IV Estimation in Finite Samples

Standard asymptotic theory can provide a very poor approximation to the finite-sample properties of IV estimates.

- Parameter estimates may be strongly biased, with distributions that are far from normal.
- Reported standard errors may be much too small.
- In consequence of these two features of IV estimation,
  - Test statistics may frequently reject hypotheses that are true.
  - Confidence intervals may severely under-cover.
- These problems are most serious when the instruments are "weak." They get worse as the number of instruments increases.

There is an enormous literature on this subject. See Isaiah Andrews, James H. Stock, and Liyang Sun, "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics*, 11, 2019, 727–753.

Consider a two-equation system with one endogenous r.h.s. variable

$$y_1 = \beta y_2 + Z\gamma + u_1, \tag{1}$$
$$y_2 = Z\pi_1 + W_2\pi_2 + u_2. \tag{2}$$

Here (1) is a **structural** equation and (2) is a **reduced-form** equation.

In (1) and (2), $W = [Z \;\; W_2]$ has l columns, and Z has $k-1$ columns, so that there are $l-k$ over-identifying restrictions.

We could also write (1) as $y_1 = X\beta + u_1$, where $X = [y_2 \;\; Z]$ and $\beta = [\beta \;\; \gamma^\top]^\top$. The coefficient we really care about is β.

The entire **unrestricted reduced form** consists of two equations:

$$y_1 = Z\pi_{11} + W_2\pi_{12} + v_1 \tag{3}$$
$$y_2 = Z\pi_{21} + W_2\pi_{22} + v_2 \tag{4}$$

Here (4) is just (2) with different notation, and (3) is the equivalent for y_1; perhaps confusingly, v_2 in (4) is u_2 in (2).

We can combine (3) and (4) as

$$\boldsymbol{Y} = \boldsymbol{W\Pi} + \boldsymbol{V}. \tag{5}$$

However, the only part of this that we care about is (2), or (4).

The reduced-form equation (2) is used to obtain the fitted values

$$\hat{\boldsymbol{y}}_2 = \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2 = \boldsymbol{Z}\hat{\boldsymbol{\pi}}_1 + \boldsymbol{W}_2\hat{\boldsymbol{\pi}}_2, \tag{6}$$

which are then used to compute the IV (or 2SLS) estimates

$$\begin{bmatrix} \hat{\beta}_{\text{IV}} \\ \hat{\gamma}_{\text{IV}} \end{bmatrix} = (\boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_1. \tag{7}$$

The formula for $\hat{\beta}_{\text{IV}}$ by itself is

$$\hat{\beta}_{\text{IV}} = (\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2)^{-1} \boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_1. \tag{8}$$

Many theoretical papers assume that $\boldsymbol{X} = \boldsymbol{y}_2$, in which case

$$\hat{\beta}_{\text{IV}} = (\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2)^{-1} \boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_1. \tag{9}$$

In this special case, assuming i.i.d. disturbances, there are $l-1$ over-identifying restrictions, and

$$\text{Var}(\hat{\beta}_{\text{IV}}) = \sigma^2(\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2)^{-1}. \tag{10}$$

Setting $\boldsymbol{X} = \boldsymbol{y}_2$ gives the impression that what matters for $\text{Var}(\hat{\beta})$ (and for finite-sample properties) is the ability of $\boldsymbol{W}$ to explain $\boldsymbol{y}_2$.

But what actually matters is the ability of $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W}_2$ to explain $\boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y}_2$. This is the additional explanatory power of $\boldsymbol{W}_2$ in the regression

$$\boldsymbol{y}_2 = \boldsymbol{Z}\boldsymbol{\pi}_1 + \boldsymbol{W}_2\boldsymbol{\pi}_2 + \boldsymbol{v}_2. \tag{11}$$

Because $\mathcal{S}(\boldsymbol{Z}) \subset \mathcal{S}(\boldsymbol{W})$, $\boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}$.

Thus we can rewrite (8) as

$$\hat{\beta}_{\text{IV}} = \left(\boldsymbol{y}_2^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}) \boldsymbol{y}_2\right)^{-1} \boldsymbol{y}_2^\top (\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}) \boldsymbol{y}_1. \tag{12}$$

To obtain a general expression for $\text{Var}(\hat{\beta})$ in the i.i.d. case, we just need to replace $\boldsymbol{P}_{\boldsymbol{W}}$ in (10) by $\boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{Z}}$.

It is surprisingly common for there to be only one endogenous regressor and only one instrument.

In this case, as we saw previously, the IV estimator $\hat{\beta}_{\text{IV}}$ is just $\boldsymbol{y}_1^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{w} / \boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{w}$, the ratio of $\hat{\gamma}_1$ to $\hat{\gamma}_2$ from the regressions

$$\boldsymbol{y}_1 = \gamma_1 \boldsymbol{w} + \boldsymbol{Z}\boldsymbol{\pi}_1 + \boldsymbol{u}_1, \tag{13}$$

$$\boldsymbol{y}_2 = \gamma_2 \boldsymbol{w} + \boldsymbol{Z}\boldsymbol{\pi}_2 + \boldsymbol{u}_2. \tag{14}$$

Since $\boldsymbol{y}_2^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{w}$ is in the denominator and can be of either sign, this ratio can be extremely large.

Thus it should not be surprising that, in this case, $\hat{\beta}_{\text{IV}}$ has no moments.

In general, the GIV estimator has at most $l - k$ moments, where $l - k$ is the number of over-identifying restrictions.

Asymptotic theory may work poorly for estimators with no or few moments, and bootstrap methods must be used with care.

Methods based on quantiles of bootstrap distributions are likely to be much more reliable than ones based on moments.

Weak Instruments

For the model (1) and (2), finite-sample properties depend on

1. the number of over-identifying restrictions, which is $l - k$, or, equivalently, the number of instruments, which is $l - k + 1$;
2. the correlation ρ between the elements of $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$;
3. the strength (or weakness) of the instruments.

A measure of how strong the instruments are is the **concentration parameter**,

$$\Lambda = \frac{1}{\sigma_2^2} \boldsymbol{\pi}_2^\top \boldsymbol{W}_2^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{W}_2 \boldsymbol{\pi}_2. \tag{15}$$

This is the variation in $\boldsymbol{W}_2 \boldsymbol{\pi}_2$ that cannot be explained by $\boldsymbol{Z}$, divided by the variance of the u_{i2}. It is a form of noncentrality parameter, or NCP.

Weak-instrument asymptotics assumes that the concentration parameter Λ stays constant as $N \to \infty$.

This means that the elements of π_2 are assumed to be $O(N^{-1/2})$.

In contrast, under conventional (strong-instrument) asymptotics, π_2 is fixed and $\Lambda = O(N)$.

There is also an extensive literature on **many weak instruments** in which $l \to \infty$ as $N \to \infty$.

The literature on IV estimation exemplifies the importance of the **asymptotic construction** that we choose to use.

We can estimate (15) by running the reduced-form regression (11).

- When the F statistic for $\pi_2 = \mathbf{0}$ in this equation is large, it is probably reasonable to rely on IV estimates and standard errors.
- When this **first-stage *F* statistic** is small, it is dangerous to do so.

Stock and Yogo (2005) suggest computing this F statistic.

They actually allow for $m \geq 1$ endogenous regressors, so there are m reduced-form regressions. Test statistic is more complicated if $m > 1$.

S & Y claim that it is safe to rely on asymptotic theory for $\hat{\beta}_{\text{IV}}$ when this F statistic is large (say, $F > 10$) and unsafe when F is small.

They provide several tables, and the required value of F can be much larger than 10. But they are all based on worst-case assumptions.

When $l - k$ is large (many over-identifying restrictions), $\boldsymbol{W}_2$ may need a lot of explanatory power to make F large enough.

Consider again the simple case in which the only regressor in the structural equation is $\boldsymbol{y}_2$.

In this case, the IV estimator is

$$\hat{\beta}_{\text{IV}} = \frac{\boldsymbol{y}_1^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2}{\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2}, \tag{16}$$

and its estimated variance, under the i.i.d. assumption, is

$$\widehat{\text{Var}}(\hat{\beta}_{\text{IV}}) = \hat{\sigma}^2 (\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}} \boldsymbol{y}_2)^{-1}. \tag{17}$$

If we replace y_1 by $\beta y_2 + u_1$ and y_2 by $W\pi + u_2$, we find that

$$\begin{aligned}\hat{\beta}_{\text{IV}} &= \frac{(\beta(W\pi + u_2) + u_1)^\top P_W (W\pi + u_2)}{(W\pi + u_2)^\top P_W (W\pi + u_2)} \\ &= \frac{\beta\pi^\top W^\top W\pi + 2\beta u_2^\top W\pi + \beta u_2^\top P_W u_2 + u_1^\top W\pi + u_1^\top P_W u_2}{\pi^\top W^\top W\pi + 2\pi^\top W^\top u_2 + u_2^\top P_W u_2}.\end{aligned}$$

Under conventional asymptotics, the quadratic form $\pi^\top W^\top W\pi$ is $O_p(N)$, the terms involving one of u_1 or u_2 are $O_p(N^{1/2})$, and the terms with P_W in the middle are $O_p(1)$.

Thus $\hat{\beta}_{\text{IV}} \stackrel{a}{=} \beta_0$, because the factors of $\pi^\top W^\top W\pi$ cancel out.

Under weak-instrument asymptotics, however, every term is $O_p(1)$.

Conventional asymptotics provides a good guide if $\pi^\top W^\top W\pi$ is large relative to the other terms.

Hence the importance of the concentration parameter Λ. The first-stage F statistic is an (inconsistent) estimate of $\Lambda/(l-k)$.

If we subtract β from $\hat{\beta}_{\text{IV}}$ and rescale everything, we obtain

$$\text{N}^{1/2}(\hat{\beta}_{\text{IV}} - \beta) = \frac{N^{-1/2}\boldsymbol{u}_1^\top \boldsymbol{W}\boldsymbol{\pi} + N^{-1/2}\boldsymbol{u}_1^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{u}_2}{N^{-1}(\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{\pi} + 2\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{u}_2 + \boldsymbol{u}_2^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{u}_2)}. \tag{18}$$

Under conventional asymptotics, only the first terms in the numerator and denominator matter asymptotically. But under weak-instrument asymptotics, they all do.

Thus, under conventional asymptotics, $N^{1/2}(\hat{\beta}_{\text{IV}} - \beta)$ asymptotically has mean zero and variance

$$\text{Var}\left(N^{1/2}(\hat{\beta}_{\text{IV}} - \beta)\right) \stackrel{a}{=} \sigma^2(N^{-1}\boldsymbol{\pi}^\top \boldsymbol{W}^\top \boldsymbol{W}\boldsymbol{\pi})^{-1}. \tag{19}$$

Compare this with $\widehat{\text{Var}}(\beta_{\text{IV}}) = \hat{\sigma}^2(\boldsymbol{y}_2^\top \boldsymbol{P}_{\boldsymbol{W}}\boldsymbol{y}_2)^{-1}$, which equals

$$\hat{\sigma}^2(\hat{\boldsymbol{\pi}}^\top \boldsymbol{W}^\top \boldsymbol{W}\hat{\boldsymbol{\pi}} + 2\hat{\boldsymbol{\pi}}^\top \boldsymbol{W}^\top \hat{\boldsymbol{u}}_2 + \hat{\boldsymbol{u}}_2^\top \boldsymbol{P}_{\boldsymbol{W}}\hat{\boldsymbol{u}}_2)^{-1}. \tag{20}$$

For this to work well, we again need Λ to be large. If not, terms other than $\hat{\boldsymbol{\pi}}^\top \boldsymbol{W}^\top \boldsymbol{W}\hat{\boldsymbol{\pi}}$ will be important.

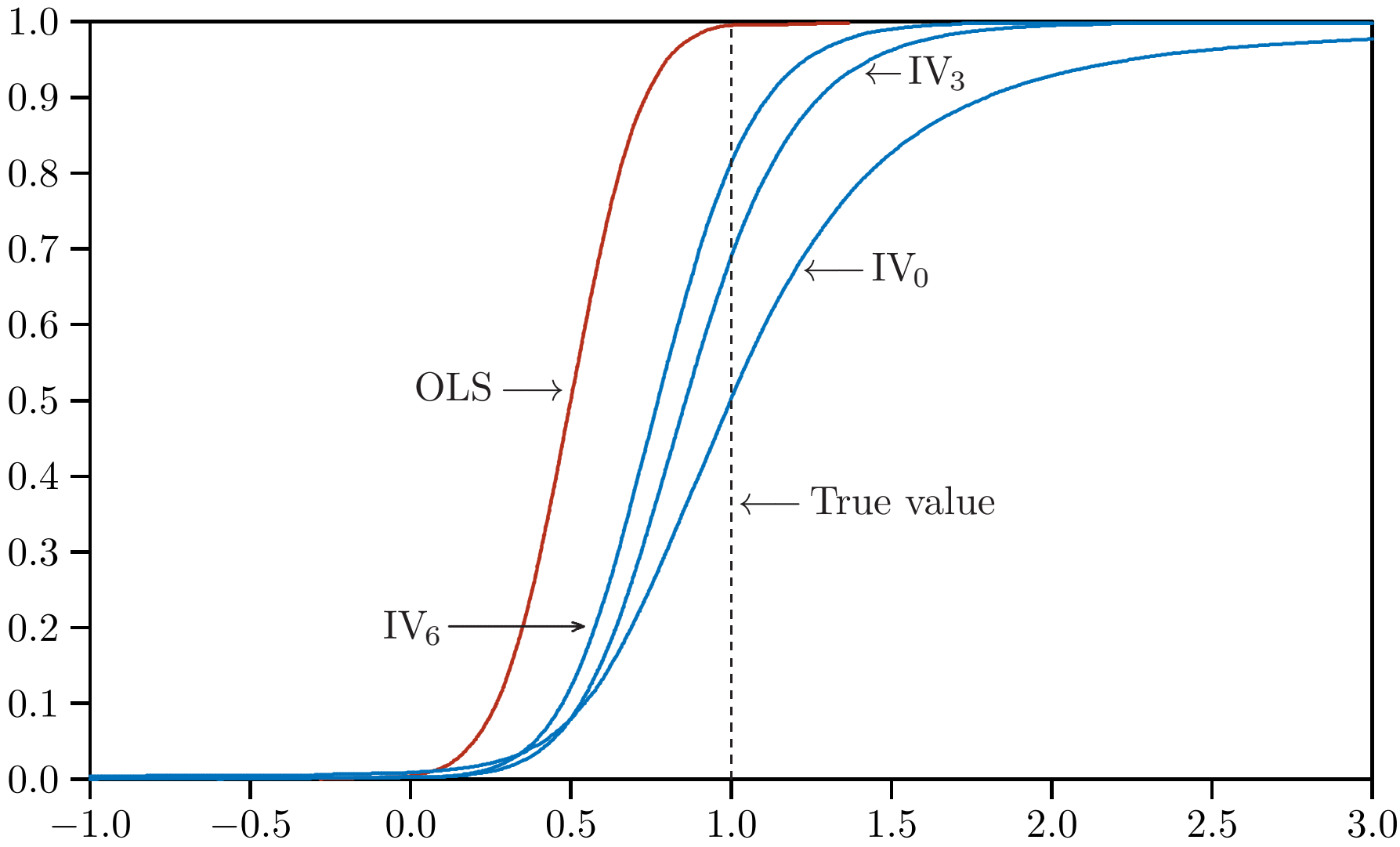
Figure 1 shows CDFs (from simulations) of OLS estimator and three IV estimators in a simple case. The three IV estimators, IV_0, IV_3, and IV_6, have $l - k$ equal to 0, 3, and 6, respectively.

We are estimating the slope parameter from an equation with one endogenous regressor and a constant term; its true value is 1. The sample size is only 25 so as to make finite-sample biases very apparent.

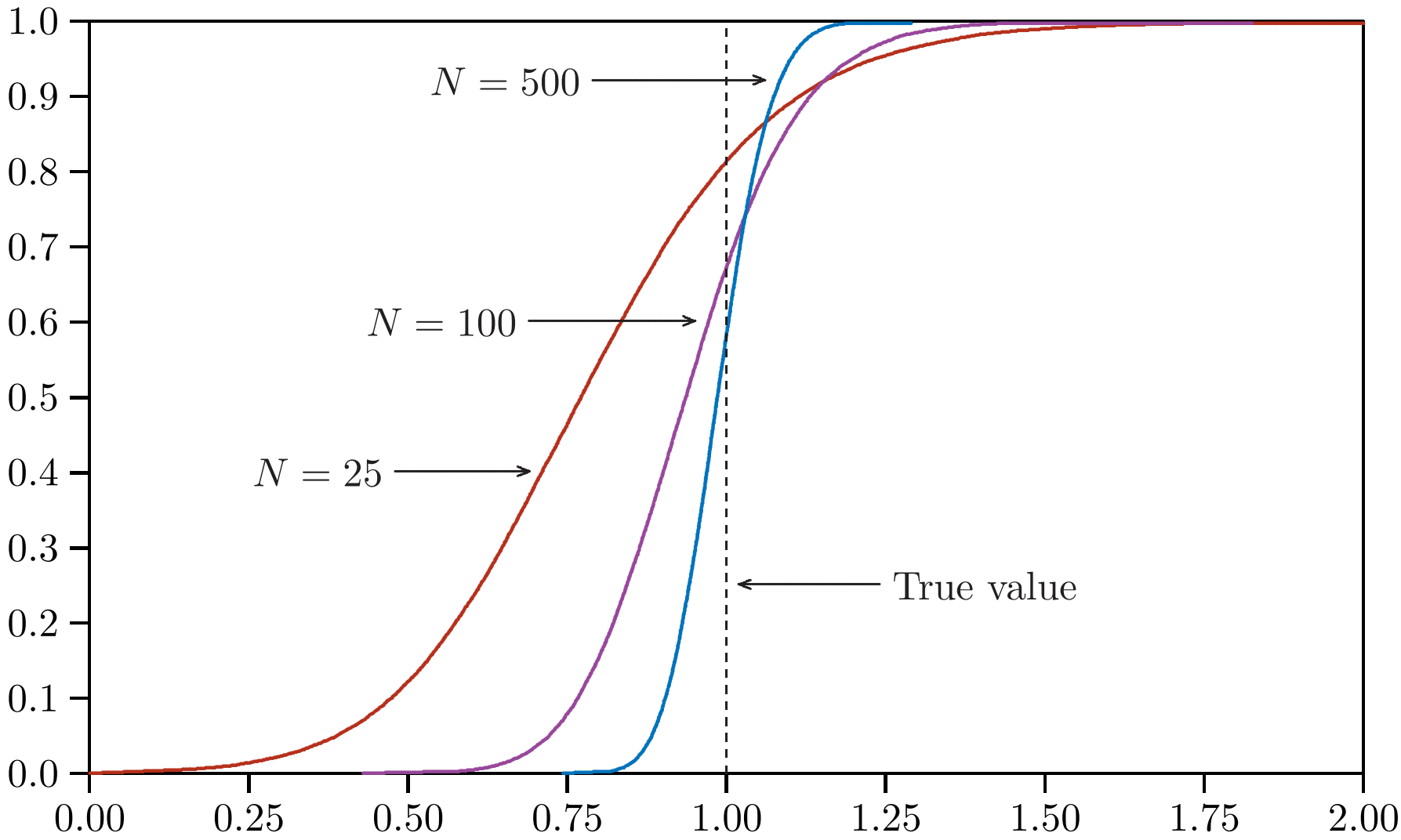
- OLS is severely biased but not very dispersed; IV_0 has approximately the right median but is extremely dispersed.
- CDFs for IV_3 and IV_6 mostly lie between those for OLS and IV_0 and have much thinner tails than the latter.

The effect of increasing the sample size is shown in Figure 2, which shows the distribution of IV_6 for $N = 25$, $N = 100$, and $N = 500$.

As N increases, the variance and bias of the estimator both decrease, as expected. However, even for $N = 500$, bias is noticeable. About 58% of the estimates are greater than the true value of 1.



1. Distributions of OLS and IV estimates, $N = 25$



2. Distributions of IV_6 estimates for several sample sizes

Since the Stock-Yogo rule of thumb ignores ρ, it can be very misleading when ρ is small.

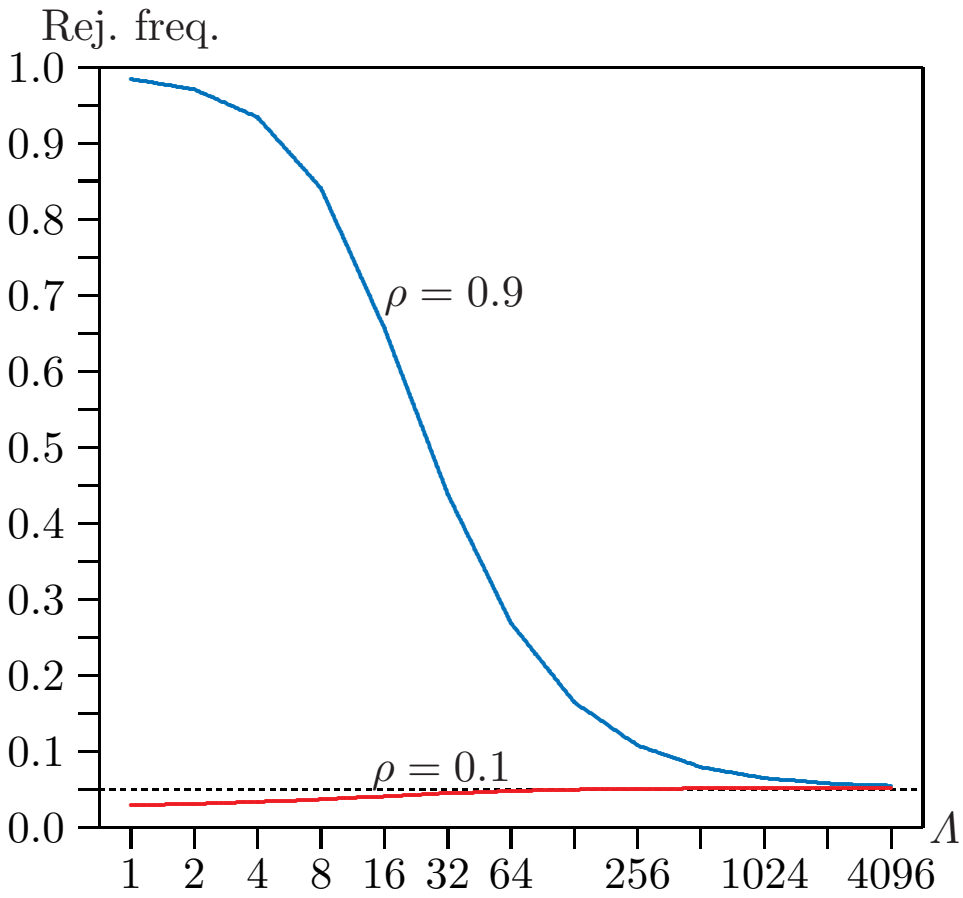
The Stock-Yogo rule of thumb is for i.i.d. disturbances. It does not work with heteroskedasticity or clustering.

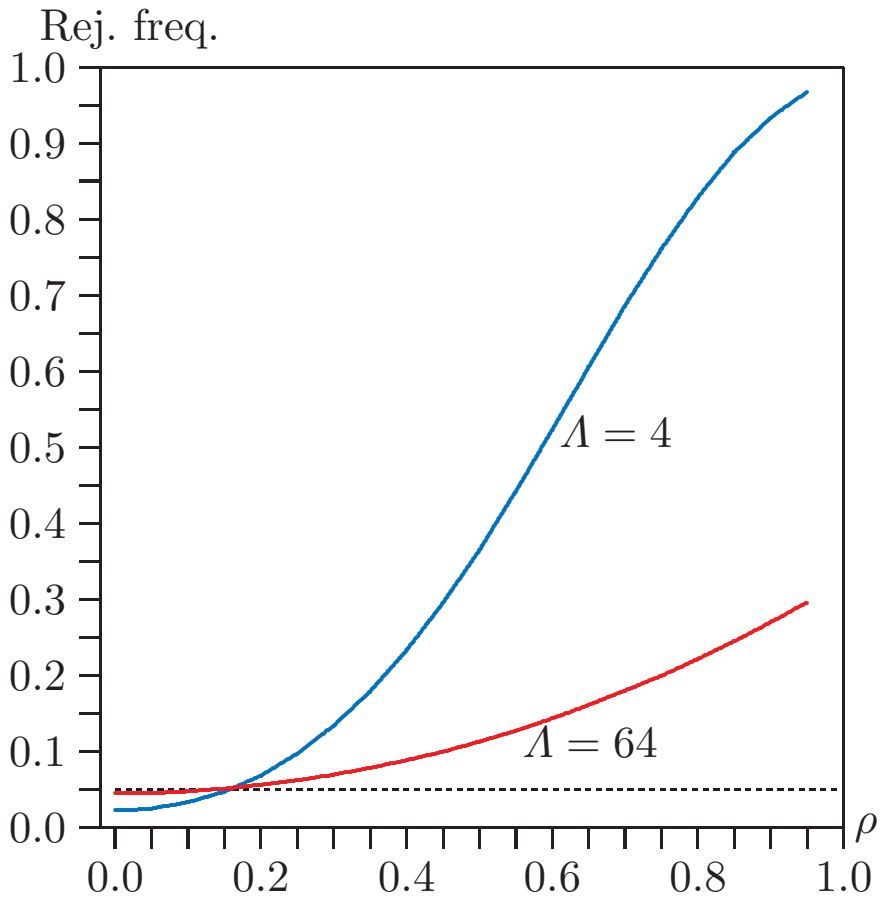
Olea and Pflueger (JBES, 2013) generalizes Stock and Yogo (2005) to these cases. Their procedure replaces the first-stage F statistic with something more complicated.

Unless the Olea-Pflueger statistic is quite large, reliable inference can be challenging. Results of Young (2022) suggest that it may need to be even larger than their paper claims.

Interest often focuses on t-tests and/or confidence intervals instead of $\hat{\beta}_{\text{IV}}$ itself. Rejection frequencies for some IV t-tests under the null hypothesis are shown in Figure 3.

There is extreme dependence on both Λ and ρ. Tests actually under-reject for small Λ when ρ is small.





The vertical axis shows rejection frequencies for asymptotic t tests. The numbers are taken from Davidson and MacKinnon (JBES, 2010).

$N = 400$ and $l - k = 11$. There are 100,000 replications.

Lee, McCrary, Moreiro, and Porter, "Valid t-ratio inference for IV," *AER*, 2022, 3260–3290 proposes procedure based on Stock-Yogo theory for the exactly identified case. Key ingredients are

$$\hat{t} = \frac{\hat{\beta} - \beta_0}{\widehat{\text{se}}(\hat{\beta})} \quad \text{and} \quad \hat{F} = \frac{\hat{\pi}^2}{\hat{V}(\hat{\pi})}. \tag{21}$$

Here $\hat{t}$ is the t-statistic on $\hat{\beta}$, the IV estimate of β in the structural equation, and $\hat{F}$ is the squared t-statistic for $\pi = 0$ in the reduced form.

Both $\widehat{\text{se}}(\hat{\beta})$ and $\hat{V}(\hat{\pi})$ may be hetero-robust or cluster-robust.

The tF procedure of LMMP rejects the null that $\beta = \beta_0$ whenever $\hat{t}$ exceeds a critical value $c_\alpha(F)$ that depends on α and $\hat{F}$. Values of $c_\alpha(F)$ differ greatly for $\alpha = .05$ and $\alpha = .01$.

For $\hat{F} = 4.00$ and $\alpha = .05$, $c_\alpha(F) = 18.66$. For $\hat{F} = 10$, $c_\alpha(F) = 3.43$.

For $\hat{F} = 6.67$ and $\alpha = .01$, $c_\alpha(F) = 91.10$. For $\hat{F} = 10$, $c_\alpha(F) = 8.86$.

These critical values are insanely conservative when ρ is small.

Alwyn Young (2022), "Consistency without inference: Instrumental variables in practical application," *European Economic Review*, 147.

It studies 1309 IV regressions in 30 published papers with heteroskedasticity and or clustering.

- First-stage F statistics greater than 10 arise frequently by chance, especially for reduced-form regressions with high-leverage observations or clusters (the HL sample).
- Although IV estimates often differ substantially from OLS ones, they rarely do so significantly, and never for the HL sample.
- For the HL sample, bootstrap DWH tests rarely reject the null that OLS estimates are consistent.
- For the HL sample, IV t-tests are unreliable and power is low.
- Bootstrap inference (both pairs and wild) and jackknife inference are much more reliable than asymptotic inference.
- Bootstrap P values based on coefficients often outperform ones based on t-statistics.

The Anderson-Rubin Test

The **Anderson-Rubin test** (A & R, 1949) is sometimes advocated for testing hypotheses about β. The test statistic is

$$\text{AR}(\beta_0) = \frac{N-l}{l-k+1} \frac{(\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2)^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2)}{(\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2)^\top \boldsymbol{M}_W (\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2)}, \tag{22}$$

where $\boldsymbol{P}_1 \equiv \boldsymbol{M}_Z - \boldsymbol{M}_W = \boldsymbol{P}_W - \boldsymbol{P}_Z$.

Under classical assumptions, with $\beta = \beta_0$, $\text{AR}(\beta_0)$ follows the $F(l-k+1, N-l)$ distribution. After all, it is just an F test.

The denominator of $\text{AR}(\beta_0)$, times $1/(N-l)$, is the SSR from a regression of $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$ on $\boldsymbol{W}$.

It is a valid estimator of σ_1^2, since the columns of $\boldsymbol{W}$ that do not belong to $\mathcal{S}(\boldsymbol{Z})$ should have no real explanatory power.

The AR test has attracted a good deal of attention because it has good finite-sample properties even when the instruments are weak.

But the AR test is really testing two different restrictions:

1. that $\beta = \beta_0$;
2. that the instruments, $\boldsymbol{W}_2$, that is, the regressors in $\boldsymbol{W}$ but not in $\boldsymbol{Z}$, have no explanatory power.

The numerator of the AR statistic can be rewritten as

$$((\hat{\beta}_{\text{IV}} - \beta_0)\boldsymbol{y}_2)^\top \boldsymbol{P}_1 ((\hat{\beta}_{\text{IV}} - \beta_0)\boldsymbol{y}_2) + (\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2)^\top \boldsymbol{P}_1 (\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2). \quad (23)$$

The first term here is testing whether $\beta = \beta_0$.

The second term is the numerator of the Sargan statistic, since

$$(\boldsymbol{M}_{\boldsymbol{Z}} - \boldsymbol{M}_{\boldsymbol{W}})(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2) = \boldsymbol{P}_{\boldsymbol{W}}(\boldsymbol{y}_1 - \hat{\beta}_{\text{IV}}\boldsymbol{y}_2), \quad (24)$$

because $\boldsymbol{Z}$ and $\boldsymbol{W}$ are orthogonal to the IV residuals.

Thus, when $l - k > 0$, large values of the Sargan statistic are associated with large values of the AR statistic.

The AR test has good finite-sample properties under the null, even with weak instruments, but it can be hard to interpret when $l - k > 0$.

- It tends to lack power against the alternative that $\beta \neq \beta_0$, especially when $l - k$ is large.
- When it does reject the null, with $l - k > 0$, it may well do so because the instruments are invalid and not because $\beta \neq \beta_0$.

Never invert an AR test (with $l - k > 0$) to obtain a confidence interval for β! It is fine to do so when there are no over-identifying restrictions.

- With weak instruments, the interval may cover the entire real line, or it may cover most of the real line but with a hole in the middle.
- When the Sargan statistic is large, the interval may be very short, or even empty.
- See Davidson & MacKinnon (2010, 2014) and Müller and Norets (2016).

A Bootstrap Method for IV Regression

Using the bootstrap to make inferences about β requires a bootstrap DGP for equations (1) and (2). A good one was proposed by Davidson and MacKinnon (JBES, 2010).

We can rewrite these two equations as

$$\boldsymbol{y}_1 = \beta \boldsymbol{y}_2 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}_1, \tag{25}$$

$$\boldsymbol{y}_2 = \boldsymbol{W}\boldsymbol{\pi} + \boldsymbol{u}_2. \tag{26}$$

Unless we are going to use the pairs bootstrap, which works poorly, we need to specify β, γ, $\boldsymbol{\pi}$, and how to generate the u_{1i}^* and u_{2i}^*.

As usual, we should impose the restriction(s) we want to test.

This means setting $\beta = \beta_0$. Then $\gamma = \tilde{\gamma}$ is obtained by an OLS regression of $\boldsymbol{y}_1 - \beta_0 \boldsymbol{y}_2$ on $\boldsymbol{Z}$.

This regression also yields residuals $\tilde{\boldsymbol{u}}_1$, which we will need.

The obvious way to estimate $\boldsymbol{\pi}$ is to regress $\boldsymbol{y}_2$ on $\boldsymbol{W}$. But this yields inefficient estimates, and the resulting bootstrap works poorly.

Instead, we regress $\boldsymbol{y}_2$ on $\boldsymbol{W}$ and $\tilde{\boldsymbol{u}}_1$. This is asymptotically equivalent to estimating (25) and (26) by FIML (which is really LIML in this case).

This regression yields $\tilde{\boldsymbol{\pi}}$ and $\tilde{\boldsymbol{u}}_2 \equiv \boldsymbol{y}_2 - \boldsymbol{W}\tilde{\boldsymbol{\pi}}$. Note that $\tilde{\boldsymbol{u}}_2$ is not the vector of residuals from the regression used to obtain $\tilde{\boldsymbol{\pi}}$.

The DGP for the **wild restricted efficient**, or **WRE**, bootstrap is

$$y_{1i}^* = \beta_0 y_{2i}^* + \boldsymbol{Z}_i\tilde{\boldsymbol{\gamma}} + \tilde{u}_{1i}v_i^* \tag{27}$$

$$y_{2i}^* = \boldsymbol{W}_i\tilde{\boldsymbol{\pi}} + \tilde{u}_{2i}v_i^*, \tag{28}$$

where v_i^* follows the Rademacher distribution. Note same v_i^*!

We may be able to improve finite-sample performance slightly by rescaling the $\tilde{u}_{1i}$ and the $\tilde{u}_{2i}$.

If there is clustering as well as heteroskedasticity, then the bootstrap DGP must generate the data cluster by cluster.

The **WCRE** bootstrap DGP is

$$\boldsymbol{y}_{1g}^* = \beta_0 \boldsymbol{y}_{2g}^* + \boldsymbol{Z}_i \tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{u}}_{1g} v_g^* \tag{29}$$

$$\boldsymbol{y}_{2g}^* = \boldsymbol{W}_g \tilde{\boldsymbol{\pi}} + \tilde{\boldsymbol{u}}_{2g} v_g^*. \tag{30}$$

As with WCR bootstrap, same value of the auxiliary random variable, v_g^*, is used for both equations and every observation in cluster g.

Important features of W(C)RE bootstrap:

1. Reduced form equation (26) is estimated efficiently.
2. Structural equation (25) uses restricted (OLS) estimates instead of unrestricted (IV) ones.
3. The same v_i^* multiplies (possibly transformed) residuals for both equations. Bootstrap disturbances retain the correlation between structural and reduced-form residuals.

These procedures are implemented efficiently in `boottest`.

Of course, if we are allowing for heteroskedasticity or clustering, we need to use hetero-robust or cluster-robust standard errors.

The usual CRVE for $\hat{\beta}$ and $\hat{\gamma}$ jointly is

$$(\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} (\boldsymbol{P}_W \boldsymbol{X})_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top (\boldsymbol{P}_W \boldsymbol{X})_g \right) (\boldsymbol{X}^\top \boldsymbol{P}_W \boldsymbol{X})^{-1}. \tag{31}$$

If we just focus on the CRVE for $\hat{\beta}$, it is

$$\frac{\sum_{g=1}^{G} \left((\boldsymbol{P}_W - \boldsymbol{P}_Z) \boldsymbol{y}_2 \right)_g^\top \hat{\boldsymbol{u}}_g \hat{\boldsymbol{u}}_g^\top \left((\boldsymbol{P}_W - \boldsymbol{P}_Z) \boldsymbol{y}_2 \right)_g}{\left(\boldsymbol{y}_2^\top (\boldsymbol{P}_W - \boldsymbol{P}_Z) \boldsymbol{y}_2 \right)^2}. \tag{32}$$

The HCCME is a special case with $G = N$.