ECON 594: Applied Economics

Instrumental Variables

Sam Norris
University of British Columbia

Where we are

So far our designs control for confounders (OLS, RD) or difference them away (DD)

- These designs each build a credible comparison group

Instrumental variables takes a different route

- Don't strip the bad variation out of the treatment
- Find a slice of the treatment that is good as-is, and use only that

You have seen the basics, so today is about making your IV analysis sing

Two ways to beat selection bias

Regression excludes the bad variation

- Hold the confounders fixed, compare units that differ only in the treatment
- You need to observe and control for what is confounding you

Instrumental variables isolates the good variation

- Find an instrument Z that shifts the treatment for reasons unrelated to the outcome
- Use only the part of the treatment that Z moves
- You do not need to observe the confounder

What IV is for

IV shows up in three recurring situations

A broken experiment

- Treatment was randomized, but people don't comply with their assignment

Omitted variables

- The treatment is chosen, so it correlates with things you can't control for

Measurement error

- A mismeasured regressor biases OLS toward zero; IV undoes it (we return to this in L8)

Fuzzy RD is a special case of an IV

Fuzzy RD was an instrumental variables design

- Crossing the cutoff doesn't switch treatment for everyone, only nudges the probability
- "Above the cutoff" is an instrument for treatment
- The jump in the outcome over the jump in treatment is the Wald ratio

Today we drop the cutoff and let the instrument be anything as-good-as-random

- Thinking through these issues in the RD case can be helpful

Running example: the KIPP Lynn lottery

KIPP is a network of oversubscribed charter schools

- Students who enroll outperform their peers
- Is that the school or selection?
- KIPP Lynn allocates its scarce seats by lottery

The lottery randomizes the *offer* of a seat, not attendance

- Winners are 74 pp more likely to attend, not 100 pp
- Attendance is a choice conditional on winning, so comparing attendees to non-attendees risks selection bias
- IV bridges from the random offer to the effect of attending

Angrist et al. (2012), "Who benefits from KIPP?", *JPAM*

What makes a variable an instrument

Relevance: the instrument moves the treatment

- Do lottery winners actually attend KIPP more often? Yes, by 74 pp
- This one you can test (first stage)

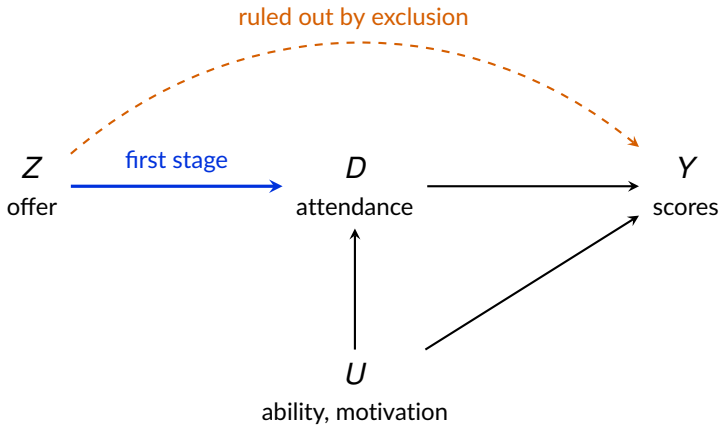
Exogeneity: the instrument is as good as randomly assigned

- The offer is unrelated to ability, motivation, family background
- Not directly testable; support it with balance checks

Exclusion: the instrument affects the outcome only through the treatment

- Winning the lottery raises scores only by getting you into KIPP, no other channel
- Calls back to “bundled treatment” idea from last class
- Empirically testable only if you observe other treatments

The three conditions



Exogeneity vs. exclusion

Same split we used for RD validity, now for the instrument

Exogeneity is about who gets the instrument

- Are winners and losers comparable before the lottery?
- Check balance on baseline characteristics, just like an RCT

Exclusion is about what the instrument does

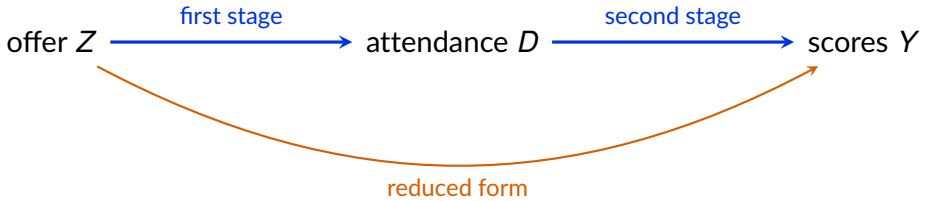
- Does the offer change anything besides attendance?
- If losing also discouraged kids from trying in school, exclusion fails

A randomized instrument gives exogeneity for free

- It never buys exclusion

IV as a chain reaction

The instrument hits the outcome only by way of the treatment



$$\underbrace{\text{effect of offer on scores}}_{\text{reduced form}} = \underbrace{\text{effect of offer on attendance}}_{\text{first stage}} \times \underbrace{\text{effect of attendance on scores}}_{\text{what we want}}$$

The Wald estimator

Rearrange the chain to solve for the thing we want

$$\text{effect of attendance on scores} = \frac{\text{effect of offer on scores}}{\text{effect of offer on attendance}} = \frac{\text{reduced form}}{\text{first stage}}$$

With a binary instrument this is the Wald estimator

$$\lambda = \frac{E[Y_i \mid Z_i{=}1] - E[Y_i \mid Z_i{=}0]}{E[D_i \mid Z_i{=}1] - E[D_i \mid Z_i{=}0]}$$

KIPP Lynn: the offer raised math scores by 0.355σ and attendance by 0.741

$$\lambda = \frac{0.355}{0.741} \approx 0.48\sigma \text{ per year of KIPP attendance}$$

The reduced form and the first stage

The reduced form/first stage is the actual experiment you ran

- The IV estimand is an interpretation of the RF

Always look at the reduced form and first stage before the IV

- If the offer doesn't move attendance, no hope to learn about the effect of attendance
- If the offer doesn't move scores, attendance didn't matter for compliers
- The 2SLS ratio can only inflate what is already in the RF

Generalizing IV with 2SLS

The Wald ratio handles one binary instrument and one binary treatment

Two-stage least squares does everything else

- Treatment can be binary, a count, or continuous
- Instruments can be binary, a count, or continuous
- You can add control variables
- You can use several instruments at once

It is the same IV idea, written as a regression

How 2SLS works

Stage one: regress the treatment on the instrument and keep the fitted values

$$D_i = \alpha_0 + \alpha_1 Z_i + u_i \qquad \Rightarrow \qquad \widehat{D}_i$$

$\widehat{D}_i$ is the part of the treatment driven only by the instrument

Stage two: regress the outcome on those fitted values

$$Y_i = \beta_0 + \beta_{2SLS}\, \widehat{D}_i + e_i$$

β_{2SLS} is the IV estimate

- With one binary instrument it is exactly the Wald ratio

Covariates and multiple instruments

Controls go in every equation

- First stage, reduced form, and second stage all carry the same covariates X
- 2SLS then uses only the instrument variation left after partialling out X

You can stack instruments for one endogenous treatment

- Quantity–quality of children: twins and same-sex siblings both shift family size
- Twins are a big shifter but rare; same-sex is a small shifter but common
- Using both at once is more efficient and lets you test them against each other

Angrist, Lavy & Schlosser (2010); Angrist & Evans (1998)

2SLS in Stata

Don’t run the two stages by hand

- Second-stage standard errors come out wrong
- Use a single command

```
ivreghdfe y x (d = z1 z2), absorb(fe) cluster(id)
```

Compare to the OLS version

```
reghdfe y d x, absorb(fe) cluster(id)
```

Bonus: `ivreghdfe` also prints first-stage F and weak-instrument diagnostics

Fuzzy RD is just 2SLS

Maimonides (12th century Spanish rabbi) wrote that class sizes should not exceed 40

- Israeli law caps class size today

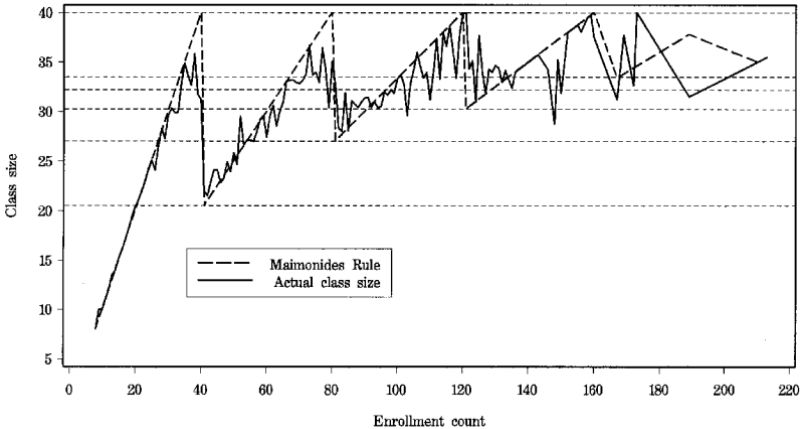
Move from 40 to 41 students enrolled and the school must split into a second class

- Can split earlier, or combine classes $\rightarrow$ imperfect compliance
- Angrist & Lavy (1999) use this setup to study the effect of class sizes on student performance

IV setup:

- Treatment: class size
- Instrument: enrollment past a multiple of 40
- Outcome: test scores
- Controls: smooth enrollment on each side

Maimonides' first stage



Class size jumps down at each multiple of 40, but not all the way

IV does not estimate the average effect

2SLS only uses people whose treatment the instrument actually moves

Sort everyone by how they respond to the instrument

- Compliers: take the treatment when offered, not otherwise
- Always-takers: take it no matter what
- Never-takers: refuse it no matter what
- Defiers: do the opposite of their assignment

Always- and never-takers contribute no first-stage variation, so IV says nothing about them

The local average treatment effect

Under one extra assumption, IV identifies the effect on compliers

Monotonicity: the instrument pushes everyone the same direction

- Winning the lottery never makes someone less likely to attend
- Rules out defiers

Then 2SLS estimates the local average treatment effect (LATE)

- The average effect among compliers, not the whole population
- “Local” to whoever the instrument happens to move

Imbens & Angrist (1994); Angrist, Imbens & Rubin (1996)

Different instruments, different compliers

Two valid instruments can give different estimates and both be right

Each instrument moves its own set of compliers

- Twins shift family size for almost everyone
- Same-sex composition shifts it only for parents who want variety
- If effects are heterogeneous, these are different populations

This reframes how we read IV

- The estimate is specific to the instrument, not a universal parameter
- For a thesis: say whose effect you are estimating, and whether that group is the policy-relevant one

When the treatment isn't binary

With a multi-valued treatment, "compliers" isn't a single group

2SLS estimates an average causal response (ACR)

- A weighted average of the per-unit effect across every margin the instrument shifts
- Weights are largest where the instrument moves the most mass

The lesson is the same: IV reweights toward where the instrument has bite

Angrist & Imbens (1995), variable treatment intensity

Why a weak first stage is dangerous

IV is a ratio with the first stage in the denominator

$$\widehat{\lambda} = \frac{\widehat{\text{reduced form}}}{\widehat{\text{first stage}}}$$

When the first stage is near zero, trouble follows

- A tiny denominator makes the estimate explode and swing wildly
- Any small violation of exogeneity gets divided by that tiny number too
- 2SLS becomes biased back toward OLS
- Standard errors understate the true uncertainty

Testing the first stage

Report the first-stage F-statistic on the excluded instruments

The old rule of thumb

- $F > 10$ was treated as “strong enough”
- Staiger & Stock (1997); Stock & Yogo (2005)

The modern correction

- $F > 10$ is not nearly enough for honest inference
- A valid 5% t-test can require an effective F above 100
- Report tF-adjusted confidence intervals, which widen smoothly as F falls
- Lee, McCrary, Moreira & Porter (2022, *AER*)

Living with weak instruments

Prefer just-identified IV

- One instrument for one treatment is approximately median-unbiased
- Piling on weak instruments makes the bias worse, not better
- “Just-ID IV is your friend”: Angrist & Kolesar (2024)

Use weak-instrument-robust inference

- Anderson–Rubin confidence sets stay valid no matter how weak the first stage
- They invert the reduced form, so they never divide by a small number

Testing exclusion

Start with institutional knowledge

- Tell a clean story for why the instrument moves only the focal treatment
- List the plausible alternative channels and rule them out one by one

Placebo reduced forms

- Find a group the instrument shouldn't affect, and estimate the first stage (ideally it is zero)
- If the reduced form is nonzero there, some other channel is live
- A nonzero effect where there is no change in treatment is exclusion failing

The overidentification test

With more instruments than treatments, you can test whether they agree

The logic

- Each instrument gives its own IV estimate
- If all are valid, the estimates should line up
- Sargan–Hansen J tests whether they do

But a rejection is ambiguous

- It could mean an instrument is invalid
- It could equally mean the instruments move different compliers (heterogeneity)

Passing the test does not prove exclusion: invalid-but-agreeing instruments slip through

- Still useful to report

Overidentification as a heterogeneity probe

The modern reading flips the test into a tool

Disagreement is information about treatment-effect heterogeneity

- Different instruments estimating different LATEs is a finding, not just a failure
- Ask which compliers each instrument is moving

You can build the heterogeneity in on purpose

- Interact the instrument with a covariate to make extra instruments
- Comparing those estimates maps how the effect varies across groups
- The overid machinery becomes a way to learn, not just to reject

What a good IV table shows

Give the reader all the pieces

Report all four pieces

- OLS, so we can see how much IV moves the answer
- The first stage, with its F-statistic
- The reduced form (or at least don't hide it)
- The 2SLS estimate, with weak-IV-robust intervals if the F is low

And say who the compliers are, so we know whose effect this is

A real example: pretrial detention

Bail judges are assigned as good as randomly and differ in leniency

- Treatment: pretrial release; instrument: the assigned judge's leniency
- A defendant's outcome shouldn't depend on which judge they drew, except through release

	Pretrial release (1)	Any guilty offense (2)
Judge leniency	0.700	−0.109
	(0.029)	(0.024)
First-stage F	569.1	
Observations	426,030	426,030

Drawing a more lenient judge raises release and lowers conviction

Dobbie, Goldin & Yang (2018, *AER*)

OLS vs 2SLS for pretrial release

Outcome: Any guilty offense		
	OLS (1)	2SLS (2)
Pretrial release	−0.050	−0.156
	(0.002)	(0.034)
First-stage F		569.1
Observations	426,030	426,030

Release cuts conviction by 16 pp for compliers; OLS sees only 5 pp

Dobbie, Goldin & Yang (2018, *AER*)

Reading the pretrial table

Why is 2SLS three times OLS here?

- OLS compares released and detained defendants, who differ in unobserved risk
- 2SLS isolates the marginal defendants whose release hinged on the judge
- Bigger IV than OLS is common

Who are these compliers?

- From subgroup first stages: about $1.3\times$ as likely to face a misdemeanor, $0.7\times$ a felony
- Roughly $1.2\times$ as likely to have a prior offense
- The estimate is the effect for marginal, lower-level cases: exactly the bail-reform margin
- Report your complier characteristics!

Extra tools for specific situations

The IV toolkit has grown well past one instrument, one treatment

Reach for these when the setting calls for it

- Examiner / judge designs: leniency of a random decision-maker as the instrument (Aizer & Doyle 2015; Dobbie, Goldin & Yang 2018; Bhuller et al. 2020); watch monotonicity (Frandsen, Lefgren & Leslie 2023)
- Variable treatment intensity: read 2SLS as an average causal response (Angrist & Imbens 1995)
- Several treatments at once: needs stronger monotonicity (Kirkeboen, Leuven & Mogstad 2016; Bhuller & Sigstad 2024)
- Nonlinear or limited-dependent models: control-function approach (Heckman & Robb 1985; Wooldridge 2015)
- Many small shocks: shift-share / Bartik instruments (Goldsmith-Pinkham, Sorkin & Swanson 2020; Borusyak, Hull & Jaravel 2022)

A checklist for your thesis

Before you call something an instrument, answer all of these

- First stage: is it strong? Report the F, not just significance
- Exogeneity: is the instrument balanced on baseline characteristics?
- Exclusion: what is the one channel, and what rules out the others?
- LATE: whose effect is this, and is that group the one you care about?
- Presentation: OLS, first stage, reduced form, 2SLS
- If the first stage is shaky: just-ID and weak-IV-robust intervals

Coming up

Next: measurement error and other empirical concerns

- Why a mismeasured regressor biases you toward zero
- How IV quietly fixes it
- And the other quiet threats: bad controls, selection, inference

Then we turn to how to present and how to write