NBER WORKING PAPER SERIES

ONE INSTRUMENT TO RULE THEM ALL:
THE BIAS AND COVERAGE OF JUST-ID IV

Joshua Angrist
Michal Kolesár

Working Paper 29417
http://www.nber.org/papers/w29417

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2021, Revised January 2022

We thank Ahmet Gulek and Luther Yap for expert research assistance. Thanks also go to Tim Armstrong, Isaiah Andrews, Brigham Frandsen, Guido Imbens, Mike Keane, Dave Lee, Whitney Newey, and Steve Pischke for helpful discussions and insightful comments. Kolesár acknowledges support by the Sloan Research Fellowship and by the National Science Foundation Grant SES-22049356. The views expressed here are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV
Joshua Angrist and Michal Kolesár
NBER Working Paper No. 29417
October 2021, Revised January 2022
JEL No. C21,C26,C31,C36,J08

ABSTRACT

We argue that in microeconometric applications, just-identified instrumental variables (IV) estimators are virtually unbiased and the usual inference strategies are likely adequate. Confidence interval undercoverage exceeds 5% only for endogeneity beyond that seen even when IV and OLS estimates differ by an order of magnitude. Three widely-cited applications are used to explain why endogeneity is likely low enough for IV estimates to be reliable. IV identification typically implies a first-stage sign restriction; most analysts probably screen their estimates accordingly. We show that screening on the estimated first stage sign halves median bias of conventional IV without reducing coverage.

Joshua Angrist
Department of Economics, E52-436
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
angrist@mit.edu

Michal Kolesár
Department of Economics
278 Julis Romo Rabinowitz Building
Princeton University
Princeton, NJ 08544-1021
mkolesar@princeton.edu

1 Introduction

The Bekker (1994) and Bound et al. (1995) critiques of the heavily over-identified two-stage least squares (2SLS) estimates reported in Angrist (1990) and Angrist and Krueger (1991, AK91) sparked a flood of interest in the finite-sample behavior of instrumental variables (IV) estimates. In the intervening three decades, attention to bias in 2SLS estimates with many weak instruments has become a staple of applied microeconometrics. The fact that the finite-sample distribution of 2SLS estimates is shifted towards the ordinary least squares (OLS) probability limit is especially worrying. IV is often motivated by fears that OLS estimates are compromised by omitted variable bias (OVB). The IV analyst hopes, therefore, that when IV estimates are close to OLS, this signals minimal OVB rather than substantial finite-sample bias in the IV estimates.

The heavily over-identified models that prompted 1990s IV critiques are a case of many weak instruments, roped together by 2SLS in an effort to estimate a single causal effect with acceptable precision. Strikingly, Bound et al. (1995) show that in the specifications reported in AK91 that interact quarter of birth dummies with 10 year-of-birth and 50 state-of-birth dummies to generate 180 instruments, replacing real quarter of birth dummmies with fake dummies randomly drawn yields 2SLS estimates and standard errors much like those generated by the real thing.[1] But most studies using IV (including Angrist (1990) and AK91) report just-identified (just-ID) IV estimates computed with a single instrument. Just-ID IV estimates are less obviously biased than the estimates generated by heavily over-identified models, and the empirical relevance of bias in just-ID IV applications remains a matter of debate. Our analysis comes in the wake of recent and renewed interest in the finite-sample properties of just-ID IV, seen in Andrews and Armstrong (2017), Lee et al. (2020), and Keane and Neal (2021), among others.

We argue here that in typical microeconometric applications of just-ID IV, conventional IV estimates and t-tests are unlikely to be compromised by failures of conventional asymptotic theory. Our analysis builds on the (approximate) finite-sample normality of reduced-form and first-stage estimators (in the argot of classical simultaneous equations models, these are both estimated "reduced forms"). Our modeling framework parallels that in Andrews et al. (2019) and earlier theoretical investigations of weak instrument problems. The normality of reduced-form estimates is justified by conventional asymptotic reasoning, as well as by the local-to-zero asymptotic sequence used in Staiger and Stock (1997) and Stock and Yogo (2005), in which the first stage shrinks to zero at a rate inversely proportional to the square root of the sample size.

Our setup has only two free parameters: the correlation between structural and first-stage residuals (henceforth, "endogeneity") and the population first-stage F statistic. This fact lends itself to the construction of easily-interpreted rejection contours characterizing conventional second-stage t-tests and confidence interval coverage rates. We see, for example, that for endogeneity less than about 0.76, 95% confidence interval coverage is distorted by no more than 5% for *any* population F. An important insight here is that, even though bias increases when the first stage gets weaker, second-stage precision

[1] This "fake instruments" simulation was originally suggested by Alan Krueger. Although not an empirical study, Bekker (1994) is likewise motivated by a heavily over-identified specification in Angrist (1990) that uses 73 draft lottery dummies plus interaction terms as instruments for Vietnam-era veteran status. This application is featured at the end of Bekker's paper, and, originally, in an Amsterdam bar in 1992, where Paul Bekker first confronted Angrist with claims of bias.

falls. In contrast with the over-identified case, conventional just-ID IV standard errors reflect this, and confidence intervals widen accordingly. This keeps coverage high unless endogeneity is extraordinarily high.[2]

Few analysts can gauge endogeneity outside of a particular empirical context. We therefore use three applications to calibrate endogeneity: the AK91 study, the Angrist and Evans (1998, AE98) IV estimates using a dummy for samesex sibships as an instrument for family size, and the Angrist and Lavy (1999, AL99) fuzzy regression discontinuity estimates of class size effects on student learning. These studies span a range of OVB scenarios, from modest (for most of the AK91 estimates), to substantial (in AE98, where OLS exceeds IV by about 50%), to dramatic (in AL99, where IV exceeds small, insignificant OLS estimates by an order of magnitude). Yet, the absolute value of estimated endogeneity is no more than 0.46 in these applications, and over 0.35 only for a single specification and sample. Although three examples do not make a theorem, we argue that the features of these studies that limit endogeneity are common to empirical strategies designed to estimate causal effects or to mitigate attenuation bias in models with measurement error.

Our theoretical case for a sanguine view of conventional just-ID IV builds on Andrews and Armstrong's (2017) observation that a contemporary analyst pursuing a just-identified IV strategy likely has a first-stage direction in mind. In particular, Andrews and Armstrong (2017) show how to use a sign restriction on the population first stage to construct an estimator, denoted $\hat{\beta}_U$, that is unbiased when the reduced-form estimates are normally distributed. Given a theoretical first-stage sign restriction, it's tempting to apply this to first-stage *estimates*. We show, however, that conditional on the sign of the *estimated* first stage, $\hat{\beta}_U$ is no longer unbiased: as it turns out, $\hat{\beta}_U$ is unbiased by virtue of the fact that it averages two conditional estimators, each biased but in opposite directions.

This discouraging result raises the question of whether and how we might use sign restrictions to mitigate the bias of IV in a manner consistent with empirical practice. Our answer to this question comes in the form of a novel theoretical result showing that the median bias of just-ID IV is roughly halved if we condition on the first-stage estimate having the expected sign (we focus on median bias because the conventional just-ID IV estimator has no moments). Analysts might justifiably worry that screening on the basis of first-stage estimates distorts inference.[3] Perhaps surprisingly, we also show that pre-screening on the *sign* of the first stage is virtually costless: rejection contours for a sign-screened estimator differ little from those obtained without screening. The upshot is that sign-screening mitigates the already-modest bias of just-ID IV without degrading coverage. To the extent that such screening is a feature of modern empirical work, reported IV estimates reflect the impressively minimal bias characteristic of sign-screened IV.

The next section details the normal just-ID IV setup and derives an expression for endogeneity in terms of OLS OVB. Section 3 reviews the relationship between t-test rejection rates and the parameters that govern the normal model. This section also explains why endogeneity in applied

[2]The spirit of this argument differs from that in Stock and Yogo (2005), which focuses on worst-case rejection rates over all possible endogeneity values. Lee et al. (2020) and Keane and Neal (2021), discussed further in Section 3 below, consider possible restrictions on endogeneity. But these studies stress the fact that the standard definition of coverage doesn't involve restrictions on parameter values.

[3]Recent years have seen growing awareness of the bias induced by screening on the basis of a first-stage pre-test, such as the first-stage F statistic. Hall et al. (1996) appear to be the first to note that pre-testing on the first-stage F-statistic for just-ID IV reduces confidence interval coverage; Andrews et al. (2019) draw on recent empirical scholarship to demonstrate the empirical relevance of pre-test bias in IV applications.

microeconometrics is unlikely to be high enough for conventional IV inference to mislead. Section 4 shows how and why conditioning on the sign of the estimated first stage strengthens the case for this view. Section 5 concludes. Proofs and details behind numerical calculations appear in the appendix.

2 Setup

We observe a sample of n units indexed by i, with data on outcome variable, Y_i, a scalar treatment variable, D_i, a vector of covariates, X_i, and a scalar instrument, Z_i. Population regressions of outcome and treatment on the instrument and covariates define the reduced form and first stage. These are written as follows:

$$Y_i = Z_i\delta + X_i'\psi_1 + u_i, \tag{1}$$

$$D_i = Z_i\pi + X_i'\psi_2 + v_i. \tag{2}$$

The parameter of interest is $\beta = \frac{\delta}{\pi}$, the ratio of the reduced-form and first-stage regression coefficients on Z_i. Provided that the instrument, Z_i, satisfies an exclusion restriction and is relevant (i.e. $\pi \neq 0$), this parameter measures the effect of D_i on Y_i. More generally, if treatment effects are heterogeneous and a monotonicity condition holds, β is a weighted average of individual causal effects (Angrist & Imbens, 1995; Imbens & Angrist, 1994).

Let $\hat{\delta} = \sum_{i=1}^n \tilde{Z}_i Y_i / \sum_{i=1}^n \tilde{Z}_i^2$ and $\hat{\pi} = \sum_{i=1}^n \tilde{Z}_i D_i / \sum_{i=1}^n \tilde{Z}_i^2$ denote OLS estimates of δ and π, where $\tilde{Z}_i$ is the residual from the regression of Z_i on X_i. Under mild regularity conditions that allow the errors (u_i, v_i) to be non-normal, heteroskedastic, and serially or cluster-dependent, $(\hat{\delta}, \hat{\pi})$ is consistent and asymptotically normal as $n \to \infty$, with an asymptotic covariance matrix that can be consistently estimated. Importantly, this holds regardless of the strength of the instrument. We therefore follow Andrews et al. (2019) and earlier analyses of weak instrument problems by assuming this large-sample approximation holds exactly. Specifically, we assume:

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \pi\beta \\ \pi \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{\hat{\delta}}^2 & \sigma_{\hat{\delta}\hat{\pi}} \\ \sigma_{\hat{\delta}\hat{\pi}} & \sigma_{\hat{\pi}}^2 \end{pmatrix} \right), \tag{3}$$

with a known covariance matrix, Σ. This distributional assumption is implied by the Staiger and Stock (1997) weak-instrument asymptotic sequence (see Andrews et al. (2019, Section 3.2) for additional discussion and references). Finite-sample results under eq. (3) can therefore be seen as asymptotic under the Staiger and Stock (1997) sequence.

Equation (3) is our only substantive restriction; this assumption allows us to focus on the weak instrument problem, separating this from other finite-same problems, such as the effect of high-leverage observations on the quality of the normal approximation to the distribution of the OLS estimators $(\hat{\delta}, \hat{\pi})$ and the challenge of standard-error estimation with clustered data.[4] With (3) as foundation, we derive finite-sample properties of the IV estimator:

$$\hat{\beta}_{IV} = \frac{\hat{\delta}}{\hat{\pi}}, \tag{4}$$

[4] Young (2021) discusses these problems in an IV context.

and the null rejection rate of the corresponding Wald test. The latter is based on the t-statistic centered at the parameter of interest, β, divided by the estimated IV standard error, $\hat{\sigma}_{IV}$:

$$t_W = \frac{\hat{\beta}_{IV} - \beta}{\hat{\sigma}_{IV}}; \qquad \hat{\sigma}^2_{IV} = \frac{\sigma^2_{\hat{\delta}} - 2\sigma_{\hat{\delta}\hat{\pi}}\hat{\beta}_{IV} + \sigma^2_{\hat{\pi}}\hat{\beta}^2_{IV}}{\hat{\pi}^2}, \tag{5}$$

where $\hat{\sigma}^2_{IV}$ estimates the asymptotic variance of $\hat{\beta}_{IV}$ under standard $n \to \infty$ asymptotics. The corresponding theoretical variance is $\sigma^2_{IV} = (\sigma^2_{\hat{\delta}} - 2\sigma_{\hat{\delta}\hat{\pi}}\beta + \sigma^2_{\hat{\pi}}\beta^2)/\pi^2$. In a homoskedastic model with constant causal effects, this simplifies to the familiar formula

$$\sigma^2_{IV} = \frac{\sigma^2_{\epsilon}}{nE[\tilde{Z}^2_i]\pi^2},$$

where σ^2_{ϵ} is the variance of the residual in the structural equation,

$$Y_i = D_i\beta + X_i'(\psi_1 - \psi_2\beta) + \epsilon_i, \tag{6}$$

that motivates IV estimation in the classic linear set-up (the structural residual is $\epsilon_i = u_i - v_i\beta$).

Given the assumption of a known covariance matrix for the first-stage and reduced-form estimates, both t_W and $\hat{\sigma}^2_{IV}$ depend on the data only through $(\hat{\delta}, \hat{\pi})$. These have distributions determined by the two unknown parameters, π and β. Rather than π and β, however, it is illuminating to characterize finite-sample behavior in terms of a pair of parameters that measures instrument strength and the degree of endogeneity (a reparameterization adopted in Staiger and Stock (1997) and Lee et al. (2020), among others). The first parameter, denoted $E[F]$, is defined as:

$$E[F] = \pi^2/\sigma^2_{\hat{\pi}} + 1.$$

Because $E[F]$ is the expectation of $F = \hat{\pi}^2/\sigma^2_{\hat{\pi}}$, the F-statistic testing $\pi = 0$, it's sometimes called the *population first-stage F-statistic*, a term adopted here.

The second parameter is defined as:

$$\rho = \mathrm{cor}(\hat{\delta} - \hat{\pi}\beta, \hat{\pi}) = \frac{\sigma_{\hat{\pi}}}{\sqrt{\sigma^2_{\hat{\delta}} - 2\beta\sigma_{\hat{\delta}\hat{\pi}} + \sigma^2_{\hat{\pi}}\beta^2}} \times (\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}} - \beta). \tag{7}$$

Under independent heteroskedastic errors, ρ is also given by $\mathrm{cor}(\tilde{Z}_i\epsilon_i, \tilde{Z}_i v_i)$. When, in addition, the errors (u_i, v_i) are homoskedastic, $\rho = \mathrm{cor}(\epsilon_i, v_i)$, where ϵ_i is the structural residual in (6). We therefore refer to ρ as (the degree of) *endogeneity*.[5]

With weak instruments as well as homoskedastic error terms, ρ is proportional to the bias of the OLS estimand. This can be seen by using the first-stage and reduced-form equations to write the OLS

[5]This simplification is obtained using the fact that, under homoskedasticity, the variance of v_i is $\sigma^2_v = \sigma^2_{\hat{\pi}} \cdot nE[\tilde{Z}^2_i]$ and the variance of ϵ_i is $\sigma^2_{\epsilon} = (\sigma^2_{\hat{\delta}} - 2\beta\sigma_{\hat{\delta}\hat{\pi}} + \sigma^2_{\hat{\pi}}\beta^2) \cdot nE[\tilde{Z}^2_i]$, with $\mathrm{cov}(v_i, \epsilon_i) = (\sigma_{\hat{\delta}\hat{\pi}} - \beta\sigma^2_{\hat{\pi}}) \cdot nE[\tilde{Z}^2_i]$. The homoskedastic formula for the variance of ϵ_i also leads yields the simplification of the formula for σ^2_{IV} noted above.

slope coefficient, β_{OLS}, as follows:

$$\beta_{OLS} = \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{E[\tilde{Z}_i^2]\pi^2\beta + E[u_i v_i]}{E[\tilde{Z}_i^2]\pi^2 + \sigma_v^2}$$

$$= R_p^2\beta + (1-R_p^2)\frac{E[u_i v_i]}{\sigma_v^2}; \qquad R_p^2 = \frac{E[\tilde{Z}_i^2]\pi^2}{E[\tilde{Z}_i^2]\pi^2 + \sigma_v^2}, \tag{8}$$

where $\tilde{D}_i$ is the residual from a regression of D_i on X_i, and $\sigma_v^2 = E[v_i^2]$. The weight multiplying β in (8), denoted R_p^2, is the population partial R^2 generated by adding the instrument to the first-stage regression. When the instrument is weak, R_p^2 is close to zero, and (8) is approximately $E[u_i v_i]/\sigma_v^2$. The OLS estimand likewise converges to $E[u_i v_i]/\sigma_v^2$ in the Staiger and Stock (1997) weak-instrument sequence (which takes $\pi \to 0$). This in turn equals $\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}}^2$ under homoskedasticity, so the second term on the right-hand side of (7),

$$\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}}^2 - \beta = \beta_{WOLS} - \beta, \tag{9}$$

is the weak-instrument OVB of OLS (where we've introduced the notation β_{WOLS} for $\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}}^2$). Moreover, when $\pi = 0$, it follows from (3) that $\beta_{WOLS} - \beta$ is the median bias of $\hat{\beta}_{IV}$ with no independence or heteroskedasticity assumptions on the errors in (1) and (2).[6] Thus, ρ also measures endogeneity in the sense that it's proportional to the median bias of the IV estimator when the instrument is irrelevant.

3 Rejection Rates in Theory and Practice

We're interested in the bias of $\hat{\beta}_{IV}$ and in t-test rejection rates when the null hypothesis is true. Beginning with the latter, the null rejection rate for a two-sided t-test with level α is the probability that the absolute value of a t-statistic, $|t_W|$, exceeds $z_{1-\alpha/2}$, the $1-\alpha/2$ quantile of a standard normal distribution. This is:

$$R_W = P_{E[F],\rho}(|t_W| > z_{1-\alpha/2}),$$

where $P_{E[F],\rho}$ is the distribution of t_W parameterized by $E[F], \rho$. As detailed in Appendix A.2 (and paralleling Stock and Yogo, 2005), R_W is evaluated by rewriting t_W in terms of the Anderson and Rubin (1949, AR) statistic,

$$t_{AR} = \frac{\hat{\delta} - \hat{\pi}\beta}{\sqrt{\sigma_{\hat{\delta}}^2 - 2\sigma_{\hat{\delta}\hat{\pi}}\beta + \sigma_{\hat{\pi}}^2\beta^2}}, \tag{10}$$

and the first-stage t-statistic, $t_1 = \hat{\pi}/\sigma_{\hat{\pi}}$, the square of which is the first-stage F statistic. Note t_{AR} differs from t_W in that it replaces $\hat{\beta}_{IV}$ with the null value of β in the formula for $\hat{\sigma}_{IV}^2$.[7]

[6] Assumption (3) implies that we can write reduced form and first stage estimates as $\hat{\delta} = \pi\beta + (\sigma_{\hat{\delta}\hat{\pi}}/\sigma_{\hat{\pi}})\mathcal{Z}_\pi + (\sigma_{\hat{\delta}}^2 - \sigma_{\hat{\delta}\hat{\pi}}^2/\sigma_{\hat{\pi}}^2)^{1/2}\mathcal{Z}_\delta$ and $\hat{\pi} = \pi + \sigma_{\hat{\pi}}\mathcal{Z}_\pi$, where $\mathcal{Z}_\delta$ and $\mathcal{Z}_\pi$ are independent standard normal variables. When $\pi = 0$, therefore, $\hat{\beta}_{IV} = \frac{1}{\sigma_{\hat{\pi}}}(\sigma_{\hat{\delta}}^2 - \sigma_{\hat{\delta}\hat{\pi}}^2/\sigma_{\hat{\pi}})^{1/2}(\mathcal{Z}_\delta/\mathcal{Z}_\pi) + \beta_{WOLS}$, the median of which is β_{WOLS} since $\mathcal{Z}_\delta/\mathcal{Z}_\pi$ has a standard Cauchy distribution with zero median.

[7] See Anderson and Rubin (1949) for the general form of the AR statistic in over-identified models with a fixed number of instruments and Mikusheva and Sun (2021) for an adaptation to models with many weak instruments.

The AR statistic often offers an attractive alternative to the usual t-statistic, t_W (as in, e.g. Andrews et al., 2019; Keane & Neal, 2021). The AR test has size undistorted by weak instruments under the Staiger and Stock (1997) sequence and, under the just-ID model in eq. (3), is optimal among unbiased tests (Moreira, 2009). When testing $\beta = 0$, t_{AR} is simply the t-statistic for the associated reduced form. We focus here on a conventional t-statistic because the conventional test is ubiquitous in applied work, while AR is seen less. This may reflect the fact that confidence intervals derived by inverting t_{AR} have infinite length when $F < z^2_{1-\alpha/2}$ (i.e., less than about 4 for 95% confidence intervals), and AR-based intervals are always wider than conventional intervals when the former are finite (see, e.g., Lee et al., 2020).

Summarizing the behavior of a conventional 5% nominal test, Panel (a) in Figure 1 depicts rejection rates for t_W as a contour plot indexed by ρ and $E[F]$. The figure shows that rejection rates greatly exceed the nominal 5% rate only if the instrument is weak (i.e., $E[F]$ is close to 1) *and* endogeneity is high. In particular, if $|\rho| < 0.76$, rejection rates are below 10%, regardless of the strength of the first stage. If $|\rho| < 0.565$, the nominal 5% test under-rejects (this cutoff is also noted in Lee et al., 2020). A simple corollary, further substantiated below, is that the coverage of conventional nominal 95% confidence intervals for $\hat{\beta}_{IV}$ is likely to be satisfactory in most applications.

The modest over-rejection seen in Figure 1 is explained by a signal feature of just-ID IV: the bias of $\hat{\beta}_{IV}$ rises as the instrument grows weaker, but precision falls apace. The IV standard error reflects this lack of precision well enough that, unless endogeneity is egregious, inference is distorted little. This contrasts with overidentified 2SLS with many weak instruments (as in Bekker (1994) and Bound et al. (1995)), where, bias notwithstanding, the usual standard errors for 2SLS remain small enough for the t-statistic to be misleading.

Our conclusions here also contrast with those drawn in Stock and Yogo (2005) and Lee et al. (2020) regarding the reliability of inference based on a conventional just-ID IV t-statistic. Although Lee et al. (2020) report a similar plot, both studies emphasize worst-case rejection rates over ρ, for a given $E[F]$. As can be seen in our Figure 1, this worst-case rejection rate occurs at $|\rho| = 1$. In the same spirit, Keane and Neal (2021) highlights simulations showing that conventional just-ID IV t-tests can be misleading when endogeneity is very high. Sections 3.1 and 3.2 explain why we are not much concerned with high values of ρ.

3.1 The Anatomy of Endogeneity

We put endogeneity in context using three applications. These are the AK91 study that launched the modern weak instruments literature, the AE98 estimates using a dummy for samesex sibships (of first- and second-born children) as an instrument for family size, and the AL99 fuzzy regression discontinuity estimates of class size effects. The AE98 and AL99 first-stage t-statistics exceed those for AK91 and are arguably out of the zone where an instrument might be considered weak. With a first-stage t-statistic of almost 8, the AK91 quarter-of-birth instrument also seems strong enough. But all three studies can be used to calibrate endogeneity and to document contextual features that constrain it.

Table 1 reports key statistics for specifications drawn from each study (some estimates in the table differ slightly from those in the original). The first row in Panel A shows estimates of the economic returns to schooling in the AK91 sample of men born 1920–29. Here, OLS and IV estimates equal 0.08 and 0.072, respectively. These are close, so endogeneity is small in this case, with an estimated ρ of only 0.043. Schooling returns estimated in the second AK91 sample, consisting of men born 1930–39, exhibit more OVB. In this sample, the IV estimate of 0.105 surprisingly exceeds the OLS estimate of 0.071 (IV estimation of the returns to schooling is usually motivated by a concern that omitted ability controls causes OLS estimates to be too large). Endogeneity is correspondingly larger at $\rho = -0.175$, but still well outside the danger zone depicted in Figure 1.[8]

The AK91, AE98, and AL99 studies span a range of OVB scenarios, from modest in the first AK91 sample, to substantial in AE98 (where OLS magnitudes consistently exceed IV by at least 50%), to dramatic in AL99 (where IV exceeds small, insignificant OLS estimates, mostly by an order of magnitude, and sometimes with a sign flip). Yet, the magnitude of endogeneity exceeds 0.40 in only one specification, that for reading scores in the AL99 discontinuity sample (which consists of classes in schools with enrollment near the cutoff that determines class size). The just-ID IV estimates in all three of these studies are therefore unlikely to be compromised by finite-sample bias.[9]

Although the consistently moderate levels of endogeneity documented in Table 1 does not make a theorem, these applications have features in common with many IV-driven microeconometric investigations of causal effects. First, measured against endogenous variable standard deviations, first-stage magnitudes are small. In particular, the range of first-stage impact in eq. (2) runs from under 4% of the endogenous variable standard deviation in AK91, to about 12% of the endogenous variable standard deviation in AE98, with the AL99 first stage in-between.

Most importantly, endogeneity in research on causal effects is often capped by the modest size of the effects of interest. To make this point, it's helpful to write ρ as a function of OVB. Using eqs. (8) and (9), we can express ρ under homoskedasticity as:

$$\begin{aligned} \rho &= \frac{\sigma_v}{\sigma_\epsilon}(\beta_{WOLS} - \beta) \\ &= \frac{\sigma_v}{\sigma_\epsilon}\left(\frac{\beta_{OLS} - \beta}{1 - R_p^2}\right) \approx \frac{\sigma_D}{\sigma_Y}(\beta_{OLS} - \beta). \end{aligned} \tag{11}$$

We can use this expression to compute ρ by replacing β with $\hat{\beta}_{IV}$. The relevance of this representation of ρ can be seen in the AE98 estimates of the effects a third child on weeks worked by women aged 21–35 in the 1980 Census. Here, the first-stage partial R-square (R_p^2) is close to zero, while the difference between the conventional OLS estimate and the corresponding IV estimate is 3.42. The term multiplying this, $\frac{\sigma_v}{\sigma_\epsilon}$, is well-approximated by the ratio of the endogenous variable standard deviation to the dependent variable standard deviation (denoted $\frac{\sigma_D}{\sigma_Y}$), a ratio of about 0.022. The product of these two terms gives 0.075, equal to the value of ρ reported in the table for this sample.

Equation (11) suggests alternative bounds on endogeneity. For starters, in the AK91 scenario,

[8] Estimates of ρ in the table are computed as described in Appendix A.1. These estimates may differ from ρ because $\hat{\beta}_{IV}$ is biased, and because of sampling variance. In the examples analyzed here, however, the instruments are not particularly weak, so the bias in estimated endogeneity is negligible. With weaker instruments, endogeneity can be bounded by features of the problem at hand; this approach is taken below.

[9] Estimates of ρ computed under homoskedasticity in the AK91 and AE98 samples are almost identical to those reported in Table 1. For the AL99 samples, endogeneity parameters computed using homoskedasticity are smaller; these are reported as ρ^* in the table.

it seems reasonable to assume that the (causal) economic returns to schooling are no more than double the OLS estimate and certainly positive. Under these restrictions, the descriptive statistics in Table 1, which approximate $\frac{\sigma_v}{\sigma_\epsilon}$ at around 5.2, suggest $|\rho|$ can be no more than about 0.41. Although substantial, this is well below the 0.76 threshold for coverage concerns. With β bounded below by zero, large magnitudes of ρ require β to far exceed β_{WOLS}. Only when the causal effect of schooling is triple the OLS estimate (so that OLS is *too small* by 0.15) does the endogeneity danger zone become relevant.

Many microeconometric IV applications involve linear probability models in which causal effects are changes in probabilities. The AE98 estimates of the effect of the birth of a third child on female labor force participation in 1980, for example, range from roughly -0.18 for OLS to -0.12 for IV. Labor force participation rates for women with only two children run around 57%. Causal effects might therefore be as large as -0.57, but no larger, since probabilities can't be negative. In this case, $\frac{\sigma_v}{\sigma_\epsilon}$ is about 1 (again, using standard deviations in the data rather than residuals), so $\beta_{OLS} - \beta$ can be no larger than $-0.18 + .57 = 0.39$, thereby bounding ρ at this value. This generous bound makes no use of the fact that selection bias is likely to make OLS estimates of family-size effects on female supply too large (in magnitude) rather than too small. Other applications with Bernoulli outcomes admit similar sorts of bounds.

A related argument, appropriate for models with continuous outcomes, shows endogeneity to be constrained by plausible values for causal effects measured in standard deviation units. This line of reasoning is especially apt for education research where standard-deviation-denominated effect sizes are widely reported. The influential Tennessee STAR class size experiment analyzed in Krueger (1999) generated a reduction of 7 students per class, roughly one standard deviation of class size in the AL99 data. The STAR experiment yielded treatment effects of about 0.2σ, an impact typical of education interventions deemed to have been effective. At the same time, education researchers often view effect sizes as large as half a standard deviation in the outcome distribution as rare, if not implausible. Using the fact that $\frac{\sigma_v}{\sigma_\epsilon}$ is about equal to $(1 - R_p^2)$ in the AL99 data, the scenario of a half-standard deviation effect size generated by a one-standard deviation reduction in class size implies $\frac{\sigma_v}{\sigma_\epsilon}\frac{\beta}{1-R_p^2} = -0.5$ on the second line of eq. (11). At the same time, OLS estimates of class size effects in AL99 are mostly zero (as is often found in class size research; see e g., Hanushek (1986)), so the magnitude of endogeneity is capped at 0.5.

Contributing to all three of these arguments is the fact that endogeneity under homoskedasticity can be split into the difference between two R-squared-like terms:

$$\rho \approx \frac{\sigma_D}{\sigma_Y}(\beta_{OLS} - \beta) = \frac{\sigma_D}{\sigma_Y}\beta_{OLS} - \frac{\sigma_D}{\sigma_Y}\beta. \tag{12}$$

The square of the first term, $(\frac{\sigma_D}{\sigma_Y}\beta_{OLS})^2$, is the variation in the dependent variable accounted for by D_i in an analysis-of-variance for Y_i. In microeconometric applications, this term is mostly small, as is the causal analog that determines the square of the second term, $(\frac{\sigma_D}{\sigma_Y}\beta)^2$. This fact limits the magnitude of the difference between them.[10]

[10] Keane and Neal (2021) consider bounds on ρ in the context of estimates of the economic returns to schooling motivated by the view that OLS returns should exceed causal effects. Although this seems defensible, it's worth noting that the literature surveyed by Card (2001) reports many IV estimates in excess of the corresponding OLS estimates, a

3.2 When Measurement Error Motivates IV

In addition to IV for causal effects, a second major arena for microeconometric IV involves models with measurement error. Suppose the regression of interest is $Y_i = D_i^*\beta + X_i'\gamma + \eta_i$, where η_i is a residual uncorrelated with (D_i^*, X_i) by definition. The regressor D_i^* is unobserved; we see only a noisy measure, $D_i = D_i^* + e_i$, where the measurement error, e_i, is assumed to be classical, that is uncorrelated with (D_i^*, X_i, η_i). Replacing D_i^* with D_i yields the structural equation to be instrumented:

$$\begin{aligned} Y_i &= D_i\beta + X_i'\gamma + (\eta_i - e_i\beta) \\ &= D_i\beta + X_i'\gamma + \epsilon_i, \end{aligned}$$

where $\epsilon_i = \eta_i - e_i\beta$ is the structural residual. Given an instrument correlated with D_i^* and uncorrelated with ϵ_i, the coefficients of interest are consistently estimated by IV. The first stage in this scenario can be written as in (2), with first-stage residual, v_i.

To calibrate endogeneity in this model, note first that, given the classical measurement error assumption, $\text{cov}(v_i, \epsilon_i) = -\sigma_e^2\beta$. Under homoskedasticity, endogeneity squared can therefore be written:

$$\rho^2 = \frac{\sigma_e^4\beta^2}{\sigma_v^2\sigma_\epsilon^2} = \frac{\sigma_e^4\beta^2}{\sigma_v^2(\sigma_\eta^2 + \beta^2\sigma_e^2)} \leq \frac{\sigma_e^2}{\sigma_v^2} = \frac{1-r}{1-R_p^2}, \tag{13}$$

where $r = \sigma^2_{\tilde{D}^*}/\sigma^2_{\tilde{D}}$ denotes the reliability (or signal-to-noise ratio) of mismeasured D_i, after partialing out covariates.[11]

Although we can't speak to reliability across all fields, labor economists have collected evidence on the reliability of key variables of interest. These include schooling, earnings, hours worked, and hourly wages. Schooling appears often on the right-hand side of wage equations, while earnings, hours, and hourly wages are used in various configurations to estimate labor supply elasticities. The Angrist and Krueger (1999) summary of reliability estimates suggests $r \approx 0.9$ for schooling and $r \approx 0.8$ for earnings, falling to about $0.65 - 0.75$ for hours worked and hourly wages. The lower end of this range may be more relevant for wage reliability after partialing out covariates.[12] With $r = 0.65$ as a reasonably conservative value, we'd need to see an R_p^2 equal to at least 0.4 for ρ to reach 0.76. But $E[F] = \frac{nR_p^2}{1-R_p^2} + 1$, so, at this level of first-stage fit, $E[F]$ is way nowhere near the trouble zone for any sample size that's empirically relevant. This suggests that, unless reliability is unusually low, microeconometric measurement error can be expected to generate parameter combinations for which conventional IV inference is trouble-free.

4 Bias Under a Good Sign

Having made an empirical case for a sanguine view of just-ID IV, we add a novel analytical argument. This builds on the idea that IV identification strategies are most credible when a strong institutional

pattern first highlighted by Lang (1993).

[11]The first equality in (13) follows from the definition of correlation, the middle inequality uses the fact that σ_η^2 must be non-negative, and the last equality uses the definition of partial R^2 in eq. (8).

[12]See Table 11 in Angrist and Krueger (1999). The reliability of variables other than wages is less sensitive to covariate adjustment.

or theoretical foundation explains the first stage. These foundations usually imply a sign for π. In the AK91 application, for example, the quarter-of-birth first stage arises from the fact that children born later in the year enter school younger, and are therefore constrained by compulsory attendance laws to stay in school longer than those born earlier. The AE98 samesex instrument for family size is predicated on parents' preference for mixed-sex sibships. The AL99 Maimonides Rule instrument for class size is derived from regulations that determine class size as a function of enrollment. In these and many other applied micro applications, institutions or preferences determine the sign of π.

Andrews and Armstrong (2017) use this insight to motivate an IV estimator that is unbiased under a first-stage sign restriction. Without loss of generality, assume $\pi > 0$, and let $\mu(x) = \frac{1-\Phi(x)}{\phi(x)}$ denote the Mills' ratio of a standard normal random variable, where $\phi(x)$ and $\Phi(x)$ are the standard normal density and cdf evaluated at x. The Andrews-Armstrong unbiased IV estimator, denoted $\hat{\beta}_U$, exploits the fact that when first-stage estimates are normally distributed,

$$E\left[\frac{\mu(t_1)}{\sigma_{\hat{\pi}}}\right] = \frac{1}{\pi}.$$

In other words, $\hat{\tau} \equiv \mu(t_1)/\sigma_{\hat{\pi}}$ is an unbiased estimator of the reciprocal of the first-stage coefficient, π. Define

$$\hat{\beta}_U \equiv \hat{\tau}(\hat{\delta} - \beta_{WOLS}\hat{\pi}) + \beta_{WOLS} = t_1\mu(t_1)\hat{\beta}_{IV} + (1 - t_1\mu(t_1))\beta_{WOLS}. \tag{14}$$

Finally, recall that β_{WOLS} is the slope from a regression of the estimated reduced form on the estimated first stage. $\hat{\delta} - \beta_{WOLS}\hat{\pi}$ and $\hat{\tau}$ are therefore uncorrelated, and since $E[\hat{\delta} - \beta_{WOLS}\hat{\pi}] = (\beta - \beta_{WOLS})\pi$, it follows that $\hat{\beta}_U$ is unbiased for β. Moreover, $\hat{\beta}_U$ is a linear combination of conventional IV estimates and β_{WOLS} with coefficients, or weights, given by $t_1\mu(t_1)$ and $1 - t_1\mu(t_1)$.

To interpret $\hat{\beta}_U$, observe that $t_1\mu(t_1) = \hat{\pi}\hat{\tau}$, so that the weights in (14) reflect the extent to which $\hat{\tau}$ differs from $1/\hat{\pi}$. If $t_1 > 0$, these weights are bounded by a classic Mills' ratio inequality (e.g. Feller, 1968, p. 175) that implies:

$$0 \leq 1 - t_1\mu(t_1) \leq \frac{1}{t_1^2}. \tag{15}$$

Thus, when the first stage is right-signed, the weights $t_1\mu(t_1)$ in eq. (14) lie between 0 and 1, and we can interpret $\hat{\beta}_U$ as shrinking the conventional IV estimate towards OLS. The amount of shrinkage is bounded by the reciprocal of t_1^2, that is, by $1/F$. When $F = t_1^2 = 10$, the unbiased estimator shrinks $\hat{\beta}_{IV}$ at most 10% of the way towards β_{WOLS}.

The shrinkage interpretation of $\hat{\beta}_U$ seems surprising: since $\hat{\beta}_{IV}$ is biased towards OLS, shrinkage towards OLS increases bias. This counterintuitive fact is reconciled with the unbiasedness of $\hat{\beta}_U$ by the following theorem:

Theorem 1. *Consider the model in (3), and suppose that $\pi > 0$. Let $\lambda = \pi/\sigma_{\hat{\pi}} = \sqrt{E[F] - 1}$. Then the relative mean bias of $\hat{\beta}_U$ conditional on $t_1 > 0$ can be written:*

$$\frac{E[\hat{\beta}_U - \beta \mid t_1 > 0]}{\beta_{WOLS} - \beta} = \sqrt{\pi/2}\frac{\phi(\lambda)}{\Phi(\lambda)},$$

while, conditional on $t_1 < 0$, relative mean bias is:

$$\frac{E[\hat{\beta}_U - \beta \mid t_1 < 0]}{\beta_{WOLS} - \beta} = -\sqrt{\pi/2}\frac{\phi(\lambda)}{(1 - \Phi(\lambda))}.$$

The estimator $\hat{\beta}_U$ is therefore unbiased because it averages conditional positive bias when $t_1 > 0$ and conditional negative bias when $t_1 < 0$. As in Stock and Yogo (2005), the theorem scales mean bias by the weak-IV OVB of the OLS estimand, defined in eq. (9). This simplifies bias formulas, while the relationship between conditional and unconditional bias stands without this rescaling.[13]

It is hard to imagine an analyst who is prepared to sign the population first stage while ignoring the sign of the estimated first stage. Such conditioning, however, strips $\hat{\beta}_U$ of its appeal. In contrast with the deleterious effects of sign-screening on $\hat{\beta}_U$, the next section shows that first-stage sign-screening has surprisingly salutary effects on the sampling distribution of conventional IV estimates.

4.1 Sign-Screened Bias and Coverage

Suppose we assume π is positive and report second-stage estimates only when first-stage estimates are positive as well (equivalently, when $t_1 > 0$). In contrast with $\hat{\beta}_U$, sign-screening reduces the bias of $\hat{\beta}_{IV}$ markedly. Moreover, screening on the estimated first-stage sign has no downside in terms of coverage: in contrast with procedures that screen on the *magnitude* of the first-stage F statistic, screening on the sign of the corresponding t-statistic is shown here to have little effect on rejection rates for a conventional second-stage t-test.

Because the expectation of a just-ID IV estimator is undefined, our investigation of sign-screening looks at median bias (the expectation of 2SLS exists only for over-identified models). As in Theorem 1, we scale bias by the weak-IV OVB of the OLS estimand. We are interested in the difference between relative median bias computed unconditionally and conditional on $\hat{\pi} > 0$.

Surprisingly, Theorem 2 below shows that the worst-case relative bias obtains in the limit as $|\rho| \to 0$ (this is not the same as relative bias when $\rho = 0$; with no endogeneity, both IV and OLS are unbiased, so that relative bias is discontinuous in ρ). The relationship between ρ and relative median bias therefore contrasts with that in Section 3, which shows higher endogeneity leading to worse coverage. This reversal reflects the fact that, although the bias of $\hat{\beta}_{IV}$ increases with endogeneity, OVB increases faster. Because modest endogeneity is empirically relevant, median relative bias is reasonably characterized by the worst case over ρ. This leads us to the following theoretical result:

Theorem 2. *Consider the model in eq. (3), and suppose that $\pi > 0$. Let $\lambda = \pi/\sigma_{\hat{\pi}} = \sqrt{E[F] - 1}$. Then, for a given λ, unconditional relative median bias of $\hat{\beta}_{IV}$ is characterized by*

$$\sup_{\rho} \left| \frac{\text{median}_{E[F],\rho}(\hat{\beta}_{IV} - \beta)}{\beta_{WOLS} - \beta} \right| = \frac{\phi(\lambda)}{\lambda[\Phi(\lambda) - 1/2] + \phi(\lambda)}.$$

[13] Stock and Yogo (2005) focus on relative mean bias for 2SLS models with over-identifying restrictions.

Moreover, if $\lambda \geq 0.84$, the relative median bias of $\hat{\beta}_{IV}$ conditional on $\hat{\pi} > 0$ satisfies

$$\sup_{\rho} \left| \frac{\text{median}_{E[F],\rho}(\hat{\beta}_{IV} - \beta \mid \hat{\pi} > 0)}{\beta_{WOLS} - \beta} \right| = \frac{\phi(\lambda)}{\lambda\Phi(\lambda) + \phi(\lambda)}.$$

Equivalently, these expressions give the limit of relative unconditional and conditional median bias as $|\rho| \to 0$.

Figure 2 plots the two bias expressions in the theorem. We see that: (i) unconditional relative bias falls rapidly as first-stage strength increases: (ii) unless $E[F]$ is exceedingly low, conditional median bias is only about half as large as the corresponding unconditional bias. This is a consequence of the fact that the ratio of conditional to unconditional median bias bounds is:

$$1 - \frac{0.5\lambda}{\lambda\Phi(\lambda) + \phi(\lambda)}.$$

For λ greater than about 1.5, this is within 1 percentage point of 0.50, since the normal cdf is then close to one and the normal density close to zero. The shaded regions in Figure 2 delineate the range of variation in relative median bias as a function of ρ, showing a surprisingly flat relationship (see Appendix A.5 for computational details). Appendix A.5 also shows that the conditional median bias of $\hat{\beta}_{IV}$ is always less than that of $\hat{\beta}_U$, and at least 50% smaller once $\lambda \geq 1$.[14]

Screening on the sign of the first-stage estimate clearly mitigates bias. Since many analysts likely pursue a just-ID IV identification strategy only when the first-stage estimates are signed as expected, the inner line in Figure 2 may be a better guide to the bias in published IV estimates. But perhaps this sort of screening is costly. Econometricians have long warned that screening on the first-stage F-statistic risks a form of pretest bias and so may do more harm than good. Intuitively, when π is truly zero, large F-statistics signal realizations in which the in-sample correlation between instruments and structural errors is largest. Consequently, when instruments are weak, large Fs come with especially misleading $\hat{\beta}_{IV}$ realizations. Perhaps sign-screening runs a similar risk.

By way of evidence on this point, panel (b) of Figure 1 plots rejection contours for a conventional (second-stage) t-test conditional on $\hat{\pi} > 0$. That is, the figure plots contours for:

$$R_W^c = P_{E[F],\rho}(|t_W| > z_{1-\alpha/2} \mid \hat{\pi} > 0).$$

Comparison of the two panels in Figure 1 shows that screening on the first-stage sign affects rejection rates little. For instance, the endogeneity cutoff required to keep rejections rates below 10% is $|\rho| \leq 0.75$, rather than the unconditional 0.76. This result is explained by the fact that, when the instrument is very weak, screening has two effects. On one hand, the bias of $\hat{\beta}_{IV}$ is reduced. At the same time, screening out wrong-signed first-stage estimates leads to an overestimate of first-stage strength, on average. These two effects are just about offsetting, so that the rejection contours in

[14] Andrews and Armstrong (2017) show numerically that the unconditional median bias of $\hat{\beta}_U$ is smaller than that of $\hat{\beta}_{IV}$ when $E[F]$ is small, while their bias ranking reverses for larger $E[F]$. They also note that the median absolute deviation of $\hat{\beta}_U$ is always smaller than that of $\hat{\beta}_{IV}$. Our numerical calculations indicate that this fails to hold for all parameter values conditional on the estimated first stage sign. An analyst evaluating estimators on the basis of conditional *mean* bias will nevertheless prefer $\hat{\beta}_U$ to $\hat{\beta}_{IV}$ since only the bias of the former is finite.

panel (a) of Figure 1 are much like those in panel (b).

What practical lesson should we draw from this? The careful analyst judges statistical power by standard errors rather than t-statistics, and hesitates to declare a finding conclusive based on $\hat{\beta}_{IV}$ alone when the reduced form is statistically uninformative. Just-ID IV is a strong hand, but the reduced form, a regression coefficient with anodyne statistical properties, is the IV analyst's ace in the hole. The good properties of the AR test statistic can be seen as justifying this view since the t-statistic for the reduced form and t_{AR} coincide for a zero null. At the same time, an empiricist hewing to conventional IV reporting strategy—showing reduced-form, first-stage, and IV estimates and the associated standard errors—can be reassured that, provided the first stage has the anticipated sign, the bias of just-ID IV is likely to be minimal and conventional confidence interval coverage adequate. Only when endogeneity exceeds a level that applied microeconomists rarely encounter is inference likely to be misleading.

In the context of the AK91, AE98, and AL99 studies, first-stage sign screening adds no action items to the empirical agenda. The first-stage estimates in these applications are robustly right-signed. The reduced forms, reported in detail, are consistent with second-stage estimates in clearly showing that the instrument moves outcomes in a manner implied by plausible first-stage values. In applications with weaker instruments than these, an empirical strategy that begins by examining the first-stage sign would seem to have no downside. Claims of credible causal evidence requires more than this, however. In AK91, for instance, the quarter-of-birth story holds water because schooling can be seen to move sharply up and down with quarter of birth as predicted by compulsory attendance laws, across 30 birth cohorts in three data sets, and because graduate degree completion that should be changed little by compulsory attendance, indeed moves little with quarter of birth. This coherence is part of what gives the AK91 first stage its strength.

5 Summary and Conclusions

Assuming reduced-form and first-stage estimates are normally distributed, null rejection rates for conventional t-tests in just-ID IV models are distorted little unless endogeneity is extraordinarily large. A corollary is that conventional IV standard errors are likely to yield confidence intervals with good coverage. Three widely-cited applications, two of which demonstrate considerable OVB in OLS estimates, are characterized by modest endogeneity and consequently fall well inside the low-distortion comfort zone. We've argued that these three examples should be seen as representative rather than idiosyncratic: the structure of much applied micro research naturally bounds endogeneity.

We've also introduced a new theoretical reason to be unconcerned with the bias of just-ID IV. This builds on the Andrews and Armstrong (2017) argument that in credible applications of just-ID IV, the analyst is rarely agnostic about the direction of the first stage. Unlike Andrews and Armstrong (2017), however, we impose the same sign restriction on the estimated as well as the theoretical first stage. Such conditioning roughly halves the median bias of the IV estimator. Moreover, in contrast with screening on the first-stage F, sign-screening generates no pretest bias. Since most analysts likely impose an estimated first-stage sign screen as a matter of course, the bias reduction sign-conditioning engenders is already reflected in published empirical work.

References

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63. https://doi.org/10.1214/aoms/1177730090

Andrews, I., & Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, *8*(2), 479–503. https://doi.org/10.3982/QE700

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, *80*(3), 313–336. https://www.jstor.org/stable/10.2307/2006669

Angrist, J. D., & Evans, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, *88*(3), 450–477. https://www.jstor.org/stable/116844

Angrist, J. D., & Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *90*(430), 431–442. https://doi.org/10.1080/01621459.1995.10476535

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (pp. 1277–1366). Elsevier. https://doi.org/10.1016/S1573-4463(99)03004-7

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575. https://doi.org/10.1162/003355399556061

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681. https://doi.org/10.2307/2951662

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450. https://doi.org/10.1080/01621459.1995.10476536

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*(5), 1127–1160. https://doi.org/10.1111/1468-0262.00237

Feller, W. (1968). *An introduction to probability theory and its application* (3rd ed., Vol. 1). Wiley.

Hall, A. R., Rudebusch, G. D., & Wilcox, D. W. (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review*, *37*(2), 283. https://doi.org/10.2307/2527324

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141–1177. https://www.jstor.org/stable/2725865

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475. https://doi.org/10.2307/2951620

Keane, M., & Neal, T. (2021). *A practical guide to weak instruments* (Working Paper). SSRN. https://doi.org/10.2139/ssrn.3846841

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics*, *114*(2), 497–532. https://doi.org/10.1162/003355399556052

Lang, K. (1993). *Ability bias, discount rate bias, and the return to education* [Unpublished manuscript, Boston University].

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. (2020). *Valid t-ratio Inference for IV*. arXiv: 2010.05058.

Mikusheva, A., & Sun, L. (2021). Inference with many weak instruments. *Review of Economic Studies*.

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics*, *152*(2), 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics - Simulation and Computation*, *9*(4), 389–419. https://doi.org/10.1080/03610918008812164

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586. https://doi.org/10.2307/2171753

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80–108). Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006

Young, A. (2021). *Leverage, heteroskedasticity and instrumental variables in practical application* [Unpublished working paper, LSE]. https://personal.lse.ac.uk/YoungA/LeverageandIV.pdf

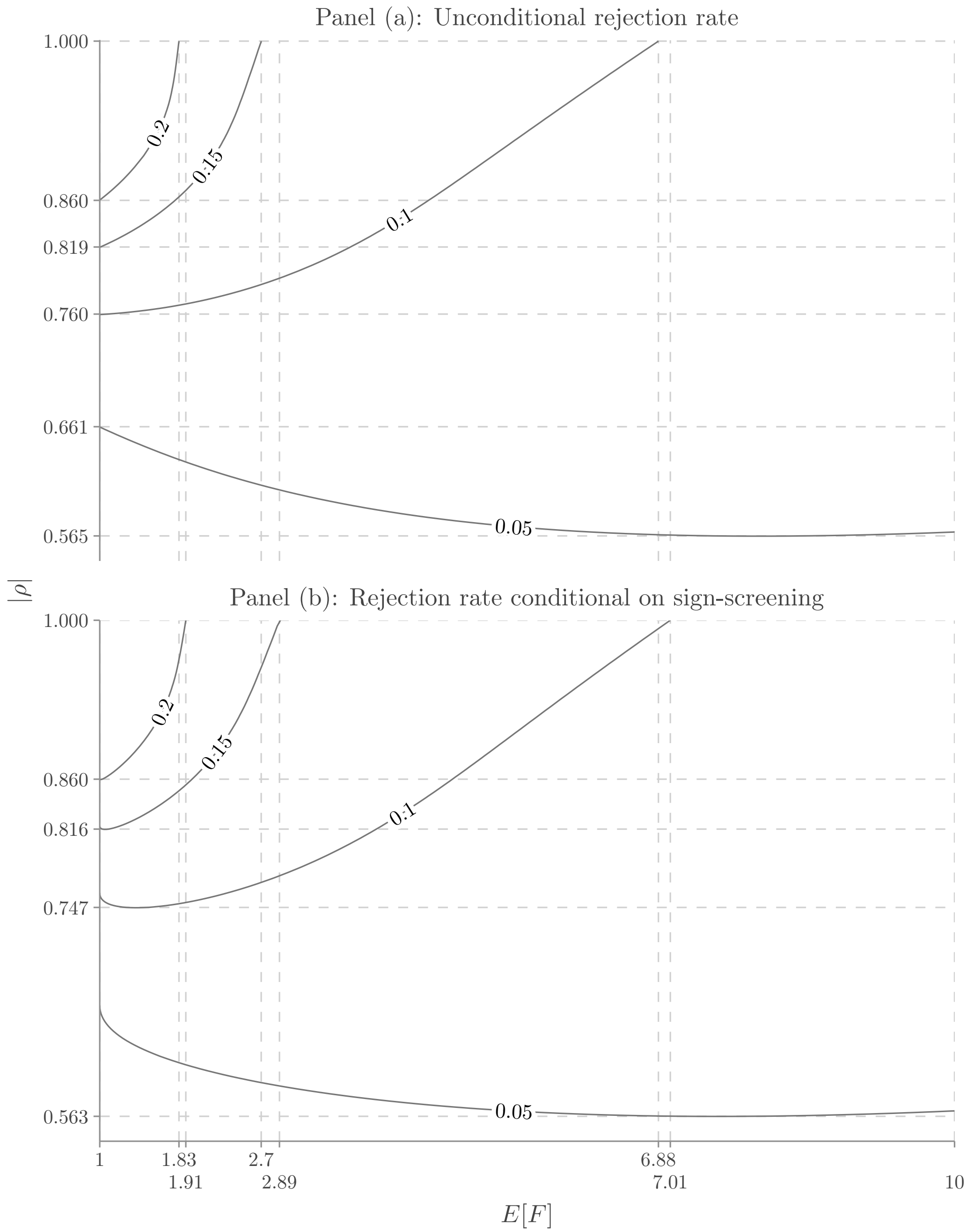

Figure 1: Contour plot of the rejection rate of conventional t-test with nominal level $\alpha = 0.05$ as function of $E[F]$ and ρ. Panel (a) plots the unconditional rejection rate R_W. Panel (b) plots the rejection rate R_W^c conditional on $\hat{\pi} > 0$. See Appendix A.2 for computational details.

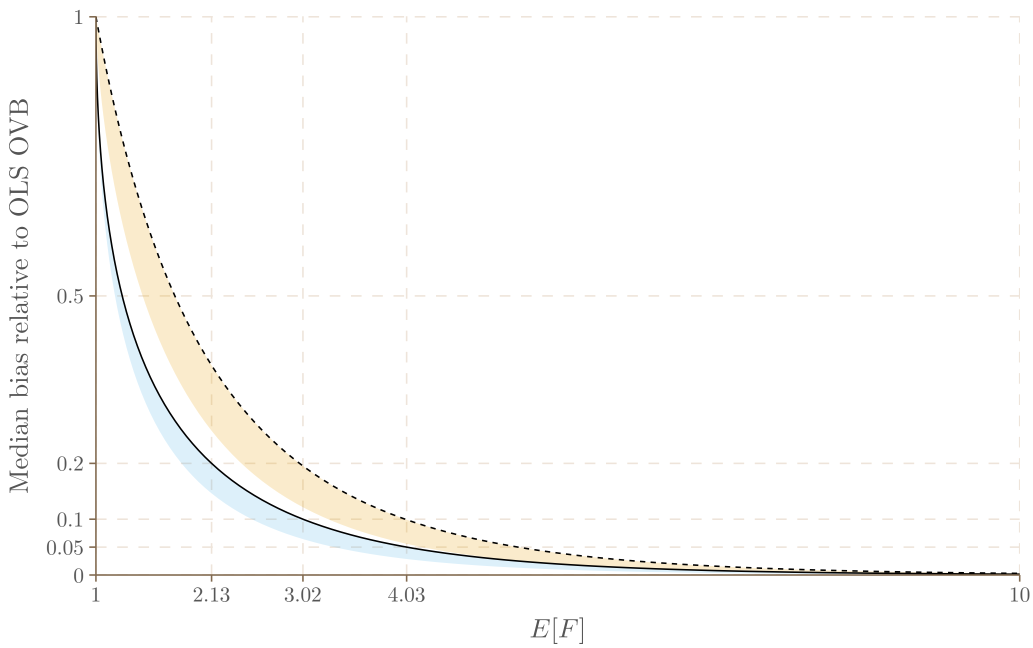

Figure 2: Bound on median bias of $\hat{\beta}_{IV}$ in units of OLS bias, i.e. as a fraction of $|\beta_{WOLS} - \beta|$. Dashed line: unconditional median bias. Solid line: median bias conditional on $\hat{\pi} > 0$. The yellow and blue shaded areas delineate the range of variation in unconditional and conditional median bias, respectively, over possible values of ρ.

Table 1: Estimates and Endogeneity in Three IV Applications

Sample	Dep. Var.	Endog. Var.	Instrument	First Stage	Estimates	
	(sd)	(sd)	(sd)		OLS	2SLS
			A. AK91			
Men born	Log weekly	Years of	QOB 1 birth	-0.122	0.0802	0.0724
1920-29	wage (0.65)	schooling	(.43)	(0.016)	(0.0004)	(0.0226)
(N= 247,199)		(3.36)			ρ= 0.043	
Men born	Log weekly	Years of	QOB 1 birth	-0.106	0.0711	0.1049
1930-39	wage (0.68)	schooling	(0.43)	(0.013)	-0.0004	(0.0246)
(N=329,509)		(3.28)			ρ = -0.175	
			B. AE98			
1980 Census	Worked for	More than	Samesex	0.061	-0.177	-0.117
mothers aged	pay (.50)	2 kids (0.49)	sibship	(0.0015)	(0.002)	(0.025)
18-35			(0.50)		ρ = -0.058	
(N=394,840)	Weeks				-8.997	-5.559
	worked				(0.071)	(1.118)
	(22.3)				ρ = -0.075	
1990 Census	Worked for	More than	Samesex	0.062	-0.164	-0.092
mothers aged	pay (.47)	2 kids (0.48)	sibship	(0.0015)	(0.002)	(0.025)
18-35			(0.50)		ρ = -0.070	
(N=380,007)	Weeks				-8.65	-5.46
	worked				(0.077)	(1.156)
	(22.8)				ρ = -0.067	
			C. AL99			
5th Grade	5th Grade	Class size	Maimonides	0.477	0.009	-0.263
classes	Reading	(6.55)	Rule (6.11)	(0.041)	(0.034)	(0.094)
(N=2,019)	(7.68)				ρ=0.352 (ρ* = 0.225)	
	5th Grade				0.036	-0.264
	Math (9.60)				(0.045)	(0.123)
					ρ =0.310 (ρ* = 0.183)	
5th Grade	5th Grade	Class size	Maimonides	0.481	-0.070	-0.410
discontinuity	Reading	(7.42)	Rule (7.50)	(0.057)	(0.050)	(0.118)
samples	(8.18)				ρ =0.460 (ρ* = 0.390)	
(N=471)	5th Grade				0.090	-0.185
	Math				(0.070)	(0.155)
	(10.20)				ρ =0.292 (ρ* = 0.224)	

Notes: This table reports IV and OLS estimates replicating the AK91, AE98, and AL99 studies discussed in the text. For each study, the table reports estimates from multiple samples, as well as the corresponding first-stage estimate and its standard error. The endogeneity parameter appears below columns showing OLS and IV. Endogeneity estimates for AL99 marked with an asterisk were computed assuming homoskedasticity. The 2nd, 3rd, and 4th columns report standard deviations in parentheses. Other columns show standard errors in parentheses. These are robust for AK91 and AE98, and clustered on school for AL99.

Appendix A Derivations and Proofs

The appendix uses the notation $\tilde{\beta}_{IV} = (\hat{\beta}_{IV} - \beta)/|\beta_{WOLS} - \beta|$ and $\tilde{\beta}_U = (\hat{\beta}_{IV} - \beta)/|\beta_{WOLS} - \beta|$ to denote the IV and the unbiased estimator, after centering and scaling by the weak-IV OVB of OLS. Also, we let $\lambda = \pi/\sigma_{\hat{\pi}} = \sqrt{E[F] - 1}$, and $s = \rho/\sqrt{1-\rho^2}$.

A.1 Estimating ρ

We estimate ρ (defined in eq. (7)) using first-stage and IV estimates, and the associated first-stage, reduced-form and IV standard errors. To see how this works, rewrite eq. (5) as

$$\sigma_{\hat{\delta}\hat{\pi}} = \frac{\sigma^2_{\hat{\pi}}\hat{\beta}^2_{IV} - \hat{\pi}^2\hat{\sigma}^2_{IV} + \sigma^2_{\hat{\delta}}}{2\hat{\beta}_{IV}}. \tag{A.1}$$

With this in hand for $\sigma_{\hat{\delta}\hat{\pi}}$, endogeneity can be computed as the sample analog of eq. (7), replacing β with $\hat{\beta}_{IV}$. The resulting estimator is:

$$\hat{\rho} = \frac{\sigma_{\hat{\pi}}}{|\hat{\pi}|\hat{\sigma}_{IV}} \times (\sigma_{\hat{\delta}\hat{\pi}}/\sigma^2_{\hat{\pi}} - \hat{\beta}_{IV}).$$

A.2 t-Test Rejection Rates

This section writes the rejection probabilities of the t-test as an integral indexed by $(E[F], \rho)$. Stock and Yogo (2005) use Monte Carlo methods to compute unconditional rejection probabilities in a similar setup. The calculation described here is faster and more accurate. More importantly, it allows us to easily compute both the unconditional rejection rates, and rejection rates conditional on sign-screening.

Using eq. (7), and the fact that $\beta_{WOLS} - \beta$ and ρ have the same sign, we may write t_{AR} as

$$t_{AR} = \frac{(\hat{\delta} - \hat{\pi}\beta)|\rho|}{\sigma_{\hat{\pi}}|\beta_{WOLS} - \beta|}. \tag{A.2}$$

Consequently,

$$\tilde{\beta}_{IV} = \frac{\hat{\delta} - \beta\hat{\pi}}{\sigma_{\hat{\pi}} t_1 |\beta_{WOLS} - \beta|} = \frac{t_{AR}}{|\rho| t_1}. \tag{A.3}$$

Thus,

$$t_W = \frac{\operatorname{sign}(t_1) t_{AR}}{\sqrt{\frac{\sigma^2_{\hat{\delta}}/\sigma^2_{\hat{\pi}} - 2\beta_{WOLS}\beta + \beta^2}{(\beta_{WOLS} - \beta)^2}\rho^2 + \frac{t^2_{AR}}{t^2_1} - 2\rho\frac{t_{AR}}{t_1}}} = \frac{\operatorname{sign}(t_1) t_{AR}}{\sqrt{1 + t^2_{AR}/t^2_1 - 2\rho t_{AR}/t_1}} \tag{A.4}$$

where the first equality uses eq. (A.3) and the definition of β_{WOLS}, and the second equality uses eq. (7). This expression for t_W implies that conditional on t_1, the rejection region $\{|t_W| \geq z_{1-\alpha/2}\}$ is quadratic in t_{AR}. Solving this quadratic inequality implies that the rejection region is given by

$$t_{AR} \in \begin{cases} \emptyset & \text{if } t_1^2 \leq (1-\rho^2) z_{1-\alpha/2}^2, \\ [a_1, a_2] & \text{if } (1-\rho^2) z_{1-\alpha/2}^2 \leq t_1^2 \leq z_{1-\alpha/2}^2. \\ (-\infty, a_2) \cup (a_1, \infty) & \text{if } t_1^2 \geq z_{1-\alpha/2}^2 \end{cases}$$

where

$$a_1 = \frac{\rho z_{1-\alpha/2}^2 t_1 - |t_1| z_{1-\alpha/2} \sqrt{t_1^2 - (1-\rho^2) z_{1-\alpha/2}^2}}{z_{1-\alpha/2}^2 - t_1^2},$$

$$a_2 = \frac{\rho z_{1-\alpha/2}^2 t_1 + |t_1| z_{1-\alpha/2} \sqrt{t_1^2 - (1-\rho^2) z_{1-\alpha/2}^2}}{z_{1-\alpha/2}^2 - t_1^2}.$$

Note that $\operatorname{cor}(t_{AR}, t_1) = \rho$, so that

$$P(t_{AR} \leq x \mid t_1) = \Phi((x - \rho(t_1 - \lambda))/\sqrt{1-\rho^2}). \tag{A.5}$$

Thus, conditional on t_1, the rejection probability is given by

$$\begin{aligned} P(|t_W| \geq z_{1-\alpha} \mid t_1) &= (P(t_{AR} \leq a_2 \mid t_1) - P(t_{AR} \leq a_1 \mid t_1)) \, \mathbb{I}\{t_1^2 \geq z_{1-\alpha/2}^2 (1-\rho^2)\} \\ &\quad + \mathbb{I}\{t_1^2 \geq z_{1-\alpha/2}^2\} \\ &= f(t_1; \lambda, \rho) \, \mathbb{I}\{t_1^2 \geq (1-\rho^2) z_{1-\alpha/2}^2\} + \mathbb{I}\{t_1^2 \geq z_{1-\alpha/2}^2\}, \end{aligned} \tag{A.6}$$

where

$$f(t_1; \lambda, \rho) = \Phi\left(\frac{a_2 - \rho(t_1 - \lambda)}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{a_1 - \rho(t_1 - \lambda)}{\sqrt{1-\rho^2}}\right).$$

Since $t_1 \sim \mathcal{N}(\lambda, 1)$, the rejection probability conditional on $t_1 \geq c$ is therefore given by

$$P(|t_W| \geq z_{1-\alpha} \mid t_1 \geq c) = \frac{\int_c^\infty (\mathbb{I}\{t_1^2 \geq (1-\rho^2) z_{1-\alpha}\} f(t_1; \lambda, \rho) + \mathbb{I}\{t_1^2 \geq z_{1-\alpha/2}^2\}) \phi(t_1 - \lambda) dt_1}{\Phi(\lambda - c)}.$$

The unconditional rejection probability R_W obtains by setting $c = -\infty$. The rejection probability conditional on sign screening, R_W^c, obtains by setting $c = 0$. The coverage contours in Figure 1 evaluate the above expression as a function of (ρ, λ) by numerical integration.

A.3 Proof of Theorem 1

We may write

$$\tilde{\beta}_U = t_1 \mu(t_1) \tilde{\beta}_{IV} + (1 - t_1 \mu(t_1)) \operatorname{sign}(\rho) = \mu(t_1) \frac{t_{AR}}{|\rho|} + (1 - t_1 \mu(t_1)) \operatorname{sign}(\rho) \tag{A.7}$$

where the first equality follows from eq. (14), and the fact that $\beta_{WOLS} - \beta$ and ρ have the same sign, and the second equality applies eq. (A.3).

Since $E[t_{AR} \mid t_1] = \rho(t_1 - \lambda)$, the relative bias conditional on t_1 is given by

$$E[\tilde{\beta}_U \mid t_1] = \text{sign}(\rho)\left[1 - \lambda\mu(t_1)\right].$$

By arguments analogous to those in the proof of Lemma 2.1 in Andrews and Armstrong (2017), we have

$$\begin{aligned}
E[\lambda\mu(t_1) \mid t_1 > 0] &= \frac{1}{\Phi(\lambda)}\int_{t=0}^{\infty}\lambda\frac{1-\Phi(t)}{\phi(t)}\phi(t-\lambda)dt = \frac{e^{-\lambda^2/2}}{\Phi(\lambda)}\int_{t=0}^{\infty}(1-\Phi(t))\cdot\lambda e^{\lambda t}dt \\
&= \frac{e^{-\lambda^2/2}}{\Phi(\lambda)}\left\{\left[e^{\lambda t}(1-\Phi(\lambda))\right]_{t=0}^{\infty} + \int_{t=0}^{\infty}\phi(t)e^{\lambda t}dt\right\} \\
&= \frac{1}{\Phi(\lambda)}\left[-\frac{1}{2}e^{-\lambda^2/2} + \int_{t=0}^{\infty}\phi(t-\lambda)dt\right] = -\frac{1}{2}\frac{e^{-\lambda^2/2}}{\Phi(\lambda)} + 1,
\end{aligned}$$

where the first line uses the definition of a Mills' ratio, second line uses integration by parts, and the third follows by "completing the square". It therefore follows that

$$\frac{E[\hat{\beta}_U - \beta \mid t_1 > 0]}{\beta_{WOLS} - \beta} = \frac{1}{2}\frac{e^{-\lambda^2/2}}{\Phi(\lambda)} = \sqrt{\pi/2}\frac{\phi(\lambda)}{\Phi(\lambda)}.$$

The second claim follows by an analogous argument.

A.4 Proof of Theorem 2

The distribution of $\tilde{\beta}_{IV}$ conditional on t_1 can then be written as

$$\begin{aligned}
P(\tilde{\beta}_{IV} \leq x \mid t_1; s) &= \begin{cases} P(t_{AR} \leq t_1 x|\rho| \mid t_1) = \Phi(s[\lambda - (1-\text{sign}(s)x)t_1]) & \text{if } t_1 \geq 0, \\ P(t_{AR} \geq t_1 x|s| \mid t_1) = \Phi(-s[\lambda - (1-\text{sign}(s)x)t_1]) & \text{if } t_1 < 0. \end{cases} \\
&= \Phi\left(s[\text{sign}(t_1)\lambda - (1-\text{sign}(s)x)|t_1|]\right).
\end{aligned} \tag{A.8}$$

where the first equality uses eq. (A.3), and the second equality follows from eq. (A.5). Observe that since $P(\tilde{\beta}_{IV} \leq x \mid t_1; -s) = 1 - P(\tilde{\beta}_{IV} \leq -x \mid t_1; -s)$, the distribution is symmetric in s. It therefore suffices to consider $s > 0$.

We first prove the claims concerning the distribution of $\tilde{\beta}_{IV}$ conditional on $t_1 > 0$. By eq. (A.8), this distribution is given by

$$P(\tilde{\beta}_{IV} \leq x \mid t_1 > 0; s) = \frac{1}{\Phi(\lambda)}\int_{-\lambda}^{\infty}\Phi\left(s[\lambda - (1-x)t_1]\right)\phi(z)dz. \tag{A.9}$$

Observe that the conditional median, denoted $m_c = m_c(s)$, of $\tilde{\beta}_{IV}$ is smaller than 1, since

$$P(\tilde{\beta}_{IV} \leq 1 \mid t_1 > 0; s) = \frac{\Phi(s\lambda)}{\Phi(\lambda)}\int_{-\lambda}^{\infty}\phi(z)dz = \Phi(s\lambda) > 1/2. \tag{A.10}$$

Next, by the mean value theorem, for some $\tilde{s} = \tilde{s}(x, s) \in [0, s]$,

$$P(\tilde{\beta}_{IV} \leq x \mid t_1 > 0; s) = \Phi(0) + \frac{s}{\Phi(\lambda)} \int_{-\lambda}^{\infty} ((x-1)z + x\lambda)\phi\left(s((x-1)z + x\lambda)\right)\phi(z)dz$$
$$= \frac{1}{2} + \frac{s}{\tilde{s}^2(1-x)\Phi(\lambda)} \int_{-\infty}^{\tilde{s}\lambda} y\phi(y)\,\phi(a+by)dy,$$

where the second line uses the change of variables $y = \tilde{s}x\lambda - \tilde{s}(1-x)z$, and we let $a = x\lambda/(1-x)$, $b = -\frac{1}{\tilde{s}(1-x)}$. By line 111 of Table 1 in Owen (1980),

$$\int x\phi(x)\phi(a+bx) = \frac{\phi(a/t)}{t^2}\left[-\phi(tx + ab/t) - \frac{ab}{t}\Phi(tx + ab/t)\right], \quad t = \sqrt{1+b^2}. \tag{A.11}$$

Applying this result to the preceding display then yields

$$P(\tilde{\beta}_{IV} \leq x \mid t_1 > 0; s) = \frac{1}{2} + \frac{s}{\tilde{s}^2(1-x)\Phi(\lambda)}\frac{\phi(a/t)}{1+b^2}\left[-\phi\left(t\tilde{s}\lambda + ab/t\right) - \frac{ab}{\sqrt{1+b^2}}\Phi(t\tilde{s}\lambda + ab/t)\right]$$
$$= \frac{1}{2} + \frac{s}{\Phi(\lambda)}\frac{\phi(a/t)(1-x)}{\tilde{s}^2(1-x)^2+1}\left[\frac{x}{|1-x|}\frac{\lambda}{\tilde{g}(x,\tilde{s})}\Phi(\lambda g(x,\tilde{s})) - \phi\left(\lambda g(x,\tilde{s})\right)\right],$$

where $g(x, \tilde{s}) = \frac{\tilde{s}^2|1-x| + \operatorname{sign}(1-x)}{\sqrt{\tilde{s}^2(1-x)^2+1}}$, and $\tilde{g}(x, \tilde{s}) = \sqrt{\tilde{s}^2(1-x)^2+1}$. When evaluated at $x = m_c$, the expression in square brackets must equal zero by definition of the median. Therefore, $m_c > 0$, and since we also know from eq. (A.10) that $m_c < 1$, the conditional median must satisfy

$$m_c = \frac{1}{\frac{\lambda}{\tilde{g}(m_c,\tilde{s}(m_c,s))}\frac{\Phi(\lambda g(m_c,\tilde{s}(m_c,s)))}{\phi(\lambda g(m_c,\tilde{s}(m_c,s)))} + 1}, \tag{A.12}$$

We have

$$\frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda g)}{\phi(\lambda g)} \geq \frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda\tilde{g})}{\phi(\lambda\tilde{g})}, \quad \text{and} \quad \frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda\tilde{g})}{\phi(\lambda\tilde{g})} \geq \lambda\frac{\Phi(\lambda)}{\phi(\lambda)} \quad \text{if } \lambda \geq 0.84.$$

Here the first inequality follows because $\Phi(x)/\phi(x)$ is increasing in x, and $g \geq \tilde{g}$, and the second inequality follows because $\frac{\Phi(x)}{x\phi(x)}$ is increasing for $x \geq 0.84$, and $\tilde{g} \geq 1$. Therefore,

$$m_c \leq \frac{\phi(\lambda)}{\lambda\Phi(\lambda) + \phi(\lambda)} = \lim_{s \downarrow 0} m_c(s),$$

where the equality follows since the right-hand side of eq. (A.12) converges to $\frac{\phi(\lambda)}{\lambda\Phi(\lambda)+\phi(\lambda)}$ as $s \to 0$.

We now prove the claims concerning the unconditional distribution of $\tilde{\beta}_{IV}$. From eq. (A.8), we obtain

$$P(\tilde{\beta}_{IV} \leq x; s) = 1 - \Phi(\lambda)$$
$$+ \int_{-\lambda}^{\infty} \Phi\left(s(x\lambda - (1-x)z)\right)\phi(z)dz - \int_{-\infty}^{-\lambda} \Phi\left(s(x\lambda - (1-x)z)\right)\phi(z)dz. \tag{A.13}$$

Observe that the median of $\tilde{\beta}_{IV}$, denoted $m_u = m_u(s)$ is smaller than 1, since

$$
\begin{aligned}
P(\tilde{\beta}_{IV} \leq 1; s) &= 1 - \Phi(\lambda) + \Phi(s\lambda)\left[\int_{-\lambda}^{\infty} \phi(z)dz - \int_{-\infty}^{-\lambda} \phi(z)dz\right] \\
&= 1 - \Phi(\lambda) + \Phi(s\lambda)\left[\Phi(\lambda) - 1 + \Phi(\lambda)\right] > 1 - \Phi(\lambda) + \left[\Phi(\lambda) - 1 + \Phi(\lambda)\right] = \Phi(\lambda) > 1/2. \quad \text{(A.14)}
\end{aligned}
$$

By arguments as in the conditional case, for $x < 1$,

$$
\begin{aligned}
P(\tilde{\beta}_{IV} \leq x; \rho) &= \frac{1}{2} + \frac{s}{\tilde{s}^2(1-x)}\left[\int_{-\infty}^{\tilde{s}\lambda} y\phi(y)\,\phi(a+by)dy - \int_{\tilde{s}\lambda}^{\infty} y\phi(y)\,\phi(a+by)dy\right] \\
&= \frac{1}{2} + \frac{s}{\tilde{s}^2(1-x)}\frac{\phi(a/t)}{1+b^2}\left[-2\phi(t\tilde{s}\lambda + ab/t) - 2\frac{ab}{t}\Phi(t\tilde{s}\lambda + ab/t) + \frac{ab}{t}\right] \\
&= \frac{1}{2} + \frac{s}{\tilde{s}^2(1-x)}\frac{\phi(a/t)}{1+b^2}\left[-2\phi(\lambda g(\tilde{s},x)) + 2\frac{x}{1-x}\frac{\lambda}{\tilde{g}}\Phi(\lambda g(\tilde{s},x)) - \frac{x}{1-x}\frac{\lambda}{\tilde{g}}\right].
\end{aligned}
$$

Here the first line follows by the mean value theorem, where $\tilde{s} = \tilde{s}(x,s) \in [0,s]$, the second line uses eq. (A.11), and the last line follows by algebra. When evaluated at $x = m_c$, the expression in square brackets must equal zero by definition of the median. Therefore, $m_u > 0$, and it must satisfy

$$
m_u = \frac{1}{\frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda g)-1/2}{\phi(\lambda g)} + 1} \tag{A.15}
$$

Now,

$$
\frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda g) - 1/2}{\phi(\lambda g)} \geq \frac{\lambda}{\tilde{g}}\frac{\Phi(\lambda\tilde{g}) - 1/2}{\phi(\lambda\tilde{g})} \geq \lambda\frac{\Phi(\lambda) - 1/2}{\phi(\lambda)}.
$$

Here the first inequality follows because $\Phi(x)/\phi(x)$ is increasing in x, and $g \geq \tilde{g}$, and the second inequality follows because $\frac{\Phi(x)-1/2}{x\phi(x)}$ is increasing for $x > 0$. As a result,

$$
m_u \leq \frac{\phi(\lambda)}{\lambda(\Phi(\lambda) - 1/2 + \phi(\lambda))} = \lim_{s\downarrow 0} m_u(s),
$$

where the equality follows since the right-hand side of eq. (A.15) converges to $\frac{\phi(\lambda)}{\lambda(\Phi(\lambda)-1/2+\phi(\lambda))}$ as $s \to 0$.

A.5 Median Bias Comparisons

To evaluate the relative median bias of $\hat{\beta}_{IV}$ as a function of both $E[F]$ and ρ conditional on $t_1 \geq c$, we first evaluate

$$
P(\tilde{\beta}_{IV} \leq x \mid t_1 \geq c; \rho, \lambda) = \frac{1}{\Phi(\lambda - c)}\int_{c-\lambda}^{\infty} f_{IV}(z; x, \lambda, s)\phi(z)dz \tag{A.16}
$$

by numerical integration. Here we use the formula $f_{IV}(z; x, \rho, \lambda) = \Phi\left(s[\operatorname{sign}(z+\lambda)\lambda - (1-x)|z+\lambda|]\right)$ from eq. (A.8) for the cdf conditional on $z = t_1 - \lambda$. We then numerically solve for the median. For unconditional median, we set $c = -\infty$, and for the median conditional on sign screening, we set $c = 0$.

The shaded regions in Figure 2 correspond to the range of the absolute value of the relative median bias as ρ varies between -1 and 1.

To compare the relative median bias to that of $\hat{\beta}_U$, it suffices to consider $\rho > 0$, since the distributions of $\hat{\beta}_U$ and $\hat{\beta}_{IV}$ are symmetric in ρ. By eq. (A.7), it follows that for $t_1 > 0$,

$$P(\tilde{\beta}_U \leq x \mid t_1; s) = P\left(\tilde{\beta}_{IV} \leq x - (1-x)\frac{(1 - t_1\mu(t_1))}{t_1\mu(t_1)} \mid t_1; s\right),$$

which for $x < 1$ is smaller than $P(\tilde{\beta}_{IV} \leq x \mid t_1; s)$. Since the median of $\tilde{\beta}_{IV}$ conditional on $t_1 > 0$ is smaller than 1, it follows that the conditional median bias of $\tilde{\beta}_{IV}$ is always smaller than that of $\tilde{\beta}_U$.

To compare the relative magnitudes of the median biases, we compute the relative median bias of $\tilde{\beta}_U$ analogously to that of $\tilde{\beta}_{IV}$, except we replace f_{IV} in eq. (A.16) with $f_U(z; x, \lambda, \rho) = \Phi(s[\lambda - (1 - \operatorname{sign}(s)x)/\mu(\lambda + z)])$ (it follows from eqs. (A.5) and (A.7) that this is the cdf $\tilde{\beta}_U$ conditional on $z = t_1 - \lambda$). We then compute the ratio $\operatorname{median}_{\lambda,\rho}(\tilde{\beta}_U \mid t_1 > 0)/\operatorname{median}_{\lambda,\rho}(\tilde{\beta}_{IV} \mid t_1 > 0)$ of the median biases on a fine grid of values of (ρ, λ). This ratio is greater than 2 if $E[F] \geq 2$, and greater than 3 if $E[F] \geq 3$, regardless of the value of ρ.