GOV1368 Section 5

Mohit Karnani

Harvard

Fall 2024

Recap

Last time we covered:

- Average Treatment Effect on the Treated
- Difference-in-Differences
- Parallel Trends Assumption

Now we can estimate the causal impact of a treatment on the treated under the parallel trends assumption. But what if the assumption does not hold and the treatment is endogenous? How can we estimate the causal impact in this case?

Today we will learn about **Instrumental Variables** as a solution to this problem.

Agenda

Contents

Motivation

Causal question: What is the (average) *causal impact* of preschooling on 4th grade test scores?

Motivation

Causal question: What is the (average) *causal impact* of preschooling on 4th grade test scores?

(Wrong) answer: just compare the average test scores of kids that had preschool education against those who didn't...

Motivation

Causal question: What is the (average) *causal impact* of preschooling on 4th grade test scores?

(Wrong) answer: just compare the average test scores of kids that had preschool education against those who didn't...

Group	Obs	Mean	[95% Conf.	Interval]
No preschool	60,030	262.5497	262.1781	262.9214
Preschool	143,467	268.3822	268.1416	268.6228
Impact		+5.832477	+6.275363	+5.389591

...and conclude that the average causal impact is 5.8 points (12.5% SD).

Why is this wrong?

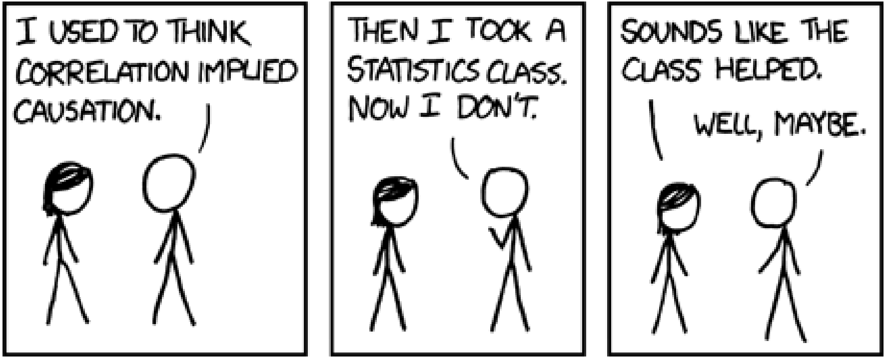
	(1)	(2)	(3)	(4)	(5)	(6)
	father_col~e	mother_col~e	high_income	belief_col~e	private_sc~l	urban_school
preschool	0.149***	0.190***	0.0669***	0.0918***	0.0840***	0.0845***
	(67.88)	(84.98)	(55.93)	(46.41)	(62.47)	(57.87)
_cons	0.230***	0.237***	0.0239***	0.704***	0.0329***	0.831***
	(125.38)	(127.00)	(23.89)	(426.15)	(29.32)	(681.17)
N	214111	214111	214111	214111	214111	214111

Why is this wrong?

	(1)	(2)	(3)	(4)	(5)	(6)
	score	score	score	score	score	score
father_col~e	25.81***					
	(123.02)					
mother_col~e		24.64***				
		(119.87)				
high_income			36.94***			
			(95.21)			
belief_col~e				27.76***		
				(114.49)		
private_sc~l					36.62***	
					(106.95)	
urban_school						14.99***
						(43.86)
_cons	257.8***	257.3***	264.0***	244.9***	263.2***	253.2***
	(2098.44)	(2031.82)	(2516.34)	(1141.93)	(2492.75)	(781.39)
N	203497	203497	203497	203497	203497	203497

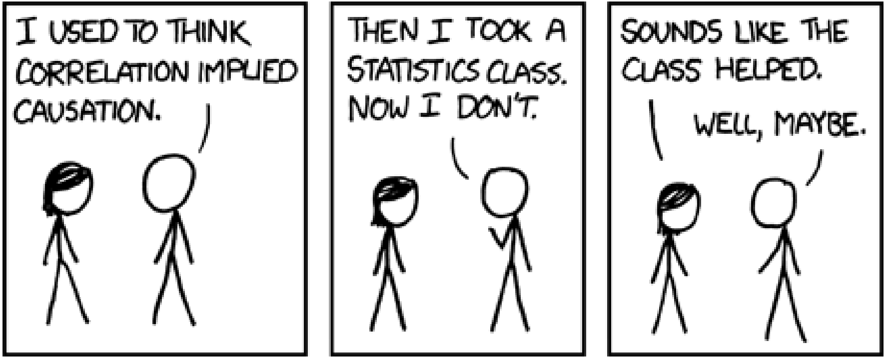
Today's Learning Goals

1. Explain with an example the endogeneity problem. $\checkmark$



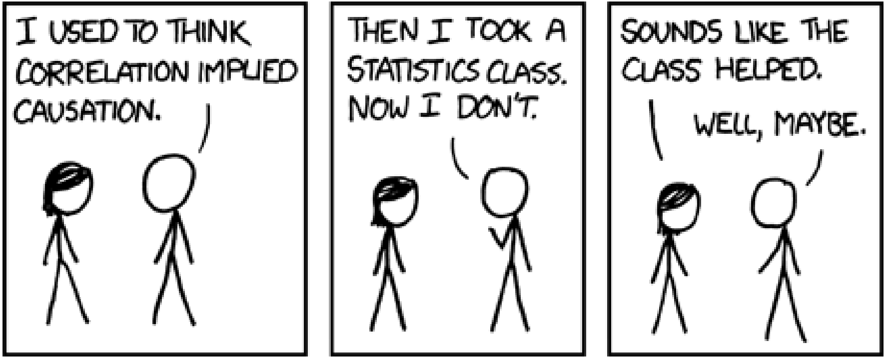
Today's Learning Goals

1. Explain with an example the endogeneity problem. $\surd$
2. Identify the Instrumental Variables (IV) model as a solution to this.



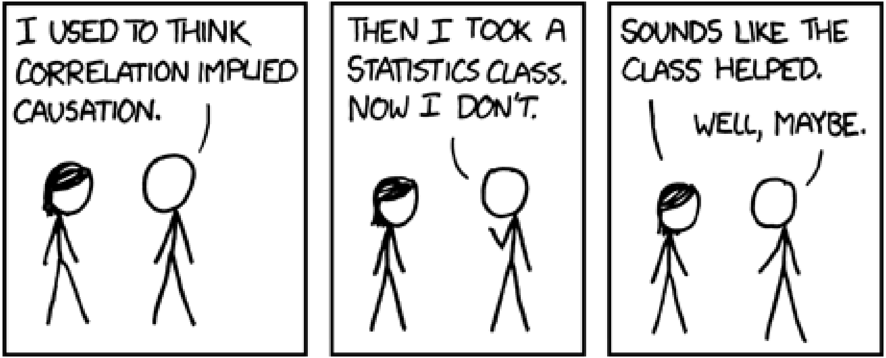
Today's Learning Goals

1. Explain with an example the endogeneity problem. $\surd$
2. Identify the Instrumental Variables (IV) model as a solution to this.
3. Write the formula for the IV-Wald estimator.



Today's Learning Goals

1. Explain with an example the endogeneity problem. $\checkmark$
2. Identify the Instrumental Variables (IV) model as a solution to this.
3. Write the formula for the IV-Wald estimator.
4. Apply this estimator to compute the impact of preschooling on scores.



Contents

Instrumental Variables

Suppose something *exogenously* affects preschool enrollment and does not *directly* impact test scores (e.g. a random voucher program).

Instrumental Variables

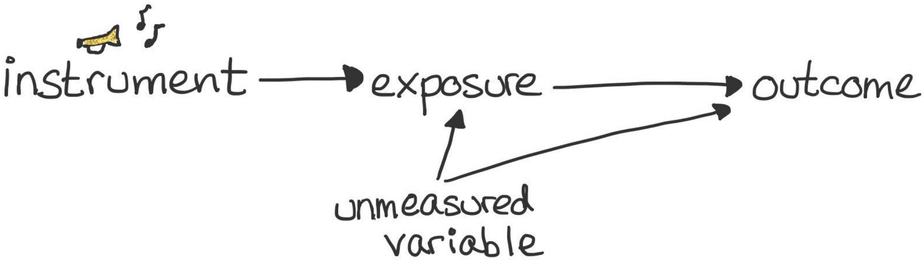
Suppose something *exogenously* affects preschool enrollment and does not *directly* impact test scores (e.g. a random voucher program).

We will call this an **instrumental variable** (or simply an **instrument**).

Instrumental Variables

Suppose something *exogenously* affects preschool enrollment and does not *directly* impact test scores (e.g. a random voucher program).

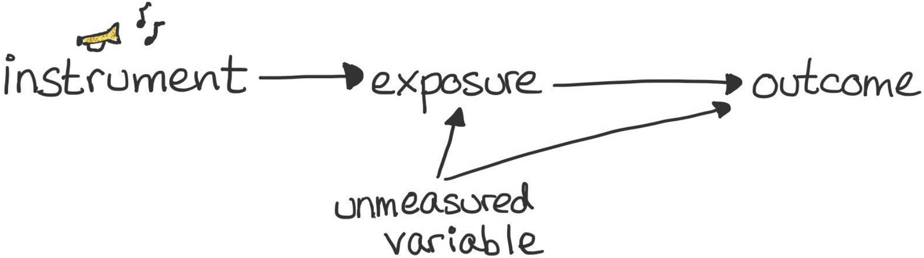
We will call this an **instrumental variable** (or simply an **instrument**).



Instrumental Variables

Suppose something *exogenously* affects preschool enrollment and does <u>not</u> *directly* impact test scores (e.g. a random voucher program).

We will call this an **instrumental variable** (or simply an **instrument**).



How can we use this to obtain a causal estimate?

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

1. Calculate the difference between the average test score of those who received the voucher and those who didn't.

$$\text{IV-Wald} \rightarrow \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} = \frac{\mathbb{E}[Score_i|Voucher]-\mathbb{E}[Score_i|NoVoucher]}{\mathbb{P}[PreK_i|Voucher]-\mathbb{P}[PreK_i|NoVoucher]}$$

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

1. Calculate the difference between the average test score of those who received the voucher and those who didn't.
2. Calculate the difference between the probability of attending preschool for those who received the voucher and those who didn't.

$$\text{IV-Wald} \rightarrow \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]} = \frac{\mathbb{E}[Score_i|Voucher] - \mathbb{E}[Score_i|NoVoucher]}{\mathbb{P}[PreK_i|Voucher] - \mathbb{P}[PreK_i|NoVoucher]}$$

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

1. Calculate the difference between the average test score of those who received the voucher and those who didn't.
2. Calculate the difference between the probability of attending preschool for those who received the voucher and those who didn't.
3. Divide the difference in outcomes by the difference in exposure.

$$\text{IV-Wald} \rightarrow \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]} = \frac{\mathbb{E}[Score_i|Voucher] - \mathbb{E}[Score_i|NoVoucher]}{\mathbb{P}[PreK_i|Voucher] - \mathbb{P}[PreK_i|NoVoucher]}$$

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

1. Calculate the difference between the average test score of those who received the voucher and those who didn't.
2. Calculate the difference between the probability of attending preschool for those who received the voucher and those who didn't.
3. Divide the difference in outcomes by the difference in exposure.

$$\text{IV-Wald} \rightarrow \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} = \frac{\mathbb{E}[Score_i|Voucher]-\mathbb{E}[Score_i|NoVoucher]}{\mathbb{P}[PreK_i|Voucher]-\mathbb{P}[PreK_i|NoVoucher]}$$

The IV-Wald Estimator

Recipe for computing the causal effect of preschooling D_i on test scores Y_i using a randomly assigned voucher Z_i as an instrument:

1. Calculate the difference between the average test score of those who received the voucher and those who didn't.
2. Calculate the difference between the probability of attending preschool for those who received the voucher and those who didn't.
3. Divide the difference in outcomes by the difference in exposure.

$$\text{IV-Wald} \rightarrow \frac{\mathbb{E}[Y_i|Z_i=1]-\mathbb{E}[Y_i|Z_i=0]}{\mathbb{E}[D_i|Z_i=1]-\mathbb{E}[D_i|Z_i=0]} = \frac{\mathbb{E}[Score_i|Voucher]-\mathbb{E}[Score_i|NoVoucher]}{\mathbb{P}[PreK_i|Voucher]-\mathbb{P}[PreK_i|NoVoucher]}$$

Another interpretation: it is the Intent To Treat (ITT) divided by the "first stage".

Contents

Local Average Treatment Effect (LATE)

Our target estimand is the **Local Average Treatment Effect** (LATE).

Local Average Treatment Effect (LATE)

Our target estimand is the **Local Average Treatment Effect** (LATE).

This is the average treatment effect on the *compliers*, i.e. the causal impact of preschooling on test scores for those who would attend preschool *only* if they received the voucher, and not otherwise:

$$LATE := \mathbb{E}[Y_i(1) - Y_i(0) | D_i(Z_i = 1) > D_i(Z_i = 0)]$$

Local Average Treatment Effect (LATE)

Our target estimand is the **Local Average Treatment Effect** (LATE).

This is the average treatment effect on the *compliers*, i.e. the causal impact of preschooling on test scores for those who would attend preschool *only* if they received the voucher, and not otherwise:

$$LATE := \mathbb{E}[Y_i(1) - Y_i(0) | D_i(Z_i = 1) > D_i(Z_i = 0)]$$

A taxonomy of the population based on the potential outcomes:

- Compliers: $D_i = Z_i$, i.e. those who attend preschool only if they receive the voucher.
- Always-takers: $D_i = 1$, i.e. those who attend preschool regardless of the voucher.
- Never-takers: $D_i = 0$, i.e. those who don't attend preschool regardless of the voucher.
- Defiers: $D_i = 1 - Z_i$, i.e. those who attend preschool only if they don't receive the voucher. Assumption: there are no defiers.

Identification Assumptions

Our identification strategy relies on the following assumptions:

1. **Relevance**: The instrument Z_i is *relevant* for the treatment D_i, i.e.

$$\mathbb{P}[D_i = 1 | Z_i = 1] \neq \mathbb{P}[D_i = 1 | Z_i = 0].$$

Identification Assumptions

Our identification strategy relies on the following assumptions:

1. **Relevance**: The instrument Z_i is *relevant* for the treatment D_i, i.e.

$$\mathbb{P}[D_i = 1 | Z_i = 1] \neq \mathbb{P}[D_i = 1 | Z_i = 0].$$

2. **Exclusion**: The instrument Z_i affects the outcome Y_i *only* through the treatment D_i:

$$\mathbb{E}[Y_i(1) - Y_i(0) | D_i = d, Z_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = d, Z_i = 0].$$

Identification Assumptions

Our identification strategy relies on the following assumptions:

1. **Relevance**: The instrument Z_i is *relevant* for the treatment D_i, i.e.

$$\mathbb{P}[D_i = 1|Z_i = 1] \neq \mathbb{P}[D_i = 1|Z_i = 0].$$

2. **Exclusion**: The instrument Z_i affects the outcome Y_i *only* through the treatment D_i:

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i = d, Z_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = d, Z_i = 0].$$

3. **Monotonicity**: There are no defiers, i.e. $D_i \neq 1 - Z_i$.

Identification Assumptions

Our identification strategy relies on the following assumptions:

1. **Relevance**: The instrument Z_i is *relevant* for the treatment D_i, i.e.

$$\mathbb{P}[D_i = 1|Z_i = 1] \neq \mathbb{P}[D_i = 1|Z_i = 0].$$

2. **Exclusion**: The instrument Z_i affects the outcome Y_i *only* through the treatment D_i:

$$\mathbb{E}[Y_i(1) - Y_i(0)|D_i = d, Z_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0)|D_i = d, Z_i = 0].$$

3. **Monotonicity**: There are no defiers, i.e. $D_i \neq 1 - Z_i$.
4. **Independence**: The instrument Z_i is *independent* of the potential outcomes $Y_i(0)$ and $Y_i(1)$.

Hypothetical Example

Suppose you want to calculate the impact of preschool education on the test scores of 4th graders.

Many years ago, someone implemented a voucher program to encourage a random set of families to send their kids to preschool at a reduced cost. As a result, 85% of the beneficiaries of this voucher enrolled in some preschool, whereas 70% of non-beneficiaries did so.

When comparing the 4th-grade test scores for beneficiaries and non-beneficiaries, the former achieve an average score of 268.5, while the latter score 267.0 on average.

What is the causal impact of preschooling on scores in this case?

Solution

1. Difference between the average test score of those who received the voucher and those who didn't: $268.5 - 267 = 1.5$
2. Difference between the probability of attending preschool for those who received the voucher and those who didn't: $0.85 - 0.7 = 0.15$
3. Quotient of the difference in outcomes by the difference in exposure:

$$LATE = \frac{\mathbb{E}[Score|Voucher] - \mathbb{E}[Score|NoVoucher]}{\mathbb{P}[Preschool|Voucher] - \mathbb{P}[Preschool|NoVoucher]} = \frac{1.5}{0.15} = 10$$

$\therefore$ the causal impact of attending preschool is an average increase of 10 points.

Contents

Two-Stage Least Squares (2SLS)

There is a more general way of implementing instrumental variables estimation when we have an endogenous variable X_i (which can now be continuous, such as school spending) being instrumented by Z_i (which can also be continuous, such as the increase in funding due to a reform, or the years of exposure to a funding reform).

The method is called **Two-Stage Least Squares** (2SLS), because it involves two regression models:

$$X_i = \gamma + \delta Z_i + u_i \qquad \text{(First Stage)}$$
$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \text{(Second Stage)}$$

We simply run the first stage regression (endogenous variable X_i on instrument Z_i) and compute the "predicted" values $\hat{X}_i$ (e.g. the predicted expenditure in Jackson et. al. 2015).

Contents

Application: School Spending and Educational, Economic Outcomes

`Stata`