Causal Data analysis

Chapter 21: Instrumental Variables and Regression Discontinuity

Álmos Telegdy [1]

[1]Corvinus University of Budapest

Advanced methods

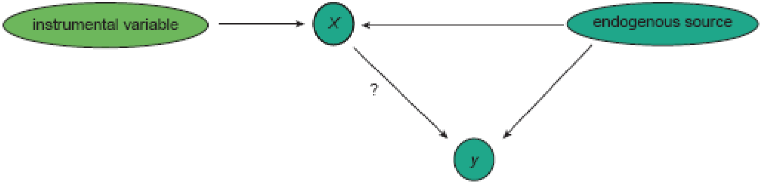
Known unknowns – what to do, if we cannot control for something important?

The data generating process can facilitate sometimes the estimation of a causal effect

- there is a variable in the data which is correlated with x but not with $y \Rightarrow$ Instrumental variables (IV)
- there is a discontinuity in the data $\Rightarrow$ Regression discontinuity design (RDD)

Instrumental variable

Sometimes there is a variable in the data, which *has an exogenous effect on* x (so it does not have an effect on y) $\longrightarrow$ **instrumental variable (IV)**



Properties of an instrumental variable

$$y^E = \alpha + \beta x \tag{1}$$

Problem: The estimated effect of x on y is biased

Properties of the instrument (z_{IV})

- The instrumental variable is correlated with $x \rightarrow cov(z_{IV}, x) \neq 0$
- The instrumental variable affects y ONLY through $x \rightarrow cov(z_{IV}, \epsilon) = 0$

Intuition: if we estimate a correlation between z_{IV} and y, that is because x and y **are related**

Example: effect of children on mothers' labor supply

Question: what is the effect of the number of children on mothers' labor supply?
Problem: the number of children is endogenous (why?)

Example: effect of children on mothers' labor supply

Question: what is the effect of the number of children on mothers' labor supply?
Problem: the number of children is endogenous (why?)
IV: first two children have the same gender

- $z_{IV} \rightarrow x$: families with two boys or girls are more likely to have a third child
- $z_{IV} \not\rightarrow y$: the gender of children is not likely to be related to the mother's labor supply preferences
  - the gender of the child (may) depend on genetics, but this will not cause a bias

Use of IV

Regression of interest: $y^E = \alpha + \beta x$

Run two regressions

- Endogenous variable on the instrument: $x = \pi_0 + \pi_1 z_{IV}$
- Outcome variable on the instrument: $y = \varphi_0 + \varphi_1 z_{IV}$

What is φ_1 composed of?

- No direct effect of z_{IV} on y (by assumption)
  - Effect of z_{IV} on $X \rightarrow \pi_1$
  - Effect of x on $y \rightarrow \beta$
- $\varphi_1 = \pi \cdot \beta \rightarrow \hat{\beta}_{IV} = \frac{\varphi_1}{\pi_1}$

Example: effect of children on mothers' labor supply

Data: USA, 1990

$y = LFP$, Female labor market participation
$x = 3CHILDREN$, $= 1$ if three children, $= 0$ if two children

$z_{IV} = SAMESEX$ first two children have the same sex

Instrumental variable estimation:

- $3CHILDREN = \pi_0 + 0.06 \cdot SAMESEX$
- $LFP = \varphi_0 - 0.008 \cdot SAMESEX$
- $\hat{\beta}_{IV} = \frac{-0.008}{0.006} = -0.13$

Other IV method: two stage least squares

1. First stage: $x = \pi_0 + \pi_1 \cdot z_{IV} + u \rightarrow \hat{x}_{IV}$
   - The predicted value of x contains only the variability which depends on z_{IV}
2. Second stage: $y = \beta_0 + \beta_{IV} \cdot \hat{x}_{IV} + v$

Where to look for an instrumental variable?

$\hat{z}_{IV}$ should usually be defined at a more aggregate level than the subject, or be something completely random (like genetics)

$\hat{z}_{IV}$ usually originates from the institutional framework

- Compulsory education legislation (USA)
  - Kids go to school in the year they become 6, obliged to stay until 16th birthday
  - z_{IV}: **birth quarter**. Start of school: Q1 – 6.5 years old; Q4 – 5.75 years old $\rightarrow$, Q1 attends less schooling than Q4
- Skill-biased technological change (Norway)
  - High-skilled workers gain more from the introduction of ICT
  - Broadband internet was introduced in different counties according to a plan $\rightarrow$ z_{IV}: **introduction of broadband internet in the region**

Weak instrument

If $cov(z_{IV}, x)$ small $\rightarrow$ **weak instrument**

- If $cov(z_{IV}, u)$ is only slightly different from 0, bias($\hat{\beta}_{IV}$) may be larger than bias($\hat{\beta}_{OLS}$)

Intuition:

- $cov(z_{IV}, x)$ small $\rightarrow$ $cov(z_{IV}, y)$ is the bias
- $\hat{\beta}_{IV} = \frac{\varphi_1}{\pi_1} \rightarrow \frac{\text{bias}}{\approx 0}$

Internal and external validity

Internal validity: high (if we believe that z_{IV} uncorrelated with the error term)

External validity usually low $\rightarrow$ we estimate from very special parts of the data

- Families with 2 vs 3 children
- Students who want to leave the school system asap

Family ownership and firm performance

Main effect of family firm: family appointed CEO

+ More knowledge of firm
+ Long-term goals
– Smaller pool of candidates – lower ability

Succession likely to be endogenous

- Badly performing firms more likely to be sold

Data: Danish, 1994-2002

TABLE I
FIRM CHARACTERISTICS BY TYPE OF CEO SUCCESSION

		Type of Succession		
Variable	All (1)	Family (2)	Unrelated (3)	Difference (4)
Ln assets	8.605	8.232	8.791	−0.559***
	(0.0240)	(0.0332)	(0.0315)	(0.0458)
	[5,334]	[1,776]	[3,558]	
Operating return on assets (OROA)	0.065	0.074	0.061	0.013***
	(0.0020)	(0.0032)	(0.0025)	(0.0041)
	[5,334]	[1,776]	[3,558]	
Net income to assets	0.033	0.038	0.031	0.007*
	(0.0019)	(0.0031)	(0.0024)	(0.0039)
	[5,334]	[1,776]	[3,558]	
Industry-adjusted OROA	−0.002	0.007	−0.006	0.014***
	(0.0020)	(0.0032)	(0.0025)	(0.0041)
	[5,334]	[1,776]	[3,558]	
Firm Age	19.417	19.826	19.213	0.613
	(0.3106)	(0.4840)	(0.3981)	(0.6267)
	[5,334]	[1,776]	[3,558]	

Instrument, first stage

Instrument: z

- cov(z, family succession) = large
- cov(z, firm performance) = 0 (performance other than through family succession)

Instrument, first stage

Instrument: z

- cov(z, family succession) = large
- cov(z, firm performance) = 0 (performance other than through family succession)

z_{IV} = gender of firstborn child

- firstborn girl/boy: 29/39% family succession

First stage: $CEO_{Family} = \alpha + \pi \cdot FirstBorn_{Male}$

Gender of the firstborn child

Description	Number of successions	Family		Unrelated	
		Number	Share	Number	Share
	(1)	(2)	(3)	(4)	(5)
All	5,334	1,776	0.333	3,558	0.667
D. By gender of first born child					
Female	2,216	652	0.294	1,564	0.706
Male	2,476	965	0.390	1,511	0.610
Difference male minus female			**0.096***		
			(0.014)		

First stage regression

Gender of the Firstborn Child, Family Successions, and Performance

	Dependent variable: family CEO						
Part A. First Stage	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Gender of the first born child is male	0.0955***	0.0404**			0.0955***	0.0927***	0.0936***
	(0.0138)	(0.0171)			(0.0136)	(0.0135)	(0.0135)
Male child indicator variable		0.1162***					
		(0.0191)					
Number of male children			0.0737***				
			(0.0077)				
Ratio male to total children				0.1436***			
				(0.0186)			
Ln assets					−0.0448***	−0.0515***	−0.0508***
					(0.0034)	(0.0036)	(0.0037)
Firm age						0.0016***	0.0015***
						(0.0003)	(0.0003)
Industry-adjusted OROA, t = −1						0.2446***	
						(0.0445)	
Industry-and-performance-adjusted OROA, t = −1							0.3374***
							(0.0792)
Year controls	No	No	No	No	Yes	Yes	Yes
F-statistic	48.058	46.566	91.768	59.494	25.590	26.506	24.662
Number of CEO transitions	4,692	4,692	4,692	4,692	4,692	4,692	4,692

Estimated effect

Diff-in-diff: $\Delta ROA = \alpha + \beta_{OLS} \cdot FamilyCEO + X + YEAR$

Second stage: $\Delta ROA = \alpha + \beta_{IV} \cdot \widehat{FamilyCEO} + X + YEAR$

Effect of family succession on ROA

- $\beta_{OLS} = -0.008^{**}$
- $\beta_{IV} = -0.093^{**}$

Regression discontinuity design (RDD)

- Subjects are divided into treated and control groups based on a rule that depends on a **running variable**
  - The running variable has a **threshold value**. On one side of the threshold subjects are treated, on the other they are untreated
  - Division does not have to be strict $\rightarrow$ **fuzzy RDD**
- At values of the running variable **close enough to the threshold**, treatment is considered exogenous
  - BUT subjects should not be able to manipulate the running variable

RDD equation

$$y^E = \alpha + \beta D + f(z_{running})$$

- y = outcome
- $z_{running}$ = running variable
- $D = z_{running}$ above discontinuity threshold value
- $f(.)$ = polynomial function

Examples of RDD

- The effect of trade unions on worker wages
  - Workers vote whether to establish a trade union at the firm. This may be endogenous (why?)
  - Running variable: share of positive votes; threshold: 50%
- The effect of school quality on student outcomes (Romania)
  - Entry to a good school depends on the points achieved at the entrance exam
  - Running variable: points achieved; threshold: the points of the last student who entered at a good school
- The long-run effects of slavery (Peru)
  - Native Americans were taken to work for mines in mountainous regions
  - Running variable: region; threshold: the boundary of the region

Limits of RDD

- Only treatment may be different at the two sides of the threshold
  - Compare the best students in the weaker school with the worst students in the better school
- Subjects must not be able to decide on which part of the boundary they are
  - Ex: firm size dependent regulation
- Limited external validity: we use only subjects close to the threshold
  - Ex: age of the subject

Effect of elite university on wages

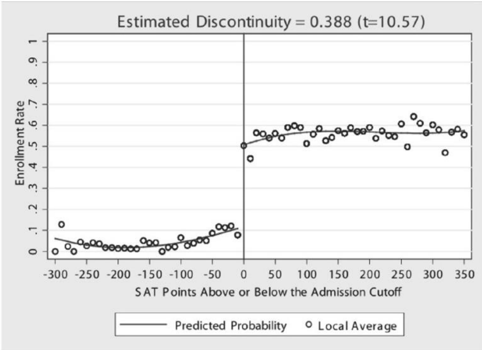
Old question: return to collage *to* is it heterogeneous?

- Is it worth to attend an elite university?

BUT: elite universities more selective → not surprising that they have higher wages after graduation

Discontinuity in acceptance

- Admission based on the SAT score
- Discontinuity at a threshold
- SAT score correlates with ability



Estimate effect on wages

- $y = \alpha + \beta \cdot ABOVE + f(SAT)$
- Do the regression in the neighborhood of the threshold
- Estimated effect: $\hat{\beta} = 0.074 - 0.11$, depending on the bandwidth