Lecture 08: Instrumental variables I

PPHA 34600

Prof. Fiona Burlig

Harris School of Public Policy
University of Chicago

From last time: selection on observables

We tried our first non-experimental estimators:

- SOO requires **extremely strong** assumptions (☠)
- ...but when you're stuck in this world, you can use:
  1. Regression adjustment
     - Controlling for stuff
  2. Matching
     - Pairing treated and untreated on observables

→ We typically **don't** believe SOO designs

Moving on from SOO

SOO requires strong assumptions:

- "I know everything about everything"
- $(Y_i(1), Y_i(0)) \perp D_i | X_i$
- Need to isolate **all** "bad" variation in D_i

We move to selection on unobservables instead:

- "I know a little bit about a little bit"
- Typically represented as $(Y_i(1), Y_i(0)) \perp Z_i$ and $Cov(Z_i, D_i) \neq 0$
- Need to find some source of "good' variation in D_i

Where does this "good variation" come from?

We turn to natural experiments:

- Rather than needing to observe everything...
- ... we observe some (quasi) random variation in D_i
- Natural experiments are "naturally" occurring random variation
  - Can come from actual nature (e.g. variation in temperature / rainfall)
  - But more commonly are policy-induced

Where does this “good variation” come from?

We turn to natural experiments:

- Rather than needing to observe everything...
- ... we observe some (quasi) random variation in D_i
- Natural experiments are “naturally” occurring random variation
  - Can come from actual nature (e.g. variation in temperature / rainfall)
  - But more commonly are policy-induced

Some new definitions:

- **Observational data:** Data generated from non-experimental settings
- **Identification strategy:** approach to using observational data to estimate causal effects
- **Identifying assumptions:** Assumptions required for the identification strategy to causally estimate impacts

Isolating good variation

Consider the following regression model:

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

where:

Y_i is our outcome of interest

D_i is treatment

X_i is a set of covariates, where $Cov(X_i, \varepsilon_i) = 0$

ε_i is the error

$\rightarrow$ What formal condition do we need to recover the causal effect of D_i?

Isolating good variation

Consider the following regression model:

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

where:
Y_i is our outcome of interest
D_i is treatment
X_i is a set of covariates, where $Cov(X_i, \varepsilon_i) = 0$
ε_i is the error

→ What formal condition do we need to recover the causal effect of D_i?

→ We need $E[\varepsilon_i | D_i] = 0 \iff Cov(D_i, \varepsilon_i) = 0$

What happens when $E[\varepsilon_i|D_i] \neq 0$?

In this case, D_i is endogenous:

$\rightarrow$ We cannot get an unbiased estimate of τ^{ATE}

- This can result from:
  - Omitted variable bias
  - Reverse casuality (simultaneity)

Isolating exognenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i into two parts:

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0)$

Isolating exognenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i into two parts:

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0)$

Then we can write:

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

Isolating exognenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i into two parts:

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0$)

Then we can write:

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

$$Y_i = \underbrace{\alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\varepsilon_i}_{\text{rearranging}}$$

Isolating exognenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i into two parts:

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0$)

Then we can write:

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

$$Y_i = \underbrace{\alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\varepsilon_i}_{\text{rearranging}}$$

$\rightarrow$ If we observed C_i, we could regress Y_i on C_i and recover τ!

Isolating exognenous variation

$$Y_i = \alpha + \tau D_i + \beta X_i + \varepsilon_i$$

Suppose we can separate D_i into two parts:

$$D_i = B_i \varepsilon_i + C_i$$

with $Cov(C_i, \varepsilon_i) = 0)$

Then we can write:

$$Y_i = \alpha + \tau(B_i \varepsilon_i + C_i) + \beta X_i + \varepsilon_i$$

$$Y_i = \underbrace{\alpha + \tau C_i + \beta X_i + (1 + \tau B_i)\varepsilon_i}_{\text{rearranging}}$$

$\rightarrow$ If we observed C_i, we could regress Y_i on C_i and recover τ!

$\rightarrow$ ...but we can't

Instrumental variables to the rescue!

We want to find an instrumental variable:

Intuitively, an IV generates variation in C_i but is uncorrelated with ε_i

Z_i is a valid instrument for D_i when the following are satisfied:

Instrumental variables to the rescue!

We want to find an instrumental variable:

Intuitively, an IV generates variation in C_i but is uncorrelated with ε_i

Z_i is a valid instrument for D_i when the following are satisfied:

1. **First stage**: $Cov(Z_i, D_i) \neq 0$
    - Z_i and D_i are related
    - Without this, you're capturing nothing
    - This is actually testable!

Instrumental variables to the rescue!

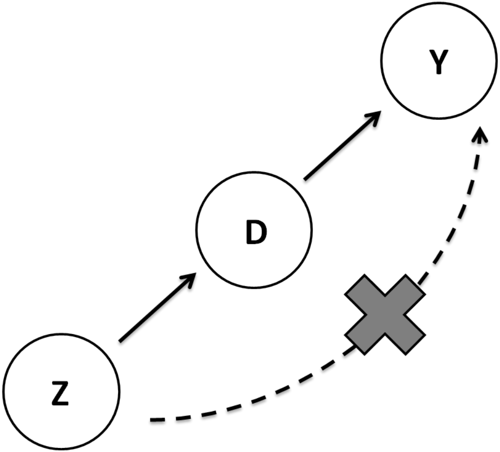
We want to find an instrumental variable:

Intuitively, an IV generates variation in C_i but is uncorrelated with ε_i

Z_i is a valid instrument for D_i when the following are satisfied:

1. **First stage**: $Cov(Z_i, D_i) \neq 0$
   - Z_i and D_i are related
   - Without this, you're capturing nothing
   - This is actually testable!
2. **Exclusion restriction**: $Cov(Z_i, \varepsilon_i) = 0$
   - Z_i and ε_i are **not** related
   - Z_i only affects Y_i through D_i
   - Fundamentally untestable! ☠

The exclusion restriction



Z_i affects Y_i only through D_i: Z_i cannot affect Y_i through any other channel, and therefore can be "excluded" from a regression of Y_i on D_i

Estimation of IV

Strictly speaking, the IV estimator is:

$$\hat{\tau}^{IV} = (Z'D)^{-1}(Z'Y)$$

Since we don't do matrix algebra in this class, without covariates, we can write:

$$\hat{\tau}^{IV} = Cov(Z_i, Y_i)/Cov(Z_i, D_i)$$

It's much easier to think through IV through its mechanics

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS, in several steps:

1. **First stage:**

   Regress endogenous D_i on all exogenous variables (including Z_i)

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

→ Store the predicted values of D_i, $\hat{D}_i$

→ Note: this provides a test of our first assumption

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS, in several steps:

1. **First stage:**

   Regress endogenous D_i on all exogenous variables (including Z_i)

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

→ Store the predicted values of D_i, $\hat{D}_i$

→ Note: this provides a test of our first assumption

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation...

- Estimates the effect of our instrument on treatment

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation...

- Estimates the effect of our instrument on treatment
- **This is isolating the "good variation" in D_i**
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - ...so a regression of D_i on Z_i only keeps the good stuff
  - When we extract $\hat{D}_i$, we've "eliminated" the "bad variation"

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation...

- Estimates the effect of our instrument on treatment
- **This is isolating the "good variation" in** D_i
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - ...so a regression of D_i on Z_i only keeps the good stuff
  - When we extract $\hat{D}_i$, we've "eliminated" the "bad variation"
- Lets us test whether $Cov(Z_i, D_i \neq 0)$:
  - → The proper way to do this is with an F test
  - → Rule of thumb: need F statistics > 20

The first stage

$$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

This equation...

- Estimates the effect of our instrument on treatment
- **This is isolating the "good variation" in** D_i
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - ...so a regression of D_i on Z_i only keeps the good stuff
  - When we extract $\hat{D}_i$, we've "eliminated" the "bad variation"
- Lets us test whether $Cov(Z_i, D_i \neq 0)$:
  - → The proper way to do this is with an F test
  - → Rule of thumb: need F statistics > 20
- Should have an intuitive sign

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS, in several steps:

1. **First stage:**

   Regress endogenous D_i on all exogenous variables (including Z_i)

   $$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

   → Store the predicted values of D_i, $\hat{D}_i$

   → Note: this provides a test of our first assumption

2. **Second stage:**

   Regress outcome Y_i on predicted $\hat{D}_i$ and other Xs:

   $$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

   → $\hat{\tau}$ in this equation is our IV estimate

   → Note: The standard errors on this will be wrong! Use a canned routine

The second stage

$$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

This equation...

- Estimates the effect of our treatment our outcomes

The second stage

$$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

This equation...

- Estimates the effect of our treatment our outcomes
- **This only uses "good variation" in D_i**
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - ...so a regression of Z_i on $\hat{D}_i$ only uses the good stuff
- **Central assumption of IV:**

→ Z_i is as good as randomly assigned!

The second stage

$$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

This equation...

- Estimates the effect of our treatment our outcomes
- **This only uses "good variation" in D_i**
  - By assumption, Z_i is unrelated to Y_i except through D_i
  - ...so a regression of Z_i on $\hat{D}_i$ only uses the good stuff
- **Central assumption of IV:**

$\rightarrow$ Z_i is as good as randomly assigned!

- You should be able to talk about the difference between this and the OLS

Two stage least squares (2SLS)

The classic way to perform IV is via 2SLS, in several steps:

1. **First stage:**

   Regress endogenous D_i on all exogenous variables (including Z_i)

   $$D_i = \alpha + \gamma Z_i + \beta X_i + \eta_i$$

   → Store the predicted values of D_i, $\hat{D}_i$

   → Note: this provides a test of our first assumption

2. **Second stage:**

   Regress outcome Y_i on predicted $\hat{D}_i$ and other Xs:

   $$Y_i = \alpha + \tau \hat{D}_i + \delta X_i + \varepsilon_i$$

   → $\hat{\tau}$ in this equation is our IV estimate

   → Note: The standard errors on this will be wrong! Use a canned routine

The reduced form

Good IV studies will also report the reduced form:

- This is a regression of Y_i on your instrument Z_i:

$$Y_i = \alpha + \theta Z_i + \pi X_i + \eta_i$$

The reduced form

Good IV studies will also report the reduced form:

- This is a regression of Y_i on your instrument Z_i:

$$Y_i = \alpha + \theta Z_i + \pi X_i + \eta_i$$

- This does **not** recover $\hat{\tau}^{IV}$
- Instead, this tells us how your outcome varies with the instrument
- This needs to have a causal interpretation!
  - If you don't believe that your Z_i is "as good as random" wrt Y_i, it definitely can't help you with D_i
  - In almost all cases, the RF should be independently interesting

Alternative ways to estimate τ^{IV}

We've estimated three total coefficients:

1. **First stage:** $\hat{\gamma}$ is the effect of our instrument on treatment
2. **Reduced form:** $\hat{\theta}$ is the effect of our instrument on our outcome
3. **Second stage:** $\hat{\tau}^{IV}$ is the effect of treatment on our outcome

Alternative ways to estimate τ^{IV}

We've estimated three total coefficients:

1. **First stage:** $\hat{\gamma}$ is the effect of our instrument on treatment
2. **Reduced form:** $\hat{\theta}$ is the effect of our instrument on our outcome
3. **Second stage:** $\hat{\tau}^{IV}$ is the effect of treatment on our outcome

→ Very Smart People realized that we can just compute:

$$\hat{\tau}^{IV} = \frac{\hat{\theta}}{\hat{\gamma}}$$

- This should be intuitive:
  - We're just *scaling* the reduced form by the first stage
  - Our IV estimate is just the effect of the instrument on our outcome, weighted by how much the instrument moves treatment
- Standard errors are still tricky (use that canned routine)

The exclusion restriction is the key to any IV

You should always ask:
What is the exclusion restriction in this analysis saying?

The exclusion restriction is the key to any IV

You should always ask:
What is the exclusion restriction in this analysis saying?

Do we believe this? Why or why not?

A few random IV details

When doing IV, you should know:

- Your standard errors will be bigger than OLS standard errors
  - → IV is "throwing out" variation!
  - → You need a canned routine to get them right in any two-step approach

A few random IV details

When doing IV, you should know:

- Your standard errors will be bigger than OLS standard errors
  - → IV is "throwing out" variation!
  - → You need a canned routine to get them right in any two-step approach
- Always, always, always include the same covariates X_i in the first stage, second stage, and reduced form
  - → Your IV is impossible to interpret if you don't do this
  - → (The real reason why is Linear Algebra)

A few random IV details

When doing IV, you should know:

- Your standard errors will be bigger than OLS standard errors
  - → IV is "throwing out" variation!
  - → You need a canned routine to get them right in any two-step approach
- Always, always, always include the same covariates X_i in the first stage, second stage, and reduced form
  - → Your IV is impossible to interpret if you don't do this
  - → (The real reason why is Linear Algebra)
- You might encounter the "forbidden regression"
  - → You need to use OLS for the first and second stage and RF
  - → No non-linear estimators!
  - → ($D_i = \alpha + \delta Z_i + \beta_1 X + \beta_2 X^2 + \varepsilon_i$ is ok)
  - → ☠

A few random IV details

When doing IV, you should know:

- Your standard errors will be bigger than OLS standard errors
  - → IV is "throwing out" variation!
  - → You need a canned routine to get them right in any two-step approach
- Always, always, always include the same covariates X_i in the first stage, second stage, and reduced form
  - → Your IV is impossible to interpret if you don't do this
  - → (The real reason why is Linear Algebra)
- You might encounter the "forbidden regression"
  - → You need to use OLS for the first and second stage and RF
  - → No non-linear estimators!
  - → ($D_i = \alpha + \delta Z_i + \beta_1 X + \beta_2 X^2 + \varepsilon_i$ is ok)
  - → ☠
- **The exclusion restriction is fundamentally untestable**
  - → ☠☠☠

Recap

TL;DR:

1. Instrumental variables are very powerful
2. ...but they require extremely strong assumptions!
3. Hashtag no free lunch