastic residuals and regressors. The "wild" bootstrap imposes parametric structure and uses transformations of the residuals to mimic a more traditional resampling of stochastic residuals alone. For example, in the regression model $Y_i = \mathbf{z}_i'\boldsymbol{\beta}_\mathbf{z} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + v_i$, to calculate the distribution of coefficients and test statistics under the null that $\boldsymbol{\beta}_\mathbf{z} = \mathbf{0}$ one estimates the restricted equation $Y_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_\mathbf{x} + \widehat{v}_i$, then generates artificial data $Y_i^{wild} = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_\mathbf{x} + \eta_i\widehat{v}_i$, where $\eta_i$ is a 50/50 iid observation or cluster level draw from the pair (-1,1), and finally runs $Y_i^{wild}$ on $\mathbf{z}_\mathrm{i}$ and $\mathbf{x}_\mathrm{i}$. The initial estimation of the parametric data generating process can involve imposing the null, as just done, or not, the transformations $\eta_i$ can be symmetric or asymmetric, and can involve the actual or delete-**i** residuals. In Monte Carlo studies reported in the on-line appendix I find that a failure to impose the null results in rejection probabilities well above nominal level and asymmetric transformations provide no advantages, even when the data generating process for the residuals $v_i$ is decidedly asymmetric. Imposing the null eliminates the negative influence of leverage on estimated residuals, allows for more accurate inference than the use of delete-**i** residuals alone, and is the method used in the remainder of the paper. Full details on how I impose the null for each separate test and how this improves the accuracy of inference using the method are provided in the on-line appendix.

For both the pairs resampling and wild transformations bootstraps I draw inferences using two methods, one based upon the distribution of bootstrapped test statistics (the bootstrap-t) and another based upon the distribution of bootstrapped coefficients (the bootstrap-c). To illustrate with the case of the resampling bootstrap, one can test whether the estimate $\boldsymbol{\beta}_1$ based on the sample $F_1$ is different from **0** by looking at the distribution of the Wald-statistics for the test that the estimates $\boldsymbol{\beta}_2$ based on the sample $F_2$ drawn from $F_1$ are different from $\boldsymbol{\beta}_1$ (the known parameter value for the data generating process), computing the probability

$$\left(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1\right)'\mathbf{V}\left(\boldsymbol{\beta}_2^i\right)^{-1}\left(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1\right) \;>\; (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}(\boldsymbol{\beta}_1)^{-1}(\boldsymbol{\beta}_1 - \mathbf{0}) \tag{13}$$

where $\boldsymbol{\beta}_1$ is the vector of coefficients estimated using the original sample $F_1$, $\boldsymbol{\beta}_2^i$ the vector of coefficients estimated in the $i^{\text{th}}$ draw of sample $F_2$ from $F_1$, and $\mathbf{V}(\boldsymbol{\beta}_1)$ and $\mathbf{V}(\boldsymbol{\beta}_2^i)$ their respective clustered/robust covariance estimates. In the case of a single coefficient, this reduces to calculating the distribution of the squared t-statistic, i.e. the probability:

$$\left[\left(\beta_2^i - \beta_1\right)/\widehat{\sigma}\left(\beta_2^i\right)\right]^2 \;>\; \left[(\beta_1 - 0)/\widehat{\sigma}(\beta_1)\right]^2 \tag{14}$$

where $\widehat{\sigma}$ is the estimated standard error of the coefficient. This method, which requires calculating an iteration by iteration covariance or standard error estimate, is the bootstrap-t. Alternatively, one can use the distribution of the bootstrapped coefficients to compute a common covariance estimate, calculating the probability:

$$\left(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1\right)'\mathbf{V}\left(F\left(\boldsymbol{\beta}_2^i\right)\right)^{-1}\left(\boldsymbol{\beta}_2^i - \boldsymbol{\beta}_1\right) \;>\; (\boldsymbol{\beta}_1 - \mathbf{0})'\mathbf{V}\left(F\left(\boldsymbol{\beta}_2^i\right)\right)^{-1}(\boldsymbol{\beta}_1 - \mathbf{0}) \tag{15}$$

where $\mathbf{V}(F(\boldsymbol{\beta}_2^i))$ denotes the covariance matrix of $\boldsymbol{\beta}_2^i$ calculated using the empirical bootstrapped distribution of the coefficients. In the case of an individual coefficient, the common variance in the denominator on both sides cancels and the method reduces to calculating the probability:

$$\left(\beta_2^i - \beta_1\right)^2 \;>\; (\beta_1 - 0)^2 \tag{16}$$

which is simply the tail probability of the squared coefficient deviation from the null hypothesis. This method is the bootstrap-c. The frequency with which the inequalities in (13) - (16) occur forms the basis of the calculation of the p-value in each test.

From the point of view of asymptotic theory, the bootstrap-t is considered superior, but in practical application it has its weaknesses. Hall (1992) showed that while coverage error in a symmetric hypothesis test of a single coefficient of the resampling bootstrap-t converges to zero at a rate $O(n^{-2})$, the coverage error of the bootstrap-c converges at a rate of only $O(n^{-1})$, i.e. no better than the convergence of asymptotic normal approximations. The intuition for this, as presented by Hall, lies in the fact that the bootstrap-t estimates an asymptotically pivotal distribution, one that does not depend upon unknowns, while the bootstrap-c estimates an asymptotically non-pivotal distribution, one that depends upon the estimated variance. As the sample expands to infinity, the

[22] Thus, in implementing the method, I follow the assumptions implicit in the authors' covariance calculation methods: resampling clusters where they cluster and resampling observations where they do not.

bootstrap-c continues to make estimates of this parameter, which results in greater error and slower convergence of rejection probabilities to nominal value. This argument, however, as recognized by Hall (1992, p. 167) himself, rests upon covariance estimates being sufficiently accurate so that the distribution of the test statistic is actually pivotal. Hall's concern is particularly relevant in the context of using asymptotically valid clustered/robust covariance estimates in highly leveraged finite samples. I find (below) that the bootstrap-c performs at least as well as the bootstrap-t in tests of IV coefficients and is by no means very much worse in tests of OLS coefficients.[23]

"Publication bias" argues in favour of comparing results using the bootstrap-c to those found using the -t in a study such as this. If results are selected for publication on the basis of statistical significance, they will have unusual t-statistics, regardless of whether the null is true or false. However, to the degree the distribution of the standard error is independent of the distribution of coefficient estimates, spuriously large t-statistics will not be perfectly correlated with spuriously large point estimates and significance rates using the bootstrap-c will be substantially lower than those found using the -t. This is precisely the pattern I find in the analysis of my sample below. Significant published IV results do not have unusually large coefficient values under the null, but they do have unusually large t-statistics, and hence appear systematically more significant when analyzed using the bootstrap-t, despite the fact that the bootstrap-c and -t have similar size and power in Monte Carlos, as shown shortly below.

**(c) Monte Carlos**

Table 10 below presents a Monte Carlo analysis of the different methods using the normal and "actual" error data generating processes described in 9.1 - 9.3 and 11.1 - 11.3 earlier (results using $chi^2$ errors show similar patterns and are given in the on-line appendix). As calculation of the jackknife and bootstrap (with 1000 draws per instance) is very costly, I only evaluate 10 realizations of each data generating process for each of the 1309 equations. With 13090 p-values per data generating process, this still allows evaluation of average size and power. For comparison, I also report results for cl/robust methods using the same 13090 realizations of data.[24] For conventional 2SLS tests of $2^{nd}$ stage instrumented coefficients and $1^{st}$ stage F-tests I see whether empirical rejection probabilities of true nulls are close to nominal level by testing whether the parameters equal those of the data generating process, which are the coefficient estimates of the original authors' estimates, and get some sense of power by testing the false null that they equal zero. For Hausman (1978) tests of the bias of OLS coefficient estimates, the equivalent tests involve data generating processes with uncorrelated and correlated errors, respectively, as these are the circumstances in which the null of no OLS bias is true or false.[25]

Three patterns are readily apparent in Table 10. First, the jackknife and all forms of the bootstrap provide Type I error rates much closer to nominal value than cl/robust methods, while raising the ratio of power to Type I errors. Results given in the on-line appendix show that the improvement in Type I error rates brought about by the use of the jackknife and bootstrap are concentrated in medium and high leverage papers, while in low leverage papers these methods are as accurate as cl/robust inference. Second, as noted earlier, the bootstrap-c is at least as accurate, and often more so, as the -t in tests of IV coefficients and is by no means systematically worse in other tests. Third, as can be seen in Table 10, no matter which method is used, there is a very substantial decline in power with non-iid error processes. This must be borne in mind when evaluating results for the actual sample further below.

While the bootstraps may provide improvement over inference using cl/robust covariance estimates, it is important to note that they are not immune to weak instrument problems. Dufour (1997) showed that for a confidence interval to have a probability 1-α of covering the true parameter value whose range is unbounded when the model is locally almost unidentified, it must have a probability of at least 1-α of delivering an unbounded confidence interval. While confidence intervals with the wild bootstrap when the null is imposed may be unbounded (Davidson and MacKinnon 2008), it is unclear whether they attain the required 1-α probability. Moreover, if confidence intervals are almost surely bounded, as is the case for the jackknife and pairs bootstrap, asymptotically the null rejection probability for some true parameter value will be 1.0! However, if extraordinary parameter values are ruled out on a priori grounds, i.e. the parameter range is bounded, such pathologies need not arise (e.g., see Gleser & Hwang 1987). Such truncation of considered values also eliminates much of the mean squared error disadvantages of 2SLS, as suggested earlier in Table 5.

## 6. Application of the Jackknife and Bootstrap to the Sample Itself

This section applies the jackknife and bootstrap to the sample itself, separately reporting on all published and only headline results. Headline results share the same instruments and hence maximum leverage shares of all published results (Table 2 earlier) and

[23] Similarly, I find that the bias corrected and accelerated bootstrap, which is another asymptotic refinement, performs very poorly in finite samples (on-line appendix). An asymptotic result I do find to be relevant is Hall's (1992) argument that, because they minimize the influence of skewness, symmetric tests (such as those described in (13) - (16) above) converge to nominal size at twice the rate of asymmetric equal tailed tests. In finite sample Monte Carlos (on-line appendix) I find that asymmetric tests are less accurate than symmetric tests.

[24] Comparing IV rejection rates for conventional cl/robust methods using 13090 iterations in the upper left-hand corner of Table 10 with the same using 1309000 iterations in upper left-hand corner of Table 4, one sees that using 10 vs 1000 iterations has very little effect on averages. The on-line appendix shows that this is true for all of the conventional cl/robust rejection rates reported in Table 10. 10 iterations per equation for 1309 equations yields reasonably accurate estimates of the average and relative performance of the different methods.

[25] I use Hausman's test based upon the cl/robust significance of the coefficient on the $1^{st}$ stage residuals entered into the $2^{nd}$ stage OLS regression. The test of the difference between the IV and OLS $2^{nd}$ stage coefficients, which is equivalent with homoskedastic variance estimates, cannot be properly adapted to non-iid circumstances as the cl/robust IV variance estimate is not always larger than the corresponding OLS estimate. When performed using homoskedastic variance estimates, I find it has large size distortions and poor power (see the on-line appendix).

**Table 10**
Improved Finite Sample Inference Using the JackKnife & Bootstrap (average within paper rejection rates at .01 and .05 levels, 10 Monte Carlos for each of 1309 equations)

| | tests of true nulls | | | | | | | | | | | | tests of false nulls | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | clustered/ robust | | jackknife | | pairs bootstrap | | | | wild bootstrap | | | | clustered/ robust | | jackknife | | pairs bootstrap | | | | wild bootstrap | | | |
| | | | | | c | | t | | c | | t | | | | | | c | | t | | c | | t | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| IV coefficients (correlated 1st and 2nd stage errors): $H_0 = \beta_{dgp}$ or 0 | | | | | | | | | | | | | | | | | | | | | | | | |
| iid normal | .028 | .081 | .018 | .050 | .009 | .042 | .021 | .065 | .009 | .046 | .011 | .052 | .455 | .588 | .391 | .518 | .312 | .482 | .384 | .544 | .257 | .434 | .376 | .551 |
| h normal | .069 | .126 | .024 | .061 | .011 | .048 | .025 | .063 | .015 | .051 | .016 | .058 | .263 | .364 | .202 | .284 | .181 | .270 | .182 | .270 | .156 | .245 | .218 | .323 |
| h cl normal | .070 | .124 | .023 | .048 | .009 | .041 | .025 | .059 | .013 | .049 | .015 | .055 | .190 | .273 | .127 | .186 | .102 | .177 | .121 | .184 | .100 | .174 | .137 | .228 |
| iid "actual" | .025 | .073 | .014 | .044 | .007 | .035 | .019 | .060 | .007 | .042 | .011 | .050 | .428 | .551 | .370 | .485 | .311 | .447 | .355 | .495 | .263 | .425 | .362 | .520 |
| h "actual" | .034 | .081 | .012 | .040 | .005 | .035 | .022 | .063 | .010 | .049 | .014 | .059 | .407 | .535 | .339 | .453 | .274 | .416 | .322 | .470 | .226 | .380 | .342 | .501 |
| h cl "actual" | .035 | .083 | .014 | .039 | .004 | .032 | .024 | .064 | .009 | .045 | .015 | .057 | .293 | .444 | .226 | .350 | .157 | .294 | .228 | .375 | .139 | .303 | .273 | .424 |
| 1st Stage F-tests (correlated errors): $H_0 = \boldsymbol{\pi}_{dgp}$ or $\mathbf{0}$ | | | | | | | | | | | | | | | | | | | | | | | | |
| iid normal | .051 | .119 | .023 | .073 | .008 | .054 | .017 | .065 | .010 | .053 | .012 | .056 | .925 | .950 | .894 | .933 | .848 | .924 | .855 | .912 | .833 | .915 | .858 | .921 |
| h normal | .085 | .162 | .034 | .081 | .020 | .081 | .015 | .059 | .018 | .072 | .017 | .065 | .759 | .825 | .693 | .772 | .688 | .787 | .547 | .655 | .699 | .790 | .668 | .758 |
| h cl normal | .091 | .171 | .030 | .078 | .023 | .088 | .012 | .056 | .023 | .076 | .017 | .065 | .647 | .729 | .562 | .658 | .571 | .680 | .434 | .551 | .576 | .683 | .540 | .636 |
| iid "actual" | .040 | .105 | .018 | .054 | .009 | .041 | .015 | .051 | .008 | .042 | .009 | .044 | .880 | .924 | .837 | .897 | .795 | .879 | .806 | .873 | .791 | .882 | .816 | .889 |
| h "actual" | .081 | .160 | .029 | .075 | .011 | .056 | .017 | .064 | .017 | .067 | .016 | .066 | .857 | .910 | .778 | .855 | .738 | .853 | .718 | .820 | .751 | .854 | .754 | .856 |
| h cl "actual" | .084 | .162 | .032 | .078 | .015 | .066 | .018 | .062 | .020 | .067 | .015 | .063 | .766 | .846 | .666 | .766 | .621 | .777 | .588 | .724 | .617 | .767 | .613 | .755 |
| Hausman tests: $H_0 = (\beta_{iv} = \beta_{ols})$ | | | | | | | | | | | | | | | | | | | | | | | | |
| | (uncorrelated errors) | | | | | | | | | | | | (correlated errors) | | | | | | | | | | | |
| iid normal | .021 | .071 | .008 | .036 | .005 | .030 | .008 | .041 | .006 | .044 | .010 | .050 | .373 | .493 | .255 | .373 | .216 | .354 | .262 | .405 | .266 | .408 | .306 | .450 |
| h normal | .065 | .145 | .011 | .047 | .006 | .036 | .007 | .035 | .013 | .051 | .014 | .068 | .268 | .378 | .153 | .211 | .147 | .216 | .146 | .202 | .158 | .242 | .191 | .279 |
| h cl normal | .076 | .157 | .008 | .029 | .003 | .025 | .005 | .025 | .011 | .050 | .020 | .070 | .210 | .316 | .098 | .147 | .089 | .139 | .085 | .138 | .103 | .178 | .135 | .219 |
| iid "actual" | .023 | .073 | .006 | .030 | .003 | .024 | .007 | .040 | .024 | .050 | .013 | .052 | .375 | .484 | .266 | .364 | .221 | .346 | .258 | .377 | .211 | .318 | .314 | .441 |
| h "actual" | .039 | .100 | .007 | .032 | .003 | .028 | .007 | .041 | .021 | .051 | .021 | .068 | .335 | .454 | .211 | .315 | .186 | .297 | .205 | .323 | .171 | .272 | .260 | .387 |
| h cl "actual" | .049 | .111 | .008 | .033 | .003 | .024 | .007 | .037 | .021 | .052 | .022 | .075 | .278 | .405 | .139 | .240 | .097 | .204 | .134 | .260 | .127 | .232 | .218 | .350 |

Notes: Average across 30 papers of the within paper average rejection rate. Bootstrap-t methods use clustered/robust covariance estimates.

**Table 11**
Significance of 2SLS Coefficients (average across papers of the fraction of coefficients rejecting the null of 0)

| | all results | | headline results | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | all | | low | | medium | | high | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| authors' methods | .365 | .558 | .522 | .788 | .617 | .867 | .604 | .754 | .344 | .742 |
| clustered/robust | .339 | .531 | .463 | .768 | .567 | .867 | .504 | .721 | .319 | .717 |
| jackknife | .250 | .401 | .382 | .537 | .517 | .767 | .367 | .467 | .262 | .376 |
| pairs bootstrap - c | .160 | .340 | .243 | .520 | .333 | .767 | .367 | .467 | .029 | .326 |
| pairs bootstrap - t | .247 | .453 | .308 | .599 | .367 | .767 | .217 | .442 | .340 | .588 |
| wild bootstrap - c | .115 | .337 | .231 | .444 | .300 | .700 | .379 | .504 | .014 | .129 |
| wild bootstrap - t | .346 | .535 | .512 | .719 | .667 | .967 | .517 | .688 | .351 | .502 |

Notes: Low/medium/high refer to papers divided by maximum leverage, as described in Table 2 earlier; bootstrap-t implemented using the clustered/robust covariance estimate; wild bootstrap using restricted efficient residuals; bootstrap p-values evaluated using 2000 draws.

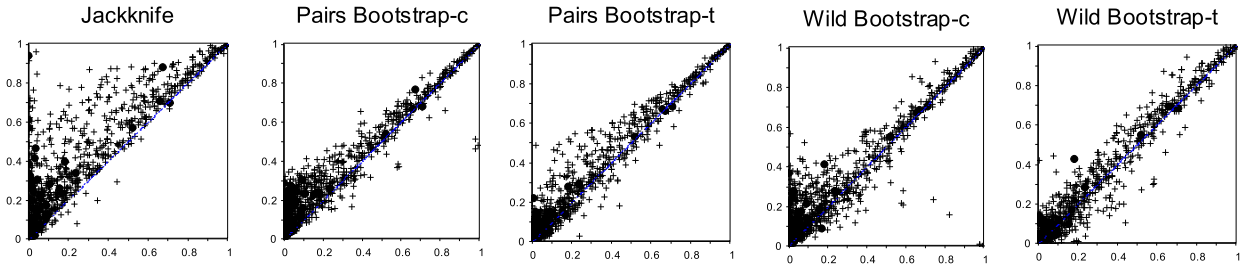


**Fig. 3.** Jackknife, Bootstrap & Clustered/Robust P-Values, Notes: X-axes = clustered/robust p-values, Y-axes = jackknife or bootstrap p-values. Solid circles = headline results, plus marks = other results.

consequently evince similar Type I error rates in simulation (e.g. Table 4 earlier). While headline results have higher first and second stage significance rates to begin with, in all tables that follow the proportional reduction in their statistical significance through the application of alternative inference methods is very similar to that found for all published results. As headline results are of particular interest to readers, further detail by leverage group is given for these. The same detail and patterns for all published results in the sample can be found in the on-line appendix.

Table 11 begins by evaluating the statistical significance of the coefficients of instrumented right-hand side variables. In the first row I report authors' p-values, using their covariance calculation methods (default or cl/robust) and chosen distribution (normal or t). The second row of the table moves things to a consistent framework, using cl/robust covariance matrices[26] and the finite sample t distribution throughout. All subsequent discussion is relative to this consistent benchmark. Fig. 3 graphs the alternative p-values against the benchmark cl/robust test of 2SLS significance.

Several patterns are apparent in the table and figure. First, while the application of the pairs and wild bootstrap-c lowers significance rates in all and headline results to ½ or less of those found using cl/robust methods at the .01 level, and $^2/_3$ or less of cl/robust methods at the .05 level, the corresponding adjustments using the pairs bootstrap-t are about .7 and .8, while the wild bootstrap-t largely leaves significance rates unchanged. As argued earlier, this large gap between -c and -t significance levels, which have similar size and power in simulations, is suggestive of publication bias. Papers do not report unusually large mean effects given the null; they report unusually large t-statistics. The role small standard errors relative to coefficient estimates[27] play in published significance is highlighted by the jackknife, which generates large changes in p-values (Fig. 3) simply by substituting an alternative standard error estimate. Notably, there is no systematic difference between bootstrap-c and -t results for OLS versions of these equations (on-line appendix), which do not form the basis for the publication decision.

Second, the differences between conventional and alternative significance rates reported in Table 11 are concentrated in medium and especially high leverage papers, where significance rates at the .01 level using the bootstrap-c are negligible and at the .05 level are substantially lower than cl/robust findings even when using the bootstrap-t. Regressions in the on-line appendix find that the

[26] I use the robust covariance estimate in a paper that used the homoskedastic estimate throughout, and cluster the sole regression that was left unclustered in a paper that otherwise clustered all other covariance estimates.

[27] In the simulations presented above, I find that IV standard error estimates are strongly positively correlated with the absolute value of the deviation of the coefficient estimate from the null (e.g. average correlations of .28 and .68 with iid normal or heteroskedastic & clustered normal errors, respectively, and .37 and .47 with iid and heteroskedastic & clustered "actual" errors). Consequently, it comes as no surprise that reported standard error estimates are not in the lowest percentiles of the bootstrapped distributions (averaging, for example, in the 49$^{th}$ percentile of the pairs bootstrap distribution of standard errors). They are, however, low given the magnitude of the coefficient estimates, as shown by the difference between -c and -t results in Table 11.

**Table 12**
Distribution of Alternative P-Values for Coefficients that are .05 Significant in 2SLS Clustered/Robust Tests of Instrumented Coefficients

| | jackknife | | pairs boot-c | | pairs boot-t | | wild boot-c | | wild boot-t | |
|---|---|---|---|---|---|---|---|---|---|---|
| | all | headline | all | headline | all | headline | all | headline | all | headline |
| < .05 | .671 | .664 | .569 | .644 | .730 | .750 | .565 | .535 | .878 | .881 |
| .05 - .10 | .161 | .120 | .183 | .125 | .187 | .151 | .164 | .209 | .095 | .060 |
| .10 - .20 | .084 | .111 | .153 | .151 | .071 | .093 | .169 | .123 | .024 | .059 |
| > .20 | .084 | .105 | .095 | .080 | .012 | .006 | .102 | .133 | .003 | .000 |

Note: average across papers of within paper distributions.

difference between cl/robust and alternative p-values are significantly related to maximum leverage, with greater effects when the p-value of the null that the errors are homoskedastic is low, which is consistent with the results found in simulations above. However, as the p-values of tests for homoskedasticity are close to 0 for ¾ of the sample, point estimates are imprecise and the coefficients on their interaction with leverage are not statistically significant when evaluated with the bootstrap.

Third, when alternative methods change a conventionally significant result, the change in the p-value is often substantive, as shown by the stacked observations at low conventional p-values (x-axis) in Fig. 3. Table 12 explores this further, reporting the distribution of alternative p-values for coefficients that are .05 significant using a cl/robust p-value. When, for either all results or headline results alone, a change in significance is recorded using a jackknifed or bootstrap-c p-value, at least half of the movement is beyond the .10 level. Thus, for example, while .183 of .05 significant cl/robust p-values lie between .05 and .10 when evaluated using the pairs bootstrap-c, an additional .25 (=.153 + .095) lie in the .10-.20 and .20+ groupings. P-value changes using the bootstrap-t are also often substantial, as shown in the table.

Published 2SLS coefficient estimates are imprecise and, for the most part, statistically indistinguishable from OLS results for the same parameters. As shown in Table 13, the conventional cl/robust .99 2SLS confidence interval contains the corresponding OLS point estimate for .870 of the regressions and .831 of the headline results of the typical paper. 95 confidence intervals are tighter, reducing these frequencies to .750 and .673, respectively. Jackknife and bootstrapped confidence intervals raise these proportions, but only in medium and high leverage papers. In the latter, at the .99 level coverage of the OLS point estimate approaches 1.0 with the bootstrap-c. These results are not a consequence of a close similarity between OLS and 2SLS point estimates. In the average paper .13 of headline 2SLS coefficient estimates are of the *opposite sign* of the OLS estimate for the same equation, while the absolute difference of the 2SLS and OLS point estimates is greater than 0.5 times the absolute value of the OLS point estimate in .73 of headline regressions and greater than 5.0 times that value in .24 of headline regressions. 2SLS and OLS point estimates often differ substantively, but statistically the IV estimator rejects the OLS value much less frequently.

The imprecision of 2SLS estimation carries over into an inability to provide statistical evidence that OLS is biased. Table 14 reports the Hausman test of OLS bias based upon the significance of the 1st stage residuals entered as regressors in OLS versions of the 2nd stage equation. The conventional cl/robust estimate rejects the null that OLS is unbiased .232 & .382 of the time at the .01 or .05 levels in the typical 2SLS regression, and somewhat more often, .309 & .445, in headline results. Jackknife and bootstrap methods lower these frequencies, down to an average of .172 and .319 at the .01 and .05 levels for headline results, with differences concentrated in medium and high leverage papers, where all bootstrap tests produce average .01 rejection rates of less than 3 percent. There may be theoretical reasons to believe that OLS estimates of parameters of interest in these papers are substantively biased, but 2SLS estimation is in most cases unable to provide strong empirical evidence to substantiate those beliefs.

Table 15 asks whether published 2SLS results are identified by testing the null that all first stage coefficients on the excluded exogenous variables are zero. Using the conventional test with the cl/robust covariance estimate, an average of .858 of 1st stage regressions in the typical paper reject the null of a rank zero first stage relation at the .01 level. This share falls to between .638 and .718, i.e. on average about .8 of the original level, using bootstrap and jackknife techniques. Headline results, which are often highlighted by authors on the basis of the first stage, start out much better, rejecting the null 100% of the time at the .01 level using cl/robust techniques, but on average suffer the exact same .8 proportional reduction in significance rates at the .01 level. Once again, differences are most pronounced in medium and high leverage papers, where bootstrap and jackknife rejection rates for headline results at the .01 level fall as low as .381. Once jackknife and bootstrap techniques are used to reduce, albeit not eliminate (Table 10), the dimensionally-increasing size distortions that appear with cl/robust covariance estimates and non-iid errors, 1/5 of all regressions which are singled out by authors as headline results, and about ½ of the same in high leverage papers, cannot present strong statistical evidence against the null that the instruments are irrelevant.

Table 16 brings the preceding results together, asking to what degree 2SLS credibly provides information that is statistically different from the biased results of OLS. Column (i) in the table reports the average fraction of 2SLS regressions that reject the null hypothesis that the IV regression is completely unidentified, a basic prerequisite for credibility, and *either* deliver point estimates that are statistically different from OLS *or* reject the null that OLS is unbiased. Using conventional cl/robust methods, only .234 of all results and .309 of headline results meet these criteria at the .01 level, these shares falling to an average of .129 and .188, respectively, when jackknife and bootstrap tests are used. Results are especially poor in high leverage papers where only .014 or .114 of headline 2SLS regressions meet these criteria using the jackknife or bootstrap. Lowering the bar by raising the level to .05 raises rejection rates, but only to an average across jackknife and bootstrap methods of .258 and .333 in all and headline results, respectively.

None of the preceding results validate OLS as an estimation and inference procedure for the problems considered in my sample papers. As noted early on in Table 4, precise estimates of biased parameters do not provide a sensible basis for statistical inference.

**Table 13**
Frequency with which IV Confidence Intervals contain OLS Point Estimates

| | all results | | headline results | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | all | | low | | medium | | high | |
| | .99 | .95 | .99 | .95 | .99 | .95 | .99 | .95 | .99 | .95 |
| clustered/robust | .870 | .750 | .831 | .673 | .900 | .800 | .800 | .475 | .793 | .743 |
| jackknife | .902 | .825 | .862 | .801 | .900 | .800 | .800 | .767 | .886 | .836 |
| pairs bootstrap - c | .934 | .852 | .895 | .790 | .900 | .700 | .800 | .800 | .986 | .871 |
| pairs bootstrap - t | .902 | .779 | .895 | .769 | .900 | .800 | .900 | .808 | .886 | .699 |
| wild bootstrap - c | .916 | .801 | .847 | .759 | .900 | .850 | .654 | .542 | .986 | .886 |
| wild bootstrap - t | .887 | .719 | .858 | .664 | .800 | .700 | .888 | .575 | .886 | .718 |

Notes: As in Table 11.

**Table 14**
Rejection Rates in Hausman Tests (tests of OLS bias)

| | all results | | headline results | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | all | | low | | medium | | high | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| clustered/robust | .232 | .382 | .309 | .445 | .250 | .400 | .446 | .638 | .232 | .296 |
| jackknife | .135 | .227 | .188 | .254 | .250 | .300 | .200 | .333 | .114 | .129 |
| pairs bootstrap - c | .098 | .200 | .138 | .249 | .200 | .400 | .200 | .233 | .014 | .114 |
| pairs bootstrap - t | .110 | .243 | .110 | .300 | .200 | .300 | .100 | .433 | .029 | .168 |
| wild bootstrap - c | .129 | .247 | .187 | .319 | .200 | .250 | .346 | .592 | .014 | .114 |
| wild bootstrap - t | .175 | .328 | .237 | .470 | .250 | .500 | .446 | .604 | .014 | .307 |

Note: Test of the significance of $\theta$ in the equation $\mathbf{y} = \mathbf{Y}\beta + \mathbf{X}\boldsymbol{\delta} + \widehat{\mathbf{v}}\theta + \mathbf{u}$, where $\widehat{\mathbf{v}} = \mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\pi}} - \mathbf{X}\widehat{\boldsymbol{\gamma}}$.

**Table 15**
Identification in the First-Stage (rejection rates in tests of instrument irrelevance)

| | all results | | headline results | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | all | | low | | medium | | high | |
| | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 | .01 | .05 |
| clustered/robust | .858 | .929 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| jackknife | .718 | .827 | .835 | .945 | 1.00 | 1.00 | .800 | .900 | .706 | .936 |
| pairs bootstrap - c | .661 | .874 | .781 | .967 | 1.00 | 1.00 | .800 | 1.00 | .542 | .900 |
| pairs bootstrap - t | .638 | .773 | .755 | .877 | 1.00 | 1.00 | .767 | .900 | .498 | .731 |
| wild bootstrap - c | .704 | .886 | .794 | .967 | 1.00 | 1.00 | .900 | 1.00 | .481 | .900 |
| wild bootstrap - t | .660 | .856 | .783 | .952 | 1.00 | 1.00 | .967 | 1.00 | .381 | .857 |

Notes: As in Table 11.

**Table 16**
Does 2SLS Provide Information that is Statistically Different from OLS? (average fraction of 2SLS regressions rejecting $\boldsymbol{\pi} = \mathbf{0}$ & $\beta_{ols} \in \text{CI}_{2sls}$ or $\beta_{ols}$ unbiased)

| | (i) at .01 level | | | | | (ii) at .05 level | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | all | headline results | | | | all | headline results | | | |
| | results | all | low | med | high | results | all | low | med | high |
| cl/robust | .234 | .309 | .250 | .446 | .232 | .378 | .445 | .400 | .638 | .296 |
| jackknife | .130 | .188 | .250 | .200 | .114 | .228 | .271 | .300 | .333 | .179 |
| pairs boot - c | .097 | .138 | .200 | .200 | .014 | .183 | .221 | .400 | .233 | .029 |
| pairs boot - t | .127 | .138 | .200 | .100 | .114 | .277 | .355 | .300 | .492 | .273 |
| wild boot - c | .116 | .187 | .200 | .346 | .014 | .249 | .319 | .250 | .592 | .114 |
| wild boot - t | .177 | .287 | .300 | .446 | .114 | .353 | .502 | .500 | .638 | .368 |

Notes: $\boldsymbol{\pi} = \mathbf{0} = 1^{st}$ stage coefficients on excluded instruments all equal 0; $\beta_{ols} \in \text{CI}_{2sls}$ = OLS point estimate in .99 or .95 2SLS confidence interval; $\beta_{ols}$ unbiased = Hausman test used in Table 4 earlier.

**Table 17**
Probability of 1st Stage F > 10 when Instruments are Irrelevant (1000 Monte Carlo simulations for each of 1309 equations)

| | default covariance estimate | | | | clustered/robust covariance estimate | | | |
|---|---|---|---|---|---|---|---|---|
| | all | low | medium | high | all | low | medium | high |
| iid normal | .001 | .001 | .002 | .001 | .012 | .002 | .010 | .023 |
| h normal | .223 | .193 | .128 | .349 | .042 | .004 | .015 | .107 |
| h & cl normal | .416 | .435 | .209 | .604 | .047 | .005 | .018 | .119 |
| iid "actual" | .004 | .005 | .002 | .003 | .011 | .006 | .007 | .021 |
| h "actual" | .057 | .036 | .058 | .078 | .030 | .002 | .031 | .056 |
| h & cl "actual" | .198 | .193 | .074 | .328 | .033 | .003 | .034 | .061 |

Notes: Average across papers of within paper average rejection rates; low, medium and high divide the sample into thirds, based upon average maximum leverage, as in Table 2 earlier.

What the results above do show, however, is that unbalanced regression design, non-iid error processes and the inherent inefficiency of 2SLS have interacted to create a published literature which fundamentally has very low power. While published 2SLS results often differ dramatically in sign and magnitude from their OLS counterparts, they are not actually statistically very informative. This does not validate OLS point estimates, but it does show that much less has been learnt than might otherwise be thought.

## 7. Conclusion

Contemporary IV practice involves the screening of reported results on the basis of the 1st 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. Table 17 reports the Monte Carlo probability of an F greater than 10 in tests of true nulls in my sample. Following Stock & Yogo's (2005) asymptotic iid based theory, an F of 10 became an important benchmark in the profession. As shown in the table, in an ideal iid normal world, using the appropriate homoskedastic/default covariance estimate, the probability of an F greater than 10 arising when the instruments are completely irrelevant is a 1 or 2 in 1000 event, whether leverage is low, medium or high. However, with clustered and heteroskedastic errors, in high leverage papers the probability of an F greater than 10 rises to 30 or 60 percent, depending upon the error process, and is still very substantial when the default covariance estimate is replaced with its clustered/robust counterpart. A benchmark F of 10, used for years by the profession, ensured that regressions in which the instruments were utterly irrelevant would regularly pass as having strong 1st stage relations. The adoption of more demanding cl/robust standards, such as that of the Olea-Pflueger test, will screen out most unidentified regressions,[28] but, as shown in the simulations above, will not guarantee the protection against relative bias sought of the test. More generally, the adoption of any one-size-fits-all standards based upon zero maximum leverage asymptotic theory selects in favour of the worst finite sample regression design, where the fixed standards have no predictive value and 2SLS is at its very worst.

This paper has highlighted a number of ways in which current practice might be improved. The reporting of the number of clusters for each regression in a table, now rarely done, is an easy starting point. Delete-one sensitivity, of t-statistics not coefficients, highlights the degree to which significant results depend upon sensitive coefficient and standard error estimates. The maximum leverage share of one cluster provides a measure of the degree to which regression design has small sample characteristics and an appeal to asymptotic theorems is less compelling. The bootstrap provides improved null rejection probabilities in a variety of tests across a range of regression designs and disturbance characteristics, although its use must still be tempered by a consideration of the pathologies that may arise in unidentified regressions. The use of the considerable talents of econometricians to develop additional methods which adjust for finite sample regression design would be an enormous boon to the profession.

Economists use 2SLS because they wish to gain a more accurate estimate of parameters of interest than is provided by biased OLS. In this regard, explicit consideration of the degree to which 2SLS results are distinguishable from OLS seems natural, a point raised early on by Sargan (1958) in his seminal paper. In the analysis of the sample, above, I find that 2SLS rarely rejects the OLS point estimate or is able to provide strong statistical evidence against OLS being unbiased, despite the fact that 2SLS point estimates are often of the opposite sign or substantially different magnitude. This is virtually always true in high leverage papers, but is even true in the low leverage sample, where on average only .23 or .35 of headline results are able to either reject the null of zero OLS bias or exclude OLS point estimates at the .01 or .05 levels in bootstrap and jackknife tests. These results need not heighten confidence in OLS point estimates, as the simulations in this paper show that heteroskedastic clustered disturbances systematically lower power, especially in 2SLS estimation; but they do show that in practical application 2SLS is sufficiently inefficient that it does not often provide meaningful information regarding the degree to which OLS point estimates are biased. Learning about the world may simply be harder than suggested by simple dichotomies between good and bad research design.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2022.104112.

[28] Thus, for example, the probabilities of a cl/robust F greater than 40 with clustered heteroskedastic normal and "actual" errors in high leverage papers when the regression is unidentified are only .016 and .0013, respectively

## References


Angrist, Joshua D., Pischke, Jörn-Steffen, 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.

Bound, John, Jaeger, David A., Baker, Regina M., 1995. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. Journal of the American Statistical Association 90 (June), 443–450.

Davidson, Russell, MacKinnon, James G., 2008. Bootstrap inference in a linear equation estimated by instrumental variables. Econometrics Journal 11 (3), 443–477.

Dufour, Jean-Marie., 1997. Some Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models. Econometrica 65 (6), 1365–1387.

Eicker, F, 1963. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. Annals of Mathematical Statistics 34 (2), 447–456.

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.

Gleser, Leon Jay, Hwang, Jiunn T, 1987. The Nonexistence of 100(1-α)% Confidence Sets of Finite Expected Diameter in Errors-in-Variables and Related Models. Annals of Statistics 15 (4), 1351–1362.

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

Hausman, Jerry A., 1978. Specification Tests in Econometrics. Econometrica 46 (6), 1251–1271.

Hinkley, David V., 1977. Jackknifing in Unbalanced Situations. Technometrics 19 (3), 285–292.

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

Koenker, Roger., 1981. A Note on Studentizing a Test for Heteroskedasticity. Journal of Econometrics 17, 107–112.

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

MacKinnon, James G., White, Halbert, 1985. Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. Journal of Econometrics 29, 305–325.

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

Olea, Jose Luis Montiel, Pflueger, Carolin, 2013. A Robust Test for Weak Instruments. Journal of Business and Economic Statistics 31 (3), 358–369.

Sargan, J.D., 1958. The Estimation of Economic Relationships using Instrumental Variables. Econometrica 26 (3), 393–415.

Staiger, Douglas, Stock, James H., 1997. Instrumental Variables Regression with Weak Instruments. Econometrica 65 (3), 557–586.

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

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

White, Halbert., 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 48 (4), 817–838.

Wooldridge, J.M., 2013. Introductory Econometrics: A Modern Approach, 5th ed. Mason, OH: South-Western.