Econometrics 2

Instrumental Variables Regression – Part II

Lasha Chochua

2026

Checking Instrument Validity

Checking Instrument Validity

- Instrumental variables (IV) regression is only useful if instruments are valid.
- Invalid instruments lead to meaningless results.
- We must check instrument validity in each application.

- There are **two conditions** to verify:

  1. **Relevance**: $\text{corr}(Z_i, X_i) \neq 0$ – testable from the data.
  2. **Exogeneity**: $\text{corr}(Z_i, u_i) = 0$ – largely untestable; requires economic reasoning.

Note

These two conditions are **not symmetric**: relevance can be checked with a first-stage regression; exogeneity is an assumption that data alone cannot verify in the exactly-identified case.

Condition 1: Instrument Relevance

Instrument Relevance

- Instrument relevance is like sample size: more relevance means more information.
- If instruments explain more variation in X, IV regression is more accurate.
- A more relevant instrument gives better estimates, just like a larger sample.
- Statistical inference in TSLS uses the normal approximation.
- This approximation works better if the instruments are highly relevant.

Example of a Weak Instrument

- Instruments that explain little variation in X are **weak instruments**.
- Example: **Distance to a cigarette factory** as an instrument for price.
  - Greater distance may increase shipping costs.
  - But cigarettes are lightweight, so shipping costs are small.
  - Price variation due to distance is likely very small.
- Therefore, distance is probably a **weak instrument** in this case.

Why Weak Instruments Are a Problem

- With weak instruments, TSLS does **not** follow a normal distribution.
- Even in large samples, standard inference becomes unreliable.
- Confidence intervals may not cover the true value 95% of the time.
- TSLS may become biased toward OLS.

Technical Problem: Irrelevant Instruments

- Recall from Part I that the TSLS estimator is:

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}$$

- If Z_i is irrelevant: $\text{cov}(Z_i, X_i) = 0 \Rightarrow s_{ZX} \xrightarrow{p} 0$
- The denominator vanishes $\Rightarrow$ TSLS breaks down:
  - The sampling distribution becomes **non-normal**.
  - It behaves like a ratio of two normal variables – not normal itself.
  - This makes inference unreliable even in large samples.

> **Note**
>
> Completely irrelevant instruments are rare, but **weak** ones – where $\text{cov}(Z_i, X_i)$ is small but nonzero – are common and cause the same qualitative problems.

How to Check for Weak Instruments

- Suppose there is a **single endogenous regressor**.
- Compute the **first-stage F-statistic** from the regression of X on the instruments.
- It tests whether the instruments are jointly significant.

Rule of Thumb

- If the first-stage F-statistic is **greater than 10**: weak instrument concerns are minimal (more on this later).
- If it is **below 10**: the instruments may be weak – inference is unreliable.

Testing for Weak Instruments

The Bias Approximation

- Weak instruments make TSLS estimates biased and inference unreliable.
- We can understand the potential **bias of TSLS** using the following approximation.
- Let:
  - β_1 = true structural coefficient
  - β_1^{OLS} = probability limit of the OLS estimator (biased if X is endogenous)
  - $\mathbb{E}(F)$ = expected value of the first-stage F-statistic
- Under many instruments, the bias of TSLS can be approximated by:

$$\mathbb{E}(\hat{\beta}_1^{TSLS}) - \beta_1 \approx \frac{\beta_1^{OLS} - \beta_1}{\mathbb{E}(F) - 1} \tag{1}$$

- The numerator is the **OLS bias**: how far OLS is from the truth.
- The denominator scales this bias by instrument strength.

What the Bias Formula Tells Us

- From (1): the larger $\mathbb{E}(F)$, the smaller the TSLS bias.

$\mathbb{E}(F)$	**TSLS bias as % of OLS bias**	**Interpretation**
1	∞	Instrument irrelevant – TSLS undefined
2	100%	TSLS as bad as OLS
10	$\approx 11\%$	Rule-of-thumb threshold
∞	0%	Perfect instrument – no bias

- At $\mathbb{E}(F) = 10$: TSLS bias $\approx \frac{1}{9}$ of OLS bias – acceptable in most applications.
- This is why $F > 10$ is used as the **threshold** for acceptable instrument strength.

Stock–Yogo Test

- Stock and Yogo (2005) provide a formal test for weak instruments.
  - H_0: instruments are weak (TSLS bias $\geq 10\%$ of OLS bias).
  - H_1: instruments are strong (TSLS bias $< 10\%$ of OLS bias).
- Compare first-stage F-statistic to a **critical value** that depends on:
  - Number of instruments m
  - Desired maximum bias level
- At 5% significance level, the critical value lies between **9.08 and 11.52** depending on m.
- So $F > 10$ is a reliable (?!) rule of thumb across most practical settings.

What to Do if Instruments Are Weak

- If you have many instruments, some might be weak.
- Best strategy: **drop weak instruments** and keep the most relevant ones.
- This improves the accuracy of TSLS and avoids misleading inference.

- If dropping instruments raises standard errors, that is acceptable.
  - The original standard errors were unreliable anyway.

- If the model is **exactly identified** or dropping instruments risks under-identification:
  1. **Find better instruments** – requires **deeper knowledge** of the **empirical problem**.
  2. **Use weak-instrument-robust methods** – the Anderson–Rubin test (covered next).

Inference Robust to Weak Instruments

Hypothesis Testing with Weak Instruments

- When instruments are weak, the TSLS t-test for $H_0 : \beta_1 = \beta_{1,0}$ is **not valid**.
- Why? TSLS has a biased and non-normal distribution in this case.

- Instead, we use a test that is valid **even if instruments are weak**:
  - **Anderson–Rubin (AR) test**

- The AR test works by **inverting a hypothesis test** to construct a confidence set:
  - Rather than estimating β_1 and building a CI around it,
  - It asks: for which values of $\beta_{1,0}$ does the data **not reject** H_0?
  - The set of non-rejected values is the confidence set for β_1.

Anderson–Rubin (AR) Test

- The AR test evaluates $H_0 : \beta_1 = \beta_{1,0}$ in two steps:
  1. Construct a transformed dependent variable:
  $$Y_i^* = Y_i - \beta_{1,0} X_i$$
  2. Regress Y_i^* on all instruments Z_i and controls W_i.
- The AR test checks **whether the instruments jointly explain** Y_i^*.
- Use the F-**statistic** from this auxiliary regression.

Anderson–Rubin (AR) Test

- The AR test evaluates $H_0 : \beta_1 = \beta_{1,0}$ in two steps:
  1. Construct a transformed dependent variable:

$$Y_i^* = Y_i - \beta_{1,0} X_i$$

  2. Regress Y_i^* on all instruments Z_i and controls W_i.

- The AR test checks **whether the instruments jointly explain** Y_i^*.
- Use the F-**statistic** from this auxiliary regression.

- If H_0 is true and instruments are valid:
  - $Y_i^* = u_i + (\beta_1 - \beta_{1,0}) X_i = u_i$ under H_0
  - So Y_i^* should be **uncorrelated with** $Z_i \Rightarrow F$**-statistic should be small**.

AR Test – Why It Works

- The key insight: $Y_i^* = Y_i - \beta_{1,0} X_i$ removes the endogenous variation **if** $\beta_{1,0} = \beta_1$.
- Under $H_0 : \beta_1 = \beta_{1,0}$:

$$Y_i^* = \beta_0 + (\beta_1 - \beta_{1,0}) X_i + u_i = \beta_0 + u_i$$

- Since instruments are exogenous (cov$(Z_i, u_i) = 0$), they should not explain Y_i^*.
- A large F-statistic $\Rightarrow$ Z_i explains Y_i^* $\Rightarrow$ reject H_0.

Confidence Set via AR Test

Repeat the AR test over a grid of $\beta_{1,0}$ values. The confidence set for β_1 is the collection of all values **not rejected** at the chosen significance level. This procedure is valid regardless of instrument strength.

Estimation Under Weak Instruments

- If instruments are irrelevant, **no consistent estimator** for β_1 exists.
- Even large samples cannot compensate without stronger assumptions.

- **Practical recommendation**:
  - Prefer **AR confidence sets** over point estimation when $F < 10$.
  - Report the confidence set alongside (or instead of) the point estimate.
  - State explicitly that identification is weak if F is below the Stock–Yogo critical value.

- Weak instruments in IV regression remain an **active area of econometric research**.

> **Note**
>
> We introduced the AR test here as a remedy for weak instruments. In the next section we will see that it is useful in **all** IV applications – even when instruments are strong.

Beyond Weak Instruments: A New Warning

The Old Weak-IV Warning (I)

- The traditional weak-instrument warning is familiar:

Warning

If the instrument is weak, then 2SLS estimates may be biased toward OLS, and the usual 2SLS t-test may reject too often.

- A common rule of thumb is:

$$F_{first-stage} > 10$$

- The idea is simple:
  - If Z strongly predicts X, then the denominator of the IV estimator is stable.
  - If Z weakly predicts X, then small random sample correlations can move the IV estimate a lot.

The Old Weak-IV Warning (II)

- But **Keane and Neal (2024)** argue that this traditional warning is incomplete.

New warning

The usual 2SLS t-test can be misleading even when instruments are not conventionally weak.

The 2SLS Estimator in One Equation

- Consider the simple model: $Y_i = \beta X_i + u_i$
  - with first stage: $X_i = \pi Z_i + e_i$
- The 2SLS estimator can be written as:

$$\hat{\beta}_{2SLS} = \beta + \frac{\widehat{\text{cov}}(Z_i, u_i)}{\widehat{\text{cov}}(Z_i, X_i)}$$

- In the population, the instrument is valid: $\text{cov}(Z_i, u_i) = 0$
- But in any finite sample: $\widehat{\text{cov}}(Z_i, u_i) \neq 0$

Note

Even with a valid instrument, the sample correlation between Z_i and u_i is almost never exactly zero.

The Root of the Problem

- The finite-sample error in 2SLS comes from $\widehat{\text{cov}}(Z_i, u_i)$.
- If this accidental sample covariance is positive, the 2SLS estimate is baised upward.
- If it is negative, the 2SLS estimate is baised downward.
- When OLS is upward biased, positive sample covariance between Z_i and u_i tends to pull 2SLS toward OLS.
- This by itself is not the surprising part.

The surprising part

The same samples in which 2SLS is pulled toward OLS are also the samples in which the 2SLS standard error becomes artificially small.

The Problem with the 2SLS t-test (I)

- After running 2SLS, we usually test $H_0 : \beta = 0$ using:

$$t_{2SLS} = \frac{\hat{\beta}_{2SLS}}{SE(\hat{\beta}_{2SLS})}$$

- A large absolute t-statistic can come from two sources:

Reason 1	Reason 2
Large estimate	Small standard error

- In OLS, this logic is usually fine.
- In 2SLS, Keane and Neal argue that the standard error is not neutral.

The Problem with the 2SLS t-test (II)

Warning

The 2SLS standard error tends to be small precisely when the 2SLS estimate is close to OLS.

Why Does the Standard Error Shrink? (I)

- In the second stage, 2SLS uses the predicted component of X_i:

$$\hat{X}_i = \hat{\pi} Z_i.$$

- But the usual 2SLS residuals used for the standard error are computed with the original regressor X_i:

$$\hat{u}_i = Y_i - \hat{\beta}_{2SLS} X_i.$$

Note

This is why we should not treat the second stage as an ordinary OLS regression of Y_i on $\hat{X}_i$ with ordinary OLS standard errors.

Why Does the Standard Error Shrink? (II)

- Now recall a basic property of OLS:

> **Note**
>
> OLS chooses the coefficient that minimizes the sum of squared residuals:
>
> $$\sum_i (Y_i - bX_i)^2.$$

- Therefore, if $\hat{\beta}_{2SLS}$ happens to be close to $\hat{\beta}_{OLS}$, then

$$\sum_i (Y_i - \hat{\beta}_{2SLS} X_i)^2$$

is relatively small.

Why Does the Standard Error Shrink? (III)

- Small residual variance produces a smaller usual 2SLS standard error.
- But this smaller standard error does **not** necessarily mean that 2SLS is genuinely precise.

Interpretation

The usual 2SLS standard error partly reflects how close $\hat{\beta}_{2SLS}$ is to $\hat{\beta}_{OLS}$, because the residual variance is smallest near the OLS coefficient.

Two Things Happen in the Same Samples (I)

- Suppose OLS is upward biased:

$$\hat{\beta}_{OLS} > \beta$$

- Now suppose that, by chance, the instrument is positively correlated with the structural error in the sample:

$$\widehat{\text{cov}}(Z_i, u_i) > 0$$

- Then two things happen at the same time:
  - $\hat{\beta}_{2SLS}$ is pulled upward, toward the OLS estimate.
  - $SE(\hat{\beta}_{2SLS})$ becomes artificially small.

Two Things Happen in the Same Samples (II)

- Therefore:

$$|t_{2SLS}| = \left| \frac{\hat{\beta}_{2SLS}}{SE(\hat{\beta}_{2SLS})} \right|$$

  becomes large.

Bad news

The t-test is **most confident** exactly in the samples where 2SLS is **most contaminated** by OLS bias.

Power Asymmetry (I)

- Keane and Neal (2024) call this problem **power asymmetry**.
- The 2SLS t-test has high power to detect effects in the direction of the OLS bias.
- But it has low power to detect effects in the opposite direction.

Estimate close to OLS	**Estimate far from OLS**
Residuals are small	Residuals are large
Standard error is small	Standard error is large
t-statistic is large	t-statistic is small
t-test often rejects	t-test often fails to reject
Estimate is more biased	Estimate may be closer to truth

- This is the opposite of what we want from a statistical test.

A Simple Numerical Illustration (I)

- Suppose the true effect is zero:

$$\beta = 0$$

- Suppose OLS is upward biased:

$$E(\hat{\beta}_{OLS}) = 0.6$$

- Keane and Neal simulate data where the instrument passes the usual rule of thumb.
- With first-stage strength around the conventional threshold, the usual 2SLS t-test rejects at roughly the correct overall rate.
- But almost all rejections occur on the positive side.

A Simple Numerical Illustration (II)

	2SLS t-test	AR test
Overall rejection rate when $\beta = 0$	About 5%	About 5%
Rejections with $\hat{\beta}_{2SLS} > 0$	Almost all	About half
Rejections with $\hat{\beta}_{2SLS} < 0$	Almost none	About half

Lesson

The simulation shows the mechanism; the AER audit (next slide) shows it matters in real applied work.

The AER Audit

- Keane and Neal (2024) examine all replicable IV papers published in the *American Economic Review* from 2011 to 2023 where the first-stage F was below 50 or unreported.

- They find **49 replicable papers**.
- In **12 of those papers (24%)**, switching from the usual 2SLS t-test to the AR test overturns a key result.
- In 11 out of 12 cases: a result significant by t-test becomes **insignificant** under AR.

The pattern

The overturned results are systematically cases where the 2SLS estimate is close to OLS – exactly where we expect the power asymmetry to create false confidence.

Why This Is Not Just a Weak-IV Problem (I)

- One might think:

Note

"If the first-stage F is large enough, this problem should disappear."

- Keane and Neal argue that this is not true.
- The reason is that the power asymmetry comes partly from the second-stage residual variance.
- Since OLS minimizes residuals, the 2SLS standard error remains mechanically smaller when 2SLS is close to OLS.
- This logic does not disappear simply because the first stage is stronger.

Why This Is Not Just a Weak-IV Problem (II)

Warning

The usual first-stage F statistic checks instrument relevance. It does not check whether the 2SLS t-test has symmetric power.

Anderson–Rubin Test: The Basic Idea

- The Anderson–Rubin test avoids this problem by testing the reduced-form implication of the null.
- Start from the structural equation: $Y_i = \beta X_i + u_i$
- Under the null hypothesis $H_0 : \beta = \beta_0$:

$$Y_i - \beta_0 X_i = u_i$$

- If the instrument is valid, then under the null: $\mathrm{cov}(Z_i, Y_i - \beta_0 X_i) = 0$
- So the AR test asks: **Is the instrument correlated with the structural residual implied by the null hypothesis?**

AR Test for $H_0 : \beta = 0$

- For the simple case with one endogenous regressor and one instrument, testing $H_0 : \beta = 0$ is especially simple.
- Under H_0, the structural equation becomes $Y_i = u_i$.
- Therefore, if Z_i is valid, Z_i should not predict Y_i.
- The AR test runs the reduced-form regression:

$$Y_i = \gamma_0 + \gamma_1 Z_i + v_i$$

and tests $H_0 : \gamma_1 = 0$.

Interpretation

If Z_i significantly predicts Y_i, then the null $\beta = 0$ is rejected.

AR Test with Exogenous Controls

- Suppose the structural equation includes exogenous controls $\mathbf{W}_i$:

$$Y_i = \beta X_i + \mathbf{W}_i'\delta + u_i$$

- To test $H_0 : \beta = 0$, run:

$$Y_i = \gamma Z_i + \mathbf{W}_i'\alpha + v_i$$

  and test $H_0 : \gamma = 0$.

> **Note**
>
> With controls, the AR test asks whether Z_i predicts Y_i after controlling for $\mathbf{W}_i$.

AR Test for a General Null

- More generally, to test $H_0 : \beta = \beta_0$, construct the transformed outcome:

$$Y_i(\beta_0) = Y_i - \beta_0 X_i$$

- Then run:

$$Y_i - \beta_0 X_i = \gamma Z_i + \mathbf{W}_i' \alpha + v_i$$

  and test $H_0 : \gamma = 0$.

- If we reject, then β_0 is not compatible with the data.

AR confidence interval

The AR confidence interval is the set of all values β_0 that are not rejected by the AR test.

Practical Recommendation

- Do not rely only on the usual 2SLS coefficient and t-statistic.
- In IV applications, report:
  1. The first-stage coefficient and first-stage F statistic.
  2. The OLS estimate.
  3. The 2SLS estimate.
  4. The usual 2SLS standard error.
  5. The Anderson–Rubin test or AR confidence interval.
- Compare the 2SLS estimate to OLS.

Warning

If 2SLS is close to OLS and significant by the usual t-test, but insignificant by AR, be very cautious.

Final Takeaway

- The old lesson was:

Note

Weak instruments create biased 2SLS estimates and invalid t-tests.

- The new lesson is stronger:

Important

Even when instruments look reasonably strong, the usual 2SLS t-test can be most confident when the IV estimate is closest to OLS bias.

- Therefore, IV inference should not be based only on the usual 2SLS t-test.

Remember

Important

If the first-stage F is very large, especially above 50, weak-IV concerns are much less severe. But the usual 2SLS t-test can still have power asymmetry.

Therefore, report Anderson–Rubin inference alongside the usual 2SLS results, especially when $\hat{\beta}_{2SLS}$ is close to $\hat{\beta}_{OLS}$.

Condition 2: Instrument Exogeneity

Instrument Exogeneity

- Instruments must be **uncorrelated with the error term** u_i.
- If not, TSLS will be **inconsistent** – it converges to the wrong value.
- Exogeneity ensures that instruments shift X in a way unrelated to omitted variables.

- The whole logic of IV relies on instruments isolating variation in X that is unrelated to u_i.

Relevance vs. Exogeneity – A Critical Asymmetry

Testability Asymmetry

- **Relevance** $[\text{corr}(Z_i, X_i) \neq 0]$: **testable** from the data.
    - We know what to do.
- **Exogeneity** $[\text{corr}(Z_i, u_i) = 0]$: **cannot be tested** from the data in general.
    - u_i is unobserved – we never see whether Z_i correlates with it.
    - In the exactly-identified case, no test exists at all.
    - Requires economic reasoning, institutional knowledge, and judgment.

- This asymmetry is fundamental: **every IV application must defend exogeneity verbally**, not just statistically.

Can We Ever Test Exogeneity?

- **Exactly identified model** $(m = k)$: Cannot test exogeneity.
  - There are just enough instruments to estimate β.
  - No redundancy – no way to check if instruments are valid.
  - Must rely on **economic theory or expert judgment**.
- **Overidentified model** $(m > k)$: Can test **overidentifying restrictions**.
  - Extra instruments allow a partial check of instrument validity.
  - But note: the test only checks **internal consistency** across instruments – not true exogeneity.
  - If all instruments are invalid in the same direction, the test will not detect it.

Overidentifying Restrictions Test – Intuition

- Suppose you have 2 instruments and 1 endogenous variable ($m = 2$, $k = 1$: overidentified).
- Estimate TSLS twice – once with each instrument alone.
- If both are valid, the two estimates should be **similar**.
- If they differ greatly $\Rightarrow$ at least one instrument may be invalid.

Note

This intuition generalizes: with $m > k$ instruments, we have $m - k$ extra restrictions that can be tested. These are the **overidentifying restrictions** – they measure whether all instruments point to the same causal estimate.

Formal Overidentifying Restrictions Test

- Compute TSLS residuals:

$$\hat{u}_i^{TSLS} = Y_i - \hat{\beta}_0^{TSLS} - \hat{\beta}_1^{TSLS} X_{1i} - \cdots - \hat{\beta}_{k+r}^{TSLS} W_{ri} \tag{2}$$

- Regress $\hat{u}_i^{TSLS}$ on **all instruments** $Z_{1i}, \ldots, Z_{mi}$ and controls $W_{1i}, \ldots, W_{ri}$.
- H_0: all instruments are exogenous $\Rightarrow$ instruments should be **uncorrelated** with $\hat{u}_i^{TSLS}$.
- The test statistic is the J-**statistic**:

$$J = m \cdot F \;\sim\; \chi^2_{m-k} \quad \text{under } H_0 \tag{3}$$

- $m =$ number of instruments, $F = F$-statistic from the auxiliary regression.
- Degrees of freedom $= m - k =$ number of **overidentifying restrictions**.

The J-Test – Limitations

- The J-test only evaluates the **extra** (overidentifying) restrictions – not the model as a whole.

What the J-Test Cannot Do

- If the model is **exactly identified** $(m = k)$: $J = 0$ by construction – no test is possible.
- If all instruments are **invalid in the same direction** (correlated with u_i similarly), the J-test will not detect it – the estimates will agree even though all are biased.
- A **non-rejection** of H_0 does not prove instruments are exogenous – it only means the overidentifying restrictions are internally consistent.

- Exogeneity ultimately rests on **economic reasoning**, not statistical testing.

Application – J-Test for Cigarette Demand

- From Part I: two-instrument specification for cigarette demand.
  - Endogenous: $\ln(P_i^{cigarettes})$
  - Instruments: $SalesTax_i$ and $CigTax_i$ ($m = 2$, $k = 1 \Rightarrow 1$ overidentifying restriction)
- Steps:
  1. Estimate TSLS with both instruments $\Rightarrow$ obtain residuals $\hat{u}_i^{TSLS}$.
  2. Regress $\hat{u}_i^{TSLS}$ on $SalesTax_i$, $CigTax_i$, and $\ln(Inc_i)$.
  3. Compute $J = 2 \cdot F$ and compare to χ_1^2 critical value (3.84 at 5%).
- If $J < 3.84$: cannot reject that both instruments are exogenous $\Rightarrow$ consistent with validity.
- If $J > 3.84$: at least one instrument is likely correlated with u_i.

Note

A failure to reject does not prove both instruments are valid – it only means they are internally consistent. Economic judgment remains essential.

Where Do Valid Instruments Come From?

Where Do Valid Instruments Come From?

- Finding valid instruments is often the **hardest part** of IV estimation.
- Valid instruments must be both:
  - **Relevant**: explain variation in the endogenous regressor X.
  - **Exogenous**: uncorrelated with the error term u.
- There are **two main approaches** to constructing instruments:
  1. **Use economic theory** – identify variables that shift one side of the market.
  2. **Find exogenous variation** – exploit natural or quasi-experimental events.

1. Using Economic Theory

- Identify variables that **shift supply but not demand** (or vice versa).
- Example: Philip Wright used **weather** in agriculture as an instrument.
    - Weather affects supply but not demand.
- Limitation: economic models may be **nonlinear** or **incomplete**.
    - They are abstractions – not always practical for real data analysis.
    - The exclusion restriction (exogeneity) must be justified by the model's assumptions.

2. Finding Exogenous Variation

- Look for **natural or random events** that affect X but not u.
- Example: Earthquake damage changes class size in schools.
  - This change is unrelated to unobserved determinants of student performance.
- These sources create **quasi-experimental variation** in X.

- Advantage: Relies on **empirical context**, not abstract models.
- Requires detailed **knowledge of the data** and institutional setting.
- These are sometimes called **natural experiments**.

Application: Do Institutions Affect Economic Development?

Do Institutions Affect Economic Development?

- Key question: Why are some countries rich and others poor?
- Theory: Strong institutions (e.g. protection of property rights) help growth.
- But we cannot simply regress GDP per capita on institutions.

- Institutions and development are **endogenously linked**:
  - Economic growth can also lead to stronger institutions.
  - This creates **simultaneous causality bias** in OLS.
  - Additionally, omitted variables (e.g., culture, geography) affect both.

Endogeneity in Institutions Example

- OLS regression of development on institutions gives **biased results**.
- Control variables (e.g. geography, history) cannot solve this alone.
- Need an **instrument**: a variable that:
    1. Affects institutions (**relevance**),
    2. Does **not directly affect development** (**exogeneity**).
- Solution: Use TSLS with a valid instrument.

Acemoglu et al. (2001): Colonial Origins as an Instrument

- Idea: Use **settler mortality rates** during colonization as an instrument for current institutions.

- The causal chain:

$$\underbrace{Z_i}_{\text{settler mortality}} \longrightarrow \underbrace{X_i}_{\text{colonial institutions}} \longrightarrow \underbrace{Y_i}_{\text{current institutions}} \longrightarrow \underbrace{\text{GDP p.c.}}_{\text{development}}$$

- High mortality → extractive institutions (no settlement, extraction of resources).
- Low mortality → European-style property rights (settlers stayed and built institutions).
- These initial choices **persist** and shape today's institutions.

- Settler mortality does **not directly affect modern GDP** – only via the institutional channel.

Is Settler Mortality a Valid Instrument?

- **Relevance**: Settler mortality explains **27% of variation** in current institutions (F-statistic well above 10).
- **Exogeneity**:
  - Acemoglu et al. control for malaria incidence, latitude, continent dummies, and other geographic factors.
  - Results are robust across many control specifications.
  - They use **3 instruments** (split causal path), enabling an overidentification test.
  - The J-test does not reject – overidentifying restrictions are consistent.

Note

Settler mortality from the 1600s–1800s cannot plausibly affect modern GDP through any channel other than the institutional path – this is the key economic argument for exogeneity.

OLS vs. TSLS – Results

Specification	Coefficient on institutions	SE
OLS	0.52	0.06
TSLS (settler mortality)	0.94	0.16

- TSLS estimate is **almost twice** the OLS estimate.
  - OLS is downward biased: reverse causality (weak institutions $\leftrightarrow$ low development) attenuates the coefficient.
- In the TSLS model:
  - Africa dummy = **not significant** once institutions are instrumented.
  - Distance from equator = **not significant** once institutions are instrumented.
- Conclusion: "Africa is poor not because of geography, but because of worse institutions."

What Makes a Good Instrument – Summary

	Relevance	Exogeneity
Requirement	corr$(Z_i, X_i) \neq 0$	corr$(Z_i, u_i) = 0$
Testable?	Yes – first-stage F-statistic	No (exactly identified); partially (overidentified)
Rule of thumb	$F > 10$: minimum floor $F > 50$: reliable threshold Always report AR test; be especially cautious if $\hat{\beta}_{2SLS} \approx \hat{\beta}_{OLS}$	Economic reasoning + J-test if overidentified
If violated	TSLS variance explodes; bias toward OLS	TSLS inconsistent – converges to wrong value
Example	Rainfall $\rightarrow$ butter supply $\rightarrow$ price	Rainfall uncorrelated with demand shifters

Required Reading

- **Stock and Watson (2020)** – Chapter 12, Sections 12.3-12.6
- **Keane, M. P. and Neal, T. (2024)** *"A Practical Guide to Weak Instruments" Annual Review of Economics*, 16:185–212 (**Optional**)

Additional Slides

Why Are 2SLS Standard Errors Computed with X_i, Not $\hat{X}_i$? (I)

- Recall the structural equation: $Y_i = \beta_0 + \beta_1 X_i + u_i$.
- The structural error is: $u_i = Y_i - \beta_0 - \beta_1 X_i$.
- It is defined using the **original endogenous regressor** X_i – not the fitted value $\hat{X}_i$.
- This is the error term whose variance matters for correct 2SLS inference.
- If we manually run the second stage by regressing Y_i on $\hat{X}_i$, the residuals are:

$$\tilde{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{X}_i.$$

Important

These are **not** the structural residuals. They are residuals from an artificial regression that replaces X_i by $\hat{X}_i$.

Why Are 2SLS Standard Errors Computed with X_i, Not $\hat{X}_i$? (II)

- Write the first-stage decomposition as $X_i = \hat{X}_i + \hat{v}_i$, so $\hat{X}_i = X_i - \hat{v}_i$, where $\hat{v}_i$ is the first-stage residual.
- The manual second-stage residual is $\tilde{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{X}_i$.
- Substitute $\hat{X}_i = X_i - \hat{v}_i$:

$$\tilde{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 (X_i - \hat{v}_i) = \underbrace{(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)}_{\hat{u}_i} + \hat{\beta}_1 \hat{v}_i$$

- Therefore: $\tilde{u}_i = \hat{u}_i + \hat{\beta}_1 \hat{v}_i$

Consequence

The manual residual $\tilde{u}_i$ is not the structural residual – it contains an additional component $\hat{\beta}_1 \hat{v}_i$. Standard errors from the manual second stage are therefore invalid.

Why Are 2SLS Standard Errors Computed with X_i, Not $\hat{X}_i$? (III)

- Correct 2SLS residuals use the original structural equation:

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- As $n \to \infty$, under IV assumptions, $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$. Therefore:

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \xrightarrow{p} Y_i - \beta_0 - \beta_1 X_i = u_i$$

- So $\hat{u}_i$ consistently estimates the structural error u_i.
- Recall the asymptotic variance formula:

$$\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n} \cdot \frac{\text{var}\left[(Z_i - \mu_Z)u_i\right]}{\left[\text{cov}(Z_i, X_i)\right]^2}$$

- This formula involves u_i – defined with respect to X_i, not $\hat{X}_i$.

Why Are 2SLS Standard Errors Computed with X_i, Not $\hat{X}_i$? (IV)

Bottom line

Software uses X_i in the residuals because this consistently estimates var(u_i). Using $\hat{X}_i$ mixes the structural error with the first-stage residual and gives inconsistent standard errors.