Treatment Effect Heterogeneity and Weak Instruments

Michal Kolesár

ECO539B, Spring 2025

March 30, 2025

Textbook model

Treatment effect heterogeneity

Weak instruments

Uses of IV

Can use instrumental variables (IV) regression to solve a number of issues:

1. Errors-in-variables (e.g., Zellner 1970);
2. Omitted variable bias: want to recover β in the projection $E[Y_i \mid D_i, A_i] = D_i\beta + A_i'\gamma$, but A_i is not observed (e.g., Chamberlain 2007);
3. Estimate a simultaneous equations model, such as a demand-and-supply system (e.g., Angrist, Graddy, and Imbens 2000); or
4. Estimate treatment effects when unconfoundedness assumption fails.

Focus on last goal, and consider:

1. Implications of treatment effect heterogeneity for estimation and inference; and
2. Weak instrument issues

Setup

- Focus on i.i.d. sampling of (Y_i, D_i, Z_i, W_i), with $\dim(Z_i) = k$. Let $X_i = (Z_i', W_i')'$.
- Reduced form and first stage projections

$$\begin{aligned} Y_i &= Z_i'\delta + W_i'\psi_Y + u_{Y,i}, && (1) \\ D_i &= Z_i'\pi + W_i'\psi_D + u_{D,i}. && (2) \end{aligned}$$

Assumptions i

Normality assumption

$$\sqrt{n}\begin{pmatrix}\hat{\delta}-\delta\\ \hat{\pi}-\pi\end{pmatrix} \Rightarrow \mathcal{N}\left(0, \mathrm{var}(u_i \otimes E[\tilde{Z}_i\tilde{Z}_i']^{-1}\tilde{Z}_i)\right), \qquad u_i = \begin{pmatrix}u_{Y,i}\\ u_{D,i}\end{pmatrix} \tag{3}$$

with asymptotic variance consistently estimable

- Fails if
  1. k is large relative to n (next lecture); or
  2. Leverages are high (Young 2022, e.g.). Analogous to ordinary least squares (OLS) diagnostics—these are just OLS regressions!

Assumptions ii

Valid IV with constant treatment

$\delta = \beta\pi$, with $\beta = E[Y(1) - Y(0)]$.

Lemma

Valid IV assumption is implied by following:

Constant treatment effects $Y(d) = Y(0) + d\beta$.

Random assignment Z mean-independent of the potential outcomes $Y(d, z)$ given W

Exclusion restriction $Y_i(d, z)$ in fact only depends on d

Linearity $E[Z \mid W]$ is linear in W (or else $E[Y(0) \mid W]$ linear)

Estimation: TSLS

- Add and subtract $D_i\beta$ from reduced form, and plug in first stage to obtain structural equation

$$\begin{aligned} Y_i &= D_i\beta + Z_i'(\delta - \pi\beta) + W_i'\gamma + \epsilon_i, \quad \epsilon_i = u_{Y,i} - u_{D,i}\beta, \quad \gamma = (\psi_Y - \psi_D\beta) \\ &= D_i\beta + W_i'\gamma + \epsilon_i \end{aligned}$$

  where second line used Valid IV assumption.

- Since $X_i = (D_i, W_i)$ orthogonal to u_i, obtain moment condition

$$E[X_i\epsilon_i] = 0$$

- When $\sigma^2(X_i) = \operatorname{var}(\epsilon_i \mid X_i)$ homoskedastic, optimal generalized method of moments (GMM) weighting matrix $\propto E[X_i X_i]^{-1}$, and solving it yields two-stage least squares (TSLS):

$$\hat{\beta}_{\text{TSLS}} = \frac{D' H_{\ddot{Z}} Y}{D' H_{\ddot{Z}} D} = \frac{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\delta}}{\hat{\pi} \ddot{Z}' \ddot{Z} \hat{\pi}}, \qquad \hat{\gamma}_{\text{TSLS}} = (W'W)^{-1} W'(Y - D\hat{\beta}_{\text{TSLS}}),$$

with $\ddot{Z}_i$ denoting sample residual from projecting off W_i.

- If $k = 1$, weighting doesn't matter, $\hat{\beta}_{\text{TSLS}} = \hat{\delta}/\hat{\pi}$.
- Standard GMM (or delta method) arguments deliver

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_T), \qquad \mathcal{V}_T = \frac{E[\sigma^2(X_i)(\tilde{Z}_i' \pi)^2]}{(\pi' E[\tilde{Z}_i \tilde{Z}_i'] \pi)^2}. \tag{4}$$

Estimation: LIML

- Anderson and Rubin (1949): assume $(\epsilon_i, u_{D,i})$ are homoskedastic and jointly normal conditional on X_i, and estimate β by maximum likelihood. This gives:

$$\hat{\beta}_{\text{LIML}} = \operatorname*{argmin}_{\beta} \frac{(\hat{\delta} - \beta\hat{\pi})'\ddot{Z}\ddot{Z}'(\hat{\delta} - \beta\hat{\pi})}{(1, -\beta)S(1, -\beta)'},$$

  where $S = [(\ddot{Y}, \ddot{D}) - \ddot{Z}(\hat{\delta}, \hat{\pi})]'[(\ddot{Y}, \ddot{D}) - \ddot{Z}(\hat{\delta}, \hat{\pi})]/(n - k - \ell)$ estimates $\text{var}(u_i)$

- Equivalent to minimum distance estimator minimizing

$$\begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}' W \begin{pmatrix} \hat{\delta} - \pi\beta \\ \hat{\pi} - \pi \end{pmatrix}.$$

  with $W = S^{-1} \otimes \ddot{Z}'\ddot{Z}/n$ optimal weighting matrix under homoskedasticity (Goldberger and Olkin 1971)

- Minimum distance objective doesn't rely on normality or homoskedasticity $\implies$ LIML asymptotically normal and consistent. In fact, can show LIML first-order asymptotically equivalent to TSLS

$$\sqrt{n}(\hat{\beta}_{\text{LIML}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_T),$$

- How to pick between these approaches? Will answer based on robustness to dropping some restrictive assumptions

What can go wrong

1. Normality assumption fails
   - Can happen due to high leverages, that also make Eicker-Huber-White (EHW) standard errors unreliable. Research question: extend small-sample corrections we talked about in previous lecture to IV
   - Can also happen because k high relative to instrument strength: next lecture.
2. Valid IV assumption fails
   - Heterogeneous treatment effects: will focus on next
   - Random assignment or exclusion restriction fails
   - Linearity fails: do diagnostic wrt functional form with which controls W_i enter, similar to OLS discussion in previous lecture.
3. Delta method underlying asymptotic normality of TSLS and LIML fails
   - This happens if $\pi = 0$. By continuity, this implies that the delta method will work poorly if π is close to zero. This is a weak instrument problem: will focus on

Diagnostics

- Check first-stage and reduced form leverages, similar to OLS diagnostics
- What are diagnostics for random assignment/exclusion restriction?

Textbook model

Treatment effect heterogeneity

Weak instruments

Main takeaways

1. TSLS estimates weighed average of local average treatment effects (LATEs), but LIML does not in general estimates an object that has a causal interpretation.
2. Implication for instrument choice (Heckman and Vytlacil 2005):

   > *The relevant question regarding the choice of instrumental variables in the general class of models studied in this paper is "What parameter is being identified by the instrument?" rather than the traditional question of "What is the efficient combination of instruments for a fixed parameter?"—the question that has traditionally occupied the attention of econometricians who study instrumental variables.*

3. The standard error for $\hat{\beta}_{\text{TSLS}}$ based on $\mathcal{V}_T$ no longer valid.

Estimands

- Imbens and Angrist (1994) show that without covariates, when Z_i binary, and $D_i(1) \geq D_i(0)$, $\hat{\beta}_{\text{TSLS}}$ estimates
$$E[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)]$$
- What if there are covariates and instrument and treatment not binary?
- Write first stage $D(z) = f(z, W, V)$, where V unobserved heterogeneity ("type")
  - Marginal treatment effect framework imposes $D(z) = \mathbb{1}\{g(z, W) \geq V\}$, but we don't require additive separability or binary D.
- Let $E[Y(d) \mid W, V] = \mu(d, W, V)$ denote marginal treatment response function (Heckman and Vytlacil 2005)

- Let $\hat{D} = \hat{D}(W, Z)$ denote constructed instrument, after covariates are partialled out
    - With linear first stage as in (2), $\hat{D} = \tilde{Z}'\pi$.
- Then

$$\beta_{\text{TSLS}} = \frac{E[\hat{D}Y]}{\text{var}(\hat{D})}$$

- Can write β_{TSLS} in terms of marginal treatment effect $\mu(1, W, V) - \mu(0, W, V)$ (binary treatment) or, more generally, derivatives of marginal treatment response function $\mu'(d, W, V) = \partial\mu(d, W, V)/\partial d$ if following hold:

Assumptions

Exclusion restriction $E[Y(d) \mid W, V, Z] = E[Y(d) \mid W, V]$

Partialling out $[\hat{D} \mid W, V] = 0$ (i.e. linearity if $\hat{D} = \tilde{Z}'\pi'$)

Theorem (Kolesár and Plagborg-Møller 2025, Appendix A.2)

Under mild regularity conditions, $\beta_{\text{TSLS}} = E[\int \omega(\epsilon, W, V)\mu'(\epsilon, W, V)d\epsilon]$, where $\omega(\epsilon, W, V)/\text{var}(\hat{D}) = E[\hat{D}1\{D \geq \epsilon\} \mid W, V]$. With binary D,

$$\beta_{\text{TSLS}} = E[\omega(W, V)\tau(W, V)], \quad \tau(W, V) = E[Y(1) - Y(0) \mid W, V],\ \omega(W, V) = \frac{E[\hat{D}D \mid W, V]}{\text{var}(\hat{D})}.$$

- Regularity conditions mild—don't need differentiability of $Y(d)$ (as in Angrist, Graddy, and Imbens 2000), only local absolute continuity of $\mu(d, W, V)$ and mild tail conditions.
- D may be continuous, discrete or mixed. With discrete D, how are derivatives μ' defined?
- Allows for misspecified first stage
- No first-stage monotonicity assumptions imposed

Interpretation

$$\beta_{\text{TSLS}} = E[\omega(W,V)\tau(W,V)], \quad \tau(W,V) = E[Y(1)-Y(0) \mid W,V],\ \omega(W,V) = \frac{\operatorname{cov}(\hat{D}, D \mid W,V)}{\operatorname{var}(\hat{D})}.$$

- Weights $\omega(W,V)$ non-zero only if $D \mid W,V$ not degenerate $\implies$ always and never-takers (i.e. $D(z)$ doesn't vary with z) receive zero weight
- Compliers receive positive weight, Defiers receive negative weight: compliers and defiers specific to instrument $\hat{D}$
- Sufficient for $\omega \geq 0$: uniform monotonicity (Imbens and Angrist 1994), i.e. no defiers, i.e. $D(z) = \{g(Z,W) \geq V\}$, $\hat{D} = g(Z,W) - E[g(Z,W) \mid W]$.

Monotonicity

- Recently a slew of papers arguing Imbens and Angrist (1994) monotonicity too strong theoretically and rejected empirically. E.g. Mogstad, Torgovitsky, and Walters (2021), Frandsen, Lefgren, and Leslie (2023), and Goff (2024).
- Papers come up with weaker monotonicity requirements
- Clear from above representation uniform monotonicity not necessary:
  - E.g. if τ depends only on W (selection only on $Y(0)$, no selection on treatment gains), then sufficient that $\operatorname{cov}(\hat{D}, D \mid W) \geq 0$.
- Results misleading only if weights $\omega(W, V)$ negative and correlated with MTE $\tau(W, V)$

LIML

- LIML generally no longer has a causal interpretation in the sense that it estimates a convex combination of LATEs (Kolesár 2013)
- Under treatment effect heterogeneity, proportionality restriction $\delta = \beta\pi$ no longer holds. But LIML imposes it with a non-diagonal weight matrix $\implies$ quite sensitive to its failure
- In contrast, TSLS constructs a single instrument $\hat{D}$—much more robust.
- Upshot: don't use LIML unless confident $\delta = \beta\pi$.

Inference

- With linear first stage, $\hat{\beta}_{\text{TSLS}} = \hat{\pi}\ddot{Z}'\ddot{Z}\hat{\delta}/\hat{\pi}\ddot{Z}'\ddot{Z}\hat{\delta}$
- Smooth function of asymptotically normal estimates $\hat{\delta}$ and $\hat{\pi}$, so delta method still works:

$$\sqrt{n}(\hat{\beta}_{\text{TSLS}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_2), \qquad \mathcal{V}_2 = \frac{E[((\tilde{Z}_i'\pi_\Delta)u_{2i} + \epsilon_i(\tilde{Z}_i'\pi))^2]}{(\pi' E[\tilde{Z}_i\tilde{Z}_i']\pi)^2}, \tag{5}$$

  Variance estimator needs to include extra term reflecting heterogeneity of LATEs:

$$\hat{V}_2 = n\frac{\sum_{i=1}^n[((\tilde{Z}_i'(\hat{\delta} - \hat{\pi}\hat{\beta}))\hat{u}_{2i} + \hat{\epsilon}_{\text{TSLS},i}(\tilde{Z}_i'\hat{\delta}))^2]}{(\sum_i \tilde{Z}_i\hat{\pi})^2},$$

  Derived in appendix to Imbens and Angrist (1994), see also Lee (2018)
- Heterogeneity correction rarely used in practice—but it should be!

Textbook model

Treatment effect heterogeneity

Weak instruments

- Interested in estimating elasticity of intertemporal substitution (EIS). Can log-linearize Euler equation based on a portfolio choice problem of an agent with Epstein-Zin preferences (e.g. Campbell 2003; Yogo 2004):

$$E_t[\Delta c_{t+1} - \mu_j - \psi r_{j,t+1}] = 0$$

  where $r_{j,t+1}$ is return on asset j, Δc_{t+1} is consumption growth, μ_j is a constant, and ψ is the EIS, and E_t denotes expectation conditional on the agent's information set at time t.

- Could estimate ψ by running the regression

$$\Delta c_{t+1} = \mu_j + \psi r_{j,t+1} + e_t.$$

  but $e_t = \Delta c_{t+1} - E_t[\Delta c_{t+1}] - \psi(r_{j,t+1} - E_t[r_{j,t+1}])$ is going to be correlated with $r_{j,t+1}$

- On the other hand, e_t will be by definition uncorrelated with any variables in the information set at time t, so we can use those as instruments.
- Alternatively, could instrument for Δc_{t+1} using same instruments, and estimate $1/\psi$. Called reverse TSLS.
- With a single instrument both standard and reverse TSLS numerically equivalent. When $k > 1$ and strong instruments, asymptotically equivalent.
- But empirical estimates of both ψ and of $1/\psi$ are small. For instance, Campbell (2003, Table 9) reports a 95% confidence interval (CI) $[-0.14, 0.28]$ for ψ, and CI $[-0.73, 2.14]$ for $1/\psi$, using quarterly U.S. data (1947–1998) on non-durable consumption and T-bill returns.
- Problem: instruments are weak because both consumption growth and asset returns are notoriously difficult to predict.

Quarter of birth

- Bound, Jaeger, and Baker (1995) give convincing evidence of weak instrument issues based on the Angrist and Krueger (1991) study:

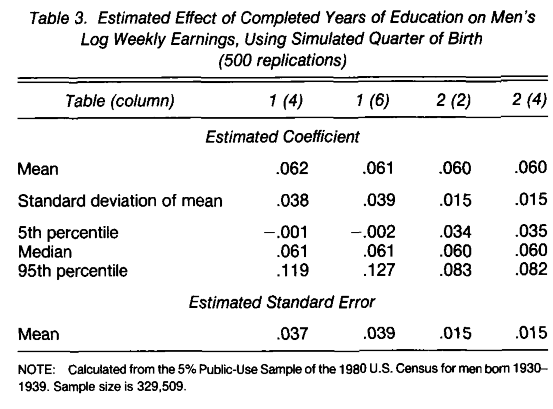


Table 3. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Using Simulated Quarter of Birth (500 replications)

Table (column)	*1 (4)*	*1 (6)*	*2 (2)*	*2 (4)*
Estimated Coefficient				
Mean	.062	.061	.060	.060
Standard deviation of mean	.038	.039	.015	.015
5th percentile	−.001	−.002	.034	.035
Median	.061	.061	.060	.060
95th percentile	.119	.127	.083	.082
Estimated Standard Error				
Mean	.037	.039	.015	.015

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509.

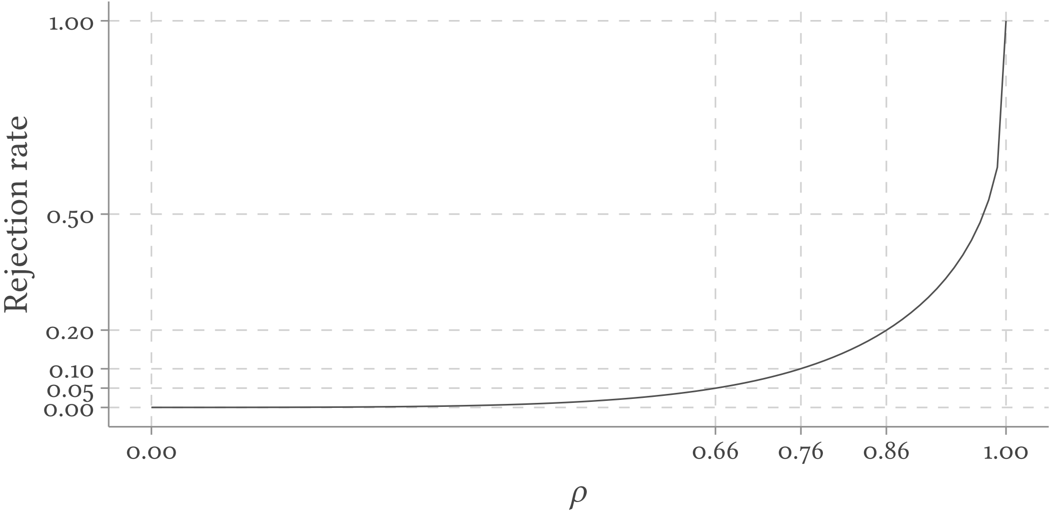
What goes wrong?

- Consider first single irrelevant instrument. Then

$$\hat{\beta}_{\text{TSLS}} = \frac{\tilde{Z}'Y}{\tilde{Z}'D} = \frac{\tilde{Z}'D\beta + \tilde{Z}'\epsilon}{\tilde{Z}'u_D} = \beta + \frac{n^{-1/2}\tilde{Z}'\epsilon}{n^{-1/2}\tilde{Z}'u_D} \Rightarrow \beta + \frac{\Sigma_{11}^{1/2}\mathcal{Z}_\epsilon}{\Sigma_{22}^{1/2}\mathcal{Z}_2} \sim \beta + \frac{\Sigma_{12}}{\Sigma_{22}} + \sqrt{\frac{(1-\rho^2)\Sigma_{11}}{\Sigma_{22}}}C$$

  where $\mathcal{Z}_\epsilon$ and $\mathcal{Z}_2$ are standard normal with covariance $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, and $\Sigma = \operatorname{var}(\tilde{Z}_i(\epsilon_i, u_{D,i})')$. C is Cauchy
- Centered at OLS plim $\beta + \Sigma_{12}/\Sigma_{22} = \beta + \rho\sqrt{\Sigma_{11}/\Sigma_{22}}$, with fat tails.
- Key parameter is endogeneity ρ. Determines asymptotic TSLS bias, as well as rejection of usual t-test (next slide)

Asymptotic rejection rate of usual t-test with nominal level 0.05



- To analyze problem, assume normality in (3) exact

$$\begin{pmatrix} \hat{\delta} - \delta \\ \hat{\pi} - \pi \end{pmatrix} \sim \mathcal{N}\left(0, \mathcal{V}\right), \tag{6}$$

  with $\mathcal{V}$ known (consistently estimable), and $\pi\beta = \delta$. But don't assume Delta method applies (which requires $\pi \neq 0$).

- Can be formally justified in various ways (see lecture notes), e.g. weak-instrument asymptotics

Detection of weak instruments

- To test for weak instruments, we first need to define what we mean by "weak". Two parameters, β and π.
- Stock and Yogo (2005) assume homoskedastic errors. Two definitions, each of which gives a set of values $\Pi_W \subseteq \mathbb{R}^k$ for π for which the instruments are weak:
    1. Bias of TSLS relative to OLS is bigger than 0.1 for some β.
    2. Wald test based on TSLS has size over 10% for some β.

- Bias actually admits closed form if $k \geq 2$ (Richardson 1968; Sawa 1972)

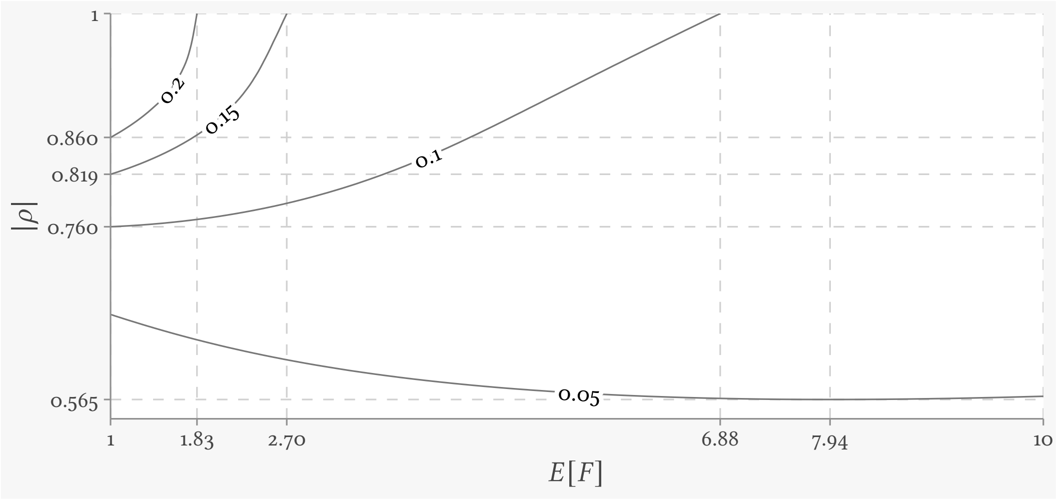
$$b_{\text{TSLS}} := \frac{E[\hat{\beta}_{\text{TSLS}} - \beta]}{\beta_{WOLS} - \beta} = 1 - \frac{k(E[F]-1)}{2} \int_0^1 x^{k/2-1} e^{(x-1)k(E[F]-1)/2} dt,$$

  where $\beta_{WOLS} = \Omega_{12}/\Omega_{22}$ is the limit of OLS under weak-instrument asymptotics

- Relative TSLS bias (relative to OLS) depends only on $E[F]$, and not β or endogeneity. Suggests (at least under homoskedasticity) $E[F]$ right quantity to measure instrument strength
- Expression only well-defined if $k \geq 2$, admits closed form for k even. E.g., with $k = 2$, $b_{\text{TSLS}} = e^{1-E[F]}$.
- By evaluating the relative bias and using the fact that $k \cdot F$ has a non-central χ^2_k distribution, we can derive critical values for testing the hypothesis $H_0 \colon b_{\text{TSLS}} \leq 0.1$.

- For example, for $k = 2$ we need $E[F] = 1 - \log(0.1) \approx 3.30$ for TSLS bias to be 10% of OLS bias, which leads to the critical value 7.85. For $k = 3, 4, 5$, we obtain critical values 9.18, 10.23, 10.78, and so on.
- Remarkable fact: critical values fluctuate between 10 and 11.5 for larger values of k. Noted in Table 5.1 in Stock and Yogo (2005). This leads to Staiger and Stock (1997) rule of thumb: $E[F] > 10$.
- In contrast, rejection rate of t-stat depends on ρ, not just $E[F]$, and increases with k. Important! Why?

Contour plot of rejection rates of t-test when $k = 1$



- Set of parameters where overrejection occurs is rather restricted.
- If willing to put a priori bounds on the value of β that would lead to $|\rho| \leq 0.76$ (for a given covariance matrix of the reduced form coefficients $\mathcal{V}$), then we never need to worry about the weak instrument problem in the sense that the Wald test will not overreject by more than 5%.
- Angrist and Kolesár (2024) argue that in many applications in labor economics, such large values of endogeneity are unlikely.
    - There is 1-1 map between β and ρ, so in given application, can often rule out extreme values of β

Homoskedasticity

- If $k > 1$ and the errors are not homoskedastic, $F > 10$ rule of thumb should not be used
- Problem: homoskedastic F doesn't correctly capture variability of first stage, while robust F, $F_{\mathrm{r}} = \frac{1}{k}\hat{\pi}'\mathcal{V}_{22}^{-1}\hat{\pi}$ not informative about TSLS denominator $\hat{\pi}'E[\tilde{Z}_i\tilde{Z}_i']\hat{\pi}$ (how close is delta method from failing)
- Montiel Olea and Pflueger (2013) propose effective F, which scales homoskedastic F by correct measure of first-stage variability. No simple rule of thumb, critical values depend on k and first-stage variance.

Takeaways so far

- $F > 10$ needs $k \geq 2$ and homoskedasticity. Doesn't tell us about inference, only tests relative TSLS bias
- But arguably, we only care about whether usual inference reliable—if usual CI has correct coverage, why care about bias? After all in say regression discontinuity (RD) or non-linear panel data, don't care mach if estimate biased, so long as inference reflects it
- Inference-based rule of thumb that allows for any ρ would be very conservative when $k > 1$ ($F \geq 16$ with $k = 1$ and increasing in k)
- Screening on the first-stage F-statistic appears to compound, rather than reduce, inferential problems arising from weak instruments (e.g. Andrews, Stock, and Sun 2019)
- World would be better if we abandoned first-stage F screening.
- What to do instead?

Weak instrument robust inference with single instrument i

- When $k = 1$, then task simplifies to ratio of means problem,

$$\begin{pmatrix} \hat{\delta} \\ \hat{\pi} \end{pmatrix} \sim \mathcal{N}_2 \left(\pi \begin{pmatrix} \beta \\ 1 \end{pmatrix}, \mathcal{V} \right).$$

- Doing inference about β equivalent to the ration of Gaussian means problem. This connection has been pointed out by Zellner (1978), and Mariano and McDonald (1979).
- Ratio of normal means is an old problem in statistics, dating back to at least Fieller (1932). Fieller (1940, 1954) observe that $\hat{\delta} - \hat{\pi}\beta_0$ is pivotal:

$$\frac{(\hat{\delta} - \beta_0 \hat{\pi})^2}{b_0' \mathcal{V} b_0} \sim_{H_0} \chi_1^2, \qquad b_0 = \begin{pmatrix} 1 \\ -\beta_0 \end{pmatrix}.$$

Weak instrument robust inference with single instrument ii

- Invert to get confidence set with exact coverage.
  1. If $F \geq z^2_{1-\alpha/2}$ (that is, we reject the null that $\pi = 0$), then the CI takes the form of an interval $[C_1, C_2]$, with endpoints given by

     $$C_j = \hat{\beta} + z^2_{1-\alpha/2} \frac{\hat{\beta} - \mathcal{V}_{12}/\mathcal{V}_{22}}{F - z^2_{1-\alpha/2}} \pm z_{1-\alpha/2} \frac{\sqrt{(\bar{F} - z^2_{1-\alpha/2})|\mathcal{V}|}}{(F - z^2_{1-\alpha/2})\mathcal{V}_{22}}$$

     Here $\bar{F} = F\mathcal{V}_{22}\hat{b}'\mathcal{V}\hat{b}/|\mathcal{V}|$ denotes the F-statistic for testing $(\delta, \pi) = 0$, with $\hat{b} = (1, -\hat{\beta})'$
  2. If $\bar{F} \geq z^2_{1-\alpha/2} \geq F$, then it takes the form $(-\infty, C_2] \cup [C_1, \infty)$, with C_j given in the previous display
  3. If $\bar{F} \leq z^2_{1-\alpha/2}$ (we don't reject the null that $(\delta, \pi) = 0$, with critical value based on χ^2_1 rather than χ^2_2), then it is given by $\mathbb{R}$.

- Fieller test known as Anderson and Rubin (1949) test in economics.
- Anderson-Rubin (AR) CI always longer than usual Wald CI, but asymptotically equivalent under standard asymptotics. Reflects price of robustness to large ρ.
- Argument for AR CI: works under weak instruments, and as efficient as Wald CI under strong instruments. One problem: not bet-proof, as discussed in Müller and Norets (2016)
- As a test, the AR test uniformly most powerful (UMP) among all unbiased tests. This was shown in Moreira (2009). If $\beta = \beta_0$ this is just test based on reduced form. Hard to argue against.
- But used little in practice. Motivated by this, Lee et al. (2022) propose an alternative procedure called tF that is based on the observation that the worst-case rejection of the Wald test occurs at $\rho = 1$. When $\rho = 1$, the t-statistic depends on the data only through the first-stage F, and they use this observation to derive critical values $c_\alpha(F)$ that depend on it. Critical values tend to infinity as $F \to z^2_{1-\alpha/2}$.

Sign screening

- No estimator can be fully immune to bias since it is impossible to construct a consistent or at least median unbiased estimator when the instruments are irrelevant.
- Can construct an unbiased estimator $\hat{\beta}_U$ if the sign of first stage known, as (Andrews and Armstrong 2017).
- Curious thing: if $t_1 \geq 0$—that is, the first-stage is right-signed—then the estimator shrinks TSLS towards OLS.
- Arises because the estimator $\hat{\beta}_U$ is unbiased by virtue of averaging a conditional positive bias when $t_1 > 0$ and with conditional negative bias when $t_1 < 0$.
- 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.

TSLS and sign-screening

- Angrist and Kolesár (2024) show that sign-screening actually halves the median bias of TSLS
- Doesn't impact coverage, in contrast to procedures that screen on the *magnitude* of the first-stage F statistic, screening on the sign of the corresponding t-statistic has to have little effect on rejection rates for a conventional Wald test.

What are your takeaways?

Overidentified case i

- Can generalize Fieller's idea:

$$\hat{\delta} - \beta\hat{\pi} \sim \mathcal{N}(0, E[\tilde{Z}_i\tilde{Z}_i']^{-1}E[b'\Omega(X_i)b\tilde{Z}_i\tilde{Z}_i']E[\tilde{Z}_i\tilde{Z}_i']^{-1}/n),$$

  Leads to Anderson and Rubin (1949) test.

- Bet-proofness problem becomes even more severe, since the resulting CI will be empty with positive probability. Furthermore, the CI is no longer efficient under standard asymptotics: it is longer than the Wald CI.
- CI may be empty because it tests joint null the TSLS estimand being equal to β_0, and δ being proportional to π.

Overidentified case ii

- conditional likelihood ratio (CLR) test suggested by Moreira (2003) is more powerful than AR, and enjoys some optimality properties under homoskedastic errors (Andrews, Moreira, and Stock 2006) (that do not, however, carry over to the heteroskedastic case)
- In just identified case, need endogeneity $|\rho|$ to be unreasonably high in cross-section applications (very close to 1) to generate severe overrejection of the Wald test: hard to come up with empirical examples where Wald CIs are substantively misleading. This changes when $k > 1$, as we've seen from the intertemporal elasticitity of substitution example.
- There is no procedure when $k > 1$ that allows for treatment effect heterogeneity (but see Luther's JMP)

References i

Anderson, Theodore W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *The Annals of Mathematical Statistics* 20, no. 1 (March): 46–63. https://doi.org/10.1214/aoms/1177730090.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74, no. 3 (May): 715–752. https://doi.org/10.1111/j.1468-0262.2006.00680.x.

Andrews, Isaiah, and Timothy B. Armstrong. 2017. "Unbiased Instrumental Variables Estimation under Known First-Stage Sign." *Quantitative Economics* 8, no. 2 (July): 479–503. https://doi.org/10.3982/QE700.

Andrews, Isaiah, James H. Stock, and Liyang Sun. 2019. "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* 11, no. 1 (August): 727–753. https://doi.org/10.1146/annurev-economics-080218-025643.

Angrist, Joshua, and Michal Kolesár. 2024. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *Journal of Econometrics* 240, no. 2 (March): 105398. https://doi.org/10.1016/j.jeconom.2022.12.012.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67, no. 3 (July): 499–527. https://doi.org/10.1111/1467-937X.00141.

Angrist, Joshua D., and Alan B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (November): 979–1014. https://doi.org/10.2307/2937954.

References ii

Bound, John, David A. Jaeger, and Regina M. Baker. 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, no. 430 (June): 443–450. https://doi.org/10.1080/01621459.1995.10476536.

Campbell, John Y. 2003. "Consumption-Based Asset Pricing." Chap. 13 in *Handbook of the Economics of Finance,* edited by G. M. Constantinides, M. Harris, and R. Stulz, vol. 1B, 803–887. Amsterdam: Elsevier. https://doi.org/10.1016/S1574-0102(03)01022-7.

Chamberlain, Gary. 2007. "Decision Theory Applied to an Instrumental Variables Model." *Econometrica* 75, no. 3 (May): 609–652. https://doi.org/10.1111/j.1468-0262.2007.00764.x.

Fieller, Edgar C. 1932. "The Distribution of the Index in a Normal Bivariate Population." *Biometrika* 24, nos. 3/4 (November): 428–440. https://doi.org/10.1093/biomet/24.3-4.428.

———. 1940. "The Biological Standardization of Insulin." *Supplement to the Journal of the Royal Statistical Society* 7 (1): 1–64. https://doi.org/10.2307/2983630.

———. 1954. "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society. Series B (Methodological)* 16, no. 2 (July): 175–185. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x.

Frandsen, Brigham, Lars Lefgren, and Emily Leslie. 2023. "Judging Judge Fixed Effects." *American Economic Review* 113, no. 1 (January): 253–277. https://doi.org/10.1257/aer.20201860.

References iii

Goff, Leonard. 2024. “A Vector Monotonicity Assumption for Multiple Instruments.” *Journal of Econometrics* 241, no. 1 (April): 105735. https://doi.org/10.1016/j.jeconom.2024.105735.

Goldberger, Arthur S., and Ingram Olkin. 1971. “A Minimum-Distance Interpretation of Limited-Information Estimation.” *Econometrica* 39, no. 3 (May): 635–639. https://doi.org/10.2307/1913273.

Heckman, James J., and Edward J. Vytlacil. 2005. “Structural Equations, Treatment Effects and Econometric Policy Evaluation.” *Econometrica* 73, no. 3 (May): 669–738. https://doi.org/10.1111/j.1468-0262.2005.00594.x.

Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62, no. 2 (March): 467–475. https://doi.org/10.2307/2951620.

Kolesár, Michal. 2013. “Estimation in an Instrumental Variables Model With Treatment Effect Heterogeneity.” Working paper, Princeton University, November. https://www.princeton.edu/~mkolesar/papers/late_estimation.pdf.

Kolesár, Michal, and Mikkel Plagborg-Møller. 2025. “Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly,” March. arXiv: 2411.10415.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. 2022. “Valid t-Ratio Inference for IV.” *American Economic Review* 112, no. 10 (October): 3260–3290. https://doi.org/10.1257/aer.20211063.

References iv

Lee, Seojeong. 2018. “A Consistent Variance Estimator for 2SLS When Instruments Identify Different LATEs.” *Journal of Business & Economic Statistics* 36, no. 3 (July): 400–410. https://doi.org/10.1080/07350015.2016.1186555.

Mariano, Roberto S., and James B. McDonald. 1979. “A Note on the Distribution Functions of LIML and 2SLS Structural Coefficient in the Exactly Identified Case.” *Journal of the American Statistical Association* 74, no. 368 (December): 847–848. https://doi.org/10.1080/01621459.1979.10481040.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R. Walters. 2021. “The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables.” *American Economic Review* 111, no. 11 (November): 3663–3698. https://doi.org/10.1257/aer.20190221.

Montiel Olea, José Luis, and Carolin Pflueger. 2013. “A Robust Test for Weak Instruments.” *Journal of Business & Economic Statistics* 31, no. 3 (July): 358–369. https://doi.org/10.1080/00401706.2013.806694.

Moreira, Marcelo J. 2003. “A Conditional Likelihood Ratio Test for Structural Models.” *Econometrica* 71, no. 4 (July): 1027–1048. https://doi.org/10.1111/1468-0262.00438.

———. 2009. “Tests with Correct Size When Instruments Can Be Arbitrarily Weak.” *Journal of Econometrics* 152, no. 2 (October): 131–140. https://doi.org/10.1016/j.jeconom.2009.01.012.

Müller, Ulrich K., and Andriy Norets. 2016. “Credibility of Confidence Sets in Nonstandard Econometric Problems.” *Econometrica* 84, no. 6 (November): 2183–2213. https://doi.org/10.3982/ECTA14023.

References v

Richardson, David H. 1968. "The Exact Distribution of a Structural Coefficient Estimator." *Journal of the American Statistical Association* 63, no. 324 (December): 1214–1226. https://doi.org/10.1080/01621459.1968.10480921.

Sawa, Takamitsu. 1972. "Finite-Sample Properties of the k-Class Estimators." *Econometrica* 40, no. 4 (July): 653–680. https://doi.org/10.2307/1912960.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65, no. 3 (May): 557–586. https://doi.org/10.2307/2171753.

Stock, James H., and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." Chap. 5 in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,* edited by Donald W. K. Andrews and James H. Stock, 80–108. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511614491.006.

Yogo, Motohiro. 2004. "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak." *Review of Economics and Statistics* 86, no. 3 (August): 797–810. https://doi.org/10.1162/0034653041811770.

Young, Alwyn. 2022. "Consistency without Inference: Instrumental Variables in Practical Application." *European Economic Review* 147 (August): 104112. https://doi.org/10.1016/j.euroecorev.2022.104112.

Zellner, Arnold. 1970. "Estimation of Regression Relationships Containing Unobservable Independent Variables." *International Economic Review* 11, no. 3 (October): 441–454. https://doi.org/10.2307/2525323.

References vi

Zellner, Arnold. 1978. “Estimation of Functions of Population Mean and Regression Coefficients Including Structural Coefficients: A Minimum Expected Loss (MELO) Approach.” *Journal of Econometrics* 8, no. 2 (October): 127–158. https://doi.org/10.1016/0304-4076(78)90024-6.