

Econometrics Learnbook for Dummies

Plain-language notes on the seven lectures: visual explanations, step-by-step reasoning, practical examples, R labs with real output, practice at three levels, two mixed exam papers and full answers. Interactive companion: sweetheartapi.github.io/learn_econometrics

CONTENTS

- | | |
|-----------------------------------|---|
| 1. How to use this learnbook | 8. Lecture 5 Instrumental variables |
| 2. Basics you need | 9. Lecture 6 Regression discontinuity |
| 3. How to read econometric output | 10. Lecture 7 Difference-in-differences |
| 4. Lecture 1 Regression basics | 11. Putting it together |
| 5. Lecture 2 Maximum likelihood | 12. Practice exam papers |
| 6. Lecture 3 The bootstrap | 13. Formula sheet |
| 7. Lecture 4 Intro to causality | 14. Answer key |

October 2026 · Based on the lecture slides, *Mastering 'Metrics* (Angrist & Pischke) and *Linear Models with R* (Faraway). R labs use simulated data with fixed seeds; output produced with R 4.3.3.

How to use this learnbook

A suggested routine: one chapter per sitting, in short steps of about 10–15 minutes, with breaks in between. Every chapter follows the same five steps, so it is always clear where one is:

1. **Understand.** Skim the "3 key ideas" box and the learning goals, then read the explanation, the "Why does this work?" boxes and the "In practice" examples, with a pen at hand. Tricky ideas come with **three ways to explain** them (simple, intuitive, technical); any one of the three can be enough.
2. **See it.** Each chapter points to an interactive demo on the companion website (sweetheartapi.github.io/learn_econometrics): moving a slider shows why a result changes.
3. **Apply.** The R lab and the Level 2 (calculation) questions turn the ideas into numbers.
4. **Challenge.** "Spot the mistake" and Level 3 questions hide a broken assumption or a wrong conclusion. The "method in five questions" box at the end of each chapter is a good summary to check against.
5. **Review.** Writing the 3 key ideas from memory on a blank sheet a few days later, then checking, shows what has stuck. The two **practice exam papers** near the end mix all chapters, closed-book.

Learning goals are marked **FOUNDATIONAL** (needed for everything after), **INTERMEDIATE** (typical exam level) or **ADVANCED** (for full marks and further study).

Box guide used throughout:

Box	What it holds
Where this fits	How the chapter connects to the previous and next lectures
3 key ideas	The minimum to remember
Learning goals	What the chapter should make possible, by difficulty
Three ways to explain	One idea at three depths: simple, intuitive, technical
Why does this work?	The reasoning between two formulas, one step at a time
In practice	A concrete example right after the theory
Careful	A common misunderstanding, corrected
R lab	Code, real output and how to read it
Spot the mistake	Wrong interpretations to correct
The method in five questions	Problem, solution, assumptions, what breaks, how to check

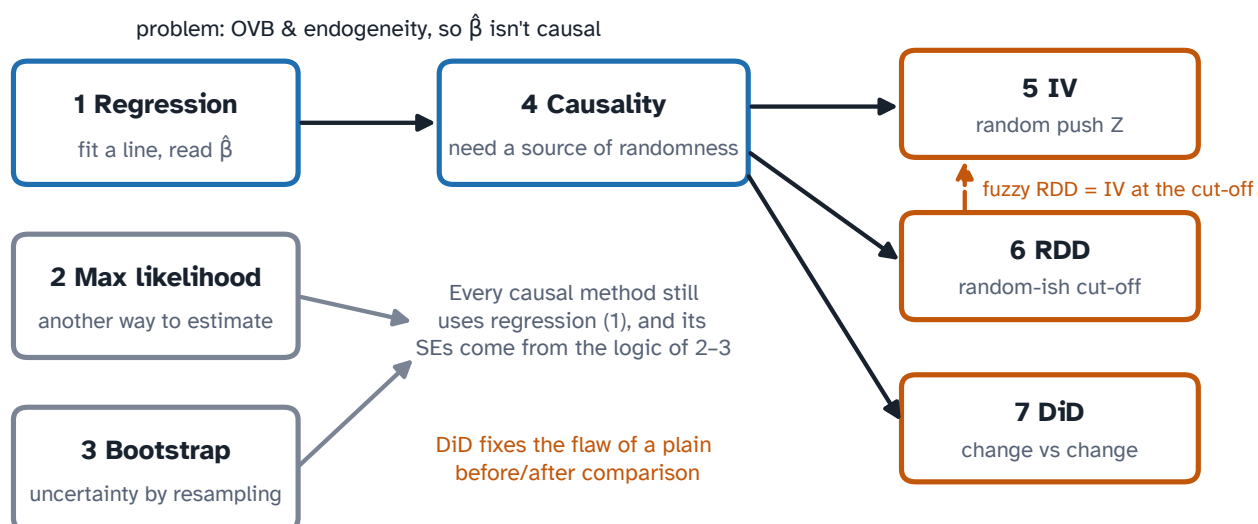
Printing tip: single-sided printing leaves the back of each page free for notes. The Basics chapter and the formula sheet work well as reference sheets alongside the other chapters.

Progress tracker

Ticking boxes is not the goal, but it shows at a glance which chapters need another look. "Practised" here means the chapter's practice questions were attempted; "mastered" means about 80 % right on the exam papers. Short reviews spaced out over time (a couple of days, a week, a month) do more for memory than one long reread.

Chapter	Studied	Practised	Mastered	Review 1 (+2 days)	Review 2 (+1 week)	Review 3 (+1 month)
1 Regression	☐	☐	☐	☐ ____	☐ ____	☐ ____
2 Maximum likelihood	☐	☐	☐	☐ ____	☐ ____	☐ ____
3 Bootstrap	☐	☐	☐	☐ ____	☐ ____	☐ ____
4 Causality	☐	☐	☐	☐ ____	☐ ____	☐ ____
5 Instrumental variables	☐	☐	☐	☐ ____	☐ ____	☐ ____
6 Regression discontinuity	☐	☐	☐	☐ ____	☐ ____	☐ ____
7 Difference-in-differences	☐	☐	☐	☐ ____	☐ ____	☐ ____

How the lectures connect



The course as one story. Regression (1) is the workhorse. Its big weakness, that $\hat{\beta}$ is not causal when something important is left out, motivates the causal part (4). Lectures 5-7 are three ways to find "as good as random" variation. Lectures 2 and 3 supply tools for estimating and measuring uncertainty that all the others use.

- **1 → 4:** Regression describes associations. Omitted variable bias and endogeneity show why a regression slope is usually not a causal effect. That gap is what Lecture 4 is about.
- **4 → 5, 6, 7:** Lecture 4 says a causal answer needs a source of randomness. IV finds it in an outside push (a lottery), RDD in a cut-off rule, DiD in a policy that hits some groups and not others.
- **5 ↔ 6:** A fuzzy RDD is literally an IV estimate at the cut-off: eligibility is the instrument.
- **4 → 7:** DiD fixes the main flaw of a plain before/after comparison: other things change over time too.
- **2 and 3 everywhere:** Logit models and many estimators are fitted by maximum likelihood; the bootstrap gives standard errors when no formula is available.

Which book goes with which chapter:

Chapter	Mastering 'Metrics (Angrist & Pischke)	Linear Models with R (Faraway)
1 Regression	ch. 2	ch. 2-4, §5.3, §6.1-6.2, ch. 8
2 Maximum likelihood	—	ch. 2 (OLS = MLE link only)
3 Bootstrap	—	§3.3 (permutation tests)
4 Causality	ch. 1	—
5 Instrumental variables	ch. 3, §6.3	—
6 Regression discontinuity	ch. 4	—
7 Difference-in-differences	ch. 5	—

Basics you need

These are the basic terms and notation every lecture assumes. This chapter serves as a reference for any word or symbol that feels unclear.

How to read an equation

Most equations in this course are built from a handful of pieces. Reading them piece by piece makes even long formulas manageable.

Piece	Example	How to read it
Subscript	y_i, x_{it}	"y for person i", "the 2nd x-variable for person i". Subscripts say <i>whose</i> or <i>which</i> value. A second subscript often means time: y_{it} = person i in period t.
Superscript	x^2, σ^2	A power: x squared, the variance σ^2 . Careful: X^T is not a power, it means "transpose", and X^{-1} means "inverse".
Hat	$\hat{\beta}, \hat{y}$	An estimate from the data, as opposed to the true unknown value (β, y).
Bar	$\bar{x}, \bar{y}$	A sample average.
Tilde or star	$\tilde{x}, \beta^*$	A modified version: demeaned x, the "true" coefficient in a longer model, a bootstrap value. Always defined nearby.
Sum	$\sum_i x_i$	Add x_i over all i: $x_1 + x_2 + \dots + x_n$.
Product	$\prod_i f(x_i)$	Multiply over all i. Appears in likelihoods.
Expectation	$E[Y]$	The population average of Y.
Conditional expectation	$E[Y X = x]$	The population average of Y among units with $X = x$. The bar " " reads "given".
Distributed as	$\varepsilon \sim N(0, \sigma^2)$	" ε follows a normal distribution with mean 0 and variance σ^2 ."
Converges	$\hat{\theta} \rightarrow^p \theta$	" $\hat{\theta}$ gets arbitrarily close to θ as the sample grows" (consistency).
Indicator	$1\{X \geq c\}$	1 if the condition is true, 0 if not.
argmax	$\operatorname{argmax}_{\theta} \ell(\theta)$	The value of θ that makes ℓ biggest (not the biggest value itself).

In practice: reading one line. $E[Y_i(1) - Y_i(0) | D_i = 1]$ reads: "the average, among the treated ($D = 1$), of the difference between each person's outcome with treatment and without it". That is the ATT from Lecture 4.

Percent vs percentage points

A change in a share or a rate can be described in two ways, and mixing them up is a classic exam error.

- **Percentage points (pp):** the plain difference between two percentages. Employment 40 % → 45 % is **+5 pp**.
- **Percent (%):** the change relative to the starting value. 40 % → 45 % is $(45 - 40)/40 = \mathbf{+12.5\%}$.

In practice. In the Austrian parental-leave study, the chance of another birth rose by 4.5 percentage points, from a mean of 34.5 %. In percent, that is $4.5/34.5 \approx +13\%$.

Logs and percent. For small changes, a change in $\log y$ of 0.05 $\approx$ a 5 % change in y . The approximation gets worse for big changes: a log change of 0.5 is really $e^{0.5} - 1 \approx 65\%$, not 50 %.

Data and describing data

Word	Plain meaning
Observation	One row of data: one person, firm or country.
Variable	One column of data: something measured for each observation.
Population vs sample	Population = everyone you care about. Sample = the part actually observed.
Parameter	A fixed, unknown number about the population (true mean μ , true slope β).
Dummy variable	A variable that is only 1 or 0 (treated or not, female or not).
Panel data	The same units observed at several points in time.
Mean, $\bar{x}$	Add everything up, divide by how many.
Median	The middle value when sorted. Not pulled around by extreme values.
Variance	Average squared distance from the mean: how spread out the data are.
Standard deviation (SD, σ)	Square root of the variance: spread in the original units.
Covariance	Do two variables move together (+) or in opposite directions (-)?
Correlation	Covariance rescaled to lie between -1 and +1.
Percentile	The value below which that % of the data falls.
Demeaning	Subtracting the mean, so the new variable averages zero.

Probability

Word	Plain meaning
Random variable	A number whose value depends on chance (a die roll).
Distribution	All possible values and how likely each is.
Probability (discrete data)	For countable outcomes (0, 1, 2 heads), $P(Y = 3)$ is a real probability between 0 and 1.
Density (continuous data)	For continuous outcomes (a waiting time), the chance of any <i>exact</i> value is 0. The density $f(y)$ is the height of the curve; probabilities are areas under it. A density can be larger than 1.
CDF	$F(x)$ = probability of being at most x . Goes from 0 to 1.
Expectation, $E[X]$	The long-run average of a random variable.
Conditional expectation, $E[Y X]$	The average of Y among units with a given X .
Independent ($\perp$)	Knowing one tells you nothing about the other.
i.i.d.	Independent and identically distributed: separate draws from the same distribution.
Normal, $N(\mu, \sigma^2)$	The bell curve. About 95 % of values lie within ± 1.96 SD of the mean.
Binomial	Number of successes in n independent yes/no trials.
Law of Large Numbers	With i.i.d. data and a finite mean, the sample mean gets closer to the true mean as n grows.
Central Limit Theorem	With i.i.d. data and a finite variance , the sample mean is approximately normal in large samples. "Large" depends on the data: very skewed or heavy-tailed data need far more observations, and with infinite variance the standard CLT fails. Dependent data need other versions.

Estimation and testing

Word	Plain meaning
Estimator vs estimate	Estimator = the recipe. Estimate = the number one sample gives.
Bias	$E[\hat{\theta}] - \theta$: how far off the estimator is on average. Unbiased = right on average.
Variance of an estimator	How much the estimate jumps around from sample to sample.
Standard error (SE)	The estimated standard deviation of an estimator across repeated samples.
SD vs SE	SD = spread of the data. SE = uncertainty of an estimate. SE shrinks with more data; SD doesn't.
Consistent	Gets to the true value as the sample grows.
Efficient	Smallest variance within a class of estimators.
Residual vs error	Error ε = true unobserved deviation. Residual $\hat{\varepsilon}$ = observed deviation from the fitted line.
Null hypothesis, H_0	The claim tested against, usually "no effect".
Test statistic	A number from the data used to judge H_0 , e.g. $t = \hat{\beta} / SE$.
p-value	If the model and H_0 were true, the chance of a result at least this extreme. Not the chance that H_0 is true.
Significance level, α	Cut-off for rejecting H_0 , usually 5 %.
Type I / II error	Type I: false alarm (reject a true H_0). Type II: missed effect.
Power	Chance of detecting an effect that is really there.
Confidence interval	A range made by a method that covers the true value in 95 % of repeated samples.
Degrees of freedom	Observations minus estimated parameters ($n - p$).

Math tools

Symbol or word	Plain meaning
log	Natural log. $\log(a \cdot b) = \log a + \log b$; $\log(a^b) = b \cdot \log a$.
exp, e	$e \approx 2.718$; $\exp(\log x) = x$.
Derivative	The slope of a curve. It is 0 at a maximum or minimum, which is how maximums are found.
Second derivative	How fast the slope changes: the curvature. Negative at a maximum.
X^T, X^{-1}	Transpose (flip rows and columns); inverse (the matrix version of dividing).
$\perp$	Orthogonal: perpendicular, uncorrelated.
Greek letters	$\alpha, \beta, \gamma, \delta$ = coefficients or effects; ε, η = errors; θ = generic parameter; λ = a rate or mean; μ = mean; σ = SD; τ = treatment effect; π = first-stage coefficient in IV; ρ, ϕ = reduced form and first stage.

How to read econometric output

Every regression table answers the same five questions: how big, how precise, how surprising under "no effect", what range is plausible, and does the size matter. This chapter walks through one real R output line by line.

The data are simulated wages for 500 workers (the same data as the Lecture 1 lab): log wage, years of education and years of experience.

R CODE • LAB0.R

```
# Reading a regression table (same simulated wage data as Lab 1)
set.seed(1)
n <- 500
ability <- rnorm(n)
educ <- round(12 + 1.5 * ability + rnorm(n, 0, 2))
exper <- round(runif(n, 0, 30))
logwage <- 1 + 0.08 * educ + 0.02 * exper + 0.10 * ability +
  rnorm(n, 0, 0.15 + 0.01 * exper)
m <- lm(logwage ~ educ + exper)
summary(m)
confint(m)
```

OUTPUT

Call:

```
lm(formula = logwage ~ educ + exper)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.2850	-0.1863	0.0006	0.2086	1.3007

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.68394	0.07723	8.86	<2e-16 ***
educ	0.10809	0.00594	18.20	<2e-16 ***
exper	0.01886	0.00173	10.93	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.341 on 497 degrees of freedom

Multiple R-squared: 0.471, Adjusted R-squared: 0.469

F-statistic: 221 on 2 and 497 DF, p-value: <2e-16

	2.5 %	97.5 %
(Intercept)	0.53220	0.83568
educ	0.09642	0.11975
exper	0.01547	0.02225

Reading the table, piece by piece

Piece of output	Value for educ	What it means
Estimate	0.108	Holding experience fixed, one more year of education is associated with about 10.8 % higher wages (log-level: $100 \cdot \beta$ %). The exact figure is $e^{0.108} - 1 \approx 11.4$ %.
Std. Error	0.00594	How much this estimate would vary across repeated samples. Small relative to the estimate = precise.
t value	18.20	Estimate $\div$ SE = $0.108 / 0.00594$. The estimate is 18 standard errors away from zero.
Pr(> t)	< 2e-16	p-value for $H_0: \beta = 0$. If education had no association with wages (and the model were right), a t this large would almost never happen.
2.5 % / 97.5 % (confint)	0.096 to 0.120	95 % confidence interval: the values of β compatible with the data, roughly estimate $\pm 2 \cdot \text{SE}$.
Stars	***	Shorthand for $p < 0.001$. They add nothing the p-value doesn't already say.

Model-level line	Value	What it means
Residual standard error	0.341	Typical size of a residual, in log points (about 34 %).
Multiple R-squared	0.471	The model explains 47 % of the variation in log wages <i>in this sample</i> .
F-statistic	221 on 2 and 497 DF	Joint test that education and experience both have zero coefficients. Rejected.

Statistical vs economic significance

The t-statistic says the education coefficient is very precisely different from zero. Whether it **matters** is a separate question answered by its size: about 11 % per year of school is economically large. A coefficient of 0.001 (0.1 % per year) could have an even bigger t in a huge sample and still be economically trivial.

Precise is not the same as correct. This table is very precise, but in the Lecture 1 lab the true effect of education is 0.08, not 0.108. Ability was left out, and that biases the estimate (omitted variable bias). A tiny standard error measures sampling noise only, never bias.

Checklist for any coefficient

1. **Sign and size** in real units (use the log table in Lecture 1).
2. **"Holding what fixed?"** List the other variables in the model.
3. **Precision:** SE and the confidence interval.
4. **Significance:** t and p, read correctly.
5. **Economic importance:** is the size meaningful?
6. **Causal or not?** What would have to be true for this to be a causal effect (Lectures 4–7)?

Regression basics

Where this fits. Regression is the workhorse of the whole course: IV, RDD and DiD are all regressions in the end. This chapter shows how the line is fitted and read. It ends with the problem that drives the second half of the course: a regression slope is usually **not** a causal effect.

3 key ideas

1. OLS picks the line that makes the **sum of squared residuals** as small as possible. The slope tells how y changes when x rises by 1, holding the other x 's fixed.
2. A regression shows **association, not causation**. Leaving out a factor that affects y and is linked to x biases the slope.
3. p-values and confidence intervals **assume the model is right**. They measure sampling uncertainty only, never bias.

After this chapter it should be possible to...

- FOUNDATIONAL** Read a slope correctly, including log forms and "holding the others fixed".
- FOUNDATIONAL** Say what R^2 measures, and when it can mislead (or go negative).
- INTERMEDIATE** Use the omitted-variable-bias formula to predict the sign and size of a bias.
- INTERMEDIATE** Choose between usual, robust and clustered standard errors.
- INTERMEDIATE** Interpret p-values and confidence intervals without the classic mistakes.
- ADVANCED** Derive $\hat{\beta} = (X^T X)^{-1} X^T y$ and read OLS as a projection.

Interactive version: the regression playground at sweetheartapi.github.io/learn_econometrics/#l1/l1-demo

Prediction or explanation?

Regression does two different jobs. **Prediction** asks "what will happen?". The test of a good predictive model is how well it predicts **new** data (out-of-sample accuracy), not just its R^2 in the data used to fit it.

Explanation asks "why?" and needs an unbiased slope; a low R^2 is not a problem in itself.

In practice. An energy company forecasting tomorrow's electricity demand judges its model by how close yesterday's forecasts came to actual demand, on days the model never saw. A study of whether a tutoring programme raises grades can have $R^2 = 0.05$ and still be useful: grades depend on many things, but the question is only whether the programme's coefficient is right.

The model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

- y = outcome. $x_1 \dots x_p$ = things that might explain it (regressors).
- β_0 = intercept: predicted y when all x 's are 0.
- β_1 = slope: change in y when x_1 rises by 1 **and the other x 's stay the same**.
- ε = error: everything not measured.
- In matrix form: $y = X\beta + \varepsilon$, where X is the data table plus a column of 1s.

A model is like a map: useful because it leaves things out. "The purpose of the model is to be useful, not to be correct."

In practice. Flat prices: $\text{price} = \beta_0 + \beta_1 \cdot \text{size} + \beta_2 \cdot \text{distance to centre} + \varepsilon$. If $\hat{\beta}_1 = 2,000$ € per m^2 , then of two flats the same distance from the centre, the one 10 m^2 bigger is predicted to cost about 20,000 € more. The view, floor and renovation sit in ε .

How the line is chosen: least squares

A **residual** is the gap between a real point and the fitted line. OLS squares every residual, adds them up, and picks the coefficients that make the total smallest.

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \text{one regressor:} \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

Why does this work? From "minimise the squares" to Cov/Var, one regressor

1. Write the total squared mistake for a candidate line $a + b \cdot x$: $S(a, b) = \sum (y_i - a - b \cdot x_i)^2$.
2. At the minimum, the slope of S in every direction is zero. Derivative with respect to a : $-2 \cdot \sum (y_i - a - b \cdot x_i) = 0$. Dividing by $-2n$ gives $\bar{y} - a - b \cdot \bar{x} = 0$, so $a = \bar{y} - b \cdot \bar{x}$. The line always passes through the point of means $(\bar{x}, \bar{y})$.
3. Derivative with respect to b : $-2 \cdot \sum x_i (y_i - a - b \cdot x_i) = 0$, so $\sum x_i (y_i - a - b \cdot x_i) = 0$.
4. Plug in a from step 2: $\sum x_i [(y_i - \bar{y}) - b(x_i - \bar{x})] = 0$, so $\sum x_i (y_i - \bar{y}) = b \cdot \sum x_i (x_i - \bar{x})$.
5. Because $\sum (y_i - \bar{y}) = 0$ and $\sum (x_i - \bar{x}) = 0$, x_i can be replaced by $(x_i - \bar{x})$ on both sides without changing anything. That gives $b = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_i - \bar{x})^2$.
6. With many regressors the same two conditions become $X^T X \hat{\beta} = X^T y$ (the "normal equations"). In words: **the residuals are uncorrelated with every regressor and average zero**. Solving gives $\hat{\beta} = (X^T X)^{-1} X^T y$.

In practice: OLS by hand. Three students studied 1, 2 and 3 hours and scored 50, 60 and 64 points.

1. Means: $\bar{x} = 2$ hours, $\bar{y} = 58$ points.
2. Deviations from the means: x : -1, 0, +1; y : -8, +2, +6.
3. Top: $(-1)(-8) + 0 \cdot 2 + 1 \cdot 6 = 14$. Bottom: $1 + 0 + 1 = 2$.
4. Slope $b = 14 / 2 = 7$. Intercept $a = 58 - 7 \cdot 2 = 44$.
5. Fitted line: score = 44 + 7·hours. Each extra hour is associated with 7 more points.

Geometry. The fitted values $\hat{y}$ are the "shadow" (projection) of y onto the columns of X ; the residuals are perpendicular to them.

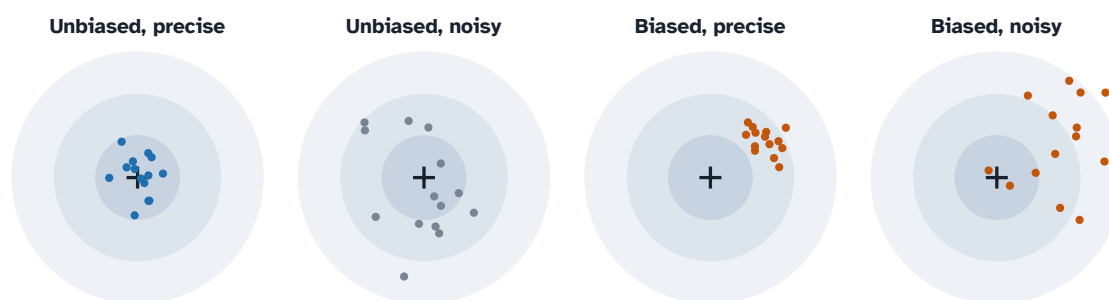
The OLS assumptions

Different properties need different assumptions. Each row adds to the ones above it.

#	Assumption	Plain meaning	What it buys
1	Linear in parameters	$y = X\beta + \varepsilon$ (x's can be logs, squares...)	—
2	Random sampling	Observations are independent draws from the population (otherwise, a model of the dependence, e.g. clusters)	—
3	No perfect multicollinearity	No regressor is an exact linear combination of others; $X^T X$ can be inverted	$\hat{\beta}$ can be computed
4	Zero conditional mean	$E[\varepsilon X] = 0$: the error is unrelated to the regressors (exogeneity)	1–4: $\hat{\beta}$ is unbiased (and consistent)
5	Homoskedasticity, no autocorrelation	$\text{Var}(\varepsilon X) = \sigma^2 I$: same error spread for everyone, errors uncorrelated	1–5: Gauss–Markov: OLS is BLUE , usual SEs correct
6	Normal errors	$\varepsilon X \sim N(0, \sigma^2 I)$	1–6: exact t and F tests in small samples

BLUE = Best Linear Unbiased Estimator: the lowest variance among all linear unbiased estimators. Normality (6) is **not** needed for BLUE; with large samples, t-tests work approximately without it.

Bias and variance



+ = true β . Each dot = the estimate from one sample. Bias = how far the cloud's centre is from +; variance = how spread out the cloud is.

Two separate ways an estimator can be bad. Bias is about being off-target on average (assumption 4 fails, e.g. an omitted variable). Variance is about noise (small samples, collinearity). A precise estimate can still be badly biased.

Multiple regression: "holding other things fixed"

With several regressors, each coefficient answers: what happens to y when this x changes **and the others don't**?

Worked example (from the R lab). Regressing log wage on education and experience gives $\hat{\beta}_{\text{educ}} = 0.108$. Among workers with the **same experience**, one more year of education is associated with about 11 % higher wages.

Why does this work? The Frisch-Waugh-Lovell (FWL) idea

How does the regression "hold experience fixed"? FWL gives an intuitive answer:

1. Regress education on experience and keep the residuals: the part of education that experience **cannot** explain.
2. Regress log wage on those residuals.
3. The slope is **exactly** the multiple-regression coefficient on education (0.1081 in the lab).

So a multiple-regression coefficient uses only the variation in x that is left after removing what the other regressors explain. This is the same "residualise, then use the simple formula" logic used for IV with controls in Lecture 5.

R^2

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS = squared residuals left over. TSS = total spread of y around its mean. For ordinary OLS **with an intercept**, judged on the data used to fit it, R^2 runs from 0 to 1: the share of y 's variation explained **in the sample**. Without an intercept, or on new data, RSS can exceed TSS and R^2 turns negative, which simply means the model predicts worse than the plain mean $\bar{y}$. Very different datasets can share the same R^2 , so plotting the data is always worth it.

R^2 and prediction. A high in-sample R^2 does not guarantee good predictions: adding variables always raises R^2 , even pure noise, and the model can overfit. Prediction quality is judged on new data. A low R^2 does not make a predictive model useless either; it may still beat every alternative.

Reading coefficients: the log table

Model	x goes up by	y changes by about
y on x (level–level)	1 unit	β_1 units
log y on x (log–level)	1 unit	$100 \cdot \beta_1$ %
y on log x (level–log)	1 %	$\beta_1/100$ units
log y on log x (log–log)	1 %	β_1 % (an elasticity)

Worked example. $\log(\text{wage}) = \beta_0 + 0.07 \cdot \text{education} + \beta_2 \cdot \text{age}$. Correct sentence: "For people of the same age, one more year of education is **associated with** about 7 % higher wages."

Interaction. In $\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{south} + \beta_3 \text{south} \cdot \text{educ}$, the education slope is β_1 outside the South and $\beta_1 + \beta_3$ in the South.

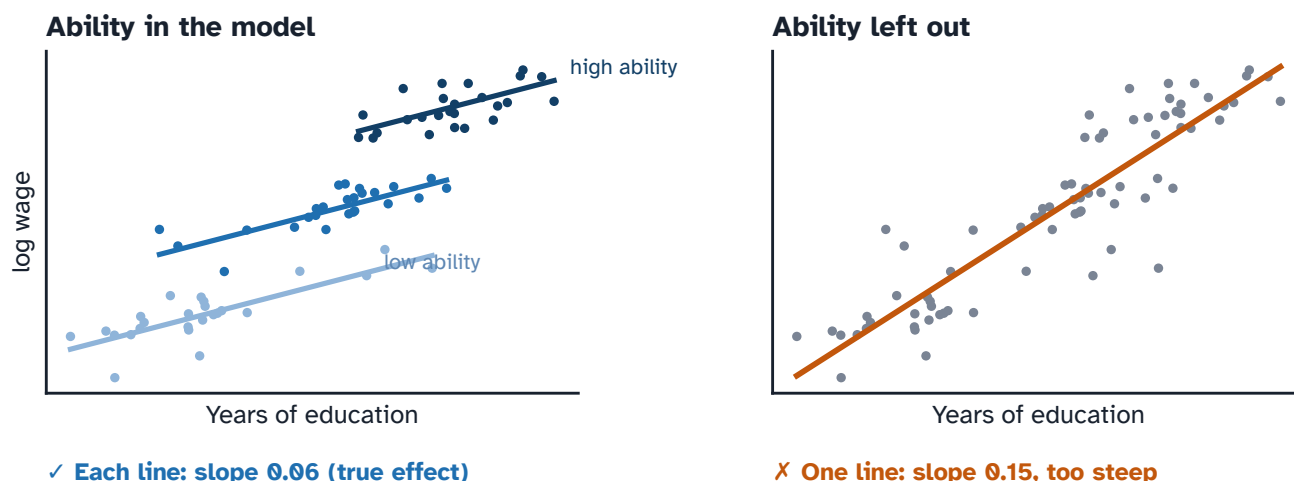
Testing and intervals

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \sim t_{n-p}$$

- Tests $H_0: \beta_i = 0$. Large $|t|$ (above about 1.96 in big samples) means significant at 5 %.
- **F-test:** do several variables together add anything? Compares the RSS of a big and a small model.
- **Confidence interval:** $\hat{\beta} \pm t \cdot se$. The parameter is fixed; the interval is the random thing.
- **Prediction interval** for one new y is wider than the interval for the average y: it adds the new observation's own error (the "1 +" under the square root).
- **t-test = regression, if coded right.** Regressing y on an intercept and *one* dummy (1 = group B) gives group A's mean as the intercept and the difference in means as the slope; its ordinary t-statistic is exactly the **pooled-variance** two-sample t-test. With unequal spreads the matching test is **Welch's**, which corresponds to the same regression with robust (HC2) standard errors. (An intercept plus two dummies would be perfectly collinear.)

In practice. A café chain tests whether customers with its loyalty app spend more per visit: $\hat{\beta} = 1.20$ €, $se = 0.50$ €. Then $t = 2.4 > 1.96$, significant at 5 %. The 95 % CI is $1.20 \pm 1.96 \cdot 0.50 = [0.22 \text{ €}, 2.18 \text{ €}]$: the effect could be small or fairly large.

What can go wrong 1: omitted variable bias



Same data, two models. Within each ability group the education slope is 0.06. Leaving ability out mixes the groups: high-ability people have both more education and higher wages, so the single line is far too steep.

$$\beta_1^{\text{short}} = \beta_1^* + \beta_3^* \gamma_1$$

β_1 = true effect of education; β_3 = effect of ability on wage; γ_1 = how ability moves with education.

Why does this work? Where the bias formula comes from

1. True model: wage = $\beta_0 + \beta_1 \cdot \text{educ} + \beta_3 \cdot \text{ability} + \varepsilon$.
2. Ability is linked to education: ability = $\gamma_0 + \gamma_1 \cdot \text{educ} + \varepsilon'$.
3. Substitute step 2 into step 1: wage = $(\beta_0 + \beta_3 \gamma_0) + (\beta_1 + \beta_3 \gamma_1) \cdot \text{educ} + (\beta_3 \varepsilon' + \varepsilon)$.
4. A regression without ability estimates the bracket in bold. The extra piece $\beta_3 \gamma_1$ is the bias.
5. **No bias** only if $\beta_3 = 0$ (ability doesn't matter for wages) or $\gamma_1 = 0$ (ability unrelated to education).
6. **Direction of bias** = sign of $\beta_3^* \times$ sign of γ_1 . Here both are positive, so the short regression overstates the effect.

In practice. Ice-cream sales and drownings rise together; the omitted variable is hot weather, which raises both. Children with private tutors may also have more motivated parents, so a tutoring coefficient picks up part of the parents' effect.

What can go wrong 2: collinearity

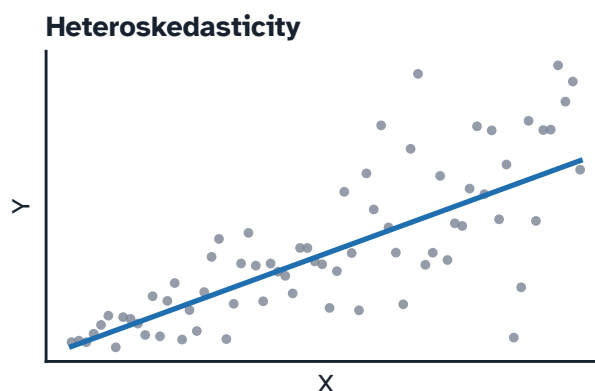
Highly correlated regressors make estimates unstable (high variance), not biased. Spot it with correlations near ± 1 , or with R_j^2 near 1 when regressing x_j on the other x 's. The variance of $\hat{\beta}_j$ is multiplied by $1/(1 - R_j^2)$ (the variance inflation factor).

In practice. Age and years of work experience move almost one-for-one for most workers. With both in the model, each gets a large standard error and can swing wildly if a few observations change.

What can go wrong 3: heteroskedasticity



✓ The spread of errors stays similar



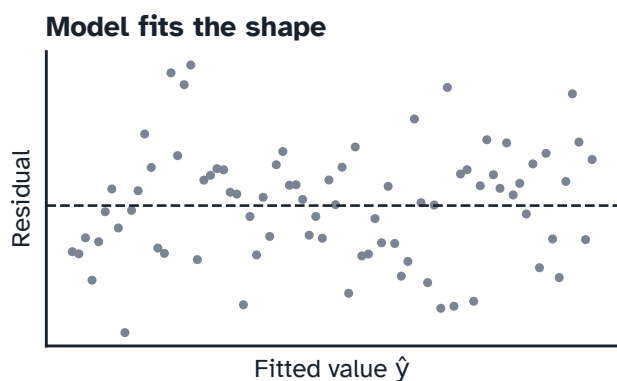
✗ The spread grows as X grows

Left: the errors have the same spread everywhere (assumption 5 holds). Right: the spread grows with X (heteroskedasticity).

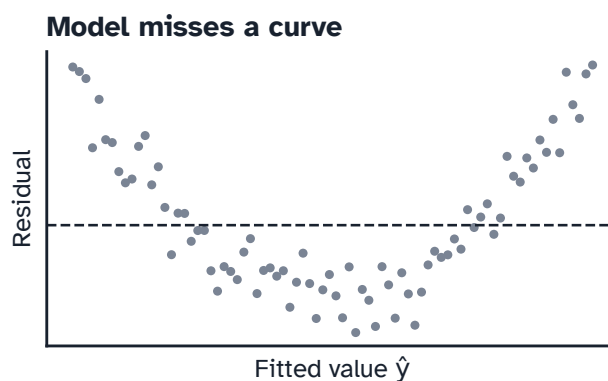
What to remember. Heteroskedasticity does **not** make OLS coefficients biased: with exogeneity (assumption 4) they stay unbiased and consistent. What goes wrong is the **conventional standard errors**, and with them t-tests and confidence intervals. Heteroskedasticity-robust standard errors (HC0–HC3) are the common fix; weighted least squares (weights = $1/\text{variance}$) is the alternative.

In practice. Household spending vs income: low-income households all spend close to their income, while high-income households range from big savers to big spenders. The errors fan out as income grows, so robust standard errors are the usual choice.

Residual plots are the everyday diagnostic. Plot the residuals against the fitted values (or against each x).



✓ Random cloud around zero



✗ U-shape: add x^2 or a log

A good model leaves a shapeless cloud around zero. A curve means the functional form is wrong (try x^2 , a log or an interaction). A fan shape means heteroskedasticity.

Clustering. Cluster standard errors when treatment is assigned to groups (schools, towns) or when groups were sampled first.

In practice. A reading programme is given to 10 of 20 schools (1,000 pupils). Pupils in the same school share teachers, so their results are not independent. There are really only 20 independent units: standard errors are clustered by school.

Three ways to explain: heteroskedasticity

Simple. The noise around the line is bigger in some places than in others.

Intuitive. Food spending against income: low-income households all spend similar amounts on food, while high-income households vary a lot (some eat out every day, some don't). The cloud of points fans out as income grows.

Technical. $\text{Var}(\varepsilon_i | x_i) = \sigma_i^2$ is not constant. If $E[\varepsilon | X] = 0$, OLS stays unbiased, but $\text{Var}(\hat{\beta} | X) = (X^T X)^{-1} X^T \Omega X (X^T X)^{-1} \neq \sigma^2 (X^T X)^{-1}$, so the usual SEs are wrong. HC estimators plug $\hat{\varepsilon}_i^2$ into Ω . OLS is no longer BLUE; GLS/WLS with the right weights is.

Choosing a model

- Prefer simpler models for explanation (Occam's razor); bigger ones can be fine for prediction.
- Automatic selection by p-values is risky: after many tests, results look better than reality.
- $\text{AIC} = -2 \cdot \log L + 2p$ and $\text{BIC} = -2 \cdot \log L + p \cdot \log n$ (p = number of parameters). Since $\log n > 2$ once $n \geq 8$, BIC penalises size more and prefers smaller models. Lower is better.
- One lecture slide writes the BIC penalty as $2p \cdot \log n$. That is not the standard definition; it is best read as a typo or a non-standard scaling, and $p \cdot \log n$ is the form used in textbooks and software.

What p-values and CIs really mean

Wrong: " $p = 0.01$ means H_0 has a 1 % chance of being true." Wrong: "this 95 % CI has a 95 % chance of containing the truth." Wrong: "overlapping CIs mean no significant difference."

Right: **assuming the model and H_0 are true**, p is how often a result at least this extreme would appear in repeated samples. A 95 % CI comes from a method that covers the truth in 95 % of repeated samples.

In practice. A drug trial reports $p = 0.03$. If the drug truly did nothing (and the model is right), a difference this large would appear in about 3 % of repeated trials. If the drug lowers blood pressure by only 0.5 mmHg, it can be "significant" in a huge trial while being medically pointless.

Three ways to explain: standard error

Simple. A standard error tells you how much your estimate would change if you collected a different sample of people.

Intuitive. Imagine surveying 100 students to estimate their average study time. If you repeatedly picked different groups of 100, their averages would differ a little each time. The standard error describes how much those averages typically vary.

Technical. The standard error is the estimated standard deviation of an estimator's sampling distribution. For an i.i.d. sample mean it is $s/\sqrt{n}$. For OLS, $se(\hat{\beta}_j) = \sqrt{[\hat{\sigma}^2(X^T X)^{-1}]_{jj}}$ under constant variance, or the square root of the sandwich $(X^T X)^{-1}(\sum \hat{\epsilon}_i^2 x_i x_i^T)(X^T X)^{-1}$ for robust SEs.

Three ways to explain: p-value

Simple. How surprising your result would be if there were really no effect, assuming your model is right. Small p = surprising.

Intuitive. A friend says a coin is fair. You flip it 20 times and get 16 heads. If the coin really were fair, a result at least that lopsided (16+ heads or 16+ tails) happens about 1.2 % of the time. So $p \approx 0.012$: hard to square with "fair", but not impossible.

Technical. $p = P(|T| \geq |t_{obs}| ; H_0 \text{ and the model})$, using T 's null distribution (e.g. t_{n-p}). If H_0 and the model hold, p is uniformly distributed on $[0, 1]$. It is not $P(H_0 | \text{data})$, and it says nothing about effect size.

The method in five questions

OLS regression in five questions. The question to ask every time: ***Why would my coefficient be biased?***

1. **What problem are we solving?** Describe how the average of Y changes with X, either to predict or to compare people who differ in X but are alike in the other regressors.
2. **How does the method solve it?** Pick the line (or plane) that makes the sum of squared residuals as small as possible: $\hat{\beta} = (X^T X)^{-1} X^T y$.
3. **What must hold?**
 - Linear in the β 's, no perfect collinearity.
 - Exogeneity, $E[\varepsilon | X] = 0$: nothing left in ε is related to X. Needed for unbiased (and causal) slopes.
 - Constant variance and independent errors: needed for the *usual* standard errors.
 - Normal errors: only for exact small-sample t and F tests.
4. **What happens when it fails?**
 - Omitted variable or reverse causality: $\hat{\beta}$ is biased, and more data doesn't help.
 - Heteroskedasticity or clustering: $\hat{\beta}$ is fine, the usual SEs are wrong.
 - Collinearity: $\hat{\beta}$ is unbiased but very noisy.
5. **How can it be checked?**
 - Plot residuals against fitted values and each x.
 - See whether the coefficient moves when plausible confounders are added.
 - Use robust (or clustered) SEs by default.
 - Ask what is missing from the model, and in which direction it would push $\hat{\beta}$.

R lab 1 • Omitted ability, robust SEs and FWL

The script simulates 500 workers where ability raises both education and wages (true education effect = 0.08). Packages: sandwich, lmtest, ggplot2.

R CODE · LAB1.R

```
# Lab 1: wages, education and the omitted "ability"
library(sandwich); library(lmtest); library(ggplot2)
set.seed(1)
n ← 500
ability ← rnorm(n) # unobserved in real life
educ ← round(12 + 1.5 * ability + rnorm(n, 0, 2)) # able people study longer
exper ← round(runif(n, 0, 30))
e ← rnorm(n, 0, 0.15 + 0.01 * exper) # noise grows with experience
logwage ← 1 + 0.08 * educ + 0.02 * exper + 0.10 * ability + e
d ← data.frame(logwage, educ, exper, ability)

long ← lm(logwage ~ educ + exper + ability, data = d) # if ability were measured
short ← lm(logwage ~ educ + exper, data = d) # what we can actually run
round(rbind(long = coef(long)[2:3], short = coef(short)[2:3]), 3)

# Omitted variable bias formula: beta3 * gamma1
gamma1 ← coef(lm(ability ~ educ + exper, data = d))["educ"]
round(c(gap = unname(coef(short)["educ"] - coef(long)["educ"]),
        formula = unname(coef(long)["ability"] * gamma1)), 4)

# Robust (heteroskedasticity-consistent) standard errors
coeftest(short, vcov = vcovHC(short, type = "HC1"))

# Frisch-Waugh-Lovell: remove exper from educ, then a simple regression
educ_resid ← resid(lm(educ ~ exper, data = d))
round(coef(lm(logwage ~ educ_resid, data = d))[2], 4)

# Residual plot: look for curves or a fan shape
ggplot(data.frame(fitted = fitted(short), resid = resid(short), exper = d$exper),
       aes(exper, resid)) +
  geom_point(alpha = 0.5) + geom_hline(yintercept = 0, linetype = 2) +
  labs(x = "Experience (years)", y = "Residual") + theme_minimal()
```

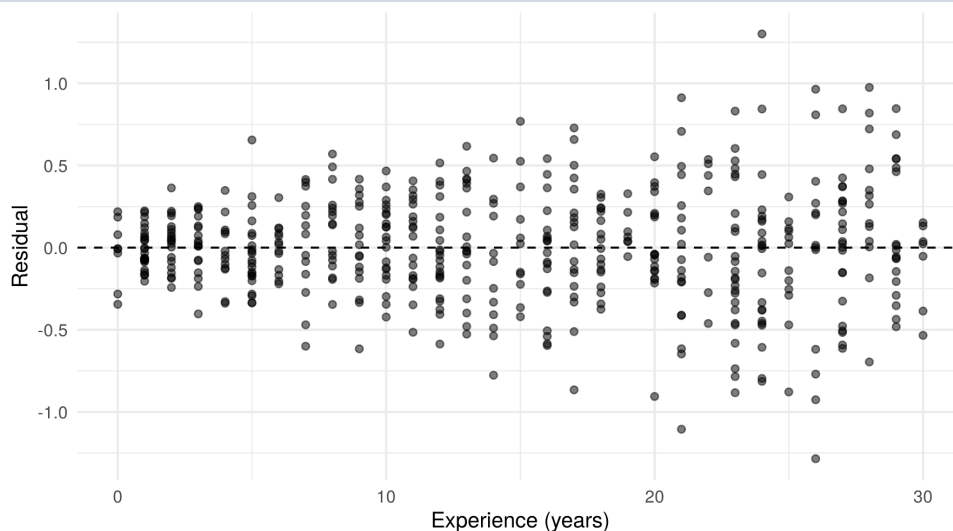
OUTPUT

```
      educ exper
long  0.086 0.019
short 0.108 0.019
      gap formula
0.0219  0.0219

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.68394    0.07470    9.16   <2e-16 ***
educ         0.10809    0.00593   18.24   <2e-16 ***
exper        0.01886    0.00172   10.95   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

educ_resid
0.1081
```



Residuals against experience: the spread grows with experience, a fan shape. The simulated noise was built to do this.

How to read it.

- **long vs short:** with ability in the model, the education coefficient is 0.086 (close to the true 0.08). Without it, 0.108. Leaving ability out inflates the estimate by about 25 %.
- **gap vs formula:** the bias (0.0219) is exactly $\hat{\beta}_3 \cdot \hat{\gamma}_1$, the omitted variable bias formula.
- **coftest with HC1:** robust standard errors. They are valid with or without heteroskedasticity; here they are close to the default ones, because the fan shape is mild.
- **educ_resid:** the FWL check reproduces the multiple-regression coefficient 0.1081.

Spot the mistake. Each statement is wrong. Correct it before checking the answer key.

1. " R^2 is only 0.12, so the education coefficient can't be trusted."
2. " $\hat{\beta} = 0.108$ with $p < 0.001$, so one more year of school causes 10.8 % higher wages."
3. "The residual plot shows a fan shape, so the OLS coefficients are biased and must be thrown away."

Practice 1

Level 1 • Understanding

1. What is the difference between $\beta_1 = 0.07$ and $\hat{\beta}_1 = 0.07$?
2. Why is a U-shaped pattern in a residual plot a problem?

Level 2 • Calculation

3. In $\log(\text{wage}) = 1.2 + 0.06 \cdot \text{educ} + 0.01 \cdot \text{age}$, interpret 0.06. What is the predicted wage difference for 2 extra years of education?
4. $\hat{\beta} = 0.42$ with $\text{se} = 0.15$ in a large sample. Compute t and the 95 % confidence interval. Is it significant at 5 %?
5. $\beta_1 = 0.05$, $\beta_3 = 0.10$, $\gamma_1 = 0.3$. What does the regression without ability give? Too high or too low?

Level 3 • Critical thinking

6. A study regresses a health index on the number of doctor visits and finds a negative coefficient. Name an omitted variable and use the bias formula to explain the sign.
7. Robust standard errors come out twice as large as the default ones. What does this suggest, and which should be reported?

R questions

8. In the lab output, what share of the short-regression coefficient is bias?
9. Change `0.10 * ability` to `0 * ability` in the script and rerun. What happens to the gap between long and short, and why?

Maximum likelihood

Where this fits. Lecture 1 found β by minimising squared mistakes. Maximum likelihood is a second, more general recipe: it works for 0/1 outcomes, counts and waiting times, where a straight line makes no sense. It also explains *why* OLS is a good idea: with normal errors, OLS **is** the maximum likelihood estimator. The logit models used with RDD and IV data later on are fitted this way.

3 key ideas

1. **Likelihood** = how probable (or, for continuous data, how dense) the observed data are, seen as a function of the parameter. The data are fixed; the parameter slides.
2. **Recipe:** write the likelihood, take logs, take the derivative, set it to zero, solve.
3. **A sharply curved peak means a precise estimate.** The curvature at the top (the information) gives the standard error.

After this chapter it should be possible to...

FOUNDATIONAL

Explain the difference between a probability and a likelihood.

FOUNDATIONALFind a simple MLE: write $\ell(\theta)$, take the derivative, set it to zero.**INTERMEDIATE**

Get a standard error from the curvature of the log-likelihood.

INTERMEDIATE

Interpret logistic-regression coefficients as changes in log-odds.

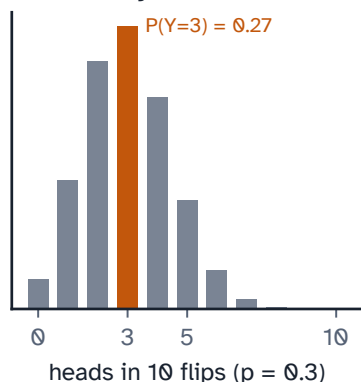
ADVANCED

Use Fisher information, the Cramér–Rao bound and the information matrix equality.

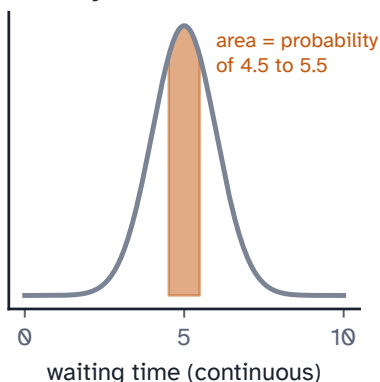
Interactive version: the likelihood explorer at sweetheartapi.github.io/learn_econometrics/#l2/l2-demo

Probability, density and likelihood

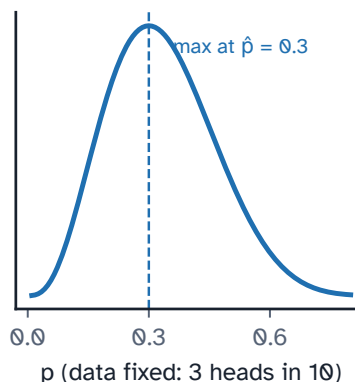
Probability



Density



Likelihood



Left: for counts, each outcome has a real probability (bar height). Middle: for continuous data, single values have probability zero; the density is a height and probabilities are areas. Right: the likelihood uses the same formula, but the data are fixed and the parameter varies.

	Probability	Density	Likelihood
Data	discrete (counts, 0/1)	continuous (times, wages)	already observed, fixed
What varies	the outcome	the outcome	the parameter
Values	between 0 and 1	any height ≥ 0 (can exceed 1)	any height ≥ 0
Sums to 1?	over outcomes, yes	area under the curve = 1	no; it is not a distribution over θ

So the likelihood is built from **probabilities** when the data are discrete (coin flips, O-ring failures) and from **densities** when the data are continuous (waiting times, normal regression errors). Either way, $L(\theta)$ is not the probability that θ is true.

Three ways to explain: likelihood

Simple. How well a particular parameter value explains the data you actually saw.

Intuitive. You saw 15 heads in 50 flips. If p were 0.5, that result would be fairly unlikely. If p were 0.3, it would be as likely as it can be. The likelihood gives every candidate p a score; the MLE is the top scorer.

Technical. $L(\theta) = \prod f(x_i | \theta)$, read as a function of θ with the data fixed. It is not a probability distribution over θ (it needn't integrate to 1). $\hat{\theta} = \operatorname{argmax} \ell(\theta)$; the observed information $-\ell''(\hat{\theta})$ gives $SE(\hat{\theta}) \approx 1/\sqrt{-\ell''(\hat{\theta})}$.

The coin example, every step shown

500 flips, 97 heads. If the coin were fair ($p = 0.5$), getting 97 heads would have probability about 9.3×10^{-46} . So which p fits best?

Why does this work? Logs, then derivative, step by step (k heads in n flips)

1. Probability of the data: $L(p) = C \cdot p^k \cdot (1 - p)^{n-k}$, where $C = "n \text{ choose } k"$ does not depend on p .
2. Take logs. A product becomes a sum and powers come down in front: $\ell(p) = \log C + k \cdot \log p + (n - k) \cdot \log(1 - p)$.
3. Differentiate each piece. $\log C$ is a constant (derivative 0); the derivative of $\log p$ is $1/p$; the derivative of $\log(1 - p)$ is $-1/(1 - p)$. So $\ell'(p) = k/p - (n - k)/(1 - p)$.
4. Set $\ell'(p) = 0$: $k/p = (n - k)/(1 - p)$.
5. Cross-multiply: $k(1 - p) = (n - k)p$, so $k - kp = np - kp$, so $k = np$.
6. Solve: $\hat{p} = k/n$. Here $97/500 = 0.194$.
7. Check it is a maximum: $\ell''(p) = -k/p^2 - (n - k)/(1 - p)^2 < 0$ (curving down).

Taking logs doesn't move the peak, because log is an increasing function. It just turns a product of many tiny numbers into a sum that is easy to differentiate.

Think of a detective with fixed evidence trying out suspects. The best suspect is the one under whom the evidence is least surprising.

The general recipe

$$L_n(\theta) = \prod_{i=1}^n f(X_i | \theta), \quad \ell_n(\theta) = \sum_{i=1}^n \log f(X_i | \theta), \quad \hat{\theta} = \arg \max_{\theta} \ell_n(\theta)$$

- $f(x | \theta)$ is a probability (discrete data) or a density (continuous data) with unknown parameter θ .
- The product needs **independent** observations; i.i.d. data can be multiplied.
- No closed formula? A computer climbs the hill numerically (R's `optim`, or `glm` for logit).

Waiting times (exponential), every step shown

Why does this work? $\hat{\lambda} = \bar{x}$ for exponential waiting times

1. Density for one waiting time with mean λ : $f(x | \lambda) = (1/\lambda) \cdot e^{-(x/\lambda)}$.
2. Log of one density: $\log f = -\log \lambda - x/\lambda$.
3. Add over n observations: $\ell(\lambda) = -n \cdot \log \lambda - (\sum x_i)/\lambda = -n \cdot \log \lambda - n \cdot \bar{x}/\lambda$.
4. Derivative: $\ell'(\lambda) = -n/\lambda + n \cdot \bar{x}/\lambda^2$.
5. Set to zero and multiply by λ^2/n : $-\lambda + \bar{x} = 0$, so $\hat{\lambda} = \bar{x}$.

In practice. A call-centre manager counts calls in five random minutes: 3, 5, 4, 2, 6. With a Poisson model, $\ell(\lambda) = \sum (y_i \cdot \log \lambda - \lambda) + \text{constant}$; setting $\sum y_i / \lambda - n = 0$ gives $\hat{\lambda} = \bar{y} = 4$ calls per minute. Staffing is planned around that rate.

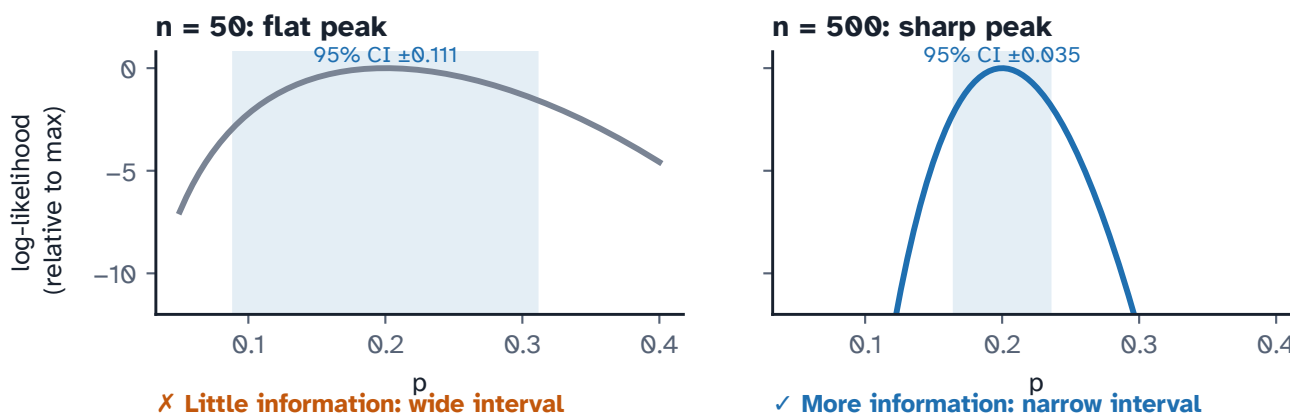
Challenger: logistic regression by ML

Each launch had 6 O-rings; the question is how the failure probability p_i depends on temperature.

- A straight line $p_i = \beta_0 + \beta_1 \cdot \text{temp}$ can give probabilities below 0 or above 1.
- So model the log-odds: $\log(p_i / (1 - p_i)) = \beta_0 + \beta_1 \cdot \text{temp}_i$.
- Failures $y_i \sim \text{Binomial}(6, p_i)$. Multiply the probabilities across launches, take logs, and maximise over β_0, β_1 .

In practice. A bank predicts loan default from income with a logit. If $\hat{\beta}$ on income (in thousands of €) is -0.05 , each extra 1,000 € multiplies the odds of default by $e^{(-0.05)} \approx 0.95$, about 5 % lower odds.

Curvature, information and standard errors



Same estimate $\hat{p} = 0.2$, different sample sizes. With 50 observations the log-likelihood is flat near the top: many values of p fit almost as well, so the interval is wide. With 500 the peak is sharp.

Term	Definition	Plain meaning
Score	first derivative of ℓ	Slope of the log-likelihood; zero at the top
Hessian	second derivative of ℓ	Curvature. At a maximum it is negative (curving down)
Observed information	minus the Hessian, $-\ell''(\hat{\theta})$	Positive number: how sharply curved the peak is
Fisher information J	variance of the score per observation (= expected $-\ell''$)	How much one observation tells about θ

Hessian vs information. Some texts, including the lecture slides, define $H_n(\theta)$ with a minus sign, so that it is positive. Both conventions appear; what matters is that the variance uses the *positive* curvature: $\text{Var}(\hat{\theta}) \approx 1 / (\text{observed information}) \approx 1/(n \cdot J)$.

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow N(0, J^{-1}) \Rightarrow \text{Var}(\hat{\theta}) \approx \frac{1}{nJ}$$

For the exponential, $J = 1/\lambda^2$, so $\text{Var}(\hat{\lambda}) \approx \lambda^2/n$ and $\text{SE} \approx \hat{\lambda}/\sqrt{n}$. With 100 waits averaging 5 minutes, $\text{SE} \approx 0.5$; with 400 waits it halves to 0.25: four times the data, twice the precision.

Why people use MLE

- **Consistent:** gets to the true θ as n grows (with a correctly specified model).
- **Asymptotically normal:** so tests and confidence intervals can be built.
- **Efficient:** reaches the Cramér–Rao lower bound, $\text{Var} \geq 1/(nJ)$, in large samples.
- **Invariant:** the MLE of $h(\theta)$ is $h(\hat{\theta})$. The MLE of p^2 is $\hat{p}^2$.

Link to Lecture 1

With $y = X\beta + \varepsilon$ and **normal** errors, the log-likelihood is a constant minus $(1/2\sigma^2) \cdot \sum (y_i - x_i\beta)^2$. Maximising it is the same as minimising squared residuals. **So OLS = MLE under normal errors.** (The ML estimate of σ^2 divides by n , not $n - p$.)

The method in five questions

Maximum likelihood in five questions. The question to ask every time: ***What distribution am I assuming for the data?***

1. **What problem are we solving?** Estimate the parameters of a model for the whole distribution of the data (coin flips, waiting times, 0/1 outcomes).
2. **How does the method solve it?** Choose the θ that makes the observed data most probable: maximise $\ell(\theta) = \sum \log f(x_i | \theta)$. The curvature at the top gives the SE.
3. **What must hold?**
 - The assumed distribution is (close to) right.
 - Independent observations, so the likelihood is a product.
 - The parameter is identified and not on the edge of its range.
 - Large n for the normal approximation and $\text{SE} \approx 1/\sqrt{\text{information}}$.
4. **What happens when it fails?**
 - Wrong distribution: $\hat{\theta}$ targets a "pseudo-true" value and Hessian-based SEs are wrong (use the sandwich $H^{-1}JH^{-1}$).
 - Small n or a boundary: the normal approximation is poor.
5. **How can it be checked?**
 - Compare fitted and observed distributions.
 - Compare the observed information with the outer product of scores (information matrix equality).
 - Try alternative distributions and report robust SEs.

R lab 2 · MLE by hand and with glm()

Base R only. Part 1 maximises the exponential likelihood numerically and compares it with the formula. Part 2 fits a Challenger-style logit (simulated: 23 launches, 6 O-rings each).

R CODE · LAB2.R

```
# Lab 2: maximum likelihood by hand, and with glm()
set.seed(2)
wait <- rexp(100, rate = 1/5)          # 100 waiting times, true mean 5 minutes

# Negative log-likelihood of the exponential with mean lambda
nll <- function(lambda) -sum(dexp(wait, rate = 1/lambda, log = TRUE))
fit <- optim(4, nll, method = "Brent", lower = 0.1, upper = 50, hessian = TRUE)
c(mle = fit$par, sample_mean = mean(wait))

# Standard error from the curvature (Hessian of the negative log-likelihood)
c(se_hessian = sqrt(1 / fit$hessian[1, 1]), se_formula = fit$par / sqrt(100))

# Logistic regression is also fitted by maximum likelihood
launch <- data.frame(temp = round(runif(23, 12, 27), 1))
p <- plogis(5 - 0.35 * launch$temp)
launch$fail <- rbinom(23, size = 6, prob = p)      # failed O-rings out of 6
m <- glm(cbind(fail, 6 - fail) ~ temp, family = binomial, data = launch)
round(summary(m)$coefficients, 3)
round(exp(coef(m)["temp"]), 3)                    # odds multiplier per +1 degree
round(predict(m, data.frame(temp = c(10, 20)), type = "response"), 3)
logLik(m)
```

OUTPUT

```
      mle sample_mean
      5.034      5.034
se_hessian se_formula
      0.5034      0.5034
      Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.569      1.089   4.195    0
temp          -0.327      0.065  -5.021    0
temp
0.721
  1    2
0.785 0.122
'log Lik.' -24.65 (df=2)
```

How to read it.

- **mle vs sample_mean:** the numerical maximum (5.034) equals the sample mean, as the derivation predicts.
- **se_hessian vs se_formula:** optim minimised the *negative* log-likelihood, so its Hessian is the observed information (positive). $1/\sqrt{(\text{information})} = 0.503$, the same as $\hat{\lambda}/\sqrt{n}$.
- **temp coefficient -0.327:** each extra degree lowers the log-odds of an O-ring failing by 0.327. **exp(coef) = 0.721:** each degree multiplies the odds of failure by 0.72 (28 % lower odds).
- **Predicted probabilities:** 0.785 at 10 °C vs 0.122 at 20 °C. Cold launches are far riskier.
- **logLik:** the maximised log-likelihood (-24.65). It is used for AIC and likelihood-ratio tests.

Spot the mistake.

1. "The likelihood at $p = 0.3$ is 0.27, so there is a 27 % chance that p equals 0.3."
2. "The MLE is the most probable value of the parameter given the data."
3. "At the maximum, the Hessian of the log-likelihood is positive, which is why the standard error is $1/\sqrt{\text{Hessian}}$."

Practice 2

Level 1 • Understanding

1. Explain the difference between a probability, a density and a likelihood in one sentence each.
2. Why does maximising $\log L$ give the same answer as maximising L ?

Level 2 • Calculation

3. A coin lands heads 3 times in 10 flips. Write the log-likelihood, differentiate it, and find $\hat{p}$.
4. Exponential data: $n = 100$ and $\hat{\lambda} = 5$. Approximate SE of $\hat{\lambda}$? What would it be with $n = 400$?
5. A logit coefficient on temperature is -0.327 . By how much do the odds of failure change per extra degree?

Level 3 • Critical thinking

6. Two studies estimate the same $\hat{p} = 0.2$. In one, the log-likelihood is almost flat around the peak. What does that say about that study, and why?
7. A classmate says a density value of 1.8 must be a mistake because probabilities can't exceed 1. Explain why it can be correct.

R questions

8. Why does the script minimise nll (the *negative* log-likelihood) instead of maximising the log-likelihood?
9. Change `rexp(100, ...)` to `rexp(400, ...)`. What happens to the standard error, and why?

The bootstrap

Where this fits. Lectures 1 and 2 gave formulas for standard errors, but they rely on assumptions (large samples, a known variance formula). The bootstrap gets uncertainty by simulation instead. It comes back in practice whenever an estimator has no easy SE formula.

3 key ideas

1. **Resample n observations from the data, with replacement.** Some values appear twice, some not at all.
2. **Recompute the estimate on each resample**, thousands of times. The spread of those estimates is the standard error; their percentiles give a confidence interval.
3. **It needs fewer assumptions, but it is not magic.** It cannot add information a small or unrepresentative sample doesn't have, and it fails in some known cases (sample maximum, boundary parameters).

After this chapter it should be possible to...

FOUNDATIONAL Run the bootstrap algorithm by hand: resample, recompute, repeat.

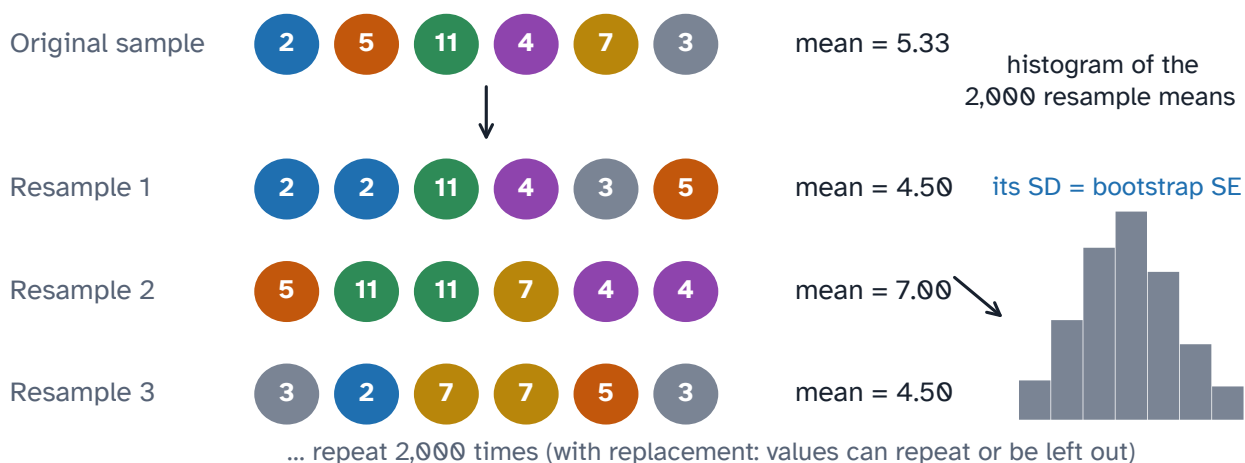
FOUNDATIONAL Read a bootstrap SE and a percentile confidence interval.

INTERMEDIATE Judge when the bootstrap helps and when it fails (tiny n , the maximum, boundaries).

ADVANCED Explain bootstrap consistency, pivotal statistics and subsampling.

Interactive version: the bootstrap demo with a coverage check at sweetheartapi.github.io/learn_econometrics/#l3/l3-try-it-yourself

The idea



Each resample has the same size as the original sample and is drawn with replacement. Recomputing the statistic on thousands of resamples gives a distribution; its spread mimics how the estimate would vary across real repeated samples.

The algorithm

1. Draw a sample of size n from the data (the empirical distribution $\hat{F}_n$), with replacement.
2. Compute the statistic on it: T^* .
3. Repeat B times (1,000–10,000).
4. SD of the B values = bootstrap SE. Their percentiles = CI. Their mean minus the original estimate = bias estimate.

Why does this work? The plug-in principle, step by step

1. The goal is the sampling distribution of an estimate: how it would vary if new samples could be drawn from the true population F .
2. F is unknown, but the sample's own distribution $\hat{F}_n$ (each observation with probability $1/n$) is the best available stand-in, and it gets closer to F as n grows.
3. Drawing n values from $\hat{F}_n$ **with replacement** is exactly "drawing a new sample from the stand-in population".
4. So the spread of the estimate across bootstrap samples approximates its spread across real samples.
5. The approximation is only as good as step 2: if the sample is tiny or unrepresentative, $\hat{F}_n$ is a poor stand-in and the bootstrap inherits that.

Worked example. Data: 2, 5, 11 (mean 6). One resample: 2, 2, 11 $\rightarrow$ mean 5. Another: 5, 11, 11 $\rightarrow$ mean 9. Another: 2, 5, 5 $\rightarrow$ mean 4. Repeat 1,000 times; the SD of those means is the bootstrap SE of the mean.

Three ways to explain: bootstrap standard error

Simple. How much your estimate jumps around when you recompute it on reshuffled copies of your own data.

Intuitive. Your 9 claim times are a mini-population. Draw 9 of them with replacement, compute the mean, and do that 1,000 times. The spread of those 1,000 means mimics how the mean would vary if you could collect new samples.

Technical. $SE_{boot} = SD$ of $T_1, \dots, T_B$, where each T^* is computed on a sample from the empirical distribution $\hat{F}_n$. It is consistent when $G_n(\cdot, \hat{F}_n) \rightarrow G_n(\cdot, F_0)$: smooth statistics of i.i.d. data. B only controls Monte Carlo error.

When it's useful, and when it isn't

Useful:

- **Complicated estimators** with no variance formula (e.g. the output of a numerical optimisation).
- **Formulas that need unknowns.** The median's variance is $1/(4n \cdot f(m)^2)$, which needs the unknown density at the median.
- **Doubtful distributional assumptions**, e.g. skewed spending data where a normal approximation is poor.
- **Moderately small samples** where large-sample formulas are shaky, *provided* the sample still represents the population reasonably well.

The bootstrap does not fix small samples. It resamples the data available; it cannot create information that isn't there. With very few observations (say 8), the empirical distribution is a crude stand-in, the bootstrap distribution is lumpy, and intervals often cover the truth less than 95 % of the time. A small sample remains a small sample.

Unreliable or needing a special version:

Situation	Problem	What to do
Very small or unrepresentative sample	$\hat{F}_n$ is a poor stand-in for F	Be cautious; report it; gather more data
Sample maximum / minimum	Bootstrap repeats the observed max $\approx 63\%$ of the time	Subsampling or extreme-value methods
Parameter on the boundary	e.g. $\mu \geq 0$ with true $\mu = 0$ (Andrews 2000)	Subsampling
Dependent data (time series)	Resampling single rows breaks the dependence	Block bootstrap
Clustered data (pupils in schools)	Rows within a cluster aren't independent	Resample whole clusters
Heavy tails (infinite variance)	The bootstrap of the mean is inconsistent	Robust statistics, subsampling

Subsampling draws *smaller* samples (size $m < n$) **without** replacement. It works more generally but is less efficient, and m must be chosen.

In practice. A small online shop has 20 customers and wants a CI for the **median** basket value (two huge orders distort the mean). There's no simple formula for the median's SE, so it bootstraps the median 2,000 times (R lab below). An engineer who wants the highest load a sensor can record should *not* bootstrap the sample maximum: most resamples simply repeat it, so the spread looks far too small.

Three kinds of bootstrap confidence interval

Interval	How it's built	Strength	Weakness
Normal	estimate $\pm 1.96 \times$ bootstrap SE (bias-adjusted in R's boot.ci)	Simple, stable	Assumes the bootstrap distribution is symmetric and bell-shaped
Percentile	2.5th and 97.5th percentiles of the B estimates	Follows skewness automatically	Ignores bias; can be off in small samples
BCa (bias-corrected and accelerated)	Percentiles shifted to correct for bias and skewness	Usually the most accurate of the three	Needs more resamples ($\geq 2,000$); can be unstable with very small n

When the three intervals roughly agree, the choice hardly matters. When they disagree, the bootstrap distribution is skewed or biased, and the percentile or BCa interval is usually preferred over the normal one.

The stamp example (from the slides)

Were stamps printed on one kind of paper or several? H_0 : the thickness distribution has one peak (mode). The test statistic is the smallest smoothing level $\hat{h}_1 = 0.0068$ that gives one peak. A **smooth bootstrap** resamples from a smoothed one-peak version of the data. With $B = 5,000$ the significance level was 0.0002, so "one paper type" was rejected.

Theory words

- **$G_n(\tau, F_0)$** : true distribution of the test statistic (unknown). Bootstrap estimate: $G_n(\tau, \hat{F}_n)$.
- **Pivotal**: the statistic's distribution doesn't depend on F (asymptotically pivotal: true for large n , like a t -statistic). Bootstrapping pivotal statistics works better.
- **Consistency**: the bootstrap distribution converges to the true one as n grows. It needs $\hat{F}_n$ to be a good estimate of F_0 and the limit to be smooth (Beran & Ducharme).

The method in five questions

The bootstrap in five questions. The question to ask every time: ***Does resampling reproduce the relevant sampling uncertainty?***

1. **What problem are we solving?** Get a standard error or confidence interval for an estimator that has no simple formula, or whose formula needs unknowns.
2. **How does the method solve it?** Treat the sample as the population: draw n observations with replacement, recompute the statistic, repeat B times, and use the spread.
3. **What must hold?**
 - The sample represents the population well (n not tiny).
 - Observations are independent (or the resampling copies the dependence, e.g. block bootstrap).
 - The statistic depends smoothly on the data (not a maximum, not a boundary parameter).
4. **What happens when it fails?**
 - Tiny n : intervals cover the truth less often than they claim.
 - Maximum or boundary: the bootstrap distribution is wrong (spike at the sample maximum).
 - Dependent data resampled as if independent: SEs too small.
5. **How can it be checked?**
 - Is the statistic smooth? Is the data independent?
 - Compare with an analytic SE where one exists.
 - Try subsampling ($m < n$, without replacement) when in doubt.

R lab 3 • Bootstrapping a median with boot

Packages: boot (ships with R), ggplot2. The data are 20 skewed basket values.

R CODE • LAB3.R

```
# Lab 3: bootstrapping a median with the boot package
library(boot); library(ggplot2)
set.seed(3)
basket <- round(rlnorm(20, meanlog = 3.5, sdlog = 0.8), 1) # 20 skewed basket values (EUR)
c(mean = mean(basket), median = median(basket))

med_fun <- function(data, idx) median(data[idx]) # statistic on a resample
b <- boot(basket, med_fun, R = 2000)
b # original, bias, std. error
boot.ci(b, type = c("norm", "perc", "bca"))

# Histogram of the 2,000 bootstrap medians
ggplot(data.frame(med = b$t[, 1]), aes(med)) +
  geom_histogram(bins = 30, fill = "grey60", colour = "white") +
  geom_vline(xintercept = b$t0, linetype = 2) +
  labs(x = "Bootstrap median (EUR)", y = "Count") + theme_minimal()

# Where the bootstrap fails: the sample maximum
x <- runif(50, 0, 10)
bmax <- boot(x, function(d, i) max(d[i]), R = 2000)
mean(bmax$t[, 1] == max(x)) # share of resamples that repeat the sample max
```

OUTPUT

```
mean median
35.23  30.05

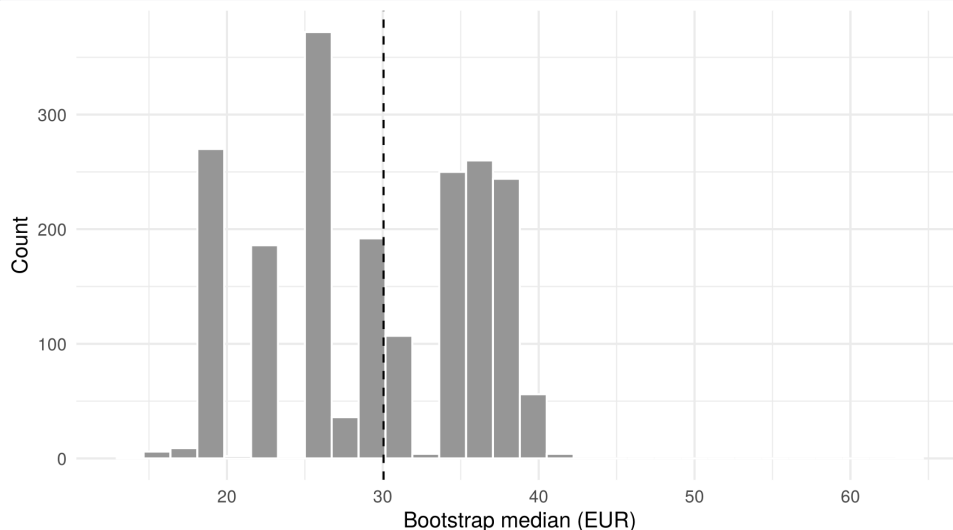
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = basket, statistic = med_fun, R = 2000)

Bootstrap Statistics :
    original    bias      std. error
t1*   30.05 -0.4944      6.562
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = b, type = c("norm", "perc", "bca"))

Intervals :
Level      Normal              Percentile          BCa
95%  (17.68, 43.41 )  (18.50, 38.80 )  (18.50, 38.70 )
Calculations and Intervals on Original Scale
[1] 0.653
```



The 2,000 bootstrap medians. The histogram is lumpy because the median of 20 values can only be one of a few numbers in the data (or the average of two neighbours). This is the small-sample warning made visible.

How to read it.

- **original 30.05, std. error 6.56:** the sample median and its bootstrap SE. **bias -0.49:** the bootstrap medians average slightly below the original.
- **Normal (17.7, 43.4)** is symmetric around the estimate. **Percentile (18.5, 38.8)** and **BCa (18.5, 38.7)** are shorter on the upper side, because the bootstrap distribution is not symmetric. Here the percentile and BCa intervals are more trustworthy than the normal one.
- **0.653:** in 65 % of resamples the bootstrap maximum equals the sample maximum, close to the theoretical $1 - (1 - 1/50)^{50} \approx 0.64$. The real sampling distribution of a maximum has no such spike, so the bootstrap misrepresents it.

Spot the mistake.

1. "With $B = 100,000$ resamples the estimate is much more precise than with $B = 1,000$."
2. "The bootstrap creates new data, so it works well even with 6 observations."
3. "To bootstrap a panel of 50 firms over 10 years, resample the 500 firm-year rows."

Practice 3

Level 1 • Understanding

1. Why must the bootstrap resample **with** replacement?
2. What does the histogram of bootstrap estimates approximate?

Level 2 • Calculation

3. Data: 1, 3, 8. Write down two possible bootstrap samples and their means.
4. From 1,000 bootstrap means, the 25th smallest is 4.1, the 975th smallest is 7.9, the SD is 1.0, and the original mean is 6. Give the percentile and the normal 95 % intervals.

Level 3 • Critical thinking

5. In the R lab the normal interval is (17.7, 43.4) but the percentile interval is (18.5, 38.8). Why do they differ, and which is more credible here?
6. A researcher with 8 observations writes: "The sample is small, so I bootstrapped; this solves the small-sample problem." Respond.

R questions

7. Why is the histogram of bootstrap medians lumpy rather than smooth?
8. The script finds that 65 % of bootstrap maxima equal the sample maximum. Why does that show the bootstrap failing here?

Intro to causality

Where this fits. Lecture 1 ended with a problem: a regression slope mixes the effect of x with everything left out (omitted variable bias). This lecture defines what a *causal* effect is, shows why randomisation solves the problem, and sets up the three quasi-experimental methods of Lectures 5–7.

3 key ideas

1. Everyone has two **potential outcomes**: $Y(1)$ with treatment and $Y(0)$ without. Only one is ever observed: the fundamental problem of causal inference.
2. **Random assignment** makes treated and control groups comparable **on average**, so a difference in means estimates the causal effect without bias.
3. Without an experiment, look for a **natural source of randomness**: a lottery, nature, a law, a cut-off. Lectures 5–7 are three ways to use one.

After this chapter it should be possible to...

- FOUNDATIONAL** Define potential outcomes and the fundamental problem of causal inference.
- FOUNDATIONAL** Explain why random assignment turns a difference in means into a causal effect.
- INTERMEDIATE** Split a naive comparison into ATT + selection bias.
- INTERMEDIATE** Decide what to control for: confounders yes, mediators and colliders no.
- ADVANCED** Compute a randomisation-inference p-value.

Interactive version: the coin flip vs self-selection demo at sweetheartapi.github.io/learn_econometrics/#l4/l4-demo

Potential outcomes

Imagine seeing both versions of each person:

Person	$Y(1)$ treated	$Y(0)$ untreated	Effect $Y(1) - Y(0)$
Adam	4	6	-2
Boris	5	7	-2
Cyril	8	6	+2
Diana	3	7	-4
Ema	3	4	-1
Filip	1	3	-2
Average	4.0	5.5	-1.5

The true average effect for these six is **-1.5**. In real data, half the cells are missing: if Adam is treated, his $Y(0)$ is never seen.

Worked example: randomisation in this class of six. Suppose a coin puts Adam, Boris and Ema in the treatment group. Observed treated mean = $(4 + 5 + 3)/3 = 4.0$; control mean (Cyril, Diana, Filip) = $(6 + 7 + 3)/3 = 5.33$. Estimate = **-1.33**. A different draw (Cyril, Diana, Filip treated) gives $4.0 - 5.67 = \mathbf{-1.67}$. Neither equals -1.5 exactly, but averaged over all 20 possible ways to split the class, the estimate is exactly -1.5.

Random assignment balances groups on average, not in every sample. In any one randomisation the groups can differ by chance (here -1.33 or -1.67 instead of -1.5). That is why studies show a balance table of pre-treatment characteristics, and why small experiments are noisy. What randomisation guarantees is no **systematic** difference: the estimator is unbiased.

$$Y(1), Y(0) \perp\!\!\!\perp D$$

In words: who gets treated has nothing to do with how they would do either way.

Why does this work? Naive comparison = effect + selection bias

1. The naive comparison is $E[Y \mid D = 1] - E[Y \mid D = 0]$.
2. For the treated, $Y = Y(1)$; for the untreated, $Y = Y(0)$. So it equals $E[Y(1) \mid D = 1] - E[Y(0) \mid D = 0]$.
3. Add and subtract $E[Y(0) \mid D = 1]$ (the treated group's outcome *without* treatment): $= \{E[Y(1) \mid D = 1] - E[Y(0) \mid D = 1]\} + \{E[Y(0) \mid D = 1] - E[Y(0) \mid D = 0]\}$.
4. The first bracket is the **ATT** (effect on the treated). The second is **selection bias**: how different the treated would have been anyway.
5. With random assignment, D is independent of $Y(0)$, so the second bracket is zero on average and the ATT equals the ATE.

In practice. Gym members are healthier than non-members, but people who join were probably healthier and more motivated to begin with: their $Y(0)$ is higher (selection bias). If a city gave free memberships by lottery, winners and losers would differ only by chance, and the difference in health would estimate the gym's effect.

Three ways to explain: selection bias

Simple. The treated group would have been different even without the treatment.

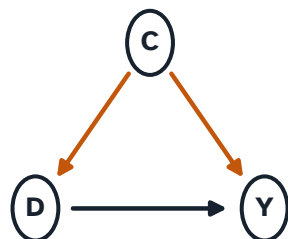
Intuitive. People who join a gym are already healthier than people who don't. Comparing their health a year later mixes the gym's effect with who chose to join.

Technical. $E[Y \mid D=1] - E[Y \mid D=0] = \text{ATT} + \{E[Y(0) \mid D=1] - E[Y(0) \mid D=0]\}$. Random assignment makes D independent of $Y(0)$, so the second term is zero.

Causal diagrams (DAGs)

A DAG draws variables as nodes and causal effects as arrows. Three shapes come up again and again:

Confounding



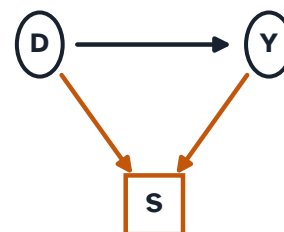
C drives both: control for C

Mediation



M is part of the effect: don't control

Selection (collider)



Studying only $S = 1$ creates bias

Confounding: C causes both D and Y, creating a misleading association; adjust for C. Mediation: M is how D affects Y; adjusting for M removes part of the effect being measured. Selection: S is caused by both D and Y; looking only at $S = 1$ (or controlling for S) creates a fake association.

Shape	Example	Control for it?
Confounder ($D \leftarrow C \rightarrow Y$)	Ability raises both education and wages	Yes: otherwise omitted variable bias
Mediator ($D \rightarrow M \rightarrow Y$)	Training $\rightarrow$ skills $\rightarrow$ wage	No, if the total effect is the goal
Collider ($D \rightarrow S \leftarrow Y$)	Studying only hired workers, when hiring depends on both education and talent	No: conditioning on it creates bias

In practice: a collider. Among people already admitted to hospital, a study may find that smokers have *better* outcomes from some disease. Hospital admission depends on both smoking-related illness and the disease's severity, so looking only at admitted patients creates a misleading link. The fix is not more control variables, it is a better sample.

Who is the effect for? ATE, ATT, ITT and LATE

Estimand	Definition	In words
ATE	$E[Y(1) - Y(0)]$	Average effect over everyone
ATT	$E[Y(1) - Y(0) \mid D = 1]$	Average effect for those who were treated
ITT	$E[Y \mid Z = 1] - E[Y \mid Z = 0]$	Effect of being offered / assigned treatment (Z), whether or not it was taken
LATE	$E[Y(1) - Y(0) \mid \text{complier}]$	Effect for compliers : those whose treatment was changed by the offer

Worked example: one programme, four different numbers. A job-training offer is assigned by lottery (Z) in a population of 100:

- 40 **compliers** train only if offered; their effect is +10.
- 20 **always-takers** train anyway; their effect is +4.
- 40 **never-takers** never train; their effect would be 0.

1. **ATE** = $(40 \cdot 10 + 20 \cdot 4 + 40 \cdot 0) / 100 = 4.8$.

2. **ATT**: the trained are the 20 always-takers plus the 20 compliers who won the lottery (half of 40).
 $ATT = (20 \cdot 4 + 20 \cdot 10) / 40 = 7.0$.

3. **First stage**: 60 % train if offered (compliers + always-takers) vs 20 % if not, so 0.4.

4. **ITT**: always- and never-takers behave the same in both groups and cancel; only compliers differ.
 $ITT = 0.4 \times 10 = 4.0$.

5. **LATE** = $ITT \div \text{first stage} = 4.0 / 0.4 = 10$, the compliers' effect.

Four correct answers to four different questions. Reporting a LATE as if it were the ATE would overstate this programme's average effect by more than double.

When there's no experiment

Many important questions can't be put in an experiment: Do veterans earn less because of the war? Does education raise wages? Does the minimum wage cost jobs?

- **Option 1**: compare similar people, $Y(1), Y(0) \perp\!\!\!\perp D \mid X$ ("as good as random once X is controlled for"). This works only if every confounder is measured.
- **Option 2**: find a **source of randomness** (a quasi-experiment).

In practice. A job-training evaluation compares participants with non-participants of the same age, education and past earnings (Option 1). The weak spot: motivation isn't in the data. If more motivated people sign up, the estimate is still too high.

Some Z pushes the treatment D and affects Y **only through D**: $Z \rightarrow D \rightarrow Y$, with no arrow $Z \rightarrow Y$.

$$\hat{\beta} = \frac{\text{effect of } Z \text{ on } Y}{\text{effect of } Z \text{ on } D} = \frac{E[Y \mid Z = 1] - E[Y \mid Z = 0]}{E[D \mid Z = 1] - E[D \mid Z = 0]}$$

This is exactly the IV estimator of Lecture 5, and the LATE from the worked example above.

Famous examples

Study	Source of randomness	Method
Vietnam veterans' earnings (Angrist 1990)	Draft lottery by birth date	IV
Returns to schooling (Angrist & Krueger 1991)	Quarter of birth; about 7 % per year of school	IV
Demand for fish (Angrist, Graddy & Imbens 2000)	Weather at sea	IV
Class size (Angrist & Lavy 1999)	Maimonides' rule: 40 pupils → 1 class, 41 → 2	RDD / IV
Minimum wage (Card & Krueger 1994)	New Jersey raised it, Pennsylvania didn't	DiD
Phones and fish markets (Jensen 2007)	Mobile coverage arrived at different times	DiD-style
Chernobyl fallout (Almond et al.)	Rainfall decided where fallout landed	Natural experiment

Stay critical

- Does the result **generalise** beyond this sample? Effects can **differ between people**.
- Is the natural experiment **really random**? (Is quarter of birth?)
- Even real experiments break: non-response, measurement error, a too-specific sample, changing conditions.
- Signs supporting causality: the effect is strong, consistent, specific, stable over time, grows with the dose, plausible, and confirmed by experiments.

The method in five questions

Randomised experiments in five questions. The question to ask every time: ***Would the two groups have had the same outcomes without the treatment?***

1. **What problem are we solving?** Measure the effect of D on Y when the people who get D might differ from those who don't.
2. **How does the method solve it?** Assign D at random, then compare average outcomes. Randomisation makes D independent of the potential outcomes, so selection bias is zero on average.
3. **What must hold?**
 - Random assignment (independence).
 - No spillovers between units (SUTVA).
 - People take the treatment they are assigned (or you analyse by assignment).
 - No differential drop-out.
4. **What happens when it fails?**
 - Selection into treatment: the difference in means mixes the effect with who chose it.
 - Non-compliance: comparing by assignment gives the ITT; use IV for the effect of treatment itself.
 - Attrition that differs by group: biased comparison.
5. **How can it be checked?**
 - Balance table on pre-treatment characteristics.
 - Compliance and attrition rates by group.
 - Randomisation inference for small experiments.

R lab 4 · Randomisation vs self-selection

Base R only. 1,000 people with known $Y(0)$ and $Y(1)$, so the true effects can be computed.

R CODE · LAB4.R

```
# Lab 4: potential outcomes, randomisation and selection bias
set.seed(4)
n <- 1000
y0 <- rnorm(n, 50, 10)           # outcome without treatment
effect <- rnorm(n, 5, 4)          # individual effects differ
y1 <- y0 + effect
c(ATE = mean(y1 - y0))

# (a) Random assignment: compare means
d <- sample(rep(0:1, n / 2))
y <- ifelse(d == 1, y1, y0)       # we only see one outcome per person
c(estimate = mean(y[d == 1]) - mean(y[d == 0]))

# Repeat the random assignment 1,000 times
est <- replicate(1000, {d <- sample(rep(0:1, n / 2))
  mean(y1[d == 1]) - mean(y0[d == 0])})
round(c(mean_of_estimates = mean(est), sd = sd(est),
  min = min(est), max = max(est)), 2)

# (b) Self-selection: people with high y0 choose treatment more often
d_self <- rbinom(n, 1, plogis((y0 - 50) / 5))
y_self <- ifelse(d_self == 1, y1, y0)
c(naive = mean(y_self[d_self == 1]) - mean(y_self[d_self == 0]),
  ATT = mean((y1 - y0)[d_self == 1]),
  selection_bias = mean(y0[d_self == 1]) - mean(y0[d_self == 0]))
```

OUTPUT

```
ATE
5.098
estimate
4.585
mean_of_estimates      sd      min      max
           5.11      0.64      3.14      7.21
      naive      ATT selection_bias
      16.780      5.179      11.601
```

How to read it.

- **ATE 5.10** is the truth, known only because the data are simulated.
- **estimate 4.59** from one random assignment: off by about 0.5 purely by chance.
- **Repeating the randomisation 1,000 times:** the estimates average 5.11 (unbiased) but range from 3.1 to 7.2 (SD 0.64). This is the "balanced on average" point in numbers.
- **Self-selection:** when people with high $Y(0)$ choose treatment, the naive comparison is 16.78. It splits exactly into ATT (5.18) + selection bias (11.60), the identity from the "Why does this work?" box.

Spot the mistake.

1. "Treatment was randomised, so the treated and control groups in our sample have exactly the same average age."
2. "The ITT of a voucher offer tells us the effect of actually using the voucher."
3. "Adding more control variables can only reduce bias."

Practice 4

Level 1 • Understanding

1. In one sentence, what is the fundamental problem of causal inference?
2. Explain the difference between the ATE and the ATT in words.

Level 2 • Calculation

3. Observed: Anna (treated) $Y = 10$, Ben (untreated) 6, Clara (treated) 8, David (untreated) 4. Compute the difference in means. When is it a causal effect?
4. A lottery-based programme: compliers are 30 % of people (effect 8), always-takers 10 % (effect 2), never-takers 60 % (effect 0). Compute the first stage, the ITT, the LATE and the ATE.
5. A naive comparison gives 12; the ATT is 5. How large is the selection bias?

Level 3 • Critical thinking

6. Training raises skills, and skills raise wages. A researcher estimates the effect of training on wages "controlling for skills". What goes wrong?
7. A study uses only data on admitted university students and finds that test scores and motivation are negatively correlated. Draw the DAG and explain.

R questions

8. Why is the single-randomisation estimate (4.59) not equal to the ATE (5.10), even though assignment was random?
9. Check the identity naive = ATT + selection bias using the numbers in the output.

Instrumental variables

Where this fits. Lecture 1 showed that OLS is biased when the regressor is tangled up with the error (omitted variables, reverse causality, measurement error). Lecture 4 said the cure is a source of randomness. IV is the first cure: an outside push Z that moves the treatment for reasons unrelated to the outcome. Lecture 6 will reuse the same idea at a cut-off (fuzzy RDD).

3 key ideas

1. A valid instrument is **relevant** (moves D), **as good as random**, and **excluded** (affects Y only through D). Only relevance can be checked in the data.
2. **IV = reduced form $\div$ first stage** = (effect of Z on Y) / (effect of Z on D). In practice this is two-stage least squares (2SLS).
3. When effects differ between people, IV gives a **LATE**: the effect for *compliers*, whose treatment the instrument changed.

After this chapter it should be possible to...

- FOUNDATIONAL** State the IV conditions: relevance, independence, exclusion, monotonicity.
- FOUNDATIONAL** Compute the Wald estimator: reduced form $\div$ first stage.
- INTERMEDIATE** Identify compliers, always-takers and never-takers, and say who the LATE is about.
- INTERMEDIATE** Recognise a weak instrument and know what to report instead.
- ADVANCED** Derive IV by the method of moments; read an overidentification test.

Interactive version: the weak and invalid instrument simulator at sweetheartapi.github.io/learn_econometrics/#l5/l5-demo

The problem: endogeneity

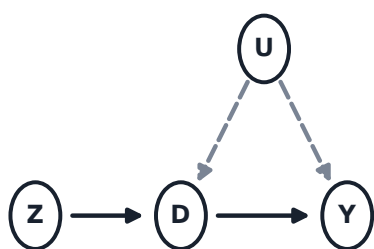
D is **endogenous** (correlated with the error) when there is an omitted variable U (like ability or motivation), reverse causality, or measurement error. Then $E[D \cdot \eta] \neq 0$ and OLS is biased.

Think of Z as a clean knob that turns D up and down without touching anything else that affects Y . Stage 1 measures how much D moved because of the knob; stage 2 asks how much Y moved along with it.

In practice. Does more police reduce crime? Cities with more crime hire more police, so a plain regression can even make police look like they "cause" crime (reverse causality). Levitt (1997) used election years as an instrument: mayors hire extra police before elections, for reasons unrelated to that year's crime trends.

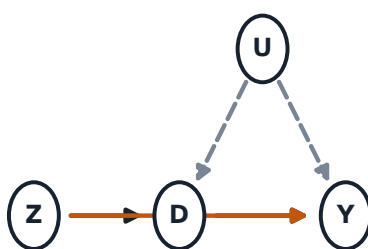
Is the instrument valid? A checklist

Valid instrument



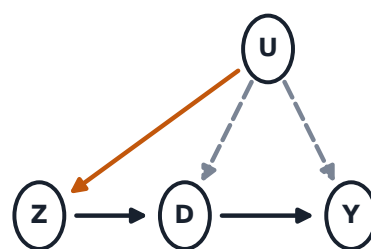
Z only reaches Y through D

Exclusion violated



Z also affects Y directly

Independence violated

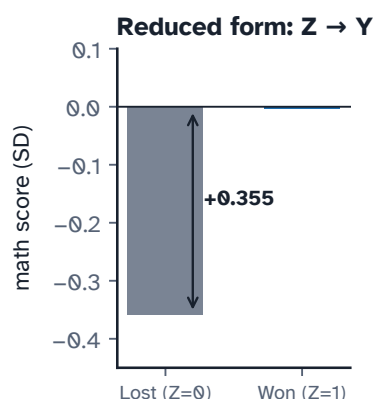
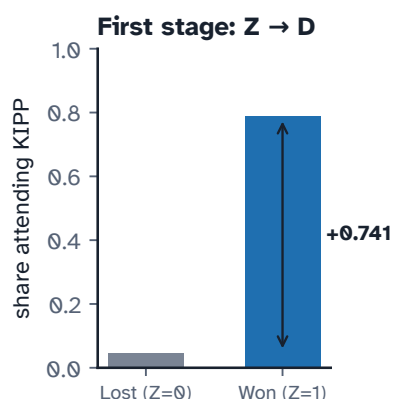


Z is linked to the hidden U

Left: a valid instrument reaches Y only through D, and has no link to the hidden U. Middle: if Z also affects Y directly, the exclusion restriction fails. Right: if Z is related to U, independence fails. In both failure cases IV is biased, and no statistical test can fully detect it.

Question	Condition	How to check
Does Z actually move D?	Relevance	First-stage regression: size, sign, F-statistic
Is Z as good as random?	Independence	Story of how Z was assigned; balance of pre-treatment characteristics across Z
Does Z affect Y only through D?	Exclusion	Argument and institutional knowledge; look for other channels; placebo outcomes Z shouldn't affect
Does Z push everyone the same way?	Monotonicity (for LATE)	Argument: are there plausible "defiers"?
With several instruments, do they agree?	Overidentification	Sargan/J-test (it can reject, but never prove validity)

First stage and reduced form



IV (LATE)

= reduced form ÷ first stage

= $0.355 \div 0.741$

= 0.48 SD

KIPP charter-school lottery. Winning raised attendance by 74 percentage points (first stage) and math scores by 0.355 SD (reduced form). Dividing the two gives the effect of actually attending for compliers.

$$\hat{\delta}_{IV} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, D)} \quad \text{binary } Z : \quad \hat{\delta}_{Wald} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}} = \frac{\text{reduced form}}{\text{first stage}}$$

Why does this work? From the model to the IV ratio

1. Model: $Y = \alpha + \delta \cdot D + \eta$, where η contains everything else (including the hidden U), so $\text{Cov}(D, \eta) \neq 0$.
2. Take the covariance of both sides with Z : $\text{Cov}(Z, Y) = \delta \cdot \text{Cov}(Z, D) + \text{Cov}(Z, \eta)$. (The constant α has zero covariance with anything.)
3. Validity (independence + exclusion) says $\text{Cov}(Z, \eta) = 0$.
4. So $\text{Cov}(Z, Y) = \delta \cdot \text{Cov}(Z, D)$, and dividing: **$\delta = \text{Cov}(Z, Y) / \text{Cov}(Z, D)$** .
5. Relevance is needed to divide: $\text{Cov}(Z, D)$ must not be 0.
6. Compare OLS: $\text{Cov}(D, Y) / \text{Var}(D) = \delta + \text{Cov}(D, \eta) / \text{Var}(D)$. The extra term is the bias that IV avoids.

In practice: Wald by hand. Instrument: growing up near a college. Hypothetical numbers: people who grew up near a college have 13.5 years of school vs 13.0 (first stage = 0.5 years) and 0.05 higher log wages (reduced form). $IV = 0.05 / 0.5 = 0.10$: about 10 % higher wages per extra year of school.

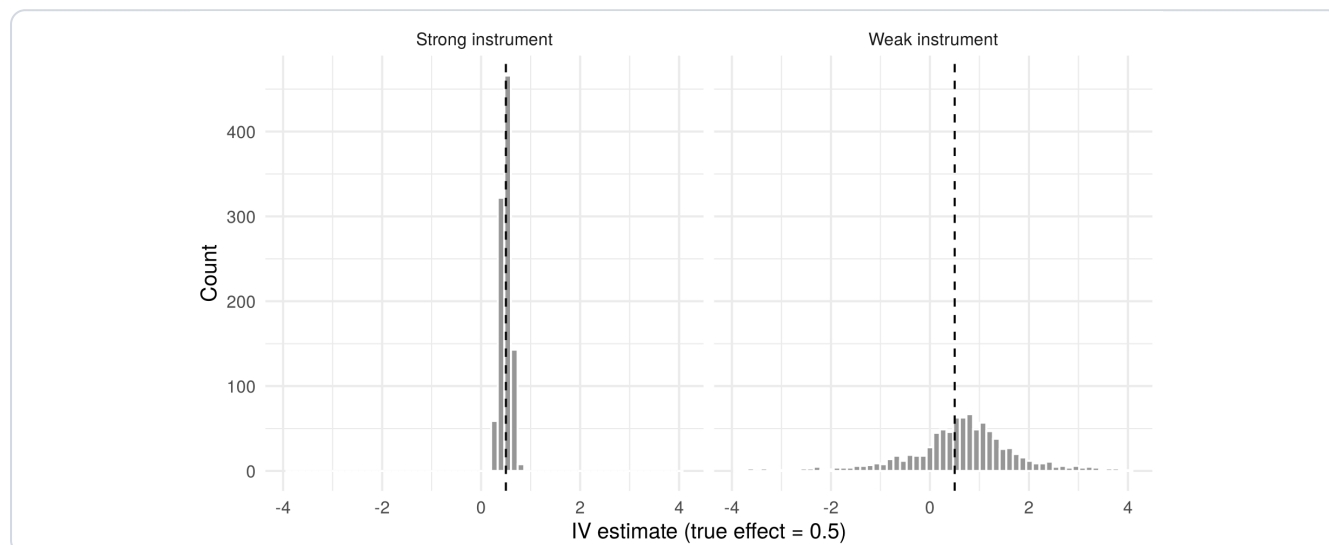
Two-stage least squares

1. **First stage:** regress D on Z (and controls X). Save the fitted values $\hat{D}$.
 2. **Second stage:** regress Y on $\hat{D}$ (and the same X , but not Z). The slope equals the IV ratio.
- **Controls** that affect both Z and Y go in **both** stages. (Same FWL logic as Lecture 1: remove X from Y , D and Z , then use the simple ratio.)
 - Running stage 2 by hand gives the right coefficient but **wrong standard errors**. An IV command (ivreg in R, ivregress 2sls in Stata) uses the actual D for the residuals.

Weak instruments

$$\hat{\delta}_{IV} - \delta = \frac{\widehat{\text{Cov}}(Z, \eta)}{\widehat{\text{Cov}}(Z, D)}$$

If Z barely moves D , the bottom is close to zero. Then even small random noise on top gets blown up: the estimate becomes erratic and conventional standard errors stop being trustworthy. In the textbook setting (one endogenous D , conventional 2SLS, a truly valid Z) the estimates also tend to be centred towards the OLS value, because the little clean variation is swamped by the same contaminated variation that biases OLS. That is a common pattern, not a law: with many instruments it gets worse, and if Z is slightly invalid a weak first stage magnifies that bias in either direction.



1,000 simulated studies each, true effect 0.5 (dashed line). With a strong instrument (median first-stage $F \approx 100$) the estimates cluster tightly around the truth. With a weak one (median $F \approx 1$) they are all over the place and centred too high, towards the biased OLS value.

"First-stage $F \geq 10$ " is a rough historical rule of thumb, not a guarantee. It comes from Staiger & Stock (1997) and Stock & Yogo (2005), and is about limiting bias. It does not ensure that conventional t-tests and confidence intervals are reliable. Lee, McCrary, Moreira & Porter (2022) show that for a standard 5 % t-test with one instrument, the first-stage F would need to be above about 104. With a moderate F , weak-instrument-robust methods (e.g. Anderson–Rubin confidence intervals) are safer.

In practice. Quarter of birth shifts schooling by only a fraction of a year, so it is a fairly weak instrument. Bound, Jaeger & Baker (1995) showed that with such weak instruments, even randomly made-up "quarters of birth" gave estimates close to OLS.

LATE and compliance types

In real experiments people don't always follow their assignment. Comparing by assignment gives the **ITT**: the effect of being *offered* treatment.

Type	Treated if Z = 1?	Treated if Z = 0?
Compliers	yes	no
Always-takers	yes	yes
Never-takers	no	no
Defiers (assumed none)	no	yes

Why does this work? Why reduced form ÷ first stage = effect for compliers

1. Z is random, so the Z = 1 and Z = 0 groups contain the same mix of types.
2. Always-takers are treated in both groups; never-takers in neither. By exclusion, Z doesn't change their outcomes either. Their contribution to the reduced form is zero.
3. Monotonicity rules out defiers. So the whole reduced form comes from compliers: reduced form = (share of compliers) × (complier effect).
4. The first stage is exactly the share of compliers: $P(D = 1 | Z = 1) - P(D = 1 | Z = 0)$.
5. Dividing: reduced form / first stage = **complier effect (LATE)**.

Worked example: KIPP charter schools (lottery).

1. 78.7 % of lottery winners and 4.6 % of losers attended. First stage = $0.787 - 0.046 = 0.741$.
2. Math scores: winners -0.003 SD, losers -0.358 SD. Reduced form (ITT) = **0.355**.
3. LATE = $0.355 / 0.741 = 0.479$ SD.
4. Shares: compliers 0.741; always-takers 0.046; never-takers $1 - 0.787 = 0.213$.

In practice. An insurer randomly mails free flu-vaccine vouchers. 60 % of receivers get vaccinated vs 20 % of others (first stage 0.40). Flu rates are 10 % vs 14 % (reduced form -0.04). LATE = $-0.04 / 0.40 = -0.10$: for people vaccinated **because of the voucher**, flu risk fell by 10 percentage points. Always-takers are not covered by this number.

Three ways to explain: LATE

Simple. The effect for the people whose treatment was changed by the instrument.

Intuitive. In the KIPP lottery, some families would send their child to KIPP whatever happened (always-takers) and some never would (never-takers). The lottery only changes what the "compliers" do, so it can only tell us about them.

Technical. Under relevance, independence, exclusion and monotonicity, $(E[Y | Z=1] - E[Y | Z=0]) / (E[D | Z=1] - E[D | Z=0]) = E[Y(1) - Y(0) | D(1) > D(0)]$ (Imbens & Angrist 1994). Different instruments have different compliers, so different LATEs.

Case studies

Kids and mothers' work (Angrist & Evans 1998). Instrument: the first two children are the same sex, so parents try for a third. 43.2 % vs 37.2 % had a third child (first stage +6 points). Effect on "worked for pay": OLS -0.176 , 2SLS -0.120 . Only about 6 % of families are compliers, so this is their effect, not every mother's.

Institutions and growth (Acemoglu, Johnson & Robinson 2001). Instrument: settler mortality. Where Europeans died, they set up extractive institutions that persisted. OLS 0.52 vs IV **0.94**. Nigeria vs Chile: OLS explains a 2-fold income gap, IV a 7-fold gap. The results survive many extra controls, but the exclusion restriction (mortality affects income only through institutions) remains debated: the disease environment could matter directly.

The method in five questions

Instrumental variables in five questions. The question to ask every time: ***Why should my instrument be independent of unobserved causes of the outcome?***

1. **What problem are we solving?** D is endogenous (omitted variables, reverse causality, measurement error), so OLS is biased.
2. **How does the method solve it?** Use only the variation in D that comes from an outside "knob" Z: effect of Z on Y divided by effect of Z on D (2SLS).
3. **What must hold?**
 - Relevance: Z moves D.
 - Independence: Z is as good as random.
 - Exclusion: Z affects Y only through D.
 - Monotonicity (for LATE): no defiers.
4. **What happens when it fails?**
 - Weak first stage: erratic estimates and unreliable usual SEs; in the textbook case they drift towards OLS.
 - Invalid Z: biased, and a weak first stage magnifies the bias.
 - Effects differ between people: you get a LATE for compliers, not an ATE.
5. **How can it be checked?**
 - First-stage F (relevance; ≥ 10 classic, ~ 104 for plain t-tests).
 - Balance of covariates across Z (partial check of independence).
 - Argue exclusion; try placebo outcomes Z shouldn't affect.
 - With several instruments, an overidentification test.

R lab 5 · A charter-school lottery

Packages: AER (for ivreg), ggplot2. In the simulation, motivated students attend more *and* score higher, so OLS is biased. The true effect of attending is 0.5 SD.

R CODE · LAB5.R

```
# Lab 5: a charter-school lottery as an instrument
library(AER); library(ggplot2)
set.seed(5)
n ← 2000
motivation ← rnorm(n)                # unobserved: affects attending AND scores
win ← rbinom(n, 1, 0.5)              # Z: lottery offer, random
```

```
# D: winners attend unless very unmotivated; losers attend only if very motivated
attend ← ifelse(win = 1, motivation > -1, motivation > 1.5) * 1
score ← 0.5 * attend + 0.6 * motivation + rnorm(n) # true effect = 0.5 SD
d ← data.frame(score, attend, win)

coef(lm(score ~ attend, data = d))["attend"] # OLS: mixes effect with motivation

fs ← mean(d$attend[d$win == 1]) - mean(d$attend[d$win == 0]) # first stage
rf ← mean(d$score[d$win == 1]) - mean(d$score[d$win == 0]) # reduced form
round(c(first_stage = fs, reduced_form = rf, wald = rf / fs), 3)

iv ← ivreg(score ~ attend | win, data = d) # 2SLS with correct SEs
summary(iv, diagnostics = TRUE)

# Weak vs strong instrument: 1,000 simulated studies each
sim ← function(strength) t(replicate(1000, {
  u ← rnorm(200); z ← rnorm(200)
  x ← strength * z + u + rnorm(200) # u makes x endogenous
  y ← 0.5 * x + u + rnorm(200)
  c(est = cov(z, y) / cov(z, x),
    F = summary(lm(x ~ z))$fstatistic[[1]]))})
strong ← sim(1); weak ← sim(0.1)
round(rbind(strong = c(median_F = median(strong[, "F"]), median_est = median(strong[, "est"])),
  weak = c(median_F = median(weak[, "F"]), median_est = median(weak[, "est"]))), 2)
sims ← data.frame(est = c(strong[, "est"], weak[, "est"]),
  type = rep(c("Strong instrument", "Weak instrument"), each = 1000))
ggplot(subset(sims, abs(est) < 4), aes(est)) +
  geom_histogram(bins = 60, fill = "grey60", colour = "white") +
  geom_vline(xintercept = 0.5, linetype = 2) + facet_wrap(~type) +
  labs(x = "IV estimate (true effect = 0.5)", y = "Count") + theme_minimal()
```

OUTPUT

```
attend
0.9456
  first_stage reduced_form      wald
      0.768      0.332      0.432

Call:
ivreg(formula = score ~ attend | win, data = d)

Residuals:
    Min       1Q   Median       3Q      Max
-4.77483 -0.79300  0.00443  0.81692  3.93551

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0706     0.0418    1.69   0.091 .
attend        0.4319     0.0698    6.19 7.4e-10 ***

Diagnostic tests:
              df1 df2 statistic p-value
Weak instruments    1 1998      2916 <2e-16 ***
Wu-Hausman          1 1997      150 <2e-16 ***
Sargan              0 NA        NA     NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.2 on 1998 degrees of freedom
Multiple R-Squared: 0.0984, Adjusted R-squared: 0.0979
Wald test: 38.3 on 1 and 1998 DF, p-value: 7.4e-10

      median_F median_est
strong    99.62      0.51
weak      1.08      0.70
```

How to read it.

- **OLS 0.946:** almost double the truth, because attendance also picks up motivation.

- **first_stage 0.768, reduced_form 0.332, wald 0.432:** the Wald ratio by hand.
- **ivreg attend 0.432 (SE 0.070):** same number with correct standard errors. The 95 % CI, about 0.43 ± 0.14 , covers the true 0.5.
- **Weak instruments test, $F = 2916$:** a very strong first stage.
- **Wu-Hausman $p < 0.001$:** OLS and IV differ by more than chance, a sign that attendance is endogenous.
- **Sargan NA:** with one instrument there is nothing to cross-check (exactly identified).
- **Simulation:** strong instrument, median estimate 0.51; weak instrument, 0.70, pulled towards the biased OLS value (about 1.0 in that simulation).

Spot the mistake.

1. "The first-stage F is 25, so the instrument is valid."
2. "The IV estimate is the average effect of the treatment for the whole population."
3. "The IV estimate (0.43) has a larger standard error than OLS (0.95), so OLS is the better estimate."

Practice 5

Level 1 • Understanding

1. List the three IV conditions. Which one can be tested, and how?
2. Define first stage and reduced form in one sentence each.

Level 2 • Calculation

3. Lottery winners: 60 % attend; losers: 10 %. Test scores differ by 0.2 between winners and losers. Compute the LATE and the shares of compliers, always-takers and never-takers.
4. $\text{Cov}(Z, Y) = 0.6$ and $\text{Cov}(Z, D) = 0.2$. What is the IV estimate?

Level 3 • Critical thinking

5. Rainfall is proposed as an instrument for farm income in a study of child health. Give one reason the exclusion restriction might fail.
6. The first-stage F is 12. A classmate says "above 10, so no weak-instrument problem". Respond.
7. In Angrist & Evans, is -0.12 the effect of a third child for all mothers with three children?

R questions

8. Why is the OLS estimate in the lab so much larger than the true 0.5?
9. In the weak-instrument simulation, why do the estimates drift towards the OLS value rather than towards zero?

Regression discontinuity

Where this fits. IV needed an outside push that is as good as random. Many policies create one through a **rule**: a cut-off in a score, an age or a date. Near the cut-off, who ends up on which side is close to random. A sharp RDD compares the two sides directly; a fuzzy RDD is IV (Lecture 5) applied right at the cut-off.

3 key ideas

1. Treatment depends on a **running variable X crossing a cut-off c**. Compare outcomes just below and just above c.
2. Key assumption: **nothing else jumps at c**, and people can't precisely sort themselves to one side. Check with a density test, covariate balance and placebo cut-offs.
3. Use **local linear regression near c**, not high-order polynomials over all the data. The effect is **local**: it applies to units at the cut-off.

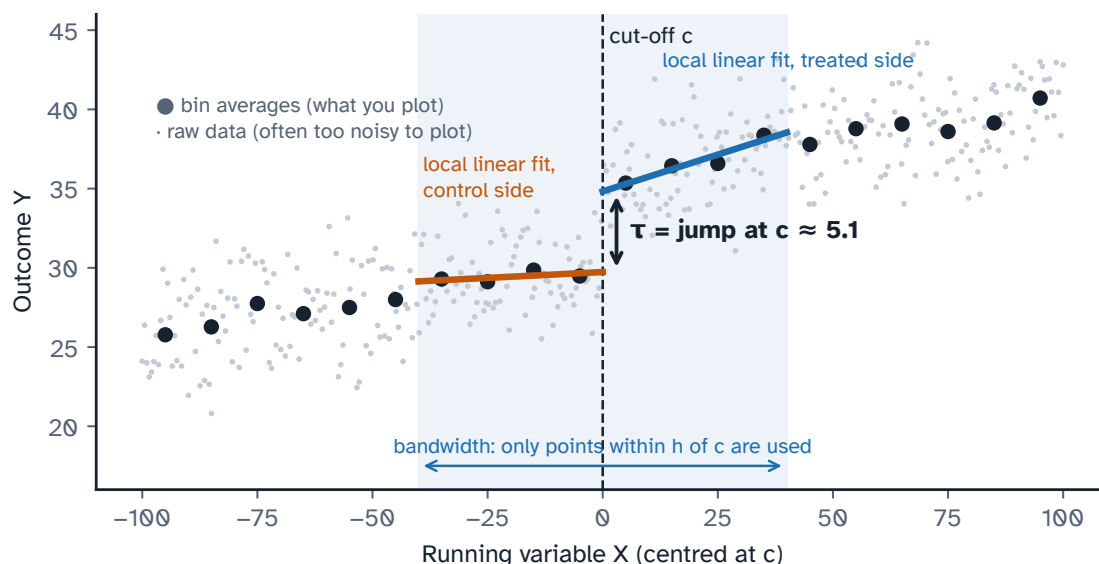
After this chapter it should be possible to...

- FOUNDATIONAL** Explain the running variable, the cut-off and the sharp/fuzzy distinction.
- INTERMEDIATE** Interpret the jump as a local effect, valid under continuity.
- INTERMEDIATE** Explain the bias-variance trade-off in choosing a bandwidth.
- INTERMEDIATE** Recognise threats (sorting, other policies at c) and run the standard checks.
- ADVANCED** Read rdrobust output, including robust bias-corrected intervals.

Interactive version: the bandwidth simulator at sweetheartapi.github.io/learn_econometrics/#l6/l6-demo

The idea

A student who scored 79 and one who scored 80 may be almost the same student. If only the 80 gets a scholarship, comparing them can be close to a coin flip. But closeness is not the assumption. The comparison works only **if** (1) nothing else changes at the cut-off (average potential outcomes are continuous through c) and (2) people cannot precisely control which side they land on. Under (2), Lee (2008) shows that RD behaves like a local randomised experiment. Both are assumptions to argue and check, never facts that follow from being near c .



An RD plot, annotated. Dots are bin averages of the raw data. Separate local linear fits are estimated on each side, using only data within the bandwidth (shaded). The vertical gap between the two fits at c is the estimated effect.

If the basic plot shows no jump, fancy regressions won't credibly find one. **Always start with the plot.**

Sharp RD

$$D_i = 1\{X_i \geq c\} \quad \tau = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x] = E[Y(1) - Y(0) \mid X = c]$$

X = running (forcing) variable: a score, age or date of birth. Knowing X tells you D exactly: that's "sharp".

Why does this work? The continuity argument, step by step

1. Just above c everyone is treated, so the average outcome there is the average of Y(1) at c.
2. Just below c no one is treated, so the average outcome there is the average of Y(0) at c.
3. **Continuity assumption:** without the treatment, average outcomes would change smoothly through c. Nothing else jumps exactly at c.
4. So E[Y(0)] just above c equals E[Y(0)] just below c, which is observed. The control side supplies the treated side's missing counterfactual.
5. The gap between the two limits is therefore E[Y(1) – Y(0)] **at X = c**. Away from c the effect could be different, and the data can't tell by how much.

In practice: the drinking age (Mastering 'Metrics ch. 4). In the US, drinking is legal from 21. Running variable = age in months, cut-off = 21st birthday. Death rates of young adults jump right at 21: a linear RD estimate is about **7.7 extra deaths per 100,000** per year, mostly from car accidents.

Threats: gaming the cut-off

Threat	What it looks like	What to do
Bunching / manipulation	Too many people just above c, too few just below	Density test (McCrary; now rddensity)
Donut hole	People avoid being right at c	Donut RD: drop the closest points and check the estimate
Heaping	X piles up at round numbers	Careful: units at heaps may differ
Mass point at c	Many units exactly at c	Count them; re-estimate without them

Validity checklist: (1) no jump in the density of X at c; (2) characteristics fixed before treatment don't jump at c; (3) no jumps at fake (placebo) cut-offs; (4) donut RD is stable; (5) the estimate is stable across bandwidths.

In practice. Babies under 1,500 g are classified "very low birth weight" and get extra care. If weights are rounded or nudged just under 1,500 g, babies at the threshold are no longer comparable; Barreca et al. (2011) checked this with a donut RD. Similarly, when a tax rate jumps at an income threshold, incomes often bunch just below it.

Estimating it

$$Y_i = \alpha + \tau D_i + \beta(X_i - c) + \gamma D_i(X_i - c) + \varepsilon_i$$

- τ = the jump at c : **the effect**.
- β = slope on the left; $\beta + \gamma$ = slope on the right.
- Centring X at c makes τ the jump measured exactly at the cut-off.
- **Avoid high-order global polynomials**: far-away points get big, erratic weight. In the slides' simulation (true $\tau = 1$), estimates ranged from 0.55 to 2.2 depending on the order (Gelman & Imbens 2019).
- **Local linear (recommended)**: use only points with $|X - c| \leq h$ (the **bandwidth**), weighting closer points more (triangular kernel).
- Use **robust bias-corrected** confidence intervals (rdrobust). Report $\hat{\tau}$, SE, h , kernel and polynomial order.

Choosing the bandwidth

Small h = less bias (only very similar units) but more variance (fewer points). Large h = more data but more bias. $MSE = \text{bias}^2 + \text{variance}$; rdrobust picks the h that minimises it. A credible result should not hinge on one particular h .

Bandwidth sensitivity (from R lab 6, true effect 0.045):

Bandwidth h (days)	Estimate	Robust 95 % CI
30	0.074	0.043 to 0.190
60	0.059	0.025 to 0.129
90	0.048	0.025 to 0.111
150	0.043	0.017 to 0.083

All four intervals are positive and all contain the truth. The smallest window is the noisiest (widest interval); the largest is the most precise but leans most on the straight-line assumption far from c . The conclusion, "a positive effect of a few percentage points", is stable.

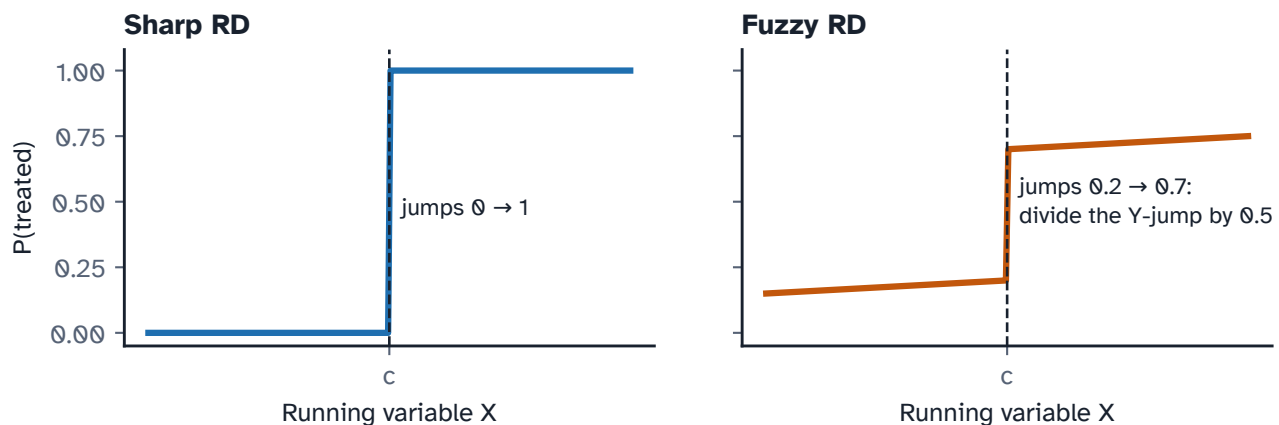
Three ways to explain: bandwidth

Simple. How far from the cut-off you look.

Intuitive. Comparing students who scored 79 with those who scored 80 compares very similar people, but there are few of them. Comparing 70–79 with 80–89 gives far more people, but they are less alike. The bandwidth is the size of that window.

Technical. Local polynomial fit with kernel weights $K((X_i - c)/h)$. For local linear, bias is of order h^2 and variance of order $1/(nh)$, so the MSE-optimal h shrinks like $n^{-1/5}$. At that h , conventional CIs under-cover; robust bias-corrected CIs (Calonico, Cattaneo & Titiunik 2014) fix this.

Sharp vs fuzzy RD



In a sharp design the probability of treatment jumps from 0 to 1 at c . In a fuzzy design it only jumps part of the way (here from 0.2 to 0.7). The jump in the outcome must then be scaled up by the jump in take-up.

	Sharp RD	Fuzzy RD
Treatment at c	jumps from 0 to 1	probability jumps by less than 1
Formula	jump in Y	jump in $Y \div$ jump in $P(D = 1)$
Whose effect	units at c	compliers at c (a LATE)
Estimation	local linear regression with $D = 1\{X \geq c\}$	2SLS in the same window: $Z = 1\{X \geq c\}$ instruments actual treatment D (rdrobust with fuzzy =)
Assumptions	continuity	continuity + relevance + exclusion + monotonicity
Extra check	—	size of the first-stage jump (a small jump = weak instrument)

Why does this work? Fuzzy RD = IV at the cut-off

1. Crossing c changes **eligibility** Z , not treatment itself. Think of Z as a lottery that is only "random" right at c .
2. The jump in Y at c is the effect of eligibility: the ITT at the cut-off.
3. The jump in $P(D = 1)$ at c is the first stage: the share of compliers at c .
4. As in Lecture 5, $\text{ITT} \div \text{first stage} = \text{the effect for compliers}$. So $\tau_{\text{FRD}} = \text{jump in } Y / \text{jump in } D$.

In practice: Boston exam schools (Mastering 'Metrics ch. 4.2). Admission to elite public schools depends on an entrance-exam cut-off. Scoring just above it sharply raises the chance of enrolling, but not from 0 to 1, so it is a fuzzy RD. The 2SLS estimate of the effect on later math scores is about **zero**: the schools' strong results reflect who gets in, not what the school adds ("the elite illusion").

Case study: Austrian parental leave (Lalive & Zweimüller 2009)

For children born on or after **1 July 1990**, parental leave went from 1 to 2 years. Running variable: the child's birth date. June 1990 mothers = control; July 1990 mothers = treated.

Outcome	Effect of July vs June birth
Leave taken in year 2	+9.3 months
Another child within 36 months	+4.5 percentage points (mean 0.345)
Back at work within 36 months	-10.9 points
Back at work within 10 years	about -3 points
Earnings, years 0-3	about -€2.7 per day
Earnings, years 4-10	no significant difference

Checks: no spike in births on 1 July; dropping births within a week of the cut-off gives 0.056; a placebo comparison in 1987 gives 0.008 (about zero).

The method in five questions

Regression discontinuity in five questions. The question to ask every time: ***Could something other than the treatment change at the cut-off?***

- 1. What problem are we solving?** Treatment is assigned by a rule at a cut-off of a score, age or date, and treated people differ from untreated ones overall.
- 2. How does the method solve it?** Compare outcomes just above and just below the cut-off, with a local linear fit on each side within a bandwidth.
- 3. What must hold?**
 - Continuity: average potential outcomes change smoothly through c (nothing else jumps there).
 - No precise sorting: people can't choose their side.
 - The local linear fit is a good approximation inside the bandwidth.
- 4. What happens when it fails?**
 - Another policy at the same cut-off: the jump mixes both.
 - Manipulation: people just above differ from those just below.
 - Too wide a bandwidth or high-order polynomials: biased; too narrow: noisy.
- 5. How can it be checked?**
 - Density test (rddensity).
 - Covariates shouldn't jump at c .
 - Placebo cut-offs and donut RD.
 - Estimates across bandwidths; robust bias-corrected CIs (rdrobust).

R lab 6 · A simulated parental-leave reform

Packages: rdrobust, rddensity, ggplot2. 40,000 births within 180 days of 1 July; the true jump in "another child within 36 months" is 4.5 percentage points.

R CODE · LAB6.R

```
# Lab 6: a parental-leave reform as a regression discontinuity
library(rdrobust); library(rddensity); library(ggplot2)
set.seed(6)
n <- 40000 # births in one year
days <- sample(-180:180, n, replace = TRUE) # birth date relative to 1 July (cut-off = 0)
long_leave <- as.numeric(days >= 0) # sharp: born on/after cut-off → 2-year leave
p <- 0.32 + 0.00015 * days + 0.045 * long_leave # true jump = 4.5 percentage points
birth36 <- rbinom(n, 1, p) # another child within 36 months?
d <- data.frame(days, long_leave, birth36)

rd <- rdrobust(d$birth36, d$days, c = 0) # local linear, MSE-optimal bandwidth
summary(rd)

# Manipulation check: is there a jump in the density of birth dates?
round(rddensity(d$days, c = 0)$test$p_jk, 3) # p-value of the density test

# Bandwidth sensitivity: same model, different windows (in days)
suppressWarnings(t(sapply(c(30, 60, 90, 150), function(h) {
  r <- rdrobust(d$birth36, d$days, c = 0, h = h)
  round(c(h = h, estimate = r$coef[1], ci_low = r$ci[3, 1], ci_high = r$ci[3, 2]), 3)})))

# Fuzzy version: eligibility only raises take-up (90 % vs 10 %)
took_leave <- rbinom(n, 1, ifelse(long_leave == 1, 0.9, 0.1))
birth36_f <- rbinom(n, 1, 0.32 + 0.00015 * days + 0.05 * took_leave) # true effect 0.05
fz <- suppressWarnings(rdrobust(birth36_f, days, c = 0, fuzzy = took_leave))
round(c(fuzzy_estimate = fz$coef[1], ci_low = fz$ci[3, 1], ci_high = fz$ci[3, 2]), 3)

# RD plot: monthly bins, separate lines on each side
bins <- aggregate(birth36 ~ cut(days, seq(-180, 180, 15), right = FALSE), data = d, mean)
bins$mid <- seq(-172.5, 172.5, 15)
ggplot(bins, aes(mid, birth36)) + geom_point() +
  geom_smooth(data = d, aes(days, birth36, group = long_leave), method = "lm",
    formula = y ~ x, se = FALSE, colour = "black") +
  geom_vline(xintercept = 0, linetype = 2) +
  labs(x = "Birth date (days from 1 July)", y = "Share with another child") + theme_minimal()
```

OUTPUT

Call: rdrobust

Sharp RD estimates using local polynomial regression.

Number of Obs.	40000
BW type	mserd
Kernel	Triangular
VCE method	NN

	Left	Right
Number of Obs.	19821	20179
Eff. Number of Obs.	5870	6039
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	54.902	54.902
BW bias (b)	94.358	94.358
rho (h/b)	0.582	0.582
Unique Obs.	180	181

```
=====
              Point      Robust Inference
              Estimate      z      P>|z|      [ 95% C.I. ]
-----
RD Effect    0.061    3.083    0.002    [0.025 , 0.111]
=====
```

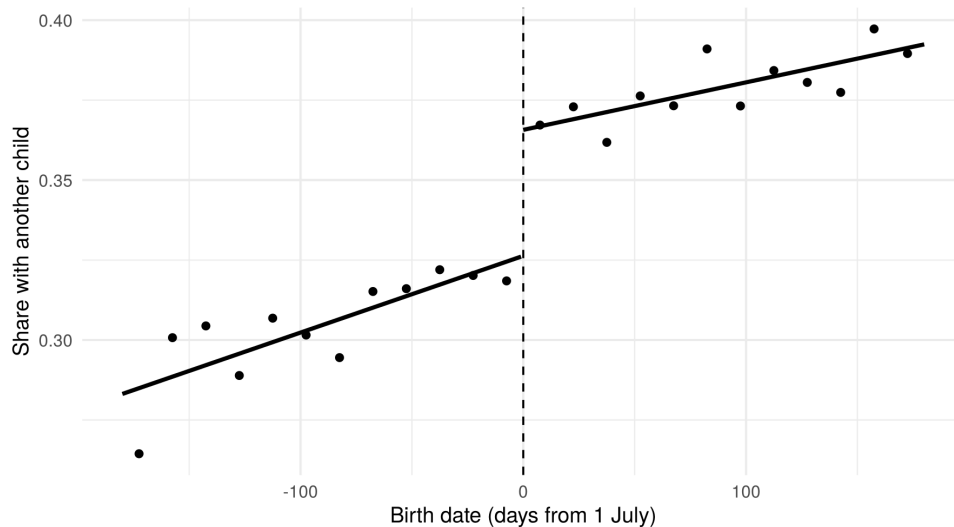
```
[1] 0.927
      h estimate ci_low ci_high
[1,] 30    0.074 0.043 0.190
[2,] 60    0.059 0.025 0.129
[3,] 90    0.048 0.025 0.111
[4,] 150   0.043 0.017 0.083
fuzzy_estimate      ci_low      ci_high
```

0.064 0.010 0.119

Warning message:

In `rdrobust(d$birth36, d$days, c = 0)` :

Mass points detected in the running variable.



RD plot from the lab: 15-day bins and a straight line on each side of 1 July. The jump at the cut-off is visible to the naked eye.

How to read it.

- **BW est. (h) 54.9:** `rdrobust` chose a window of ± 55 days. **Eff. Number of Obs.** (5,870 + 6,039) are the births inside it; the other births are not used for the estimate.
- **RD Effect 0.061, robust CI [0.025, 0.111], p = 0.002:** a clear positive jump. The interval contains the true 0.045.
- **Density test p = 0.927:** no sign of bunching in birth dates around 1 July (as built into the simulation).
- **Bandwidth table:** estimates from 0.043 to 0.074, all positive (see the box above).
- **fuzzy_estimate 0.064 [0.010, 0.119]:** with incomplete take-up, `rdrobust` scales the jump by the jump in take-up. It estimates the effect of actually taking long leave for compliers (true value 0.05).
- **"Mass points detected":** birth dates are whole days, so many births share each value of X. `rdrobust` adjusts for this automatically; the warning is a reminder that X is discrete.

Spot the mistake.

1. "The RD estimate of 0.061 shows that longer leave raises fertility by 6 points for all mothers, whenever their child was born."
2. "A 6th-order polynomial on the whole sample is the most flexible choice, so it gives the most reliable estimate."
3. "The density test p-value is 0.93, which proves that nobody manipulated birth dates."

Practice 6

Level 1 • Understanding

1. A scholarship is awarded to students scoring 80 or more. Why are students scoring 79 and 80 useful for estimating the scholarship's effect?
2. What is the difference between a sharp and a fuzzy RD?

Level 2 • Calculation

3. The average outcome immediately above the cut-off is 75 and immediately below is 68. What is the estimated sharp RD effect?
4. At a cut-off, the outcome jumps by 6 and the share treated jumps from 20 % to 70 %. Estimate the effect for compliers.
5. In the Austrian study, +4.5 percentage points on a mean of 34.5 % is roughly what percentage increase?

Level 3 • Critical thinking

6. A density test reveals a large jump in the number of students just above the cut-off. What does this mean for the conclusions?
7. Across bandwidths of 30, 60, 90 and 150 days, the estimates are 0.074, 0.059, 0.048 and 0.043. What does this pattern suggest, and what should be reported?

R questions

8. What do "BW est. (h)" and "Eff. Number of Obs." mean in the rdrobust output?
9. Why does the fuzzy estimate have a wider confidence interval than the sharp one?

Difference-in-differences

Where this fits. The simplest causal claim is "before vs after": things changed when the policy started. Lecture 4 explains why that fails: everything else that changed over time is mixed in. DiD fixes this with a control group that experienced the same time changes but not the policy. It is a regression with group and time effects (Lecture 1), and its standard errors need clustering.

3 key ideas

1. **DiD = (treated after – treated before) – (control after – control before).**
2. Key assumption: **parallel trends**. Without treatment, both groups would have changed by the same amount. Levels may differ; trends may not. It can't be tested directly.
3. With **staggered timing** and effects that change over time, plain two-way fixed effects (TWFE) can be badly wrong, because it uses already-treated units as controls. Clean comparisons (Callaway & Sant'Anna) avoid this.

After this chapter it should be possible to...

- FOUNDATIONAL** Compute a 2×2 DiD from four means and as a regression.
- FOUNDATIONAL** State parallel trends, no anticipation and no spillovers.
- INTERMEDIATE** Read an event-study plot: what leads can and can't show.
- INTERMEDIATE** Cluster standard errors at the right level.
- ADVANCED** Explain why TWFE fails with staggered timing and compute ATT(g, t).

Interactive version: the parallel-trends demo at sweetheartapi.github.io/learn_econometrics/#l7/l7-demo

John Snow's cholera "Grand Experiment"

London, 1852: the Lambeth water company moved its intake upriver, above the sewage. Southwark & Vauxhall (S&V) kept drawing dirty water. Both companies' pipes ran down the same streets, so neighbours breathed the same air but drank different water.

Cholera deaths per 10,000 households	1849	1854	Change
Lambeth (treated: clean water)	62	14	-48
S&V (control: dirty water)	565	614	+49

Why does this work? DiD step by step, with Snow's numbers

1. **Target:** the effect of clean water on Lambeth customers in 1854 = (actual 1854 Lambeth deaths) – (Lambeth deaths in 1854 *had the intake not moved*). The second number is never observed.
2. **Before/after alone:** $14 - 62 = -48$. This mixes the water effect with everything else that changed between 1849 and 1854.
3. **What else changed?** S&V customers had no water change, and their deaths rose by $614 - 565 = +49$: the epidemic got worse overall.
4. **Parallel trends:** without the move, Lambeth's deaths would also have risen by 49, to $62 + 49 = 111$. That is the counterfactual.
5. **Effect** = $14 - 111 = -97$ deaths per 10,000 households. That equals $(-48) - (+49)$, the difference in differences.
6. The levels (62 vs 565) never enter; only the changes do. That's why DiD needs parallel **trends**, not equal **levels**.

The 2×2 formula and regression

$$\widehat{DiD} = (\mu_{1,1} - \mu_{1,0}) - (\mu_{0,1} - \mu_{0,0}) \quad Y_{it} = \alpha + \beta G_i + \gamma Post_t + \delta (G_i \times Post_t) + \varepsilon_{it}$$

$\mu_{\{g,t\}}$ = average outcome for group g (1 = treated, 0 = control) in period t (1 = after, 0 = before).

- α = control group before. β = baseline gap between groups. γ = common change over time.
- **δ = the DiD effect:** the extra change for the treated group after treatment.

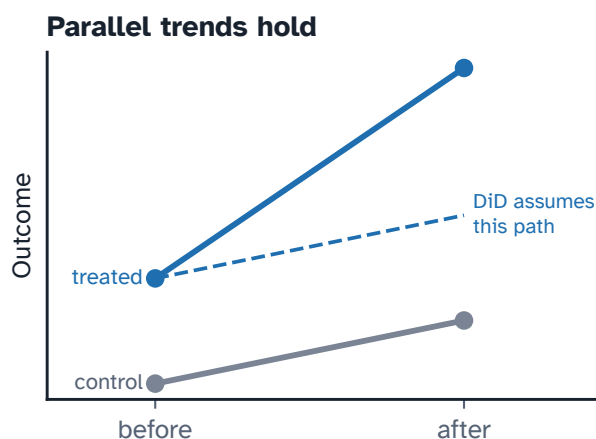
In practice: DiD by hand. City A introduces a sugar tax in January; neighbouring City B doesn't.

Weekly soda sales per shop: A goes $20 \rightarrow 15$ (-5), B goes $22 \rightarrow 21$ (-1, e.g. colder weather). $DiD = (-5) - (-1) = -4$. In the regression: $\alpha = 22$, $\beta = -2$, $\gamma = -1$, $\delta = -4$. The R lab estimates exactly this design.

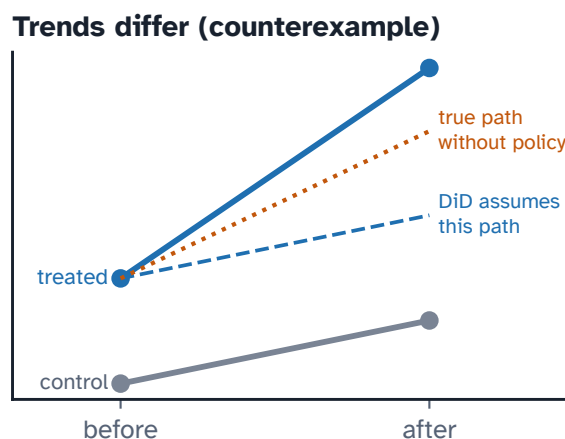
What you get and what you assume

Target: the **ATT**, the average effect on the treated group after treatment. It needs:

- **Consistency / SUTVA:** a well-defined treatment and no spillovers onto the control group.
- **No anticipation:** the treated don't react before treatment starts.
- **Parallel trends:** $E[Y_1(0) - Y_0(0) \mid \text{treated}] = E[Y_1(0) - Y_0(0) \mid \text{control}]$.



✓ **DiD = 10 – 3 = 7 = true effect**



✗ **DiD still says 7, true effect is 25 – 22 = 3**

Same observed data in both panels: treated 15 → 25, control 10 → 13. If trends are parallel (left), the treated would have reached 18 without the policy, and DiD = 7 is right. If the treated group was on a steeper path anyway (right, true counterfactual 22), DiD still reports 7 but the true effect is only 3.

In practice: when assumptions fail. In the sugar-tax example, some City A residents start buying soda in City B. B's sales rise because of the tax (a **spillover**), so the effect looks bigger than it is. If the tax was announced in November and people stocked up, A's "before" sales are inflated (**anticipation**). If B is a growing tourist town while A is shrinking, the trends were never **parallel**.

With covariates: if trends are parallel only within groups with the same X, compute DiD within each X group and average using the treated group's mix. In a regression, add $X \times \text{Post}$ (adding X alone does nothing). The slides' example: the correct ATT is 1.8 ($0.2 \cdot 1 + 0.8 \cdot 2$), while ignoring X gives 3.0.

Three ways to explain: parallel trends

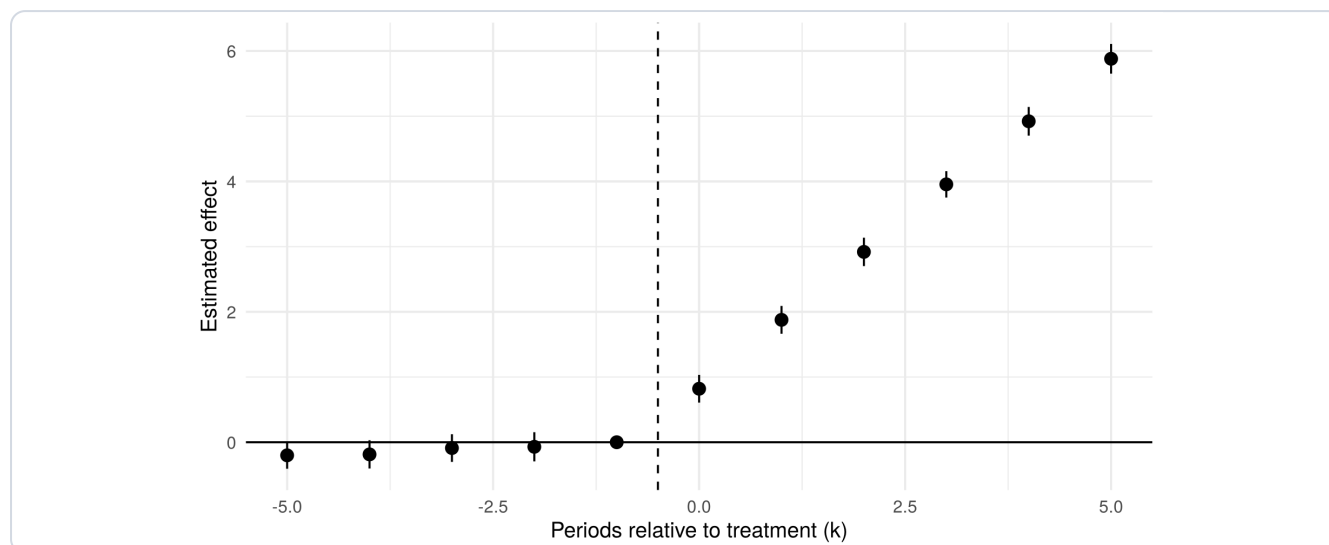
Simple. Without the policy, both groups would have changed by the same amount.

Intuitive. Two cities' café sales both grow about 3 % a year. City A introduces a sugar tax, and its sales then grow only 1 % while city B's still grow 3 %. That suggests the tax cost about 2 %, but only if A would otherwise have kept growing 3 % too.

Technical. $E[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid \text{treated}] = E[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid \text{control}]$. It concerns an unobserved post-period counterfactual, so it can't be tested directly, and it depends on the scale (parallel in levels $\neq$ parallel in logs).

Event studies

Replace the single treated $\times$ after dummy with one dummy per period relative to treatment ($k = -5 \dots 5$), leaving out $k = -1$ as the reference.



Event-study coefficients from R lab 7, with 95 % confidence intervals (clustered by region). Before treatment ($k < 0$) the estimates are close to zero; after it, the effect grows by about 1 per period.

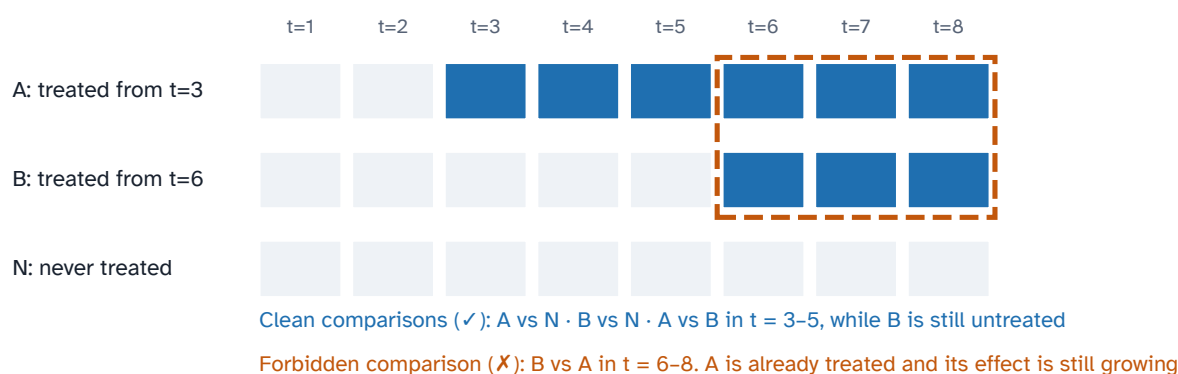
- **Leads** ($k < 0$) show whether the groups were already drifting apart before treatment.
- **Lags** ($k \geq 0$) show how the effect evolves. One average "after" number would hide a growing or fading effect.

Flat leads are evidence, not proof. Parallel trends is a claim about the *post*-treatment counterfactual, which is never observed. Pre-trend tests often have low power: a real but modest pre-trend can easily go undetected (Roth 2022). And choosing a specification *because* it passes the pre-trend test distorts inference. Leads near zero make parallel trends more plausible; they cannot establish it.

In practice. A retail chain opens a new store and studies sales at its nearby existing stores. Leads are close to zero; after opening, nearby stores lose 8 % of sales in month 1, shrinking to 3 % by month 12 as the novelty fades.

Staggered adoption: why TWFE breaks

When groups start treatment at different times, the usual regression is TWFE: $Y_{it} = \alpha_i + \lambda_t + \beta \cdot D_{it} + u_{it}$ (a fixed effect for each unit and each period). Goodman-Bacon (2021) shows that β is a weighted average of every 2×2 comparison in the data, including some that use **already-treated** units as the control group.



Three cohorts. Comparing A or B with the never-treated N is clean, and so is A vs B while B is still untreated (t = 3–5). Using A as the "control" for B in t = 6–8 is not: A is already treated, and its own effect is still growing.

Why does this work? Why the forbidden comparison gives the wrong answer

Take the slides' example (mean outcomes; A treated from t = 2, B from t = 3, N never):

	t = 1	t = 2	t = 3	t = 4
A (from t = 2)	10	16	21	24
B (from t = 3)	20	22	27	29
N (never)	30	32	35	36

- Clean:** B vs never-treated N, from t = 2 to t = 3: $(27 - 22) - (35 - 32) = 5 - 3 = 2$, B's true effect.
- Forbidden:** B vs already-treated A over the same periods: $(27 - 22) - (21 - 16) = 5 - 5 = 0$.
- Why 0? A's change from t = 2 to t = 3 includes its own effect growing from 4 to 6. Using A as a control subtracts that growth from B's effect.
- If effects grow fast enough, these forbidden comparisons can push the overall TWFE estimate down, even below zero.

The fix (Callaway & Sant'Anna 2021): never use already-treated units as controls. For each cohort g (first treated in period g) and each later period t, compute a clean 2 × 2 DiD, $ATT(g, t)$: the cohort's change from g – 1 to t, minus the change over the same periods for units not yet treated at t (or never treated). Then average these cells by event time, calendar time or overall.

Worked example: $ATT(2, 3)$ in the table above. Cohort A, from t = 1 to t = 3: $21 - 10 = +11$. Controls at t = 3: only N (B is treated by then), $35 - 30 = +5$. **$ATT(2, 3) = 11 - 5 = 6$.**

In practice. Countries introduce paid parental leave in different years, some in 2010 and others in 2016. If the effect on mothers' employment builds up over several years, TWFE compares the 2016 adopters with the 2010 adopters as if the latter were a clean control. Their growing effect gets subtracted, and the overall estimate is pulled down.

Standard errors

Outcomes of the same unit are correlated over time, so ordinary standard errors are far too small (Bertrand, Duflo & Mullainathan 2004). **Cluster at the level where treatment is assigned** (e.g. state). With very few treated clusters, even clustered SEs are unreliable.

In practice. A study of a state law uses survey data on 2 million workers in 50 states. The law varies by state, so there are only 50 independent pieces of information about it, not 2 million. Clustering by state gives honest (larger) standard errors.

Case study: minimum wage (Card & Krueger 1994)

April 1992: New Jersey raised the minimum wage from \$4.25 to \$5.05; Pennsylvania did not. 410 fast-food restaurants were surveyed before and after. NJ employment **rose by 2.33 full-time-equivalent workers per store** relative to PA, against the textbook prediction. Meal prices rose about 3.3 %. Neumark & Wascher (2000) found negative effects with payroll data; Card & Krueger (2000) found none with administrative data. The original paper showed no parallel-trends evidence, and with one treated and one control state, inference is fragile.

The method in five questions

Difference-in-differences in five questions. The question to ask every time: ***Why would the groups have followed parallel trends without treatment?***

1. **What problem are we solving?** A policy reaches some groups and not others. A before–after comparison mixes the effect with everything else that changed over time.
2. **How does the method solve it?** Subtract the control group's change from the treated group's change. With staggered timing, build clean 2×2 comparisons (ATT(g, t)).
3. **What must hold?**
 - Parallel trends: same change without treatment (levels may differ).
 - No anticipation before the start date.
 - No spillovers onto the control group.
 - Staggered timing with TWFE: effects must not change over time (otherwise use Callaway & Sant'Anna).
4. **What happens when it fails?**
 - Non-parallel trends: the bias equals the gap in trends.
 - Anticipation: changes appear before the policy.
 - TWFE with staggered, growing effects: wrong, sometimes wrong sign.
5. **How can it be checked?**
 - Event study: leads near zero (evidence, not proof).
 - Sensitivity to plausible violations (Rambachan & Roth).
 - Compare TWFE with clean estimators.
 - Cluster SEs where treatment is assigned.

R lab 7 · 2 × 2, event study and staggered adoption

Packages: sandwich, lmtest, dplyr, ggplot2. The regressions use base R's `lm()` with dummy variables, which gives exactly the same coefficients as the `fixest` package.

R CODE · LAB7.R

```
# Lab 7: difference-in-differences, event study and staggered adoption
library(sandwich); library(lmtest); library(dplyr); library(ggplot2)
set.seed(7)

# (a) 2x2: sugar tax in city A, not in city B; 100 shops each, before and after
shops <- expand.grid(shop = 1:200, post = 0:1)
shops$cityA <- as.numeric(shops$shop ≤ 100)
shop_fe <- rnorm(200, 0, 3)
shops$sales <- 22 - 2 * shops$cityA - 1 * shops$post - 4 * shops$cityA * shops$post +
  shop_fe[shops$shop] + rnorm(400)
m <- lm(sales ~ cityA * post, data = shops)
coeftest(m, vcov = vcovCL(m, cluster = ~shop)) # clustered by shop

# (b) Event study: 200 regions, half treated from period 0, effect grows over time
es <- expand.grid(unit = 1:200, t = -5:5)
es$treated <- as.numeric(es$unit ≤ 100)
es$y <- rnorm(200)[es$unit] + 0.3 * es$t + es$treated * pmax(es$t + 1, 0) + rnorm(nrow(es), 0, 0.5)
for (k in setdiff(-5:5, -1)) # one dummy per event time, k = -1 left out
  es[[paste0("k", ifelse(k < 0, paste0("m", -k), k))]] <- as.numeric(es$treated == 1 & es$t == k)
ev <- lm(as.formula(paste("y ~ factor(unit) + factor(t) +",
  paste(grep("^k", names(es), value = TRUE), collapse = " + "))), data = es)
ct <- coeftest(ev, vcov = vcovCL(ev, cluster = ~unit))[grep("^k", rownames(coef(summary(ev))))], 1:2]
plot_df <- data.frame(k = setdiff(-5:5, -1), est = ct[, 1], se = ct[, 2])
round(t(plot_df), 2)
ggplot(rbind(plot_df, data.frame(k = -1, est = 0, se = 0)), aes(k, est)) +
  geom_hline(yintercept = 0) + geom_vline(xintercept = -0.5, linetype = 2) +
  geom_pointrange(aes(ymin = est - 1.96 * se, ymax = est + 1.96 * se)) +
  labs(x = "Periods relative to treatment (k)", y = "Estimated effect") + theme_minimal()

# (c) Staggered adoption: cohort A treated from t = 3, B from t = 6, N never; effects grow
st <- expand.grid(unit = 1:150, t = 1:8)
st$g <- c(3, 6, Inf)[ceiling(st$unit / 50)]
st$D <- as.numeric(st$t ≥ st$g)
st$effect <- ifelse(st$D == 1, 2 * (st$t - st$g + 1), 0) # 2, 4, 6, ... after adoption
st$y <- rnorm(150)[st$unit] + 0.5 * st$t + st$effect + rnorm(nrow(st), 0, 0.5)
twfe <- coef(lm(y ~ factor(unit) + factor(t) + D, data = st))["D"]

att_gt <- function(gg, tt) { # clean 2x2: cohort gg vs not-yet/never treated at
  tt
  base <- gg - 1
  ctrl <- st$g > tt
  chg <- function(sel) mean(st$y[sel & st$t == tt]) - mean(st$y[sel & st$t == base])
  chg(st$g == gg) - chg(ctrl)
}
cells <- do.call(rbind, lapply(c(3, 6), function(gg)
  data.frame(g = gg, t = gg:8, att = sapply(gg:8, function(tt) att_gt(gg, tt)))))
true_avg <- mean(st$effect[st$D == 1])
round(cells, 2)
round(c(TWFE = unname(twfe), clean_average = mean(cells$att), true_average = true_avg), 2)
```

OUTPUT

t test of coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   22.341      0.321   69.62 < 2e-16 ***
cityA         -1.960      0.441   -4.45 1.1e-05 ***
post          -0.949      0.153   -6.20 1.4e-09 ***
cityA:post     -3.975      0.209  -18.99 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
      km5 km4 km3 km2 k0 k1 k2 k3 k4 k5
k    -5.0 -4.00 -3.00 -2.00 0.00 1.00 2.00 3.00 4.00 5.00
est -0.2 -0.19 -0.09 -0.07 0.82 1.88 2.92 3.95 4.92 5.88
se   0.1  0.11  0.11  0.11 0.11 0.11 0.11 0.10 0.11 0.12

  g t att
1 3 3 1.98
2 3 4 4.10
3 3 5 6.00
4 3 6 7.78
5 3 7 9.79
6 3 8 12.05
7 6 6 2.24
8 6 7 4.11
9 6 8 6.23

      TWFE clean_average true_average
      3.65           6.03           6.00
```

How to read it.

- **cityA:post –3.975 (clustered SE 0.209):** the DiD estimate; the true effect is –4. **cityA –1.96** is the baseline gap and **post –0.95** the common time change, matching α , β , γ in the sugar-tax box.
- **Event study:** leads between –0.07 and –0.20, lags rising from 0.82 to 5.88 (true: 1 to 6). There is no pre-trend in the simulation; the small negative leads come from noise in the reference period $k = -1$, which every coefficient is measured against. The one at –0.20 (SE 0.10) is borderline, a reminder that a single lead is not decisive either way.
- **Cell table:** each ATT(g, t) is a clean 2×2 and recovers the true growing effects (2, 4, 6, ...) closely.
- **TWFE 3.65 vs clean average 6.03 vs truth 6.00:** with growing effects and staggered timing, TWFE understates the average effect by about 40 %.

Same analysis with the specialised packages (not run here; these packages weren't available in this environment). In Positron:

R CODE • NOT RUN HERE

```
library(fixest) # TWFE and event studies, clustered SEs built in
feols(sales ~ cityA:post | shop + post, data = shops, cluster = ~shop)
library(did)    # Callaway & Sant'Anna
att_gt(yname = "y", tname = "t", idname = "unit", gname = "g_did", data = st_did)
# g_did: first treated period, with 0 for never-treated units (the did package convention)
```

Spot the mistake.

1. "The pre-treatment coefficients are not statistically significant, so parallel trends is proven."
2. "TWFE always averages the treatment effects, so with staggered adoption it can't get the sign wrong."
3. "With 2 million workers in 50 states, standard errors clustered by worker are fine for a state-level policy."

Practice 7

Level 1 • Understanding

1. Explain parallel trends in your own words. Why don't the groups need the same level?
2. What do the leads and lags in an event study show?

Level 2 • Calculation

3. Treated group: $50 \rightarrow 70$. Control group: $40 \rightarrow 50$. Compute the DiD.
4. In $Y = \alpha + \beta G + \gamma \text{Post} + \delta(G \times \text{Post})$, $\alpha = 10$, $\beta = 5$, $\gamma = 3$, $\delta = 2$. Write the four cell means and check that $\text{DiD} = \delta$.
5. Using the A/B/N table, compute $\text{ATT}(2, 4)$ and $\text{ATT}(2, 2)$. (Hint: at $t = 2$, B is not yet treated.)

Level 3 • Critical thinking

6. Before a policy, the treated region's outcome was already growing faster than the control region's. What does this do to the DiD estimate, and what could be done?
7. A city bans cars in its centre; the "control" is the neighbouring district. Why might SUTVA fail?

R questions

8. Interpret the coefficient on `cityA:post` and its clustered standard error.
9. Why is the TWFE estimate (3.65) so far below the true average effect (6.00) in part (c)?

Putting it together

Which method?

The choice of method follows from where the randomness comes from. The questions below are asked in order; the first "yes" gives the method.

1. **Can treatment be randomised?** → Experiment: compare means (check balance).
2. **Is there something random that pushes the treatment** (a lottery, weather, birth timing) but affects Y only through the treatment? → **IV**.
3. **Does a rule switch treatment on at a cut-off** in a score, age or date? → **RDD** (fuzzy RDD if the cut-off only changes the odds of treatment).
4. **Did a policy reach some groups and not others, with data before and after?** → **DiD** (clean comparisons if adoption is staggered).
5. **None of these?** → Regression with controls, read as association unless every important confounder is measured.

How the causal methods fit together

One causal question: what would have happened without the treatment? The best case is to randomise (Lecture 4). Otherwise each method finds a credible counterfactual somewhere else:

- **IV:** something random that pushes the treatment (a lottery, the weather, birth timing). Gives a LATE for compliers.
- **RDD:** a rule that switches treatment on at a cut-off. Gives the effect at the cut-off; a fuzzy RD is IV at the cut-off.
- **DiD:** a change over time that hits some groups and not others. Gives the ATT for the treated.

In every case the identification assumption decides whether the causal reading is credible: **exclusion** (IV), **continuity** (RDD), **parallel trends** (DiD).

How to recognise assumptions

Method	Key assumption	What a violation looks like	What can be checked in the data
OLS (causal reading)	No omitted confounders: $E[\varepsilon X] = 0$	Coefficient changes a lot when controls are added; a plausible confounder is missing	Sensitivity to controls; nothing can prove there is no unmeasured confounder
OLS (inference)	Homoskedastic, independent errors	Fan-shaped residuals; units grouped in clusters	Residual plots; use robust / clustered SEs anyway
Experiment	Random assignment, compliance, no attrition	Imbalanced baseline characteristics; people dropping out differently by group	Balance table; attrition rates by group
IV	Relevance	Small first-stage jump, low F	First stage (testable)
IV	Independence	Z related to background characteristics	Balance of covariates across Z (partial check)
IV	Exclusion	A plausible direct channel from Z to Y	Not testable; argue it; placebo outcomes
RDD	Continuity / no sorting	Bunching just above c; covariates jump at c; another policy uses the same c	Density test, covariate balance, placebo cut-offs, donut RD
RDD	Local functional form	Estimate swings with bandwidth or polynomial order	Bandwidth sensitivity
DiD	Parallel trends	Groups were already diverging before treatment	Event-study leads (diagnostic, not proof)
DiD	No anticipation / no spillovers	Changes before the start date; control group affected	Leads just before treatment; choose distant controls
Staggered DiD	Homogeneous effects (for TWFE)	Effects grow over time; TWFE and clean estimators disagree	Event study; Bacon decomposition; compare with Callaway & Sant'Anna

One question, four methods: does schooling raise wages?

The same research question can be answered with each method from this course. The answers differ, and the differences are informative rather than contradictory.

Method	Design	Typical result	What it estimates, and its weak spot
OLS	Regress log wage on years of school with controls	e.g. 0.10 per year	Association for everyone; biased upward if ability is omitted (Lecture 1 lab: 0.108 without ability vs 0.086 with it; truth 0.08)
IV	Quarter of birth or college proximity as instrument	Often similar or even larger than OLS (Angrist & Krueger: about 7 %)	LATE for people whose schooling was changed by the instrument, e.g. those who would have left school early; weak-instrument concerns
RDD	Compulsory-schooling cut-off by birth date, or an exam cut-off for an extra year	Effect at the cut-off	Only for people near the cut-off; needs no sorting around it
DiD	A reform that raised the minimum school-leaving age in some regions first	ATT for affected cohorts	Needs parallel trends between reform and non-reform regions; staggered rollout needs clean comparisons

Why the numbers differ:

- **Different populations:** OLS averages over everyone; IV over compliers; RDD over people near a cut-off; DiD over treated cohorts. If returns to school differ between people, these are different, legitimate numbers (see the ATE/ATT/ITT/LATE example in Lecture 4).
- **Different biases:** OLS suffers from omitted ability; IV from weak or invalid instruments; RDD from sorting; DiD from non-parallel trends.
- **What to do:** state the estimand, check each design's assumptions, and treat agreement across methods as stronger evidence than any single estimate.

Choose the design: seven research questions

Each case names a question and the data, but not the method. The useful habit is to ask how treatment was assigned, what comparison the data allow, and which estimand would result, before picking a method.

Answers, with the estimand and what to check, are in the answer key.

D1 RESEARCH DESIGN

Does a higher minimum wage affect employment? State A raises its minimum wage in 2025; neighbouring state B doesn't. You have quarterly employment for both states, 2022–2027. Which design fits best?

- Randomised experiment
- Difference-in-differences
- Regression with controls
- Instrumental variables
- Regression discontinuity

D2 RESEARCH DESIGN

Does a merit scholarship improve graduation? Every applicant with an entrance score of 75 or more gets the scholarship; nobody below does. You have scores and graduation for all applicants.

- a. Instrumental variables
- b. Difference-in-differences
- c. Regression with controls
- d. Fuzzy regression discontinuity
- e. Sharp regression discontinuity

D3 RESEARCH DESIGN

Does an extra year of school raise earnings? A compulsory-schooling law lets children leave at 16, so those born early in the year can leave with less schooling. You have quarter of birth, schooling and earnings for 300,000 men.

- a. Difference-in-differences
- b. Randomised experiment
- c. Instrumental variables with quarter of birth
- d. Regression of earnings on schooling with controls

D4 RESEARCH DESIGN

Does attending a charter school raise test scores? The school is oversubscribed and admits by lottery. 80 % of winners attend; 5 % of losers get in later anyway.

- a. Compare students who attend with those who don't.
- b. Compare winners with losers and call it the effect of attending.
- c. IV: the lottery as an instrument for attending (Wald = reduced form $\div$ 0.75)
- d. Regression discontinuity

D5 RESEARCH DESIGN

Does a new store layout raise sales? A retailer with 200 stores can choose which stores get the new layout next month.

- a. Difference-in-differences
- b. Instrumental variables
- c. Regression with controls
- d. Randomised experiment

D6 RESEARCH DESIGN

Free tutoring. Students scoring below 50 on a test are offered free tutoring. About 60 % of eligible students use it; about 10 % of ineligible students find a way in.

- a. Compare students who used tutoring with those who didn't.
- b. Sharp RD
- c. Fuzzy RD (IV at the cut-off)
- d. Difference-in-differences

Training and wages, no experiment. A cross-section of 5,000 workers: whether they did a training course, their wage, and many background variables. No policy change, lottery or cut-off.

- a. Regression with controls, read as association unless the controls plausibly capture everything that drives selection; add a sensitivity analysis.
- b. Difference-in-differences
- c. Instrumental variables with "did the course" as the instrument
- d. Regression discontinuity

Common exam traps

1. **Correlation as causation.** "Associated with" for regressions; "causes" only with a design that justifies it, and its assumptions stated.
2. **Forgetting "holding others fixed".** Every multiple-regression coefficient is conditional on the other regressors.
3. **Log interpretations.** log-level: $100 \cdot \beta$ %; level-log: $\beta/100$ units per 1 %; log-log: elasticity. Percent $\neq$ percentage points.
4. **Misreading p-values.** p is not the probability that H_0 is true, and $p > 0.05$ doesn't show "no effect".
5. **Misreading confidence intervals.** A computed 95 % CI doesn't have a 95 % chance of containing the truth; overlapping CIs don't imply no significant difference.
6. **Significance as importance.** A tiny effect can be highly significant in a large sample. Always discuss size.
7. **Precise means unbiased.** A small SE says nothing about omitted variable bias.
8. **Heteroskedasticity biases $\hat{\beta}$.** It doesn't (under exogeneity); it breaks the conventional SEs.
9. **Instrument validity from the F-statistic.** F only checks relevance; exclusion is never tested by it.
10. **IV, RDD and DiD give "the" effect.** They give a LATE, an effect at the cut-off and an ATT respectively.
11. **A robustness check as proof.** Flat event-study leads, a passed density test or a stable estimate across bandwidths make assumptions *more plausible*; none proves them.
12. **Controlling for everything.** Controlling for a mediator or a collider adds bias.
13. **Randomisation means identical groups.** Groups are balanced on average, not in every sample.
14. **Bootstrap as a small-sample cure.** It can't add information that isn't in the data.

Practice exam papers

How these papers work. Two mixed papers of 22 questions each, covering all seven lectures in random order, with calculation, conceptual and applied (R output) questions. A paper is most useful a few days after studying the chapters, closed-book, ideally timed: about 40 minutes. The answer key gives the right answer, why it is right, and **why each wrong option is wrong**, since the wrong options are the classic misconceptions. A chapter can count as mastered when about 80 % of its questions across both papers are right. The same questions, with instant feedback and a timer, are in exam mode on the companion website (sweetheartapi.github.io/learn_econometrics/#exam).

Paper A

A1 DID · CHALLENGE · CONCEPTUAL

Spot the mistake. Event-study leads ($k = -5$ to -2): -0.12 , -0.09 , -0.06 , -0.03 , each with SE 0.10 , none significant. "This proves parallel trends."

- a. Correct: insignificant leads prove parallel trends.
- b. The leads should be significant for DiD to work.
- c. Leads are evidence, not proof: pre-trend tests often lack power (here a steady drift is visible), and the assumption concerns the post-period. Report sensitivity to violations.
- d. Leads don't matter; only lags matter.

A2 REGRESSION · CHALLENGE · CONCEPTUAL

Spot the mistake. "My residuals fan out as x grows. So heteroskedasticity has biased my $\hat{\beta}$, and I can't use this regression."

- a. The student should drop the observations with large x .
- b. Nothing is wrong: heteroskedasticity biases $\hat{\beta}$ upwards.
- c. $\hat{\beta}$ is still unbiased; what breaks are the usual standard errors. Use robust SEs (or WLS).
- d. The fix is to add more regressors until the fan disappears.

A3 CAUSALITY · CHALLENGE · CONCEPTUAL

A study with 2 million observations finds a very precise association between X and Y (tiny standard error). Explain why the sample size doesn't settle whether X causes Y .

A4 CAUSALITY · CHALLENGE · CONCEPTUAL

Spot the mistake. Among hospital patients, two diseases A and B are negatively correlated. A paper concludes that disease A protects against disease B.

- a. The sample is too small to see the positive correlation.
- b. The sample conditions on a collider: being in hospital is caused by A and by B, which creates a negative link even if none exists.
- c. There is no problem: a correlation in hospital data is causal.
- d. The paper should also control for hospitalisation.

A5 REGRESSION · CHALLENGE · CONCEPTUAL

In your own words: why doesn't a highly significant coefficient prove that X causes Y?

A6 IV · APPLY · CONCEPTUAL

Same lottery. Who are the 20 % of *losers* who attend anyway?

- a. Defiers
- b. Always-takers
- c. Compliers
- d. Never-takers

A7 REGRESSION · APPLY · APPLIED

R output for `lm(logwage ~ educ + exper)` (Lab 1, $n = 500$):

R CODE · NOT RUN HERE

	Estimate	Std. Error	t value	Pr(> t)
educ	0.10809	0.00594	18.20	<2e-16 ***

Which is (approximately) the 95 % confidence interval for the educ coefficient?

- a. It can't be computed without the p-value.
- b. [0.096, 0.120]
- c. [0.102, 0.114]
- d. [0.049, 0.167]

A8 CAUSALITY · APPLY · CALCULATION

Three people. Adam: $Y(1) = 5$, $Y(0) = 3$. Bea: $Y(1) = 4$, $Y(0) = 4$. Cem: $Y(1) = 2$, $Y(0) = 5$. What is the average treatment effect (ATE)? (Two decimals.)

Answer: _____

A9 BOOTSTRAP · CHALLENGE · CONCEPTUAL

You bootstrap the *maximum* of 50 uniform observations. 65 % of bootstrap maxima equal the sample maximum exactly. What does this tell you?

- a. The data cannot be uniform.
- b. The bootstrap fails here: the real sampling distribution has no such spike. Use subsampling instead.
- c. B was too small; increase it to 10,000.
- d. The estimate is very precise, because it barely varies.

A10 IV · CHALLENGE · CONCEPTUAL

You use distance to the nearest college as an instrument for years of schooling. A critic notes that families near colleges tend to live in cities with better-paid jobs. Which condition is threatened?

- a. Monotonicity
- b. Relevance
- c. None: controlling for schooling solves it.
- d. Validity: independence/exclusion. Distance may affect wages through local labour markets, not only through schooling.

A11 BOOTSTRAP · APPLY · CALCULATION

You sorted $B = 2,000$ bootstrap medians from smallest to largest. For a 95 % percentile interval, which position (rank) gives the lower end, approximately?

Answer: _____

A12 MAX LIKELIHOOD · APPLY · APPLIED

A logistic regression of O-ring failure on launch temperature gives a coefficient of -0.327 on temp, and $\exp(-0.327) = 0.72$. Which reading is right?

- a. Warmer launches cause 72 % fewer failures.
- b. Each extra degree lowers the probability of failure by 0.327.
- c. Each extra degree lowers the probability of failure by 28 percentage points.
- d. Each extra degree multiplies the odds of failure by about 0.72 (about 28 % lower odds).

A13 RDD · CHALLENGE · CONCEPTUAL

Spot the mistake. A student fits a 6th-order polynomial to the whole range and reports a jump of 2.1. A local linear fit near c gives 0.9.

- a. Prefer local linear near c ; high-order global polynomials are erratic at the boundary. Report bandwidth sensitivity.
- b. The polynomial uses more data, so 2.1 is more reliable.
- c. Use an 8th-order polynomial to check.
- d. Average the two: 1.5.

A14 REGRESSION · APPLY · APPLIED

You estimate $\log(\text{wage}) = 1.2 + 0.06 \cdot \text{educ} + 0.02 \cdot \text{exper}$. What is the right reading of 0.06?

- a. One more year of education causes wages to rise by 6 %.
- b. One more year of education is associated with wages about 6 % higher, for people with the same experience.
- c. One more year of education raises wages by €0.06.
- d. A 1 % rise in education is associated with 0.06 % higher wages.

A15 MAX LIKELIHOOD · APPLY · CALCULATION

100 waiting times, modelled as exponential with mean λ . The sample mean is 4.0 minutes. Using $SE(\hat{\lambda}) \approx \hat{\lambda}/\sqrt{n}$, what is the standard error?

Answer: _____ minutes

A16 IV · CHALLENGE · APPLIED

Spot the mistake. A student regresses D on Z, saves the fitted values $\hat{D}$, runs $\text{lm}(y \sim \text{Dhat})$ and reports that SE.

- a. The coefficient is right, but the SE is wrong: the residuals use $\hat{D}$ instead of D. Use `ivreg()`.
- b. Both the coefficient and the SE are wrong.
- c. Nothing is wrong; this is what `ivreg()` does internally.
- d. The coefficient is wrong because Z should be in the second stage too.

A17 DID · APPLY · CALCULATION

Staggered adoption. Mean Y: cohort B (treated from $t = 3$): $t=2$: 22, $t=4$: 29. Never-treated N: $t=2$: 32, $t=4$: 36. Compute $ATT(3, 4)$ using N as the control.

Answer: _____

A18 RDD · APPLY · APPLIED

Lab 6 output from `rdrobust(birth36, days, c = 0)`:

R CODE · NOT RUN HERE

```
BW est. (h)          54.902
Eff. Number of Obs.  5870 (left) 6039 (right)
      Estimate      Robust 95% C.I.
RD Effect      0.061    [0.025 , 0.111]
```

Which reading is right?

- a. The CI isn't centred on 0.061, so something is broken.
- b. The reform raised the chance of another birth within 36 months by about 6 percentage points for all mothers.
- c. The estimate uses all 40,000 births.
- d. For mothers giving birth near the cut-off, longer leave raised the chance of another birth by about 6 pp; the robust CI excludes zero.

A19 BOOTSTRAP · CHALLENGE · CONCEPTUAL

Spot the mistake. "I only have 8 observations, too few for the CLT. So I'll bootstrap, which gives exact 95 % intervals in small samples."

- a. The student should resample without replacement instead.
- b. The bootstrap is not a small-sample cure: with 8 values the intervals are lumpy and often cover the truth less than 95 % of the time.
- c. Nothing: the bootstrap is designed for tiny samples.
- d. The student should use $B = 100,000$ resamples to make up for the small n .

A20 MAX LIKELIHOOD · CHALLENGE · CONCEPTUAL

When does OLS give exactly the maximum likelihood estimate of β in $y = X\beta + \varepsilon$?

- a. Only when n is large.
- b. When the errors are heteroskedastic.
- c. Always, for any error distribution.
- d. When the errors are independent, normal and have constant variance.

A21 RDD · CHALLENGE · CONCEPTUAL

A scholarship goes to students who score at least 80. The histogram shows 40 students at 79 and 120 at 80. What should worry you?

- a. The cut-off should be moved to 79.
- b. The bandwidth is too large.
- c. Bunching just above the cut-off suggests manipulation (regrading, retakes). Students at 80 may differ from those at 79. Run a density test.
- d. Nothing: more students at 80 just means the test was easy.

A22 DID · CHALLENGE · CONCEPTUAL

Lab 7, staggered adoption with effects that grow over time: TWFE gives 3.65; the average of clean ATT(g, t) comparisons gives 6.03 (true 6.00). Why is TWFE so far off?

- a. TWFE has too few observations.
- b. TWFE also uses already-treated early units as controls for later units, which subtracts the early cohort's growing effect.
- c. The clean estimator ignores the never-treated group.
- d. TWFE is biased because of heteroskedasticity.

Paper B

B1 REGRESSION · APPLY · CONCEPTUAL

A model without an intercept, judged on new data, reports $R^2 = -0.2$. What does that tell you?

- a. 20 % of the variation is explained in the opposite direction.
- b. There must be a bug: R^2 can never be negative.
- c. The slope is negative.
- d. The model predicts worse than simply using the average of y .

B2 DID · APPLY · APPLIED

Lab 7, sugar tax in city A, $\text{lm}(\text{sales} \sim \text{cityA} * \text{post})$ with SEs clustered by shop:

R CODE · NOT RUN HERE

	Estimate	Std. Error
(Intercept)	22.341	0.321
cityA	-1.960	0.441
post	-0.949	0.153
cityA:post	-3.975	0.209

Which is the DiD estimate of the tax's effect?

- a. The total fall in city A, $-0.949 - 3.975 = -4.92$.
- b. $\text{cityA} = -1.96$
- c. $\text{post} = -0.95$
- d. $\text{cityA:post} = -3.98$: sales in city A fell about 4 units more than in city B.

B3 REGRESSION · CHALLENGE · CONCEPTUAL

The coefficient on educ has $p = 0.03$. Which statement is correct?

- a. If the model is right and $\beta = 0$, a t-statistic at least this extreme would turn up in about 3 % of samples.
- b. There is a 3 % chance that the null hypothesis ($\beta = 0$) is true.
- c. There is a 97 % chance that the effect is real.
- d. The effect of education is large and important.

B4 MAX LIKELIHOOD · CHALLENGE · CONCEPTUAL

Spot the mistake. "At the MLE the likelihood is only 0.0001, so there is just a 0.01 % chance that $\hat{p}$ is the right value."

- a. The likelihood must always be above 0.5 at the maximum.
- b. The likelihood is the probability of the data given p , not the probability of p . Its absolute size is meaningless; only comparisons across p matter.
- c. The student should have used the log-likelihood, which would give a probability of $\log(0.0001)$.
- d. Nothing: a likelihood of 0.0001 means $\hat{p}$ is a poor estimate.

B5 CAUSALITY · APPLY · CONCEPTUAL

You study whether a training programme (D) raises earnings (Y). Which variable should you *not* control for if you want the total effect?

- a. Education completed before the programme.
- b. The type of job people hold after the programme, which training helps them get.
- c. Age.
- d. Earnings in the year before the programme.

B6 MAX LIKELIHOOD · APPLY · CONCEPTUAL

Two log-likelihood curves both peak at $\hat{\theta} = 2$. Curve A is sharply peaked, curve B is flat around the top. Which is true?

- a. A has the larger SE because its log-likelihood changes faster.
- b. B gives the more precise estimate, because a flat curve is more stable.
- c. They are equally precise because the peak is in the same place.
- d. A gives the more precise estimate: more curvature means more information and a smaller SE.

B7 BOOTSTRAP · APPLY · APPLIED

Lab 3 output for the median of 20 basket values:

R CODE · NOT RUN HERE

```
original bias std. error
t1*    30.05 -0.4944    6.562
Percentile 95%: (18.50, 38.80)
```

Which report is best?

- a. "The median is €30.05 ± €0.49."
- b. "The median basket is €30.05 (bootstrap SE €6.6; 95 % percentile CI €18.5 to €38.8)."
- c. "The true median is 30.05 – (–0.49) = €30.54."
- d. "There is a 95 % probability that the true median is between €18.5 and €38.8."

B8 BOOTSTRAP · CHALLENGE · CONCEPTUAL

Raising the number of resamples B from 1,000 to 100,000 will...

- a. be equivalent to collecting 100 times more data.
- b. make the 95 % interval cover the truth more often.
- c. shrink the standard error by a factor of 10.
- d. only reduce simulation noise in the bootstrap SE and interval.

B9 RDD · APPLY · CALCULATION

Students below a cut-off score are offered free tutoring. At the cut-off, the share using tutoring jumps from 0.2 to 0.7, and test scores jump by 0.15 SD. What is the fuzzy RD estimate?

Answer: _____ SD

B10 DID · CHALLENGE · APPLIED

A state-level law, 50 states, individual-level data over 10 years. At what level should you cluster the standard errors?

- a. State
- b. Year
- c. Individual
- d. No clustering needed with this many observations.

B11 MAX LIKELIHOOD · APPLY · CALCULATION

You flip a coin 150 times and get 60 heads. What is the maximum likelihood estimate of p (the chance of heads)?

Answer: _____

B12 RDD · APPLY · CONCEPTUAL

Narrowing the bandwidth from 90 to 30 days changes the robust CI from [0.025, 0.111] to [0.043, 0.190]. Why is it wider?

- a. Because the kernel changed.
 - b. The effect is larger, and larger effects have wider CIs.
 - c. Narrow bandwidths are always biased.
 - d. Fewer observations fall inside the window, so the estimate is noisier (more variance), even if bias is smaller.
-

B13 CAUSALITY · APPLY · CALCULATION

Treated people average $Y = 80$, untreated average $Y = 70$. A perfect study tells you the ATT is 4. How large is the selection bias?

Answer: _____

B14 BOOTSTRAP · APPLY · CONCEPTUAL

You have $n = 40$ observations. How do you draw one bootstrap sample?

- a. Draw 20 values without replacement.
 - b. Draw 40 values from your data with replacement.
 - c. Draw 1,000 values with replacement.
 - d. Draw 40 values from your data without replacement.
-

B15 RDD · CHALLENGE · CONCEPTUAL

At age 65 people become eligible for both a public pension and public health insurance. A study uses an RD at 65 to estimate the effect of the health insurance on health. What is the main problem?

- a. The effect is local.
 - b. Age can't be a running variable.
 - c. People can manipulate their age.
 - d. Another treatment changes at the same cut-off, so the jump mixes the effects of both policies.
-

B16 IV · CHALLENGE · CONCEPTUAL

Spot the mistake. The first-stage F-statistic is 4. The student reports the 2SLS estimate with an ordinary 95 % confidence interval.

- a. Nothing: $F = 4$ is above 1, so the instrument works.
- b. Run the second stage by hand to get correct SEs.
- c. The instrument is weak, so conventional SEs and CIs are unreliable. Use weak-instrument-robust (Anderson–Rubin) inference and say the instrument is weak.
- d. Add several more weak instruments to raise F .

B17 DID · CHALLENGE · CONCEPTUAL

Spot the mistake. "DiD is invalid here because unemployment was 12 % in the treated region and 5 % in the control region before the policy."

- a. Use the ratio 12/5 instead of the difference.
- b. Correct: groups must start at the same level.
- c. The levels must differ by less than 5 percentage points.
- d. Different levels are fine; parallel trends only requires the same change without treatment.

B18 IV · APPLY · APPLIED

Lab 5 output, `ivreg(score ~ attend | win)`:

R CODE · NOT RUN HERE

```
              Estimate Std. Error
attend          0.4319      0.0698
Weak instruments F = 2916    p < 2e-16
OLS coefficient on attend: 0.9456
```

Which conclusion is justified?

- a. IV and OLS disagree, so the IV estimate must be wrong.
- b. The huge F shows the instrument is valid, so the IV estimate is causal.
- c. The effect is 0.43 SD for every student.
- d. The instrument is strong; for compliers attending raises scores by about 0.43 SD; OLS (0.95) is inflated by motivation.

B19 CAUSALITY · CHALLENGE · CONCEPTUAL

Spot the mistake. In a randomised experiment with 60 people, the treated are on average 41 years old and the controls 39. "Randomisation failed, so the experiment is useless."

- a. Randomisation balances groups on average, not in every sample. Small differences are expected; check with a balance test and you can control for age.
- b. Correct: randomised groups must be identical.
- c. Age differences can't matter for outcomes.
- d. The researcher should re-randomise until the ages match exactly, then drop the people that don't fit.

B20 IV · APPLY · CALCULATION

Lottery winners: 70 % attend the school. Losers: 20 % attend anyway. Winners score 0.30 SD higher on average. What is the LATE?

Answer: _____ SD

B21 DID · APPLY · CALCULATION

A company trains some employees. Productivity: trained 50 before, 75 after; untrained 40 before, 55 after. What is the DiD estimate?

Answer: _____ productivity units

True model: $\log(\text{wage}) = \dots + 0.08 \cdot \text{educ} + 0.10 \cdot \text{ability}$. Regressing ability on education gives $\gamma_1 = 0.15$. Ability is not in your data. What slope on education do you expect from the short regression?

Answer: _____

Formula sheet

Every formula from the course, with its meaning in words. This page works well printed separately as a reference sheet.

Formula	In words	Lecture
$y = X\beta + \varepsilon$	Outcome = coefficients $\times$ regressors + error	1
$\hat{\beta} = (X^T X)^{-1} X^T y$	OLS: the β with the smallest sum of squared residuals	1
$\hat{\beta}_1 = \text{Cov}(x, y) / \text{Var}(x)$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	One-regressor slope and intercept	1
$R^2 = 1 - \text{RSS}/\text{TSS}$	Share of y's variation explained in the sample	1
$t = \hat{\beta} / \text{se}(\hat{\beta})$	How many standard errors the estimate is from 0	1
$\text{CI} = \hat{\beta} \pm 1.96 \cdot \text{se}$ (large n)	95 % confidence interval	1
$\beta_1(\text{short}) = \beta_1 + \beta_3 \gamma_1$	Omitted variable bias; direction = $\text{sign}(\beta_3^*) \times \text{sign}(\gamma_1)$	1
$\text{Var}(\hat{\beta}_i) \propto 1/(1 - R_i^2)$	Collinearity inflates variance	1
$\text{AIC} = -2 \log L + 2p$	Fit minus complexity penalty; lower is better	1
$\ell(\theta) = \sum \log f(X_i \theta)$	Log-likelihood (f = probability or density)	2
$\hat{\theta} = \text{argmax } \ell(\theta)$; set $\ell'(\theta) = 0$	MLE	2
$\text{Var}(\hat{\theta}) \approx 1/(-\ell''(\hat{\theta})) \approx 1/(n \cdot J)$	Precision from curvature (information)	2
$\hat{p} = \text{heads} / \text{flips}$; $\hat{\lambda} = \bar{x}$	Two MLEs worth remembering	2
$\text{SE}_{\text{boot}} = \text{SD of the B bootstrap estimates}$	Bootstrap standard error	3
Percentile CI = 2.5th and 97.5th percentiles	Bootstrap interval (BCa adjusts for bias and skew)	3
$\text{ATE} = E[Y(1) - Y(0)]$	Average effect for everyone	4
$\text{ATT} = E[Y(1) - Y(0) D = 1]$	Average effect for the treated	4
$\text{naive} = \text{ATT} + \{E[Y(0) D=1] - E[Y(0) D=0]\}$	Naïve comparison = effect + selection bias	4
$\text{ITT} = E[Y Z=1] - E[Y Z=0]$	Effect of the offer/assignment	4, 5
$\hat{\delta}_{\text{IV}} = \text{Cov}(Z, Y) / \text{Cov}(Z, D)$	IV estimator	5
$\text{LATE} = \text{reduced form} / \text{first stage}$	Wald estimator; effect for compliers	5
$\text{first stage} = P(D=1 Z=1) - P(D=1 Z=0)$	Share of compliers	5
$D = 1\{X \geq c\}$	Sharp RD treatment rule	6
$\tau = \text{jump in } E[Y X] \text{ at } c$	Sharp RD effect at the cut-off	6
$\tau_{\text{FRD}} = \text{jump in } Y / \text{jump in } P(D = 1)$	Fuzzy RD = IV at the cut-off	6
$Y = \alpha + \tau D + \beta(X - c) + \gamma D(X - c) + \varepsilon$	Local linear RD regression	6
$\text{MSE} = \text{bias}^2 + \text{variance}$	Bandwidth trade-off	6

Formula	In words	Lecture
$DiD = (\mu_{1,1} - \mu_{1,0}) - (\mu_{0,1} - \mu_{0,0})$	Treated change minus control change	7
$Y = \alpha + \beta G + \gamma Post + \delta(G \times Post) + \varepsilon$	DiD regression; δ is the effect	7
$ATT(g, t) = \Delta \text{ cohort } g (g-1 \rightarrow t) - \Delta \text{ not-yet/never-treated}$	Clean staggered-DiD cell (Callaway & Sant'Anna)	7

Log interpretation: log y on x $\rightarrow 100 \cdot \beta$ % per unit; y on log x $\rightarrow \beta/100$ units per 1 %; log-log $\rightarrow \beta$ % per 1 % (elasticity). **Percentage points** = difference of two percentages; **percent** = change relative to the start.

Answer key

Answers to the "Spot the mistake" exercises and practice questions. They are most useful after a question has been attempted, with a reread of the matching section after any mistake.

Lecture 1 • Regression

Spot the mistake

1. R^2 measures fit, not trustworthiness. A coefficient can be precise and unbiased with a low R^2 ; what matters is its standard error and whether the regressor is exogenous.
2. It is an association, holding experience fixed, not a causal effect: ability is omitted (in the lab the true effect is 0.08). Also, in a log-level model the exact change is $e^{0.108} - 1 \approx 11.4\%$.
3. A fan shape means heteroskedasticity. Under exogeneity the coefficients are still unbiased; the conventional standard errors are wrong. Use robust SEs.

Practice

1. β_1 is the true, unknown population slope; $\hat{\beta}_1$ is the estimate computed from one sample. $\hat{\beta}_1$ would change with a different sample.
2. It means the model misses a non-linear pattern (wrong functional form). Predictions are systematically off and coefficients misleading; try x^2 , a log or an interaction.
3. For people of the same age, one more year of education is associated with about 6 % higher wages. Two years: about 12 % (exactly $e^{0.12} - 1 \approx 12.7\%$).
4. $t = 0.42/0.15 = 2.8 > 1.96$, so significant at 5 %. $CI = 0.42 \pm 1.96 \cdot 0.15 = [0.13, 0.71]$.
5. $0.05 + 0.10 \times 0.3 = 0.08$. Too high: it overstates by 0.03.
6. Underlying sickness: it lowers health ($\beta_3^* < 0$) and raises doctor visits ($\gamma_1 > 0$). Bias = negative $\times$ positive = negative, so the coefficient on visits is pushed down and can look harmful even if visits help.
7. Heteroskedasticity (or another violation of assumption 5). Report the robust SEs: they are valid either way.
8. $0.108 - 0.086 = 0.022$, about 20 % of the short coefficient.
9. The gap becomes about 0: ability no longer affects wages ($\beta_3^* = 0$), so leaving it out causes no bias, even though it is still correlated with education.

Lecture 2 · Maximum likelihood

Spot the mistake

1. The likelihood is not a probability over p . 0.27 is the probability of the observed data *if* $p = 0.3$. It says nothing like "27 % chance that $p = 0.3$ ".
2. The MLE is the value under which the observed **data** are most probable (or most dense). In frequentist statistics θ is fixed, not random, so "most probable θ " has no meaning here.
3. At a maximum the Hessian (second derivative of ℓ) is **negative**. The positive quantity used for SEs is the observed information, $-\text{Hessian}$: $\text{SE} \approx 1/\sqrt{-\ell''}$.

Practice

1. Probability: the chance of a discrete outcome, between 0 and 1. Density: the height of the curve for continuous data; areas give probabilities. Likelihood: the same formula seen as a function of the parameter, with the data fixed.
2. log is strictly increasing, so whatever maximises L also maximises $\log L$. Logs turn the product into a sum, which is easier to differentiate.
3. $\ell(p) = \text{const} + 3 \log p + 7 \log(1 - p)$; $\ell'(p) = 3/p - 7/(1 - p) = 0 \Rightarrow 3(1 - p) = 7p \Rightarrow \hat{p} = 0.3$.
4. $\text{SE} \approx 5/\sqrt{100} = 0.5$. With $n = 400$: $5/20 = 0.25$.
5. Odds are multiplied by $e^{(-0.327)} \approx 0.72$ per degree, about 28 % lower.
6. A flat log-likelihood means many values of p fit the data almost equally well: little information, a large standard error and a wide interval, usually because of a small sample.
7. A density is not a probability; it is a height, and probabilities are areas. A density concentrated on a narrow range (e.g. values between 0 and 0.5) must be higher than 1 for the area to equal 1.
8. optim minimises by default. Minimising the negative log-likelihood is the same as maximising the log-likelihood, and its Hessian is then the (positive) observed information.
9. The SE roughly halves (about 0.25): it shrinks with $1/\sqrt{n}$, so four times the data gives twice the precision.

Lecture 3 · Bootstrap

Spot the mistake

1. B only reduces simulation noise in the bootstrap distribution. The precision of the estimate depends on the sample size n , which doesn't change.
2. The bootstrap only resamples the data available; it can't add information. With 6 observations the empirical distribution is a crude stand-in, and intervals are unreliable.
3. Firm-years within a firm are correlated. Resample whole firms (all 10 years together): a cluster bootstrap.

Practice

1. Without replacement, every resample would be the same values reshuffled, so the statistic would never change.
2. The sampling distribution of the estimate: how it would vary across repeated samples from the population.
3. Any two samples of size 3 from {1, 3, 8} with repeats allowed, e.g. 1, 1, 8 (mean 3.33) and 3, 8, 8 (mean 6.33).
4. Percentile: [4.1, 7.9]. Normal: $6 \pm 1.96 \cdot 1.0 = [4.04, 7.96]$.
5. The bootstrap distribution of the median is skewed, so a symmetric normal interval is misleading. The percentile (or BCa) interval follows the skew and is more credible here.
6. The bootstrap does not solve small-sample problems. With 8 observations, the empirical distribution is a poor stand-in for the population and coverage can be poor. The small sample should be acknowledged, not presented as fixed.
7. The median of 20 numbers can only be one of the data values, or the average of two neighbouring ones. So bootstrap medians pile up on a few values.
8. The true sampling distribution of a maximum is continuous and below the true upper limit; it has no 65 % spike at one value. The bootstrap can't produce values above the observed maximum, so it misrepresents the uncertainty.

Lecture 4 · Causality

Spot the mistake

1. Randomisation balances groups **on average** across possible assignments, not exactly in one sample. Small chance differences are normal; a balance table shows them.
2. The ITT is the effect of being **offered** the voucher. If not everyone uses it, the effect of using it is $ITT \div \text{take-up difference}$ (the LATE).
3. Controlling for a mediator removes part of the effect; controlling for a collider creates bias. More controls are not always better.

Practice

1. For each unit only one potential outcome is observed ($Y(1)$ or $Y(0)$), so individual causal effects can never be seen directly.
2. ATE: the average effect if everyone were treated vs no one. ATT: the average effect among those who actually were treated.
3. Treated mean 9, untreated mean 5, difference 4. Causal if treatment was assigned (as good as) randomly.
4. First stage = 0.30. $ITT = 0.30 \times 8 = 2.4$. $LATE = 2.4/0.30 = 8$. $ATE = 0.3 \cdot 8 + 0.1 \cdot 2 + 0.6 \cdot 0 = 2.6$.
5. Selection bias = $12 - 5 = 7$.
6. Skills are a mediator (training $\rightarrow$ skills $\rightarrow$ wage). Controlling for them removes the main channel, so the estimate shows only the effect *not* through skills.
7. Admission depends on both test scores and motivation (both arrows point into "admitted"). Among the admitted, low scores must be compensated by high motivation and vice versa, creating a negative correlation that doesn't exist in the population: collider bias.
8. Chance: any one random split gives groups that differ a little in their potential outcomes. Only on average over many randomisations is the estimate equal to the ATE.
9. $5.179 + 11.601 = 16.780$ ✓.

Lecture 5 • Instrumental variables

Spot the mistake

1. The F-statistic only checks relevance. Validity (independence and exclusion) can't be tested this way; it must be argued.
2. IV estimates the LATE: the effect for compliers. It equals the population average only under strong assumptions (e.g. the same effect for everyone).
3. OLS is precise but biased (0.95 vs the true 0.5). A precise wrong answer is not better than a less precise right one.

Practice

1. Relevance, independence, exclusion. Only relevance is testable, through the first-stage regression of D on Z.
2. First stage: effect of Z on the treatment D. Reduced form: effect of Z on the outcome Y.
3. First stage = $0.6 - 0.1 = 0.5$; LATE = $0.2/0.5 = 0.4$. Always-takers 0.10, never-takers 0.40, compliers 0.50.
4. $0.6/0.2 = 3$.
5. Rainfall may affect child health directly (e.g. malaria, water-borne disease, food prices), not only through farm income.
6. $F = 12$ passes the old rule of thumb but is far below the level needed for reliable standard t-tests (about 104). Use weak-instrument-robust inference (e.g. Anderson–Rubin intervals) and interpret with caution.
7. No. It is the LATE for compliers: mothers who had a third child only because the first two were the same sex (about 6 % of families). It says nothing about always-takers.
8. Motivated students attend more and also score higher. OLS attributes the motivation effect to attending, so it is biased upward.
9. With a weak first stage, the IV ratio is dominated by noise and by the same endogenous variation that biases OLS, so in this standard setting (one instrument, valid Z, conventional 2SLS) its distribution is centred towards the OLS value. This is a typical finite-sample pattern rather than a universal rule; the conventional standard errors are unreliable as well.

Lecture 6 · RDD

Spot the mistake

1. The RD estimate is local: it applies to mothers giving birth around 1 July 1990, at the cut-off. Extending it further needs extra assumptions.
2. High-order global polynomials give far-away points large, erratic weight and are sensitive to the order chosen. Local linear regression near c is preferred.
3. Failing to reject is not proof. The test may lack power, and manipulation could also show up in covariates rather than the density. It is supporting evidence only.

Practice

1. They are almost identical in ability and background, and differ mainly in eligibility. Comparing them is close to a randomised experiment at the cut-off.
2. Sharp: treatment jumps from 0 to 1 at c . Fuzzy: only the probability of treatment jumps, so the outcome jump is divided by the take-up jump (IV at c).
3. $75 - 68 = 7$.
4. $6/(0.70 - 0.20) = 6/0.5 = 12$.
5. $4.5/34.5 \approx 13\%$.
6. It suggests manipulation: people sorted themselves just above the cut-off. They are probably different from those just below, so continuity fails and the RD estimate may be biased. Investigate how sorting happened; a donut RD or covariate checks can help, but conclusions should be cautious.
7. All estimates are positive with overlapping intervals, so the finding is robust in sign and rough size. The drift (higher with small h) is partly noise, partly curvature. Report the MSE-optimal estimate with robust CIs, and show the sensitivity table.
8. h : the window (± 55 days) used for estimation, chosen to minimise MSE. Effective observations: the births inside that window, the only ones contributing to the estimate.
9. It divides the outcome jump by the take-up jump (0.8). Dividing by a number below 1 scales up both the estimate and its uncertainty, and the first stage is itself estimated with noise.

Lecture 7 • DiD

Spot the mistake

1. Insignificant leads support parallel trends but cannot prove them: the assumption concerns the unobserved post-treatment counterfactual, and pre-trend tests often have low power.
2. With staggered adoption and changing effects, TWFE includes comparisons with already-treated units, which can produce estimates far from any average effect, even with the wrong sign.
3. The policy varies by state, so errors are correlated within states. Cluster by state; clustering by worker gives far too small standard errors.

Practice

1. Without the policy, the treated group's outcome would have changed by the same amount as the control group's. DiD compares changes, so constant level differences cancel out.
2. Leads: differences before treatment (a check on pre-trends). Lags: how the effect develops after treatment.
3. $(70 - 50) - (50 - 40) = 10$.
4. Control before = 10; control after = 13; treated before = 15; treated after = 20. $\text{DiD} = (20 - 15) - (13 - 10) = 2 = \delta$.
5. $\text{ATT}(2, 4) = (24 - 10) - (36 - 30) = 14 - 6 = 8$. $\text{ATT}(2, 2) = (16 - 10) - \text{average change of B and N from } t = 1 \text{ to } 2 = 6 - 2 = 4$.
6. Parallel trends is likely violated, so DiD would attribute part of the pre-existing faster growth to the policy (overstating it). Options: an event study to show the pre-trend, a better-matched control group, or methods that allow for trends, interpreted with caution.
7. Traffic and shoppers may move to the neighbouring district, changing its outcomes because of the treatment. The control group is contaminated, so DiD is biased.
8. City A's sales fell by about 3.98 units per shop relative to City B after the tax (true -4). The clustered SE of 0.21 accounts for repeated observations of the same shop.
9. With staggered timing, TWFE uses A (already treated, with a growing effect) as a control for B. Subtracting A's growing effect pulls the estimate down.

Choose the design

D1 b. Yes: a policy hits one group, with data before and after for both.

Estimand: ATT for state A. Key assumption: parallel trends in employment. Check pre-period quarters (event study), worry about spillovers across the border, and note that one treated state makes inference fragile (Card & Krueger 1994).

Why the other options are wrong:

- **a** You can't randomise minimum wages across states, and the policy has already happened.
- **c** With a policy change and a comparison group over time, DiD uses much weaker assumptions than "we controlled for everything".
- **d** There is no obvious instrument that shifts the minimum wage and affects employment only through it.
- **e** There is no running variable with a cut-off (a spatial RD along the border is an advanced variant, but the data here are state totals).

D2 e. Yes: treatment switches from 0 to 1 at a known cut-off of the running variable.

Estimand: the effect for students with scores around 75. Check: density of scores at 75 (could students retake?), balance of background characteristics, bandwidth sensitivity.

Why the other options are wrong:

- **a** With full compliance at the cut-off, the design is a sharp RD (a fuzzy RD would be IV at the cut-off).
- **b** There is no before/after policy change here; treatment is set by a score.
- **c** Scholarship holders have higher scores, so a simple comparison mixes the scholarship with ability. The cut-off gives a cleaner comparison.
- **d** Fuzzy applies when crossing the cut-off only changes the *probability* of treatment. Here everyone above gets it.

D3 c. Yes: quarter of birth shifts schooling and is (arguably) unrelated to ability (Angrist & Krueger 1991).

Estimand: LATE for men whose schooling was changed by the law (those who would have left early). Concerns: a weak first stage, and whether season of birth affects earnings in other ways (exclusion).

Why the other options are wrong:

- **a** There is no group-by-time policy change in the description.
- **b** You can't randomise years of schooling.
- **d** Ability is unobserved, so this suffers from omitted variable bias.

D4 c. Yes. First stage $0.80 - 0.05 = 0.75$; the result is a LATE for compliers.

Report both: ITT (winners vs losers) and LATE (ITT $\div$ first stage). This is the KIPP design from lecture 5.

Why the other options are wrong:

- **a** Families who choose to attend differ (motivation), so this has selection bias.
- **b** That gives the ITT (the effect of an *offer*). Not every winner attends and some losers do.
- **d** There is no score with a cut-off; admission is by lottery.

D5 d. Yes: if you can randomise, do. Assign the layout by lottery across stores.

Randomise across stores, check balance on past sales, cluster SEs by store if you use weekly data, and watch for spillovers between nearby stores. Combining randomisation with before/after data (a randomised DiD) adds precision.

Why the other options are wrong:

- **a** DiD is what you'd do if the retailer had already picked stores non-randomly. Randomising is cleaner.
- **b** No need for an instrument when you control assignment.
- **c** Choosing stores yourself (e.g. the best managers first) would create selection bias.

D6 c. Yes. Jump in outcome $\div$ jump in take-up (about $0.6 - 0.1 = 0.5$).

Estimand: LATE for compliers at the cut-off. Extra exclusion worry: being labelled "below 50" might itself affect motivation.

Why the other options are wrong:

- **a** Those who use tutoring choose to; selection bias.
- **b** Treatment doesn't jump from 0 to 1 at the cut-off; only the probability does.
- **d** No before/after change is described.

D7 a. Yes. Be honest about the assumption (selection on observables).

Estimand: the conditional association (causal only under "no unmeasured confounders"). Look for a design: e.g. course capacity limits, distance to training centres, or a rollout over time.

Why the other options are wrong:

- **b** You need data before and after a change for a treated and a comparison group.
- **c** The instrument must be something other than the treatment itself.
- **d** No cut-off assigns training.

Practice exam papers

Each answer gives the correct option first, then the reasoning, then what each wrong option gets wrong.

Paper A

A1 c. Yes (Roth 2022; Rambachan & Roth 2023).

A steady pattern like $-0.12, -0.09, -0.06, -0.03$ suggests the groups were converging before treatment. If that continued, it would contaminate the post-period estimates.

Why the other options are wrong:

- **a** Parallel trends is about the unobserved post-period. And here the leads drift steadily, which hints at a pre-trend the test is too weak to flag.
- **b** We hope leads are near zero; significant leads would be a warning sign.
- **d** Leads are the main diagnostic for pre-trends. They just can't prove the assumption.

A2 c. Yes. The slope is fine; t-tests and CIs built on the constant-variance formula are not.

Under exogeneity OLS stays unbiased with heteroskedastic errors; it is just no longer BLUE and the conventional SE formula is wrong. The regression playground on the companion website shows this with heteroskedastic noise and 500 repeated samples.

Why the other options are wrong:

- **a** Throwing away data is not a fix, and would lose the most informative points about the slope.
- **b** Heteroskedasticity does not bias $\hat{\beta}$ as long as $E[\epsilon | X] = 0$. It changes the spread of the errors, not their average.
- **d** The fan is about the error variance. Adding regressors is not the standard fix; robust SEs or weighted least squares are.

A3 Model answer. Sample size reduces *sampling* uncertainty, so the estimate becomes precise. But precision is about statistical inference; causation is about whether the comparison is fair. If the people with high X differ in other ways (selection bias, confounders, reverse causality), the estimate is biased, and a bias doesn't shrink with n. Two million rows give a very precise estimate of the wrong number. You need a design (randomisation, IV, RDD, DiD) or a convincing argument that nothing important is left out.

A self-check: an answer that names the same mechanism (in any words) has the key points.

A4 b. Right. This is Berkson's paradox: selecting on a common effect invents a correlation.

A collider K has arrows from both D and Y. Conditioning on K (by controlling for it or by selecting the sample on it) opens a path between D and Y that is not causal.

Why the other options are wrong:

- **a** The problem is who is in the sample, not how many. More hospital patients give a more precise wrong answer.
- **c** Correlation in a selected sample is not causal, and here the selection itself creates it.
- **d** Everyone in the sample is already hospitalised; that conditioning is exactly what causes the problem.

A5 Model answer. Significance only says the estimate is unlikely to be pure sampling noise *under the model*. It says nothing about whether the model is right. If X is linked to something left out (ability, motivation), if Y also affects X, or if people selected into X, the coefficient is biased, and more data just makes the biased number more precise. Causality needs a design or a convincing argument that $E[\varepsilon | X] = 0$. A self-check: an answer that names the same mechanism (in any words) has the key points.

A6 b. Yes: they attend whatever the lottery says.

Shares here: always-takers 20 % (losers who attend), never-takers 30 % (winners who don't), compliers 70 – 20 = 50 %.

Why the other options are wrong:

- **a** Defiers would attend only if they *lose*. We assume there are none (monotonicity), and these losers would surely also attend if they won.
 - **c** Compliers attend if they win and don't if they lose.
 - **d** Never-takers don't attend even if they win.
-

A7 b. Right: $0.108 \pm 1.96 \times 0.0059 \approx 0.108 \pm 0.012$.

R's `confint(m)` gives [0.0964, 0.1198]. Remember: this is the interval from one sample. The recipe covers the true value in about 95 % of samples, if the model is right.

Why the other options are wrong:

- **a** The estimate and the SE are enough: estimate $\pm 1.96 \times$ SE. The p-value is a different summary of the same information.
 - **c** That is estimate ± 1 SE, which is only about a 68 % interval. A 95 % interval uses about ± 1.96 SE.
 - **d** That uses ± 0.0594 , ten times the actual SE. Read the SE column carefully: 0.00594.
-

A8 –0.333.

Effects: $5 - 3 = 2$, $4 - 4 = 0$, $2 - 5 = -3$. ATE = $(2 + 0 - 3)/3 = -0.33$. In real data we would only ever see one of the two columns for each person.

Common wrong answers:

- **0.333** Check the sign: Cem loses 3, which outweighs Adam's gain of 2.
 - **-1** -1 is the sum of the effects divided by 1. Divide the sum (-1) by 3 people.
-

A9 b. Correct. Probability $\approx 1 - (1 - 1/n)^n \rightarrow 1 - 1/e \approx 63$ % (Lab 3: 65 %).

The bootstrap needs the statistic to depend smoothly on the distribution. The maximum sits on the edge of the data, so it breaks. Subsampling ($m < n$, without replacement) still works.

Why the other options are wrong:

- **a** This happens precisely because of how the maximum behaves under resampling, for uniform data too.
- **c** The spike stays at about 63 % for any B. It is a property of resampling the maximum.
- **d** The spike is an artefact of resampling, not a feature of the real sampling distribution of the maximum.

A10 d. Right. A path $Z \rightarrow (\text{city jobs}) \rightarrow Y$ breaks exclusion; Z linked to family background breaks independence.

Card's college-proximity papers control for region and urban residence for exactly this reason. Whether that is enough is an argument, not a test.

Why the other options are wrong:

- **a** Monotonicity is about everyone being pushed the same way by Z . The critic's point is a separate path to wages.
- **b** Relevance is about distance moving schooling. That is plausible, and testable with the first stage.
- **c** You can't control for the endogenous variable itself. You could control for region or urban status, if that closes the path.

A11 50.

Lower end $\approx 2.5\% \times 2,000 = 50\text{th value}$; upper end $\approx 97.5\% \times 2,000 = 1,950\text{th value}$.

Common wrong answers:

- **100** is 5% of 2,000: that would be a 90% interval.
- **25** would be right for $B = 1,000$. Here $B = 2,000$.
- **2.5** is the percentage. The question asks for the position in the sorted list.

A12 d. Correct: $\exp(\beta)$ is an odds ratio per unit of x .

$\log(p/(1-p)) = \beta_0 + \beta_1 \cdot \text{temp}$, so +1 degree adds β_1 to the log-odds, i.e. multiplies the odds by e^{β_1} . Lab 2 shows the predicted failure probability falling from 0.785 at 10 °C to 0.122 at 20 °C.

Why the other options are wrong:

- **a** Two problems: 0.72 is an odds ratio, and "cause" needs more than a fitted model.
- **b** Logit coefficients are on the log-odds scale, not the probability scale.
- **c** 0.72 multiplies the odds, not the probability. The change in probability depends on where you start.

A13 a. Yes (Gelman & Imbens 2019).

Local linear regression within a bandwidth, with robust bias-corrected inference (rdrobust), is the standard approach.

Why the other options are wrong:

- **b** High-order global polynomials give far-away points large, erratic weight at c .
- **c** Higher orders make the problem worse.
- **d** Averaging a good and a bad estimator doesn't fix the bad one.

A14 b. Yes: log-level, so multiply by 100 for a percentage, and "same experience" because exper is held fixed.

$\log y$ on level x : $\beta \times 100 \approx \%$ change in y per one-unit change in x , holding the other regressors fixed. For larger coefficients the exact change is $100 \cdot (e^{\beta} - 1)\%$.

Why the other options are wrong:

- **a** The arithmetic is right, but "causes" is too strong. Leaving out ability (or anything else linked to both) would bias the slope. A regression on its own shows association.
- **c** That is the level-level reading. Here y is in logs, so the coefficient is (about) a proportional change: $0.06 \times 100 = 6\%$.
- **d** That is the log-log (elasticity) reading. Education is measured in years, not in logs.

A15 0.4 minutes.

$\hat{\lambda} = \bar{X} = 4.0$ and $SE \approx 4.0/\sqrt{100} = 0.4$. This comes from the curvature: $\text{Var}(\hat{\lambda}) \approx 1/(n \cdot J)$ with $J = 1/\lambda^2$.

Common wrong answers:

- **4** 4.0 is $\hat{\lambda}$ itself (the MLE equals the sample mean). The SE divides it by $\sqrt{n}$.
 - **0.04** You divided by $n = 100$. The formula divides by $\sqrt{n} = 10$.
-

A16 a. Correct.

In R: `AER::ivreg(y ~ D | Z)`, or `fixest::feols(y ~ 1 | D ~ Z)`.

Why the other options are wrong:

- **b** The coefficient from the manual second stage is exactly the 2SLS estimate.
 - **c** `ivreg()` computes the coefficient this way but builds the residuals with the real D.
 - **d** Z must not be in the second stage; that would violate the exclusion logic.
-

A17 3.

B: $29 - 22 = 7$. N: $36 - 32 = 4$. $ATT(3, 4) = 7 - 4 = 3$. A clean 2x2: only not-yet- or never-treated units serve as controls.

Common wrong answers:

- **7** is B's own change from $t = 2$ to 4. Subtract N's change over the same period.
 - **4** is N's change. You need B's change minus N's change.
-

A18 d. Yes. Local effect, robust inference, and the size in percentage points.

Report the estimate, the robust CI, the bandwidth, and who it applies to. The simulated true jump was 4.5 pp, so this sample landed a bit high, which the wide CI reflects.

Why the other options are wrong:

- **a** Robust bias-corrected intervals are built around a bias-corrected estimate, so they needn't be centred on the conventional point estimate.
 - **b** RD gives the effect *at the cut-off*: mothers giving birth around 1 July.
 - **c** Only births within $h = 55$ days of the cut-off are used: about $5,870 + 6,039$.
-

A19 b. Right. Try the coverage check in the bootstrap demo with $n = 9$.

The bootstrap replaces the unknown population F with the sample's distribution $\hat{F}_n$. That works when $\hat{F}_n$ is close to F, i.e. when n is reasonably large.

Why the other options are wrong:

- **a** Without replacement (and same size) every resample is the original data.
 - **c** Its guarantees are large-sample results too. With 8 values the resamples can only recombine those 8 numbers.
 - **d** B only reduces simulation noise. It can't add information that isn't in the 8 observations.
-

A20 d. Then the log-likelihood is a constant minus $\sum (y_i - x_i\beta)^2/2\sigma^2$, so maximising it = least squares.

OLS = MLE under i.i.d. normal errors. The one difference: the MLE of σ^2 divides by n , OLS by $n - p$.

Why the other options are wrong:

- **a** Large n is about the sampling distribution, not about which estimator maximises the likelihood.
 - **b** With unequal variances the normal MLE weights observations (weighted least squares), so it differs from OLS.
 - **c** With, say, Laplace errors the MLE minimises absolute deviations, not squares.
-

A21 c. Yes. Sorting breaks the "as good as random near c" logic.

Validity checklist: density test (rddensity), covariate balance, placebo cut-offs, donut RD, bandwidth sensitivity.

Why the other options are wrong:

- **a** Moving the cut-off doesn't remove the sorting; it just makes a placebo test.
- **b** The problem is in the data, not the estimation window.
- **d** An easy test would shift the whole distribution, not create a pile-up exactly at the cut-off.

A22 b. Yes (Goodman-Bacon 2021). It can even flip the sign.

Use Callaway & Sant'Anna (did::att_gt), Sun & Abraham, or similar estimators when timing is staggered and effects may change over time.

Why the other options are wrong:

- **a** Same data for both. The difference is in which comparisons are used.
- **c** It uses never- and not-yet-treated units as controls; it only drops the already-treated.
- **d** Heteroskedasticity affects SEs, not the point estimate.

Paper B

B1 d. Exactly. $R^2 = 1 - \text{RSS}/\text{TSS}$, and TSS is the error you'd make by always predicting $\bar{y}$.

R^2 compares your model's squared errors with those of the "always guess the mean" model. Below zero means worse than that baseline.

Why the other options are wrong:

- **a** There is no such thing. A negative R^2 just means $\text{RSS} > \text{TSS}$.
- **b** The 0-to-1 range only holds for in-sample OLS with an intercept. Without an intercept, or out of sample, RSS can exceed TSS.
- **c** R^2 says nothing about the sign of the slope. A model with a negative slope can have $R^2 = 0.9$.

B2 d. Yes. The interaction is the extra change for the treated group after.

In $Y = \alpha + \beta G + \gamma \text{Post} + \delta(G \times \text{Post}) + \epsilon$, δ is the DiD estimate. The simulated true effect was -4 .

Why the other options are wrong:

- **a** That is city A's raw before-after change, which still contains the common time change.
- **b** That is the gap between the cities before the tax (baseline difference).
- **c** That is the common change over time, seen in the control city.

B3 a. Correct. It measures how surprising the data are under model + null.

p-values describe the data under an assumed model and null. They say nothing about the size of the effect or whether the model (e.g. no omitted variables) is right.

Why the other options are wrong:

- **b** A p-value is calculated *assuming* the null is true, so it can't be the probability that the null is true.
- **c** Same mistake turned around. The p-value doesn't give the probability of any hypothesis.
- **d** Significance is not size. With enough data a tiny effect is "significant". Always look at the estimate itself.

B4 b. Right. With many observations the probability of any exact dataset is tiny, even under the best p .

Density and likelihood are the same formula read two ways: as a function of the data (θ fixed) or of θ (data fixed). The likelihood ranks parameter values; it is not a probability distribution over them.

Why the other options are wrong:

- **a** No such rule. In the coin example the likelihood at $p = 0.5$ was about 10^{-46} , and even at the peak it is small.
- **c** Taking logs doesn't turn it into a probability of p either.
- **d** The likelihood of the exact data is tiny for every p when n is large. The MLE is still the best-fitting value.

B5 b. Yes. That is a mediator ($D \rightarrow \text{job} \rightarrow Y$). Controlling for it removes part of the effect you want.

Rule of thumb: control for things decided *before* treatment that affect both D and Y (confounders). Don't control for things caused by D (mediators) or caused by both D and Y (colliders).

Why the other options are wrong:

- **a** Also pre-treatment and a plausible confounder. Fine to control for.
- **c** Age is fixed before treatment and plausibly affects both D and Y : a confounder.
- **d** Pre-treatment earnings likely affect both joining and later earnings: a confounder. Controlling for it helps.

B6 d. Yes. $SE \approx 1/\sqrt{(\text{observed information})}$, and information is the curvature at the top.

Observed information = minus the second derivative of ℓ at $\hat{\theta}$. The likelihood explorer on the companion website shows it: more flips with the same share of heads make the hill sharper.

Why the other options are wrong:

- **a** Fast change away from the peak means the data clearly prefer $\hat{\theta}$ over nearby values: that is a small SE.
- **b** A flat top means many values of θ fit the data almost equally well, so the data pin θ down poorly.
- **c** The location gives the estimate; the shape around it gives the precision.

B7 b. Clear and complete: estimate, uncertainty, and how the interval was built.

The bootstrap gives a standard error and an interval for an estimator that has no simple formula (the median). The bias column is a diagnostic, not something to report as " $\pm$ ".

Why the other options are wrong:

- **a** -0.49 is the estimated bias, not the uncertainty. The SE is 6.56.
- **c** Bias estimates are noisy; correcting with them adds variance and is rarely done unless the bias is large relative to the SE.
- **d** This interval is from one sample. The 95 % describes how often the method covers the truth, not the chance for this one interval.

B8 d. Yes. More resamples make the Monte Carlo approximation smoother, nothing more.

Precision comes from the real sample size n . B just needs to be big enough (1,000–10,000) that the simulation noise is small.

Why the other options are wrong:

- **a** Resampling the same data again and again can't create new information.
- **b** Coverage depends on whether the bootstrap works for this statistic and n , not on B (once B is reasonably large).
- **c** The SE reflects sampling variation from the n real observations. B doesn't change n .

B9 0.3 SD.

$\tau_{FRD} = 0.15 / (0.7 - 0.2) = 0.3 \text{ SD}$, the effect of tutoring for compliers at the cut-off. Same Wald logic as lecture 5.

Common wrong answers:

- **0.15** is the jump in scores (a reduced form). In a fuzzy design you divide by the jump in take-up.
- **0.214** You divided by 0.7. The first stage is the *jump* in take-up: $0.7 - 0.2 = 0.5$.
- **0.075** You multiplied by the jump. Divide instead.

B10 a. Yes: the level where treatment is assigned.

Cluster at the level of treatment assignment. With very few treated clusters, even clustered SEs are unreliable.

Why the other options are wrong:

- **b** Treatment varies by state, and outcomes within a state are correlated over time.
- **c** Individuals in the same state share state-level shocks; clustering by person ignores that.
- **d** More individuals don't fix correlated errors within states (Bertrand, Duflo & Mullainathan 2004).

B11 0.4.

Score: $60/p - 90/(1 - p) = 0 \Rightarrow \hat{p} = 60/150 = 0.4$, the share of heads.

Common wrong answers:

- **0.6** is the share of tails ($90/150$). p is the chance of heads.
- **60** That is the number of heads. The MLE of p is a probability between 0 and 1.

B12 d. Right: the bias-variance trade-off.

$MSE = \text{bias}^2 + \text{variance}$. Small h : little bias, high variance. Large h : low variance, more bias. The bandwidth simulator on the companion website shows this.

Why the other options are wrong:

- **a** Only the bandwidth changed. The kernel matters much less than the bandwidth anyway.
- **b** CI width depends on the standard error, which depends on how much data you use, not on the size of the effect.
- **c** The opposite: narrower windows usually reduce bias, at the cost of variance.

B13 6.

$80 - 70 = 10 = ATT + \text{selection bias} = 4 + 6$. The treated would have scored 6 points higher even without treatment: $E[Y(0)|D=1] - E[Y(0)|D=0] = 6$.

Common wrong answers:

- **10** is the naive difference. Subtract the ATT to get the selection bias.
- **-6** Check the sign: $10 = 4 + \text{bias}$, so the bias is +6. The treated would have done better even without treatment.
- **4** is the ATT. The question asks for the rest of the naive difference.

B14 b. Yes: same size as the data, with replacement, so some values repeat and some are left out.

Bootstrap sample = n draws with replacement from the empirical distribution. Repeat B times and recompute the statistic each time.

Why the other options are wrong:

- **a** That is subsampling, a different method (used when the ordinary bootstrap fails).
- **c** 1,000 is a typical number of *resamples* (B). Each resample has the original size $n = 40$.
- **d** That just shuffles the original data, so every statistic comes out identical.

B15 d. Right. Continuity requires that nothing else jumps at c.

Always ask: could something other than the treatment change at the cut-off?

Why the other options are wrong:

- **a** True, but that limits generalisation; it doesn't make the estimate wrong. The bigger issue is the second policy.
- **b** Age is a common running variable (e.g. the minimum drinking age).
- **c** Hard to manipulate. The issue here is a second policy at the same cut-off.

B16 c. Yes. Try the IV simulator with low strength: the estimates scatter widely and drift towards OLS.

A weak first stage means dividing by a number close to zero: the estimate becomes erratic and conventional inference breaks down.

Why the other options are wrong:

- **a** The classic threshold is $F \geq 10$, and for reliable conventional t-tests F needs to be far higher (around 104).
- **b** By-hand second stages give wrong SEs even with strong instruments.
- **d** Many weak instruments make 2SLS bias worse, not better.

B17 d. Right. Snow's Lambeth and S&V differed by 62 vs 565 deaths.

Very different levels can still make parallel trends less plausible in practice (e.g. mean reversion), so it is worth discussing. But equal levels are not required.

Why the other options are wrong:

- **a** That changes the scale (it would be parallel trends in logs), but level gaps are not the problem.
- **b** DiD differences out level gaps. Only the changes need to match without treatment.
- **c** There is no such rule.

B18 d. Yes, assuming the lottery is valid. The data were simulated with a true effect of 0.5.

Report the first stage, the IV estimate with IV standard errors, and who the estimate is about (compliers). Validity is an argument, not a test result.

Why the other options are wrong:

- **a** They are supposed to differ when D is endogenous: OLS picks up motivation, IV doesn't.
- **b** F only checks relevance. Independence and exclusion must be argued; here they rest on the lottery being random and affecting scores only through attendance.
- **c** IV gives a LATE: the effect for compliers. Always- and never-takers contribute nothing.

B19 a. Yes. The estimate is still unbiased, and adjusting for age can improve precision.

Over all possible random assignments the difference in means is unbiased for the ATE. In one draw, chance differences appear; a balance table shows whether they are larger than chance would explain.

Why the other options are wrong:

- **b** Randomisation balances groups *on average over many possible assignments*, not exactly in one sample.
- **c** They can. The point is that randomisation makes such differences chance, not systematic.
- **d** Dropping people after seeing outcomes or characteristics breaks the randomisation.

B20 0.6 SD.

First stage = $0.70 - 0.20 = 0.50$. LATE = $0.30 / 0.50 = 0.6 \text{ SD}$, the effect for compliers (children who attend only if they win).

Common wrong answers:

- **0.3** 0.30 is the reduced form (the ITT). Divide by the first stage, $0.70 - 0.20$.
- **0.43** You divided by 0.70. The first stage is the *difference* in attendance: $0.70 - 0.20 = 0.50$.
- **0.15** You multiplied. LATE divides the reduced form by the first stage.

B21 10 productivity units.

$(75 - 50) - (55 - 40) = 25 - 15 = 10$. Reading it as the causal effect of training needs parallel trends: without training, both groups would have improved by 15.

Common wrong answers:

- **25** is only the trained group's before–after change. DiD also subtracts the untrained group's change.
- **15** is the untrained group's change. You need the difference between the two changes.
- **20** Check both before–after differences separately ($75 - 50$ and $55 - 40$) before subtracting them.
- **35** = $75 - 40$ compares the trained after with the untrained before. Compare changes, not mixed levels.

B22 0.095.

$\beta_1(\text{short}) = 0.08 + 0.10 \times 0.15 = 0.08 + 0.015 = 0.095$. The slope is too big because part of ability's effect is credited to education.

Common wrong answers:

- **0.08** is the true effect. The short regression adds the bias term $\beta_3 \cdot \gamma_1$ on top.
- **0.015** is only the bias (0.10×0.15). Add it to the true effect 0.08.
- **0.065** Check the sign: ability raises wages (+0.10) and moves with education (+0.15), so the bias is positive.