2. Instrumental Variables

PhD Applied Methods

Duncan Webb
NovaSBE

Spring 2026

Treatment effects models: 2. Advanced instrumental variables

Why instrumental variables?

The problem: Treatment is often endogenous — people who choose treatment differ in unobservable ways

Example: Does college education increase wages?

- Selection bias: $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(1) - Y_i(0)]$

Why instrumental variables?

The problem: Treatment is often endogenous — people who choose treatment differ in unobservable ways

Example: Does college education increase wages?

- Selection bias: $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(1) - Y_i(0)]$

Solution: Find an **instrument** — a source of exogenous variation in treatment

- Affects treatment but not the outcome directly
- Examples: Draft lottery, proximity to college, scholarship eligibility

Why instrumental variables?

The problem: Treatment is often endogenous — people who choose treatment differ in unobservable ways

Example: Does college education increase wages?

- Selection bias: $\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1] \neq \mathbb{E}[Y_i(1) - Y_i(0)]$

Solution: Find an **instrument** — a source of exogenous variation in treatment

- Affects treatment but not the outcome directly
- Examples: Draft lottery, proximity to college, scholarship eligibility

This week's goal: Master IV theory and practice, including three key challenges:

1. Justifying the **exclusion restriction**
2. Understanding the **Local Average Treatment Effect** (LATE)
3. Dealing with **weak instruments**

Roadmap for today

1. **Basics of instrumental variables**: What is an instrument? 2SLS and the Wald estimator
2. **The exclusion restriction challenge**: Why ”as-good-as-random” is not enough
3. **Heterogeneous treatment effects and LATE**: Compliers, always-takers, never-takers — why IV estimates effects only for those who respond to the instrument
4. **Weak instruments**: Finite-sample bias and the first-stage F-statistic

Outline

What is an instrumental variable?

- Let's start with the definition in the context of a DAG
- Consider an effect we are interested in identifying: T on Y
  - In this setting, we know it is not identifiable

$$
\begin{tikzcd}
U \arrow[d] \arrow[dr] & \\
T \arrow[r] & Y
\end{tikzcd}
$$

What is an instrumental variable?

- Now, we have a variable Z which can identify two effects:
  - Z on T
  - Z on Y
- What is the content of this instrumental variable, Z?
  - It affects Y (**Relevance**)
  - It only affects Y through T (**Exclusion**)
- Without further assumptions, it won't be possible to identify the effect of T on Y using this, but it highlights the features of an IV
  - We'll discuss why shortly

$$
\begin{tikzcd}
 & U \arrow[d] \arrow[dr] & \\
Z \arrow[r] & T \arrow[r] & Y
\end{tikzcd}
$$

2SLS

$$Y = T\beta + \varepsilon \tag{1}$$

$$T = Z\pi + v \tag{2}$$

Reduced form

$$Y = Z(\pi \times \beta) + (u\beta + \varepsilon) \tag{3}$$

2SLS

$$Y = [Z\hat{\pi}] \times \beta + (u\beta + \varepsilon) \tag{4}$$

$$Y = \hat{T} \times \beta + (u\beta + \varepsilon) \tag{5}$$

Remember the reduced form

- Z has no *direct* effect on Y
- If Z is found to be correlated with Y, can only result from Z affecting T and T affecting Y
- Therefore, reveals that T affects Y
- To quantify: divide by π

Wald estimator

Z is a dummy $(= 0/1)$

$$E(Y|Z=0) = E(T|Z=0)\beta + E(\varepsilon|Z=0) \tag{6}$$
$$E(Y|Z=1) = E(T|Z=1)\beta + E(\varepsilon|Z=1) \tag{7}$$

and $E(\varepsilon|Z) = 0$

$$\beta = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(T|Z=1) - E(T|Z=0)} \tag{8}$$

Same interpretation: reduced form, divided by impact of Z on T

NB: the Wald estimator is an instrumental variable estimator

Getting close to random draws

Random events

- Actual random draw: Angrist (1990) Vietnam veterans randomly designated based on birth day
- Natural randomness: Angrist & Krueger (1991) quarter of birth affects school duration
- Natural randomness: Paxton (1992) climate shock affect income
- Natural randomness: Angrist & Evans (1998): have same-sex or different-sex children

Institutional rules that should have no relation with the outcome variable

- Levitt (1997) Local election to estimate impact of police on crime
- Duflo (2001) School building program / returns to schooling

The necessary assumptions so far

- So far, we need the following assumptions (and this is what you should always discuss when writing a paper on IV):
  1. relevance $cov(X, Z) \neq 0$
  2. exclusion $E(\varepsilon|Z) = 0$

- Tricky part starts now. Two main issues with this setup:
  1. **Exclusion restriction** - it's challenging to think about whether the exclusion restriction is true in terms of potential outcomes
  2. **Homogenous effects**: we have assumed homogeneous effects. E.g. β is the same for all individuals.
     - This is fixable in the model, but question is what estimand do we have?

Outline

Why is the exclusion restriction challenging?

- Recall the key (untestable) feature for IV: exclusion restriction

- In the context of the DAG, the intuition is that Z only affects Y through T

- Intuitively, it feels like something randomly assigned or nearly random should satisfy this, so long as it affects T

- This is not sufficient
    - You need to think critically about the IV

$$\begin{tikzcd} & U \arrow[d] \arrow[dr] & \\ Z \arrow[r] & T \arrow[r] & Y \end{tikzcd}$$

Why is the exclusion restriction challenging?

- Consider two examples. First, using Vietnam war lottery numbers as an IV for military service, studying the impact on mortality.
  - Y: death, T: vietnam vet, Z: lottery number
- Lottery number was randomly assigned as a function of birthdate
  - Well-defined design-based view of Z allocation!
- Does that necessarily satisfy exclusion restriction? Seems like a pretty slam dunk IV
  - Clearly affects veteran status
  - Clearly random!

Why is the exclusion restriction challenging?

- Does that necessarily satisfy exclusion restriction?
  - Not necessarily!
- Why? Consider one simple example: being drafted induces you to change your behavior to avoid the draft
  - Stay in school
  - Flee to Canada
- This would violate the exclusion restriction!

Why is the exclusion restriction challenging?

- Second, consider rainfall as an instrument for income in agriculture environments (many crops are heavily dependent on it)
  - This is not uncommon in development papers, as Sarsons (2015) points out
  - Y: conflict, T: income, Z: rainfall
- Exclusion restriction is that rainfall has no effect on conflict beyond income
  - While the logic seems reasonable, Sarsons (2015) shows that places with dams (which protect against the income shocks due to rain) have similar conflict to those without dams
- Plausible that while rain is "random", it might have many channels

Exclusion Restrictions

- Even with a variable that is near-random in its allocation, the exclusion restriction is not always satisfied
  - Worse yet, it's a fundamentally untestable restriction
- Using an IV requires thinking carefully about justifying the exclusion restriction
  - It can also be useful to think about what violations in the restriction implies

Outline

We are now going to discuss a fundamental concern about IV methods.

That IVs do not in general the **Average Treatment Effect**, but the “**Local Average Treatment Effect**” on a subpopulation of individuals called compliers.

Hypothetical model

- Estimate the wage impact of college rather than high school education
- Education is endogenous
- Instrument for education: whether the individuals are eligible to a college scholarship (to fix ideas, assume this has been randomized)

Notations:

$$W : \text{wage} \tag{9}$$
$$S = 1 \text{ if college} \tag{10}$$
$$Z = 1 \text{ if scholarship} \tag{11}$$

Z=0 (no scholarship)	**Z=1** (scholarship)
100 High school	0 High school
0 College	100 College
Average wage: 100	Average wage: 130

Causal impact = ITT / [E(X—Z=1) - E(X—Z=0)] = (130-100) / (1 - 0) = 30

Causal impact = ITT / [E(X—Z=1) - E(X—Z=0)] = (130-110) / (1 - 0.2) = 25

Z=0 (no scholarship)	**Z=1** (scholarship)
80 High school 20 College	0 High school 100 College
Average wage: 110	Average wage: 130

ATE = (122 - 110) / (0.8 - 0.2) = 20

Z=0 (no scholarship)	**Z=1** (scholarship)
80 High school 20 College	20 High school 80 College
Average wage: 110	Average wage: 122

	Z=0 (no scholarship)	**Z=1** (scholarship)
HIGH SCHOOL PARENTS		
	80 High school	0 High school
	0 College	80 College
	Average wage: 100	Average wage: 120
COLLEGE PARENTS		
	0 High school	0 High school
	20 College	20 College
	Average wage: 125	Average wage: 125

Z=0 (no scholarship)	**Z=1** (scholarship)
HIGH SCHOOL PARENTS	
80 High school	0 High school
0 College	80 College
COLLEGE PARENTS	
0 High school	0 High school
20 College	20 College
Average wage: 105	Average wage: 121

Sum up

- When we have separate information for High school parents and College parents: High school parents: +20 effect College parents: not identified
- When we have global estimation stacking HS and C parents: +20 effect

College parents population does not seem to contribute

Sum up

There is no way we can learn something on the impact among college parents population because there is no experiment actually going on in that population

- In this example, all the reduced form effect comes from HS parents population: 121-105=16
- And they represent a change in college participation in 80% of the sample
- Thus, the effect $16/0.8 = 20$ results only from HS parents population

Let's call them **compliers** because they comply with the treatment assignment

Z=0 (no scholarship)	**Z=1** (scholarship)
High school 80 (HS parents)	College 80 (HS parents)
College 20 (College parents)	College 20 (College parents)
Average wage: 105	Average wage: 121

Impact is identified on the share of population who moves from HS to College

Now add Never takers:

Z=0 (no scholarship)	Z=1 (scholarship)
High school 10 (Never takers)	High school 10 (Never takers)
High school 80 (Compliers)	College 80 (Compliers)
College 10 (Always takers)	College 10 (Always takers)
Average wage: 105	Average wage: 121

All the change in the reduced form: 121-105 is due to compliers What is the share of compliers in the sample? 80% Thus impact: $16/0.8 = 20$

- **20 is the effect on the compliers** (should it be different for the other populations)
- Information: 90 HS, 10 College for $Z = 0$ and 10 HS, 90 College for $Z = 1$
- How do we know there are 80% compliers? Can we name them?

Now add Defiers:

Z=0 (no scholarship)	**Z=1** (scholarship)
High school 5 (Never takers)	High school 5 (Never takers)
High school 80 (Compliers)	College 80 (Compliers)
College 10 (Always takers)	College 10 (Always takers)
College 5 (Defiers)	High school 5 (Defiers)
Average wage: 105	Average wage: 121

Formalizing

$T(Z)$ is a random variable that assigns an individual response T to the value of the instrument Z

Every person may respond differently to the instrument

$$\begin{aligned}
\text{Compliers} \quad & T_i(0) = 0 \quad T_i(1) = 1 && (12) \\
\text{Never-takers} \quad & T_i(0) = 0 \quad T_i(1) = 0 && (13) \\
\text{Always-takers} \quad & T_i(0) = 1 \quad T_i(1) = 1 && (14) \\
\text{Defiers} \quad & T_i(0) = 1 \quad T_i(1) = 0 && (15)
\end{aligned}$$

Note: can generalize to more values of the instrument than just $(0, 1)$

Hypothesis 1 (Independence)

Z is independent from $(Y_0, Y_1, T(0), T(1))$

In particular implies that people with some sensitivity to the instrument (described by the set $\{T(0), T(1)\}$) are not more or less likely to draw a specific value of z

Hypothesis 2 (Monotonicity)

either $T_i(0) \geq T_i(1) \quad \forall i$ or $T_i(0) \leq T_i(1) \quad \forall i$

i.e.: all agents' response to the instrument is (weakly) in the same direction

For instance: a mother with one boy-one girl who has a third child would also have a third child if she had two boys (the effect of same-sex is never to reduce fertility)

Monotonicity is equivalent to the absence of defiers

Reduced form

$$
\begin{aligned}
E(Y|Z=1) &= E(Y_0 + T(Y_1 - Y_0)|Z=1) && (16)\\
&= E(Y_0 + T(1)(Y_1 - Y_0)) && (17)
\end{aligned}
$$

Thus

$$
\begin{aligned}
E(Y|Z=1) - E(Y|Z=0) &= E(Y_0 + T(1)(Y_1 - Y_0)) - E(Y_0 + T(0)(Y_1 - Y_0)) && (18)\\
&= E[(T(1) - T(0))(Y_1 - Y_0)] && (19)
\end{aligned}
$$

Reduced form

$$
\begin{aligned}
E[(T(1) - T(0))(Y_1 - Y_0)] = & \quad (20)\\
& E[(Y_1 - Y_0)|T(1) - T(0) = 1]P(T(1) - T(0) = 1) \quad (21)\\
& + E[0 \times (Y_1 - Y_0)|T(1) - T(0) = 0]P(T(1) - T(0) = 0) \quad (22)\\
& + E[-1 \times (Y_1 - Y_0)|T(1) - T(0) = -1]P(T(1) - T(0) = -1) \quad (23)
\end{aligned}
$$

$$
\begin{aligned}
& = E[(Y_1 - Y_0)|C]P(C) \quad (24)\\
& + E[0 \times (Y_1 - Y_0)|A \text{ or } N]P(A \text{ or } N) \quad (25)\\
& + E[-1 \times (Y_1 - Y_0)|D]P(D) \quad (26)
\end{aligned}
$$

Role of monotonicity

Assume $T(1) \geq T(0)$; then $T(1) - T(0) = -1$ is impossible; there are no defiers

Thus:

$$E(Y|Z=1) - E(Y|Z=0) = E[(Y_1 - Y_0)|T(1) - T(0) = 1]P(T(1) - T(0) = 1) \tag{27}$$

with

$$\begin{aligned} P(T(1) - T(0) = 1) &= E(T(1) - T(0)) && (28) \\ &= E(T|Z=1) - E(T|Z=0) && (29) \\ &= P(T=1|Z=1) - P(T=1|Z=0) && (30) \end{aligned}$$

LATE

Under hypothesis 1 (*Independence*) and 2 (*Monotonicity*), the Wald estimator is:

$$W = \frac{E(Y|Z=1) - E(Y|Z=0)}{P(T=1|Z=1) - P(T=1|Z=0)} \tag{31}$$

$$= E[(Y_1 - Y_0)|T(1) - T(0) = 1] = LATE \tag{32}$$

Local Average Treatment Effect: treatment effect on those that change their behavior (T) under the instrument (compliers)

LATE with more than 2 values

When instrument takes more than 2 values, $LATE_{Z_1,Z_2}$ can be defined for each pair of values of the instrument (Z_1, Z_2).

The IV estimator uses all values of Z at a time: can be interpreted as a weighted sum of the LATEs, where the weights depend on the local impact of the instrument

What about the ATE?

So we cannot use IVs to estimate the ATE if:

1. There is treatment heterogeneity ($E(Y_1 - Y_0)$ is not constant), and
2. **This heterogeneity is related to treatment behavior**:

$$E(Y_1 - Y_0) \neq E(Y_1 - Y_0 | \text{Compliers}) \tag{33}$$

This is called "essential heterogeneity".

In this case, $LATE \neq ATE$.

Implications

- IV has no clear interpretation if there is essential heterogeneity or if there are defiers
- Different instruments can identify different parameters because they estimate the impacts on different populations
- The gap between OLS and IV mix the result of bias reduction and change in the populations that contribute to the estimation

This is a **major reason** why IV estimations have fallen out of favor among economists, along with the difficulty of justifying the exclusion restriction

Thinking about the LATE: examples

1. Scholarship $\rightarrow$ secondary education $\rightarrow$ wage at 25
2. Vietnam draft $\rightarrow$ military service $\rightarrow$ death
3. Rainfall shocks $\rightarrow$ household agricultural income $\rightarrow$ civil conflict

Outline

- So far we have been thinking about **identification**, but less about **estimation**
- Now let's discuss the main issue regarding estimation of IVs - **weak instruments**

An instrument is said to be **weak** if it explains little of the endogenous variable

Weak instruments

$$Y_i = T_i\beta + \epsilon_i$$
$$T_i = Z_i\pi_1 + u_i$$

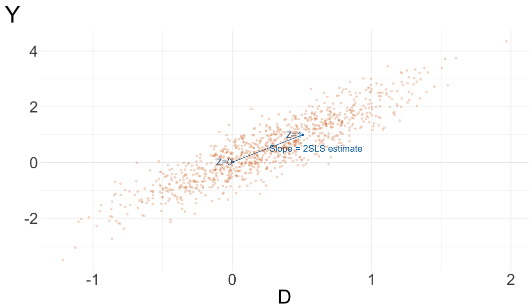
- Recall that one of the key assumptions for our estimation procedure was relevance
  - $\pi_1 \neq 0$, or $Cov(Z_i, T_i) \neq 0$
- Why is this necessary? Consider the 2SLS estimator for β_{IV} in the simplest case:

$$\hat{\beta} = \frac{Cov(Y_i, Z_i)}{Cov(D_i, Z_i)}$$

- If $Cov(D_i, Z_i) = 0$, this estimate is obviously undefined! But what about if it's very small?
  - Small variations in it will move around $\hat{\beta}$ in a big way. That's what statistical uncertainty will do
  - One easy way to see this: graphically

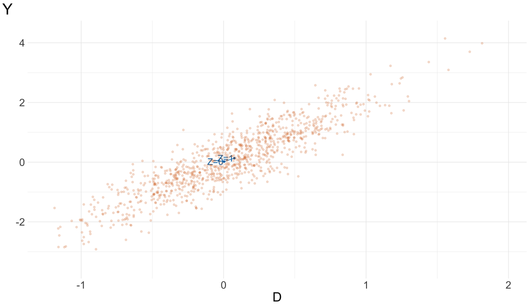
Weak instruments

- Simple 2SLS simulation, with binary instrument
  - First stage coef = 0.5, true beta = 2
- Note that the estimation on the x-axis comes from variation in the first stage
- The larger this is, the stronger the first stage
- However, if the first stage is weak, this interval is quite short, even if the variation in D stays the same



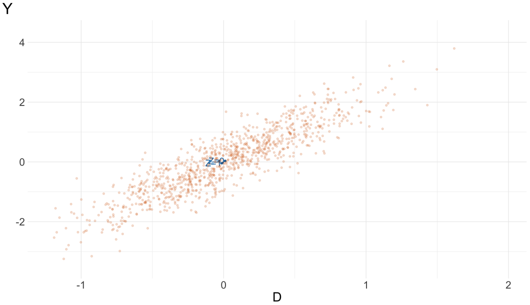
Weak instruments

- With a first stage coefficient of 0.1, it becomes hard to distinguish the points
  - Note: I hold fixed the overall variance of D here to keep the correct comparison!
- Given that the model is correctly specified, with enough data it should converge to the right β
- But small shifts in the x-axis will massively swing the estimate!



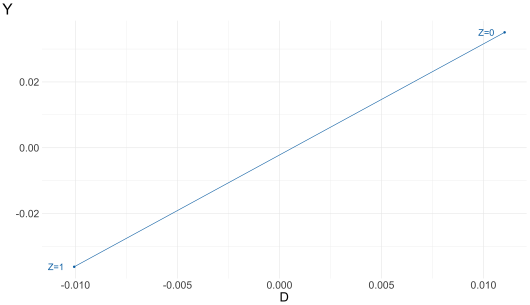
Weak instruments

- With a first stage coefficient of 0.01, the problem is even worse



Weak instruments

- With a first stage coefficient of 0.01, the problem is even worse
- We see that the relevant variation being exploited is tiny
- A small change in the x-axis points would even flip the sign!
- What does that do to our estimation procedure?



Examples of weak instruments?

1. Birth month $(Z) \to$ years of schooling $(T) \to$ adult wages (Y)
   [Angrist & Krueger 1991, later critiqued by Bound, Jaeger & Baker 1995]
2. Colonial settler mortality $(Z) \to$ institutional quality $(T) \to$ modern-day GDP (Y)
   [Acemoglu, Johnson, Robinson 2001]

What concretely happens if we have a weak instrument?

1. **Loss of precision**
2. **Bias in finite samples**

1. Loss of precision

Recall expression for variance of an IV estimator in the simplest case:

$$Var(\hat{\beta}_{IV}) = \frac{V(u)}{N} \cdot \frac{1}{Var(T)} \cdot \frac{1}{\pi_1^2}$$

when $Y = \alpha + \beta T + u$ and $T = \pi_1 Z + v$

So as π_1 gets smaller, $Var(\hat{\beta}_{IV})$ increases more than linearly

2. Bias in finite samples

Even though an IV estimator is consistent, it is still biased in finite samples.

2. Bias in finite samples

Even though an IV estimator is consistent, it is still biased in finite samples.

- **Consistency**: $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$
- **Biased**: $\mathbb{E}[\hat{\beta}_{2SLS}] \neq \beta$

The bias is towards β_{OLS}

$$T = Z\pi + v \tag{34}$$

We want to replace T with what is in the 2nd stage. We need to estimate π using $\hat{\pi}$

- We would require $\hat{T} = Z\pi$
- But in finite sample $\hat{\pi} \neq \pi$ so $\hat{T} \neq Z\pi$
- The least square criteria to estimate $\hat{\pi}$ ”get $\hat{T}$ close to T”
- The mistake is towards ”$\hat{T}$ looks like T too much”: ”overfit”
- So $\hat{\beta}_{2SLS}$ looks too much like $\hat{\beta}_{OLS}$

Determinants of bias

Expression for bias of IV estimator:

$$E(\hat{\beta}_{2SLS}) - \beta \approx \frac{\text{cov}(\varepsilon, v)}{\sigma_v^2} \left[\frac{1}{1+F} \right] \tag{35}$$

where F is an F-test statistic of the regression of T on Z, i.e.,

$$F = \frac{R^2_{T,Z}/K}{(1 - R^2_{T,Z})/(N-K)} \tag{36}$$

where K is the number of instruments (usually $K = 1$), $R^2_{T,Z}$ is the R^2 in the regression of T and Z

Determinants of bias

$$E(\hat{\beta}_{2SLS}) - \beta \approx \frac{\text{cov}(\varepsilon, v)}{\sigma_v^2}\left[\frac{1}{1+F}\right] \tag{37}$$

$$F = \frac{R^2_{T,Z}/K}{(1-R^2_{T,Z})/(N-K)} \tag{38}$$

- Correlation between ε and v (source of bias)
- F (measure of weak instruments), mostly driven by how much the instruments explain T ($R^2_{T,Z}$) (weak instrument when R^2 is small)

Determinants of bias

$$E(\hat{\beta}_{2SLS}) - \beta \approx \frac{\text{cov}(\varepsilon, v)}{\sigma_v^2} \left[\frac{1}{1+F} \right] \tag{39}$$

$$F = \frac{R^2_{T,Z}/K}{(1 - R^2_{T,Z})/(N-K)} \tag{40}$$

Implications:

- If $R^2_{T,z}$ is small enough, even large n cannot impede strong bias
- Adding instruments is a bad idea if instruments are weak (increase K but hardly increases $R^2_{T,Z}$)

Testing for weak instruments

The F-stat for $\pi = 0$ (significance of excluded instruments in first step) is proportional to the bias

But also depends on other parameters K, N and cov(ε, v)

Stock & Yogo (2005) derive formal tests: Roughly, if $F > 10$, reject that the 2SLS bias will be more than 10% of the OLS bias

Summing up

We covered the basics of IVs. They are a way of estimating causal effects that don't rely on an experimenter randomly allocating treatments.

But they come with a number of very important challenges:

1. Justifying the **exclusion restriction**
2. Understanding what the **LATE** is really measuring
3. Dealing with **weak instruments**