Ec1123

Section 8

Andrea Passalacqua

andreapassalacqua@g.harvard.edu

November 8th 2017

Table of Contents

Threats to External Validity

External Validity

Our estimates are externally valid if inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

Potential Issues:

- Nonrepresentative sample
- Nonrepresentative program or policy (e.g., duration and scale)
- Other factors may not be held constant in other settings. General equilibrium effects may make experimental estimates not useful for policy guidance.

Threats to Internal Validity

Internal Validity

Our estimates are internally valid if statistical inferences about causal effects are valid for the population being studied.

Potential Issues:

- Omitted variable bias
- Simultaneous causality bias
- Measurement Error (Errors-in-variables bias)
- Sample selection bias
- Wrong functional form

Threats to Internal Validity in IV Regressions

If the instruments are valid, IV takes care of

- Omitted variable bias
- Simultaneous causality
- Measurement error

Instead, have to worry about whether the instruments are valid:

IV conditions

1. **Relevance:** $\text{Corr}(Z, X) \neq 0$
2. **Exogeneity:** $\text{Corr}(Z, u) = 0$

Treatment Effects

We are interested in causal effects. Put differently, we are interested in estimating the **treatment effect** of X on Y

Treatment Effect

the causal impact on Y of switching from $X = 0$ to $X = 1$

Ex: What is the treatment effect on health of receiving the drug?

Ex: What is the treatment effect on smoking rates of a ban on bar smoking?

Heterogeneous Treatment Effects

Typically, treatment effects vary across entities

Ex: The effect of attending college on earnings differs across students

Ex: The effect of a state-wide smoking ban on smoking rates varies across states

$$Y_i = \beta_{0,i} + \beta_{1,i} X_i + u_i$$

Mathematically: $\beta_{1,i}$ differs across different i (e.g.students or states)

Thus far, we've been discussing **Average Treatment Effects**

Average Treatment Effect

If **conditional mean independence (CMI) is satisfied** (i.e. $E(u|X,W) = E(u|W)$, OLS estimates the ATE:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

so

$$\begin{aligned} \beta_1 &= \text{ATE} \\ &= \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] \\ &= \mathbb{E}[\beta_{1,i}] = \text{Average effect of a unit change in } X \end{aligned}$$

However, Instrument Variable regression generally does **NOT** estimate the ATE

Local Average Treatment Effect (LATE)

IV estimates the **Local Average Treatment Effect** (LATE)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Mathematically:

$$\beta_{1,IV} = \text{LATE} = \frac{\mathbb{E}[\beta_{1,i} \times \Pi_{1,i}]}{\mathbb{E}[\Pi_{1,i}]}$$

where $\beta_{1,i}$ is the true treatment effect of X on individual i, and $\Pi_{1,i}$ is the first-stage relationship for agent i

Two Stage Least Squares (2SLS) estimates this LATE

```
ivregress 2sls y w (x=z), robust
```

LATE – Intuition

$$\textbf{First-Stage} \qquad X_i = \Pi_{0,i} + \Pi_{1,i} Z_i + v_i$$

$$\textbf{Second-Stage} \qquad Y_i = \beta_{0,i} + \beta_{1,i} \widehat{X}_i + u_i$$

LATE is the average treatment effect for entities affected by the instrument (i.e. for whom $\Pi_{1,i} \neq 0$)

The word *local* indicates the LATE is the average for this affected group known as **compliers**. Compliers are those affected by the instrument (i.e. they **complied** with Z)

LATE – Example 1

We are investigating the causal impact of studying on grades.

$$\text{GPA}_i = \beta_{0,i} + \beta_{1,i}\text{Hours Studied}_i + u_i$$

Suppose the dataset has **75% industrious ants** and **25% slacker grasshoppers**, who respond differently to both X and Z

Let $Z =$ whether the roommate brought a video game to school

We'll study two cases:

- Suppose ants do not respond to the instrument Z at all
- Suppose ants do react to Z but not as much as grasshoppers

Average Treatment Effect

		Ants	Grasshoppers
β_{1i}	Δ GPA from $+1$ hr studying	0.5	1
	Sample %	75%	25%

What is the **average treatment effect**?

Average Treatment Effect

		Ants	Grasshoppers
β_{1i}	Δ GPA from +1 hr studying	0.5	1
	Sample %	75%	25%

What is the **average treatment effect**?

$$\textbf{ATE} = (75\%) \times 0.5 + (25\%) \times 1 = 0.625$$

However, OLS won't identify this ATE, because it suffers from omitted variable bias

LATE – Example 1

Suppose we use our instrument Z = roommate with a video game

		Ants	**Grasshoppers**
β_{1i}	Δ GPA from $+1$ hr studying	0.5	1
Π_{1i}	Δ Hours Studied b/c roommate w/ video game	0 hr	-0.8 hr

Without doing any math, you should know what the LATE is

$$\text{LATE} = \beta_{IV} = \frac{\mathbb{E}[\beta_{1,i} \times \Pi_{1,i}]}{\mathbb{E}[\Pi_{1,i}]} = ?$$

LATE – Example 1

		Ants	Grasshoppers
β_{1i}	Δ GPA from $+1$ hr studying	0.5	1
Π_{1i}	Δ Hours Studied b/c roommate w/ video game	0 hr	-0.8 hr

$$\widehat{\beta}_{IV} = \frac{\mathbb{E}[\beta_{1,i} \times \Pi_{1,i}]}{\mathbb{E}[\Pi_{1,i}]} = \frac{75\% \times (0.5 \times 0) + 25\% \times (1 \times -0.8)}{75\% \times 0 + 25\% \times -0.8} = 1$$

Notice LATE $= \beta_{1,\text{grasshopper}}$ since grasshoppers are the only compliers

LATE – Example 2

Now suppose that ants too respond to Z but less than grasshoppers

		Ants	**Grasshoppers**
β_{1i}	Δ GPA from $+1$ hr studying	0.5	1
Π_{1i}	Δ Hours Studied b/c roommate w/ video game	-0.2 hr	-0.8 hr

Notice that the ATE has not changed!

LATE – Example 2

		Ants	Grasshoppers
β_{1i}	Δ GPA from +1 hr studying	0.5	1
Π_{1i}	Δ Hours Studied b/c roommate w/ video game	-0.2 hr	-0.8 hr

Using the same Z, do we expect the new LATE to be different than the LATE in our previous case? Will it be greater or smaller than 1?

LATE – Example 2

		Ants	Grasshoppers
β_{1i}	Δ GPA from $+1$ hr studying	0.5	1
Π_{1i}	Δ Hours Studied b/c roommate w/ video game	-0.2 hr	-0.8 hr

$$\widehat{\beta}_{IV} = \frac{\mathbb{E}[\beta_{1,i} \times \Pi_{1,i}]}{\mathbb{E}[\Pi_{1,i}]} = \frac{75\%(0.5 \times -0.2) + 25\%(1 \times -0.8)}{75\% \times -0.2 + 25\% \times -0.8} = 0.786$$

Notice β_{IV} is a weighted average of $\beta_{1,\text{ants}}$ and $\beta_{1,\text{grasshoppers}}$

LATE – Recap

Recall from last section that IV is identified off of the **"as-if random" variation in X induced by Z**

So intuitively, $\widehat{\beta}_{IV}$ only captures the causal effect of X on Y for **compliers** whose X vary by Z

$\widehat{\beta}_{IV}$ is a weighted average of the treatment effect for compliers, with more weight given to more compliant groups (e.g. grasshoppers)

In this goofy example, we knew $\beta_{1,i}$ and $\Pi_{1,i}$. This is not usually the case especially with more heterogeneous populations. Hence, we rely on `ivregress 2sls` for estimation

More LATE

Suppose we are estimating the causal effect of X on Y. We have two valid instruments Z_1 and Z_2.

- We just use Z_1 and run 2SLS to estimate $\widehat{\beta}_{2SLS}$
- We just use Z_2 and run 2SLS to estimate $\tilde{\beta}_{2SLS}$

Should we expect our estimates to equal?

$$\widehat{\beta}_{2SLS} \stackrel{?}{=} \tilde{\beta}_{2SLS}$$

More LATE

Suppose we are estimating the causal effect of X on Y. We have two valid instruments Z_1 and Z_2.

- We just use Z_1 and run 2SLS to estimate $\widehat{\beta}_{2SLS}$
- We just use Z_2 and run 2SLS to estimate $\tilde{\beta}_{2SLS}$

Should we expect our estimates to equal?

$$\widehat{\beta}_{2SLS} \stackrel{?}{=} \tilde{\beta}_{2SLS}$$

No. Different groups may respond differently to the different instruments Z_1 and Z_2. Each instrument may have a different group of compliers and therefore different LATEs.

LATE – Mathematically

$$\text{LATE} = \frac{\mathbb{E}[\beta_{1,I} \times \Pi_{1,i}]}{\mathbb{E}[\Pi_{1,i}]} = \text{ATE} + \frac{\text{Cov}(\beta_{1,i},\, \Pi_{1,i})}{\mathbb{E}[\Pi_{1,i}]}$$

where

$$\begin{aligned} \text{Cov}(\beta_{1,i}, \Pi_{1,i}) &= \mathbb{E}\Big[\big(\beta_{1,i} - \mathbb{E}[\beta_{1,i}]\big)\big(\Pi_{1,i} - \mathbb{E}[\Pi_{1,i}]\big)\Big] \\ &= \mathbb{E}\big[\beta_{1,i}\Pi_{1,i}\big] - \mathbb{E}[\beta_{1,i}]\mathbb{E}[\Pi_{1,i}] \end{aligned}$$

ATE v. LATE

$$\beta_{1,i} = \text{ causal impact of } X \text{ on } Y \text{ for individual } i$$

$$\Pi_{1,i} = \text{ correlation between } X \text{ and } Z \text{ for individual } i$$

LATE = **ATE** if any of the following is true

- no heterogeneity in treatment effects

$$\beta_{1,i} = \beta_1 \text{ for all } i$$

- no heterogeneity in first-stage responses to the instrument Z

$$\Pi_{1,i} = \Pi_1 \text{ for all } i$$

- no correlation between response to instrument Z and response to treatment X

$$\text{Cov}(\beta_{1,i}, \Pi_{1,i}) = 0$$

ATE v. LATE

Which do we care about: ATE or LATE?

ATE v. LATE

Which do we care about: ATE or LATE?

Depends on the context.

- if proposed policy is to give everyone the treatment, then ATE
- if proposed policy only affects a subset, then maybe LATE is more appropriate