Instrumental Variables

EC 425/525, Set 8

Edward Rubin

08 May 2019

Prologue

Schedule

Last time

Mathching and propensity-score methods

- Conditional independence
- Overlap

Today

Instrumental variables (and two-stage least squares)

Upcoming

- Admin: Assignment/project proposal this weekend
- Admin: Midterm very soon

Research designs

Selection on observables and/or unobservables

We've been focusing on ***selection-on-observables* designs**, *i.e.*,

$$(\mathrm{Y}_{0i},\, \mathrm{Y}_{1i}) \perp\!\!\!\perp \mathrm{D}_i | \mathrm{X}_i$$

for **observable** variables X_i.

***Selection-on-unobservable* designs** replace this assumption with two new (but related) assumptions

1. $(\mathrm{Y}_{0i},\, \mathrm{Y}_{1i}) \perp \mathrm{Z}_i$
2. $\mathrm{Cov}(\mathrm{Z}_i,\, \mathrm{D}_i) \neq 0$

Selection on observables and/or unobservables

Our main goal in causal-inference minded (applied) econometrics boils down to isolating **"good" variation** in $\mathbf{D}_i$ (exogenous/as-good-as-random) from **"bad" variation** (the part of $\mathbf{D}_i$ correlated with $\mathbf{Y}_{0i}$ and $\mathbf{Y}_{1i}$).

(We want to avoid selection bias.)

- **Selection-on-observables designs** assume that we can control for all *bad variation* (selection) in $\mathbf{D}_i$ through a known (observed) $\mathbf{X}_i$.
- **Selection-on-unobservables designs** assume that we can extract part of the *good variation* in $\mathbf{D}_i$ (generally using some $\mathbf{Z}_i$) and then use this *good* part of $\mathbf{D}_i$ to estimate the effect of $\mathbf{D}_i$ on $\mathbf{Y}_i$. We throw away the *bad variation* in $\mathbf{D}_i$ (it's bad).

Which route?

So set of research designs is more palatable?

1. There are plenty of bad applications of both sets.
   Violated assumptions, bad controls, *etc.*

1. **Selection on observables** assumes we know *everything* about selection into treatment—we can identify *all* of the good (or bad) variation in $\mathbf{D}_i$.
   Tough in non-experimental settings. Difficult to validate in practice.

1. **Selection on unobservables** assumes we can isolate *some* good/clean variation in $\mathbf{D}_i$, which we then use to estimate the effect of $\mathbf{D}_i$ on $\mathbf{Y}_i$.
   Seems more plausible. Possible to validate. May be underpowered.

Instrumental variables

Introduction

Instrumental variables (IV)[†] is the canonical selection-on-unobservables design—isolating *good variation* in $\mathbf{D}_i$ via some magical instrument $\mathbf{Z}_i$.

Consider some model (structural equation)

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{D}_i + \varepsilon_i \tag{1}$$

To guarantee consistent OLS estimates for β_1, want $\mathbf{Cov}(\mathbf{D}_i,\, \varepsilon_i) = \mathbf{0}$.
In general, this is a heroic assumption.

Alternative: Estimate β_1 via instrumental variables.

† For the moment, we're lumping together IV and two-stage least squares (2SLS) together—as many people do—even though they are technically different.

Instrumental variables

Definition

For our model

$$\text{Y}_i = \beta_0 + \beta_1 \text{D}_i + \varepsilon_i \tag{1}$$

A valid **instrument** is a variable Z_i such that

1. $\textbf{Cov}(\text{Z}_i,\ \textbf{D}_i) \neq \textbf{0}$
   our instrument correlates with treatment (so we can keep part of $\textbf{D}_i$)

1. $\textbf{Cov}(\text{Z}_i,\ \varepsilon_i) = \textbf{0}$
   our instrument is uncorrelated with other (non-$\textbf{D}_i$) determinants of $\textbf{Y}_i$,
   i.e., Z_i is excludable from equation (1). **(exclusion restriction)**

Example

Back to the returns to a college degree,

$$\text{Income}_i = \beta_0 + \beta_1 \text{Grad}_i + \varepsilon_i$$

OLS is likely biased.

What if that state conducts a (random) **lottery** for scholarships?

Let Lottery_i denote an indicator for whether i won a lottery scholarship.[†]

1. $\text{Cov}(\text{Lottery}_i, \text{Grad}_i) \neq 0$ (> 0) if scholarships increase grad. rates.

1. $\text{Cov}(\text{Lottery}_i, \varepsilon_i) = 0$ since the lottery is randomized.

† We'll have to focus on families who were eligible/who applied.

The IV estimator

The IV estimator for our model

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{D}_i + \varepsilon_i \tag{1}$$

with (valid) instrument Z_i is

$$\hat{\beta}_{\mathrm{IV}} = \left(\mathrm{Z}'\mathrm{D}\right)^{-1}\left(\mathrm{Z}'\mathrm{Y}\right)$$

If you have no covariates, then

$$\hat{\beta}_{\mathrm{IV}} = \frac{\mathrm{Cov}(\mathrm{Z}_i,\ \mathrm{Y}_i)}{\mathrm{Cov}(\mathrm{Z}_i,\ \mathrm{D}_i)}$$

The IV estimator

The IV estimator for our model

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{D}_i + \varepsilon_i \tag{1}$$

with (valid) instrument Z_i is

$$\hat{\beta}_{\mathrm{IV}} = \left(\mathrm{Z}'\mathrm{D}\right)^{-1}\left(\mathrm{Z}'\mathrm{Y}\right)$$

If you have additional (exogenous) covariates $\mathbf{X}_i$, then

$$\mathbf{Z} = \begin{bmatrix} \mathrm{Z}_i & \mathbf{X}_i \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} \mathrm{D}_i & \mathbf{X}_i \end{bmatrix}$$

Proof: Consistency

With a valid instrument $\mathbf{Z}_i$, $\hat{\beta}_{\text{IV}}$ is a consistent estiamtor for β_1 in

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_i + \varepsilon_i \tag{1}$$

$$\begin{aligned}
&\text{plim}\left(\hat{\beta}_{IV}\right) \\
&\quad = \text{plim}\left(\left(\text{Z}'\text{D}\right)^{-1}\left(\text{Z}'\text{Y}\right)\right) \\
&\quad = \text{plim}\left(\left(\text{Z}'\text{D}\right)^{-1}\left(\text{Z}'\text{D}\beta + \text{Z}'\varepsilon\right)\right) \\
&\quad = \text{plim}\left(\left(\text{Z}'\text{D}\right)^{-1}\left(\text{Z}'\text{D}\right)\beta\right) + \text{plim}\left(\frac{1}{N}\text{Z}'\text{D}\right)^{-1}\text{plim}\left(\frac{1}{N}\text{Z}'\varepsilon\right) \\
&\quad = \beta \quad ✔
\end{aligned}$$

Two-stage least squares

Two-stage least squares

Setup

You'll commonly see IV implemented as a two-stage process known as **two-stage least squares** (2SLS).

First stage Estimate the effect of the instrument Z_i on our endogenous variable $\mathbf{D}_i$ and (predetermined) covariates $\mathbf{X}_i$. Save $\widehat{\mathrm{D}}_i$.

$$\mathbf{D}_i = \gamma_1 \mathrm{Z}_i + \gamma_2 \mathbf{X}_i + u_i$$

Second stage Estimate model we wanted—but only using the variation in $\mathbf{D}_i$ that correlates with Z_i, *i.e.*, $\widehat{\mathrm{D}}_i$.

$$\mathbf{Y}_i = \beta_1 \widehat{\mathrm{D}}_i + \beta_2 \mathbf{X}_i + \varepsilon_i$$

Note The controls $\mathbf{X}_i$ must match in the first and second stages.

Two-stage least squares

IV estimation

This two-step procedure, with a valid instrument, produces an estimator $\hat{\beta}_1$ that is consistent for β_1.

$$\hat{\beta}_{\text{2SLS}} = \left(\text{D}'\text{P}_\text{Z}\text{D}\right)^{-1} \left(\text{D}'\text{P}_\text{Z}\text{Y}\right)$$

$$\text{P}_\text{Z} = \text{Z}\left(\text{Z}'\text{Z}\right)^{-1}\text{Z}'$$

where $\mathbf{D}$ is a matrix of our treatment and predetermined covariates $(\mathbf{X}_i)$ and Z is a matrix of our instrument and our predetermined covariates.

Two-stage least squares

IV estimation

Important notes

- The controls $(\mathbf{X}_i)$ must match in the first and second stages.
- If you have exactly **one instrument** and exactly **one endogenous variable**, then 2SLS and IV are identical.
- Your second-stage standard errors are not correct.

Table of contents