Experimental Analysis of Public Policies

Seance 0

Econometric Intuitions for Applied Work

Etienne Dagorn - Universite de Lille

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Why Causal Inference Matters

- → Public policies are choices under scarcity: training slots, subsidies, inspections, classrooms, clinics.
- → Administrative data often show strong correlations between policies and outcomes.
- → The policy question is not only what happened to treated people.
- → It is what would happen if we changed the treatment rule.

Econometric task

Move from observed comparisons to credible counterfactual comparisons.

Why This Refresher?

Objective

Reconnect familiar econometric tools with applied empirical reasoning and causal questions.

→ You may have heard about OLS, inference, panel data, and IV elsewhere.

→ This session asks what these tools mean in applied work.

→ The central distinction: estimate, inference, identification.

→ Correlations are often useful, but they rarely settle policy questions by themselves.

→ Running example: a job-training programme and employment.

The Econometric Approach

1. Start from an **economic or policy question**.
2. State the **comparison** that would answer it.
3. Identify the **variation** that makes the comparison informative.
4. Estimate the relationship and quantify **uncertainty**.
5. Interpret the result, assumptions, and limits.

Key idea

A sophisticated econometric model cannot compensate for a weak identification strategy.

The Three Questions

Idea	Question
Estimate	What number did we compute from the data?
Inference	How uncertain is that number?
Identification	Can the number be interpreted causally?

Course reflex

A precise estimate is not necessarily a causal estimate.

Applied work combines all three: compute the estimate, quantify uncertainty, and justify the comparison.

What Is Identification?

Identification

Identification asks whether the target parameter can, **in principle**, be recovered from the data variation under stated assumptions.

→ What variation moves the treatment, exposure, or comparison?

→ Why is that variation separated from other determinants of the outcome?

→ Which assumption lets the observed comparison stand in for the missing counterfactual?

Key idea

More observations help only after the comparison is informative about the target.

Opening Poll: Which Claim Is Causal?

1. Neighbourhoods with more police have more recorded crime.
2. Students in smaller classes have higher test scores.
3. Workers who complete training are more often employed.

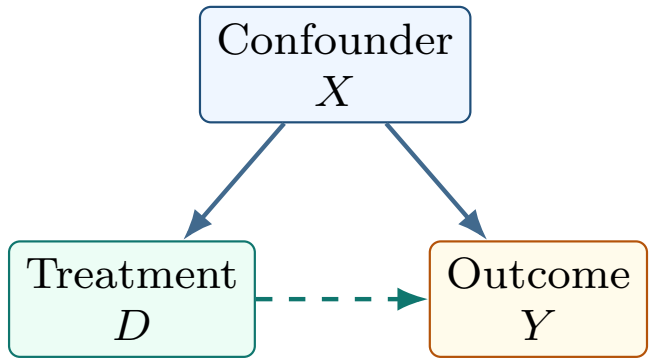
Show of hands

For each claim: causal effect, selection/confounding, or impossible to tell?

Follow-up: What extra comparison would make the claim more credible?

How to Read the DAGs

- → Each node is a variable or concept.
- → Each arrow means "may causally affect."
- → A backdoor path is a noncausal path linking treatment and outcome.
- → The goal is intuition, not graphical formalism.



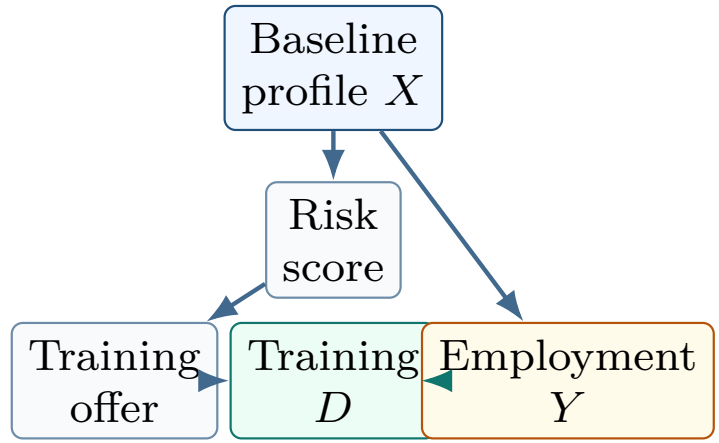
Teaching point

The picture separates the causal question from the observed association.

Example and DAG: Prediction vs Causation

Example: the agency builds a risk score to target training.

- → Prediction: who is likely to be unemployed?
- → Causation: whose employment would training change?
- → The same variables can help prediction and still confound treatment.



Discussion

Would a perfect prediction model automatically tell us whether training works?

Running Example: Training and Employment

- → A public employment agency offers a voluntary training programme.
- → Outcome: employed six months later.
- → Treatment: participation in the programme.
- → Data: unemployed workers, baseline characteristics, participation, employment.
- → Policy question: does training increase employment?

Question	**Training example**
Descriptive	Are trainees more often employed?
Predictive	Which workers are likely to find a job?
Causal	Would trainees have found jobs without training?

Activity: Translate the Policy Question

Element	Training example
Treatment	participates in training vs does not participate
Outcome	employed six months later
Observed comparison	trainees vs non-trainees
Missing counterfactual	trainees if they had not trained
Policy decision	fund, scale, redesign, or stop the programme

Pairs - 2 minutes

Write the missing counterfactual in one sentence.

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Five Descriptive Statistics to Keep Straight

Concept	What it measures	Key intuition
Mean	Central level	What is the average value of Y?
Variance	Dispersion	How far are observations spread around the mean?
Standard deviation	Original units	Typical distance from the mean.
Covariance	Joint variation	Do X and Y tend to move together?
Correlation	Standardised joint variation	Strength of linear co-movement.

Course reflex

Descriptive statistics summarise the data we observe. They do **not**, by themselves, tell us why the data look that way.

Mean: Central Level

Mean: central level

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- → The mean is the average value of Y in the sample.
- → It is measured in the same units as Y.
- → It is the first observed comparison, not a causal explanation.

Training example

If 60% of trainees are employed six months later, the mean of the employment indicator among trainees is 0.60.

Dispersion: Variance and Standard Deviation

Variance: dispersion

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

→ Measures dispersion around the mean.

Standard deviation

$$s_Y = \sqrt{s_Y^2}$$

The standard deviation measures dispersion in the **same units as** Y, which usually makes it easier to interpret than the variance.

Applied point

More outcome dispersion usually makes effects harder to estimate precisely.

Why Sample Variance Uses $n - 1$

Example: three observations

Deviations from the sample mean must sum to zero.

Suppose two deviations are

$$+2 \qquad \text{and} \qquad -1.$$

What must the third deviation be?

Key idea

Once the mean is fixed, the last deviation is determined by the others.

$$\boxed{\text{degrees of freedom} = n - 1}$$

Why Sample Variance Uses $n - 1$

Example: three observations

Deviations from the sample mean must sum to zero.

Suppose two deviations are

$$+2 \qquad \text{and} \qquad -1.$$

What must the third deviation be?

$$\boxed{-1}$$

Key idea

Once the mean is fixed, the last deviation is determined by the others.

$$\boxed{\text{degrees of freedom} = n - 1}$$

Historical Example: Why "Student" Cared About Small Samples

Gosset at Guinness (1908)

How much can we learn from only a few observations when the variance is unknown?

→ Large experiments were costly.

→ The population variance had to be estimated from the same small sample.

→ That extra uncertainty matters most when n is small.

Imagine $n = 5$: you observe five experimental yields and estimate both $\bar{Y}$ and s_Y. How confident should you be about the population mean?

The key insight

The t-distribution depends on degrees of freedom: $df = n - 1$.

Covariance: Direction of Co-Movement

Sample covariance

For two variables X and Y,

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

- $\rightarrow$ $s_{XY} > 0$: X and Y tend to move together.
- $\rightarrow$ $s_{XY} < 0$: high X tends to accompany low Y.
- $\rightarrow$ $s_{XY} \approx 0$: little linear co-movement.

Reading

Covariance gives direction in units that are often hard to interpret.

Correlation: Standardised Co-Movement

Sample correlation

Standardise the covariance:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}.$$

Therefore,

$$-1 \leq r_{XY} \leq 1.$$

- → Sign: direction of the linear association.
- → Magnitude: strength of the linear association.
- → No units: correlation is scale-free.

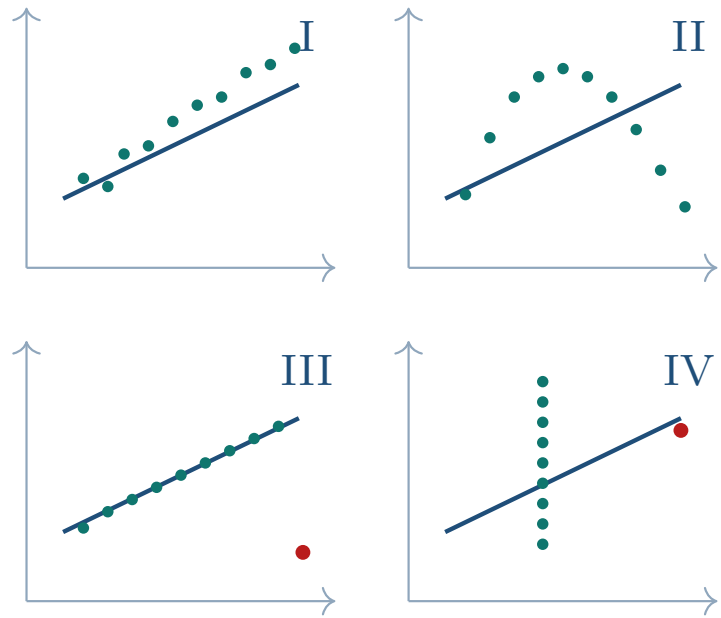
Historical Example: Anscombe's Quartet

Anscombe (1973)

Four datasets can share nearly the same means, variances, correlation, and fitted line.

Lesson

Never read a regression table without asking what the data look like.

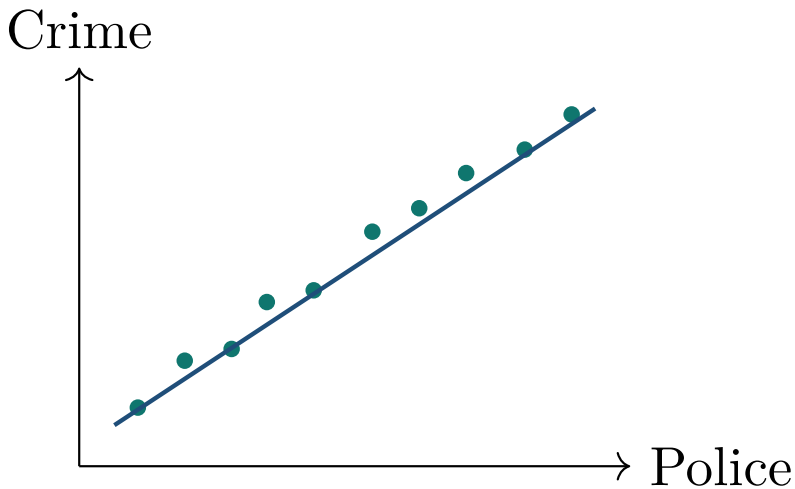


Same regression line; very different stories.

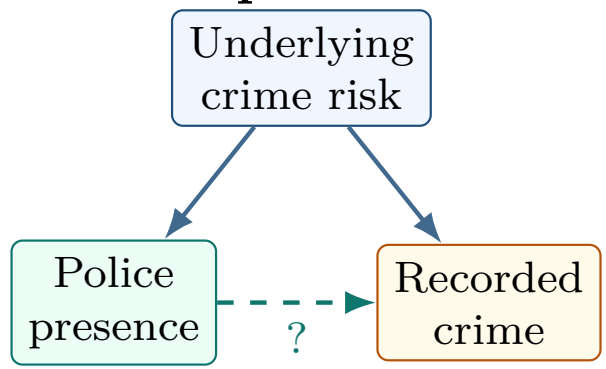
Question: Which dataset would make you distrust the linear summary most?

Example and DAG: Correlation

Example: police and recorded crime



Simple DAG



Teaching point

A positive slope can reflect where police are sent, not what police cause.

Discussion: Three Stories for One Correlation

Observed fact: trainees have higher employment than non-trainees.

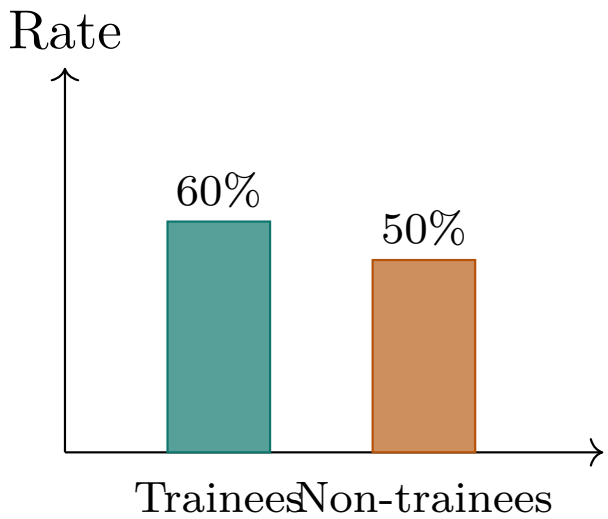
Story	**What it implies**
Training builds useful skills	positive causal effect
Motivated workers choose training	upward selection bias
Caseworkers send hardest cases to training	downward selection bias

Short discussion

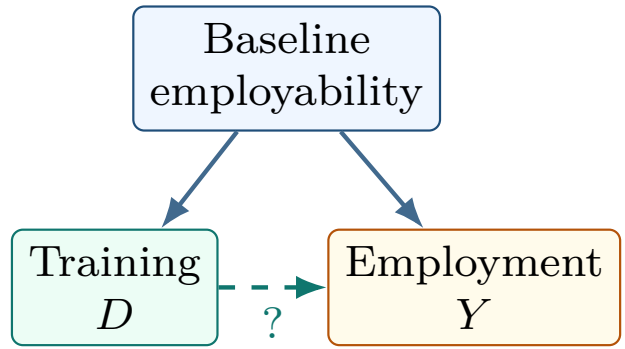
Which story is most plausible? What data would help distinguish them?

Example and DAG: Descriptive Difference

Figure: observed employment rates



DAG behind the comparison



Teaching point

The bar chart is descriptive. The DAG asks what else produced the bars.

A First Comparison

$$\Delta = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

- $D_i = 1$: worker participated in training.
- $Y_i = 1$: worker is employed after six months.
- Δ compares employment rates for trainees and non-trainees.
- It summarises observed outcomes, not the missing counterfactual.

Question

Is this difference the effect of training, or a difference between who selects into training?

What OLS Does Intuitively

OLS

OLS chooses the line that makes residuals as small as possible in squared distance.

$$Y_i = \alpha + \beta X_i + u_i$$

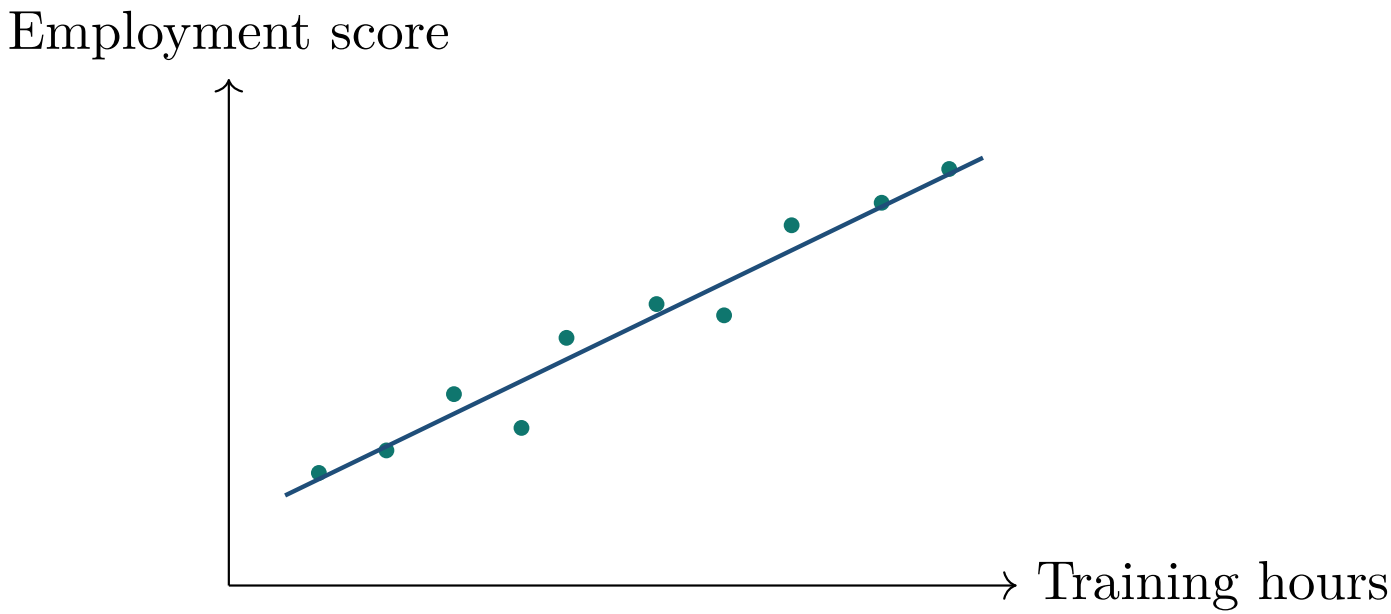
$$\widehat{\beta}_1 = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}$$

→ Y_i: outcome.

→ X_i: explanatory variable.

→ $\widehat{\beta}_1$: slope of the fitted linear association.

The Regression Line



→ The line summarises the average relationship.

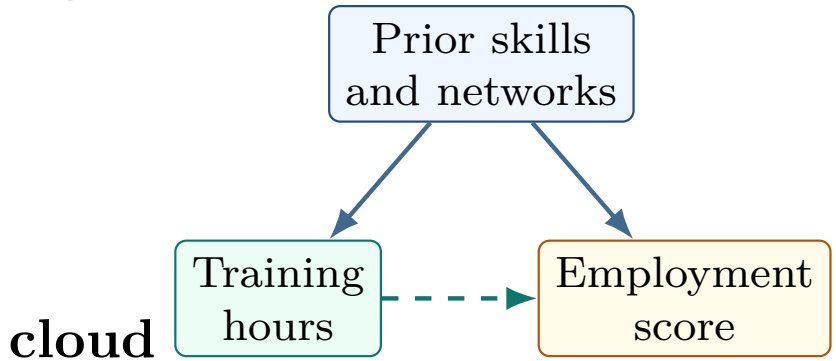
→ Points above or below the line are not perfectly predicted.

Example and DAG: OLS Slope

Example: training hours

- → OLS slope: employment score rises with training hours.
- → But stronger workers may both take more hours and find jobs.
- → The slope summarises a conditional pattern.

Causal structure under the cloud



Teaching point

OLS gives a slope. The DAG tells us why the slope may not be causal.

Fitted Values and Residuals

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i, \qquad \widehat{u}_i = Y_i - \widehat{Y}_i$$

- → Fitted value: model prediction for unit i.
- → Residual: observed outcome minus fitted value.
- → Residuals show what the model did not explain.

Reading a regression

The regression line is a summary; residuals remind us that individuals differ.

Quick Calculation: Fitted Value and Residual

$$\widehat{Employment}_i = 0.20 + 0.05\, TrainingHours_i$$

→ Worker has $TrainingHours_i = 4$.

→ Observed employment score is 0.55.

1. What is the fitted value?
2. What is the residual?
3. Is the residual a causal effect?

Individual - 2 minutes

Compute the prediction and the leftover.

Residuals vs Error Terms

Object	Meaning
Error term u_i	unobserved part in the population model
Residual $\widehat{u}_i$	leftover part after estimating the model

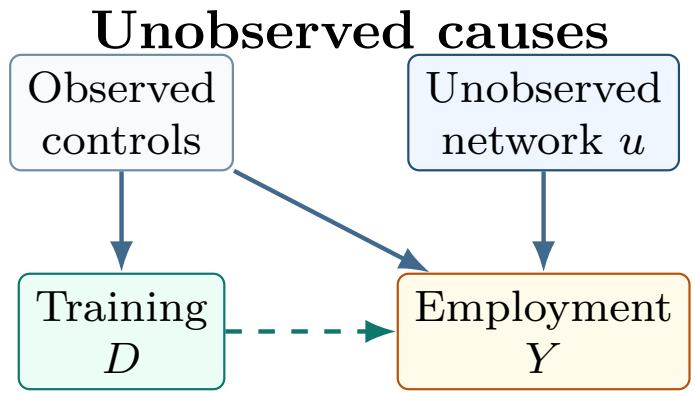
Vocabulary

The error term is theoretical; the residual is computed from the sample.

Example and DAG: What Lives in u_i

Example: one worker above the line

- → Predicted employment score: 0.40.
- → Observed employment score: 0.55.
- → Residual: 0.15.
- → Possible reason: an informal network not in the data.



Teaching point

Residuals are computed leftovers; error terms represent omitted forces in the model.

Multiple Regression

$$Y_i = \alpha + \beta D_i + \gamma X_i + u_i$$

- D_i: training participation.
- X_i: observed pre-training characteristics.
- β: the conditional association between training and employment after accounting linearly for X_i.

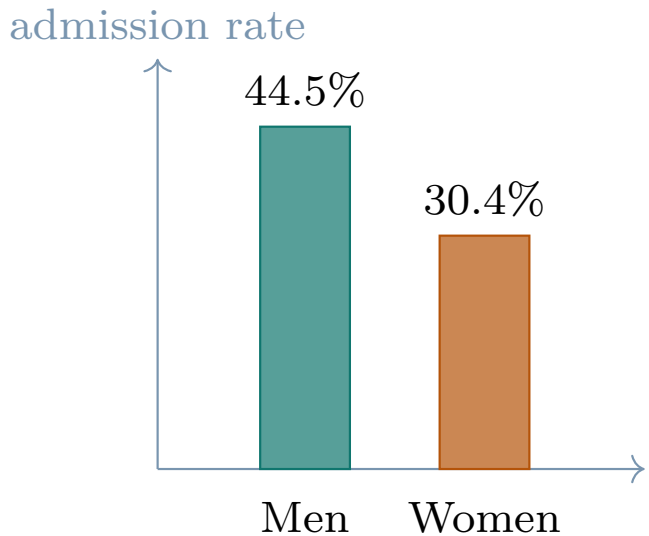
Important

"Controlling for X" changes the fitted comparison; it is causal only under additional assumptions.

Historical Example: Berkeley Admissions

Bickel et al. (1975)

Graduate admissions at Berkeley in 1973 looked biased against women in the aggregate.



Prediction: what happens if we compare applicants *within departments*?

Berkeley Admissions: Conditioning

Dept.	Men admitted	Women admitted
A	62%	82%
B	60%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%

Lesson
Conditioning can change an association because groups choose different contexts.

Interpretation question
Does the within-department table settle every fairness question?

Controls and Conditional Associations

With controls, a regression coefficient is a fitted conditional association:

$$Y_i = \alpha + \beta D_i + \gamma X_i + u_i.$$

- → X_i can include education, prior earnings, and local labour-market conditions.
- → $\widehat{\beta}$ summarises how Y differs with D, after accounting linearly for X.
- → This is not literal one-to-one matching of identical workers.
- → The causal interpretation comes later, from assumptions about treatment assignment.

Regression reading

Say what $\widehat{\beta} = 0.06$ means without using the word "causes."

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Estimate vs True Parameter

Object	Meaning
β	true population parameter or target relationship
$\widehat{\beta}$	estimate computed from one sample

Key idea

The estimate changes across samples; the target parameter is the object we want to learn about.

Bias, Consistency, and Efficiency

Property	Question
Bias	Does the procedure miss the target on average?
Consistency	Does the estimator approach the target in very large samples?
Efficiency	Among appropriate estimators, does it have low variance?

Identification reminder

An efficient estimator of the wrong target is still the wrong target.

Example: Target vs Estimate

Example: what are we targeting?

Object	Training example
Target population	all eligible unemployed workers in Lille
Target parameter	average effect of training for that population
Observed sample	240 workers in one intake wave
Estimate	$\widehat{\beta}$ computed from that sample

Teaching point

An estimate is sample-specific; the parameter is the object we want to learn about.

Sampling Variation

- → Different samples would give different estimates.
- → Sampling variation is normal, not a mistake.
- → Larger and more informative samples usually reduce variation.
- → Clustered or noisy data can keep uncertainty large.

Intuition

Inference asks how much $\widehat{\beta}$ would move if the study were repeated.

Estimate vs Inference

Estimate	$\widehat{\beta} = 0.06$
Inference	$SE = 0.03$, CI, p-value, test

Core distinction

The coefficient gives the estimated magnitude. Inference quantifies uncertainty around it.

Identification comes first

Inference does not tell us whether the parameter has the causal interpretation we want.

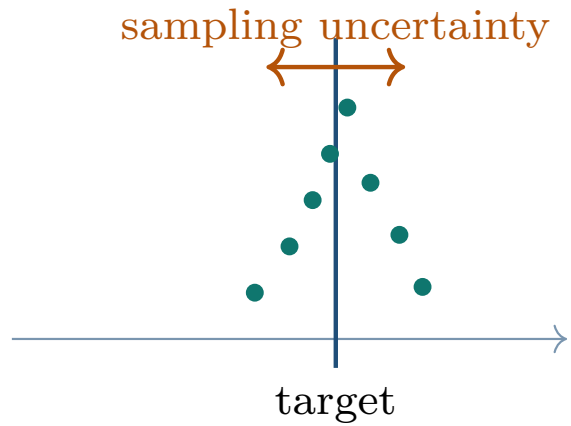
Standard Errors: How Much Would the Estimate Move?

Suppose we could repeat the same study many times.

Each sample would give a slightly different estimate:

$$\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}, \widehat{\beta}^{(3)}, \ldots$$

Standard error: the typical sampling movement of $\widehat{\beta}$.



Question: what would make these estimates more spread out?

Same Estimate, Same Evidence?

Study	Workers	$\widehat{\beta}$	Standard error
A	5,000	0.06	0.01
B	80	0.06	0.08

Read the results

Same estimate; different standard errors. Would you describe the evidence in the same way?

Same Estimate, Same Evidence?

Study	Workers	$\widehat{\beta}$	Standard error
A	5,000	0.06	0.01
B	80	0.06	0.08

Read the results

Same estimate; different standard errors. Would you describe the evidence in the same way?

Reading

Study A is precise; Study B is uncertain. Same magnitude is not the same information. A small standard error does not make an estimate causal.

Confidence Intervals: Calculation

Suppose we estimate:

$$\widehat{\beta} = 0.06, \qquad SE(\widehat{\beta}) = 0.03.$$

For a large sample, an approximate 95% confidence interval is

$$\widehat{\beta} \pm 1.96 \times SE(\widehat{\beta})$$

$$0.06 \pm 1.96(0.03) \approx \boxed{[0.00,\ 0.12]}.$$

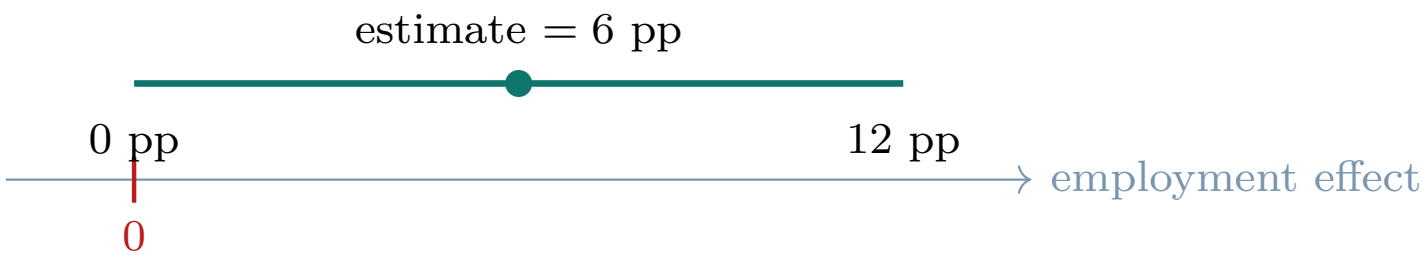
Interpretation discipline

The interval describes uncertainty in the estimation procedure; it does not prove causal identification.

Confidence Intervals: Policy Interpretation

Read the interval in substantive units

The data are compatible with effects ranging from **almost no increase** to about a **12 percentage-point increase**.



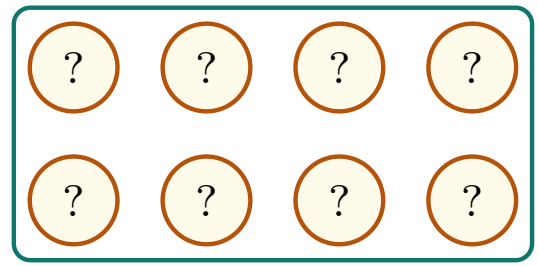
Policy interpretation

Would all of these imply the same policy decision?

Historical Example: Fisher's Tea Experiment

Fisher (1935)

A subject claims she can tell whether milk or tea was poured first.



choose exactly 4 cups as milk-first

- → 8 cups: 4 milk-first and 4 tea-first.
- → Random order.
- → Under guessing, all allocations of 4 cups are equally likely.

Null calculation: perfect classification has probability $1/\binom{8}{4} = 1/70 \approx 0.014$ under guessing.

Class vote

Would perfect classification be surprising under random guessing?

P-Values and Statistical Significance

- → A p-value summarises evidence against a null hypothesis.
- → Common null: $\beta = 0$.
- → "Significant" usually means unlikely under the null.
- → It does not measure importance or causality.

Do not say

"The p-value is the probability that the null is true."

Optional Preview: Type I and Type II Errors

	Reject H_0	**Do not reject H_0**
H_0 **true**	Type I error	Correct non-rejection
H_0 **false**	Correct rejection	Type II error

→ Type I error: finding an effect when the null is true.

→ Type II error: missing an effect when the null is false.

→ Larger samples often reduce Type II errors by improving precision.

Quick check

If a study is underpowered, which error becomes especially likely?

Statistical vs Economic Significance

Scenario: a large dataset estimates that training raises employment by 0.5 percentage points with $SE = 0.1$ percentage points. The programme is expensive.

Short debate

Would you fund it? Separate statistical evidence from policy relevance.

Policy reading

A tiny precise effect may be statistically significant but economically unimportant.

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Thought Experiment: The Same Worker Twice

Impossible observation

Amina participates in training and is employed six months later. Did training cause her employment?

1. Which outcome do we observe?
2. Which outcome is missing?
3. Who could serve as a credible stand-in for the missing outcome?

Thought Experiment: The Same Worker Twice

Impossible observation

Amina participates in training and is employed six months later. Did training cause her employment?

1. Which outcome do we observe?
2. Which outcome is missing?
3. Who could serve as a credible stand-in for the missing outcome?

Teaching point

Causal inference is the art of replacing the missing path with a credible comparison.

Potential Outcomes: One Worker, Two Worlds

For each worker i:

$$Y_i(1) = \text{outcome if treated}, \qquad Y_i(0) = \text{outcome if untreated}.$$

$$\tau_i = Y_i(1) - Y_i(0).$$

Worker	D	$Y(1)$	$Y(0)$
Amina	1	observed	missing
Boris	0	missing	observed

Fundamental problem

We never observe both potential outcomes for the same unit at the same time.

Identification vs Estimation

Identification

Can the causal parameter be recovered in principle from the available variation and assumptions?

Question about the **research design**.

$$\text{Data} + \text{assumptions} \Longrightarrow ATE?$$

Estimation

Given an identified parameter, how precisely can we estimate it from a finite sample?

Question about **sampling uncertainty**.

$$\widehat{\theta}_N \approx \theta$$

Important

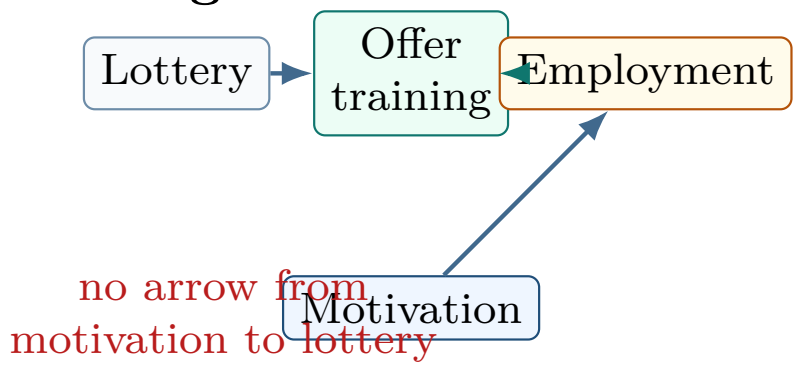
More observations can improve precision, but they do **not** solve an identification problem.

Example and DAG: Identification

Example: lottery offer

- → More eligible workers apply than the programme can serve.
- → Seats are allocated by lottery.
- → Compare workers offered a seat to workers not offered a seat.
- → The design creates a credible comparison at baseline.

Design breaks selection



Teaching point

Identification comes from why treatment variation is separated from untreated potential outcomes.

Why Assumptions Matter

Assumption

An assumption links an observed comparison to an unobserved counterfactual.

→ OLS with controls: selection is captured by observed pre-treatment covariates.

→ IV: an instrument provides exogenous variation in treatment.

→ DiD: treated and control units would have followed parallel trends.

Question

Which assumption would make non-trainees a credible comparison for trainees?

Selection on Observables

For OLS with controls to have a causal interpretation, a common assumption is:

$$(Y_i(1), Y_i(0)) \perp D_i \mid X_i.$$

Plain language

Conditional on the included pre-treatment covariates, there must be no remaining factor jointly determining treatment assignment and potential outcomes.

Check

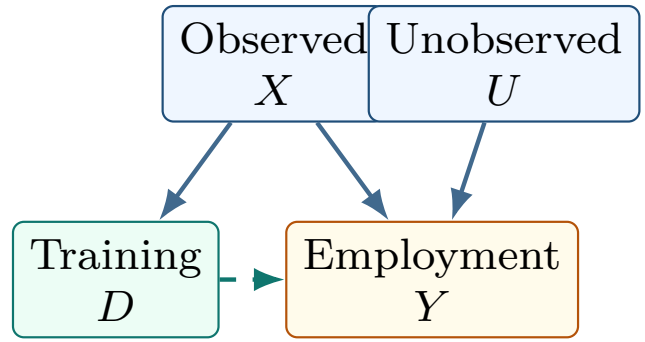
Why can we not learn much about training effects for workers who are always assigned to training?

Example and DAG: Assumptions

Example: OLS with controls

- → $\widehat{\beta}$ summarises the conditional association after accounting linearly for X.
- → X: education, prior earnings, local labour market.
- → Causal claim: no remaining unobserved selection into training after conditioning on X.

Assumption as a missing arrow



assume no arrow from U to D

Teaching point

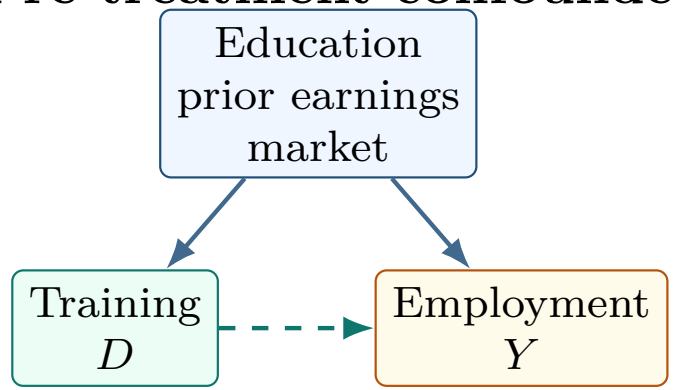
An identification assumption is a claim about blocked or absent non-causal paths.

Confounders and Good Controls

Example: compare better-adjusted workers

- → Education, prior earnings, and local market are measured before training.
- → They may affect both training participation and employment.
- → Controlling for them can block a backdoor path.

Pre-treatment confounder



Teaching point

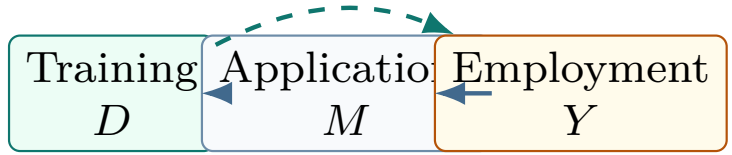
Good controls are pre-treatment common causes of treatment and outcome.

Mediators: When Controls Change the Estimand

Example: job applications

- → Training improves CV quality and search confidence.
- → Workers then send more job applications.
- → Applications increase employment.
- → Controlling for applications changes the question.

Mechanism



Teaching point

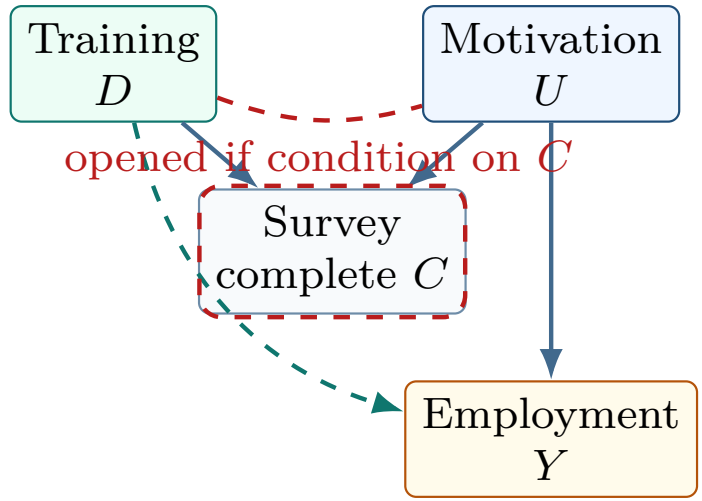
If M is part of the treatment effect, controlling for M blocks part of the total effect.

Colliders: When Controls Create Bias

Example: analysing completers only

- → Training can raise survey completion.
- → Motivation can raise survey completion and employment.
- → Restricting to completers conditions on C.
- → That can create a misleading association.

Conditioning on a collider



Teaching point

Do not condition mechanically: some variables create bias when selected on.

Activity: Which Controls Belong?

Candidate controls for the training regression:

- → age, education, prior earnings
- → local unemployment rate before training
- → number of job applications after training
- → confidence in interviews after training
- → motivation

Sort the list

Pre-treatment confounder, mediator, collider or selection variable, or still unobserved?

Roadmap

From Economic Questions to Empirical Evidence

Describing Data and Reading Regressions

Estimation and Statistical Inference

Causal Identification and Counterfactuals

Threats to Identification and What Economists Do About Them

Four Common Threats to Causal Interpretation

A regression coefficient can fail to answer the causal question for several distinct reasons.

- → **Omitted-variable bias**: a missing factor affects both X and Y.
- → **Reverse causality**: Y also affects X.
- → **Measurement error**: the variable we observe differs from the variable in the causal question.
- → **Selection bias**: we only observe or compare a selected group.

Key idea

More data can make a biased estimate very precise.

Omitted-Variable Bias: Two Ingredients

Suppose the true model is:

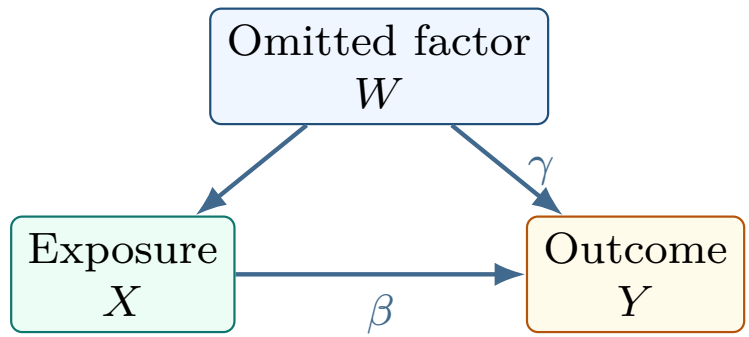
$$Y_i = \alpha + \beta X_i + \gamma W_i + \varepsilon_i,$$

but we estimate the shorter regression that omits W_i.

Teaching point

Bias needs two ingredients:

- $\rightarrow$ W predicts Y;
- $\rightarrow$ W is correlated with X.



One minute check

Can an omitted variable matter for prediction but create no bias in $\widehat{\beta}$?

Omitted-Variable Bias: Direction

The sign of the bias depends on two signs.

Effect of omitted W on Y	Correlation of W with X	Bias in $\widehat{\beta}$
$+$	$+$	Upward
$+$	$-$	Downward
$-$	$+$	Downward
$-$	$-$	Upward

Short calculation

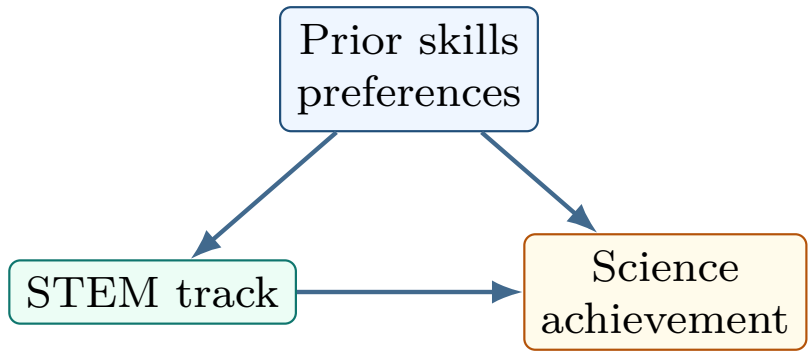
Education is positively related to wages. Ability raises wages and is positively related to education. In which direction is the simple education coefficient biased?

Expected answer: upward, because ability is positively related to both education and wages.

Example and DAG: STEM Track Choice

Question: Do students in a STEM track score higher in science because of the track itself?

- → Observed difference: STEM-track students often score higher in science.
- → Causal concern: prior preparation and preferences may affect both track choice and later achievement.
- → The naive comparison may exaggerate the effect of the track.

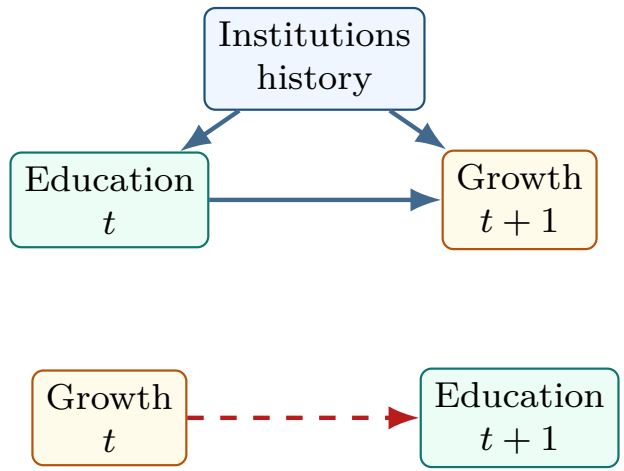


Pair question

What data would make this comparison more credible?

Reverse Causality: Education and Growth

Question: Does education cause economic growth, or do richer economies invest more in education later?



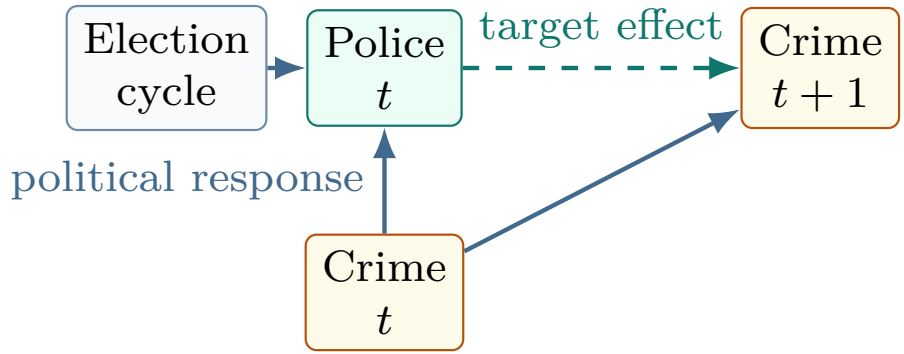
Discussion prompt

If education reflects earlier growth shocks omitted from the regression, why is X correlated with the error term?

Historical Example: Police and Crime

Levitt (1997)

Cross-city correlations between police and crime are hard to interpret because crime may itself lead cities to hire more police.



Lesson: a causal design tries to isolate one direction of causality.

→ Naive association: more police where crime is high.

→ Reverse causality: crime affects police hiring.

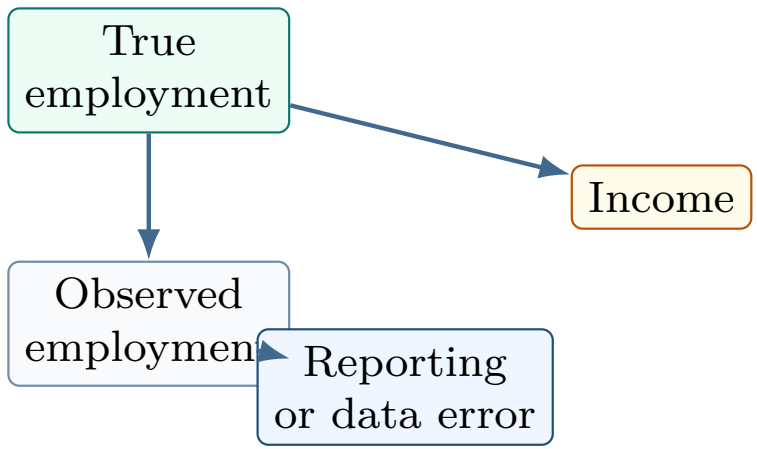
→ IV idea: electoral cycles shift police hiring.

Measurement Error: Why It Matters

Sometimes the variable in the dataset is only an imperfect proxy for the variable in the causal question.

Example: informal employment may be missing from administrative data.

- → True employment is the causal variable.
- → Observed employment is recorded with error.
- → Systematic error can create bias; random error can reduce precision or attenuate slopes.



Quick question

Would this problem disappear if the sample were very large?

Measurement Error: Regressor vs Outcome

Problem	Typical consequence
Classical error in an explanatory variable	slope often attenuated towards zero
Classical error in the outcome	more noise and larger standard errors
Differential measurement by treatment status	possible bias in less predictable directions

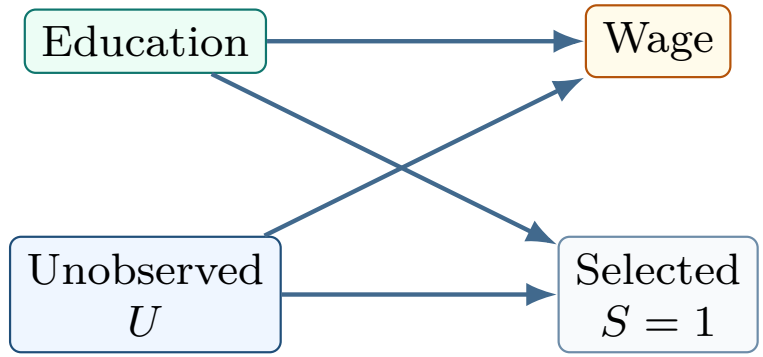
Show of hands

Which is worse for causal estimation: random outcome noise or outcome reporting that changes with treatment?

Selection Bias: Observed Samples

Example: estimate returns to education using wages, but wages are observed only when $S_i = 1$, where S_i means employed or observed in the wage sample.

- → The wage regression is estimated conditional on $S_i = 1$.
- → Education may affect selection into observed wages.
- → Unobserved constraints or motivation may affect both selection and wages.
- → Conditioning on selected observations can bias the comparison.



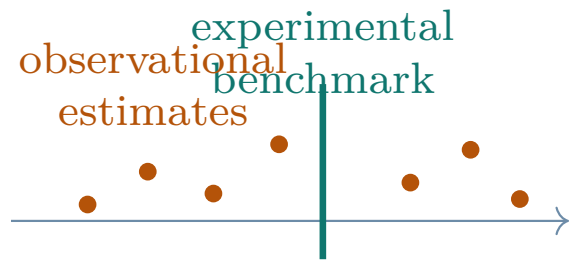
Thought experiment

What changes if we could observe potential wages for everyone, not only those with $S = 1$?

Historical Example: LaLonde's Training Benchmark

LaLonde (1986): experimental training data provide a benchmark for observational methods.

- → Treatment: participation in the programme.
- → Outcome: post-training earnings.
- → Benchmark: randomised treatment vs control.



Lesson: Regression adjustment is not automatically causal; the comparison must be credible.

What Can We Do About These Biases?

There is no automatic statistical fix for a weak comparison. The solution depends on the source of bias.

Threat	Typical response
Omitted variables	Controls, fixed effects, research design
Reverse causality	Timing, instruments, natural experiments
Measurement error	Validation data, repeated measures
Selection bias	Model or redesign the comparison

Bridge to the course

For each later method, ask: which comparison does it create, and which bias is it trying to remove?

Which Bias Is It?

For each case, name the main threat to causal interpretation.

1. Cities with more police officers have more crime.
2. Self-reported income contains substantial reporting noise.
3. We estimate returns to education only among employed workers.
4. Students who attend tutoring obtain higher grades, but struggling students are more likely to seek tutoring.
5. We estimate the effect of class size on grades but omit prior achievement, which affects placement and later grades.

Possible answers

Reverse causality; measurement error; sample selection; selection or omitted variables; omitted-variable bias.

Reusable Empirical Checklist

1. What is the treatment?
2. What is the outcome, and over what horizon?
3. What is the comparison group?
4. What is the missing counterfactual?
5. What creates the identifying variation?
6. What assumption makes that variation causal?
7. What could still bias the estimate?
8. To what population, setting, or policy context does the result apply?

External validity

External validity concerns whether an estimated effect generalises or transports to the target population, setting, or policy context.

Preview of Identification Strategies

Strategy	Variation	Intuition	Main assumption
RCT	randomised assignment	randomisation makes treatment groups comparable	assignment is random and follow-up is comparable
IV	instrument Z shifts treatment D	use exogenous variation in treatment	relevance $Z \to D$; exogeneity and exclusion for Y
DiD	policy changes over time across groups	compare changes, not levels	parallel untreated trends
RDD	cutoff rule assigns treatment near a threshold	compare units just around the cutoff	continuity of potential outcomes at the cutoff

Preview question

Which strategy creates variation by design, which uses timing, and which uses a threshold?

Historical Example: Card and Krueger's Minimum Wage Study

Card and Krueger (1994)

New Jersey raised its minimum wage from \$4.25 to \$5.05 in 1992; Pennsylvania did not.

- → Same sector: fast-food restaurants.
- → Two states.
- → Before and after the policy.
- → Identification idea: compare changes.

	Before	After
New Jersey	\$4.25	\$5.05
Pennsylvania	\$4.25	\$4.25

$$\widehat{\tau}_{DiD} = (\bar{Y}_{NJ,after} - \bar{Y}_{NJ,before}) - (\bar{Y}_{PA,after} - \bar{Y}_{PA,before}).$$

Question

What must be true about Pennsylvania for it to be a credible comparison?

Five Key Distinctions

Distinction	Why it matters
Description vs modelling	summary is not explanation
Prediction vs causation	useful forecasts need not be causal
Estimate vs parameter	data number vs target object
Estimate vs inference	magnitude vs uncertainty
Estimation vs identification	computation vs causal interpretation

One sentence

Applied econometrics is disciplined comparison, uncertainty, and causal reasoning.

Connection to the Rest of the Course

- → Session 1 formalises counterfactuals and potential outcomes.
- → Session 2 studies regression, matching, and selection on observables.
- → Sessions 3–4 use experiments, compliance, IV, and LATE.
- → Sessions 5–7 use time variation and threshold rules.
- → Session 8 synthesises method choice, robustness, and replication.

Course habit

For every empirical result, ask: estimate, inference, identification.