Weak Instruments in IV Regression: Theory and Practice

Isaiah Andrews, James Stock, and Liyang Sun*

November 20, 2018

Abstract

When instruments are weakly correlated with endogenous regressors, conventional methods for instrumental variables estimation and inference become unreliable. A large literature in econometrics develops procedures for detecting weak instruments and constructing robust confidence sets, but many of the results in this literature are limited to settings with independent and homoskedastic data, while data encountered in practice frequently violate these assumptions. We review the literature on weak instruments in linear IV regression with an emphasis on results for non-homoskedastic (heteroskedastic, serially correlated, or clustered) data. To assess the practical importance of weak instruments, we also report tabulations and simulations based on a survey of papers published in the American Economic Review from 2014 to 2018 that use instrumental variables. These results suggest that weak instruments remain an important issue for empirical practice, and that there are simple steps researchers can take to better handle weak instruments in applications.

Keywords: weak instruments, heteroskedasticity, F-statistic

1 Introduction

In instrumental variables (IV) regression, the instruments are called weak if their correlation with the endogenous regressors, conditional on any controls,

*Andrews and Stock, Department of Economics, Harvard University, Cambridge, MA, 02138. Sun, Department of Economics, MIT, Cambridge, MA, 02139.

is close to zero. When this correlation is sufficiently small, conventional approximations to the distribution of IV estimators, such as two-stage least squares, are generally unreliable. In particular, IV estimators can be badly biased, while t-tests may fail to control size and conventional IV confidence intervals may cover the true parameter value far less often than we intend.

A recognition of this problem has led to a great deal of work on econometric methods applicable to models with weak instruments. Much of this work, especially early in the literature, focused on the case where the data are independent and the errors in the reduced-form and first-stage regressions are homoskedastic. Homoskedasticity implies that the variance matrix for the reduced-form and first-stage regression estimates can be written as a Kronecker product, which substantially simplifies the analysis of many procedures. As a result, there are now strong theoretical results for both detection of weak instruments and construction of identification-robust confidence sets in the homoskedastic case.

More recently, much of the theoretical literature on weak instruments has considered the more difficult case where the data may be dependent and/or the errors heteroskedastic. In this setting, which we refer to as the non-homoskedastic case, the variance of the reduced-form and first-stage estimates no longer has Kronecker product structure in general, rendering results based on such structure inapplicable. Because homoskedasticity is rarely a plausible assumption in practice, procedures applicable to the non-homoskedastic case have substantial practical value.

This survey focuses on the effects of weak instruments in the non-homoskedastic case. We concentrate on detection of weak instruments and weak-instrument robust inference. The problem of detection is relevant because weak-instrument robust methods can be more complicated to use than standard two-stage least squares, so if instruments are plausibly strong it is convenient to report two-stage least squares estimates and standard errors. If instruments are weak, on the other hand, then practitioners are advised to use weak-instrument robust methods for inference, the second topic of this survey. We do not survey estimation, an area in which less theoretical progress has been made.[1]

[1]Two notable exceptions are Hirano & Porter (2015) and I. Andrews & Armstrong (2017). Hirano & Porter (2015)'s contribution is a negative result: they prove that, if one includes the possibility that instruments can be arbitrarily weak, then no unbiased estimator of the coefficient of interest exists without further restrictions. I. Andrews & Armstrong (2017) show, however, that if one imposes correctly the sign of the first-stage

In addition to surveying the theoretical econometrics literature, we examine the role of weak instruments in empirical practice using a sample of 230 specifications gathered from 17 papers published in the American Economic Review (AER) from 2014-2018 that use the word "instrument" in their abstract. For details on the sample of papers, please see the online appendix. We use this sample for two purposes. The first is to learn what empirical researchers are actually doing when it comes to detecting and handling weak instruments. The second is to develop a collection of specifications that we use to assess the importance of weak instrument issues and the performance of weak instrument methods in data generating processes reflective of real-world settings.

Figure 1 displays a histogram of first-stage F-statistics reported in specifications in our AER sample with a single endogenous regressor, truncated above at 50 for visibility. The first-stage F-statistic for testing the hypothesis that the instruments are unrelated to the endogenous regressor is a standard measure of the strength of the instrument. Many of the first-stage F-statistics in our AER sample are in a range that, based on simulations and theoretical results, raise concerns about weak instruments, including many values less than 10. This suggests that weak instruments are frequently encountered in practice.

Another noteworthy feature of the data underlying Figure 1 is that 15 of the 17 papers in our sample reported at least one first-stage F-statistic. Evidently and reassuringly, there is widespread recognition by empirical economists that one needs to be attentive to the potential problems caused by weak instruments. This said, our review of these papers leads us to conclude that there is room for improving current empirical methods.

Specifically, in the leading case with a single endogenous regressor, we recommend that researchers judge instrument strength based on the effective F-statistic of Montiel Olea & Pflueger (2013). If there is only a single instrument, we recommend reporting identification-robust Anderson-Rubin confidence intervals. These are efficient regardless of the strength of the instruments, and so should be reported regardless of the value of the first stage F. Finally, if there are multiple instruments, the literature has not yet converged on a single procedure, but we recommend choosing from among the several available robust procedures that are efficient when the instruments

regression coefficient then asymptotically unbiased estimation is possible, and they derive unbiased estimators.

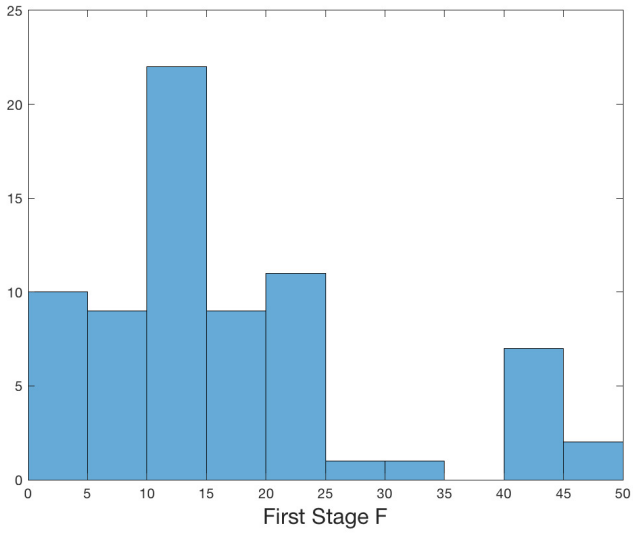

Figure 1: Distribution of reported first-stage F-statistics (and their non-homoskedastic generalizations) in 72 specifications with a single endogenous regressor and first-stage F smaller than 50. Total number of single endogenous regressor specifications reporting F-statistics is 108.

are strong.

The paper is organized as follows. Section 2 lays out the instrumental variables model and notation. Section 3 describes the weak instruments problem. Section 4 reviews methods for detecting weak instruments, Section 5 reviews weak-instrument robust inference, and Section 6 concludes with a discussion of open questions in the literature on weak instruments. In the online appendix we discuss our AER sample, the details of our simulation designs, and available Stata implementations of the procedures we discuss in the main text.

2 The Instrumental Variables Model

We study the linear instrumental variables (IV) model with a scalar outcome Y_i, a $p \times 1$ vector of potentially endogenous regressors X_i, a $k \times 1$ vector of instrumental variables Z_i, and an $r \times 1$ vector of exogenous regressors W_i.

This yields the linear constant effects instrumental variables model

$$Y_i = X_i'\beta + W_i'\kappa + \varepsilon_i, \tag{1}$$

$$X_i' = Z_i'\pi + W_i'\gamma + V_i, \tag{2}$$

where $E[Z_i\varepsilon_i] = 0$, $E[Z_iV_i'] = 0$, $E[W_i\varepsilon_i] = 0$, and $E[W_iV_i'] = 0$. We are interested in β, but X_i is potentially endogenous in the sense that we may have $E[\varepsilon_iV_i] \neq 0$. Consequently we may have $E[X_i\varepsilon_i] \neq 0$, so regression of Y_i on X_i and W_i may yield biased estimates. This model nests a wide variety of IV specifications encountered in practice. We allow the possibility that the errors (ε_i, V_i) are conditionally heteroskedastic given the exogenous variables (Z_i, W_i), so $E[(\varepsilon_i, V_i')'(\varepsilon_i, V_i')|Z_i, W_i]$ may depend on (Z_i, W_i). We further allow the possibility that (Y_i, X_i, Z_i, W_i) are dependent across i, for example due to clustering or time-series correlation. Finally, the results we discuss generalize to the case where the data are non-identically distributed across i, though for simplicity of notation we do not pursue this extension.

Substituting for X_i in (1), we obtain the equation

$$Y_i = Z_i'\delta + W_i'\tau + U_i \tag{3}$$

with $\delta = \pi\beta$. In a common abuse of terminology, we will refer to (1) as the structural form, (2) as the first-stage, and (3) as the reduced-form (for the older meaning of these terms, see e.g. Hausman (1983)). We can equivalently express the model as (1)-(2) or as (2)-(3), since each set of equations is an invertible linear transformation of the other. Likewise, the errors $(U_i, V_i) = (\varepsilon_i + \beta V_i, V_i)$ are an invertible linear transformation of (ε_i, V_i).

For ease of exposition we focus primarily on the case with a scalar endogenous regressor X_i, and so assume $p = 1$ unless noted otherwise. In our AER sample 211 of the 230 specifications have $p = 1$, so this appears to be the leading case in practice. Further, for most of this section we assume that the instruments Z_i are orthogonal to the control variables W_i, and so drop the controls from our notation. We discuss how to handle non-orthogonal control variables at the end of this section.

In this survey, we focus on estimators and tests that are functions of the reduced-form least squares coefficient $\hat{\delta}$, the first-stage least squares coefficient $\hat{\pi}$, and matrices that can be consistently estimated from the first-stage and reduced-form (e.g. variance and weighting matrices). Estimators in this class include two-stage least squares, which for $\hat{Q}_{ZZ} = \frac{1}{n}\sum Z_iZ_i'$ can be

written as

$$\hat{\beta}_{2SLS} = \left(\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}\right)^{-1}\hat{\pi}'\hat{Q}_{ZZ}\hat{\delta}, \tag{4}$$

as well as efficient-two-step GMM $\hat{\beta}_{2SGMM} = \left(\hat{\pi}'\hat{\Omega}\left(\hat{\beta}^1\right)^{-1}\hat{\pi}\right)^{-1}\hat{\pi}'\hat{\Omega}\left(\hat{\beta}^1\right)^{-1}\hat{\delta}$, for $\hat{\Omega}(\beta)$ an estimator for the variance of $\hat{\delta}-\hat{\pi}\beta$ and $\hat{\beta}^1$ a first-step estimator. Limited information maximum likelihood and continuously updated GMM likewise fall into this class.

Under mild regularity conditions (and, in the time-series case, stationarity), $(\hat{\delta}, \hat{\pi})$ are consistent and asymptotically normal in the sense that

$$\sqrt{n}\begin{pmatrix} \hat{\delta}-\delta \\ \hat{\pi}-\pi \end{pmatrix} \to_d N\left(0, \Sigma^*\right) \tag{5}$$

for

$$\Sigma^* = \begin{pmatrix} \Sigma^*_{\delta\delta} & \Sigma^*_{\delta\pi} \\ \Sigma^*_{\pi\delta} & \Sigma^*_{\pi\pi} \end{pmatrix} = \begin{pmatrix} Q_{ZZ}^{-1} & 0 \\ 0 & Q_{ZZ}^{-1} \end{pmatrix}\Lambda^*\begin{pmatrix} Q_{ZZ}^{-1} & 0 \\ 0 & Q_{ZZ}^{-1} \end{pmatrix}$$

where $Q_{ZZ} = E[Z_iZ_i']$ and

$$\Lambda^* = \lim_{n\to\infty} Var\left(\left(\frac{1}{\sqrt{n}}\sum_i U_iZ_i', \frac{1}{\sqrt{n}}\sum_i V_iZ_i'\right)'\right).$$

Hence, the asymptotic variance of $\sqrt{n}(\hat{\delta}-\delta, \hat{\pi}-\pi)$ has the usual sandwich form. Under standard assumptions the sample-analog estimator $\hat{Q}_{ZZ}$ will be consistent for Q_{ZZ}, and we can construct consistent estimators $\hat{\Lambda}^*$ for Λ^*. These results imply the usual asymptotic properties for IV estimators. For example, assuming the constant-effect IV model is correctly specified (so $\delta = \pi\beta$) and π is fixed and nonzero, the delta method together with (5) implies that $\sqrt{n}(\hat{\beta}_{2SLS}-\beta) \to_d N\left(0, \Sigma^*_{\beta,2SLS}\right)$ for $\Sigma^*_{\beta,2SLS}$ consistently estimable. We can likewise use (5) to derive the asymptotic distribution for limited information maximum likelihood as well as for two-step and continuously updated GMM.

Homoskedastic and Non-Homoskedastic Cases A central distinction in the literature on weak instruments, and in the historical literature on IV more broadly, is between what we term the homoskedastic and non-homoskedastic cases. In the homoskedastic case, we assume that the data (Y_i, X_i, Z_i, W_i) are iid across i and the errors (U_i, V_i) are homoskedastic, so

$E[(U_i, V_i')'(U_i, V_i')|Z_i, W_i]$ does not depend on (Z_i, W_i). Whenever these conditions fail, whether due to heteroskedasticity or dependence (e.g. clustering or time-series dependence), we will say we are in the non-homoskedastic case.

Two-stage least squares is efficient in the homoskedastic case but not, in general, in the non-homoskedastic case. Whether homoskedasticity holds also determines the structure of Λ^*. Specifically, in the homoskedastic case we can write

$$\Lambda^* = E\left[\begin{pmatrix} U_i^2 & U_iV_i \\ U_iV_i & V_i^2 \end{pmatrix} \otimes (Z_iZ_i')\right] = E\left[\begin{pmatrix} U_i^2 & U_iV_i \\ U_iV_i & V_i^2 \end{pmatrix}\right] \otimes Q_{ZZ}$$

where the first equality follows from the assumption of iid data, while the second follows from homoskedasticity. Hence, in homoskedastic settings the variance matrix Ω^* can be written as the Kronecker product of a 2×2 matrix that depends on the errors with a $k \times k$ matrix that depends on the instruments. The matrix Σ^* inherits the same structure, which as we note below simplifies a number of calculations. By contrast, in the non-homoskedastic case Σ^* does not in general have Kronecker product structure, rendering these simplifications inapplicable.

Dealing with Control Variables If the controls W_i are not orthogonal to the instruments Z_i, we need to take them into account. In this more general case, let us define $(\hat{\delta}, \hat{\pi})$ as the coefficients on Z_i from the reduced-form and first-stage regressions of Y_i and X_i, respectively, on (Z_i, W_i). By the Frisch-Waugh theorem these are the same as the coefficients from regressing Y_i and X_i on $Z_i^\perp$, the part of Z_i orthogonal to W_i. One can likewise derive estimators for the asymptotic variance matrix Σ^* in terms of $Z_i^\perp$ and suitably defined regression residuals. Such estimators, however, necessarily depend on the assumptions imposed on the data generating process (for example whether we allow heteroskedasticity, clustering, or time-series dependence).

A simple way to estimate Σ^* in practice when there are control variables is to jointly estimate $(\hat{\delta}, \hat{\pi})$ in a seemingly unrelated regression with whatever specification one would otherwise use (including fixed effects, clustering or serial-correlation robust standard errors, and so on). Appropriate estimates of Σ^* are then generated automatically by standard statistical software.

3 The Weak Instruments Problem

Motivated by the asymptotic approximation (5), let us consider the case where the reduced-form and first-stage regression coefficients are jointly normal

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} \delta \\ \pi \end{pmatrix}, \Sigma\right) \tag{6}$$

with $\Sigma = \frac{1}{n}\Sigma^*$ known (and, for ease of exposition, full-rank). Effectively, (6) discards the approximation error in (5) as well as the estimation error in $\hat{\Sigma}^*$ to obtain a finite-sample normal model with known variance. This suppresses any complications arising from non-normality of the OLS estimates or difficulties with estimating Σ and focuses attention solely on the weak instruments problem. Correspondingly, results derived in the model (6) will provide a good approximation to behavior in applications where the normal approximation to the distribution of $(\hat{\delta}, \hat{\pi})$ is accurate and Σ is well-estimated. By contrast, in settings where the normal approximation is problematic or $\hat{\Sigma}$ is a poor estimate of Σ results derived based on (6) will be less reliable (see Section 6 below, and Young (2018)).

Since the IV model implies that $\delta = \pi\beta$, the IV coefficient is simply the constant of proportionality between the reduced-form coefficient δ and the first-stage parameter π. In the just-identified setting matters simplify further, with the IV coefficient becoming $\beta = \delta/\pi$, and the usual IV estimators, including two-stage least squares and GMM, simplifying to $\hat{\beta} = \hat{\delta}/\hat{\pi}$. Just-identified specifications with a single endogenous variable constitute a substantial fraction of the specifications in our AER sample (101 out of 230), highlighting the importance of this case in practice.

It has long been understood (see e.g. Fieller (1954)) that ratio estimators like $\hat{\beta}$ can behave badly when the denominator is close to zero. The weak instruments problem is simply the generalization of this issue to potentially multidimensional settings. In particular, when the first-stage coefficient π is close to zero relative to the sampling variability of $\hat{\pi}$, the normal approximations to the distribution of IV estimates discussed in the last section may be quite poor. Nelson & Startz (1990a) and Nelson & Startz (1990b) provided early simulation demonstrations of this issue, while Bound et al. (1995) found similar issues in simulations based on Angrist & Krueger (1991).

The usual normal approximation to the distribution of $\hat{\beta}$ can be derived by linearizing $\hat{\beta}$ in $(\hat{\delta}, \hat{\pi})$. Under this linear approximation, normality of $(\hat{\delta}, \hat{\pi})$

implies approximate normality of $\hat{\beta}$. This normal approximation fails in settings with weak instruments because $\hat{\beta}$ is highly nonlinear in $\hat{\pi}$ when the latter is close to zero. As a result, normality of $(\hat{\delta}, \hat{\pi})$ does not imply approximate normality of $\hat{\beta}$. Specifically, the IV coefficient $\hat{\beta} = \hat{\delta}/\hat{\pi}$ is distributed as the ratio of potentially correlated normals, and so is non-normal. If π is large relative to the standard error of $\hat{\pi}$, however, then $\hat{\pi}$ falls close to zero with only very low probability and the nonlinearity of $\hat{\beta}$ in $(\hat{\delta}, \hat{\pi})$ ceases to matter. Hence, we see that non-normality of the instrumental variables estimate arises when the first-stage parameter π is small relative to its sampling variability. The same issue arises in the overidentified case with $p = 1 < k$, where the weak instruments problem arises when the $k \times 1$ vector π of first-stage coefficients is close to zero relative to the variance of $\hat{\pi}$. Likewise, in the general $1 \leq p \leq k$ case, the weak instruments problem arises when the $k \times p$ matrix π of first-stage coefficients is close to having reduced rank relative to the sampling variability of $\hat{\pi}$.

Failure of the Bootstrap A natural suggestion for settings where conventional asymptotic approximations fail is the bootstrap. Unfortunately, the bootstrap (and its generalizations, including subsampling and the m-out-of-n bootstrap) do not in general resolve weak instruments issues. See D. Andrews & Guggenberger (2009). For intuition, note that we can view the bootstrap as simulating data based on estimates of the data generating process. In the model (6), the worst case for identification is $\pi = 0$, since in this case β is totally unidentified. In the normal model (6), however, we never estimate π perfectly, and in particular estimate $\hat{\pi} = 0$ with probability zero. Hence, the bootstrap incorrectly "thinks" β is identified with probability one. More broadly, the bootstrap can make systematic errors in estimating the strength of the instruments, which suggests why it can yield unreliable results. None of the IV specifications in our AER sample used the bootstrap.

Motivation of the Normal Model The normal model (6) has multiple antecedents. A number of papers in the early econometric literature on simultaneous equations assumed fixed instruments and exogenous variables along with normal errors, which leads to the homoskedastic version of (6), sometimes with Σ unknown (Anderson & Rubin, 1949; Sawa, 1969; Mariano & Sawa, 1972).

More recently, a number of papers in the literature on weak instruments

including Kleibergen (2002), Moreira (2003), D. Andrews et al. (2006), and Moreira & Moreira (2015) derive results in the normal model (6), sometimes with the additional assumption that the underlying data are normal. While here we have motivated the normal model (6) heuristically based on the asymptotic normality (5) of the reduced-form and first-stage estimates, this connection is made precise elsewhere in the literature. Staiger & Stock (1997) show that the normal model (6) arises as an approximation to the distribution of the scaled reduced-form and first-stage regression coefficients under weak-instrument asymptotics where first-stage shrinks at a $\sqrt{n}$ rate. As discussed in Staiger & Stock (1997), these asymptotics are intended to capture situations in which the true value of the first-stage is on the same order as sampling uncertainty in $\hat{\pi}$, so issues associated with small π cannot be ignored. Finite sample results for the model (6) then translate to weak-instrument asymptotic results via the continuous mapping theorem. Many other authors including Kleibergen (2005), D. Andrews et al. (2006), I. Andrews (2016), and I. Andrews & Armstrong (2017) have built on these results to prove validity for particular procedures under weak-instrument asymptotics.

More recently D. Andrews & Guggenberger (2015), I. Andrews & Mikusheva (2016), D. Andrews & Guggenberger (2017), D. Andrews (2018), and D. Andrews et al. (2018a) have considered asymptotic validity uniformly over values of the first-stage parameter π and distributions for (U_i, V_i, W_i, Z_i). These authors show that some, though not all, procedures derived in the normal model (6) are also uniformly asymptotically valid, in the sense that e.g. the probability of incorrectly rejecting true null hypotheses converges to the nominal size uniformly over a large class of data generating processes as the sample size increases. D. Andrews et al. (2018a) discuss general techniques to establishing uniform asymptotic validity, but the argument for a given procedure is case-specific. Hence, in this review we focus on the normal model (6) which unites much of the weak-instruments literature, and refer readers interested in questions of uniformity to the papers cited above.

Simulated Distribution of t-Statistics While we know from theory that weak instruments can invalidate conventional inference procedures, whether weak instruments are a problem in a given application is necessarily case-specific. To examine the practical importance of weak instruments in recent applications of instrumental variables methods, we report simulation results

calibrated to our AER sample.

Specifically, we calibrate the normal model (6) to each of the 124 specifications in the sample for which we can estimate the full variance matrix Σ of the reduced-form and first-stage estimates, based either on results reported in the paper or replication files. We drop four specifications where our estimate of Σ is not positive definite. It happens to be the case that all remaining specifications have only a single endogenous regressor ($p = 1$). Hence, our simulation results only address this case. In each specification, we set the first-stage parameter π to the estimate $\hat{\pi}$ in the data, and set δ to $\hat{\pi}\hat{\beta}_{2SLS}$, the product of the first-stage with the two-stage least squares estimates. We set Σ equal to the estimated variance matrix for $(\hat{\delta}, \hat{\pi})$, maintaining whatever assumptions were used by the original authors (including the same controls, clustering at the same level, and so on).

In each specification we repeatedly draw first-stage and reduced-form parameter estimates $(\hat{\delta}^*, \hat{\pi}^*)$ and for each draw calculate the two-stage least squares estimate, along with the t-statistic for testing the true value of β (that is, the value used to simulate the data). In the left panels of Figures 2 and 3, we plot the median t-statistic and the frequency with which nominal 5% two-sided t-tests reject on the vertical axis, and the average of the effective F-statistic of Montiel Olea & Pflueger (2013), which we introduce in the next section, on the horizontal axis. This statistic is equivalent to the conventional first-stage F-statistic for testing $\pi = 0$ in models with homoskedastic errors, but adds a multiplicative correction in models with non-homoskedastic errors. For visibility, we limit attention to the 106 out of 124 specifications where there average first-stage F-statistic is smaller than 50 (the remaining specifications exhibit behavior very close to those with F-statistics between 40 and 50).

Several points emerge clearly from these results. First, there are a non-trivial number of specifications with small first-stage F-statistics (e.g. below 10, the rule of thumb cutoff for weak instruments proposed by Staiger & Stock (1997)) in the AER data. Second, even for specifications with essentially the same first-stage F-statistic, the median t-statistic and the size of nominal 5% t-tests can vary substantially due to other features (for example the true value β and the matrix Σ). Third, we see that among specifications with a small average F-statistic, behavior can deviate substantially from what we would predict under conventional (strong-instrument) asymptotic approximations. Specifically, conventional approximations imply that the median t-statistic is zero and 5% t-tests should reject 5% of the time. In

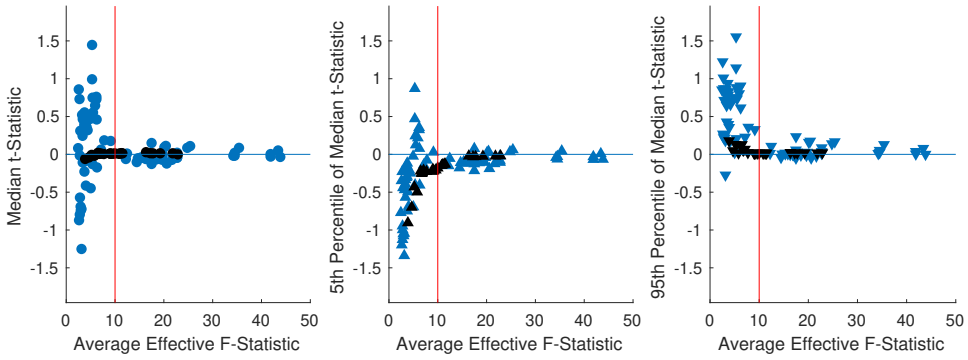

Figure 2: Median of t-statistic for testing true value of β plotted against the average effective F-statistic of Montiel Olea & Pflueger (2013) in calibrations to AER data, limited to the 106 out of 124 specifications with average F smaller than 50. Just-identified specifications are plotted in black, while over-identified specifications are in blue. Left panel plots median at parameter values estimated from AER data, while middle and right panels plot, respectively, the 5th and 95th percentiles of the median t-statistic under the Bayesian exercise described in the text. Red line corresponds to a first-stage F of 10.

our simulations, by contrast, we see that the median t-statistic sometimes has absolute value larger than one, while the size of 5% t-tests can exceed 30%. These issues largely disappear among specifications where the average F-statistic exceeds 10, and in these cases conventional approximations appear to be more accurate.

These results suggest that weak-instrument issues are relevant for modern applications of instrumental variables methods. It is worth emphasizing that these simulations are based on the normal model (6) with known variance Σ, so these results arise from the weak instruments problem alone and not from e.g. non-normality of $(\hat{\delta}, \hat{\pi})$ or difficulties estimating the variance matrix Σ.

These results are sensitive to the parameter values considered (indeed, this is the reason the bootstrap fails). Since we estimate (β, π) with error, it is useful to quantify the uncertainty around our estimates for the median t-statistic and the size of t-tests. To do so, we adopt a Bayesian approach consistent with the normal model (6), and simulate a posterior distribution for the median t-statistic and the size of 5% t-tests. Specifically, we calculate the posterior distribution on (δ, π) after observing $(\hat{\delta}, \hat{\pi})$ using the normal

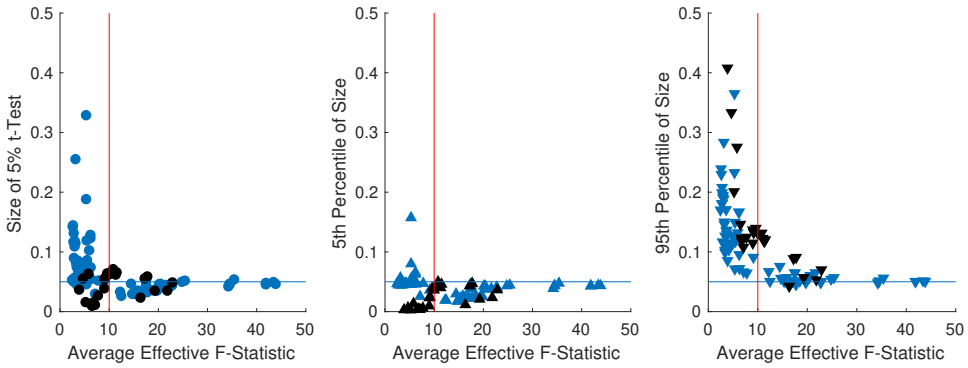

Figure 3: Rejection probability for nominal 5% two-sided t-tests plotted against the average effective F-statistic of Montiel Olea & Pflueger (2013) in calibrations to AER data, limited to the 106 out of 124 specifications with average F smaller than 50. Just-identified specifications are plotted in black, while over-identified specifications are in blue. Left panel plots size at parameter values estimated from AER data, while middle and right panels plot, respectively, the 5th and 95th percentiles of the size under the Bayesian exercise described in the text. Red line corresponds to a first-stage F of 10.

likelihood from (6) and a flat prior. We draw values

$$\begin{pmatrix} \tilde{\delta} \\ \tilde{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix}, \Sigma \right)$$

for the reduced-form and first-stage parameters from this posterior, calculate the implied two-stage least squares coefficient $\tilde{\beta}$, and repeat our simulations taking $(\tilde{\beta}, \tilde{\pi})$ to be the true parameter values (setting the reduced-form coefficient to $\tilde{\pi}\tilde{\beta}$). The middle panels of Figures 2 and 3 report the 5th percentiles of the median t-statistic and size, respectively, across draws $(\tilde{\beta}, \tilde{\pi})$, while the right panels report the 95th percentiles. As these results suggest, there is considerable uncertainty about the distribution of t-statistics in these applications. As in our baseline simulations, however, poor performance for conventional approximations is largely, though not exclusively, limited to specifications where the average effective F-statistic is smaller than 10.

Finally, it is interesting to consider behavior when we limit attention to the subset of specifications that are just-identified (i.e. that have $k = 1$), which are plotted in black in Figures 2 and 3. Interestingly, when we simulate behavior at parameter estimates from the AER data in these cases, we

find that the largest absolute median t-statistic is 0.06, while the maximal size for a 5% t-test is just 7.1%. If, on the other hand, we consider the bounds from our Bayesian approach, the worst-case absolute median t-statistic is 0.9 while the worst-case size for the t-test is over 40%. Hence, t-statistics appear to behave much better in just-identified specifications when we consider simulations based on the estimated parameters, but this is no longer the case once we incorporate uncertainty about the parameter values.

4 Detecting Weak Instruments

The simulation results in the last section suggest that weak instruments may render conventional estimates and tests unreliable in a non-trivial fraction of published specifications. This raises the question of how to detect weak instruments in applications. A natural initial suggestion is to test the hypothesis that the first-stage is equal to zero, $\pi = 0$. As noted in Stock & Yogo (2005), however, conventional methods for inference on β are unreliable not only for $\pi = 0$, but also for π in a neighborhood of zero. Hence, we may reject that $\pi = 0$ even when conventional inference procedures are unreliable. To overcome this issue, we need formal procedures for detecting weak instruments, rather than tests for total non-identification.

Tests for Weak Instruments with Homoskedastic Errors Stock & Yogo (2005) consider the problem of testing for weak instruments in cases with homoskedastic errors. They begin by formally defining the set of values π they will call weak. They consider two different definitions, the first based on the bias of IV estimates relative to OLS and the second based on the size of Wald- or t-tests. In each case they include a value of π in the weak instrument set if the worst-case bias or size over all possible values of β exceeds a threshold (they phrase this result in terms of the correlation between the errors ε and V in (1) and (2), but for Σ known this is equivalent). They then develop formal tests for the null hypothesis that the instruments are weak (that is, that π lies in the weak instrument set), where rejection allows one to conclude that the instruments are strong.

In settings with a single endogenous regressor, Stock & Yogo (2005)'s tests are based on the first-stage F-statistic. Their critical values for this statistic depend on the number of instruments, and tables are available in Stock & Yogo (2005). If we define the instruments as weak when the worst-case bias

of two-stage least squares exceeds 10% of the worst case bias of OLS, the results of Stock and Yogo show that for between 3 and 30 instruments the appropriate critical value for a 5% test of the null of weak instruments ranges from 9 to 11.52, and so is always close to the Staiger & Stock (1997) rule of thumb cutoff of 10. By contrast, if we define the instruments as weak when the worst-case size of a nominal 5% t-test based on two-stage least squares exceeds 15%, then the critical value depends strongly on the number of instruments, and is equal to 8.96 in cases with a single instrument but rises to 44.78 in cases with 30 instruments.

Stock & Yogo (2005) also consider settings with multiple endogenous variables. For such cases they develop critical values for use with the Cragg & Donald (1993) statistic for testing the hypothesis that π has reduced rank. Building on these results, Sanderson & Windmeijer (2016) consider tests for whether the instruments are weak for the purposes of estimation and inference on one of multiple endogenous variables.

Tests for Weak Instruments with Non-Homoskedastic Errors The results of Stock & Yogo (2005) rely heavily on the assumption of homoskedasticity. As discussed above, in homoskedastic settings the variance matrix Σ for $(\hat{\delta}, \hat{\pi})$ can be written as the Kronecker product of a 2×2 matrix with a $k \times k$ matrix, which Stock & Yogo (2005) use to obtain their results. As noted in Section 2, by contrast, Σ does not in general have Kronecker product structure in non-homoskedastic settings, and the tests of Stock & Yogo (2005) do not apply. Specifically, in the non-homoskedastic case the homoskedastic first-stage F-statistic is inapplicable, and should not be compared to the Stock & Yogo (2005) critical values (Montiel Olea & Pflueger, 2013).

Despite the inapplicability of Stock & Yogo (2005)'s results, F-statistics are frequently reported in non-homoskedastic settings with multiple instruments. In such cases, some authors report non-homoskedasticity-robust F-statistics

$$F^R = \frac{1}{k}\hat{\pi}'\hat{\Sigma}_{\pi\pi}^{-1}\hat{\pi}, \tag{7}$$

while others report traditional, non-robust F-statistics

$$F^N = \frac{1}{k}\hat{\pi}'\hat{\Sigma}_{\pi\pi,N}^{-1}\hat{\pi} = \frac{n}{k\hat{\sigma}_V^2}\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi} \tag{8}$$

for $\hat{\Sigma}_{\pi\pi,N} = \frac{\hat{\sigma}_V^2}{n}\hat{Q}_{ZZ}^{-1}$ and $\hat{\sigma}_V^2$ an estimator for $E[V_i^2]$. In our AER data, for instance, none of the 52 specifications that both have multiple instruments and

report first-stage F-statistics assume homoskedasticity to calculate standard errors for $\hat{\beta}$, but at least six report F-statistics do assume homoskedasticity (we are unable to determine the exact count because most authors do not explicitly describe how they calculate F-statistics, and not all papers provide replication data). To illustrate, the left panel of Figure 4 plots the distribution of F-statistics reported in papers in our AER sample, broken down by the method (robust or non-robust) used, when we can determine this. Given the mix of methods, we use "F-statistic" as a generic term to refer both to formal first-stage F-statistics F^N (which assume homoskedasticity and single endogenous regressor) and to generalizations of F-statistics to non-homoskedastic settings, cases with multiple endogenous regressors, and so on.

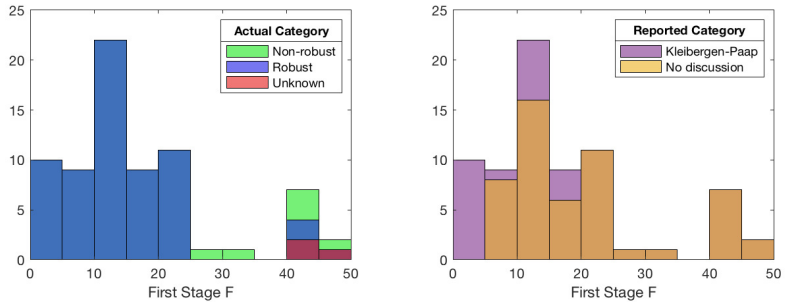

Figure 4: Distribution of reported first-stage F-statistics (and their non-homoskedastic generalizations) in 72 specifications with a single endogenous regressor and first-stage F smaller than 50. 36 other specifications (not shown) have a single endogenous regressor but first-stage F-statistic larger than 50. Left panel decomposes by statistic computed (either non-robust F-statistic F^N, robust F-statistic F^R, or unknown). Note that in settings with a single endogenous regressor, the Kleibergen-Paap F-statistic reduces to the robust F-statistic, so we categorize papers reporting this statistic accordingly. Right panel decomposes by label used by authors in text (either Kleibergen-Paap or not explicitly discussed).

Use of F-statistics in non-homoskedastic settings is built into common statistical software. When run without assuming homoskedastic errors the popular ivreg2 command in Stata automatically reports the Kleibergen & Paap (2007) Wald statistic for testing that π has reduced rank along with critical values based on Stock & Yogo (2005) (Baum et al., 2007), though

the output warns users about Stock & Yogo (2005)'s homoskedasticity assumption. In settings with a single endogenous variable the Kleibergen & Paap (2007) Wald statistic is equivalent to a non-homoskedasticity-robust F-statistic F^R for testing $\pi = 0$, while in settings with multiple endogenous regressors it is a robust analog of the Cragg & Donald (1993) statistic. Interestingly, despite the equivalence of Kleibergen-Paap statistics and robust F-statistics in settings with a single endogenous variable, the distribution of published F-statistics appears to differ depending on what label the authors use. In particular, as shown in the right panel of Figure 4, published F-statistics labeled by authors as Kleibergen-Paap statistics tend to be smaller.

We are unaware of theoretical justification for the use of either F^N or F^R to gauge instrument strength in non-homoskedastic settings. As an alternative, Montiel Olea & Pflueger (2013) propose a test for weak instruments based the effective first-stage F-statistic

$$F^{Eff} = \frac{\hat{\pi}'\hat{\Sigma}_{N,\pi\pi}^{-1}\hat{\pi}}{tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})} = \frac{tr(\hat{\Sigma}_{\pi\pi,N}\hat{Q}_{ZZ})}{tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})}F^N = \frac{k\hat{\sigma}_V^2}{tr(\hat{\Sigma}_{\pi\pi}^*\hat{Q}_{ZZ})}F^N. \qquad (9)$$

In cases with homoskedastic errors F^{Eff} reduces to F^N, while in cases with non-homoskedastic errors it adds a multiplicative correction that depends on the robust variance estimate. Likewise, in the just-identified case F^{Eff} reduces the F^R (and so coincides with the Kleibergen & Paap (2007) Wald statistic), while in the non-homoskedastic case it weights $\hat{\pi}$ by $\hat{Q}_{ZZ}$ rather than $\hat{\Sigma}_{\pi\pi}^{-1}$.

The expressions for the two-stage least squares estimator in (4), F^R in (7), F^N in (8), and F^{Eff} in (9) provide some intuition for why F^{Eff} is an appropriate statistic for testing instrument strength when using two-stage least squares in the non-homoskedastic case while F^R and F^N are not. Two-stage least squares behaves badly when its denominator, $\hat{\pi}'\hat{Q}_{ZZ}\hat{\pi}$, is close to zero. The statistic F^N measures this same object but, because it is non-robust, it "gets the standard error wrong" and so does not have a noncentral F distribution as in Stock & Yogo (2005). Indeed, in the non-homoskedastic case F^N can be extremely large with high probability even when $\pi' Q_{ZZ}\pi$ is small. By contrast, the statistic F^R measures the wrong population object, $\pi'\Sigma_{\pi\pi}^{-1}\pi$ rather than $\pi' Q_{ZZ}^{-1}\pi$, so while it has a noncentral F distribution its noncentrality parameter does not correspond to the distribution of $\hat{\beta}_{2SLS}$.[2] Finally, F^{Eff} measures the right object and "gets the standard er-

[2]The inapplicability of F^R and F^N in the non-homoskedastic case is illustrated by the

rors right on average." More precisely, F^{Eff} is distributed as a weighted average of noncentral χ^2 variables where the weights, given by the eigenvalues of $\hat{\Sigma}_{\pi\pi}^{\frac{1}{2}}\hat{Q}_{ZZ}\hat{\Sigma}_{\pi\pi}^{\frac{1}{2}}/tr(\hat{\Sigma}_{\pi\pi}\hat{Q}_{ZZ})$, are positive and sum to one. Montiel Olea & Pflueger (2013) show that the distribution of F^{Eff} can be approximated by a non-central χ^2 distribution, and formulate tests for weak instruments as defined based on the Nagar (1959) approximation to the bias of two-stage least squares and limited information maximum likelihood. Their test rejects when the effective F-statistic exceeds a critical value. Note, however, that their argument is specific to two-stage least squares and limited information maximum likelihood, so if one were to use a different estimator, a different test would be needed.

For $k = 1$, $\Sigma_{\pi\pi}$, $\Sigma_{\pi\pi,N}$, and Q_{ZZ} are all scalar, and $F^R = F^{Eff}$. Both statistics have a noncentral F distribution with the same noncentrality parameter that governs the distribution of the IV estimator. Thus, in settings with $k = 1$, $F^R = F^{Eff}$ can be used with the Stock & Yogo (2005) critical values based on t-test size (the mean of the IV estimate does not exist when $k = 1$).

For $k > 1$, as noted above the theoretical results of Montiel Olea & Pflueger (2013) formally concern only the Nagar (1959) approximation to the bias. Our simulations based on the AER data reported in the last section suggest, however, that effective F-statistics may convey useful information about instrument strength more broadly, since we saw that conventional asymptotic approximations appeared reasonable in specifications where the average effective F-statistic exceeded 10. This is solely an empirical observation about a particular dataset, but a study of why this is the case in these

following example, which builds on an example in Montiel Olea & Pflueger (2013). Let $k = 2$, $Q_{ZZ} = I_2$, and

$$\Sigma_{\pi\pi} = E\left[\begin{pmatrix} U_i^2 & U_iV_i \\ U_iV_i & V_i^2 \end{pmatrix}\right] \otimes \begin{pmatrix} \omega^2 & 0 \\ 0 & \omega^{-2} \end{pmatrix}.$$

Under weak instrument asymptotics with $\pi = C/\sqrt{n}$ for C fixed with both elements nonzero, as $\omega^2 \to \infty$ one can show that the distribution of the two-stage least squares estimate is centered around the probability limit of ordinary least squares, which is what we expect in the fully unidentified case. Hence, from the perspective of two-stage least squares the instruments are irrelevant asymptotically. At the same time, both F^N and F^R diverge to infinity, and so will indicate that the instruments are strong with probability one. By contrast, F^{Eff} converges to a χ_1^2 and so correctly reflects that the instruments are weak for the purposes of two-stage least squares estimation.

data, and whether this finding generalizes to a broader range of empirically-relevant settings, is an interesting question for future research.

The main conclusion from this section is that F^{Eff}, not F^R or F^N, is the preferred statistic for detecting weak instruments in the over-identified, non-homoskedastic setting when using two-stage least squares or limited information maximum likelihood. F^{Eff} can be compared to Stock & Yogo (2005) critical values for $k = 1$ and to Montiel Olea & Pflueger (2013) critical values for $k > 1$, or to the rule-of-thumb value of 10. It appears that none of the papers in our AER sample computed F^{Eff} (except for the $k = 1$ case where it equals F^R), but we hope to see the wider use of this statistic in the future.

4.1 Screening on the First-Stage F-Statistic

Given a method for detecting weak instruments, there is a question of what to do if we decide the instruments are weak. Anecdotal evidence and our AER data suggest that in some instances, researchers or journals may decide that specifications with small first-stage F-statistics should not be published. Specifically, Figure 1 shows many specifications just above the Staiger & Stock (1997) rule of thumb cutoff of 10, consistent with selection favoring F-statistics above this threshold.

It is important to note that Figure 1 limits attention to specifications where the original authors report first-stage F-statistics, and uses the F-statistics as reported by the authors. By contrast, in our simulation results we calculate effective F-statistics for all specifications in our simulation sample (i.e. where we can obtain a full-rank estimate of the variance matrix Σ), including in specifications where the authors do not report F-statistics, and match the assumptions used to calculate F-statistics to those used to calculate standard errors on $\hat{\beta}$. So, for example, in a paper that assumed homoskedastic errors to calculate F-statistics, but non-homoskedastic errors to calculate standard errors on $\hat{\beta}$, we use a non-homoskedasticity-robust estimator $\hat{\Sigma}_{\pi\pi}$ to compute the effective F-statistic in our simulations, but report the homoskedastic F-statistic F^N in Figure 1. We do this because the F-statistic reported by the original authors seems the relevant one when thinking about selection on F-statistics.

While selection on first-stage F-statistics is intuitively reasonable, it can unfortunately result in bias in published estimates and size distortions in published tests. This point was made early in the weak instruments literature

by Nelson et al. (1998), and relates to issues of pretesting and publication bias more generally. To illustrate the impact of these issues, we consider simulations calibrated to our AER data in which we drop all simulation draws where the effective F-statistic is smaller than 10. Figure 5 plots the size of nominal 5% t-tests in this setting against the average effective F-statistic (where the average effective F-statistic is calculated over all simulation draws, not just those with $F^{Eff} > 10$).

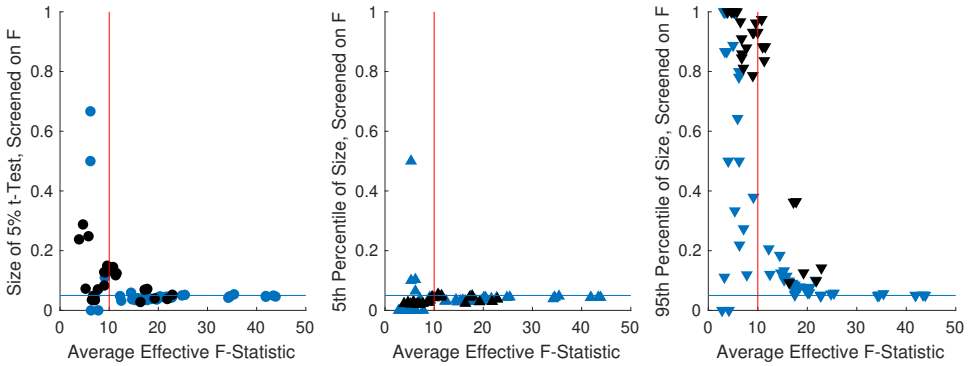

Figure 5: Rejection probability for nominal 5% two-sided t-tests after screening on $F^{Eff} > 10$, plotted against the average effective F-statistic in calibrations to AER data. Limited to the 106 out of 124 specifications with average effective F smaller than 50. Just-identified specifications are plotted in black, while over-identified specifications are in blue. Left panel plots size at parameter values estimated from AER data, while middle and right panels plot, respectively, the 5th and 95th percentiles of the size under the Bayesian exercise described in Section 3. Red line corresponds to a first-stage F of 10.

The results in Figure 5 highlight that screening on the F-statistics can dramatically increase size distortions. This is apparent even in simulations based on reported parameter estimates (shown in the left panel), where the maximal size exceeds 70%, as compared to a maximal size of less than 35% for t-tests without screening on the F-statistic. Matters look still worse when considering the upper bound for size (shown in the right panel), where many specifications have size close to one. Moreover, these upper bounds suggest that non-negligible size distortions may arise even in specifications with an average first-stage F-statistic of nearly 20, and thus that screening on first-stage F-statistics undermines the good properties we previously found for specifications with average first-stage F larger than 10. Hence, screening on

the first-stage F-statistic appears to compound, rather than reduce, inferential problems arising from weak instruments. This problem is not specific to the effective F-statistic F^{Eff}, and also appears if we screen on F^N or F^R. Likewise, if we move the threshold from 10 to some other value, we continue to see size distortions in a neighborhood of the new threshold.

If we are confident our instruments are valid, but are concerned they may be weak, screening on F-statistics is unappealing for another reason: it unnecessarily eliminates specifications of potential economic interest. In particular, as we discuss in the next section a variety of procedures for identification-robust inference on β have been developed in the literature. By using these procedures we may gain insight from the data even in settings where the instruments are weak. Hence, weak instruments alone are not a reason to discard applications.

5 Inference with Weak Instruments

The literature on weak instruments has developed a variety of tests and confidence sets that remain valid whether or not the instruments are weak, in the sense that their probability of incorrectly rejecting the null hypothesis and covering the true parameter value, respectively, remains well-controlled. Since instrumental variables estimates are non-normally distributed when the instruments are weak, these procedures do not rely on point estimates and standard errors but instead use test inversion.

The idea of test inversion is that if we are able to construct a size-α test of the hypothesis $H_0 : \beta = \beta_0$ for any value β_0, then we can construct a level $1 - \alpha$ confidence set for β by collecting the set of non-rejected values. Formally, let us represent a generic test of $H_0 : \beta = \beta_0$ by $\phi(\beta_0)$, where we write $\phi(\beta_0) = 1$ if the test rejects and $\phi(\beta_0) = 0$ otherwise. We say that $\phi(\beta_0)$ is a size-α test of $H_0 : \beta = \beta_0$ in the normal model (6) if

$$\sup_{\pi} E_{\beta_0,\pi} \left[\phi(\beta_0) = 1\right] \leq \alpha,$$

so the maximal probability of rejecting the null hypothesis, assuming the null is true, is bounded above by α no matter the value of π. If $\phi(\beta_0)$ is a size-α test of $H_0 : \beta = \beta_0$ for all values β_0 then $CS = \{\beta : \phi(\beta) = 0\}$, the set of values not rejected by ϕ, is a level $1 - \alpha$ confidence set

$$\inf_{\beta,\pi} Pr_{\beta,\pi} \left\{\beta \in CS\right\} \geq 1 - \alpha. \tag{10}$$

In practice, we can implement test inversion by taking a grid of potential values β, evaluating the test ϕ at all values in the grid, and approximating our confidence set by the set of non-rejected values.

When the instruments can be arbitrarily weak, correct coverage (10) turns out to be a demanding requirement. Specifically, the results of Gleser & Hwang (1987) and Dufour (1997) imply that in the normal model (6) without restrictions on (β, π) any level $1-\alpha$ confidence set for β must have infinite length with positive probability. For intuition, consider the case in which $\pi = 0$, so β is unidentified. In this case, the data are entirely uninformative about β, and to ensure coverage $1-\alpha$ a confidence set CS must cover each point in the parameter space with at least this probability, which is impossible if CS is always bounded. That the confidence set must be infinite with positive probability for all (β, π) then follows from the fact that the normal distribution has full support. Hence, if the event $\{CS \text{ infinite length}\}$ has positive probability under $\pi = 0$, the same is true under all (β, π). This immediately confirms that we cannot obtain correct coverage under weak instruments by adjusting our (finite) standard errors, and so points to the need for a different approach such as test inversion.

To fix ideas we first discuss test inversion based on the Anderson-Rubin (AR) statistic, which turns out be efficient in just-identified models with a single instrument. We then turn to alternative procedures developed for over-identified models and inference on subsets of parameters. Finally, we discuss the effect of choosing between robust and non-robust procedures based on a pre-test for instrument strength. Since we base our discussion on the OLS estimates $(\hat{\delta}, \hat{\pi})$, the procedures we discuss here can be viewed as minimum-distance identification-robust procedures as in Magnusson (2010).

5.1 Inference for Just-Identified Models: the Anderson-Rubin Test

Test inversion offers a route forward in models with weak instruments because the IV model with parameter β implies testable restrictions on the distribution of the data regardless of the strength of the instruments. Specifically, the IV model implies that $\delta = \pi\beta$. Hence, under a given null hypothesis $H_0 : \beta = \beta_0$ we know that $\delta - \pi\beta_0 = 0$, and hence that

$$g(\beta_0) = \hat{\delta} - \hat{\pi}\beta_0 \sim N(0, \Omega(\beta_0)) \text{ for } \Omega(\beta_0) = \Sigma_{\delta\delta} - \beta(\Sigma_{\delta\pi} + \Sigma_{\pi\delta}) + \beta^2\Sigma_{\pi\pi}$$

where $\Sigma_{\delta\delta}$, $\Sigma_{\pi\pi}$, and $\Sigma_{\delta\pi}$ denote the variance of $\hat{\delta}$, the variance of $\hat{\pi}$, and their covariance, respectively. Hence the AR statistic (Anderson & Rubin, 1949), defined as $AR(\beta) = g(\beta)'\Omega(\beta)^{-1}g(\beta)$, follows a χ^2_k distribution under $H_0 : \beta = \beta_0$ no matter the value of π. Note that Anderson & Rubin (1949) considered the case with homoskedastic normal errors, so the AR statistic as we define it here is formally a generalization of their statistic that allows for non-homoskedastic errors.

Using the AR statistic, we can form an AR test of $H_0 : \beta = \beta_0$ as $\phi_{AR}(\beta_0) = 1\left\{AR(\beta_0) > \chi^2_{k,1-\alpha}\right\}$ for $\chi^2_{k,1-\alpha}$ the $1-\alpha$ quantile of a χ^2_k distribution. As noted by Staiger & Stock (1997) this yields a size-α test that is robust to weak instruments. Hence, if we were to re-compute Figure 3 for the AR test, the size would be flat at 5% for all specifications. We can thus form a level $1-\alpha$ weak-instrument-robust confidence set CS_{AR} by collecting the non-rejected values. In the case with homoskedastic errors (or with non-homoskedastic errors but a single instrument) as noted by e.g. Mikusheva (2010) one can derive the bounds of CS_{AR} analytically, avoiding the need for numerical test inversion.

Since AR confidence sets have correct coverage regardless of the strength of the instruments, we know from Gleser & Hwang (1987) and Dufour (1997) that they have infinite length with positive probability. Specifically, as discussed in Dufour & Taamouti (2005) and Mikusheva (2010) CS_{AR} can take one of three forms in settings with a single instrument: (i) a bounded interval $[a, b]$, (ii) the real line $(-\infty, \infty)$, and (iii) the real line excluding an interval $(-\infty, a] \cup [b, \infty)$. In settings with more than one instrument but homoskedastic errors, the AR confidence set can take the same three forms, or may be empty. These behaviors are counter-intuitive, but have simple explanations.

First, as noted by Kleibergen (2007), as $|\beta| \to \infty$, $AR(\beta)$ converges to the Wald statistic for testing that $\pi = 0$ (which is equal to k times the robust first-stage F-statistic). Hence, the level-α AR confidence set has infinite length if and only if a robust F-test cannot reject that $\pi = 0$, and thus that β is totally unidentified. Thus, infinite-length confidence sets arise exactly in those cases where the data do not allow us to conclude that β is identified at all.

Second, CS_{AR} may be empty only in the over-identified setting. In this case, the AR approach tests that $\delta = \pi\beta_0$, which could fail either because $\delta = \pi\beta$ for $\beta \neq \beta_0$, or because there exists no value β such that $\delta = \pi\beta$. In

the latter case the over-identifying restrictions of the IV model fail. Hence, the AR test has power against both violations of our parametric hypothesis of interest and violations of the IV model's overidentifying restrictions, and an empty AR confidence set can be interpreted as a rejection of the overidentifying restrictions. The overidentifying restrictions could fail due either to invalidity of the instruments or to treatment effect heterogeneity as in Imbens & Angrist (1994), but either way imply that the constant-effect IV model is misspecified.

The power of Anderson-Rubin tests against violations of the IV model's overidentifying restrictions means that if we care only about power for testing the parametric restriction $H_0 : \beta = \beta_0$, AR tests and confidence sets can be inefficient. In particular, in the strongly-identified case with $\|\pi\|$ large one can show that the usual Wald statistic $(\hat{\beta} - \beta_0)^2/\hat{\sigma}^2_{\hat{\beta}}$ is approximately noncentral-χ^2_1 distributed with the same noncentrality as $AR(\beta_0)$, so tests based on the Wald statistic (or equivalently, two-sided t-tests) have higher power than tests based on AR. Strong identification is important for this result. Chernozhukov et al. (2009) show that the AR test is admissible (i.e. not dominated by any other test) in settings with homoskedastic errors and weak instruments.

Efficiency of AR in Just-Identified Models In just-identified models there are no overidentifying restrictions and the AR test has power only against violations of the parametric hypothesis. In this setting, Moreira (2009) shows that the AR test is uniformly most accurate unbiased. We say that a size-α test ϕ is unbiased if $E_{\beta,\pi}[\phi(\beta_0)] \geq \alpha$ for all $\beta \neq \beta_0$ and all π, so the rejection probability when the null hypothesis is violated is at least as high as the rejection probability when the null is correct. The Anderson-Rubin test is unbiased, and Moreira (2009) shows that for any other size-α unbiased test ϕ, $E_{\beta,\pi}[\phi_{AR}(\beta_0) - \phi(\beta_0)] \geq 0$ for all $\beta \neq \beta_0$ and all π. Hence, the AR test has (weakly) higher power than any other size-α unbiased test no matter the true value of the parameters. In the strongly-identified case the AR test is asymptotically efficient in the usual sense, and so does not sacrifice power relative to the conventional t-test.

Practical Performance of AR Confidence Sets Since Anderson-Rubin confidence sets are robust to weak identification, and are efficient in the just-identified case, there is a strong case for using these procedures in just-

identified settings. To examine the practical impact of using AR confidence sets, we return to our AER dataset, limiting attention to just-identified specifications with a single endogenous variable where we can estimate the joint variance-covariance matrix of $(\hat{\pi}, \hat{\delta})$. In the sample of 34 specifications meeting these requirements, we find that AR confidence sets are quite similar to t-statistic confidence sets in some cases, but are longer in others. Specifically, in two specifications the first-stage is not distinguishable from zero at the 5% level so AR confidence sets are infinite. In the remaining 32 specifications AR confidence sets are 56.5% longer than t-statistic confidence sets on average, though this difference drops to 20.3% if we limit attention to specifications that report a first-stage F-statistic larger than 10, and to 0.04% if we limit attention to specifications that report a first-stage F-statistic larger than 50. Complete results are reported in Section D of the online appendix.

5.2 Tests for Over-Identified Models

In contrast to the just-identified case, in over-identified settings the AR test is robust but inefficient under strong identification. This has led to a large literature seeking procedures that perform better in over-identified models.

Towards this end, note that in the normal model (6) the Anderson-Rubin statistic for testing $H_0 : \beta = \beta_0$ depends on the data only through $g(\beta_0) = \hat{\delta} - \hat{\pi}\beta_0$. To construct procedures that perform as well as the t-test in the strongly-identified case, it is valuable to incorporate information from $\hat{\pi}$, which is informative about which deviations of $\delta - \pi\beta_0$ from zero correspond to violations of the parametric restrictions of the model, rather than the overidentifying restrictions. Specifically, under alternative parameter value β, $\hat{\delta} - \hat{\pi}\beta_0 \sim N(\pi(\beta - \beta_0), \Omega(\beta_0))$. See I. Andrews (2016) for discussion. Hence, to construct procedures that perform as well as the t-test in well-identified, over-identified cases, a number of authors have considered test statistics that that depend on $(\hat{\delta}, \hat{\pi})$ through more than $\hat{\delta} - \hat{\pi}\beta_0$.

Once we seek to construct weak-instrument-robust tests that depend on the data through more than $g(\beta_0)$, however, we encounter an immediate problem: even under the null $H_0 : \beta = \beta_0$, the distribution of $(\hat{\delta}, \hat{\pi})$ depends on the (unknown) first-stage parameter π. Hence, for a generic test statistic $s(\beta_0)$ that depends on $(\hat{\delta}, \hat{\pi})$, the distribution of $s(\beta_0)$ under the null will typically depend on π. For example, if we take $s(\beta_0)$ to be the absolute t-statistic $|\hat{\beta} - \beta_0|/\hat{\sigma}_{\hat{\beta}}$, we know that the distribution of t-statistics under the null depends on the strength of the instruments. One could in principle find

the largest possible 1-α quantile for $s(\beta_0)$ over the null consistent with some set of values for π, for example an initial confidence set as in the Bonferroni approach of Staiger & Stock (1997). For many statistics $s(\beta_0)$, however, this requires extensive simulation and will be computationally intractable, and moreover typically entails a loss of power.

An alternative approach eliminates dependence on π through conditioning. Specifically, under $H_0 : \beta = \beta_0$

$$\begin{pmatrix} g(\beta_0) \\ \hat{\pi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{pmatrix} \Omega(\beta_0) & \Sigma_{\delta\pi} - \Sigma_{\pi\pi}\beta_0 \\ \Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0 & \Sigma_{\pi\pi} \end{pmatrix}\right).$$

Thus, if we define

$$D(\beta) = \hat{\pi} - (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta)\,\Omega(\beta)^{-1} g(\beta),$$

we see that $(g(\beta), D(\beta))$ is a one-to-one transformation of $(\hat{\delta}, \hat{\pi})$, and under $H_0 : \beta = \beta_0$

$$\begin{pmatrix} g(\beta_0) \\ D(\beta_0) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ \pi \end{pmatrix}, \begin{pmatrix} \Omega(\beta_0) & 0 \\ 0 & \Psi(\beta_0) \end{pmatrix}\right)$$

for $\Psi(\beta_0) = \Sigma_{\pi\pi} - (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0)\,\Omega(\beta_0)^{-1}(\Sigma_{\delta\pi} - \Sigma_{\pi\pi}\beta_0)$. Thus, under the null the nuisance parameter π enters the distribution of the data only through the statistic $D(\beta_0)$, while $g(\beta_0)$ is independent of $D(\beta_0)$ and has a known distribution. Hence, the conditional distribution of $g(\beta_0)$ (and thus of $(\hat{\delta}, \hat{\pi})$) given $D(\beta_0)$ doesn't depend on π. This conditioning approach was initially introduced to the weak-instruments literature by Moreira (2003), who studied the homoskedastic case. In settings with homoskedastic errors $g(\beta_0)$ and $D(\beta_0)$ are transformations of the statistics S and T introduced by Moreira (2003) – see I. Andrews & Mikusheva (2016).

We can simulate the conditional distribution of any statistic $s(\beta_0)$ given $D(\beta_0)$ under the null by drawing $g(\beta_0)^* \sim N(0, \Omega(\beta_0))$, constructing $(\hat{\delta}^*, \hat{\pi}^*)$ as

$$\begin{pmatrix} \hat{\delta}^* \\ \hat{\pi}^* \end{pmatrix} = \begin{pmatrix} I + \beta_0 (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0)\,\Omega(\beta_0)^{-1} & \beta_0 I \\ (\Sigma_{\pi\delta} - \Sigma_{\pi\pi}\beta_0)\,\Omega(\beta_0)^{-1} & I \end{pmatrix} \begin{pmatrix} g(\beta_0)^* \\ D(\beta_0) \end{pmatrix}$$

for given $D(\beta_0)$, and tabulating the resulting distribution of $s^*(\beta_0)$ calculated based on $(\hat{\delta}^*, \hat{\pi}^*)$. If we denote the conditional $1-\alpha$ quantile as $c_\alpha(D(\beta_0))$, we can then construct a conditional test based on s as $\phi_s =$

$1\{s(\beta_0) > c_\alpha(D(\beta_0))\}$, and provided $s(\beta_0)$ is continuously distributed conditional on $D(\beta_0)$ this test has rejection probability exactly α under the null, $E_{\beta_0,\pi}[\phi_s(\beta_0)] = \alpha$ for all π, while if the conditional distribution of $s(\beta_0)$ has point masses the test has size less than or equal to α. As noted by Moreira (2003) for the homoskedastic case, this allows us to construct a size-α test based on any test statistic s. For further discussion of the simulation approach described above and a formal size control result applicable to the non-homoskedastic case, see I. Andrews & Mikusheva (2016).

Tests which have rejection probability exactly α for all parameter values consistent with the null are said to be similar. Theorem 4.3 of Lehmann & Romano (2005) implies that if the set of values of π is unrestricted then all similar size-α tests of $H_0 : \beta = \beta_0$ are conditional tests in the sense that their conditional rejection probability given $D(\beta_0)$ under the null is equal to α. Moreover, in the present setting the power functions for all tests are continuous, so if the set of values (β, π) is unrestricted then all unbiased tests are necessarily similar. Thus, the class of conditional tests nests the class of unbiased tests. Together, these results show that in cases where (β, π) is unrestricted, the class of conditional tests has attractive properties. Within this class, however, there remains a question of what test statistics $s(\beta_0)$ to use. In the homoskedastic case we recommend using the likelihood ratio statistic as proposed by Moreira (2003). In the non-homoskedastic case, however, the literature has not yet converged on a recommendation, other than to use one of several procedures that is efficient under strong instruments.

Tests for Homoskedastic Case A wide variety of test statistics have been proposed in the literature. Kleibergen (2002) proved that a particular score statistic (Breusch & Pagan, 1980) had correct size in the model with homoskedastic errors, while in the same model Moreira (2003) proposed the general conditioning approach for homoskedastic models and noted that both the AR test and Kleibergen (2002)'s score test are conditional tests (trivially, since their conditional critical values do not depend on $D(\beta_0)$). Moreira (2003) further proposed conditional Wald and likelihood ratio tests, based on comparing the Wald and likelihood ratio statistics to a conditional critical value. Unlike AR, both the score and likelihood ratio statistics depends on both $(\hat{\delta}, \hat{\pi})$, and conditional tests based on these statistics are efficient in the well-identified case.

D. Andrews et al. (2006) found that the conditional likelihood ratio (CLR) test of Moreira (2003) has very good power properties in the homoskedastic case with a single endogenous variable. The Kronecker product structure of the variance matrix Σ in this setting means that the problem is unchanged by linear transformations of the instruments. It is therefore natural to limit attention to tests that are likewise invariant, in the sense that their value is unchanged by linear transformations of the instruments. D. Andrews et al. (2006) showed, however, that the power of such invariant tests depends only on the correlation between the errors (U, V), the (variance-normalized) length of the first-stage π, and the true parameter value β. Imposing an additional form of invariance to limit attention to two-sided tests, D. Andrews et al. (2006) showed numerically that the CLR test has power close to the upper bound for the power of any invariant similar test over a wide range of parameter values, where the calculation is made feasible by the low dimension of the invariant parameter space. D. Andrews et al. (2008) extended this result by showing that the power envelope for invariant non-similar tests is close to that for invariant similar tests, and thus that (a) there is not a substantial power cost to imposing similarity in the homoskedastic setting if one limits attention to invariant tests, and (b) that the CLR test performs well even in comparison to non-similar tests. Building on these results, Mikusheva (2010) proved a near-optimality property for CLR confidence sets. D. Andrews et al. (2018b) added a note of caution, showing that there exist parameter values not explored by D. Andrews et al. (2006) where the power of the CLR test is further from the power envelope, but still recommend the CLR test for the homoskedastic, single endogenous regressor setting.

Tests for Non-Homoskedastic Case The simplifications obtained using Kronecker structure of Σ are no longer available in the non-homoskedastic case, introducing substantial complications.

Motivated by the positive results for the CLR test, a number of authors have explored analogs and generalizations of the CLR test for non-homoskedastic settings. The working paper version of D. Andrews et al. (2006), D. Andrews et al. (2004), introduced a version of the CLR test applicable to the non-homoskedastic case, while Kleibergen (2005) introduced the conditioning statistic $D(\beta_0)$ for the non-homoskedastic case and developed score and quasi-CLR statistics applicable in this setting. D. Andrews & Guggenberger (2015) introduced two alternative quasi-CLR tests for non-

homoskedastic settings that allow a singular covariance matrix Σ. I. Andrews (2016) studied tests based on linear combinations of AR and score statistics, noting that the CLR test can be expressed in this way. Finally, Moreira & Moreira (2015) and I. Andrews & Mikusheva (2016) introduce a direct generalization of the CLR test to settings with non-homoskedastic errors, which again compares the likelihood ratio statistic to a conditional critical value.

All of these extensions of the CLR test are efficient under strong identification, and all but the proposal of I. Andrews (2016) reduce to the CLR test of Moreira (2003) in the homoskedastic, single endogenous variable setting where the results of D. Andrews et al. (2006) apply. At the same time, however, while these generalizations are intended for the non-homoskedastic case, evidence on their performance in the weakly identified case has largely been limited to simulation results.

To derive tests with provable optimality properties in the weakly-identified non-homoskedastic case, a recent literature has focused on optimizing weighted average power, meaning power integrated with respect to weights on (β, π). Specifically the similar test maximizing weighted average power with respect to the weights ν, $\int E_{\beta,\pi}[\phi] d\nu(\beta, \pi)$, rejects when

$$s(\beta_0) = \int f(\hat{\delta}, \hat{\pi}; \beta, \pi) d\nu(\beta, \pi) / f(\hat{\delta}, \hat{\pi} | D(\beta_0); \beta_0)$$

exceeds its conditional critical value. Intuitively, this weighted average power optimal test rejects when the observed data is sufficiently more likely to have arisen under the weighted alternative $H_1 : \beta \neq \beta_0$, weighted by ν, than under the null $H_0 : \beta = \beta_0$. As this description suggests, the choice of the weight ν plays an important role in determining the power and other properties of the resulting test, though the use of conditional critical values ensures size control for all choices of ν.

Moreira & Moreira (2013) and Montiel Olea (2017) show that weighted average power optimal similar tests can attain essentially any admissible power function through an appropriate choice of weights. Montiel Olea (2017) further proposed particular choices of weights ν for the homoskedastic and non-homoskedastic cases, while Moreira & Moreira (2015) showed that unless the weights are chosen carefully weighted average power optimal similar tests may have poor power even in the homoskedastic case, and that the problem can be still worse in the non-homoskedastic case. To remedy this, they modify the construction of weighted average power optimal tests to enforce a sufficient condition for local unbiasedness, and showed that these tests

performed well in simulation and are asymptotically efficient in the case with strong instruments. Finally, Moreira & Ridder (2017) proposed weights ν motivated by invariance considerations. They further showed that there exist parameter configurations in the non-homoskedastic case where tests that depend only on the AR and score statistics, like those of Kleibergen (2005) and I. Andrews (2016), have poor power.

To summarize, in settings with a single endogenous regressor and homoskedastic errors, the literature to date establishes good properties for the CLR test of Moreira (2003). In settings with non-homoskedastic errors, by contrast, a large number of procedures have been proposed, but a consensus has not been reached on what procedures to use in practice, beyond the recommendation that researchers use procedures that are efficient when the instruments are strong. Consequently, it is not yet clear what procedure(s) to recommend in this case.

5.3 Inference with Multiple Endogenous Regressors

The methods we have so far discussed for models with a single endogenous regressor can all be generalized to test of hypotheses on the $p \times 1$ vector β in settings with multiple endogenous variables (as in 19 of the 230 specifications in our AER sample). By inverting such tests, we can form simultaneous confidence sets for β. Test inversion with multiple endogenous variables becomes practically difficult for moderate or high-dimensional β, since the number of grid points at which we need to evaluate our test grows exponentially in the dimension. See the supplementary materials to I. Andrews (2016) for a discussion of this issue. On the other hand, high-dimensional settings do not appear common in practice, and no specification in our AER data has more than four endogenous regressors. It is in any event rare to report confidence sets for the full vector β in multidimensional settings with strong instruments. Instead, it is far more common to report standard errors or confidence sets for one element of β at a time.

Formally, suppose we decompose $\beta = (\beta_1, \beta_2)$ and are interested in tests or confidence sets for the subvector β_1 alone. This is known as the subvector inference problem. One possibility for subvector inference is the projection method. In the projection method, we begin with a confidence set CS^β for the full parameter vector β, and then form a confidence set for β_1 by

collecting the implied set of values

$$CS^{\beta_1} = \left\{\beta_1 : \text{ there exists } \beta_2 \text{ such that } (\beta_1, \beta_2) \in CS^{\beta}\right\}.$$

This is called the projection method because we can interpret CS^{β_1} as the projection of CS^{β} onto the linear subspace corresponding to β_1. The projection method was advocated for the weak instruments problem by Dufour (1997), Dufour & Jasiak (2001), and Dufour & Taamouti (2005). Dufour & Taamouti (2005) derived analytic expressions for projection-based confidence sets using the AR statistic in the homoskedastic case.

Unfortunately, the projection method frequently suffers from poor power. When used with the AR statistic, for example, we can interpret the projection method as minimizing $AR(\beta_1, \beta_2)$ with respect to the nuisance parameter β_2, and then comparing $\min_{\beta_2} AR(\beta_1, \beta_2)$ to the same χ^2_k critical value we would have used without minimization. As a result, projection-method confidence sets often cover the true parameter value with probability strictly higher than the nominal level, and so are conservative.

If the instruments are strong for the purposes of estimating β_2 (so if β_1 were known, estimation of β_2 would be standard), these problems have a simple solution: we can reduce our degrees of freedom to account for minimization over the nuisance parameter. Results along these lines for different tests are discussed by Stock & Wright (2000), Kleibergen (2005), and I. Andrews & Mikusheva (2016).

If we cannot assume β_2 is strongly identified, matters are unfortunately more complicated. Guggenberger et al. (2012) show that in the setting with homoskedastic errors one can reduce the degrees of freedom for the AR statistic to mitigate projection conservativeness (using a $\chi^2_{k-p_2}$ critical value for p_2 the dimension of β_2), and Guggenberger et al. (2018) propose a further modification to improve power. On the other hand, Guggenberger et al. (2012) show that the analog of their result fails for the score statistic of Kleibergen (2002). Moreover, Lee (2015) shows that even the results of Guggenberger et al. (2012) for the AR statistic does not extend to the general non-homoskedastic case.

To improve the power of the projection method without assuming the nuisance parameter β_2 is strongly identified, Chaudhuri & Zivot (2011) propose a modified projection approach which chooses the initial confidence set CS^{β} to ensure improved performance for CS^{β_1} in the case with strong instruments. In particular, Chaudhuri & Zivot (2011) base CS^{β} on the combination of a

modified score statistic with an Anderson-Rubin statistic, and show that the resulting CS^{β_1} comes arbitrarily close to efficiency in the case with strong instruments. I. Andrews (2018) proposes a variant of this approach for constructing confidence sets for functions $f(\beta)$ of the parameter vector other than subvectors, while D. Andrews (2018) generalizes Chaudhuri & Zivot (2011) in several directions, introducing a variety of test statistics and deriving confidence sets that are asymptotically efficient in the strongly identified case. Finally, Zhu (2015) introduces a Bonferroni approach for subvector inference that provides an alternative to projection.

5.4 Two-Step Confidence Sets

Weak-instrument-robust confidence sets are not widely reported in practice. For instance, only two papers in our AER sample report robust confidence sets. When such confidence sets are reported, it often appears to be because the authors have uncovered evidence that their instruments are weak. For example, in a survey of 35 empirical papers that reported confidence sets based on Moreira (2003), I. Andrews (2018) found that 29 had at least one specification reporting a first-stage F-statistic smaller than 10.

Used in this way, robust confidence sets may act as an alternative to dropping specifications altogether, which as discussed in Section 4.1 can result in large size distortions. In particular one can consider constructing a two-step confidence set, where one first assesses instrument strength and then reports conventional confidence sets if the instruments appear strong and a robust confidence set if they appear weak. As discussed in I. Andrews (2018), the results of Stock & Yogo (2005) imply bounds on the size of two-step confidence sets based on the first-stage F-statistic in homoskedastic or just-identified settings. In overidentified non-homoskedastic settings, by contrast, Andrews (2018) shows that in general two-step confidence sets based on the robust first-stage F-statistic F^R and conventional cutoffs can have large size distortions.

The implications of the negative results of I. Andrews (2018) for two-step confidence sets in empirically relevant settings, or for two-step confidence sets based on F^{Eff}, are not clear. To examine this issue, Figure 6 plots the size of two-step tests based on the effective F-statistic (which use a t-test if $F^{Eff} > 10$ and an AR test if $F^{Eff} \leq 10$) against the average effective F-statistic in simulations based on our AER data.

The results of Figure 6 show that two-step confidence sets based on the

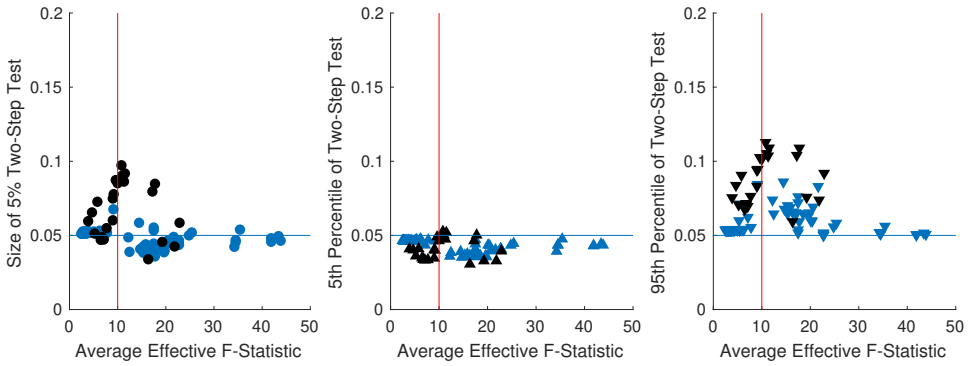

Figure 6: Rejection probability for nominal 5% two-step test that uses 5% t-test and 5% AR test when the effective F-statistic is larger than and smaller than 10, respectively. Limited to the 106 out of 124 specifications with average effective F smaller than 50. Just-identified specifications are plotted in black, while over-identified specifications are in blue. Left panel plots size at parameter values estimated from AER data, while middle and right panels plot, respectively, the 5th and 95th percentiles of the size under the Bayesian exercise described in Section 3. Red line corresponds to a first-stage F of 10.

effective F-statistic have at most mild size distortions in simulations calibrated to our AER data. Specifically, no specification yields size exceeding 10%, and even when we consider upper bounds no specification yields size exceeding 11.5%.

6 Open Questions

While considerable progress has been made on both detecting weak instruments and developing identification-robust confidence sets, a number of important open questions remain. As suggested in the last section, no consensus has been reached on what inference procedures to use in over-identified models with non-homoskedastic errors. Likewise, existing optimality results for weak-instrument-robust inference on subsets of parameters only address behavior in the strongly-identified case.

Simulation results calibrated to our AER sample raise additional questions. First, we found that conventional t-tests appears to perform reasonably well in specifications where the average effective F-statistic is larger than 10, even in over-identified, non-homoskedastic cases. Likewise, we found that

two-step confidence sets based on the effective F-statistic appear to have well-controlled size distortions. These results suggest that the effective F-statistic might provide a useful gauge of identification strength in a wider range of cases than is suggested by the current theoretical literature, but a more extensive and formal exploration of whether this is in fact the case and if so, why, is needed.

Another set of open questions concerns model misspecification. Existing weak instrument-robust procedures assume the constant-effect linear instrumental variables model holds. If the model is instead misspecified, for example because the instruments are invalid, then as noted by Guggenberger (2012) existing weak-instrument robust confidence sets do not have correct coverage for the true parameter value. Of course, the same is also true for conventional confidence sets with strong but invalid instruments, so this issue is not unique to robust confidence sets. In over-identified settings with weak instruments, however, the arguments for size control of existing procedures break down even if one considers inference on pseudo-true parameter values (e.g. the population analog of the two-stage least squares or GMM coefficient). This issue was noted and corrected for two-stage least squares estimates in strong-instrument settings with heterogeneous treatment effects in the appendix to Imbens & Angrist (1994), and more recently by Lee (2018). To the best of our knowledge, however, analogous results have not been developed for settings with weak instruments.

Concern about model misspecification could also interact with the practice of screening on instrument strength: if one thinks that many instruments used in practice are slightly invalid (in the spirit of e.g. Conley et al. (2012)), then while this will result in size distortions, it typically will not qualitatively change results when the instruments are strong. On the other hand, when the instruments are weak, even a small degree of instrument invalidity could account for most of the relationship between Z and X, and so lead to qualitatively quite different conclusions. To address this, researchers may wish to limit attention to settings where the instruments are sufficiently strong for them to be confident that results will be qualitatively robust to low levels of instrument invalidity. How to make this argument precise and conduct inference, however, we leave to future work.

Another important open question concerns the validity of the normal approximation to the distribution of the reduced-form and first-stage coefficients. In this review, including in our simulations, we have used the model (6) which takes the reduced-form and first-stage coefficients $(\hat{\delta}, \hat{\pi})$ to

be normally distributed with known variance. While this approximation can be justified with asymptotic arguments, whether or not it is reasonable in a given application is necessarily case-specific. Important recent work by Young (2018) casts serious doubt on the quality of this normal approximation in many applications.

Using a sample of studies published in the journals of the American Economic Association which overlaps with but is substantially larger than our AER sample, Young (2018) finds that many reported results are heavily influenced by a small number of observations or clusters. Since the Central Limit Theorem used to derive the limiting normal distribution (5) for the reduced-form and first-stage coefficients assumes that the influence of each observation is small, this suggests that the normal approximation may be unreasonable. Moreover, Young (2018) notes that variance estimates $\hat{\Sigma}$ for settings with non-homoskedastic data (which Young calls the non-iid case) can be extremely noisy in finite samples. In simulations that account for these factors, Young finds large size distortions for both conventional and AR tests, with particularly severe distortions for AR tests in over-identified settings. Young (2018) further finds that first-stage F-statistics do not appear to provide a reliable guide to the performance of conventional inference procedures, and that we may spuriously observe large first-stage F-statistics even when the instruments are irrelevant, though he finds somewhat better behavior for the tests of Montiel Olea & Pflueger (2013). To address these issues, Young (2018) suggests using the bootstrap for inference.

We know that bootstrap procedures based on IV estimates or t-statistics are generally invalid when the instruments are weak, and so are not a satisfactory solution in settings with weak instruments. However, appropriately constructed bootstrap procedures based on identification-robust statistics may remain valid. For example, Moreira et al. (2009) show validity of bootstrapped score and Anderson-Rubin tests under weak instruments in the homoskedastic case, where it important for their results that the bootstrap be recentered to ensure that $\hat{\delta} - \hat{\pi}\beta$ has mean zero under the bootstrap distribution. Subsequently, Davidson & MacKinnon (2014) proposed additional bootstrap procedures, but did not establish their validity when the instruments are weak. We expect that is should be possible to extend the results of Moreira et al. (2009) showing validity of bootstrap-based identification-robust tests to the non-homoskedastic case, and to other identification-robust procedures. At the same time, even when used with identification-robust test statistics the bootstrap is not a panacea, and Wang & Tchatoka (Forth-

coming) show that the bootstrap does not ensure size control for subvector inference based on the AR statistic. Given the concerns raised by Young (2018), and the practical importance of the non-homoskedastic case, such an extension seems like an important topic for future work.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to Emily Oster and Jesse Shapiro for suggesting the tabulations and simulations based on published results, among other helpful comments, and to Michal Kolesar from bringing the results in the appendix of Imbens & Angrist (1994) to our attention. We are also grateful to the participants in the NBER 2018 Summer Institute Methods Lectures (where we used a draft of this review as lecture notes) and to Donald W.K. Andrews, Adam McCloskey, Marcelo Moreira, and Carolin Pflueger for helpful comments. Isaiah Andrews gratefully acknowledges support from the National Science Foundation under grant number 1654234.

References

Anderson T, Rubin H. 1949. Estimators for the parameters of a single equation in a complete set of stochastic equations. *Annals of Mathematical Statistics* 21:570–582

Andrews D. 2017. Identification-robust subvector inference. Unpublished Manuscript

Andrews D, Cheng X, Guggenberger P. 2018a. Generic results for establishing the asymptotic size of confidence sets and tests. Cowles Foundation working paper 1813

Andrews D, Guggenberger P. 2009. Asymptotic size and a problem with subsampling and the m out of n bootstrap. *Econometric Theory* 26:426–468

Andrews D, Guggenberger P. 2015. Identification- and singularity-robust inference for moment condition models. Unpublished Manuscript

Andrews D, Guggenberger P. 2017. Asymptotic size of kleibergen's lm and conditional lr tests for moment condition models. *Econometric Theory* 33:1046–1080

Andrews D, Marmer V, Yu Z. 2018b. On optimal inference in the linear iv model. *Quantitative Economics* Forthcoming

Andrews D, Moreira M, Stock J. 2004. Optimal invariant similar tests of instrumental variables regression. Unpublished Manuscript

Andrews D, Moreira M, Stock J. 2006. Optimal two-sided invariant similar tests of instrumental variables regression. *Econometrica* 74:715–752

Andrews D, Moreira M, Stock J. 2008. Efficient two-sided nonsimilar invariant tests in iv tegression with weak instruments. *Journal of Econometrics* 146:241–254

Andrews I. 2016. Conditional linear combination tests for weakly identified models. *Econometrica* 84:2155–2182

Andrews I. 2018. Valid two-step identification-robust confidence sets for gmm. *Review of Economics and Statistics* 100:337–348

Andrews I, Armstrong TB. 2017. Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics* 8:479–503

Andrews I, Mikusheva A. 2016. Conditional inference with a functional nuisance parameter. *Econometrica* 84:1571–1612

Angrist J, Krueger A. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106:979–1014

Baum C, Schaffer M, Stillman S. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal* 7:465–506

Bound J, Jaeger D, Baker R. 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* 90:443–450

Breusch T, Pagan A. 1980. The lagrange multiplier test and its applications to model specifications in econometrics. *Econometrica* 47:239–253

Chaudhuri S, Zivot E. 2011. A new method of projection-based inference in gmm with weakly identified nuisance parameters. *Journal of Econometrics* 164:239–251

Chernozhukov V, Jansson M, Hansen C. 2009. Admissible invariant similar tests for instrumental variables regression. *Econometric Theory* 25:806–818

Conley T, Hansen C, Rossi P. 2012. Plausibly exogenous. *Review of Economics and Statistics* 94:260–272

Cragg J, Donald S. 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory* 9:222–240

Davidson R, MacKinnon J. 2014. Bootstrap confidence sets with weak instruments. *Econometric Reviews* 33:651–675

Dufour J. 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* 65:1365–1387

Dufour J, Jasiak J. 2001. Finite sample limited information inference methods for structural equations and models with generated regressors. *International Economic Review* 42:815–844

Dufour J, Taamouti M. 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica* 73:1351–1365

Fieller E. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* 16:175–185

Gleser L, Hwang J. 1987. The nonexistence of 100(1-α)% confidence sets of finite expected diameter in errors-in-variables and related models. *Journal of the American Statistical Association* 15:1341–1362

Guggenberger P. 2012. On the asymptotic size distortion of tests when instruments locally violate the exogenity assumption. *Econometric Theory* 28:387–421

Guggenberger P, Kleibergen F, Mavroeidis S. 2018. A more powerful subvector anderson rubin test in linear instrumental variable regression. Unpublished Manuscript

Guggenberger P, Kleibergen F, Mavroeidis S, Chen L. 2012. On the asymptotic sizes of subset anderson–rubin and lagrange multiplier tests in linear instrumental variables regression. *Econometrica* 80:2649–2666

Hausman JA. 1983. Specification and estimation of simultenous equation models. In *Handbook of Econometrics*, eds. Z Grilliches, M Intriligator, vol. 1. North-Holland

Hirano K, Porter J. 2015. Location properties of point estimators in linear instrumental variables and related models. *Econometric Reviews* 34:720–733

Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467–475

Kleibergen F. 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica* 70:1781–1803

Kleibergen F. 2005. Testing parameters in gmm without assuming they are identified. *Econometrica* 73:1103–1123

Kleibergen F. 2007. Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices, and identification statistics. *Journal of Econometrics* 139:181–216

Kleibergen F, Paap R. 2007. Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics* 133:97–126

Lee J. 2015. Asymptotic sizes of subset anderson-rubin tests with weakly identified nuisance parameters and general covariance structure. Unpublished Mauscript

Lee S. 2018. A consistent variance estimator for 2sls when instruments identify different lates. *Journal of Business and Economic Statistics* Forthcoming

Lehmann E, Romano J. 2005. Testing statistical hypotheses. Springer, 3rd ed.

Magnusson L. 2010. Inference in limited dependent variable models robust to weak identification. *Econometrics Journal* 13:S56–S79

Mariano R, Sawa T. 1972. The exact finite-sample distribution of the limited-information maximum likelihood estimator in the case of two included endogenous variables. *Journal of the American Statistical Association* 67:159–163

Mikusheva A. 2010. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics* 157:236–247

Montiel Olea J. 2017. Admissible, similar tests: A characterization. Unpublished Manuscript

Montiel Olea J, Pflueger C. 2013. A robust test for weak instruments. *Journal of Business and Economic Statistics* 31:358–369

Moreira H, Moreira M. 2013. Contributions to the theory of optimal tests. Unpublished Manuscript

Moreira H, Moreira M. 2015. Optimal two-sided tests for instrumental variables regression with heteroskedastic and autocorrelated errors. Unpublished Manuscript

Moreira M. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71:1027–1048

Moreira M. 2009. Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics* 152:131–140

Moreira M, Porter J, Suarez G. 2009. Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics* 149:52–64

Moreira M, Ridder G. 2017. Optimal invariant tests in an instrumental variables regression with heteroskedastic and autocorrelated errors. Unpublished Manuscript

Nagar A. 1959. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica* 27:575–595

Nelson C, Startz D, Zivot E. 1998. Valid confidence regions and inference in the presence of weak instruments. *International Economic Review* 39:1119–46

Nelson C, Startz R. 1990a. The distribution of the instrumental variable estimator and its t-ratio when the instrument is a poor one. *Journal of Business* 63:5125–5140

Nelson C, Startz R. 1990b. Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58:967–976

Sanderson E, Windmeijer F. 2016. A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics* 190:212–221

Sawa T. 1969. The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association* 64:923–937

Staiger D, Stock J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65:557–586

Stock J, Wright J. 2000. Gmm with weak identification. *Econometrica* 68:1055–1096

Stock J, Yogo M. 2005. Identification and inference for econometric models: Essays in honor of thomas rothenberg, chap. Testing for Weak Instruments in Linear IV Regression. Cambridge University Press, 80–108

Wang W, Tchatoka FD. Forthcoming. On bootstrap inconsistency and bonferroni-based size-correction for the subset anderson–rubin test under conditional homoskedasticity. *Journal of Econometrics*

Young A. 2018. Consistency without inference: Instrumental variables in practical application. Unpublished Manuscript

Zhu Y. 2015. A new method for uniform subset inference of linear instrumental variables models. Unpublished Manuscript