HE2003 Econometrics I – Revision Notes

Quantitative Research Society @NTU

Academic Year 2026–2027

Abstract

Comprehensive revision notes for HE2003 Econometrics I, rebuilt lecture by lecture from the AY2026/27 slides. Each chapter develops the theory with full derivations, then consolidates it with key formulas, worked examples drawn from the lectures and tutorials, common mistakes, and exam-focused checklists. Part I covers causality and the simple linear regression model: the OLS estimator, the Least Squares Assumptions, measures of fit, unbiasedness and efficiency, binary regressors, and inference with a single regressor. Part II (multiple regression, joint hypotheses, nonlinear models, and specification issues) is added as the lectures are delivered.

For academic reference only. This document is provided for educational purposes and does not constitute an official question paper, answer key, or any formally authorised academic material. Its content is not guaranteed to be complete or accurate.

Chapter	Week	Lecture	Tutorial	Key Concepts
Ch. 1: Introduction and Causality	1	L1	T1	Correlation vs. causation; conditional mean; probability toolkit; Type I/II errors; RCTs
Ch. 2: The OLS Estimator	2	L2	T2	SLR model; least-squares problem; FOCs; closed-form $\hat{\beta}_0, \hat{\beta}_1$; fitted values and residuals
Ch. 3: The Least Squares Assumptions	3	L3	T3	Assumptions 1–3; zero conditional mean; RCT and ATE; residual identities; selection bias
Ch. 4: Sampling, Fit and Distribution	4	L4	T4	Random sampling; R^2, TSS = ESS + SSR; CAPM beta; sampling distribution of $\bar{Y}$
Ch. 5: Unbiasedness, Efficiency and Binary Regressors	5	L5	T5	Unbiasedness; efficiency; BLUE; Gauss–Markov; dummy variables; testing set-up
Ch. 6: Inference with a Single Regressor	6	L6	T6	SE($\hat{\beta}_1$); t-statistic; confidence intervals; p-values; applications
Ch. 7: Part I Summary	–	L1–L6	T1–T6	Synthesis for the Week 7 Quiz: comparison tables, method patterns, traps, integrated examples.
Ch. 8–13: Part II	8–12	L7–L12	T7–T11	Multiple regression, OVB, heteroskedasticity, F-tests, nonlinear and interaction models, specification (added on release).
Ch. 14: Part II Summary	–	L7–L12	T7–T11	Synthesis for the Final Examination.

"This revision guide follows the official HE2003 lecture sequence, one chapter per lecture, with part-level synthesis chapters before each assessment."

Contents

How to Use This Document 5

I Main Notes 6

1 Lecture 1: Introduction 6
- 1.1 Why Study Econometrics? . . . 6
- 1.2 Correlation versus Causation . . . 6
- 1.3 Conditional Mean and the Population–Sample Distinction . . . 8
- 1.4 Probability Toolkit . . . 10
- 1.5 Hypothesis Testing Vocabulary . . . 14
- 1.6 Randomized Controlled Trials and the Case for Econometrics . . . 16
- 1.7 Preview: The Simple Regression Model . . . 16
- 1.8 Chapter Summary . . . 19

2 Lecture 2: The OLS Estimator 20
- 2.1 The Simple Linear Regression Model . . . 20
- 2.2 Three Kinds of "Beta": Parameters, Candidates, Estimators . . . 20
- 2.3 Summation Notation and Two Algebraic Identities . . . 21
- 2.4 The OLS Minimization Problem . . . 23
- 2.5 Deriving the Closed-Form OLS Estimator . . . 23
- 2.6 Fitted Values, Residuals, and the Mean Point . . . 26
- 2.7 A Load-Bearing Figure: Fitting a Line Through Data . . . 27
- 2.8 Worked Examples: California Test Scores . . . 28
- 2.9 Worked Example: A Hand-Computed OLS Regression . . . 30
- 2.10 Worked Examples: Class Size and Test Scores (Tutorial 2) . . . 31
- 2.11 Interpretation in Words, and Further Traps . . . 31
- 2.12 Chapter Summary . . . 32

3 Lecture 3: The Least Squares Assumptions 34
- 3.1 From an Estimated Line Back to the Population: CEO Salary and ROE . 34
- 3.2 True Values, OLS Estimates, and the Five Objects of Simple Regression . 35
- 3.3 The Residual Identities: Consequences of the FOCs . . . 36
- 3.4 Key Concept: The Least Squares Assumptions . . . 38
- 3.5 Assumption 1 in Depth: Zero Conditional Mean . . . 39
- 3.6 The RCT Benchmark and the Average Treatment Effect . . . 42
- 3.7 Worked Example (Tutorial 3): Wages of SSS versus SOH Graduates . . . 42
- 3.8 A Counterexample from Question 3(c): $E[u] = 0$ Does Not Imply $E[u \mid X] = 0$ . . . 46
- 3.9 Return to Education: A Case Where Assumption 1 May Fail . . . 47
- 3.10 Policy Application: Survivorship Bias and Wald's Bombers . . . 48
- 3.11 Assumption 2: I.I.D. Sampling . . . 48
- 3.12 Assumption 3: Large Outliers Are Unlikely . . . 48
- 3.13 Why All Three Assumptions Matter . . . 50
- 3.14 Chapter Summary . . . 51

4 Lecture 4: Sampling, Measures of Fit and the Sampling Distribution 52
4.1 Selection Bias Revisited: When the Sample Misleads . . . 52
4.2 Least Squares Assumption 2 in Depth: The Mechanism of Random Sampling 56
4.3 Why the Three Least Squares Assumptions Matter . . . 58
4.4 Measures of Fit: The Coefficient of Determination R^2 . . . 59
4.5 Worked Example: Predicting Weight from Height . . . 66
4.6 The CAPM and the "Beta" of a Stock . . . 67
4.7 The Sampling Distribution of the OLS Estimators . . . 69
4.8 Chapter Summary . . . 71

5 Lecture 5: Unbiasedness, Efficiency and Binary Regressors 73
5.1 Estimators, Unbiasedness, and Efficiency . . . 73
5.2 Linear Estimators of a Population Mean . . . 74
5.3 Application: Comparing Two Independent Sample Means . . . 79
5.4 The Gauss–Markov Theorem . . . 81
5.5 Regression with a Binary Regressor . . . 84
5.6 Common Slope, Different Intercepts . . . 87
5.7 Multiple Dummy Variables: The Law School Ranking Model . . . 88
5.8 Setting Up Hypothesis Tests: A First Look . . . 88
5.9 Chapter Summary . . . 89

6 Lecture 6: Inference with a Single Regressor 91
6.1 From the Sampling Distribution of $\hat{\beta}_1$ to Its Standard Error . . . 91
6.2 The t-Statistic and Hypothesis Testing . . . 94
6.3 Confidence Intervals for β_1 . . . 101
6.4 The p-Value . . . 104
6.5 Chapter Summary . . . 107

7 Part I Summary: Simple Linear Regression (Lectures 1–6) 108
7.1 Comparison of Central Results . . . 110
7.2 Standard Problem Types and Solution Patterns . . . 112
7.3 Common Mistakes and Exam Traps . . . 116
7.4 Integrated Worked Review Examples . . . 117
7.5 Part I Revision Checklist . . . 129

A Formula & Key Results Reference 131

B Notation 135

C Exam & Admin Information 137

How to Use This Document

Structure of the Notes

Lectures provide the core theory, definitions, assumptions, and examinable derivations. Tutorials are folded in as labelled worked examples ("Worked Example (Tutorial n)") where they sharpen technique; full tutorial solutions live in the separate Tutorial Solutions series. Each chapter is organised into:

1. **Core Theory** — definitions, assumptions, theorems, and the intuition behind them.
2. **Derivations** — every estimator and result derived step by step, not merely stated.
3. **Worked Examples** — lecture and tutorial applications with full numerical working.
4. **Exam Focus** — high-yield results, interpretation templates, and typical traps.

Suggested Revision Flow

- First pass: read Core Theory + Key Formula boxes.
- Second pass: rederive every Key Formula without notes; attempt Worked Examples *before* reading solutions.
- Final pass: Summary chapters, then Common Mistake and Exam Focus boxes in the week before each test.

Colour and Box System (Legend)

Colour	Box Name	Use This For
Green	**Definition**	Formal definitions and assumptions (numbered).
Blue (band)	**Theorem**	Formal results; a proof/derivation follows immediately.
Blue	**Key Formula**	Equations to memorise or rederive quickly.
Brown	**Worked Example**	Fully worked questions. Try on your own first.
Red	**Common Mistake**	Typical errors and conceptual traps.
Yellow	**Policy Application**	Economic interpretation and empirical applications.
Purple	**Exam Focus**	High-yield checklists and exam strategy.
Cyan	**Extension**	Beyond-syllabus depth (non-examinable unless stated).

Part I
Main Notes

1 Lecture 1: Introduction

This chapter answers three questions: (1) *why do we need econometrics at all, given that economic theory already tells us the sign of most relationships?*; (2) *when two variables move together, how do we tell whether one causes the other?*; (3) *what vocabulary and probability tools does the rest of this course assume we already have?* It has no predecessor chapter; it prepares Chapter 2 (*the OLS estimator*) and Chapter 3 (*the Least Squares Assumptions*), both of which rely on the conditional-mean and probability-toolkit machinery built here.

1.1 Why Study Econometrics?

Economic theory is usually good at predicting the *sign* of a relationship but silent on its *magnitude*. Macroeconomic theory says a central-bank rate hike slows growth, but a policymaker needs to know by how much. Microeconomic theory says a smaller class size should help students, but a school board deciding whether to hire another teacher needs a number, not a direction. Econometrics is the discipline that turns economic theory plus real-world data into quantitative, testable answers to questions such as:

- What is the effect on economic growth of a 1 percentage point increase in the policy interest rate?
- What is the effect of reducing class size by 5 (or 10) students on student test scores?
- What is the "return to education": how much does one extra year of schooling raise a worker's wage?
- Do cigarette taxes actually reduce smoking, and by how much?
- Do high-speed railways (HSR) promote local economic growth?

Several of these questions are not just about association but about *causal effect*: if a person is given one more year of education, how much will their wage rise, holding everything else about that person fixed? Observing that education and wages are positively correlated in the data is not, by itself, evidence of a causal link — people with more education may also have higher innate ability, and ability itself raises wages. Separating causation from mere correlation is the organising theme of this chapter, and indeed of most of this course.

1.2 Correlation versus Causation

Definition 1.1: Three sources of correlation between X and Y

Suppose two variables X and Y are observed to be correlated (positively or negatively). This can arise from (at least) three distinct underlying structures, which are observationally similar in a simple correlation coefficient but economically very different:

1. **Causality:** $X \to Y$, i.e. X causally affects Y.
2. **Reverse causality:** $Y \to X$, i.e. Y causally affects X, and the correlation is mislabelled if we regress Y on X.
3. **Common factor:** there is a third variable C with $C \to X$ and $C \to Y$, so X and Y move together only because both respond to C, with no direct causal link between them.

Intuition. A raw correlation coefficient $\text{Corr}(X, Y)$ cannot by itself distinguish these three cases — all three produce a nonzero correlation. Telling them apart requires either institutional knowledge of the mechanism (which variable could plausibly move first) or a research design, such as a randomized experiment (Section 1.6), that rules out reverse causality and common factors by construction.

Common Mistake: beer and air-conditioner sales

Beer sales and air-conditioner sales are positively correlated. It would be wrong to conclude that buying an air conditioner causes people to drink more beer, or that drinking beer causes people to buy air conditioners. Both are driven by a common factor: hot temperature. This is Case 3 (common factor) above, and running a regression of beer sales on air-conditioner sales (or vice versa) would produce a badly misleading "causal" estimate.

Worked Example (Tutorial 1): classifying correlation into the three cases

Method 1: Mechanism (timing) argument for each pair.
(a) Windmills and wind.
- Wind is air in motion; it pushes on the blades and turns them, so a faster wind (in m/s) makes a windmill spin faster (in revolutions per minute).
- The reverse channel is absent: a spinning windmill does not generate the regional wind that drives it, so this is Case 1 with $X =$ wind and $Y =$ windmill speed.

Hence the dependent variable is windmill speed, and the population model and slope are

$$\text{windmill}_i = \beta_0 + \beta_1 \,\text{wind}_i + u_i, \qquad \beta_1 = \frac{\Delta\,\text{windmill}}{\Delta\,\text{wind}}.$$

For instance, if $\beta_1 = 3$ rpm per m/s, a wind that is 2 m/s faster raises windmill speed by $3 \times 2 = 6$ rpm.
(b) Ice cream sales and drowning deaths.
- Hot weather raises ice cream sales (units per day), and hot weather also sends more people swimming, which raises drowning deaths (deaths per day).
- Eating ice cream does not make a swimmer drown, and a drowning does not make others buy ice cream, so there is no direct arrow between them; this is Case 3 with $C =$ temperature.

Hence each variable is regressed on the common factor:

$$\text{icecream}_i = \beta_0 + \beta_1 \,\text{temp}_i + u_i, \qquad \text{drowning}_i = \gamma_0 + \gamma_1 \,\text{temp}_i + v_i.$$

(c) Masks and online-working hours.
- The COVID-19 outbreak raised the demand for masks (boxes bought) and, through lockdowns and office closures, raised online-working hours (hours per week).
- Buying masks does not move work online and working online does not create a need for masks (a person at home needs fewer masks), so this is Case 3 with $C =$ COVID-19.

$$\boxed{\text{(a) causality } X \to Y; \quad \text{(b), (c) common factor } C \to X,\ C \to Y}$$

Method 2: Hold-the-common-factor-fixed (intervention) test. Ask two questions for each pair: if X were changed by an outside intervention, would Y move; and if a candidate common factor C were held fixed, would the X–Y correlation survive?
- (a) Shielding a windmill from the wind (intervening on wind) stops it, while driving its blades with a motor (intervening on the windmill) creates no wind, so only $X \to Y$ survives.
- (b) Fix the temperature at, say, 30°C. Suppose $\beta_1 = 5$ units and $\gamma_1 = 0.2$ deaths per °C; a 4°C warmer day then raises ice cream sales by $5 \times 4 = 20$ units and drownings by $0.2 \times 4 = 0.8$ deaths together, while on days of equal temperature the two series carry no link, so the correlation is produced entirely by C.
- (c) Within one week of fixed COVID-19 conditions, banning mask sales would not change online-working hours, so the correlation is produced by C.

$$\boxed{\text{(a) causality } X \to Y; \quad \text{(b), (c) common factor } C \to X,\ C \to Y}$$

Policy Application: high-speed rail and local growth

China's HSR network offers an example of how hard case-classification can be in practice. Of 149 planned HSR lines, 83 were open by 2016 and 66 were not. It is tempting to treat the 83 opened lines as a "treatment group" and the 66 unbuilt lines as a "control group" and simply compare local economic growth between them. But route selection is rarely random: planners are more likely to build HSR lines connecting cities that are *already* growing quickly or are strategically important. If so, the comparison confounds the causal effect of HSR with the pre-existing growth trajectories of selected cities — a preview of the *selection bias* theme developed formally in Chapter 3.

1.3 Conditional Mean and the Population–Sample Distinction

Definition 1.2: Conditional mean

For random variables Y and X, the *conditional mean* of Y given $X = x$, written

$$E(Y \mid X = x),$$

is the average value of Y among the sub-population for which X takes the value x. It generally varies with x; comparing conditional means at different values of X is one of the most basic tools for describing how Y relates to X.

QRS@NTU — qrs.ntu.edu.sg

Worked Example (Lecture 1): the gender wage gap as a difference in conditional means

Let Y = wage and X = gender. The gender wage gap is defined as the difference in conditional means,

$$E(\text{wage} \mid \text{gender} = \text{male}) - E(\text{wage} \mid \text{gender} = \text{female}).$$

- The gap is a population object: it compares the average monthly wage (in S$) of *all* male workers with that of *all* female workers; a sample of workers only estimates it by the difference of the two group sample means.
- The gap is descriptive, not causal: it may reflect a common factor (occupation, hours worked) as well as discrimination.
- For example, if male workers earn S$5,000 per month on average and female workers S$4,400, the gap is $5000 - 4400 = 600$ S$ per month.

Write $\mu_1 = E(Y \mid D = 1)$, $\mu_0 = E(Y \mid D = 0)$ for $Y_i = \beta_0 + \beta_1 D_i + u_i$ with $D_i = 1$ for male and 0 for female, and impose Least Squares Assumption 1, $E(u_i \mid D_i) = 0$.

Method 1: Conditional means at $D = 1$ and $D = 0$. Take the conditional mean of the model at each value of D:

$$E(Y \mid D = 1) = E(\beta_0 + \beta_1 \cdot 1 + u \mid D = 1) = \beta_0 + \beta_1 + E(u \mid D = 1) = \beta_0 + \beta_1 + 0,$$
$$E(Y \mid D = 0) = E(\beta_0 + \beta_1 \cdot 0 + u \mid D = 0) = \beta_0 + E(u \mid D = 0) = \beta_0 + 0.$$

Subtract the second line from the first:

$$E(Y \mid D = 1) - E(Y \mid D = 0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

In the example, $\beta_0 = \mu_0 = 4400$ and $\beta_1 = \mu_1 - \mu_0 = 600$ S$ per month.

$$\boxed{\beta_1 = E(Y \mid D = 1) - E(Y \mid D = 0)}$$

Method 2: Population slope $\text{Cov}(D, Y)/\text{Var}(D)$. By the LIE, $E(u) = E[E(u \mid D)] = 0$ and $E(Du) = E[D\,E(u \mid D)] = 0$, so $\text{Cov}(D, u) = 0 - E(D) \cdot 0 = 0$. By bilinearity,

$$\text{Cov}(D, Y) = \text{Cov}(D, \beta_0 + \beta_1 D + u) = 0 + \beta_1 \text{Cov}(D, D) + \text{Cov}(D, u) = \beta_1 \text{Var}(D).$$

Let $p = P(D = 1)$ with $0 < p < 1$. Then $E(D) = p$, $E(D^2) = p$, $\text{Var}(D) = p - p^2 = p(1 - p)$, $E(Y) = p\mu_1 + (1 - p)\mu_0$ and $E(DY) = p\mu_1$ (LIE), so

$$\text{Cov}(D, Y) = p\mu_1 - p\big(p\mu_1 + (1 - p)\mu_0\big) = p(1 - p)\mu_1 - p(1 - p)\mu_0 = p(1 - p)(\mu_1 - \mu_0),$$
$$\beta_1 = \frac{\text{Cov}(D, Y)}{\text{Var}(D)} = \frac{p(1 - p)(\mu_1 - \mu_0)}{p(1 - p)} = \mu_1 - \mu_0.$$

With $p = 0.5$: $\text{Cov}(D, Y) = 0.25 \times 600 = 150$, $\text{Var}(D) = 0.25$, $\beta_1 = 150/0.25 = 600$.

$$\boxed{\beta_1 = E(Y \mid D = 1) - E(Y \mid D = 0)}$$

Definition 1.3: Population versus sample

The *population* is the entire collection of units the researcher wants to learn about (e.g. the wages of *all* workers in Singapore). The *sample* is the (typically much smaller) group of units actually observed, usually obtained by drawing a subset of the population at random (e.g. 1,000 or 5,000 randomly selected workers).

Treating an observation's value (e.g. a worker's wage) as a random variable does not mean the value itself fluctuates randomly once the worker is sampled; it means the *selection procedure* is random — a different random draw would have produced a different sample, hence different sample statistics. This

is the conceptual foundation for everything in Chapter 4 onward: sample averages, OLS estimators, and their standard errors are all random variables precisely because the sample itself is the outcome of a random draw from the population.

1.4 Probability Toolkit

This section collects, with proofs, the expectation/variance/covariance results that Chapters 2–6 use repeatedly (unbiasedness proofs, variance-of-the-OLS-estimator derivations, the Gauss–Markov theorem). These facts are prerequisite material (first-year probability), so the proofs are short; what matters is having the exact statements available for later citation.

Definition 1.4: Covariance

For random variables X, Y with means $E[X], E[Y]$,

$$\operatorname{Cov}(X,Y) := E\big[(X - E[X])(Y - E[Y])\big] = E[XY] - E[X]E[Y].$$

Theorem 1.1: Linearity of expectation

For random variables X, Y and constants a, b,

$$E[aX + b] = aE[X] + b, \qquad E[X + Y] = E[X] + E[Y].$$

Proof. By definition of expectation as a weighted (probability-weighted) sum/integral, the expectation operator is linear: $E[aX+b] = \sum_x (ax+b)P(X=x) = a\sum_x xP(X=x) + b\sum_x P(X=x) = aE[X]+b$, using $\sum_x P(X=x) = 1$. For the sum of two random variables, $E[X+Y] = \sum_{x,y}(x+y)P(X=x, Y=y) = \sum_{x,y} xP(X=x,Y=y) + \sum_{x,y} yP(X=x,Y=y) = E[X] + E[Y]$, splitting the double sum and summing out the variable not appearing in each term. □

Theorem 1.2: Variance of a linear transformation

For a random variable X and constants a, b,

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X).$$

Proof. By Theorem 1.1, $E[aX + b] = aE[X] + b$. Hence

$$\operatorname{Var}(aX+b) = E\big[(aX + b - aE[X] - b)^2\big] = E\big[a^2(X - E[X])^2\big] = a^2 E\big[(X - E[X])^2\big] = a^2 \operatorname{Var}(X).$$

□

Theorem 1.3: Bilinearity of covariance

For random variables X, Y, Z and constants a, b,

$$\operatorname{Cov}(aX + b,\, Y) = a\, \operatorname{Cov}(X,Y), \qquad \operatorname{Cov}(X,\, Y + Z) = \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z).$$

Proof. Using Definition 1.4 and Theorem 1.1,

$$\begin{aligned}
\operatorname{Cov}(aX+b, Y) &= E[(aX+b)Y] - E[aX+b]E[Y] \\
&= aE[XY] + bE[Y] - \big(aE[X] + b\big)E[Y] \\
&= a\big(E[XY] - E[X]E[Y]\big) = a\operatorname{Cov}(X,Y).
\end{aligned}$$

For the second identity,

$$\begin{aligned}\mathrm{Cov}(X, Y+Z) &= E[X(Y+Z)] - E[X]E[Y+Z] \\ &= E[XY] + E[XZ] - E[X]E[Y] - E[X]E[Z] \\ &= \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z).\end{aligned}$$

□

Theorem 1.4: Variance of a sum

For random variables X, Y,

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$

Proof. Let $\mu_X = E[X]$, $\mu_Y = E[Y]$. By linearity, $E[X+Y] = \mu_X + \mu_Y$, so

$$\begin{aligned}\mathrm{Var}(X+Y) &= E\big[\big((X-\mu_X) + (Y-\mu_Y)\big)^2\big] \\ &= E\big[(X-\mu_X)^2\big] + E\big[(Y-\mu_Y)^2\big] + 2\,E\big[(X-\mu_X)(Y-\mu_Y)\big] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y),\end{aligned}$$

expanding the square and applying linearity of expectation term by term. □

Theorem 1.5: Independence implies zero covariance

If X and Y are independently distributed, then

$$E[XY] = E[X]E[Y] \qquad \text{and hence} \qquad \mathrm{Cov}(X,Y) = 0.$$

Proof. Independence means the joint probability (density or mass) factors as $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all x, y. Then

$$E[XY] = \sum_x \sum_y xy\, f_{X,Y}(x,y) = \sum_x \sum_y xy\, f_X(x) f_Y(y) = \Big(\sum_x x f_X(x)\Big)\Big(\sum_y y f_Y(y)\Big) = E[X]E[Y].$$

By Definition 1.4, $\mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0$. □

Common Mistake: zero covariance does not imply independence

Theorem 1.5 only goes one way. $\mathrm{Cov}(X,Y) = 0$ (equivalently $\mathrm{Corr}(X,Y) = 0$) does *not* imply X and Y are independent — there could be a nonlinear relationship (e.g. $Y = X^2$ with X symmetric around 0) that produces zero *linear* covariance while X and Y are obviously dependent. Keep this asymmetry in mind: independence is the stronger statement.

Theorem 1.6: Properties of conditional expectation

Let X, Z be random variables and $g(\cdot)$ any function.

1. $E[g(X) \mid X] = g(X)$: conditioning on X turns any function of X alone into a known constant.
2. If Z and X are independently distributed, then $E[Z \mid X] = E[Z]$: conditioning on a variable independent of Z gives no information about Z, so the conditional mean collapses to the unconditional mean.

Proof. (1) Once we condition on $X = x$, the quantity $g(X)$ is no longer random — it takes the fixed value $g(x)$ with probability 1, so its conditional expectation equals its value, $g(x)$; writing this for general X gives $E[g(X) \mid X] = g(X)$.

(2) If X, Z are independent, the conditional distribution of Z given $X = x$ equals the (unconditional) marginal distribution of Z for every x, since independence means $f_{Z \mid X}(z \mid x) = f_Z(z)$. Averaging z against this conditional distribution therefore gives the same number regardless of x, namely $E[Z]$. □

Theorem 1.7: Law of Iterated Expectations (LIE)

For random variables X, Y,

$$E\big[E(Y \mid X)\big] = E[Y].$$

Proof (discrete case). Let X take values $x_1, \dots, x_k$ with probabilities $P(X = x_j)$, and let Y take values y with probability mass function $P(Y = y)$. By definition of conditional expectation,

$$E(Y \mid X = x_j) = \sum_y y\, P(Y = y \mid X = x_j).$$

Then

$$\begin{aligned}
E\big[E(Y \mid X)\big] &= \sum_{j=1}^{k} P(X = x_j)\, E(Y \mid X = x_j) \\
&= \sum_{j=1}^{k} P(X = x_j) \sum_y y\, P(Y = y \mid X = x_j) \\
&= \sum_{j=1}^{k} \sum_y y\, P(X = x_j) P(Y = y \mid X = x_j) \\
&= \sum_{j=1}^{k} \sum_y y\, P(X = x_j,\, Y = y) \qquad \text{(definition of conditional probability)} \\
&= \sum_y y \sum_{j=1}^{k} P(X = x_j,\, Y = y) \\
&= \sum_y y\, P(Y = y) \qquad \text{(summing the joint mass over all } x_j \text{ recovers the marginal of } Y) \\
&= E[Y].
\end{aligned}$$

□

Theorem 1.8: Moments of the standard normal distribution

If $X \sim N(0, 1)$, then

$$E[X] = 0, \qquad E[X^2] = 1, \qquad E[X^3] = 0.$$

Proof. The standard normal density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is symmetric about 0: $\phi(-x) = \phi(x)$. For any odd function h (i.e. $h(-x) = -h(x)$), the integrand $h(x)\phi(x)$ is odd, so $\int_{-\infty}^{\infty} h(x)\phi(x)\, dx = 0$ by symmetry (the negative and positive halves of the real line cancel exactly). Taking $h(x) = x$ gives $E[X] = 0$; taking $h(x) = x^3$ gives $E[X^3] = 0$. For the second moment, $\mathrm{Var}(X) = E[X^2] - (E[X])^2$ by definition of variance, and $\mathrm{Var}(X) = 1$ by the definition of the standard normal, so $E[X^2] = 1 + 0^2 = 1$. □

Key Formula: Probability Toolkit Recap

$$E[aX+b] = aE[X]+b, \qquad \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X),$$
$$\mathrm{Cov}(aX+b, Y) = a\,\mathrm{Cov}(X,Y), \qquad \mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y),$$
$$E\big[E(Y\,|\,X)\big] = E[Y] \text{ (LIE)}, \qquad X \perp Y \;\Rightarrow\; \mathrm{Cov}(X,Y) = 0.$$

Independence also gives $E[XY] = E[X]E[Y]$. For $X \sim N(0,1)$: $E[X] = 0$, $E[X^2] = 1$, $E[X^3] = 0$.

Worked Example (Tutorial 1): X, Z i.i.d. $N(0,1)$, $Y = X^2 + Z$

Let X and Z be independently distributed standard normal random variables, and let $Y = X^2 + Z$. Show $E[Y\,|\,X] = X^2$, $E[Y] = 1$, and $E[XY] = 0$.

Method 1: Linearity, independence and normal moments.

(a) Substitute $Y = X^2 + Z$ and split the conditional mean. Since X^2 is a function of X alone, Theorem 1.6(1) gives $E[X^2\,|\,X] = X^2$; since Z is independent of X, part (2) gives $E[Z\,|\,X] = E[Z]$:

$$\begin{aligned} E[Y\,|\,X] &= E[X^2+Z\,|\,X] = E[X^2\,|\,X] + E[Z\,|\,X] && \text{(linearity)} \\ &= X^2 + E[Z] = X^2 + 0 = X^2. && (Z \sim N(0,1)) \end{aligned}$$

(b) By linearity and Theorem 1.8 ($E[X^2] = 1$, $E[Z] = 0$):

$$E[Y] = E[X^2+Z] = E[X^2] + E[Z] = 1 + 0 = 1.$$

(c) Multiply out $XY = X(X^2+Z) = X^3 + XZ$. Independence gives $E[XZ] = E[X]E[Z]$ (Theorem 1.5), and Theorem 1.8 gives $E[X^3] = 0$, $E[X] = E[Z] = 0$:

$$E[XY] = E[X^3] + E[XZ] = E[X^3] + E[X]E[Z] = 0 + 0\cdot 0 = 0.$$

$$\boxed{E[Y\,|\,X] = X^2, \qquad E[Y] = 1, \qquad E[XY] = 0.}$$

Method 2: Conditional distribution and the LIE. (a) Fix $X = x$. Because Z is independent of X, $Z\,|\,X = x$ is still $N(0,1)$, so $Y\,|\,X = x$ equals the constant x^2 plus an $N(0,1)$ draw, i.e. $Y\,|\,X = x \sim N(x^2, 1)$. Its mean is x^2, so $E[Y\,|\,X] = X^2$.

(b) Apply the LIE to the result of (a):

$$E[Y] = E\big[E(Y\,|\,X)\big] = E[X^2] = 1.$$

(c) Given X, the factor X is known (Theorem 1.6(1)), so it comes out of the conditional mean; then apply the LIE:

$$\begin{aligned} E[XY\,|\,X] &= X\,E[Y\,|\,X] = X\cdot X^2 = X^3, && \text{(by (a))} \\ E[XY] &= E\big[E(XY\,|\,X)\big] = E[X^3] = 0. && \text{(LIE)} \end{aligned}$$

$$\boxed{E[Y\,|\,X] = X^2, \qquad E[Y] = 1, \qquad E[XY] = 0.}$$

Method 3: Covariance bilinearity (part (c)). By Definition 1.4, $E[XY] = \mathrm{Cov}(X,Y) + E[X]E[Y] = \mathrm{Cov}(X,Y) + 0\cdot 1 = \mathrm{Cov}(X,Y)$. By Theorem 1.3, with $\mathrm{Cov}(X,Z) = 0$ from independence (Theorem 1.5),

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,X^2) + \mathrm{Cov}(X,Z) = \big(E[X^3] - E[X]E[X^2]\big) + 0 = (0 - 0\cdot 1) + 0 = 0.$$

$$\boxed{E[XY] = 0.}$$

Exam Focus: the LIE "two-route" trick

- Whenever a question asks for $E[Y]$ or $E[g(X)Y]$ and Y is defined through a conditional structure (as in the worked example above), computing $E[Y \mid X]$ *first* and then applying the LIE is usually faster than expanding everything unconditionally — and doing both routes is a fast way to check your algebra.
- This exact question (Question 1 Q3 / Tutorial 1 Q3) is a template: expect close variants with different functional forms of Y in X and Z.

1.5 Hypothesis Testing Vocabulary

Chapters 5–6 build formal t-tests for regression coefficients; this section introduces the vocabulary in the abstract so it is available from the start.

Definition 1.5: Null and alternative hypotheses

The *null hypothesis* H_0 is the default claim being tested (e.g. "this policy has no effect"); the *alternative hypothesis* H_1 is the claim we entertain if the evidence contradicts H_0. A hypothesis test uses sample data to decide whether to reject H_0 in favour of H_1.

Definition 1.6: Type I and Type II errors

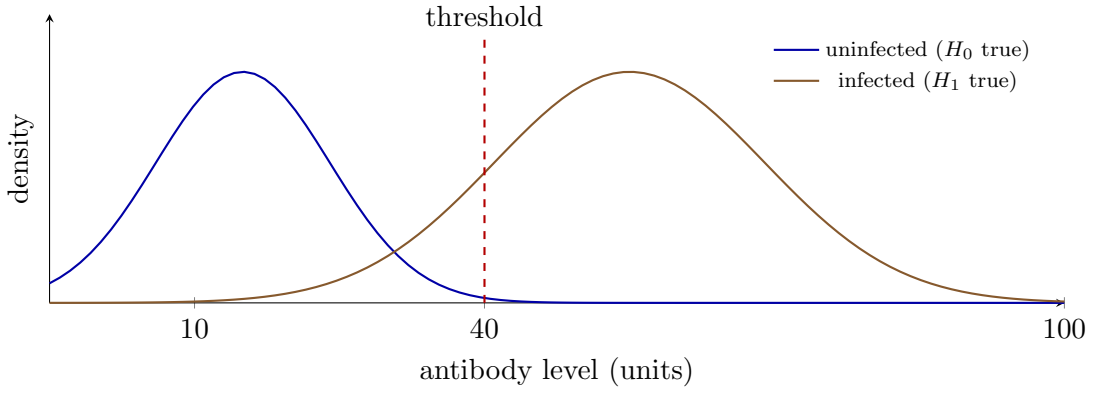
	Accept H_0	**Reject** H_0
H_0 **true**	Correct decision	**Type I error** (false positive)
H_1 **true**	**Type II error** (false negative)	Correct decision

A *Type I error* rejects H_0 when H_0 is actually true. A *Type II error* accepts H_0 when H_1 is actually true. The *significance level* of a test is defined as

$$\text{significance level} := P(\text{Type I error}),$$

conventionally set at 10%, 5%, or 1% in economics (occasionally 0.1% in legal or medical settings).

Intuition. We cannot generally minimise both error probabilities simultaneously: making the test stricter (harder to reject H_0) mechanically lowers the Type I error rate but raises the Type II error rate, and vice versa. The standard testing procedure therefore fixes the Type I error rate at the significance level first, and then evaluates the test's Type II error properties (its *power*) subject to that constraint.



Worked Example (Tutorial 1): the Type I / Type II trade-off in a COVID-19 antibody test

An antibody rapid test (ART) declares a person infected (reject H_0) if their measured antibody level exceeds a threshold. Take H_0: the individual is healthy (no COVID); H_1: the individual has COVID. Suppose the original threshold is 10 units.

Method 1: Nesting of the rejection regions. Let L be the measured antibody level (units) and c the threshold, so the test rejects H_0 when $L > c$. Then

$$P(\text{Type I}) = P(L > c \,|\, H_0 \text{ true}), \qquad P(\text{Type II}) = P(L \le c \,|\, H_1 \text{ true}).$$

For two thresholds $c < c'$, every level above c' is also above c, and every level at or below c is also at or below c':

$$\{L > c'\} \subseteq \{L > c\}, \qquad \{L \le c\} \subseteq \{L \le c'\}.$$

A smaller event has a smaller (or equal) probability under either hypothesis, hence

$$P(L > c' \,|\, H_0) \le P(L > c \,|\, H_0), \qquad P(L \le c' \,|\, H_1) \ge P(L \le c \,|\, H_1).$$

- Raising the threshold ($c = 10 \to c' = 100$ units) shrinks the rejection region, so fewer healthy people are flagged as infected: the Type I rate falls.
- The same move means infected people whose level lies between 10 and 100 units are now cleared, so the Type II rate (infected people missed) rises.
- Lowering the threshold ($c' = 10 \to c = 1$ unit) runs the same inclusions backwards: almost everyone is flagged, so the Type II rate falls and the Type I rate rises.
- Both error rates are population probabilities over all healthy (or all infected) people; a validation sample of tested people only estimates them by the share of false positives (or false negatives).

$$\boxed{\text{threshold} \uparrow \Rightarrow \text{Type I} \downarrow,\ \text{Type II} \uparrow; \qquad \text{threshold} \downarrow \Rightarrow \text{Type I} \uparrow,\ \text{Type II} \downarrow.}$$

Method 2: Tail probabilities under normal densities. Take the stylised curves in the figure above as $L \,|\, H_0 \sim N(15, 9^2)$ and $L \,|\, H_1 \sim N(55, 14^2)$, and let Φ be the standard normal CDF. Then

$$P(\text{Type I}) = 1 - \Phi\left(\frac{c-15}{9}\right), \qquad P(\text{Type II}) = \Phi\left(\frac{c-55}{14}\right).$$

c (units)	$(c-15)/9$, $(c-55)/14$	P(Type I)	P(Type II)
1	-1.556, -3.857	0.9401	0.0001
10	-0.556, -3.214	0.7107	0.0007
100	9.444, 3.214	0.0000	0.9993

Moving from $c = 10$ to $c = 100$ lowers the Type I rate from 0.7107 to about 0 and raises the Type II rate from 0.0007 to 0.9993; moving from $c = 10$ to $c = 1$ raises the Type I rate to 0.9401 and lowers the Type II rate to 0.0001.

$$\boxed{\text{threshold} \uparrow \Rightarrow \text{Type I} \downarrow,\ \text{Type II} \uparrow; \qquad \text{threshold} \downarrow \Rightarrow \text{Type I} \uparrow,\ \text{Type II} \downarrow.}$$

Common Mistake: misreading the significance level

The significance level is $P(\text{Type I error})$ — the probability of rejecting H_0 *given that* H_0 *is true*. It is *not* the probability that H_0 is true, and it is *not* the probability of making an error in general (it says nothing about the Type II error rate). Keep the conditioning event straight: significance level $= P(\text{reject } H_0 \,|\, H_0 \text{ true})$.

1.6 Randomized Controlled Trials and the Case for Econometrics

Definition 1.7: Randomized controlled trial (RCT)

A randomized controlled trial randomly assigns subjects to a *treatment group* (which receives some intervention, e.g. a job-training program) and a *control group* (which does not). Because assignment is random, the two groups are, *on average*, identical in every other respect — observed and unobserved — before treatment. Consequently, any difference in average outcomes between the two groups after the intervention can be attributed to the treatment itself:

$$\text{causal effect} \approx E(Y \mid \text{treatment}) - E(Y \mid \text{control}).$$

Why random assignment is the key step. If assignment were not random — say, subjects could choose whether to enter the training program — then people who expect to benefit most (e.g. the most motivated) might disproportionately select into treatment. The resulting difference in outcomes would then mix the true causal effect with this *selection effect*, and Case 3 (common factor: motivation drives both self-selection into treatment and the outcome) would contaminate the comparison. Random assignment rules this out by construction, since the coin flip that assigns treatment cannot be correlated with any pre-existing characteristic of the subject.

Policy Application: why RCTs are often infeasible in economics

Many of the causal questions economists care about cannot be studied with an RCT. A social planner cannot randomly assign years of education to people to study the return to education, nor randomly assign HSR lines to cities to study their effect on local growth — both are far too expensive, slow, or simply impossible to implement as a controlled experiment. Consequently, most data in economics is *observational* (nonexperimental): it is generated by people's and firms' own choices, not by a randomizer. Uncovering causal effects from observational data is much harder than from an RCT — this is precisely the HSR selection concern raised in Section 1.2 — but a large part of econometrics (Chapters 3 onward: the Least Squares Assumptions, omitted variable bias, and beyond) is devoted to getting as close as possible to the credibility of a real experiment even when one cannot be run.

Extension: the 2021 Nobel Prize and Difference-in-Differences (non-examinable preview)

The 2021 Nobel Memorial Prize in Economic Sciences was awarded (jointly) for methodological contributions to exactly this problem: extracting credible causal conclusions from observational data using natural experiments. One such technique, briefly named in this lecture but not yet developed, is *Difference-in-Differences*, which compares the *change* in outcomes over time between a group exposed to some event and a group that is not, rather than a single pre/post or treatment/control snapshot. It is not examinable in this course at this stage and is flagged here purely so the name is not unfamiliar if it resurfaces later.

1.7 Preview: The Simple Regression Model

Returning to the class-size example: the superintendent's question — "if I cut class size, what is the causal effect on test scores?" — can be written as a linear relationship. Writing $\text{TestScore} = \beta_0 + \beta_1 \, \text{STR}$ ignores every other factor that affects test scores (economic background, school facilities, the particular day of testing, ...), so we augment it with an error term:

$$\text{TestScore} = \beta_0 + \beta_1 \, \text{STR} + u.$$

In general, for observations $i = 1, \ldots, n$,

Definition 1.8: Simple linear regression model (preview)

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

where Y_i is the *dependent variable*, X_i the single *regressor*, β_0, β_1 are (unknown, population) *parameters*, and u_i is the *error term* representing every factor affecting Y_i other than X_i. This is called a *simple* linear regression model because it has exactly one regressor X.

We are typically most interested in the slope parameter,

$$\beta_1 = \frac{\Delta Y}{\Delta X},$$

the change in Y for a one-unit change in X. In the class-size example, $\beta_1 = \Delta\,\text{TestScore}/\Delta\,\text{STR}$, the change in average test score for a one-student change in the student–teacher ratio.

Worked Example (Lecture 1): interpreting β_1 numerically

Suppose (purely hypothetically, before we have estimated anything) that $\beta_1 = -2$. If a school district reduces its student–teacher ratio by 2 (i.e. $\Delta\text{STR} = -2$), we want the change in test score that the model predicts.

Method 1: Rearranging the slope definition. Multiply both sides of $\beta_1 = \Delta\text{TestScore}/\Delta\text{STR}$ by ΔSTR:

$$\Delta\text{TestScore} = \beta_1 \times \Delta\text{STR} = (-2) \times (-2) = 4.$$

- $\beta_1 = -2$ is measured in test-score points per student: each extra student per teacher lowers the district average score by 2 points.
- A cut of 2 students per teacher therefore raises the district average score by 4 points, holding every other factor in u fixed.
- β_1 is a population parameter; with data, Chapter 2 replaces it by the sample estimate $\hat{\beta}_1$, and the predicted change becomes $\hat{\beta}_1 \times \Delta\text{STR}$.

$$\boxed{\Delta\text{TestScore} = 4 \text{ points}}$$

Method 2: Difference of two levels of the regression line. Take a district moving from STR = 20 to STR = 18 with the same u. Its scores before and after are

$$\text{TestScore}_{\text{before}} = \beta_0 + \beta_1 \cdot 20 + u = \beta_0 - 40 + u,$$
$$\text{TestScore}_{\text{after}} = \beta_0 + \beta_1 \cdot 18 + u = \beta_0 - 36 + u.$$

Subtract; β_0 and u cancel, so the answer does not depend on the unknown intercept:

$$\Delta\text{TestScore} = (\beta_0 - 36 + u) - (\beta_0 - 40 + u) = -36 + 40 = 4.$$

With $\beta_0 = 700$ and $u = 0$, for instance, the score goes from 660 to 664 points.

$$\boxed{\Delta\text{TestScore} = 4 \text{ points}}$$

Worked Example (Tutorial 1): the meaning of β_0, β_1, and u in the class-size model

For $\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$:
Method 1: Term-by-term reading of the model.

- TestScore_i is the dependent variable: the average test score (in points) of district i; STR_i is the regressor: the district's average number of students per teacher.
- β_0 is the intercept (in points): the height of the population regression line at STR $= 0$, which only anchors the line, since no district has zero students per teacher.
- β_1 is the slope (in points per student): the change in the district average test score when STR rises by one student per teacher, other factors held fixed, $\beta_1 = \Delta\text{TestScore}/\Delta\text{STR}$.
- u_i is the error term (in points): all other factors affecting TestScore_i, e.g. students' economic background, school facilities, the particular day of testing.
- β_0, β_1, u_i are population objects and are unknown; the sample counterparts $\hat{\beta}_0, \hat{\beta}_1, \hat{u}_i$ come only in Chapter 2.

$$\boxed{\beta_0: \text{ intercept;} \quad \beta_1 = \Delta\text{TestScore}/\Delta\text{STR;} \quad u_i: \text{ everything else affecting TestScore}_i}$$

Method 2: Reading through the conditional mean. (Uses Least Squares Assumption 1 from Chapter 3.) If $E(u_i \,|\, \text{STR}_i) = 0$, take conditional means of the model:

$$E(\text{TestScore}_i \,|\, \text{STR}_i = x) = \beta_0 + \beta_1 x + E(u_i \,|\, \text{STR}_i = x) = \beta_0 + \beta_1 x.$$

Setting $x = 0$ gives $\beta_0 = E(\text{TestScore} \,|\, \text{STR} = 0)$, the intercept. Subtracting the values at $x+1$ and x gives the slope:

$$E(\text{TestScore} \,|\, \text{STR} = x+1) - E(\text{TestScore} \,|\, \text{STR} = x) = \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1.$$

Subtracting the conditional mean from the model gives everything else:

$$u_i = \text{TestScore}_i - E(\text{TestScore}_i \,|\, \text{STR}_i).$$

For instance, with $\beta_0 = 700$ and $\beta_1 = -2$, a district with STR $= 20$ has conditional mean $700 - 2 \times 20 = 660$ points; if it scores 670 points, then $u_i = 670 - 660 = 10$ points come from factors other than STR.

$$\boxed{\beta_0: \text{ intercept;} \quad \beta_1 = \Delta\text{TestScore}/\Delta\text{STR;} \quad u_i: \text{ everything else affecting TestScore}_i}$$

Common Mistake: treating the preview equation as already estimated

At this stage β_0, β_1 are *unknown population parameters* — we have not yet used any data to pin them down, and we have made no assumption yet about u (that is Chapter 3's job, via the Least Squares Assumptions). Do not write $\hat{\beta}_0, \hat{\beta}_1$ or $\hat{u}_i$ in this chapter: those hatted objects are the OLS *estimates* and *residuals* that Chapter 2 constructs from a sample. Keep "true, population" (β_0, β_1, u_i) and "estimated, sample" $(\hat{\beta}_0, \hat{\beta}_1, \hat{u}_i)$ strictly separate.

Exam Focus: Lecture 1 takeaways

- Be able to name and give a fresh example of each of the three correlation cases (causality, reverse causality, common factor) — Question 1/Tutorial 1 Q1 is a direct template for this.
- Be able to write down $Y_i = \beta_0 + \beta_1 X_i + u_i$ for a new applied context and state, in words, what β_0, β_1, and u_i mean — Question 1/Tutorial 1 Q2 is a direct template.
- Be comfortable applying linearity of expectation, independence $\Rightarrow$ zero covariance, and the Law of Iterated Expectations to a $Y = f(X, Z)$ style question with X, Z independent standard normal — Question 1/Tutorial 1 Q3 is a direct template, and this exact toolkit reappears in the unbiasedness proofs of Chapter 5.
- Be able to state the significance level as $P(\text{Type I error} \mid H_0 \text{ true})$ and explain, with an example, why lowering it raises the Type II error rate.

1.8 Chapter Summary

Key Formula: Chapter 1 Recap

Correlation between X, Y arises from causality ($X \to Y$), reverse causality ($Y \to X$), or a common factor ($C \to X$ and $C \to Y$); conditional mean $E(Y \mid X = x)$; significance level $= P(\text{Type I error})$; RCT causal effect $\approx E(Y \mid \text{treat}) - E(Y \mid \text{control})$.

$$E\big[E(Y \mid X)\big] = E[Y] \text{ (LIE)}, \qquad \operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y + 2\operatorname{Cov}(X,Y),$$
$$X \perp Y \Rightarrow \operatorname{Cov}(X,Y) = 0.$$

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \text{(preview; Chapter 2 estimates it, Chapter 3 justifies it).}$$

Exam Focus: what carries forward

- The three correlation cases and the RCT logic are qualitative but recur throughout the course as the justification for why Least Squares Assumption 1 ($E[u \mid X] = 0$) can fail (Chapter 3: selection bias, survivorship bias, omitted variables).
- The probability toolkit (Section 1.4) is not optional background — Chapters 4–6 cite Theorems 1.1–1.7 by name when deriving $\operatorname{Var}(\bar{Y})$, unbiasedness of $\hat{\beta}_0, \hat{\beta}_1$, and the Gauss–Markov theorem. Memorise the covariance bilinearity and variance-of-a-sum formulas cold.
- $Y_i = \beta_0 + \beta_1 X_i + u_i$ is only a *preview*: no assumption has yet been placed on u_i, and no estimator has yet been constructed. That is the entire content of Chapters 2–3.

2 Lecture 2: The OLS Estimator

This chapter answers three questions: (1) *how do we write down a linear relationship between one outcome Y and one explanatory variable X, and what does each piece of that equation mean?* (2) *given a sample of data, how do we choose the "best" line through the scatterplot?* (3) *once that line is chosen, how do we use it to predict Y and to measure the mistakes it makes?* It builds on Chapter 1 (*correlation, causality, and the informal statement of the class-size/test-score relationship*) and prepares Chapter 3 (*the assumptions under which the estimator derived here is a good one*).

2.1 The Simple Linear Regression Model

Chapter 1 introduced, informally, the idea that cutting class size might raise test scores. We now write this relationship precisely.

Definition 2.1: Simple linear regression model

For $i = 1, \ldots, n$ observations (X_i, Y_i),

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

Here Y_i is the *dependent variable* (or *explained variable*); X_i is the *independent variable* (or *explanatory variable*, or *regressor*); $\beta_0 + \beta_1 X$ is the *population regression line*; β_0 is its intercept and β_1 its slope; and u_i is the *error term*, which collects every other factor that affects Y_i besides X_i.

In the running example, $Y_i = \text{TestScore}_i$ and $X_i = \text{STR}_i$ (the student–teacher ratio) for school district i, so

$$\text{TestScore}_i = \beta_0 + \beta_1 \,\text{STR}_i + u_i.$$

The slope is the object of real interest for policy:

$$\beta_1 = \frac{\Delta\,\text{TestScore}}{\Delta\,\text{STR}} = \text{the change in test score for a one-unit change in STR.}$$

If, hypothetically, $\beta_1 = -2$, then reducing STR by 2 students per teacher would raise the predicted test score by $(-2) \times (-2) = 4$ points. We expect $\beta_1 < 0$ (smaller classes, higher scores), but the exact magnitude is an empirical question — this is precisely what OLS estimates.

Common Mistake: the intercept need not be meaningful

β_0 is simply the value of the population regression line *at* $X = 0$ — the point where the line crosses the Y-axis. In some applications this has a real interpretation (e.g., X = years of experience, β_0 = starting wage). In others it does not: STR = 0 describes a class with no students, and "predicted test score with zero students" has no sensible reading. When $X = 0$ lies far outside the range of the data, do not attach economic meaning to β_0 — report it as a mathematical necessity of the line, not a policy-relevant number. This is usually why applied work is "typically more interested in the value of β_1".

The error term u_i is not one specific omitted variable; it is the sum of *everything else* that makes district i's actual test score differ from what the population line predicts — overall family background, quality of school facilities, teacher quality, and so on, all bundled together. It is convenient, though not yet justified, to think of u_i as "noise" around the line; Chapter 3 states precisely what we must assume about u_i for OLS to recover β_0, β_1 in a meaningful sense.

2.2 Three Kinds of "Beta": Parameters, Candidates, Estimators

The lecture slides move between three different pieces of notation — (β_0, β_1), (b_0, b_1), and $(\hat{\beta}_0, \hat{\beta}_1)$ — without ever pausing to contrast them side by side (Tutorial 2, Question 2, asks for exactly this contrast, and the official tutorial slides for that question repeat the prompt four times with no answer recorded).

Getting this distinction straight is essential: almost every derivation in this course, and every proof about unbiasedness or efficiency in Chapters 3 and 5, turns on knowing *which* of these three objects a symbol refers to at each step.

Definition 2.2: True parameters, candidate values, and OLS estimators

- (β_0, β_1) — the **true, unknown population parameters**. They are fixed numbers (not random) that describe the actual population regression line $\beta_0 + \beta_1 X$. We never observe them directly; the whole point of the course is to learn about them from a sample.
- (b_0, b_1) — **generic candidate values**. These are the dummy arguments of the minimization problem in Section 2.4 below: we plug in *some* numbers b_0, b_1, compute how badly the resulting line $b_0 + b_1 X$ fits the data, and then search over all possible $(b_0, b_1) \in \mathbb{R}^2$ for the pair that fits best. (b_0, b_1) are not attached to any particular sample or estimator; they are placeholders used only to set up and solve the optimization problem.
- $(\hat{\beta}_0, \hat{\beta}_1)$ — the **OLS estimators**. These are the *specific* values of (b_0, b_1) that solve the minimization problem for a given sample $\{(X_i, Y_i)\}_{i=1}^n$ — i.e., the minimizers. Because they are computed from the sample, and the sample is random, $(\hat{\beta}_0, \hat{\beta}_1)$ are themselves **random variables**: draw a different sample of n districts and you get different numbers $\hat{\beta}_0, \hat{\beta}_1$, even though the underlying (β_0, β_1) never change. $(\hat{\beta}_0, \hat{\beta}_1)$ are *estimates of* the fixed unknowns (β_0, β_1).

Symbol	What it is	Random or fixed?
(β_0, β_1)	true population parameters	fixed, unknown constants
(b_0, b_1)	generic candidate values in $\min_{b_0, b_1}$	dummy arguments, not attached to data
$(\hat{\beta}_0, \hat{\beta}_1)$	OLS estimators = minimizers of $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$	random – functions of the sample $\{(X_i, Y_i)\}$

Common Mistake: treating $\hat{\beta}_1$ as fixed

It is tempting to treat $\hat{\beta}_1 = -2.28$ (the California estimate) as "the" slope, full stop. But $\hat{\beta}_1$ is only *one realisation* of a random variable: a different sample of 420 California districts drawn from the same population would give a slightly different $\hat{\beta}_1$. Chapters 4–6 build the sampling distribution of $\hat{\beta}_1$ precisely because of this randomness; keep (β_0, β_1) (fixed target) and $(\hat{\beta}_0, \hat{\beta}_1)$ (random estimator) conceptually separate from the outset.

2.3 Summation Notation and Two Algebraic Identities

The derivation of $\hat{\beta}_0, \hat{\beta}_1$ leans on a handful of purely mechanical summation manipulations. For a constant a and sequences $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$\sum_{i=1}^n a\, x_i = a \sum_{i=1}^n x_i, \qquad \sum_{i=1}^n (x_i + y_i) = \sum_{i=1}^n x_i + \sum_{i=1}^n y_i, \qquad \sum_{i=1}^n a = n\, a.$$

These three rules — linearity of the sum and the sum of a constant — are all that is needed to expand every squared or cross-product term below.

A first, small consequence is used repeatedly and is worth isolating on its own:

$$\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - \sum_{i=1}^n \bar{X} = \sum_{i=1}^n X_i - n\bar{X} = \sum_{i=1}^n X_i - n\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = 0,$$

i.e. *deviations from the sample mean always sum to zero*, for any sequence of numbers whatsoever (this uses no assumption about the regression model at all).

We now prove the two identities behind Question 2, Q4, which rewrite a sum of *cross-deviations* $(X_i - \bar{X})(Y_i - \bar{Y})$ and a sum of *squared deviations* $(X_i - \bar{X})^2$ in terms of raw sums $\sum_{i=1}^n X_i Y_i$ and $\sum_{i=1}^n X_i^2$. These are exactly the identities used to collapse the normal equations into the closed-form OLS estimator in Section 2.5.

Theorem 2.1: Two summation identities

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}, \qquad \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2.$$

Equivalently, dividing both sides of each identity by n,

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}, \qquad \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2.$$

Proof. Expand the product $(X_i - \bar{X})(Y_i - \bar{Y})$ term by term:

$$(X_i - \bar{X})(Y_i - \bar{Y}) = X_i Y_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}.$$

Sum both sides over $i = 1, \ldots, n$ and apply linearity of the sum:

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

Now use $\sum_{i=1}^n X_i = n\bar{X}$, $\sum_{i=1}^n Y_i = n\bar{Y}$, and $\sum_{i=1}^n \bar{X}\bar{Y} = n\bar{X}\bar{Y}$ (the last since $\bar{X}\bar{Y}$ is a constant, summed n times):

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

The last three terms combine to $-n\bar{X}\bar{Y}$, giving

$$\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y},$$

which proves the first identity. The second identity is the special case $Y_i = X_i$ (so $\bar{Y} = \bar{X}$): substituting into what was just proved,

$$\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n X_i X_i - n\bar{X}\bar{X} \quad \Longrightarrow \quad \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2.$$

Dividing each identity through by n gives the second (equivalent) display in the statement. □

Common Mistake: a missing $1/n$ on the official question sheet

Question 2, Q4, states the first identity as

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

i.e. with a $\frac{1}{n}$ on the right-hand side but *not* on the left. As Theorem 2.1 shows, this is only correct once the left-hand side is *also* divided by n:

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

As printed (without the $\frac{1}{n}$ on the left), the equation is off by a factor of n — for the California data with $n = 420$, that is a two-orders-of-magnitude error. The version with $\frac{1}{n}$ on *both* sides, or with $\frac{1}{n}$ on *neither* side (Theorem 2.1's first display, which is the form actually used inside the OLS derivation), are the only two internally consistent statements. Always check which one you are using before substituting into $\hat{\beta}_1$.

2.4 The OLS Minimization Problem

Since (β_0, β_1) and the population regression line are unknown, we must use the sample to construct an estimate. A scatterplot of the $n = 420$ California districts suggests a rough downward-sloping cloud of points, but different people, drawing a line "by eye", would draw different lines. We need an objective, reproducible rule for choosing the best line.

Definition 2.3: Ordinary Least Squares (OLS) estimator

For candidate values b_0, b_1, the *prediction error* (or mistake) in predicting Y_i from X_i using the line $b_0 + b_1 X$ is

$$Y_i - (b_0 + b_1 X_i).$$

The *sum of squared errors* for the whole sample is

$$\sum_{i=1}^{n}\big[Y_i - (b_0 + b_1 X_i)\big]^2 = \big[Y_1 - (b_0 + b_1 X_1)\big]^2 + \cdots + \big[Y_n - (b_0 + b_1 X_n)\big]^2.$$

The **OLS estimators** $\hat{\beta}_0, \hat{\beta}_1$ are the values of b_0, b_1 that solve

$$\min_{b_0,\, b_1} \sum_{i=1}^{n}\big[Y_i - (b_0 + b_1 X_i)\big]^2.$$

Intuitively, we search over every possible pair (b_0, b_1) and keep the one that makes the fitted line $b_0 + b_1 X$ as close as possible to the observed data, where "closeness" is measured by the total squared vertical distance from each point to the line. Squaring serves two purposes: it makes positive and negative errors count equally (an error of $+3$ is penalised as heavily as -3), and it penalises large errors disproportionately more than small ones. In practice, we do not search blindly; we solve the first-order conditions of this minimization problem.

2.5 Deriving the Closed-Form OLS Estimator

Write $S(b_0, b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$ for the objective function. This is the single most exam-heavy derivation in the course (Lecture 2; Question 2, Q1 and Q3; Tutorial 2, Q1 and Q3) and is reproduced

here in full, every algebraic step on its own line.

Theorem 2.2: The OLS estimator, closed form

The values $(\hat{\beta}_0, \hat{\beta}_1)$ that minimise $S(b_0, b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$ are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

provided $\sum_{i=1}^n (X_i - \bar{X})^2 > 0$ (equivalently, X is not the same number for every observation).

Proof. **Step 1: first-order conditions.** Differentiate $S(b_0, b_1)$ with respect to each candidate value using the chain rule:

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^n 2(Y_i - b_0 - b_1 X_i)(-1) = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i),$$

$$\frac{\partial S}{\partial b_1} = \sum_{i=1}^n 2(Y_i - b_0 - b_1 X_i)(-X_i) = -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) X_i.$$

The minimisers $(\hat{\beta}_0, \hat{\beta}_1)$ are the values of (b_0, b_1) at which both partial derivatives equal zero (Section 2.5 below confirms this stationary point is in fact the unique global minimum, not merely a critical point). Setting $\partial S/\partial b_0 = 0$ and $\partial S/\partial b_1 = 0$ and dropping the harmless factor of -2:

$$\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \qquad \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0.$$

Step 2: reduce to the two normal equations. Expand the first condition using linearity of the sum:

$$\sum_{i=1}^n Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n X_i = 0.$$

Divide through by n, and use $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$:

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0. \tag{4}$$

Expand the second condition similarly:

$$\sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \sum_{i=1}^n X_i - \hat{\beta}_1 \sum_{i=1}^n X_i^2 = 0.$$

Divide through by n:

$$\frac{1}{n}\sum_{i=1}^n X_i Y_i - \hat{\beta}_0 \bar{X} - \hat{\beta}_1 \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) = 0. \tag{5}$$

Equations (4) and (5) are the *normal equations*: two linear equations in the two unknowns $\hat{\beta}_0, \hat{\beta}_1$.

Step 3: solve (4) for $\hat{\beta}_0$. Equation (4) rearranges immediately to

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

This already gives the intercept formula, in terms of the still-unknown $\hat{\beta}_1$.

Step 4: substitute into (5) and isolate $\hat{\beta}_1$. Replace $\hat{\beta}_0$ in (5) by $\bar{Y} - \hat{\beta}_1\bar{X}$:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - (\bar{Y} - \hat{\beta}_1\bar{X})\bar{X} - \hat{\beta}_1\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = 0. \tag{6}$$

Expand the bracket $(\bar{Y} - \hat{\beta}_1\bar{X})\bar{X} = \bar{X}\bar{Y} - \hat{\beta}_1\bar{X}^2$:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y} + \hat{\beta}_1\bar{X}^2 - \hat{\beta}_1\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = 0. \tag{7}$$

Collect the two terms in $\hat{\beta}_1$ on one side and everything else on the other:

$$\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y} = \hat{\beta}_1\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right).$$

Since $\Sigma(X_i - \bar{X})^2 > 0$ by hypothesis, Theorem 2.1 guarantees $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$ as well, so this coefficient of $\hat{\beta}_1$ is strictly positive and we may divide by it:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2}. \tag{8}$$

Step 5: convert to deviation form. Multiply numerator and denominator of (8) by n:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}.$$

By Theorem 2.1, the numerator equals $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ and the denominator equals $\sum_{i=1}^{n}(X_i - \bar{X})^2$, so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Finally, multiplying numerator and denominator by the same constant $\frac{1}{n-1}$ leaves the ratio unchanged and turns each sum into a sample covariance/variance,

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2},$$

where $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is the sample covariance of X and Y, and $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance of X. Substituting this $\hat{\beta}_1$ back into Step 3's expression $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ completes the proof. □

Key Formula: the OLS estimator

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}.$$

Requires $\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$: X must vary across the sample, or the slope is not identified (dividing by zero).

Extension: second-order condition – why the stationary point is a minimum

The first-order conditions in Step 1 above locate a *stationary point* of $S(b_0, b_1)$, but stationarity alone does not distinguish a minimum from a maximum or a saddle. Because S is quadratic in (b_0, b_1), its second derivatives are constants:

$$\frac{\partial^2 S}{\partial b_0^2} = 2n, \qquad \frac{\partial^2 S}{\partial b_0\, \partial b_1} = 2\sum_{i=1}^{n} X_i, \qquad \frac{\partial^2 S}{\partial b_1^2} = 2\sum_{i=1}^{n} X_i^2.$$

The Hessian matrix is therefore

$$H = 2\begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix}.$$

Its leading principal minor is $2n > 0$. Its determinant is

$$\det H = 4\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right) = 4n\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) = 4n\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

using $\sum_{i=1}^{n} X_i = n\bar{X}$ and Theorem 2.1. Whenever $\sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$ (the same non-degeneracy condition needed for $\hat{\beta}_1$ to exist), $\det H > 0$, so H is **positive definite**: $S(b_0, b_1)$ is strictly convex in (b_0, b_1), and the stationary point found from the first-order conditions is the unique **global** minimum. This is why solving the FOCs is enough — there is no need to separately rule out maxima or saddle points.

2.6 Fitted Values, Residuals, and the Mean Point

Once $\hat{\beta}_0, \hat{\beta}_1$ are computed from the sample, the estimated (or *OLS*) regression line $\hat{\beta}_0 + \hat{\beta}_1 X$ can be used to predict Y at any value of X, and to measure how far actual data points fall from that prediction.

Definition 2.4: OLS predicted values and residuals

For $i = 1, \ldots, n$, the *OLS predicted value* (or *fitted value*) is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

and the *OLS residual* is the difference between the actual and predicted value,

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i).$$

Common Mistake: confusing u_i with $\hat{u}_i$

u_i is the *true, unobservable* population error in $Y_i = \beta_0 + \beta_1 X_i + u_i$ — we never see it, because we never see β_0, β_1. $\hat{u}_i$ is the *observable OLS residual*, computed from the *estimated* line. $\hat{u}_i$ is the sample's estimate of u_i, but the two are not the same object, and $\hat{u}_i$ depends on which sample happened to be drawn. Never write "$\sum_{i=1}^{n} u_i = 0$" — only $\sum_{i=1}^{n} \hat{u}_i = 0$ holds (and it holds automatically, by construction, from the first first-order condition of Section 2.5). Chapter 3 develops this residual identity, together with $\sum_{i=1}^{n} \hat{u}_i X_i = 0$ and $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$, as formal results in their own right.

A useful and immediate consequence of the derivation above is that the OLS line always passes through the point of sample means $(\bar{X}, \bar{Y})$ – a fact used directly in Tutorial 2, Q5(c).

Theorem 2.3: The OLS line passes through $(\bar{X}, \bar{Y})$

The fitted value of Y at $X = \bar{X}$ equals $\bar{Y}$:

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y}.$$

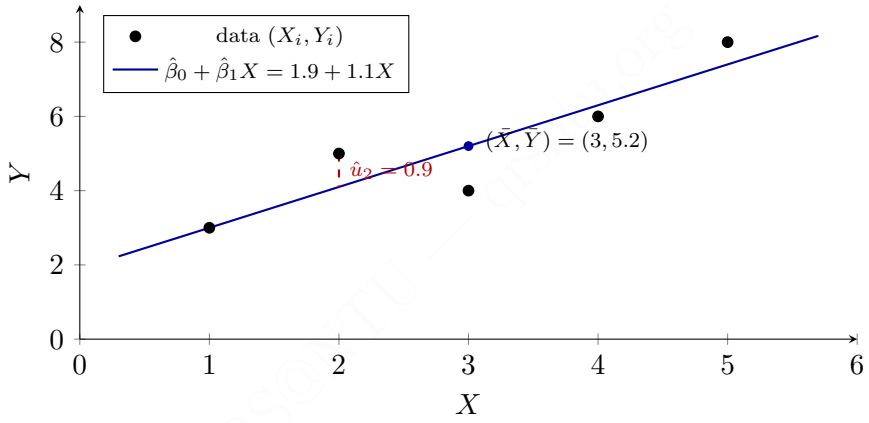
Proof. By Step 3 of the proof of Theorem 2.2, the intercept satisfies $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Adding $\hat{\beta}_1 \bar{X}$ to both sides gives directly

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \bar{X} = \bar{Y},$$

as claimed. Geometrically, the fitted line is anchored at the centroid $(\bar{X}, \bar{Y})$ of the scatterplot and rotates about that point as $\hat{\beta}_1$ changes. □

2.7 A Load-Bearing Figure: Fitting a Line Through Data

The picture below uses the small hand-worked dataset of Section 2.9 ($n = 5$; $\hat{\beta}_0 = 1.9$, $\hat{\beta}_1 = 1.1$) to show what the minimization problem actually does: it chooses the line that makes the sum of the squared *vertical* distances (one of which, at $X = 2$, is marked) as small as possible.



Every other candidate line through this scatter – steeper, flatter, shifted up or down – gives a strictly larger sum of squared vertical distances than $1.9 + 1.1X$ does; this is exactly the content of Theorem 2.2 and the convexity argument in Section 2.5.

2.8 Worked Examples: California Test Scores

Worked Example (Lecture 2): the California regression line

Fitting the model $\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + u_i$ by OLS to the $n = 420$ California school districts (`caschool.dta`) gives, reported in S&W style with standard errors in parentheses beneath each coefficient,

$$\widehat{\text{TestScore}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28}\,\text{STR}.$$

(a) Prediction at STR $= 20$.
Method 1: Substitution into the fitted line. Set STR $= 20$, multiply out the slope term, then subtract it from the intercept:

$$\begin{aligned}\widehat{\text{TestScore}}\big|_{\text{STR}=20} &= 698.9 - 2.28 \times 20 \\ &= 698.9 - 45.6 \\ &= 653.3.\end{aligned}$$

$$\boxed{\widehat{\text{TestScore}}\big|_{\text{STR}=20} = 653.3.}$$

(b) The Antelope, CA district. This district has STR $= 19.33$ and an actual test score of 657.75.
Method 1: Substitution, then actual minus predicted. The slope term is $2.28 \times 19.33 = 44.0724$, so the prediction is the intercept minus this term, and the residual is the actual score minus the prediction:

$$\begin{aligned}\widehat{\text{TestScore}}_{\text{Antelope}} &= 698.9 - 44.0724 = 654.8276 = 654.83 && (2\text{ d.p.}),\\ \hat{u}_{\text{Antelope}} &= 657.75 - 654.8276 = 2.9224 = 2.92 && (2\text{ d.p.}).\end{aligned}$$

$$\boxed{\widehat{\text{TestScore}}_{\text{Antelope}} = 654.83, \qquad \hat{u}_{\text{Antelope}} = 2.92.}$$

Method 2: Moving along the line from the STR $= 20$ prediction. Two fitted values differ by the slope times the change in STR, because the intercept cancels:

$$\begin{aligned}\widehat{\text{TestScore}}\big|_{19.33} - \widehat{\text{TestScore}}\big|_{20} &= \big(698.9 - 2.28 \times 19.33\big) - \big(698.9 - 2.28 \times 20\big) \\ &= -2.28 \times (19.33 - 20) = -2.28 \times (-0.67) = 1.5276.\end{aligned}$$

Add this to the part (a) prediction, then subtract from the actual score:

$$\begin{aligned}\widehat{\text{TestScore}}_{\text{Antelope}} &= 653.3 + 1.5276 = 654.8276 = 654.83 && (2\text{ d.p.}),\\ \hat{u}_{\text{Antelope}} &= (657.75 - 653.3) - 1.5276 = 4.45 - 1.5276 = 2.9224 = 2.92 && (2\text{ d.p.}).\end{aligned}$$

$$\boxed{\widehat{\text{TestScore}}_{\text{Antelope}} = 654.83, \qquad \hat{u}_{\text{Antelope}} = 2.92.}$$

Policy Application: what a school board can and cannot conclude

The fitted slope $\hat{\beta}_1 = -2.28$ says: *across the 420 sampled districts, a district with one more student per teacher has, on average, a test score* 2.28 *points lower.* A school board cutting STR by, say, 2 (e.g. by hiring more teachers) might be tempted to read this as "test scores will rise by $2 \times 2.28 = 4.56$ points if we do this." That causal reading is only valid under the Least Squares Assumptions of Chapter 3 (in particular, that u is unrelated to STR on average, i.e. no confounding factor – such as district wealth – simultaneously drives both class size and test scores). Lecture 2 and this chapter establish only the *mechanics* of fitting the line; whether $\hat{\beta}_1$ is a believable estimate of the *causal* effect of class size is a separate question taken up next chapter.

QRS@NTU – qrsntu.org

2.9 Worked Example: A Hand-Computed OLS Regression

Worked Example (Lecture 2): computing $\hat{\beta}_0, \hat{\beta}_1$ by hand

Consider a toy sample of $n = 5$ observations:

i	1	2	3	4	5
X_i	1	2	3	4	5
Y_i	3	5	4	6	8

Method 1: Closed form in deviation form. The sample means are

$$\bar{X} = \tfrac{1}{5}(1+2+3+4+5) = \tfrac{15}{5} = 3, \qquad \bar{Y} = \tfrac{1}{5}(3+5+4+6+8) = \tfrac{26}{5} = 5.2.$$

Subtract the means observation by observation and form the products:

$X_i - \bar{X}$	$1-3=-2$	$2-3=-1$	$3-3=0$	$4-3=1$	$5-3=2$
$Y_i - \bar{Y}$	$3-5.2=-2.2$	$5-5.2=-0.2$	$4-5.2=-1.2$	$6-5.2=0.8$	$8-5.2=2.8$
$(X_i-\bar{X})(Y_i-\bar{Y})$	$(-2)(-2.2)=4.4$	$(-1)(-0.2)=0.2$	$0(-1.2)=0$	$1(0.8)=0.8$	$2(2.8)=5.6$
$(X_i-\bar{X})^2$	$(-2)^2=4$	$(-1)^2=1$	$0^2=0$	$1^2=1$	$2^2=4$

Sum each row, then divide and back-substitute:

$$\sum_{i=1}^{n}(X_i-\bar{X}) = -2-1+0+1+2 = 0, \qquad \sum_{i=1}^{n}(X_i-\bar{X})^2 = 4+1+0+1+4 = 10,$$

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = 4.4+0.2+0+0.8+5.6 = 11.0,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{11.0}{10} = 1.1,$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 5.2 - 1.1(3) = 5.2 - 3.3 = 1.9.$$

The fitted values $\hat{Y}_i = 1.9 + 1.1X_i$ and residuals $\hat{u}_i = Y_i - \hat{Y}_i$ are

$\hat{Y}_i$	$1.9+1.1=3.0$	$1.9+2.2=4.1$	$1.9+3.3=5.2$	$1.9+4.4=6.3$	$1.9+5.5=7.4$
$\hat{u}_i$	$3-3.0=0$	$5-4.1=0.9$	$4-5.2=-1.2$	$6-6.3=-0.3$	$8-7.4=0.6$

so that $\sum_{i=1}^{n}\hat{u}_i = 0+0.9-1.2-0.3+0.6 = 0$ and $\hat{Y}\big|_{X=\bar{X}=3} = 1.9+1.1(3) = 5.2 = \bar{Y}$ (Theorem 2.3).

$$\boxed{\hat{\beta}_0 = 1.9, \qquad \hat{\beta}_1 = 1.1, \qquad \hat{Y}_i = 1.9 + 1.1X_i.}$$

Method 2: Normal equations from the first-order conditions. The raw sums are $\sum_{i=1}^{n}X_i = 15$, $\sum_{i=1}^{n}Y_i = 26$ and

$$\sum_{i=1}^{n}X_i^2 = 1+4+9+16+25 = 55,$$

$$\sum_{i=1}^{n}X_iY_i = 1(3)+2(5)+3(4)+4(6)+5(8) = 3+10+12+24+40 = 89.$$

The first-order conditions expand to $\sum_{i=1}^{n}Y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}X_i$ and $\sum_{i=1}^{n}X_iY_i = \hat{\beta}_0\sum_{i=1}^{n}X_i + \hat{\beta}_1\sum_{i=1}^{n}X_i^2$. Substitute the sums, multiply the first by 3 and subtract it from the second:

$$26 = 5\hat{\beta}_0 + 15\hat{\beta}_1, \qquad 89 = 15\hat{\beta}_0 + 55\hat{\beta}_1,$$

$$89 - 3(26) = (15\hat{\beta}_0 + 55\hat{\beta}_1) - (15\hat{\beta}_0 + 45\hat{\beta}_1) \implies 11 = 10\hat{\beta}_1 \implies \hat{\beta}_1 = 1.1,$$

$$5\hat{\beta}_0 = 26 - 15(1.1) = 26 - 16.5 = 9.5 \implies \hat{\beta}_0 = 1.9.$$

Here $89 - 78 = \sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}$ and $55 - 45 = \sum_{i=1}^{n}X_i^2 - n\bar{X}^2$ (Theorem 2.1).

$$\boxed{\hat{\beta}_0 = 1.9, \qquad \hat{\beta}_1 = 1.1, \qquad \hat{Y}_i = 1.9 + 1.1X_i.}$$

2.10 Worked Examples: Class Size and Test Scores (Tutorial 2)

Worked Example (Tutorial 2): predictions from a second regression

A researcher, using data on class size CS and average test scores from 50 third-grade classes, estimates

$$\widehat{\text{TestScore}} = 640.3 - 4.93 \times CS.$$

(a) A classroom with $CS = 25$ students.
Method 1: Substitution into the fitted line.

$$\widehat{\text{TestScore}}\big|_{CS=25} = 640.3 - 4.93 \times 25 = 640.3 - 123.25 = 517.05.$$

$$\boxed{\widehat{\text{TestScore}}\big|_{CS=25} = 517.05.}$$

(b) Predicted change from $CS = 21$ (last year) to $CS = 24$ (this year).
Method 1: Difference of two predictions.

$$\begin{aligned}
\widehat{\text{TestScore}}\big|_{CS=21} &= 640.3 - 4.93 \times 21 = 640.3 - 103.53 = 536.77,\\
\widehat{\text{TestScore}}\big|_{CS=24} &= 640.3 - 4.93 \times 24 = 640.3 - 118.32 = 521.98,\\
\Delta\widehat{\text{TestScore}} &= 521.98 - 536.77 = -14.79.
\end{aligned}$$

$$\boxed{\Delta\widehat{\text{TestScore}} = -14.79 \text{ points.}}$$

Method 2: Slope times change in CS. Subtract the two fitted values; the intercept cancels:

$$\Delta\widehat{\text{TestScore}} = (\hat\beta_0 + 24\hat\beta_1) - (\hat\beta_0 + 21\hat\beta_1) = \hat\beta_1(24 - 21) = -4.93 \times 3 = -14.79.$$

$$\boxed{\Delta\widehat{\text{TestScore}} = -14.79 \text{ points.}}$$

(c) Sample average test score, given $\overline{CS} = 22.8$.
Method 1: Mean point of the fitted line. By Theorem 2.3 the line passes through $(\overline{CS}, \overline{\text{TestScore}})$:

$$\overline{\text{TestScore}} = \hat\beta_0 + \hat\beta_1\,\overline{CS} = 640.3 - 4.93 \times 22.8 = 640.3 - 112.404 = 527.896.$$

$$\boxed{\overline{\text{TestScore}} = 527.896.}$$

Method 2: Averaging fitted values and residuals. Write $\text{TestScore}_i = \widehat{\text{TestScore}}_i + \hat u_i$ and average over the $n = 50$ classes, using $\sum_{i=1}^n \hat u_i = 0$ (the first first-order condition):

$$\begin{aligned}
\overline{\text{TestScore}} &= \frac{1}{n}\sum_{i=1}^n (\hat\beta_0 + \hat\beta_1 CS_i) + \frac{1}{n}\sum_{i=1}^n \hat u_i = \frac{1}{n}(n\hat\beta_0) + \hat\beta_1\Big(\frac{1}{n}\sum_{i=1}^n CS_i\Big) + 0\\
&= \hat\beta_0 + \hat\beta_1\,\overline{CS} = 640.3 - 4.93 \times 22.8 = 640.3 - 112.404 = 527.896.
\end{aligned}$$

$$\boxed{\overline{\text{TestScore}} = 527.896.}$$

2.11 Interpretation in Words, and Further Traps

In prose, the California result reads: *"On average, districts with one more student per teacher have test scores that are* 2.28 *points lower."* Two features of this sentence are worth flagging every time a slope coefficient is interpreted:

- the units of $\hat{\beta}_1$ are always (*units of* Y) per (*unit of* X) – here, test-score points per student-per-teacher;
- the word "on average" matters: $\hat{\beta}_1$ describes the *average* relationship across districts in the sample, not a guarantee for any single district (Antelope, above, deviates from the line by +2.92 points).

Common Mistake: extrapolating $\hat{\beta}_0$ or the line beyond the data

Two related traps recur in exam answers. First, quoting $\hat{\beta}_0 = 698.9$ as "the predicted test score when STR= 0" without noting that no district in the sample has anywhere near STR= 0 – the fitted line is only a reliable description of the data over the *range of* X *actually observed* (roughly STR $\in [14, 25]$ for the California data); extrapolating to $X = 0$ treats the line as if it held far outside where it was estimated. Second, reporting $\hat{\beta}_1$ without its standard error and treating it as if it were the exact, fixed population slope β_1 – recall from Section 2.2 that $\hat{\beta}_1$ is a random estimator, not β_1 itself.

Exam Focus: reproducing the derivation

- Be able to reproduce Theorem 2.2's five-step derivation from a blank page: set up $S(b_0, b_1)$, take the two FOCs, form normal equations (4)–(5), solve (4) for $\hat{\beta}_0$, substitute into (5), and simplify to $\hat{\beta}_1 = s_{XY}/s_X^2$ using Theorem 2.1. This exact sequence is asked for directly in Question 2, Q3, and appears again inside Chapter 5's unbiasedness proofs.
- Know the three-way notation distinction of Section 2.2 cold – (β_0, β_1) vs. (b_0, b_1) vs. $(\hat{\beta}_0, \hat{\beta}_1)$ – since the official Tut 2 slides leave this unanswered and it is an easy place to lose marks by conflating symbols.
- Given any estimated line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, be ready to (i) predict $\hat{Y}$ at a stated X; (ii) compute a *change* in predicted Y two ways (difference of two predictions, or $\hat{\beta}_1 \times \Delta X$) and confirm they agree; and (iii) recover $\overline{Y}$ from $\overline{X}$ using Theorem 2.3, without needing the raw data – exactly the three tasks in Tutorial 2, Q5.

2.12 Chapter Summary

Key Formula: Chapter 2 recap

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad S(b_0, b_1) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \qquad \hat{u}_i = Y_i - \hat{Y}_i,$$

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y}.$$

Exam Focus: Chapter 2 takeaways

- (β_0, β_1) = fixed unknown truth; (b_0, b_1) = dummy candidates in the minimization; $(\hat{\beta}_0, \hat{\beta}_1)$ = the random, sample-dependent minimizers – keep all three separate.
- OLS minimizes the sum of squared vertical prediction errors $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$; the FOCs give the normal equations (4)–(5), which solve to $\hat{\beta}_1 = s_{XY}/s_X^2$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and this stationary point is a genuine minimum because $S(b_0, b_1)$ is strictly convex whenever X varies.
- The fitted line always passes through $(\bar{X}, \bar{Y})$ – a shortcut for recovering $\bar{Y}$ from $\bar{X}$ and the fitted coefficients without new data.
- u_i (true, unobserved error) and $\hat{u}_i$ (observed OLS residual) are different objects; only $\sum_{i=1}^{n} \hat{u}_i = 0$ holds, and only by construction of the OLS FOCs – the full residual-identity toolkit ($\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = 0$, $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$) is developed formally in Chapter 3.

QRS@NTU – qrsntu.org

3 Lecture 3: The Least Squares Assumptions

This chapter answers three questions: (1) *what, precisely, do "population" and "OLS" versions of the regression line, the error, and the fitted value mean, and how do they relate*; (2) *what mechanical identities does the OLS residual automatically satisfy, purely because* $\hat{\beta}_0, \hat{\beta}_1$ *solve the first-order conditions*; (3) *under what conditions on the data-generating process does* $\hat{\beta}_1$ *actually estimate the causal slope* β_1*, rather than some contaminated mixture of causation and selection.* It builds on Chapter 2 (*the OLS estimator and its FOC derivation*) and prepares Chapter 4 (*random sampling,* R^2*, and the sampling distribution of the OLS estimators*), which leans on the residual identities proved here.

3.1 From an Estimated Line Back to the Population: CEO Salary and ROE

A firm's return on equity (ROE) is

$$\text{ROE} = \frac{\text{net income in a fiscal year}}{\text{shareholders' investment}},$$

a standard measure of how much profit a firm generates per dollar invested; ROEs of 15–20% are considered good, and firms such as Google, Amazon, Facebook, and Apple can reach ROEs near 30%. To study how firm performance (`roe`) affects the CEO's pay, consider the simple linear regression model

$$\texttt{salary}_i = \beta_0 + \beta_1 \texttt{roe}_i + u_i,$$

where `salary` is measured in \$1,000s and `roe` in percentage points, so $\beta_1 = \Delta\texttt{salary}/\Delta\texttt{roe}$ is the change in CEO salary (in \$1,000s) for a one-percentage-point increase in ROE.

Using the dataset `ceosal1.dta` ($n = 209$ US CEOs, from *Business Week*), OLS gives

$$\widehat{\texttt{salary}} = 963.191 + 18.501\,\texttt{roe}, \qquad n = 209.$$

Worked Example (Lecture 3): predicting CEO salary from ROE

(a) Interpretation. Compare the fitted line at two values of `roe` one percentage point apart:

$$\begin{aligned}\Delta\widehat{\texttt{salary}} &= \left[963.191 + 18.501(\texttt{roe} + 1)\right] - \left[963.191 + 18.501\,\texttt{roe}\right] \\ &= 18.501 \text{ (\$1,000s)}.\end{aligned}$$

- $\hat{\beta}_1 = 18.501$ is the sample (OLS) estimate of the population slope β_1: in this sample of $n = 209$ CEOs, a firm whose ROE is one percentage point higher is predicted to pay its CEO 18.501 thousand dollars, i.e. about \$18,500, more per year.
- The effect scales linearly: an ROE of 20% instead of 10% ($\Delta\texttt{roe} = 10$) raises predicted salary by $18.501 \times 10 = 185.01$, i.e. about \$185,000.
- $\hat{\beta}_0 = 963.191$ is the predicted salary at $\texttt{roe} = 0$, i.e. about \$963,000; it is the height of the OLS line, not an observed CEO's pay.
- $\hat{\beta}_1$ estimates the unknown population β_1; it is read as the causal effect of ROE on pay only if Least Squares Assumption 1 holds, which this chapter examines below.

(b) Predicted value and residual for the first observation. The first CEO in the sample has $\texttt{roe}_1 = 14.1$ and actual salary $\texttt{salary}_1 = 1095$ (i.e. \$1,095,000).

Method 1: Substitution into the OLS line, then $\hat{u}_1 = Y_1 - \hat{Y}_1$**.** Substitute $\texttt{roe}_1 = 14.1$ into the OLS line. The slope term is

$$18.501 \times 14.1 = 18.501 \times 14 + 18.501 \times 0.1 = 259.014 + 1.8501 = 260.8641.$$

Hence

$$\begin{aligned}\widehat{\texttt{salary}}_1 &= 963.191 + 18.501 \times 14.1 \\ &= 963.191 + 260.8641 \\ &= 1224.0551.\end{aligned}$$

$$\boxed{\widehat{\texttt{salary}}_1 \approx 1224.06}$$

The residual is the actual salary minus the predicted salary:

$$\begin{aligned}\hat{u}_1 &= \texttt{salary}_1 - \widehat{\texttt{salary}}_1 \\ &= 1095 - 1224.0551 \\ &= -129.0551.\end{aligned}$$

$$\boxed{\hat{u}_1 \approx -129.06}$$

The OLS line over-predicts this CEO's pay by about \$129,060; $\hat{u}_1$ is the sample residual, not the unobservable population error $u_1 = \texttt{salary}_1 - (\beta_0 + \beta_1 \times 14.1)$.

Method 2: Residual first, from the decomposition $Y_i = \hat{Y}_i + \hat{u}_i$. By definition $\hat{u}_1 = Y_1 - \hat{\beta}_0 - \hat{\beta}_1 X_1$, so subtract the intercept from the actual salary first:

$$1095 - 963.191 = 131.809.$$

Then subtract the slope term $\hat{\beta}_1 X_1 = 18.501 \times 14.1 = 260.8641$:

$$\begin{aligned}\hat{u}_1 &= 131.809 - 260.8641 \\ &= -129.0551.\end{aligned}$$

Every observation splits as $Y_i = \hat{Y}_i + \hat{u}_i$, so the fitted value is recovered as

$$\begin{aligned}\widehat{\texttt{salary}}_1 &= \texttt{salary}_1 - \hat{u}_1 \\ &= 1095 - (-129.0551) \\ &= 1224.0551.\end{aligned}$$

$$\boxed{\widehat{\texttt{salary}}_1 \approx 1224.06} \qquad \boxed{\hat{u}_1 \approx -129.06}$$

3.2 True Values, OLS Estimates, and the Five Objects of Simple Regression

Every quantity in the simple regression model has a "true" (population, unobservable) version and a "hat" (OLS, sample, computable) analogue. Question 3.1 asks for exactly this glossary; the recurring exam trap is to blur the two columns.

Definition 3.1: The five objects of simple regression

For the model $Y_i = \beta_0 + \beta_1 X_i + u_i$, with OLS estimates $\hat{\beta}_0, \hat{\beta}_1$ from a sample of size n:

- **Population regression line**: $\beta_0 + \beta_1 X$, the (unknown) line describing the relationship that holds *on average* between X and Y across the whole population.
- **OLS regression line**: $\hat{\beta}_0 + \hat{\beta}_1 X$, the fitted line computed from one sample; it is the sample analogue (estimate) of the population regression line.
- **Error term**: $u_i = Y_i - (\beta_0 + \beta_1 X_i)$, the unobserved population deviation collecting every factor other than X_i that affects Y_i.
- **OLS residual**: $\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i$, the observable sample analogue of u_i – what is "left over" after fitting the line to this particular sample.
- **OLS predicted value**: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, the fitted value of Y_i read off the OLS line at $X = X_i$; the sample analogue of the (unobservable) systematic part $\beta_0 + \beta_1 X_i$ of Y_i.

Population (true, unknown)	**Sample (OLS, computable)**
Parameters β_0, β_1	Estimators $\hat{\beta}_0, \hat{\beta}_1$
Error term $u_i = Y_i - (\beta_0 + \beta_1 X_i)$	Residual $\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
Population regression line $\beta_0 + \beta_1 X$	OLS regression line $\hat{\beta}_0 + \hat{\beta}_1 X$
Systematic part $\beta_0 + \beta_1 X_i$ (no sample counterpart, unobservable)	OLS predicted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Reading the table. Every row on the right is something you can compute from data; every row on the left is a fixed but unknown feature of the population that the right-hand column estimates. This is precisely the population-versus-sample distinction that Chapter 1 introduced for $E(Y \mid X)$ in general, now specialised to the linear model.

Common Mistake: u_i is not $\hat{u}_i$

u_i is the true, unobservable population error; $\hat{u}_i$ is the observable OLS residual computed from one particular sample. They are numerically different objects: u_i depends on the unknown β_0, β_1, while $\hat{u}_i$ depends on the estimated $\hat{\beta}_0, \hat{\beta}_1$. In particular $\sum_{i=1}^n u_i$ need *not* equal zero (it is a sum of random population deviations), whereas $\sum_{i=1}^n \hat{u}_i = 0$ *always* holds by construction (Theorem 3.1 below).

3.3 The Residual Identities: Consequences of the FOCs

Chapter 2 showed that $\hat{\beta}_0, \hat{\beta}_1$ minimise $\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$, and that this minimisation has first-order conditions

$$-2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0, \tag{1}$$

$$-2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)X_i = 0. \tag{2}$$

Because $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ by definition, these two equations are really statements about $\hat{u}_i$. Three identities follow immediately, purely as algebraic consequences of $\hat{\beta}_0, \hat{\beta}_1$ being *the* least-squares solution – they hold in *every* sample, regardless of whether the Least Squares Assumptions (introduced below)

are true. This is the content of Question 3.2 and is one of the most exam-tested derivation chains in the course, since Chapter 4 reuses it to prove TSS = ESS + SSR and $R^2 \leq 1$.

Theorem 3.1: Sum of OLS residuals is zero

$$\sum_{i=1}^{n} \hat{u}_i = 0.$$

Proof. Divide FOC (1) by -2:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0.$$

By definition $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$, so the left-hand side is exactly $\sum_{i=1}^{n} \hat{u}_i$. Hence

$$\sum_{i=1}^{n} \hat{u}_i = 0.$$

□

Theorem 3.2: Residuals are orthogonal to the regressor

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0.$$

Proof. Divide FOC (2) by -2:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0.$$

The summand is $(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = \hat{u}_i X_i$, since $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$. Hence

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0.$$

□

Theorem 3.3: Residuals are orthogonal to the fitted values

$$\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0.$$

Proof. Write $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and expand:

$$\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = \sum_{i=1}^{n} \hat{u}_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i X_i.$$

The first term is $\hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i = \hat{\beta}_0 \cdot 0 = 0$ by Theorem 3.1; the second is $\hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i X_i = \hat{\beta}_1 \cdot 0 = 0$ by Theorem 3.2. Hence

$$\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0 + 0 = 0.$$

□

Corollary (sample covariance of $\hat{u}$ and X is zero). The sample covariance between the residuals and the regressor is

$$s_{\hat{u}X} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})(X_i - \bar{X}),$$

where $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$ by Theorem 3.1. Substituting $\bar{\hat{u}} = 0$,

$$s_{\hat{u}X} = \frac{1}{n-1}\sum_{i=1}^{n}\hat{u}_i(X_i - \bar{X}) = \frac{1}{n-1}\Big[\sum_{i=1}^{n}\hat{u}_i X_i - \bar{X}\sum_{i=1}^{n}\hat{u}_i\Big] = \frac{1}{n-1}\big[0 - \bar{X}\cdot 0\big] = 0,$$

using Theorem 3.2 for $\sum_{i=1}^{n}\hat{u}_i X_i = 0$ and Theorem 3.1 for $\sum_{i=1}^{n}\hat{u}_i = 0$. So the OLS residuals are, by construction, uncorrelated with X in *every* sample – a purely mechanical fact, not an assumption. (A fourth consequence, that the average fitted value equals $\bar{Y}$, i.e. $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$, follows the same way and is used to define R^2 in Chapter 4.)

Key Formula: the three residual identities

$$\sum_{i=1}^{n}\hat{u}_i = 0, \qquad \sum_{i=1}^{n}\hat{u}_i X_i = 0, \qquad \sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = 0,$$

all three following purely from the OLS first-order conditions – no assumption on the data-generating process is needed.

Exam Focus: the residual-identity proof chain

- Reproduce Theorems 3.1–3.3 from the two FOCs *from scratch*; this is Question 3.2 / Tutorial 3 Q2 verbatim, and Question/Tutorial 5 Q2 reuses exactly this chain to prove TSS = ESS + SSR and $R^2 \leq 1$ (Chapter 4).
- Keep the logic straight: $\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = 0$ is *derived from* the first two identities, not an independent FOC.
- These identities hold in every sample by construction; do not confuse them with the Least Squares *Assumptions* below, which are conditions on the population, not automatic algebra.

3.4 Key Concept: The Least Squares Assumptions

The residual identities above hold no matter what. But for $\hat{\beta}_0, \hat{\beta}_1$ to be *good estimators* of the true β_0, β_1 – for OLS to recover something meaningful about the population, not just fit whatever sample happened to be drawn – three assumptions on the data-generating process are needed.

Definition 3.2: The Least Squares Assumptions

For the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, $i = 1, \ldots, n$:

ssumption 1. The error term has conditional mean zero given X_i:

$$E[u_i \mid X_i] = 0$$

(zero conditional mean).

ssumption 2. (X_i, Y_i), $i = 1, \ldots, n$, are independently and identically distributed (i.i.d.).

ssumption 3. Large outliers are unlikely (formally, X and Y have finite fourth moments, stated informally in this course).

These may look abstract, but each has a direct interpretation and a direct failure mode: Assumption 1 is about whether X and the "other factors" u have been successfully separated; Assumption 2 is about how the sample was drawn; Assumption 3 is about whether a handful of extreme points can hijack the squared-error objective. Understanding exactly when each assumption does and does not hold is essential for knowing when $\hat{\beta}_0, \hat{\beta}_1$ give useful estimates of β_0, β_1 – this is the conceptual heart of the simple-regression unit, and Assumption 1 in particular is examined from many angles (diagrams, the RCT benchmark, return to education, selection bias) across this chapter and the next.

3.5 Assumption 1 in Depth: Zero Conditional Mean

Assumption 1 says: for *any* fixed value of X, the "other factors" collected in u average out to zero. It asserts that u is not systematically related to X – exactly the separation needed to read β_1 off the data as the causal effect of X, rather than as a mixture of the effect of X and whatever else is correlated with X.

Theorem 3.4: Zero conditional mean implies the conditional mean of Y is the population line

Under Assumption 1 ($E[u_i \mid X_i] = 0$),

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i.$$

Proof. Start from the model and take the conditional expectation given X_i of both sides:

$$E[Y_i \mid X_i] = E\big[\beta_0 + \beta_1 X_i + u_i \mid X_i\big].$$

Conditional expectation is linear, so it distributes across the sum:

$$E[Y_i \mid X_i] = E[\beta_0 \mid X_i] + E[\beta_1 X_i \mid X_i] + E[u_i \mid X_i].$$

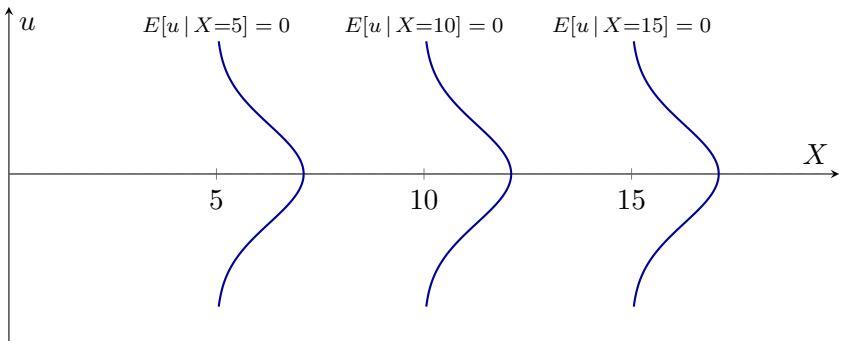
β_0 is a constant, so $E[\beta_0 \mid X_i] = \beta_0$; conditional on X_i, the term $\beta_1 X_i$ is itself a known constant, so $E[\beta_1 X_i \mid X_i] = \beta_1 X_i$; and $E[u_i \mid X_i] = 0$ by Assumption 1. Substituting,

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i + 0 = \beta_0 + \beta_1 X_i.$$

□

Reading Theorem 3.4. Given any value of X, the average value of Y lands exactly on the population regression line – there is no need to know u to predict the average of Y, since u cancels out on average. Equivalently: at a fixed X (say $X = 20$ students per teacher), individual districts scatter above and below the line because of district-specific factors ($u_i > 0$ pushes Y_i above, $u_i < 0$ pushes it below), but the average across all districts with that X sits exactly on the line.

Figure: what Assumption 1 looks like



At each fixed value of X (here $X = 5, 10, 15$), the density of u is a bump centred on zero: values of u near zero are most likely, and the density is symmetric enough that the conditional *average* is exactly zero at every one of these X values – compactly, $E[u_i \mid X_i] = 0$. (Redrawing u as Y instead, by shifting each bump up onto the population regression line $\beta_0 + \beta_1 X$, gives the companion picture for $E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$; this is the construction requested by Question 3.3(b) for Lecture 3 slide 11.)

Extension: zero conditional mean implies zero correlation (via the Law of Iterated Expectations)

The slides state, without requiring the proof in this course, that $E[u \mid X] = 0$ implies $\text{Cov}(X, u) = 0$ (equivalently $\text{Corr}(X, u) = 0$). Here is the argument, using the Law of Iterated Expectations (L.I.E.), $E[Z] = E\big[E[Z \mid X]\big]$ for any random variable Z.
Step 1: $E[u] = 0$**.** By the L.I.E. applied to u,

$$E[u] = E\big[E[u \mid X]\big] = E[0] = 0,$$

using Assumption 1, $E[u \mid X] = 0$, in the last-but-one step.
Step 2: $E[Xu] = 0$**.** By the L.I.E. applied to Xu,

$$E[Xu] = E\big[E[Xu \mid X]\big].$$

Conditional on X, the factor X is a known constant, so it can be pulled out of the inner conditional expectation:

$$E[Xu \mid X] = X\, E[u \mid X] = X \cdot 0 = 0.$$

Hence $E[Xu] = E[0] = 0$.
Step 3: $\text{Cov}(X, u) = 0$**.** By definition,

$$\text{Cov}(X, u) = E[Xu] - E[X]E[u].$$

Substituting $E[Xu] = 0$ (Step 2) and $E[u] = 0$ (Step 1),

$$\text{Cov}(X, u) = 0 - E[X] \cdot 0 = 0.$$

Step 4: $\text{Corr}(X, u) = 0$**.** Since $\text{Corr}(X, u) = \text{Cov}(X, u)/\big(\sqrt{\text{Var}(X)}\sqrt{\text{Var}(u)}\big)$ and the numerator is zero (Step 3), $\text{Corr}(X, u) = 0$ whenever $\text{Var}(X), \text{Var}(u) > 0$.

Common Mistake: zero conditional mean is stronger than zero correlation

The Extension above shows $E[u \mid X] = 0 \Rightarrow \text{Cov}(X, u) = 0$, but the converse is *false*: $\text{Cov}(X, u) = 0$ does *not* imply $E[u \mid X] = 0$. Zero conditional mean is a much stronger requirement – it must hold at *every* value of X, not just "on average" in the correlation sense.

Worked Example: a counterexample where $\text{Corr}(X, u) = 0$ but Assumption 1 fails

Let $X \sim N(0, 1)$ and define $u = X^2 - 1$.
Method 1: Direct moment computation. Recall the moments of $X \sim N(0, 1)$: $E[X] = 0$, $\text{Var}(X) = 1$, $E[X^3] = 0$, $E[X^4] = 3$. The odd moments vanish because the density $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is even, so $x\phi(x)$ and $x^3\phi(x)$ are odd functions whose integrals over $(-\infty, \infty)$ are zero.
Step 1: $E[u] = 0$. From $\text{Var}(X) = E[X^2] - (E[X])^2$,

$$E[X^2] = \text{Var}(X) + (E[X])^2 = 1 + 0^2 = 1.$$

By linearity of expectation,

$$E[u] = E[X^2 - 1] = E[X^2] - 1 = 1 - 1 = 0.$$

Step 2: $\mathrm{Cov}(X,u)=0$. Expand the product $Xu=X(X^2-1)=X^3-X$:

$$\begin{aligned}\mathrm{Cov}(X,u)&=E[Xu]-E[X]E[u]\\&=E[X^3-X]-0\cdot 0\\&=E[X^3]-E[X]\\&=0-0=0.\end{aligned}$$

Step 3: $\mathrm{Corr}(X,u)=0$. The correlation needs $\mathrm{Var}(u)>0$. Expanding $u^2=(X^2-1)^2=X^4-2X^2+1$,

$$\begin{aligned}\mathrm{Var}(u)&=E[u^2]-(E[u])^2\\&=E[X^4]-2E[X^2]+1-0^2\\&=3-2(1)+1=2>0.\end{aligned}$$

Hence

$$\mathrm{Corr}(X,u)=\frac{\mathrm{Cov}(X,u)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(u)}}=\frac{0}{\sqrt{1}\,\sqrt{2}}=0.$$

Step 4: $E[u\,|\,X]\neq 0$. Once $X=x$ is known, $u=x^2-1$ is a known number, so

$$E[u\,|\,X=x]=x^2-1.$$

Factor $x^2-1=(x-1)(x+1)$ and close each case:

- $|x|>1$: both factors have the same sign, so $E[u\,|\,X=x]>0$ (e.g. $x=2$ gives 3);
- $|x|<1$: the factors have opposite signs, so $E[u\,|\,X=x]<0$ (e.g. $x=0$ gives -1);
- $x=\pm 1$: one factor is zero, so $E[u\,|\,X=x]=0$.

Therefore Assumption 1 fails at every $x\neq\pm 1$, although X and u are uncorrelated. In a regression $Y=\beta_0+\beta_1X+u$ with this u, the population conditional mean is $E[Y\,|\,X]=\beta_0-1+\beta_1X+X^2$, which is not the population regression line.

$$\boxed{\mathrm{Corr}(X,u)=0\ \text{ yet }\ E[u\,|\,X]=X^2-1\neq 0\ (\text{except at } X=\pm 1)}$$

Method 2: Law of Iterated Expectations with odd/even symmetry. Once $X=x$ is known, $u=x^2-1$ is a known number, so the conditional mean function is $m(x)=E[u\,|\,X=x]=x^2-1$. Also $E[X^2]=\mathrm{Var}(X)+(E[X])^2=1+0=1$. By the Law of Iterated Expectations,

$$\begin{aligned}E[u]&=E\big[E[u\,|\,X]\big]=E[m(X)]=E[X^2]-1=1-1=0,\\E[Xu]&=E\big[E[Xu\,|\,X]\big]=E\big[X\,m(X)\big]=E\big[g(X)\big],\qquad g(x)=x(x^2-1).\end{aligned}$$

The function g is odd:

$$g(-x)=(-x)\big((-x)^2-1\big)=-x(x^2-1)=-g(x).$$

$X\sim N(0,1)$ is symmetric about 0, so X and $-X$ have the same distribution, and

$$E[g(X)]=E[g(-X)]=-E[g(X)]\implies 2E[g(X)]=0\implies E[Xu]=0.$$

The same symmetry gives $E[X]=0$. Hence

$$\mathrm{Cov}(X,u)=E[Xu]-E[X]E[u]=0-0\cdot 0=0,$$

and with $\mathrm{Var}(X)=1>0$ and $\mathrm{Var}(u)=E[X^4]-2E[X^2]+1-0^2=3-2+1=2>0$, $\mathrm{Corr}(X,u)=0/(\sqrt{1}\,\sqrt{2})=0$. The covariance vanishes because $m(x)$ is even (it treats x and $-x$ alike) while x is odd: a U-shaped conditional mean has no *linear* association with X, which is all that Corr measures. Assumption 1 instead requires $m(x)=0$ for every x, and $m(x)=x^2-1$ is zero only at $x=\pm 1$.

$$\boxed{\mathrm{Corr}(X,u)=0\ \text{ yet }\ E[u\,|\,X]=X^2-1\neq 0\ (\text{except at } X=\pm 1)}$$

Common Mistake: "$E[u] = 0$" is enough

It is not. $E[u] = 0$ (the *unconditional* mean of u is zero) is true essentially by convention – it can always be achieved by absorbing any nonzero constant into β_0. What Assumption 1 actually requires is $E[u \mid X] = 0$ for *every* value of X, which is a much stronger statement, as both the Extension above and Question 3.3(c) (worked out below) demonstrate: it is entirely possible to have $E[u] = 0$ while $E[u \mid X] \neq 0$.

3.6 The RCT Benchmark and the Average Treatment Effect

A useful benchmark for Assumption 1 is an ideal randomised controlled trial (RCT). When X (a treatment or intervention) is assigned by a computer's random number generator, X is mechanically unrelated to every pre-existing characteristic of the subject – education, ability, health, motivation – all of which sit inside u. Formally, random assignment makes X and u *fully independent*, and full independence is strong enough to give

$$\text{Full independence between } X \text{ and } u \implies E[u \mid X] = E[u] = 0,$$

because (i) independence implies $E[u \mid X] = E[u]$ (conditioning on an independent variable changes nothing), and (ii) $E[u] = 0$ always, by the convention that any nonzero constant mean of u is absorbed into β_0. So Assumption 1 typically holds by design in an RCT; with observational data, by contrast, whether $E[u \mid X] = 0$ holds requires careful, case-by-case justification, since X was not handed out at random.

Theorem 3.5: The slope on a randomly assigned binary regressor is the average treatment effect

Let $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $X_i \in \{0, 1\}$ is a treatment indicator ($X_i = 1$ if subject i is assigned to treatment, $X_i = 0$ if assigned to control) determined by a randomised experiment, so that Assumption 1 holds. Then

$$\beta_1 = E[Y_i \mid X_i = 1] - E[Y_i \mid X_i = 0],$$

the difference between the average treatment-group outcome and the average control-group outcome, i.e. the average treatment effect (ATE).

Proof. By Theorem 3.4 (which only used Assumption 1),

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i.$$

Evaluate this at the two possible values of the binary X_i. At $X_i = 1$:

$$E[Y_i \mid X_i = 1] = \beta_0 + \beta_1(1) = \beta_0 + \beta_1.$$

At $X_i = 0$:

$$E[Y_i \mid X_i = 0] = \beta_0 + \beta_1(0) = \beta_0.$$

Subtracting the second equation from the first,

$$E[Y_i \mid X_i = 1] - E[Y_i \mid X_i = 0] = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

□

Because random assignment makes the treatment and control groups statistically similar in every other respect, comparing their average outcomes isolates the causal effect of the treatment – exactly what Theorem 3.5 formalises. This is also the template used repeatedly below: whenever X is binary, β_0 is the conditional mean of Y in the $X = 0$ group, and β_1 is the *difference* in conditional means between the $X = 1$ and $X = 0$ groups.

3.7 Worked Example (Tutorial 3): Wages of SSS versus SOH Graduates

Worked Example (Tutorial 3, Question 4): SSS vs. SOH wages

Question. Is there a difference in average wages between graduates of the School of Social Sciences (SSS) and the School of Humanities (SOH)?

(a) The model.

$$\text{wage}_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1, \ldots, n,$$

with wage_i the dependent variable; $X_i = 1$ if i graduated from SSS, $X_i = 0$ if from SOH; u_i the other factors affecting wage (ability, experience, family connections).

(b) Assumption 1 here. X_i takes only the values 0 and 1, so "$E[u_i \mid X_i] = 0$ for every value of X_i" is two equations:

$$E[u_i \mid X_i = 1] = 0 \quad \text{and} \quad E[u_i \mid X_i = 0] = 0.$$

Since $E[u_i] = 0$ by the convention that any constant is absorbed into β_0, this is the same as

$$E[u_i \mid \text{SSS}] = E[u_i \mid \text{SOH}] = E[u_i] = 0.$$

- In words: the population average of everything else that drives wages (ability, experience, family connections) is the same for SSS graduates as for SOH graduates.
- It is plausible if admission to SSS versus SOH is effectively random with respect to ability and earnings potential, as in the RCT benchmark above.
- It fails if one programme's intake is systematically higher-ability. For example, if SSS graduates have $E[u_i \mid \text{SSS}] = +200$ and SOH graduates $E[u_i \mid \text{SOH}] = -200$ (dollars per month), then school type and ability are dependent and a \$400 part of the SSS–SOH wage gap is ability, not school.
- Assumption 1 is a statement about the population; a sample can never verify it, because u_i is unobserved and the residuals $\hat{u}_i$ average to zero in each group by construction.

(c) β_0, β_1 as conditional means.

Method 1: Conditional expectation of the model at $X_i = 0$ and $X_i = 1$. Take the conditional expectation of the model given X_i and use linearity:

$$\begin{aligned} E[\text{wage}_i \mid X_i] &= E[\beta_0 \mid X_i] + E[\beta_1 X_i \mid X_i] + E[u_i \mid X_i] \\ &= \beta_0 + \beta_1 X_i + 0 && \text{(Assumption 1)} \\ &= \beta_0 + \beta_1 X_i, \end{aligned}$$

which is Theorem 3.4. Evaluate at the two values of X_i:

$$\begin{aligned} E[\text{wage}_i \mid X_i = 1] &= \beta_0 + \beta_1(1) = \beta_0 + \beta_1 && \text{(SSS)} \\ E[\text{wage}_i \mid X_i = 0] &= \beta_0 + \beta_1(0) = \beta_0 && \text{(SOH)} \end{aligned}$$

The second line gives β_0 directly. Subtract the second line from the first:

$$E[\text{wage}_i \mid X_i = 1] - E[\text{wage}_i \mid X_i = 0] = (\beta_0 + \beta_1) - \beta_0 = \beta_1.$$

For a numerical illustration, if the population average monthly wage is \$3,500 for SOH graduates and \$4,000 for SSS graduates, then $\beta_0 = 3500$ and $\beta_1 = 4000 - 3500 = 500$ dollars per month. The sample analogues are $\hat{\beta}_0 =$ the SOH sample mean wage and $\hat{\beta}_1 =$ the SSS sample mean minus the SOH sample mean.

$$\boxed{\beta_0 = \text{average wage of SOH graduates}, \qquad \beta_1 = E[\text{wage} \mid \text{SSS}] - E[\text{wage} \mid \text{SOH}]}$$

Method 2: Population slope formula $\beta_1 = \text{Cov}(X, Y)/\text{Var}(X)$ for a binary regressor. Write $Y_i = \text{wage}_i$, $p = P(X_i = 1)$ with $0 < p < 1$, $\mu_1 = E[Y_i \mid X_i = 1]$ and $\mu_0 = E[Y_i \mid X_i = 0]$.

Assumption 1 gives $E[u_i] = 0$ and $\text{Cov}(X_i, u_i) = 0$ (Extension above). Take the covariance of both sides of the model with X_i:

$$\begin{aligned}\text{Cov}(X_i, Y_i) &= \text{Cov}(X_i, \beta_0 + \beta_1 X_i + u_i) \\ &= 0 + \beta_1 \,\text{Var}(X_i) + \text{Cov}(X_i, u_i) \\ &= \beta_1 \,\text{Var}(X_i),\end{aligned}$$

so $\beta_1 = \text{Cov}(X_i, Y_i)/\,\text{Var}(X_i)$. Now compute both moments for a binary X_i. Since $X_i^2 = X_i$,

$$\text{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2 = p(1-p).$$

By the Law of Total Expectation, $E[Y_i] = p\mu_1 + (1-p)\mu_0$, and $X_i Y_i$ equals Y_i when $X_i = 1$ and 0 otherwise, so $E[X_i Y_i] = p\mu_1$. Hence

$$\begin{aligned}\text{Cov}(X_i, Y_i) &= E[X_i Y_i] - E[X_i]E[Y_i] \\ &= p\mu_1 - p\big[p\mu_1 + (1-p)\mu_0\big] \\ &= (p - p^2)\mu_1 - p(1-p)\mu_0 \\ &= p(1-p)\mu_1 - p(1-p)\mu_0 \\ &= p(1-p)(\mu_1 - \mu_0).\end{aligned}$$

Therefore

$$\beta_1 = \frac{p(1-p)(\mu_1 - \mu_0)}{p(1-p)} = \mu_1 - \mu_0.$$

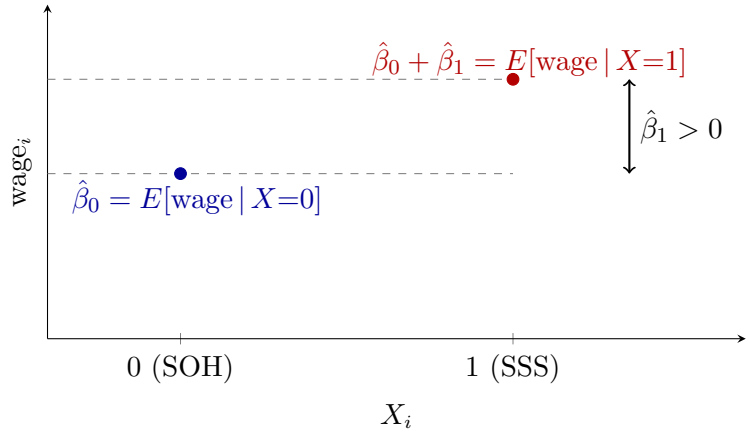
Taking expectations of the model, $E[Y_i] = \beta_0 + \beta_1 E[X_i]$, so

$$\begin{aligned}\beta_0 &= E[Y_i] - \beta_1 E[X_i] \\ &= p\mu_1 + (1-p)\mu_0 - p(\mu_1 - \mu_0) \\ &= p\mu_1 + \mu_0 - p\mu_0 - p\mu_1 + p\mu_0 \\ &= \mu_0.\end{aligned}$$

With the illustrative numbers $\mu_0 = 3500$, $\mu_1 = 4000$ and any $p \in (0,1)$, this gives $\beta_1 = 500$ and $\beta_0 = 3500$ dollars per month, the same as Method 1. The sample version of this route is the OLS formula $\hat{\beta}_1 = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^n (X_i - \bar{X})^2$, which for a dummy X_i reduces to the difference in sample group means.

$$\boxed{\beta_0 = \text{average wage of SOH graduates}, \qquad \beta_1 = E[\text{wage} \,|\, \text{SSS}] - E[\text{wage} \,|\, \text{SOH}]}$$

(d) Diagram. Below, assuming SSS graduates earn more on average ($\beta_1 > 0$). The two dots sit at heights $E[\text{wage} \,|\, X{=}0]$ (SOH) and $E[\text{wage} \,|\, X{=}1]$ (SSS); the vertical gap between them is the coefficient on X_i.



Common Mistake: reading a binary-regressor $\hat{\beta}_1$ as a slope

When $X_i \in \{0, 1\}$, it is tempting to keep calling $\hat{\beta}_1$ "the slope," as if a one-unit increase in X were a meaningful marginal change. It is not: there are only two points on the X-axis. Here $\hat{\beta}_1$ is a *difference in group means* (SSS mean wage minus SOH mean wage), not a rate of change along a continuum. The same reading applies to any dummy regressor (region indicators, treatment indicators, etc.).

QRS@NTU — qrs.ntu.edu.sg

3.8 A Counterexample from Question 3(c): $E[u] = 0$ Does Not Imply $E[u \mid X] = 0$

Worked Example (Question 3, Q3c): constructing $E[u] = 0$ with $E[u|X] \neq 0$

To build the simplest possible counterexample, let X take only two values, $X \in \{0, 1\}$, each with probability $\frac{1}{2}$, and suppose

$$E[u \mid X = 0] = -1, \qquad E[u \mid X = 1] = +1.$$

Method 1: Law of Total Expectation over the two values of X. Assumption 1 requires $E[u \mid X = x] = 0$ at every value x the regressor takes. Here

$$E[u \mid X = 0] = -1 \neq 0 \quad \text{and} \quad E[u \mid X = 1] = +1 \neq 0,$$

so Assumption 1 fails at both values of X. The unconditional mean averages the conditional means, weighted by the probabilities of the two values. By the Law of Total Expectation,

$$\begin{aligned} E[u] &= P(X{=}0)\, E[u \mid X{=}0] + P(X{=}1)\, E[u \mid X{=}1] \\ &= \tfrac{1}{2}(-1) + \tfrac{1}{2}(+1) \\ &= -\tfrac{1}{2} + \tfrac{1}{2} = 0. \end{aligned}$$

In economic terms, with $X = 1$ for SSS graduates, $X = 0$ for SOH graduates and u containing ability, SSS graduates would be one unit above average ability and SOH graduates one unit below, yet ability averages to zero over the whole population. A regression of wage on X would then attribute the ability gap of $(+1) - (-1) = 2$ units to school type.

$$\boxed{E[u] = 0 \;\text{ while }\; E[u \mid X{=}0] = -1 < 0 \;\text{ and }\; E[u \mid X{=}1] = +1 > 0}$$

Method 2: Explicit functional form $u = (2X - 1) + v$ with v independent of X. Build u from X directly. Let $v \sim N(0, 1)$ be independent of X and set

$$u = (2X - 1) + v.$$

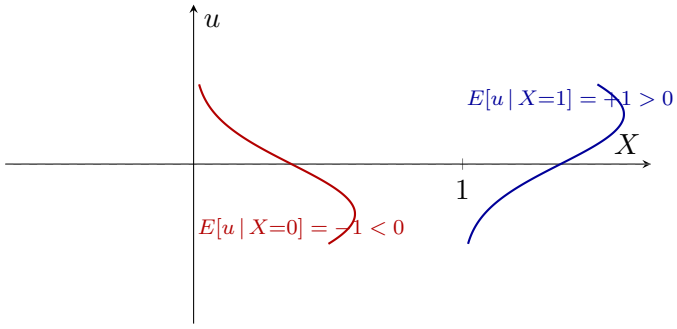
At $X = 0$ the systematic part is $2(0) - 1 = -1$ and at $X = 1$ it is $2(1) - 1 = +1$, so u is a bell-shaped variable centred at -1 or at $+1$, which is the figure below. Since $E[X] = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$ and $E[v] = 0$, linearity gives

$$\begin{aligned} E[u] &= 2E[X] - 1 + E[v] \\ &= 2 \cdot \tfrac{1}{2} - 1 + 0 = 0. \end{aligned}$$

Conditional on $X = x$, the term $2x - 1$ is a known constant, and $E[v \mid X] = E[v] = 0$ because v is independent of X. Hence

$$\begin{aligned} E[u \mid X = 0] &= 2(0) - 1 + E[v \mid X = 0] = -1 + 0 = -1 < 0, \\ E[u \mid X = 1] &= 2(1) - 1 + E[v \mid X = 1] = +1 + 0 = +1 > 0. \end{aligned}$$

$$\boxed{E[u] = 0 \;\text{ while }\; E[u \mid X{=}0] = -1 < 0 \;\text{ and }\; E[u \mid X{=}1] = +1 > 0}$$



The density of u is centred at -1 when $X = 0$ and at $+1$ when $X = 1$: at neither value of X is the conditional mean zero, even though averaging over both groups (equally weighted) gives exactly zero.

3.9 Return to Education: A Case Where Assumption 1 May Fail

With experimental data, Assumption 1 is close to automatic. With *observational* data, it must be argued for, and it can fail. The canonical example is the "return to education":

$$\text{wage} = \beta_0 + \beta_1 \,\text{educ} + u,$$

where β_1 is meant to capture the causal effect of one more year of schooling on wages. Suppose u is dominated by innate ability, $u = \text{abil}$ (plus other factors). Assumption 1 then requires

$$E[\text{abil} \,|\, \text{educ} = 8] = E[\text{abil} \,|\, \text{educ} = 9] = \cdots = E[\text{abil} \,|\, \text{educ} = 16] = \cdots,$$

i.e. average ability must be the same at *every* level of education. If, instead, people with higher innate ability tend to acquire more education – because ability makes further study easier, because higher-ability students are more likely to be admitted to competitive programmes, or because higher ability leads people to expect a bigger payoff from an extra degree – then

$$E[\text{abil} \,|\, \text{educ} = 10] < E[\text{abil} \,|\, \text{educ} = 16],$$

and Assumption 1 fails: `educ` and `abil` are dependent, so u is not "separated" from X. This self-sorting of higher-ability people into more schooling is exactly the **selection bias** (also called **survivorship bias** in this context) introduced in Lecture 3.

Direction of the bias, intuitively. Since more-able people both acquire more education *and* tend to earn more for reasons unrelated to schooling, the raw OLS slope $\hat{\beta}_1$ picks up part of ability's effect on wages along with education's own effect – $\hat{\beta}_1$ tends to *overstate* the true causal return to education, because the "return to education" coefficient is contaminated by an "effect of being the kind of person who gets more education." (The exact algebraic decomposition of this bias, via the omitted-variable-bias formula, is developed once multiple regression is introduced later in the course; here the point is qualitative: know that the bias exists, and in which direction it plausibly runs.)

The same logic reappears verbatim whenever a regressor is partly a *choice*: does the model $\text{TestScore} = \beta_0 + \beta_1 \text{STR} + u$ satisfy Assumption 1? School districts are not randomly assigned a student-teacher ratio – wealthier districts can afford smaller classes *and* may have other advantages (better facilities, more parental involvement) that also raise test scores, so STR and unobserved district quality may well be correlated, exactly the same threat to Assumption 1 as in the education example.

3.10 Policy Application: Survivorship Bias and Wald's Bombers

Policy Application: Abraham Wald and the returning bombers

During World War II, the statistician Abraham Wald was asked by the US Navy how to armour bombers most effectively against enemy fire. Engineers had mapped where returning aircraft showed the most bullet damage and wanted to reinforce exactly those areas. Wald reasoned the opposite: the data on hand – damage patterns on *returning* planes – was not a random sample of all planes hit by enemy fire. Planes hit in the areas showing heavy damage on the survivors were evidently able to sustain damage there and still make it home; planes hit in the areas that appear *undamaged* on the survivors were most likely the ones that never returned at all, because a hit there was fatal. Wald therefore recommended reinforcing the areas with the *least* observed damage.

Econometric reading. The observed sample (returning bombers) is not drawn at random from the population of interest (all bombers that took enemy fire); it is selected on an outcome (survival) that is itself correlated with the variable being studied (location of damage). Any regression of "damage location" on "outcome" run only on the survivors would badly mis-estimate the true relationship – exactly the selection-bias mechanism from the return-to-education example, here in its starkest form: the missing observations are not missing at random.

3.11 Assumption 2: I.I.D. Sampling

Assumption 2 – that (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. – is satisfied automatically by **simple random sampling**: each unit of the population is equally likely to be drawn (so the (X_i, Y_i) are *independent* of one another), and every unit is drawn from the same population (so the (X_i, Y_i) are *identically distributed*).

Example. Let X be a worker's age and Y his or her earnings. If a sample of n workers is drawn purely at random from the population of workers in Singapore, then (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. draws from the same (Singapore) age-earnings distribution. If, instead, some workers are drawn from the Singapore population and others from the US population, the sample mixes two different age-earnings distributions – the observations are no longer *identically* distributed, and Assumption 2 fails. (Chapter 4 develops this mechanism further, including the case of non-random sampling and time-series data, where consecutive observations are not independent.)

Exam Focus: two distinct ways Assumption 2 can fail

- **Not identically distributed**: pooling observations from different populations (e.g. SG and US earners, as above), or from the same population at very different points in time.
- **Not independent**: sampling in a way that makes one observation's inclusion depend on another's – e.g. surveying whole households (siblings' outcomes are related) or time-series data (this quarter's revenue is correlated with last quarter's). Chapter 4 discusses this second failure mode, and the general mechanism of simple random sampling, in depth.

3.12 Assumption 3: Large Outliers Are Unlikely

Assumption 3 rules out extreme values of X or Y being too likely, because the OLS objective function is a *sum of squares*: a single very large residual is squared, so it contributes disproportionately to $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$ and can pull the fitted line a long way from where it would sit without that point.

Worked Example (Lecture 3): one outlier can dominate the OLS line

Consider five hypothetical observations. The first four lie exactly on a line of slope 1:

$$(X_i, Y_i) \in \{(1,1), (2,2), (3,3), (4,4)\}.$$

Fitting OLS to just these four points gives a perfect fit, $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = 1$ (every residual is zero). Now add a fifth, outlying point, $(X_5, Y_5) = (5, 20)$, so $n = 5$.

Method 1: Closed-form OLS formulas in deviations from the means. *Four points.* The means are

$$\bar{X} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5, \qquad \bar{Y} = \frac{1+2+3+4}{4} = 2.5.$$

The deviations are $X_i - \bar{X} = Y_i - \bar{Y} = -1.5, -0.5, 0.5, 1.5$, so

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 2.25 + 0.25 + 0.25 + 2.25 = 5,$$
$$\sum (X_i - \bar{X})^2 = 2.25 + 0.25 + 0.25 + 2.25 = 5,$$
$$\hat{\beta}_1 = \frac{5}{5} = 1, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2.5 - 1(2.5) = 0.$$

Five points. The means are

$$\bar{X} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3, \qquad \bar{Y} = \frac{1+2+3+4+20}{5} = \frac{30}{5} = 6.$$

The deviations are $X_i - \bar{X} = -2, -1, 0, 1, 2$ and $Y_i - \bar{Y} = -5, -4, -3, -2, 14$. Hence

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = (-2)(-5) + (-1)(-4) + (0)(-3) + (1)(-2) + (2)(14)$$
$$= 10 + 4 + 0 - 2 + 28 = 40,$$
$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 4 + 1 + 0 + 1 + 4 = 10,$$
$$\hat{\beta}_1 = \frac{40}{10} = 4,$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 6 - 4(3) = -6.$$

The outlier alone contributes 28 of the 40 in the numerator, so one observation out of five quadruples the estimated slope.

Adding one outlier moves the fitted line from $\hat{Y} = X$ to $\hat{Y} = -6 + 4X$

Method 2: Solving the two first-order conditions (normal equations). Dividing the FOCs (1)–(2) by -2 and expanding the sums gives

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i,$$
$$\sum X_i Y_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2.$$

Four points. $\sum X_i = 10$, $\sum Y_i = 10$, $\sum X_i Y_i = 1 + 4 + 9 + 16 = 30$, $\sum X_i^2 = 30$:

$$10 = 4\hat{\beta}_0 + 10\hat{\beta}_1 \implies \hat{\beta}_0 = 2.5 - 2.5\hat{\beta}_1,$$
$$30 = 10(2.5 - 2.5\hat{\beta}_1) + 30\hat{\beta}_1 = 25 + 5\hat{\beta}_1 \implies \hat{\beta}_1 = 1,\ \hat{\beta}_0 = 2.5 - 2.5 = 0.$$

Five points. $\sum X_i = 15$, $\sum Y_i = 30$, $\sum X_i Y_i = 1+4+9+16+100 = 130$, $\sum X_i^2 = 1+4+9+16+25 = 55$:

$$30 = 5\hat{\beta}_0 + 15\hat{\beta}_1 \implies \hat{\beta}_0 = 6 - 3\hat{\beta}_1,$$
$$130 = 15(6 - 3\hat{\beta}_1) + 55\hat{\beta}_1 = 90 - 45\hat{\beta}_1 + 55\hat{\beta}_1 = 90 + 10\hat{\beta}_1 \implies \hat{\beta}_1 = 4,$$
$$\hat{\beta}_0 = 6 - 3(4) = -6.$$

Adding one outlier moves the fitted line from $\hat{Y} = X$ to $\hat{Y} = -6 + 4X$

Method 3: Linearity of OLS in Y (old fit plus the outlier's deviation). Write every Y_i as the old line plus a deviation, $Y_i = X_i + d_i$, with $d = (0, 0, 0, 0, 15)$ since $20 - 5 = 15$. For the five points, $\bar{X} = 15/5 = 3$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 4 + 1 + 0 + 1 + 4 = 10$. The OLS slope $\sum_{i=1}^{n}(X_i - \bar{X})Y_i / \sum_{i=1}^{n}(X_i - \bar{X})^2$ is linear in the Y_i, so it splits into the slope from regressing X_i on X_i (which is 1) plus the slope from regressing d_i on X_i:

$$\hat{\beta}_1 = 1 + \frac{\sum_{i=1}^{n}(X_i - \bar{X})d_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 1 + \frac{(5-3)(15)}{10} = 1 + 3 = 4.$$

The intercept splits the same way, with $\bar{d} = 15/5 = 3$:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = (\bar{X} + \bar{d}) - (1+3)\bar{X} = \bar{d} - 3\bar{X} = 3 - 9 = -6.$$

The added slope 3 is the outlier's vertical distance 15 times its horizontal distance from $\bar{X}$, 2, divided by $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 10$: a point far from the line and far from $\bar{X}$ has the largest pull. The new residuals $\hat{u}_i = Y_i - (-6 + 4X_i)$ are $3, 0, -3, -6, 6$ (they sum to 0), so three of the four points that lay exactly on $Y = X$ now carry nonzero residuals; only $(2, 2)$ still lies on the new line.

Adding one outlier moves the fitted line from $\hat{Y} = X$ to $\hat{Y} = -6 + 4X$

Extension: why squared loss, and not absolute loss (non-examinable)

OLS minimises $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ rather than, say, $\sum_{i=1}^{n}|Y_i - b_0 - b_1X_i|$. Squaring is what makes the objective differentiable everywhere, which is exactly what lets the FOCs used throughout this chapter be solved in closed form for $\hat{\beta}_0, \hat{\beta}_1$ – absolute-value loss has no such closed form and must be solved numerically (this is "least absolute deviations" regression). The same squaring is precisely why Assumption 3 is needed: absolute-value loss barely notices an outlier, while squared loss notices it enormously, as the worked example above shows. This also explains why the fix for an influential outlier is never simply to delete it and rerun OLS – a genuine outlier may be the most informative observation in the sample, and silently dropping inconvenient points is itself a form of selection bias.

3.13 Why All Three Assumptions Matter

The three Least Squares Assumptions are not three unrelated technical conditions; each protects a different property of $\hat{\beta}_0, \hat{\beta}_1$ that later chapters build on. Assumption 1 (zero conditional mean) is what makes $\hat{\beta}_0, \hat{\beta}_1$ **unbiased** for β_0, β_1 – without it, $\hat{\beta}_1$ converges to some mixture of the true causal slope and the correlation between X and u, not to β_1 itself. Assumption 2 (i.i.d. sampling) is what makes the variance calculations behind $\text{SE}(\hat{\beta}_1)$ and the large-sample normal approximation valid – Chapter 4's derivation of $\text{Var}(\bar{Y}) = \sigma^2/n$, and its extension to $\text{Var}(\hat{\beta}_1)$, relies on independence across observations exactly as this chapter's residual identities relied on the FOCs. Assumption 3 (no large outliers) is what keeps the sampling distribution of $\hat{\beta}_0, \hat{\beta}_1$ well-behaved enough for the large-sample normal approximation in Chapter 6 to kick in. Chapter 5 states this formally: under Assumptions 1–3 (plus a technical condition), $\hat{\beta}_0, \hat{\beta}_1$ are unbiased, are the most efficient linear unbiased estimators (Gauss–Markov), and are approximately normally distributed in large samples – three separate guarantees, one from each assumption.

3.14 Chapter Summary

Key Formula: Chapter 3 recap

$$\sum_{i=1}^{n} \hat{u}_i = 0, \qquad \sum_{i=1}^{n} \hat{u}_i X_i = 0, \qquad \sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0 \qquad \text{(residual identities, always true)}$$

$$E[u_i \mid X_i] = 0 \quad \text{(Assumption 1)}$$
$$(X_i, Y_i) \text{ i.i.d.} \quad \text{(Assumption 2)}$$
$$\text{large outliers unlikely} \quad \text{(Assumption 3)}$$

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i \qquad \text{(Assumption 1)}$$
$$\beta_1 = E[Y_i \mid X_i{=}1] - E[Y_i \mid X_i{=}0] \qquad \text{(binary } X\text{)}$$

Exam Focus: Chapter 3 takeaways

- Be able to state, from memory, the five objects (population/OLS line, error/residual, predicted value) and re-derive $\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = 0$, $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$ from the two FOCs without notes – this chain is reused in Chapter 4 to prove $R^2 \leq 1$.
- State the three Least Squares Assumptions *by number* (never "SLR.1"–"SLR.5") and know which one each scenario is testing: Assumption 1 for causal/selection questions, Assumption 2 for sampling-design questions, Assumption 3 for outlier questions.
- For any binary regressor, translate β_0, β_1 into conditional means immediately: $\beta_0 = E[Y \mid X{=}0]$, $\beta_1 = E[Y \mid X{=}1] - E[Y \mid X{=}0]$ – and say explicitly that β_1 is a difference in means, not a slope.
- Be ready to construct a diagram or a two-point numerical example showing $E[u] = 0$ with $E[u \mid X] \neq 0$ (Question 3.3c) – this is a favourite way of testing whether Assumption 1 is genuinely understood.

4 Lecture 4: Sampling, Measures of Fit and the Sampling Distribution

This chapter answers three questions: (1) *when does a sample that looks perfectly ordinary still mislead us about the population regression line?*; (2) *once we have fitted a regression line, how much of the variation in Y does it actually explain, and how do we measure that?*; (3) *why do we treat $\hat{\beta}_0$ and $\hat{\beta}_1$ themselves as random variables, and what does their distribution look like when the sample is large?* It builds directly on Chapter 3, which stated the Least Squares Assumptions and proved the residual identities $\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = 0$ used repeatedly below; it prepares Chapter 5 (unbiasedness, efficiency, Gauss–Markov) and Chapter 6 (inference), both of which take the properties previewed here and prove or operationalise them in full.

4.1 Selection Bias Revisited: When the Sample Misleads

Chapter 3 introduced Least Squares Assumption 1, $E[u_i \,|\, X_i] = 0$, and showed that it fails whenever units select themselves into the sample (or into a value of X) on the basis of something correlated with u. This section collects further, sharper illustrations of that failure, because the exam draws on this material from many angles: diagrams, short-answer prompts, and applied regression stories.

Definition 4.1: Selection bias

Selection bias arises when the process that determines whether a unit appears in the observed sample – or takes a particular value of X – is itself correlated with the error term u. Formally, if inclusion in the sample (or the value of X_i) depends on u_i, then $E[u_i \,|\, X_i] \neq 0$ in the observed sample even if it would hold in the full population, and OLS applied to the selected sample no longer estimates the causal parameter of interest.

Worked Example (Lecture 4): the hospital survey

A US national hospital interview survey reports:

Group	Sample size	Mean health status	Std. error
Hospital	7,774	3.21	0.014
Non-hospital	90,049	3.93	0.003

Here "hospital" means the respondent visited a hospital during the survey period, and health status is self-reported on a scale where higher is healthier. The naive reading of the table – hospital visits make people sicker – treats the group difference $3.93 - 3.21 = 0.72$ as if it were the causal effect of hospitalisation on health.

Method 1: Least Squares Assumption 1 diagnosis via $\mathrm{Cov}(\mathbf{hospital}_i, u_i)$

Write the regression as

$$\mathrm{health}_i = \beta_0 + \beta_1 \mathrm{hospital}_i + u_i, \qquad \mathrm{hospital}_i \in \{0, 1\},$$

where β_1 is the causal effect of a hospital visit on health status (in scale points) and u_i collects every other determinant of health, for example pre-existing conditions. Recall Least Squares Assumption 1, $E[u_i \,|\, \mathrm{hospital}_i] = 0$. By the law of iterated expectations it implies

$$\begin{aligned} E[u_i] &= E\big[E[u_i \,|\, \mathrm{hospital}_i]\big] = E[0] = 0, \\ \mathrm{Cov}(\mathrm{hospital}_i, u_i) &= E[\mathrm{hospital}_i\, u_i] - E[\mathrm{hospital}_i]\, E[u_i] \\ &= E\big[\mathrm{hospital}_i\, E[u_i \,|\, \mathrm{hospital}_i]\big] - E[\mathrm{hospital}_i] \cdot 0 \\ &= E[\mathrm{hospital}_i \cdot 0] = 0. \end{aligned}$$

Hence, by the contrapositive, $\mathrm{Cov}(\mathrm{hospital}_i, u_i) \neq 0$ implies $E[u_i \,|\, \mathrm{hospital}_i] \neq 0$. People in poor underlying health (low u_i) are the ones who visit a hospital, so $\mathrm{hospital}_i = 1$ occurs together with low u_i:

$$\mathrm{Cov}(\mathrm{hospital}_i, u_i) < 0 \quad \Longrightarrow \quad E[u_i \,|\, \mathrm{hospital}_i] \neq 0.$$

Therefore Least Squares Assumption 1 fails. For the precision of the gap, the two group means come from independent samples, so their variances add:

$$\begin{aligned}\mathrm{SE}(\bar{Y}_{\text{non}} - \bar{Y}_{\text{hosp}}) &= \sqrt{0.003^2 + 0.014^2} = \sqrt{0.000009 + 0.000196} \\ &= \sqrt{0.000205} = 0.0143,\end{aligned}$$

and the gap is $0.72/0.0143 = 50.3$ standard errors from zero.

- The 0.72-point gap is a sample difference in mean health status between two self-selected groups; the causal effect β_1 is a population parameter, and the two coincide only if Least Squares Assumption 1 holds.
- Hospital visitors are, on average, sicker before they arrive, so the hospital group would report lower health status even if a visit had no effect at all.
- The gap is estimated very precisely (standard error 0.0143 points, 50.3 standard errors from zero), but a standard error measures sampling noise only; it does not remove the selection component, and a larger survey would not either.
- The table therefore cannot tell how much of the 0.72 points is the effect of hospitalisation and how much is the pre-existing difference in health between the groups.

The 0.72-point gap is not a valid estimate of the causal effect of hospitalisation; Assumption 1 fails by self-selection.

Method 2: Conditional-mean decomposition of the group gap
Take the conditional expectation of the model within each group:

$$\begin{aligned}E[\text{health}_i \,|\, \text{hospital}_i = 1] &= \beta_0 + \beta_1 + E[u_i \,|\, \text{hospital}_i = 1], \\ E[\text{health}_i \,|\, \text{hospital}_i = 0] &= \beta_0 + E[u_i \,|\, \text{hospital}_i = 0].\end{aligned}$$

Subtract the first line from the second:

$$\begin{aligned}&E[\text{health}_i \,|\, \text{hospital}_i = 0] - E[\text{health}_i \,|\, \text{hospital}_i = 1] \\ &\quad = -\beta_1 + \underbrace{\big(E[u_i \,|\, \text{hospital}_i = 0] - E[u_i \,|\, \text{hospital}_i = 1]\big)}_{\text{selection term } S}.\end{aligned}$$

The group means 3.93 and 3.21 estimate the two conditional expectations, so

$$0.72 \approx -\beta_1 + S \qquad \Longrightarrow \qquad -\beta_1 \approx 0.72 - S.$$

Sicker people select into the hospital group, so $E[u_i \,|\, \text{hospital}_i = 1] < E[u_i \,|\, \text{hospital}_i = 0]$ and $S > 0$. Hence the causal harm $-\beta_1$ is smaller than 0.72 points, and the gap measures $-\beta_1$ only when $S = 0$, which holds under Least Squares Assumption 1 because both conditional means of u_i are then zero. For a hypothetical numerical illustration, if the selection term were $S = 0.90$ points, then $-\beta_1 = 0.72 - 0.90 = -0.18$, i.e. $\beta_1 = 0.18$: a hospital visit would *raise* health status by 0.18 points, although the raw table shows the hospital group 0.72 points less healthy.

The 0.72-point gap is not a valid estimate of the causal effect of hospitalisation; Assumption 1 fails by self-selection.

The same mechanism recurs across applied econometrics, and each context is worth being able to name on the exam:

- **Wage regressions.** Wages are observed only for people who hold a wage-paying job, so a regression of wages on, say, education implicitly conditions on labour-force participation – itself a choice correlated with unobserved earnings potential.

- **Programme and policy evaluation.** When individuals *volunteer* for a training programme rather than being randomly assigned, those who volunteer may have unobserved ability or motivation correlated with the outcome, biasing the estimated programme effect.
- **Survey non-response.** Even a randomly distributed survey suffers selection bias if the decision to respond is non-random – response rates are routinely well below 100%, and non-responders are rarely a random subset.
- **Ratings.** Product and service ratings are given disproportionately by people with unusually strong (often negative) experiences, so the observed rating distribution is not representative of all users.

Worked Example (Tutorial 4, Q1a): survivorship in investment returns

Consider $\text{gain}_i = \beta_0 + \beta_1 \text{experience}_i + u_i$, where gain_i is investor i's realised gain in the financial market, experience_i is years of investing experience, and u_i includes unobserved investor ability. Investors with low ability tend to lose money early and exit the market, so only able investors accumulate many years of experience.

Method 1: Least Squares Assumption 1 diagnosis and the slope decomposition

Write $X_i = \text{experience}_i$ (years) and $Y_i = \text{gain}_i$ (percent per year). Recall Least Squares Assumption 1, $E[u_i \mid X_i] = 0$. By the law of iterated expectations,

$$\begin{aligned} E[u_i] &= E\big[E[u_i \mid X_i]\big] = E[0] = 0, \\ \operatorname{Cov}(X_i, u_i) &= E[X_i u_i] - E[X_i]\,E[u_i] \\ &= E\big[X_i\,E[u_i \mid X_i]\big] - E[X_i]\cdot 0 = E[X_i \cdot 0] = 0. \end{aligned}$$

Hence, by the contrapositive, $\operatorname{Cov}(X_i, u_i) \neq 0$ implies $E[u_i \mid X_i] \neq 0$. Low-ability investors exit early (survivorship), so among the investors observed, a large X_i occurs together with a large u_i:

$$\operatorname{Cov}(X_i, u_i) > 0 \quad \Longrightarrow \quad E[u_i \mid X_i] \neq 0,$$

and Least Squares Assumption 1 fails. For the direction of the bias, average the model over i,

$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}, \qquad \bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i,$$

and subtract it from $Y_i = \beta_0 + \beta_1 X_i + u_i$:

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u}).$$

Multiply by $(X_i - \bar{X})$ and sum over i:

$$\begin{aligned} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) &= \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u}) \\ &= \beta_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X}) u_i - \bar{u} \sum_{i=1}^{n} (X_i - \bar{X}). \end{aligned}$$

The last sum is zero:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0.$$

Divide by $\sum_{i=1}^{n} (X_i - \bar{X})^2$ and use the OLS slope formula:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}) u_i}{\frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

The numerator of the last fraction is the sample covariance of X_i and u_i (up to the divisor), which is positive under survivorship; the denominator is positive. Hence $\hat{\beta}_1$ tends to exceed β_1.

- β_1 is the population causal return to one more year of experience, in percentage points of annual gain per year; $\hat{\beta}_1$ is its sample estimate from the investors who survived.
- The survivors with many years of experience are mostly high-ability investors, so part of their higher gain is ability (u_i), not experience.
- OLS attributes that ability premium to experience, so $\hat{\beta}_1$ is biased upward and overstates the return to experience.
- More data drawn from the same surviving population leave this bias in place, because the problem is the selection rule, not the sample size.

$\hat{\beta}_1$ overstates the true causal return to experience, because both reflect unobserved investor ability.

Method 2: Conditional mean in the surviving sample
Let $S_i = 1$ if investor i is still active (and so appears in the data). Taking expectations of the model among survivors,

$$E[Y_i \mid X_i, S_i = 1] = \beta_0 + \beta_1 X_i + E[u_i \mid X_i, S_i = 1].$$

Survival depends on ability, so the last term rises with X_i, and the slope of the survivors' conditional mean is β_1 plus that rise. For a hypothetical numerical illustration take $\beta_0 = \beta_1 = 0$ (experience has no causal effect), so $Y_i = u_i$, with $u_i = +4$ percent per year for high-ability and $u_i = -4$ for low-ability investors. Two cohorts of 100 start half and half; one is observed after 1 year, the other after 10 years, by which time 45 of its 50 low-ability investors have exited. The survivors' mean gains are

$$E[Y_i \mid X_i = 1, S_i = 1] = \frac{50(4) + 50(-4)}{100} = \frac{200 - 200}{100} = 0,$$
$$E[Y_i \mid X_i = 10, S_i = 1] = \frac{50(4) + 5(-4)}{55} = \frac{200 - 20}{55} = \frac{180}{55} = 3.27.$$

With X_i taking only the two values 1 and 10, the line through the two conditional means fits each group's mean exactly, so it is the population regression line of the survivors, with slope

$$\frac{3.27 - 0}{10 - 1} = \frac{180/55}{9} = 0.36$$

percentage points of annual gain per year of experience, although the causal slope is $\beta_1 = 0$. The entire 0.36 is the ability of the survivors.

$\hat{\beta}_1$ overstates the true causal return to experience, because both reflect unobserved investor ability.

Common Mistake: a group-mean difference is not automatically $\hat{\beta}_1$'s causal meaning

Computing Mean$(Y \mid X = 1) -$ Mean$(Y \mid X = 0)$ from raw data always produces a number, but that number equals the *causal* effect of X on Y only if Assumption 1 holds for that sample. Before interpreting any such gap causally, ask: could units have selected themselves into the groups on the basis of something correlated with the outcome? If yes, the gap conflates the causal effect with a selection effect, and no amount of additional data on the same variables fixes this – the identifying assumption itself has failed.

Exam Focus: selection bias

- State the mechanism, not just the label: which unobserved factor drives both the selection (into the sample, or into a value of X) *and* the outcome Y?
- Keep a stock of contexts ready to cite: the hospital survey, wage regressions conditioned on labour-force participation, self-selection into training programmes, survey/rating non-response, and survivorship in financial-market experience (Tutorial 4) or Chapter 3's bomber-armour example.
- Selection bias is always an Assumption-1 failure; it is a different diagnosis from Assumption 2 (sampling mechanism) or Assumption 3 (outliers), even though all three concern how the data were generated.

4.2 Least Squares Assumption 2 in Depth: The Mechanism of Random Sampling

Chapter 3 stated Assumption 2 – that (X_i, Y_i), $i = 1, \ldots, n$, are independently and identically distributed – without explaining *why* a sample would have this property. This section supplies that mechanism.

Population versus sample. Suppose we want the distribution of wages across all workers in Singapore (its mean, its spread). Interviewing every worker is infeasible, so we instead draw a group of, say, 1,000 or 5,000 workers at random. The wages of *all* workers form the *population*; the wages of the 1,000 (or 5,000) selected workers form the *sample*. Treating a sampled worker's wage Y_i as a random variable does not mean her wage fluctuates – it means the *selection procedure* that put her in the sample was random: had a different sample been drawn, a different (but similarly distributed) set of wages would have appeared.

Definition 4.2: Mechanism of simple random sampling

Suppose n observations $Y_1, \ldots, Y_n$ are drawn from a population by *simple random sampling*: each member of the population is equally likely to be drawn.

- Because each member is equally likely to be drawn, knowing the value of Y_1 conveys no information about $Y_2, Y_3, \ldots$; the draws are made independently of one another, so $Y_1, \ldots, Y_n$ are **independent**.
- Because every Y_i is drawn from the *same* population, each Y_i has the **same** marginal distribution.

Independence together with identical distribution gives Least Squares Assumption 2: (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d.

Why "simple random" matters – and what breaks it. Suppose we want to estimate the monthly national unemployment rate by interviewing working-age adults sitting in city parks at 10am on Wednesdays. This is not simple random sampling: people who hold a 9-to-5 job are systematically absent from parks at that hour, so the unemployed are heavily over-represented. The resulting sample is not drawn with every population member equally likely to appear, the i.i.d. mechanism above does not apply, and any estimate of the unemployment rate from this sample is biased – regardless of how large the sample is.

Worked Example (Lecture 4): the US Current Population Survey

The US Census Bureau and Bureau of Labor Statistics conduct the Current Population Survey (CPS) every month, surveying approximately 6,000 households. Addresses are sampled randomly from a database of addresses. The CPS earnings data specifically use full-time workers, defined as employed more than 35 hours per week for at least 48 weeks in the previous year.

Method 1: Mechanism of simple random sampling (Definition 4.2)

Let Y_i be the annual earnings of the full-time worker reached at the i-th sampled address, $i = 1, \ldots, n$, with $n \approx 6{,}000$ households per month.

- The population is every full-time US worker (more than 35 hours per week for at least 48 weeks) living at an address in the database; the full-time definition fixes *which* population is studied, not *how* units are drawn from it.
- Every address in the database has the same chance of being drawn, so learning the earnings Y_1 at the first address tells us nothing about which address comes next; hence $Y_1, \dots, Y_n$ are independent.
- Every address is drawn from the same national database, so every Y_i has the same distribution of earnings, in US dollars per year; hence $Y_1, \dots, Y_n$ are identically distributed.
- In the park example the chance of being sampled depends on employment status, the very variable being measured, so the draws are neither equally likely nor representative; the CPS avoids this because its selection rule (a random address) is unrelated to anyone's earnings or employment.

Independence together with identical distribution gives Least Squares Assumption 2 for the CPS cross-section.

Random address selection (not convenience sampling, as in the park example) is what makes the CPS sample i.i.d.

Method 2: Direct probability calculation for the address draws
Label the addresses in the database $j = 1, \dots, N$ and let y_j be the earnings at address j. Let J_i be the address chosen on draw i, so $Y_i = y_{J_i}$ and $P(J_i = j) = 1/N$ for every j.
Identical distribution. For any earnings level y,

$$P(Y_i \le y) = \sum_{j:\, y_j \le y} P(J_i = j) = \sum_{j:\, y_j \le y} \frac{1}{N} = \frac{\#\{j : y_j \le y\}}{N} = F(y),$$

where F is the population distribution function of earnings. The right-hand side does not involve i, so every Y_i has distribution F.
Independence. If draws are made with replacement, for any addresses j, k,

$$P(J_1 = j,\, J_2 = k) = \frac{1}{N} \cdot \frac{1}{N} = P(J_1 = j)\, P(J_2 = k),$$

and summing over all j with $y_j \le y$ and all k with $y_k \le y'$ gives

$$P(Y_1 \le y,\, Y_2 \le y') = F(y)\, F(y'),$$

which is independence of Y_1 and Y_2 (the same argument applies to every pair and every larger group of draws). If draws are made without replacement, $P(J_2 = k \,|\, J_1 = j) = 1/(N-1)$ for $k \ne j$ instead of $1/N$, a factor

$$\frac{1/(N-1)}{1/N} = \frac{N}{N-1}.$$

For any database of at least one million addresses this factor is at most $10^6/(10^6 - 1) \approx 1.000001$, so any two of the 6,000 draws are independent to a very close approximation.

Random address selection (not convenience sampling, as in the park example) is what makes the CPS sample i.i.d.

Time series as the main non-i.i.d. case. Simple random sampling describes *cross-sectional* data well, but the mechanism breaks down for *time series* data recorded from the same unit over time. For example, a firm's quarterly revenue and the interest rate at which it borrows, recorded over 30 years, are *not* independent draws from a fixed population: this quarter's revenue is correlated with last quarter's

revenue (the data are *autocorrelated*), because the same firm persists across time and carries momentum, seasonality, and macroeconomic conditions forward. Time series is flagged here as the leading real-world violation of the independence half of Assumption 2; the identical-distribution half can also fail if the underlying relationship drifts over time (e.g. structural breaks).

Exam Focus: diagnosing i.i.d. violations

- To argue *for* i.i.d., name the random-selection mechanism explicitly (e.g. "addresses drawn at random from a national database"), not just the word "random."
- To argue *against* i.i.d., identify either (a) a non-random selection filter that changes who is likely to be sampled (the parks-at-10am example), or (b) sequential dependence across observations of the same unit (time series).
- "i.i.d." is a property of the *sampling process*, not of the underlying population – a population can be perfectly well-behaved while a bad sampling scheme still produces a biased, non-i.i.d. sample from it.

Extension: what non-i.i.d. data requires (not examinable here)

Time series regressions are not thereby unusable – they simply require different tools than the ones built on Assumption 2. Autocorrelation-robust standard errors, dynamic models with lagged regressors, and stationarity conditions all exist precisely to handle data that violate the i.i.d. assumption. None of this machinery appears in the course content list until later weeks (if at all); for now it is enough to recognise time series as the standing counterexample to Assumption 2 and to know that the inference tools of Chapter 6 (built assuming i.i.d. sampling) do not automatically carry over to it.

4.3 Why the Three Least Squares Assumptions Matter

Chapter 3 introduced Assumptions 1–3 as conditions under which OLS "behaves well"; this section previews precisely what "behaves well" means, so that the promises made here can be cashed in later. Under Assumptions 1–3, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ have three properties:

- **Unbiasedness (measure of center).** $E[\hat{\beta}_1] = \beta_1$ and $E[\hat{\beta}_0] = \beta_0$: the sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$ are centred exactly at the true population parameters. This is proved in full generality in Chapter 5.
- **Efficiency (measure of spread).** $\text{Var}[\hat{\beta}_1]$ and $\text{Var}[\hat{\beta}_0]$ are small relative to the variance of any other linear unbiased estimator built from the same data – OLS makes efficient use of the sample information. This is the content of the Gauss–Markov theorem, proved in Chapter 5.
- **Large-sample normality (shape of the distribution).** When the sample size is large ($n > 100$ is the usual rule of thumb), $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately normally distributed, by the Central Limit Theorem. This underlies the hypothesis tests and confidence intervals built in Chapter 6.

None of these three properties is proved here – that is deliberate. Proving them for $\hat{\beta}_0, \hat{\beta}_1$ directly requires machinery (the Gauss–Markov theorem, a variance formula for a ratio of sums) that is easiest to absorb after seeing the argument once in the simplest possible setting. Section 1.7 of this chapter supplies exactly that simplest setting: it derives the mean and variance of the sample average $\bar{Y}$ from first principles, as a rehearsal for the parallel (but more algebraically involved) argument Chapter 5 makes for $\hat{\beta}_1$ itself.

Common Mistake: "unbiased" does not mean "exactly right in this sample"

$E[\hat{\beta}_1] = \beta_1$ is a statement about the *sampling distribution* of $\hat{\beta}_1$ – averaged across all possible samples of size n, the estimator is centred on the truth. It does *not* mean that any particular sample's $\hat{\beta}_1$ equals β_1 exactly, nor even that it is close: a single unbiased estimate can be far from β_1 by chance. Unbiasedness is a property of the estimation *procedure*, checked before the data are drawn, not a guarantee about the one dataset in front of you.

In short: the distribution of the OLS estimator is centred at the truth, has small spread relative to competitors, and is well approximated by a normal bell curve once n is reasonably large. The remainder of this chapter develops the tools needed to make these claims precise – measures of fit for the regression itself, and the idea of a sampling distribution for an estimator, illustrated first with the simplest possible estimator, $\bar{Y}$.

4.4 Measures of Fit: The Coefficient of Determination R^2

Having a fitted line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is not the same as having a *good* line. This section defines a single number, R^2, that measures how much of the sample variation in Y the regression on X accounts for.

Definition 4.3: TSS, ESS, SSR, and R^2

Every observation decomposes exactly into its fitted value and its residual,

$$Y_i = \hat{Y}_i + \hat{u}_i.$$

Define

$$\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad \text{(total sum of squares)}, \qquad \text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \quad \text{(explained sum of squares)},$$

$$\text{SSR} = \sum_{i=1}^{n}\hat{u}_i^2 \quad \text{(sum of squared residuals)}.$$

The *coefficient of determination* is

$$R^2 = \frac{\text{ESS}}{\text{TSS}}.$$

Intuition. TSS measures the total sample variation in Y around its mean; ESS measures the part of that variation attributable to the fitted line moving with X; R^2 is the fraction of Y's total variation that X, through the regression, explains. The next four results make this precise and pin down exactly how the three sums of squares relate to one another.

Theorem 4.1: The average fitted value equals $\bar{Y}$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}.$$

Proof. Route 1 – via the residual identity. By definition of the residual,

$$Y_i = \hat{Y}_i + \hat{u}_i.$$

Summing both sides over $i = 1, \ldots, n$,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i.$$

Chapter 3 established, from the first OLS first-order condition, that $\sum_{i=1}^{n} \hat{u}_i = 0$. Substituting,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$

Dividing both sides by n gives $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$, as claimed.

Route 2 – via the intercept formula. By definition of the fitted value,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Averaging over i, and using that the average of a constant is itself and that the average of $\hat{\beta}_1 X_i$ is $\hat{\beta}_1 \bar{X}$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}.$$

Chapter 2 gives the OLS intercept formula $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Substituting,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \bar{X} = \bar{Y}.$$

Both routes reach the same conclusion. □

Corollary: the OLS line passes through $(\bar{X}, \bar{Y})$. Evaluating the fitted line at $X_i = \bar{X}$ gives $\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \bar{X} = \bar{Y}$, so the point $(\bar{X}, \bar{Y})$ always lies exactly on the OLS regression line, regardless of the data. This is the same intercept substitution used in Route 2 above, viewed at a single point rather than averaged, and it is a quick way to sanity-check a fitted line by eye: it must cross the sample means of X and Y.

Theorem 4.2: A flat fit gives zero explained variation

If $\hat{\beta}_1 = 0$, then $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$ for every i, and consequently ESS $= 0$ and $R^2 = 0$.

Proof. Suppose $\hat{\beta}_1 = 0$. By definition of the fitted value,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{\beta}_0,$$

since the term $\hat{\beta}_1 X_i$ vanishes. By the OLS intercept formula $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and again using $\hat{\beta}_1 = 0$,

$$\hat{\beta}_0 = \bar{Y} - 0 \cdot \bar{X} = \bar{Y}.$$

Combining the two displays, $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$ for every $i = 1, \ldots, n$. Then

$$\text{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\bar{Y} - \bar{Y})^2 = 0,$$

and therefore $R^2 = \text{ESS}/\text{TSS} = 0/\text{TSS} = 0$ (assuming TSS > 0, i.e. not every Y_i is identical). □

Economically, Theorem 4.2 says exactly what it should: if the fitted line is completely flat (slope zero), X predicts nothing beyond the unconditional mean of Y, and the regression explains none of Y's variation.

Theorem 4.3: Total sum of squares decomposes as TSS = ESS + SSR

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

Proof. Starting from $Y_i = \hat{Y}_i + \hat{u}_i$, subtract $\bar{Y}$ from both sides and rearrange:

$$Y_i - \bar{Y} = \hat{u}_i + (\hat{Y}_i - \bar{Y}).$$

Square both sides:

$$(Y_i - \bar{Y})^2 = \hat{u}_i^2 + 2\,\hat{u}_i(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2.$$

Sum this identity over $i = 1, \ldots, n$:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2.$$

By Definition 4.3, the first and third sums on the right are SSR and ESS, so

$$\text{TSS} = \text{SSR} + 2\Big[\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}\hat{u}_i\Big] + \text{ESS}.$$

It remains to show the bracketed cross term vanishes. Writing $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$,

$$\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i X_i.$$

Chapter 3 proved, from the two OLS first-order conditions, that $\sum_{i=1}^{n}\hat{u}_i = 0$ and $\sum_{i=1}^{n}\hat{u}_i X_i = 0$. Substituting both into the display above,

$$\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0.$$

Hence the cross term is

$$2\Big[\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}\hat{u}_i\Big] = 2[\,0 - \bar{Y}\cdot 0\,] = 0.$$

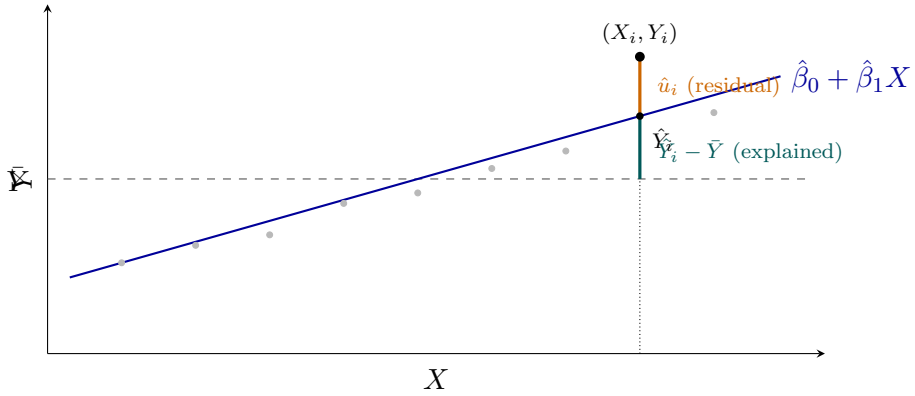
Substituting back,

$$\text{TSS} = \text{SSR} + 0 + \text{ESS} = \text{ESS} + \text{SSR}.$$

□

A Load-Bearing Figure: the ESS/SSR Decomposition

The identity of Theorem 4.3 is easiest to see for a single observation. The figure below shows one data point (X_i, Y_i) together with the fitted line and the horizontal line at $\bar{Y}$; the total deviation $Y_i - \bar{Y}$ splits into an *explained* piece, $\hat{Y}_i - \bar{Y}$, that tracks the fitted line's movement away from $\bar{Y}$, and a *residual* piece, $\hat{u}_i$, that the line misses.



Tutorial 4 makes the same point across *all* observations at once: summing the squared blue segment (ESS) and the squared orange segment (SSR) over every i recovers the squared red segment (TSS) exactly – even though for an *individual* observation the residual square can exceed the explained square, the cross term always cancels in aggregate, which is precisely the content of Theorem 4.3.

Theorem 4.4: $0 \le R^2 \le 1$, with $R^2 = 1$ if and only if every residual is zero

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}}, \qquad 0 \le R^2 \le 1,$$

and $R^2 = 1$ if and only if $\hat{u}_i = 0$ for every $i = 1, \ldots, n$.

Proof. Both ESS and SSR are sums of squares, so

$$\text{ESS} \ge 0, \qquad \text{SSR} \ge 0.$$

Divide the identity $\text{TSS} = \text{ESS} + \text{SSR}$ (Theorem 4.3) through by TSS (assuming $\text{TSS} > 0$):

$$1 = \frac{\text{ESS}}{\text{TSS}} + \frac{\text{SSR}}{\text{TSS}} = R^2 + \frac{\text{SSR}}{\text{TSS}}.$$

Rearranging,

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}}.$$

Since $\text{SSR} \ge 0$ and $\text{TSS} > 0$, the ratio $\text{SSR}/\text{TSS} \ge 0$, so

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} \le 1.$$

Since $\text{ESS} \ge 0$ and $\text{TSS} > 0$,

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \ge 0.$$

This establishes $0 \le R^2 \le 1$. For the equality case: $R^2 = 1$ iff $\text{SSR}/\text{TSS} = 0$ iff $\text{SSR} = 0$ (as $\text{TSS} > 0$), i.e. iff $\sum_{i=1}^{n} \hat{u}_i^2 = 0$. A sum of nonnegative terms (each $\hat{u}_i^2 \ge 0$) equals zero only if every term is zero, so $\text{SSR} = 0$ iff $\hat{u}_i = 0$ for every i. □

Key Formula: measures of fit

$$Y_i = \hat{Y}_i + \hat{u}_i, \qquad \text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \qquad \text{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \qquad \text{SSR} = \sum_{i=1}^{n} \hat{u}_i^2,$$

$$\text{TSS} = \text{ESS} + \text{SSR}, \qquad R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} \in [0, 1].$$

Worked Example (Lecture 4): R^2 for the test score regression

The California test score regression from Chapters 2–3,

$$\widehat{\text{TestScore}} = 698.9 - 2.28\,\text{STR}, \qquad n = 420,$$

has $R^2 = 0.051$.

Method 1: ESS/TSS share and the decomposition $\text{TSS} = \text{ESS} + \text{SSR}$

By Definition 4.3,

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 0.051 \qquad \Longrightarrow \qquad \text{ESS} = 0.051\,\text{TSS}.$$

By Theorem 4.3, $\text{TSS} = \text{ESS} + \text{SSR}$, so

$$\text{SSR} = \text{TSS} - \text{ESS} = \text{TSS} - 0.051\,\text{TSS} = (1 - 0.051)\,\text{TSS} = 0.949\,\text{TSS},$$

and dividing by TSS,

$$\frac{\text{ESS}}{\text{TSS}} = 0.051 = 5.1\%, \qquad \frac{\text{SSR}}{\text{TSS}} = 1 - R^2 = 0.949 = 94.9\%.$$

- TSS is the total sample variation of the 420 district test scores around their mean, in squared test-score points; ESS is the part traced out by the fitted line as STR (students per teacher) changes, and SSR is the part left in the residuals.
- Since the mean of the fitted values equals $\bar{Y}$ (Theorem 4.1), $\text{ESS}/(n-1)$ is the sample variance of $\widehat{\text{TestScore}}_i$ and $\text{TSS}/(n-1)$ that of TestScore_i; so the predicted scores vary only 5.1% as much (in variance) as the actual scores.
- The remaining 94.9% is due to the factors absorbed in u (for example family background and other school inputs), which the one-regressor line does not capture.
- R^2 is a sample statistic describing fit in these 420 districts; it is not the causal effect of STR, which is measured by the slope -2.28 points per student and depends on Least Squares Assumption 1.

$R^2 = 0.051$: STR explains 5.1% of the variance in TestScore; 94.9% is unexplained.

Method 2: Squared sample correlation

In a one-regressor regression with an intercept, $R^2 = [\text{Corr}(X, Y)]^2$ (Extension below). The correlation has the sign of $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$, and so does $\hat{\beta}_1 = S_{XY}/S_{XX}$ because $S_{XX} > 0$. Here $\hat{\beta}_1 = -2.28 < 0$, so

$$\text{Corr}(\text{STR}, \text{TestScore}) = -\sqrt{R^2} = -\sqrt{0.051} = -0.226,$$

a weak negative linear association between class size and test scores. Squaring it returns the explained share, and $\text{TSS} = \text{ESS} + \text{SSR}$ gives the unexplained share:

$$[\text{Corr}(\text{STR}, \text{TestScore})]^2 = (-0.2258)^2 = 0.051 = 5.1\%,$$
$$1 - [\text{Corr}(\text{STR}, \text{TestScore})]^2 = 1 - 0.051 = 0.949 = 94.9\%.$$

A correlation of -0.226 thus translates into only 5.1% of the variation in test scores being accounted for by STR.

$R^2 = 0.051$: STR explains 5.1% of the variance in TestScore; 94.9% is unexplained.

Common Mistake: a low R^2 does not mean a "bad" regression

A low R^2 does *not* imply that a regression is poorly specified or that $\hat{\beta}_1$ is unreliable. R^2 only measures how much of Y's variation X explains, not whether $\hat{\beta}_1$ is an accurate, unbiased estimate of the causal effect of X on Y. The test-score regression's $R^2 = 0.051$ simply tells us that class size is far from the only determinant of test scores – it says nothing about whether -2.28 is a good estimate of the causal effect of STR, which instead depends on whether Assumption 1 holds. A regression can have a tiny R^2 and a precisely estimated, causally valid $\hat{\beta}_1$, or a huge R^2 (e.g. from an omitted-variable-driven correlation) and a badly biased $\hat{\beta}_1$.

QRS@NTU — qrsntu.org

Extension: R^2 equals the squared sample correlation between X and Y

Let $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$, $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$, and $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{TSS}$, so that the OLS slope of Chapter 2 is $\hat{\beta}_1 = S_{XY}/S_{XX}$.
For any i, using $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$,

$$\hat{Y}_i - \bar{Y} = (\hat{\beta}_0 + \hat{\beta}_1 X_i) - \bar{Y} = \hat{\beta}_1 (X_i - \bar{X}).$$

Squaring and summing over i,

$$\text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 = \hat{\beta}_1^2 S_{XX}.$$

Substituting $\hat{\beta}_1 = S_{XY}/S_{XX}$,

$$\text{ESS} = \left(\frac{S_{XY}}{S_{XX}}\right)^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}}.$$

Hence

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{S_{XY}^2/S_{XX}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}.$$

The sample correlation coefficient is $\text{Corr}(X,Y) = \dfrac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$ (the $(n-1)$ divisors in the usual sample covariance/variance formulas cancel between numerator and denominator, so this sum-of-squares form is equivalent to the familiar $s_{XY}/(s_X s_Y)$). Therefore

$$R^2 = \left(\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\right)^2 = \big[\text{Corr}(X,Y)\big]^2.$$

In simple (one-regressor) linear regression with an intercept, R^2 is exactly the square of the sample correlation between X and Y – consistent with $R^2 \in [0,1]$, since any correlation coefficient lies in $[-1,1]$. This equivalence is specific to the one-regressor case and does not extend directly to multiple regression.
Numerical check. Take the toy sample $X = (1,2,3,4,5)$, $Y = (2,4,5,4,5)$, so $n = 5$, $\bar{X} = 3$, $\bar{Y} = 4$. Direct computation gives $S_{XX} = 10$, $S_{XY} = 6$, $S_{YY} = \text{TSS} = 6$, hence

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{6}{10} = 0.6, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 4 - 0.6(3) = 2.2.$$

The fitted values are $\hat{Y} = (2.8, 3.4, 4.0, 4.6, 5.2)$, giving residuals $\hat{u} = (-0.8, 0.6, 1.0, -0.6, -0.2)$, so $\text{ESS} = 3.6$ and $\text{SSR} = 2.4$, confirming $\text{TSS} = \text{ESS} + \text{SSR} = 3.6 + 2.4 = 6$ (Theorem 4.3). Then

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{3.6}{6} = 0.6, \qquad \text{Corr}(X,Y) = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{6}{\sqrt{10 \times 6}} \approx 0.7746,$$

$$[\text{Corr}(X,Y)]^2 \approx 0.6,$$

matching R^2 up to rounding, exactly as the general derivation predicts.

4.5 Worked Example: Predicting Weight from Height

Worked Example (Tutorial 4, Q2): weight on height

A random sample of 100 20-year-old men yields the OLS regression

$$\widehat{\text{Weight}} = -79.24 + 4.16\,\text{Height},$$

where Weight is in pounds and Height is in inches.

Method 1: Direct substitution into the sample regression line

(a) Predicted weight at 64, 68, and 72 inches. Substitute each height into $\widehat{\text{Weight}} = -79.24 + 4.16\,\text{Height}$, multiply out the slope term, then add the intercept:

$$\begin{aligned}
\text{Height} = 64: &\quad -79.24 + 4.16(64) = -79.24 + 266.24 = 187.00\\
\text{Height} = 68: &\quad -79.24 + 4.16(68) = -79.24 + 282.88 = 203.64\\
\text{Height} = 72: &\quad -79.24 + 4.16(72) = -79.24 + 299.52 = 220.28
\end{aligned}$$

$$\boxed{\widehat{\text{Weight}}(64) = 187.00, \quad \widehat{\text{Weight}}(68) = 203.64, \quad \widehat{\text{Weight}}(72) = 220.28 \text{ pounds}}$$

(b) Predicted weight gain from a 2-inch growth spurt. Let a man's height grow from H to $H + 2$ inches. The two predictions are

$$\begin{aligned}
\widehat{\text{Weight}}(H+2) &= -79.24 + 4.16(H+2) = -79.24 + 4.16H + 8.32,\\
\widehat{\text{Weight}}(H) &= -79.24 + 4.16H.
\end{aligned}$$

Subtracting, the intercept and the $4.16H$ terms cancel:

$$\Delta\widehat{\text{Weight}} = \widehat{\text{Weight}}(H+2) - \widehat{\text{Weight}}(H) = 8.32,$$

whatever H is; e.g. $\widehat{\text{Weight}}(70) = -79.24 + 291.20 = 211.96$ and $211.96 - 203.64 = 8.32$.

$$\boxed{\Delta\widehat{\text{Weight}} = 8.32 \text{ pounds for a 2-inch increase in height}}$$

Method 2: Slope times change in height

On a straight line the predicted change is the slope $\hat{\beta}_1 = 4.16$ pounds per inch times the change in height: $\Delta\widehat{\text{Weight}} = \hat{\beta}_1\,\Delta\text{Height}$.

(a) Start from $-79.24 + 4.16(64) = -79.24 + 266.24 = 187.00$ and step up 4 inches at a time, each step adding $4.16 \times 4 = 16.64$ pounds:

$$\begin{aligned}
\widehat{\text{Weight}}(68) &= \widehat{\text{Weight}}(64) + 4.16 \times 4 = 187.00 + 16.64 = 203.64,\\
\widehat{\text{Weight}}(72) &= \widehat{\text{Weight}}(68) + 4.16 \times 4 = 203.64 + 16.64 = 220.28.
\end{aligned}$$

$$\boxed{\widehat{\text{Weight}}(64) = 187.00, \quad \widehat{\text{Weight}}(68) = 203.64, \quad \widehat{\text{Weight}}(72) = 220.28 \text{ pounds}}$$

(b) With $\Delta\text{Height} = 2$ inches,

$$\begin{aligned}
\Delta\widehat{\text{Weight}} &= \hat{\beta}_1\,\Delta\text{Height} && \text{(linear fit)}\\
&= 4.16 \times 2 = 8.32.
\end{aligned}$$

$$\boxed{\Delta\widehat{\text{Weight}} = 8.32 \text{ pounds for a 2-inch increase in height}}$$

4.6 The CAPM and the "Beta" of a Stock

A further application of simple regression, drawn from finance, is the Capital Asset Pricing Model (CAPM). Investors require compensation for bearing risk, so the expected return on a risky asset R should exceed the return on a safe asset R_f (often proxied by the yield on a short-term US government bond); the expected *excess return* $R - R_f$ should be positive.

Definition 4.4: The CAPM regression

The Capital Asset Pricing Model relates an asset's excess return to the excess return on the market portfolio (e.g. the S&P 500) via

$$R - R_f = \alpha + \beta(R_m - R_f) + u,$$

where $R - R_f$ is the asset's excess return, $R_m - R_f$ is the market's excess return, α is the asset's abnormal return (zero under the strict CAPM), and β (the "market beta") measures the asset's sensitivity to market-wide movements. In practice α, β are estimated by OLS, exactly as β_0, β_1 are estimated in the generic model $Y_i = \beta_0 + \beta_1 X_i + u_i$, with $Y = R - R_f$ and $X = R_m - R_f$.

Under CAPM theory, a stock with $\beta < 1$ has *less* systematic risk than the market portfolio and, correspondingly, a *lower* expected excess return than the market; a stock with $\beta > 1$ has *more* systematic risk and a *higher* expected excess return. Estimated betas for individual stocks are a standard workhorse output on investment-research websites, produced simply by regressing the stock's excess return (Y) on the market's excess return (X). Reading such a table is a direct application of the interpretation above: a technology stock reported with $\hat{\beta} = 1.4$ moves, on average, 1.4% for every 1% move in the market and commands a higher expected excess return than the market; a defensive utility stock reported with $\hat{\beta} = 0.6$ moves less than the market and correspondingly commands a lower expected excess return. Each such $\hat{\beta}$ is nothing more than the OLS slope $\hat{\beta}_1$ of Chapter 2, applied to a pair of return series instead of a test-score/class-size pair.

Worked Example (Tutorial 4, Q4): comparing two assets with the same β but different α

Consider two assets (or funds) with excess returns following the CAPM regression, $R - R_f = \alpha + \beta(R_m - R_f) + u$, sharing the *same* market beta β but different intercepts $\alpha_1 \neq \alpha_2$. Assume, as in the population version of Assumption 1, that $\text{Cov}(R_m - R_f, u) = 0$, so the idiosyncratic term u is uncorrelated with the market factor.

Method 1: Expectation and variance rules

Write $Y = R - R_f$ and $X = R_m - R_f$, so asset $k = 1, 2$ has $Y_k = \alpha_k + \beta X + u_k$ with the same β. The variance comparison needs one further assumption, made throughout: the two assets have the same idiosyncratic variance, $\text{Var}(u_1) = \text{Var}(u_2)$. Least Squares Assumption 1 in population form, $E[u_k \mid X] = 0$, gives by the law of iterated expectations

$$E[u_k] = E\big[E[u_k \mid X]\big] = E[0] = 0.$$

Expected excess return. By linearity of expectation, with α_k and β constants,

$$E[Y_k] = E[\alpha_k + \beta X + u_k] = \alpha_k + \beta\, E[X] + E[u_k] = \alpha_k + \beta\, E[R_m - R_f].$$

For the two assets,

$$E[Y_2] - E[Y_1] = \big(\alpha_2 + \beta\, E[X]\big) - \big(\alpha_1 + \beta\, E[X]\big) = \alpha_2 - \alpha_1,$$

which is positive when $\alpha_2 > \alpha_1$: the expected excess return is strictly increasing in α for fixed β.

Variance (risk). Adding a constant does not change a variance, and $\text{Var}(aZ + W) = a^2\,\text{Var}(Z) +$

$\mathrm{Var}(W) + 2a\,\mathrm{Cov}(Z, W)$ for a constant a. Hence

$$\begin{aligned}\mathrm{Var}(Y_k) &= \mathrm{Var}(\alpha_k + \beta X + u_k) = \mathrm{Var}(\beta X + u_k) && (\alpha_k \text{ constant})\\ &= \beta^2\,\mathrm{Var}(X) + \mathrm{Var}(u_k) + 2\beta\,\mathrm{Cov}(X, u_k)\\ &= \beta^2\,\mathrm{Var}(R_m - R_f) + \mathrm{Var}(u_k) + 2\beta\cdot 0 && (\mathrm{Cov} = 0)\\ &= \beta^2\,\mathrm{Var}(R_m - R_f) + \mathrm{Var}(u_k).\end{aligned}$$

For the two assets,

$$\mathrm{Var}(Y_2) - \mathrm{Var}(Y_1) = \big[\beta^2\,\mathrm{Var}(X) + \mathrm{Var}(u_2)\big] - \big[\beta^2\,\mathrm{Var}(X) + \mathrm{Var}(u_1)\big] = \mathrm{Var}(u_2) - \mathrm{Var}(u_1) = 0.$$

- α shifts the expected excess return one for one: asset 2 earns $\alpha_2 - \alpha_1$ more per period on average, in the same units as the returns (percent per month in the illustration below).
- α does not enter the variance: risk is $\beta^2\,\mathrm{Var}(R_m - R_f)$ (market risk, set by β) plus $\mathrm{Var}(u)$ (idiosyncratic risk), and both pieces are equal for the two assets.
- With equal β and equal $\mathrm{Var}(u)$, the higher-α asset offers a higher expected return for the same risk; under the strict CAPM $\alpha = 0$, so a positive α is a premium the CAPM does not explain.
- α, β and $\mathrm{Var}(u)$ are population parameters; in a sample they are estimated by OLS, and a positive $\hat{\alpha}$ is evidence of a positive α only if it is statistically significant (Chapter 6).

For a hypothetical numerical illustration take monthly data with $E[R_m - R_f] = 0.6\%$, $\mathrm{sd}(R_m - R_f) = 4\%$, $\beta = 1.2$, $\mathrm{sd}(u_k) = 2\%$, $\alpha_1 = 0$ and $\alpha_2 = 0.3\%$:

$$\begin{aligned}\mathrm{Var}(Y_k) &= (1.2)^2(0.04)^2 + (0.02)^2 = 1.44 \times 0.0016 + 0.0004\\ &= 0.002304 + 0.0004 = 0.002704,\\ \mathrm{sd}(Y_k) &= \sqrt{0.002704} = 0.0520 = 5.20\% \quad \text{(both assets)},\\ E[Y_1] &= 0 + 1.2(0.6\%) = 0.72\%, \qquad E[Y_2] = 0.3\% + 1.2(0.6\%) = 0.3\% + 0.72\% = 1.02\%.\end{aligned}$$

Asset 2 earns $1.02\% - 0.72\% = 0.30$ percentage points more per month at the identical standard deviation of 5.20%.

Same β, $\alpha_2 > \alpha_1 \Rightarrow E[R - R_f]_2 > E[R - R_f]_1$ while $\mathrm{Var}(R - R_f)_1 = \mathrm{Var}(R - R_f)_2$.

Method 2: Variance from the definition $E\big[(Y - E[Y])^2\big]$
Write $Y = R - R_f$ and $X = R_m - R_f$, so asset $k = 1, 2$ has $Y_k = \alpha_k + \beta X + u_k$, and assume as in Method 1 that $\mathrm{Var}(u_1) = \mathrm{Var}(u_2)$. Taking expectations of $Y_k = \alpha_k + \beta X + u_k$ and using $E[u_k] = E\big[E[u_k \mid X]\big] = 0$,

$$E[Y_k] = \alpha_k + \beta\,E[X] + E[u_k] = \alpha_k + \beta\,E[X],$$

so $E[Y_2] - E[Y_1] = \alpha_2 - \alpha_1 > 0$. The deviation of Y_k from its mean is

$$Y_k - E[Y_k] = \big(\alpha_k + \beta X + u_k\big) - \big(\alpha_k + \beta\,E[X]\big) = \beta\big(X - E[X]\big) + u_k.$$

The constant α_k cancels at this step, so it cannot appear in the variance. Square the deviation:

$$\big(Y_k - E[Y_k]\big)^2 = \beta^2\big(X - E[X]\big)^2 + 2\beta\big(X - E[X]\big)u_k + u_k^2.$$

Take expectations term by term:

$$\mathrm{Var}(Y_k) = \beta^2\,E\big[(X - E[X])^2\big] + 2\beta\,E\big[(X - E[X])u_k\big] + E[u_k^2].$$

The first expectation is $\mathrm{Var}(X)$ by definition. For the second, $E[u_k] = 0$ gives

$$E\big[(X - E[X])u_k\big] = E[Xu_k] - E[X]\,E[u_k] = \mathrm{Cov}(X, u_k) = 0.$$

For the third,

$$E[u_k^2] = \mathrm{Var}(u_k) + \big(E[u_k]\big)^2 = \mathrm{Var}(u_k) + 0^2 = \mathrm{Var}(u_k).$$

Therefore

$$\mathrm{Var}(Y_k) = \beta^2\,\mathrm{Var}(R_m - R_f) + 2\beta \cdot 0 + \mathrm{Var}(u_k) = \beta^2\,\mathrm{Var}(R_m - R_f) + \mathrm{Var}(u_k),$$

which is the same for $k = 1, 2$ because the assets share β and $\mathrm{Var}(u_1) = \mathrm{Var}(u_2)$, while α_k never entered. With the hypothetical monthly numbers of Method 1, $\mathrm{Var}(Y_k) = 0.002304 + 0 + 0.0004 = 0.002704$ for both assets, and $E[Y_2] - E[Y_1] = 0.30$ percentage points per month.

Same β, $\alpha_2 > \alpha_1 \Rightarrow E[R - R_f]_2 > E[R - R_f]_1$ while $\mathrm{Var}(R - R_f)_1 = \mathrm{Var}(R - R_f)_2$.

Policy Application: "beating the market"

Under the strict CAPM, $\alpha = 0$ for every asset: excess returns are fully explained by exposure to market risk, $\beta(R_m - R_f)$. If an estimated $\hat{\alpha}$ for a stock or fund is significantly positive (in the sense developed formally in Chapter 6's t-tests), that asset has historically earned more than the CAPM predicts given its market risk – informally, it has "beaten the market." Fund managers and analysts prize a reliably positive $\hat{\alpha}$ precisely because Tutorial 4's comparison above shows it represents extra expected return *without* extra market-beta risk, though the worked example above compares only the market-risk component of variance; realised alpha over any particular historical window can still reflect sampling noise or non-market risks rather than a persistent skill.

4.7 The Sampling Distribution of the OLS Estimators

Section 1.3 asserted that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, efficient, and approximately normal in large samples. All three claims describe the *sampling distribution* of $\hat{\beta}_0, \hat{\beta}_1$ – so before those claims can be proved (Chapters 5–6), we need to be precise about what a sampling distribution is.

Definition 4.5: Sampling distribution

Because $\hat{\beta}_0, \hat{\beta}_1$ are computed from a randomly drawn sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, they are themselves random variables: had a different random sample been drawn, $\hat{\beta}_0, \hat{\beta}_1$ would generally have taken different numerical values. The probability distribution describing the values $\hat{\beta}_0$ (or $\hat{\beta}_1$) could take across repeated random samples of size n from the same population is its *sampling distribution*.

This is exactly analogous to the sample mean $\bar{Y}$ as an estimator of a population mean μ: $\bar{Y}$ is calculated from a randomly drawn sample, so it too varies from one sample to the next, and that variation is summarised by $\bar{Y}$'s own sampling distribution. Because the algebra of $\hat{\beta}_0, \hat{\beta}_1$ is more involved than that of $\bar{Y}$, this chapter develops the mean and variance of the sampling distribution of $\bar{Y}$ in full, as the template for what Chapter 5 does for $\hat{\beta}_0, \hat{\beta}_1$.

Theorem 4.5: Mean and variance of the sample average

Let $Y_1, \ldots, Y_n$ be i.i.d. with $E[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma^2$ for every i, and let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then

$$E[\bar{Y}] = \mu, \qquad \text{Var}[\bar{Y}] = \frac{\sigma^2}{n}.$$

Proof. Mean. By definition of the sample average and linearity of expectation,

$$E[\bar{Y}] = E\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i].$$

Since every Y_i has the same mean μ (identical distribution),

$$\frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \frac{1}{n}(n\mu) = \mu.$$

Hence $E[\bar{Y}] = \mu$.

Variance. By definition of the sample average and the scaling rule $\text{Var}(cZ) = c^2\,\text{Var}(Z)$ for a constant c,

$$\text{Var}[\bar{Y}] = \text{Var}\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i\Big] = \frac{1}{n^2}\,\text{Var}\Big[\sum_{i=1}^{n} Y_i\Big].$$

The variance of a sum expands into the sum of variances plus every pairwise covariance,

$$\text{Var}\Big[\sum_{i=1}^{n} Y_i\Big] = \sum_{i=1}^{n}\text{Var}(Y_i) + \sum_{i\neq j}\text{Cov}(Y_i, Y_j).$$

Because $Y_1, \ldots, Y_n$ are independent (Definition 4.2), every cross term vanishes: $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$. Since every Y_i has the same variance σ^2,

$$\text{Var}\Big[\sum_{i=1}^{n} Y_i\Big] = \sum_{i=1}^{n}\sigma^2 + 0 = n\sigma^2.$$

Substituting back,

$$\text{Var}[\bar{Y}] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

□

Two features of this proof generalise directly. First, $E[\bar{Y}] = \mu$ used only that the Y_i have a common mean – *identical* distribution – not independence; the analogous unbiasedness proof for $\hat{\beta}_1$ in Chapter 5 will similarly hinge on Assumption 1, not Assumption 2. Second, $\text{Var}[\bar{Y}] = \sigma^2/n$ collapsed to a clean $1/n$ shrinkage precisely because *independence* killed every covariance term; Chapter 5's derivation of $\text{Var}[\hat{\beta}_1]$ leans on the same mechanism.

The $1/n$ shrinkage also gives content to the word "efficient" from Section 1.3: as n grows, $\text{Var}[\bar{Y}]$ shrinks toward zero, so $\bar{Y}$'s sampling distribution concentrates ever more tightly around μ. Section 1.3 asserted the same qualitative shrinkage for $\text{Var}[\hat{\beta}_1]$ as n grows – more data pin down the OLS slope more precisely – and Chapter 5 makes this precise by deriving $\text{Var}[\hat{\beta}_1]$ explicitly and showing it is the smallest attainable variance among linear unbiased estimators (Gauss–Markov).

Large-sample normality. The Central Limit Theorem (CLT) states that, regardless of the shape of the distribution of Y_i itself, the standardised sample average $(\bar{Y} - \mu)/\sqrt{\sigma^2/n}$ converges to a standard normal distribution as $n \to \infty$; in practice, $\bar{Y}$ is already well approximated by $N(\mu, \sigma^2/n)$ once $n > 100$. Under the Least Squares Assumptions, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the outputs of formulas built from sample averages of (X_i, Y_i) and their products, so the same CLT logic applies to them: for large n (again, $n > 100$ is the

standard rule of thumb), $\hat{\beta}_0$ and $\hat{\beta}_1$ are approximately normally distributed around β_0, β_1 respectively. This chapter states the result; Chapter 5 supplies the unbiasedness and efficiency half of the picture, and Chapter 6 puts the normal approximation to work in t-tests and confidence intervals.

Key Formula: sampling distribution of $\bar{Y}$

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2) \implies E[\bar{Y}] = \mu, \qquad \text{Var}[\bar{Y}] = \frac{\sigma^2}{n}, \qquad \bar{Y} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ for } n > 100.$$

By analogy, under Assumptions 1–3: $E[\hat{\beta}_1] = \beta_1$, $\text{Var}[\hat{\beta}_1]$ small, $\hat{\beta}_1 \overset{\text{approx.}}{\sim} N(\beta_1, \text{Var}[\hat{\beta}_1])$ for $n > 100$ (proved in Chapters 5–6).

Exam Focus: sampling distribution questions

- Be able to reproduce the $E[\bar{Y}] = \mu$, $\text{Var}[\bar{Y}] = \sigma^2/n$ derivation from scratch, and to state *exactly* which property (identical distribution vs. independence) each step uses – a very similar computation appears again in Chapter 5 for $\hat{\beta}_1$, and in Tutorial 5 for a two-sample estimator $\hat{d} = \bar{Y}_1 - \bar{Y}_2$.
- Do not confuse the *sampling distribution* of $\hat{\beta}_1$ (across hypothetical repeated samples) with the *distribution of* u_i within one sample – they are different objects, and Assumption 1–3 constrain the latter in order to characterise the former.
- Know the $n > 100$ rule of thumb as the informal threshold at which the normal approximation for $\hat{\beta}_0, \hat{\beta}_1$ is considered reliable.

4.8 Chapter Summary

Key Formula: Chapter 4 recap

$$Y_i = \hat{Y}_i + \hat{u}_i, \qquad \text{TSS} = \text{ESS} + \text{SSR}, \qquad R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} = [\text{Corr}(X, Y)]^2,$$

$$E[\bar{Y}] = \mu, \qquad \text{Var}[\bar{Y}] = \frac{\sigma^2}{n}, \qquad R - R_f = \alpha + \beta(R_m - R_f) + u.$$

Exam Focus: Chapter 4 takeaways

- Selection bias is an Assumption-1 problem, not a data-quantity problem: name the self-selection mechanism (hospital visits, survivorship, non-response) and explain why it correlates X (or sample inclusion) with u.
- Assumption 2 (i.i.d.) follows from the mechanism of simple random sampling: equal selection probability gives independence, common population gives identical distribution; time series is the standard counterexample.
- Master the full proof chain for measures of fit: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i \Rightarrow \hat{\beta}_1 = 0 \Rightarrow R^2 = 0 \Rightarrow \text{TSS} = \text{ESS} + \text{SSR} \Rightarrow 0 \leq R^2 \leq 1$ – each step reuses the residual identities $\sum_{i=1}^{n}\hat{u}_i = 0$, $\sum_{i=1}^{n}\hat{u}_i X_i = 0$ from Chapter 3, and exam questions test this chain from any starting point.
- A low R^2 says nothing about whether $\hat{\beta}_1$ is a good *causal* estimate – that question belongs to Assumption 1, not to R^2.
- $\hat{\beta}_0, \hat{\beta}_1$ have a sampling distribution because they are functions of the random sample, exactly like $\bar{Y}$; be able to reproduce $E[\bar{Y}] = \mu$, $\text{Var}[\bar{Y}] = \sigma^2/n$ and to state the $n > 100$ normal-approximation rule of thumb.

QRS@NTU – qrsntu.org

5 Lecture 5: Unbiasedness, Efficiency and Binary Regressors

This chapter answers three questions: (1) *what does it mean for an estimator to be "good," and how do we compare two candidate estimators of the same parameter?* (2) *why is OLS, among all linear unbiased estimators, the one we should use — the content of the Gauss–Markov Theorem?* (3) *how do we build and interpret a regression when the regressor is not continuous but a 0/1 indicator?* It builds on Chapter 4 (the sampling distribution of $\bar{Y}$) and prepares Chapter 6 (formal hypothesis testing and confidence intervals for $\hat{\beta}_1$).

5.1 Estimators, Unbiasedness, and Efficiency

Chapter 4 showed that $\bar{Y}$, computed from a random sample, is itself a random variable with a sampling distribution: different samples give different numerical values of $\bar{Y}$. This raises a natural question. If several different formulas could all serve as "estimators" of the same unknown population parameter, how do we decide which one to use? This section formalises two criteria — unbiasedness and efficiency — that any respectable estimator should satisfy, using the sample mean $\bar{Y}$ as the running example before extending the same logic to the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ later in the chapter.

Definition 5.1: Estimator

An *estimator* of a population parameter θ is a rule — a function of the sample data $Y_1, \ldots, Y_n$ — that produces a numerical guess $\hat{\theta}$ for θ. Because the sample itself is random, different samples drawn from the same population generally produce different numerical values of $\hat{\theta}$; the estimator $\hat{\theta}$ is therefore a random variable with its own *sampling distribution*, as opposed to an *estimate*, which is the single numerical value $\hat{\theta}$ takes for one particular sample.

Definition 5.2: Unbiasedness

An estimator $\hat{\theta}$ of a parameter θ is *unbiased* if the mean (centre) of its sampling distribution equals the true parameter value:

$$E[\hat{\theta}] = \theta.$$

If instead $E[\hat{\theta}] \neq \theta$, the estimator is *biased*, with bias $E[\hat{\theta}] - \theta$.

Intuition. Suppose you evaluate an estimator many times over repeated, randomly drawn samples. For one sample the result may exceed the true value; for another it may fall short. Unbiasedness says that, averaged over infinitely many such samples, the estimator gets it exactly right — there is no systematic tendency to over- or under-estimate θ.

Definition 5.3: Efficiency

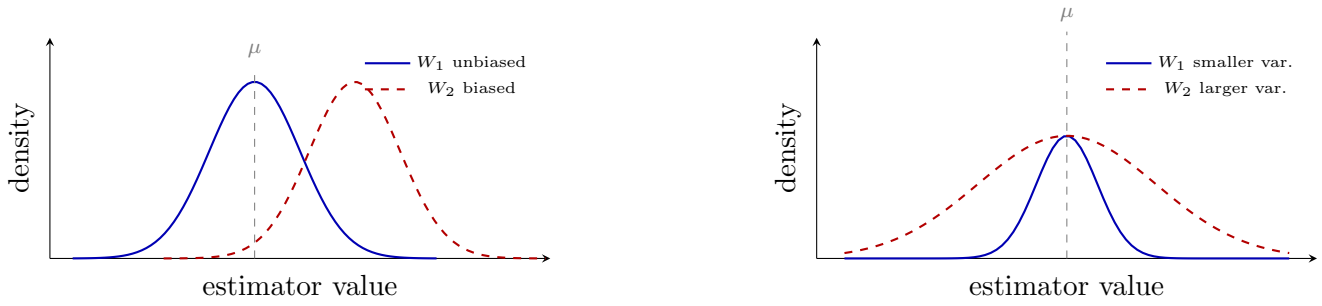
Let $\hat{\theta}$ and $\tilde{\beta}$ (or any two estimators, written generically) both be *unbiased* estimators of the same parameter θ. Then $\hat{\theta}$ is *more efficient* than $\tilde{\beta}$ if

$$\mathrm{Var}(\hat{\theta}) < \mathrm{Var}(\tilde{\beta}).$$

Efficiency is a comparison *among unbiased estimators only*: it says nothing about a biased estimator, however small its variance, because a biased estimator's sampling distribution is not even centred at the right place.

Combining the two criteria: among all unbiased estimators of a parameter, we prefer the one whose sampling distribution is most tightly concentrated around the true value, i.e., the one with the smallest

variance.



Left panel: W_1 is unbiased (centred at μ) and W_2 is biased (centred away from μ); no comparison of variances can rescue W_2, since it targets the wrong value on average. **Right panel:** W_1 and W_2 are both unbiased (both centred at μ), but W_1's sampling distribution is tighter, so W_1 is more efficient than W_2.

5.2 Linear Estimators of a Population Mean

Let $Y_1, \ldots, Y_n$ be i.i.d. draws with $E[Y_i] = \mu$ and $\text{Var}(Y_i) = \sigma^2$ for every i. A *linear estimator* of μ is any weighted sum

$$\hat{\mu} = \sum_{i=1}^{n} a_i Y_i,$$

where $a_1, \ldots, a_n$ are constants (not random, and not depending on $Y_1, \ldots, Y_n$) chosen before the data are observed. Different choices of the weights a_i nest familiar estimators: $a_1 = 1$ and $a_2 = \cdots = a_n = 0$ gives $\hat{\mu} = Y_1$; $a_i = 1/n$ for every i gives $\hat{\mu} = \bar{Y}$. Tutorial 5 Question 1(c) asks which weights make $\hat{\mu}$ unbiased in general.

Theorem 5.1: Unbiasedness of a linear estimator

The linear estimator $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$ is an unbiased estimator of μ if and only if

$$\sum_{i=1}^{n} a_i = 1.$$

Proof. Take expectations of $\hat{\mu}$:

$$E[\hat{\mu}] = E\Big[\sum_{i=1}^{n} a_i Y_i\Big] = \sum_{i=1}^{n} a_i \, E[Y_i]$$

(linearity of expectation; the a_i are non-random constants, so they pass through the expectation operator). Since $Y_1, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu$ for every i,

$$\sum_{i=1}^{n} a_i \, E[Y_i] = \sum_{i=1}^{n} a_i \mu = \mu \sum_{i=1}^{n} a_i.$$

So $E[\hat{\mu}] = \mu \sum_{i=1}^{n} a_i$. *(Sufficiency.)* If $\sum_{i=1}^{n} a_i = 1$, then $E[\hat{\mu}] = \mu \cdot 1 = \mu$, so $\hat{\mu}$ is unbiased. *(Necessity.)* Unbiasedness requires $E[\hat{\mu}] = \mu$ to hold for the (unknown, but arbitrary) population mean μ; substituting the expression just derived,

$$\mu \sum_{i=1}^{n} a_i = \mu \quad \Longleftrightarrow \quad \mu\Big(\sum_{i=1}^{n} a_i - 1\Big) = 0.$$

Because this must hold identically in μ — $a_1, \ldots, a_n$ are fixed weights chosen before μ's value is known, not tailored to one numerical μ — the bracketed coefficient must itself vanish:

$$\sum_{i=1}^{n} a_i - 1 = 0 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} a_i = 1. \qquad \square$$

Theorem 5.2: Variance of a linear estimator, and $\bar{Y}$ as BLUE

Under the same setup, the variance of the linear estimator $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$ is

$$\mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2.$$

Moreover, among all weights satisfying the unbiasedness condition $\sum_{i=1}^{n} a_i = 1$, the quantity $\sum_{i=1}^{n} a_i^2$ (and hence $\mathrm{Var}(\hat{\mu})$) is minimised uniquely at $a_i = 1/n$ for every i, i.e., at $\hat{\mu} = \bar{Y}$. Consequently

$$\mathrm{Var}(\bar{Y}) \leq \mathrm{Var}(\hat{\mu})$$

for every linear unbiased estimator $\hat{\mu}$ of μ: $\bar{Y}$ is the *Best Linear Unbiased Estimator* (BLUE) of μ.

Proof. **Step 1 (variance formula).** By the general variance-of-a-sum identity,

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\Big(\sum_{i=1}^{n} a_i Y_i\Big) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(Y_i) + 2\sum_{i<j} a_i a_j \,\mathrm{Cov}(Y_i, Y_j).$$

Since $Y_1, \ldots, Y_n$ are i.i.d. and hence pairwise independent, $\mathrm{Cov}(Y_i, Y_j) = 0$ for every $i \neq j$, so all cross terms vanish:

$$\mathrm{Var}(\hat{\mu}) = \sum_{i=1}^{n} a_i^2 \,\mathrm{Var}(Y_i) = \sum_{i=1}^{n} a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} a_i^2$$

(using $\mathrm{Var}(Y_i) = \sigma^2$ for every i).

Step 2 (minimising $\sum_{i=1}^{n} a_i^2$ subject to $\sum_{i=1}^{n} a_i = 1$). Decompose each weight as its deviation from $1/n$ plus $1/n$:

$$a_i = \Big(a_i - \tfrac{1}{n}\Big) + \tfrac{1}{n}.$$

Square and sum over i:

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2 + \frac{2}{n} \sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big) + \sum_{i=1}^{n} \frac{1}{n^2}.$$

The middle term vanishes because

$$\sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big) = \sum_{i=1}^{n} a_i - n \cdot \tfrac{1}{n} = 1 - 1 = 0$$

(using the unbiasedness constraint $\sum_{i=1}^{n} a_i = 1$ from Theorem 5.1). The last term is

$$\sum_{i=1}^{n} \frac{1}{n^2} = n \cdot \frac{1}{n^2} = \frac{1}{n}.$$

Hence

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2 + \frac{1}{n}.$$

Step 3 (conclude the minimisation). The first term on the right is a sum of squares, so $\sum_{i=1}^{n} (a_i - 1/n)^2 \geq 0$, with equality if and only if $a_i = 1/n$ for every i. Therefore

$$\sum_{i=1}^{n} a_i^2 \geq \frac{1}{n},$$

with the minimum value $1/n$ attained uniquely when $a_i = 1/n$ for all i, i.e., when $\hat{\mu} = \sum_{i=1}^{n} \frac{1}{n} Y_i = \bar{Y}$.

Step 4 (BLUE). Combining Steps 1 and 3, for *every* choice of weights satisfying $\sum_{i=1}^n a_i = 1$,

$$\mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2 \geq \sigma^2 \cdot \frac{1}{n} = \mathrm{Var}(\bar{Y}).$$

No linear unbiased estimator of μ has smaller variance than $\bar{Y}$, so $\bar{Y}$ is BLUE. □

Key Formula: Linear Estimators of μ

$$\hat{\mu} = \sum_{i=1}^n a_i Y_i \text{ unbiased} \iff \sum_{i=1}^n a_i = 1, \qquad \mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2, \qquad \sum_{i=1}^n a_i^2 = \sum_{i=1}^n \left(a_i - \tfrac{1}{n}\right)^2 + \frac{1}{n} \geq \frac{1}{n}.$$

Minimised at $a_i = 1/n \ \forall i \Rightarrow \hat{\mu} = \bar{Y}$, so $\bar{Y}$ is **BLUE**: $\mathrm{Var}(\bar{Y}) = \sigma^2/n \leq \mathrm{Var}(\hat{\mu})$ for any linear unbiased $\hat{\mu}$.

Worked Example (Tutorial 5 Q1a): Y_1, $\bar{Y}$, and $\tilde{Y} = \frac{1}{2}(Y_1 + Y_n)$

Let $Y_1, \ldots, Y_n$ i.i.d. with $E[Y_i] = \mu$, $\mathrm{Var}(Y_i) = \sigma^2$. Consider three candidate estimators of μ: Y_1 (the first observation), $\bar{Y}$, and $\tilde{Y} = \frac{1}{2}(Y_1 + Y_n)$ (the average of the first and last observations).

Method 1: Weights representation (Theorems 5.1 and 5.2)
Write each estimator as a linear estimator $\hat{\mu} = \sum_{i=1}^n a_i Y_i$ and read off its weights:

$$\begin{aligned}
Y_1 &= 1 \cdot Y_1 + 0 \cdot Y_2 + \cdots + 0 \cdot Y_n && \Longrightarrow\ a_1 = 1,\ a_i = 0\ (i \geq 2),\\
\bar{Y} &= \tfrac{1}{n} Y_1 + \tfrac{1}{n} Y_2 + \cdots + \tfrac{1}{n} Y_n && \Longrightarrow\ a_i = \tfrac{1}{n}\ \text{(every } i),\\
\tilde{Y} &= \tfrac{1}{2} Y_1 + 0 \cdot Y_2 + \cdots + 0 \cdot Y_{n-1} + \tfrac{1}{2} Y_n && \Longrightarrow\ a_1 = a_n = \tfrac{1}{2},\ a_i = 0\ \text{(else)}.
\end{aligned}$$

By Theorem 5.1, $E[\hat{\mu}] = \mu \sum_{i=1}^n a_i$, so an estimator is unbiased exactly when its weights sum to 1. Sum each set of weights:

$$\begin{aligned}
Y_1: &\quad \sum_{i=1}^n a_i = 1 + \underbrace{0 + \cdots + 0}_{n-1 \text{ terms}} = 1,\\
\bar{Y}: &\quad \sum_{i=1}^n a_i = \underbrace{\tfrac{1}{n} + \cdots + \tfrac{1}{n}}_{n \text{ terms}} = n \cdot \tfrac{1}{n} = 1,\\
\tilde{Y}: &\quad \sum_{i=1}^n a_i = \tfrac{1}{2} + \underbrace{0 + \cdots + 0}_{n-2 \text{ terms}} + \tfrac{1}{2} = 1.
\end{aligned}$$

Hence

$$E[Y_1] = E[\bar{Y}] = E[\tilde{Y}] = \mu \cdot 1 = \mu,$$

so all three are unbiased, and comparing their variances is a legitimate efficiency comparison.
By Theorem 5.2, $\mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2$. Square and sum each set of weights:

$$\begin{aligned}
\mathrm{Var}(Y_1) &= \sigma^2\left[1^2 + (n-1) \cdot 0^2\right] = \sigma^2,\\
\mathrm{Var}(\tilde{Y}) &= \sigma^2\left[\left(\tfrac{1}{2}\right)^2 + (n-2) \cdot 0^2 + \left(\tfrac{1}{2}\right)^2\right] = \sigma^2\left[\tfrac{1}{4} + \tfrac{1}{4}\right] = \frac{\sigma^2}{2},\\
\mathrm{Var}(\bar{Y}) &= \sigma^2\left[n \cdot \left(\tfrac{1}{n}\right)^2\right] = \sigma^2 \cdot \frac{n}{n^2} = \frac{\sigma^2}{n}.
\end{aligned}$$

Compare $\bar{Y}$ with $\tilde{Y}$. Divide by $\sigma^2 > 0$, then multiply by $2n > 0$:

$$\mathrm{Var}(\bar{Y}) < \mathrm{Var}(\tilde{Y}) \iff \frac{\sigma^2}{n} < \frac{\sigma^2}{2} \iff \frac{1}{n} < \frac{1}{2} \iff 2 < n.$$

Compare $\tilde{Y}$ with Y_1. Divide by $\sigma^2 > 0$:

$$\mathrm{Var}(\tilde{Y}) < \mathrm{Var}(Y_1) \iff \frac{\sigma^2}{2} < \sigma^2 \iff \frac{1}{2} < 1,$$

which holds for every n. When $n = 2$, $\tilde{Y} = \frac{1}{2}(Y_1 + Y_2) = \bar{Y}$, so the first inequality becomes an equality; that is why the ranking needs $n > 2$.

For $n > 2$: $\mathrm{Var}(\bar{Y}) < \mathrm{Var}(\tilde{Y}) < \mathrm{Var}(Y_1)$, so $\bar{Y}$ is the most efficient of the three.

Method 2: Expectation and variance rules applied directly

Use linearity of expectation, $\mathrm{Var}(cW) = c^2\,\mathrm{Var}(W)$, and $\mathrm{Var}(W+V) = \mathrm{Var}(W) + \mathrm{Var}(V) + 2\,\mathrm{Cov}(W,V)$, with $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ because the draws are independent.
Expectations:

$$\begin{aligned} E[Y_1] &= \mu, \\ E[\bar{Y}] &= E\Big[\tfrac{1}{n}\sum_{i=1}^{n} Y_i\Big] = \tfrac{1}{n}\sum_{i=1}^{n} E[Y_i] = \tfrac{1}{n}\cdot n\mu = \mu, \\ E[\tilde{Y}] &= \tfrac{1}{2}\big(E[Y_1] + E[Y_n]\big) = \tfrac{1}{2}(\mu + \mu) = \mu. \end{aligned}$$

Variances:

$$\begin{aligned} \mathrm{Var}(Y_1) &= \sigma^2, \\ \mathrm{Var}(\tilde{Y}) &= \big(\tfrac{1}{2}\big)^2 \mathrm{Var}(Y_1 + Y_n) = \tfrac{1}{4}\big[\mathrm{Var}(Y_1) + \mathrm{Var}(Y_n) + 2\,\mathrm{Cov}(Y_1, Y_n)\big] = \tfrac{1}{4}\big[\sigma^2 + \sigma^2 + 0\big] = \frac{\sigma^2}{2}, \\ \mathrm{Var}(\bar{Y}) &= \big(\tfrac{1}{n}\big)^2 \mathrm{Var}\Big(\sum_{i=1}^{n} Y_i\Big) = \frac{1}{n^2}\Big[\sum_{i=1}^{n} \mathrm{Var}(Y_i) + 2\sum_{i<j} \mathrm{Cov}(Y_i, Y_j)\Big] = \frac{1}{n^2}\big[n\sigma^2 + 0\big] = \frac{\sigma^2}{n}. \end{aligned}$$

Compare through variance ratios:

$$\frac{\mathrm{Var}(\tilde{Y})}{\mathrm{Var}(\bar{Y})} = \frac{\sigma^2/2}{\sigma^2/n} = \frac{n}{2} > 1 \iff n > 2, \qquad \frac{\mathrm{Var}(Y_1)}{\mathrm{Var}(\tilde{Y})} = \frac{\sigma^2}{\sigma^2/2} = 2 > 1.$$

For $n > 2$: $\mathrm{Var}(\bar{Y}) < \mathrm{Var}(\tilde{Y}) < \mathrm{Var}(Y_1)$, so $\bar{Y}$ is the most efficient of the three.

Method 3: Excess variance over $\bar{Y}$ via the decomposition $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}(a_i - \frac{1}{n})^2 + \frac{1}{n}$

Step 2 of the proof of Theorem 5.2 gives, for any weights with $\sum_{i=1}^{n} a_i = 1$,

$$\mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2 + \frac{\sigma^2}{n}, \qquad \text{so} \qquad \mathrm{Var}(\hat{\mu}) - \mathrm{Var}(\bar{Y}) = \sigma^2 \sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2.$$

For Y_1 (one weight equal to 1, $n-1$ weights equal to 0):

$$\sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2 = \Big(1 - \tfrac{1}{n}\Big)^2 + (n-1)\Big(\tfrac{1}{n}\Big)^2 = \frac{(n-1)^2}{n^2} + \frac{n-1}{n^2} = \frac{(n-1)(n-1+1)}{n^2} = \frac{n-1}{n}.$$

For $\tilde{Y}$ (two weights equal to $\frac{1}{2}$, $n-2$ weights equal to 0), using $\frac{1}{2} - \frac{1}{n} = \frac{n-2}{2n}$:

$$\sum_{i=1}^{n} \Big(a_i - \tfrac{1}{n}\Big)^2 = 2\Big(\frac{n-2}{2n}\Big)^2 + (n-2)\Big(\tfrac{1}{n}\Big)^2 = \frac{(n-2)^2}{2n^2} + \frac{2(n-2)}{2n^2} = \frac{(n-2)(n-2+2)}{2n^2} = \frac{n-2}{2n}.$$

Hence

$$\mathrm{Var}(\tilde{Y}) - \mathrm{Var}(\bar{Y}) = \sigma^2 \frac{n-2}{2n} > 0 \iff n > 2,$$

and

$$\mathrm{Var}(Y_1) - \mathrm{Var}(\tilde{Y}) = \sigma^2\Big[\frac{n-1}{n} - \frac{n-2}{2n}\Big] = \sigma^2 \frac{2(n-1)-(n-2)}{2n} = \sigma^2 \frac{n}{2n} = \frac{\sigma^2}{2} > 0.$$

For $n > 2$: $\mathrm{Var}(\bar{Y}) < \mathrm{Var}(\tilde{Y}) < \mathrm{Var}(Y_1)$, so $\bar{Y}$ is the most efficient of the three.

Worked Example (Lecture 5 / Tutorial 5 Q1d): simulation evidence for $\bar{Y}$'s efficiency

To make the efficiency comparison concrete, 20 independent random samples of size $n = 10$ were drawn from $N(\mu = 2, \sigma^2 = 1)$. For each sample, two estimators of μ were computed: Y_1 (the first draw) and $\bar{Y} = \frac{1}{10}(Y_1 + \cdots + Y_{10})$.

- *Spread.* Y_1 ranges over $[-0.64, 4.27]$ (width 4.91); $\bar{Y}$ over $[1.16, 2.58]$ (width 1.42).
- *Closeness.* $\bar{Y}$ is closer to $\mu = 2$ than Y_1 in 16 of 20 samples.
- *Centring.* Means over the 20 samples: $Y_1 \approx 1.89$, $\bar{Y} \approx 1.96$, both near $\mu = 2$ (both unbiased).

Method 1: Sampling-variance formula $\mathrm{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2$ (Theorem 5.2)
Here $n = 10$, $\mu = 2$, $\sigma^2 = 1$. The weights are $a_1 = 1$, $a_2 = \cdots = a_{10} = 0$ for Y_1, and $a_i = \frac{1}{10}$ for every i for $\bar{Y}$. Both weight sets sum to 1:

$$1 + 9 \cdot 0 = 1, \qquad 10 \cdot \tfrac{1}{10} = 1,$$

so by Theorem 5.1, $E[Y_1] = E[\bar{Y}] = \mu = 2$. The variances are

$$\mathrm{Var}(Y_1) = \sigma^2\big[1^2 + 9 \cdot 0^2\big] = 1 \cdot 1 = 1,$$
$$\mathrm{Var}(\bar{Y}) = \sigma^2\Big[10 \cdot \big(\tfrac{1}{10}\big)^2\Big] = 1 \cdot \frac{10}{100} = 0.1,$$

so

$$\frac{\mathrm{Var}(Y_1)}{\mathrm{Var}(\bar{Y})} = \frac{1}{0.1} = 10, \qquad \frac{\sqrt{\mathrm{Var}(Y_1)}}{\sqrt{\mathrm{Var}(\bar{Y})}} = \frac{1}{\sqrt{0.1}} = \sqrt{10} \approx 3.16.$$

The simulation evidence reads as follows (all values in the units of Y).

- Both estimators are unbiased, so both sampling distributions are centred at $\mu = 2$; the 20-sample averages 1.89 and 1.96 miss 2 by 0.11 and 0.04 only because 20 samples is a finite number of repetitions.
- The variance of $\bar{Y}$ is 0.1, one tenth of the variance 1 of Y_1, so the standard deviation of $\bar{Y}$ is $\sqrt{0.1} \approx 0.316$ against 1 for Y_1.
- The observed ranges have widths $4.27 - (-0.64) = 4.91$ and $2.58 - 1.16 = 1.42$, a ratio of $4.91/1.42 \approx 3.46$, close to the theoretical standard-deviation ratio $\sqrt{10} \approx 3.16$.
- Because $\bar{Y}$ is more tightly concentrated around 2, it lands closer to μ than Y_1 in most samples: 16 of 20, or 80%, in this simulation.

$\boxed{\text{Var}(Y_1) = 1, \quad \text{Var}(\bar{Y}) = 0.1;\ \text{both unbiased, but } \bar{Y} \text{ is ten times more efficient at } n = 10.}$

Method 2: Variance rules for a sum of independent draws
Each draw satisfies $Y_i \sim N(2, 1)$, so directly $E[Y_1] = 2$ and $\text{Var}(Y_1) = 1$. For the sample mean, use linearity of expectation:

$$E[\bar{Y}] = E\Big[\tfrac{1}{10}\sum_{i=1}^{10} Y_i\Big] = \tfrac{1}{10}\sum_{i=1}^{10} E[Y_i] = \tfrac{1}{10} \cdot 10 \cdot 2 = 2.$$

For the variance, pull out the constant as its square, then expand the variance of the sum; every covariance is 0 because the ten draws are independent:

$$\text{Var}(\bar{Y}) = \Big(\tfrac{1}{10}\Big)^2 \text{Var}\Big(\sum_{i=1}^{10} Y_i\Big) = \frac{1}{100}\Big[\sum_{i=1}^{10} \text{Var}(Y_i) + 2\sum_{i<j} \text{Cov}(Y_i, Y_j)\Big] = \frac{1}{100}\big[10 \cdot 1 + 0\big] = 0.1.$$

Both estimators have mean $2 = \mu$, and the variance ratio is $\text{Var}(Y_1)/\,\text{Var}(\bar{Y}) = 1/0.1 = 10$.

$\boxed{\text{Var}(Y_1) = 1, \quad \text{Var}(\bar{Y}) = 0.1;\ \text{both unbiased, but } \bar{Y} \text{ is ten times more efficient at } n = 10.}$

5.3 Application: Comparing Two Independent Sample Means

The same weighted-sum logic extends immediately to comparing means from *two* independent samples — a common empirical task (e.g., comparing an outcome across two countries, or a treatment and control group).

Worked Example (Tutorial 5 Q3): weekly working hours, Singapore vs. Hong Kong

Weekly working hours in Singapore (SG) have population mean μ_1 and variance σ_1^2; N_1 i.i.d. observations are drawn from this population. Weekly working hours in Hong Kong (HK) have population mean μ_2 and variance σ_2^2; N_2 i.i.d. observations are drawn from this population, and the HK sample is drawn *independently* of the SG sample. Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the two sample means, and let $\hat{d} = \bar{Y}_1 - \bar{Y}_2$ estimate $\mu_1 - \mu_2$.
Write the SG observations as Y_{1i}, $i = 1, \ldots, N_1$, and the HK observations as Y_{2j}, $j = 1, \ldots, N_2$, so that $\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_{1i}$ and $\bar{Y}_2 = \frac{1}{N_2}\sum_{j=1}^{N_2} Y_{2j}$. All quantities are in hours per week; variances are in (hours per week)^2.

Method 1: Weights theorems for each mean, then the variance of a difference
(a) Mean and variance of each sample average. $\bar{Y}_1$ is a linear estimator with weights $a_i = 1/N_1$ for $i = 1, \ldots, N_1$. The weights sum to

$$\sum_{i=1}^{N_1} \frac{1}{N_1} = N_1 \cdot \frac{1}{N_1} = 1,$$

so by Theorems 5.1–5.2 (applied to the SG population, mean μ_1, variance σ_1^2):

$$E[\bar{Y}_1] = \mu_1 \sum_{i=1}^{N_1} \frac{1}{N_1} = \mu_1 \cdot 1 = \mu_1,$$

$$\text{Var}(\bar{Y}_1) = \sigma_1^2 \sum_{i=1}^{N_1} \Big(\tfrac{1}{N_1}\Big)^2 = \sigma_1^2 \cdot N_1 \cdot \frac{1}{N_1^2} = \frac{\sigma_1^2}{N_1}.$$

$\bar{Y}_2$ has weights $1/N_2$ for $j = 1, \ldots, N_2$, which sum to $N_2 \cdot \frac{1}{N_2} = 1$; applied to the HK population (mean μ_2, variance σ_2^2):

$$E[\bar{Y}_2] = \mu_2 \sum_{j=1}^{N_2} \frac{1}{N_2} = \mu_2 \cdot 1 = \mu_2,$$

$$\mathrm{Var}(\bar{Y}_2) = \sigma_2^2 \sum_{j=1}^{N_2} \Big(\tfrac{1}{N_2}\Big)^2 = \sigma_2^2 \cdot N_2 \cdot \frac{1}{N_2^2} = \frac{\sigma_2^2}{N_2}.$$

(b) Mean and variance of $\hat{d}$. By linearity of expectation and part (a),

$$E[\hat{d}] = E[\bar{Y}_1 - \bar{Y}_2] = E[\bar{Y}_1] - E[\bar{Y}_2] = \mu_1 - \mu_2.$$

$\bar{Y}_1$ is a function of the SG observations only and $\bar{Y}_2$ is a function of the HK observations only; the two samples are drawn independently, so $\bar{Y}_1$ and $\bar{Y}_2$ are independent and

$$\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2) = 0.$$

Apply $\mathrm{Var}(W - V) = \mathrm{Var}(W) + (-1)^2 \mathrm{Var}(V) + 2(1)(-1)\,\mathrm{Cov}(W, V)$ with $W = \bar{Y}_1$, $V = \bar{Y}_2$:

$$\begin{aligned}\mathrm{Var}(\hat{d}) &= \mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_2) - 2\,\mathrm{Cov}(\bar{Y}_1, \bar{Y}_2) \\ &= \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} - 2 \cdot 0 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}.\end{aligned}$$

$$\boxed{E[\hat{d}] = \mu_1 - \mu_2, \qquad \mathrm{Var}(\hat{d}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}.}$$

Method 2: $\hat{d}$ as one linear estimator on the pooled sample

Stack the $N_1 + N_2$ observations into one list. Then $\hat{d}$ is a single weighted sum with weight $+\frac{1}{N_1}$ on each SG observation and $-\frac{1}{N_2}$ on each HK observation:

$$\hat{d} = \sum_{i=1}^{N_1} \frac{1}{N_1} Y_{1i} + \sum_{j=1}^{N_2} \Big(-\frac{1}{N_2}\Big) Y_{2j}.$$

Take expectations term by term, using $E[Y_{1i}] = \mu_1$ and $E[Y_{2j}] = \mu_2$:

$$E[\hat{d}] = \sum_{i=1}^{N_1} \frac{1}{N_1} \mu_1 - \sum_{j=1}^{N_2} \frac{1}{N_2} \mu_2 = N_1 \cdot \frac{\mu_1}{N_1} - N_2 \cdot \frac{\mu_2}{N_2} = \mu_1 - \mu_2.$$

All $N_1 + N_2$ observations are mutually independent (i.i.d. within each country, and the two samples independent of each other), so every covariance between two different observations is 0, and the variance of the weighted sum is the sum of squared weights times variances:

$$\begin{aligned}\mathrm{Var}(\hat{d}) &= \sum_{i=1}^{N_1} \Big(\frac{1}{N_1}\Big)^2 \mathrm{Var}(Y_{1i}) + \sum_{j=1}^{N_2} \Big(-\frac{1}{N_2}\Big)^2 \mathrm{Var}(Y_{2j}) \\ &= N_1 \cdot \frac{\sigma_1^2}{N_1^2} + N_2 \cdot \frac{\sigma_2^2}{N_2^2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}.\end{aligned}$$

The minus sign on the HK weights disappears on squaring, which is why the two sampling variances add even though the means are subtracted.

$$\boxed{E[\hat{d}] = \mu_1 - \mu_2, \qquad \mathrm{Var}(\hat{d}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}.}$$

5.4 The Gauss–Markov Theorem

Sections above established that $\bar{Y}$ is BLUE for the population mean μ. The Gauss–Markov Theorem extends exactly this idea to the OLS regression estimators $\hat{\beta}_0, \hat{\beta}_1$: among all linear *and* unbiased estimators of β_0, β_1, OLS has the smallest variance.

Theorem 5.3: Gauss–Markov Theorem

Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ under Least Squares Assumptions 1–3:

- **Assumption 1:** $E[u_i \mid X_i] = 0$;
- **Assumption 2:** (X_i, Y_i), $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution;
- **Assumption 3:** large outliers are unlikely (finite fourth moments);

together with the additional technical condition of *homoskedasticity*,

$$\mathrm{Var}(u_i \mid X_i) = \sigma_u^2 \quad \text{for every } i$$

(constant conditional variance of the error term). Then the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the *Best Linear Unbiased Estimators* (BLUE) of β_0 and β_1: for any other linear unbiased estimators $\tilde{\beta}_0, \tilde{\beta}_1$ constructed from $(Y_1, \ldots, Y_n)$,

$$\mathrm{Var}(\hat{\beta}_0) \le \mathrm{Var}(\tilde{\beta}_0), \qquad \mathrm{Var}(\hat{\beta}_1) \le \mathrm{Var}(\tilde{\beta}_1).$$

On the "technical assumptions." The lecture slides state this result as holding under "the OLS assumptions and some other technical assumptions," without listing the latter explicitly. The technical condition that makes the standard proof go through (and the one used below) is homoskedasticity, $\mathrm{Var}(u_i \mid X_i) = \sigma_u^2$ for all i; we make this explicit here as a refinement of the slides' informal phrasing. Without homoskedasticity, OLS remains unbiased but is generally *not* the most efficient linear unbiased estimator (a weighted least squares estimator can do better) — this is why Gauss–Markov is stated as a conditional result.
The full proof (Tutorial 5 Q1b asks only for the statement) mirrors the $\bar{Y}$-is-BLUE argument almost exactly, replacing constant weights $1/n$ with weights that depend on X_i. It is non-examinable and set out in the extension box below.

Extension: $\hat{\beta}_1$ as a linear estimator, and its unbiasedness (non-examinable)

Recall from the derivation of the OLS estimator that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Write $\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i$ with weights

$$w_i = \frac{X_i - \bar{X}}{S_{XX}}, \qquad S_{XX} := \sum_{i=1}^n (X_j - \bar{X})^2.$$

(This is a valid rewriting because $\sum_{i=1}^n (X_i - \bar{X})\bar{Y} = \bar{X} \cdot 0 \cdot \ldots$ contributes nothing extra — formally, $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y}\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})Y_i$, since $\sum_{i=1}^n (X_i - \bar{X}) = 0$.)

Two weight identities. Using $\sum_{i=1}^n X_i = n\bar{X}$,

$$\sum_{i=1}^n w_i = \frac{1}{S_{XX}} \sum_{i=1}^n (X_i - \bar{X}) = \frac{1}{S_{XX}} (n\bar{X} - n\bar{X}) = 0.$$

Adding and subtracting $\bar{X}$ inside the sum,

$$\sum_{i=1}^n w_i X_i = \frac{1}{S_{XX}} \sum_{i=1}^n (X_i - \bar{X})X_i = \frac{1}{S_{XX}} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{S_{XX}}{S_{XX}} = 1$$

(using $\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})^2$ and $\bar{X}\sum_{i=1}^n (X_i - \bar{X}) = 0$).

(a) Unbiasedness of $\hat{\beta}_1$ given X. Substituting $Y_i = \beta_0 + \beta_1 X_i + u_i$,

$$\hat{\beta}_1 = \sum_{i=1}^n w_i Y_i = \sum_{i=1}^n w_i(\beta_0 + \beta_1 X_i + u_i) = \beta_0 \sum_{i=1}^n w_i + \beta_1 \sum_{i=1}^n w_i X_i + \sum_{i=1}^n w_i u_i = \beta_1 + \sum_{i=1}^n w_i u_i$$

(using the two weight identities just derived). Taking $E[\cdot \,|\, X_1, \ldots, X_n]$ and using Assumption 1 together with Assumption 2 (so $E[u_i \,|\, X_1, \ldots, X_n] = E[u_i \,|\, X_i] = 0$):

$$E[\hat{\beta}_1 \,|\, X] = \beta_1 + \sum_{i=1}^n w_i \, E[u_i \,|\, X] = \beta_1 + 0 = \beta_1.$$

So $\hat{\beta}_1$ is (conditionally, hence unconditionally) unbiased for β_1.

Extension: Proof that $\hat{\beta}_1$ is BLUE (non-examinable, continued)

(b) Any competitor must satisfy the same identities. Let $\tilde{\beta}_1 = \sum_{i=1}^n c_i Y_i$ be any other linear estimator, with weights w_i as constructed above for $\hat{\beta}_1$. By the identical substitution as in part (a), $\tilde{\beta}_1 = \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i X_i + \sum_{i=1}^n c_i u_i$; for $E[\tilde{\beta}_1 \mid X] = \beta_1$ to hold for *every* value of β_0 and β_1 (not just the true ones), the coefficients on β_0 and β_1 must match 0 and 1 exactly:

$$\sum_{i=1}^n c_i = 0, \qquad \sum_{i=1}^n c_i X_i = 1.$$

(c) Variance comparison. Let $d_i := c_i - w_i$. From (b) and the weight identities,

$$\sum_{i=1}^n d_i = \sum_{i=1}^n c_i - \sum_{i=1}^n w_i = 0 - 0 = 0, \qquad \sum_{i=1}^n d_i X_i = \sum_{i=1}^n c_i X_i - \sum_{i=1}^n w_i X_i = 1 - 1 = 0.$$

Expand $c_i^2 = (w_i + d_i)^2 = w_i^2 + 2 w_i d_i + d_i^2$ and sum:

$$\sum_{i=1}^n c_i^2 = \sum_{i=1}^n w_i^2 + 2 \sum_{i=1}^n w_i d_i + \sum_{i=1}^n d_i^2.$$

The cross term vanishes: using $w_i = (X_i - \bar{X})/S_{XX}$,

$$\sum_{i=1}^n w_i d_i = \frac{1}{S_{XX}} \sum_{i=1}^n (X_i - \bar{X}) d_i = \frac{1}{S_{XX}} \Big[\sum_{i=1}^n X_i d_i - \bar{X} \sum_{i=1}^n d_i \Big] = \frac{1}{S_{XX}} [0 - 0] = 0.$$

Hence

$$\sum_{i=1}^n c_i^2 = \sum_{i=1}^n w_i^2 + \sum_{i=1}^n d_i^2 \geq \sum_{i=1}^n w_i^2,$$

with equality if and only if $d_i = 0$ for every i, i.e., $c_i = w_i$ for every i.
By the same algebra as in (a), $\tilde{\beta}_1 - \beta_1 = \sum_{i=1}^n c_i u_i$, so, conditioning on X and using Assumption 2 (independence across i, so covariances between different u_i's are zero) together with homoskedasticity ($\mathrm{Var}(u_i \mid X_i) = \sigma_u^2$ for every i),

$$\mathrm{Var}(\tilde{\beta}_1 \mid X) = \mathrm{Var}\Big(\sum_{i=1}^n c_i u_i \,\Big|\, X \Big) = \sum_{i=1}^n c_i^2 \,\mathrm{Var}(u_i \mid X_i) = \sigma_u^2 \sum_{i=1}^n c_i^2 \geq \sigma_u^2 \sum_{i=1}^n w_i^2 = \mathrm{Var}(\hat{\beta}_1 \mid X).$$

Thus $\mathrm{Var}(\hat{\beta}_1 \mid X) \leq \mathrm{Var}(\tilde{\beta}_1 \mid X)$ for every linear unbiased $\tilde{\beta}_1$, i.e., $\hat{\beta}_1$ is BLUE. An analogous (more tedious, since $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ mixes both weight sets) argument establishes the same conclusion for $\hat{\beta}_0$.

Exam Focus: Gauss–Markov

- State the theorem precisely: it compares OLS only against *linear and unbiased* competitors, and it requires Assumptions 1–3 *plus* homoskedasticity.
- Be ready to reproduce the $\bar{Y}$-is-BLUE proof (Theorem 5.2) in full — this is the version of the argument the course actually expects you to derive from scratch (Tutorial 5 Q1c); the OLS version in the extension box is conceptually identical but marked non-examinable.
- Do not confuse "BLUE" with "the estimator with the smallest possible variance among *all* estimators" — a nonlinear or biased estimator could in principle have smaller variance; Gauss–Markov only rules out better *linear unbiased* alternatives.

5.5 Regression with a Binary Regressor

Every regressor considered so far (test scores, STR, CEO ROE) has been continuous. Regressors can also be *binary* — taking only the values 0 or 1 — known as a *dummy* or *indicator* variable. Examples: a worker's gender ($D_i = 1$ if female, 0 if male), whether a school district is urban ($D_i = 1$) or rural ($D_i = 0$), or whether a class is small ($D_i = 1$) or large ($D_i = 0$).

Definition 5.4: Simple regression with a binary regressor

Consider

$$\text{wage}_i = \beta_0 + \beta_1 D_i + u_i, \qquad D_i = \begin{cases} 1 & \text{if person } i \text{ is female,} \\ 0 & \text{if person } i \text{ is male,} \end{cases}$$

for $i = 1, \ldots, n$, where Assumption 1 holds: $E[u_i \mid D_i] = 0$.

Theorem 5.4: Interpretation of β_0, β_1 under a binary regressor

Under the model of Definition 5.4,

$$\beta_0 = E[\text{wage}_i \mid D_i = 0], \qquad \beta_1 = E[\text{wage}_i \mid D_i = 1] - E[\text{wage}_i \mid D_i = 0].$$

Proof. Take the conditional expectation of $\text{wage}_i = \beta_0 + \beta_1 D_i + u_i$ given $D_i = 0$:

$$E[\text{wage}_i \mid D_i = 0] = \beta_0 + \beta_1(0) + E[u_i \mid D_i = 0] = \beta_0 + 0 = \beta_0$$

(using Assumption 1, $E[u_i \mid D_i] = 0$, evaluated at $D_i = 0$). Now condition on $D_i = 1$ instead:

$$E[\text{wage}_i \mid D_i = 1] = \beta_0 + \beta_1(1) + E[u_i \mid D_i = 1] = \beta_0 + \beta_1$$

(again by Assumption 1, now evaluated at $D_i = 1$). Subtracting the first result from the second,

$$E[\text{wage}_i \mid D_i = 1] - E[\text{wage}_i \mid D_i = 0] = (\beta_0 + \beta_1) - \beta_0 = \beta_1. \qquad \square$$

So β_0 is the average wage in the *base group* ($D_i = 0$, males) and β_1 is the difference in average wages between the group coded $D_i = 1$ (females) and the base group — the coefficient on a binary regressor is a *group mean difference*. Chapter 3 derived the same equality, $\beta_1 = E[Y \mid D = 1] - E[Y \mid D = 0]$, for a randomly assigned treatment indicator in the context of RCTs and the average treatment effect; the derivation here is identical in form but does not require random assignment of D_i — it is a statement about the population regression function, valid whenever Assumption 1 holds for whatever D_i happens to be.

The population interpretation in Theorem 5.4 has an exact sample-data counterpart: OLS applied to a binary regressor recovers the two sample group means directly.

Theorem 5.5: OLS with a binary regressor

Let n_0 observations have $D_i = 0$ and $n_1 = n - n_0$ have $D_i = 1$. The OLS estimates of the model in Definition 5.4 are

$$\hat{\beta}_0 = \bar{Y}_{D=0}, \qquad \hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0},$$

where $\bar{Y}_{D=0}$ and $\bar{Y}_{D=1}$ are the sample average of wage_i within each group.

Proof. OLS minimises the sum of squared residuals, which splits into a sum over the two groups since every observation has $D_i = 0$ or $D_i = 1$:

$$\text{SSR}(b_0, b_1) = \sum_{i=1}^{n} (\text{wage}_i - b_0 - b_1 D_i)^2 = \sum_{i:D_i=0} (\text{wage}_i - b_0)^2 + \sum_{i:D_i=1} (\text{wage}_i - b_0 - b_1)^2.$$

Differentiate with respect to b_1 and set the first-order condition to zero; only the second (group-1) sum involves b_1:

$$\frac{\partial \,\mathrm{SSR}}{\partial b_1} = -2 \sum_{i:D_i=1} (\mathrm{wage}_i - b_0 - b_1) = 0 \quad \Longrightarrow \quad \sum_{i:D_i=1} (\mathrm{wage}_i - b_0 - b_1) = 0.$$

Dividing by n_1 and rearranging,

$$b_0 + b_1 = \frac{1}{n_1} \sum_{i:D_i=1} \mathrm{wage}_i = \bar{Y}_{D=1}. \tag{i}$$

Differentiate with respect to b_0 and set the first-order condition to zero:

$$\frac{\partial \,\mathrm{SSR}}{\partial b_0} = -2 \sum_{i:D_i=0} (\mathrm{wage}_i - b_0) - 2 \sum_{i:D_i=1} (\mathrm{wage}_i - b_0 - b_1) = 0.$$

The second sum is exactly zero by (i) (the first-order condition in b_1 already forces it to vanish), leaving

$$\sum_{i:D_i=0} (\mathrm{wage}_i - b_0) = 0 \quad \Longrightarrow \quad b_0 = \frac{1}{n_0} \sum_{i:D_i=0} \mathrm{wage}_i = \bar{Y}_{D=0}. \tag{ii}$$

So $\hat{\beta}_0 = \bar{Y}_{D=0}$. Substituting (ii) into (i),

$$\hat{\beta}_1 = \bar{Y}_{D=1} - \hat{\beta}_0 = \bar{Y}_{D=1} - \bar{Y}_{D=0}. \qquad \square$$

Worked Example (Lecture 5): test scores and a small-class indicator

Define $D_i = 1$ if the student-teacher ratio (STR) in district i is less than 20, and $D_i = 0$ otherwise. Regressing test scores on D_i using the California data set yields

$$\widehat{\mathrm{TestScore}} = 650.0 + 7.4D.$$

Method 1: Group means (Theorem 5.5)

Theorem 5.5 gives $\hat{\beta}_0 = \bar{Y}_{D=0}$ and $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$. The base group ($D = 0$, STR ≥ 20) therefore has sample mean

$$\bar{Y}_{D=0} = \hat{\beta}_0 = 650.0.$$

Adding the two coefficients recovers the other group mean:

$$\hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_{D=0} + \left(\bar{Y}_{D=1} - \bar{Y}_{D=0}\right) = \bar{Y}_{D=1},$$

so the small-class group ($D = 1$, STR < 20) has sample mean

$$\bar{Y}_{D=1} = 650.0 + 7.4 = 657.4.$$

The same two numbers are the fitted values at $D = 0$ and $D = 1$:

$$\widehat{\mathrm{TestScore}}\big|_{D=0} = 650.0 + 7.4 \cdot 0 = 650.0, \qquad \widehat{\mathrm{TestScore}}\big|_{D=1} = 650.0 + 7.4 \cdot 1 = 657.4,$$

and their difference is $657.4 - 650.0 = 7.4 = \hat{\beta}_1$.

- $\hat{\beta}_0 = 650.0$ points is the sample average test score of districts with STR ≥ 20, the base group coded $D = 0$.
- $\hat{\beta}_0 + \hat{\beta}_1 = 657.4$ points is the sample average test score of districts with STR < 20, the group coded $D = 1$.

- $\hat{\beta}_1 = 7.4$ points is the sample difference in group means, small-class districts minus large-class districts; it estimates the population difference $\beta_1 = E[\text{TestScore} \mid D = 1] - E[\text{TestScore} \mid D = 0]$ of Theorem 5.4.
- The regression on its own reports a difference in averages; reading 7.4 points as the causal effect of moving a district below STR = 20 requires that D_i be unrelated to the other determinants of test scores in u_i.

Average score: 650.0 (STR $\geq$ 20) vs. 657.4 (STR $<$ 20);
gap $= \hat{\beta}_1 = 7.4$ points higher for small classes.

Method 2: OLS slope and intercept formulas with a 0/1 regressor

Let n_1 districts have $D_i = 1$ and $n_0 = n - n_1$ have $D_i = 0$; both groups are non-empty in the data, so $n_0 \geq 1$ and $n_1 \geq 1$. Then

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{n_1}{n}, \qquad 1 - \bar{D} = \frac{n - n_1}{n} = \frac{n_0}{n}.$$

Denominator of $\hat{\beta}_1$. Split the sum by group ($D_i - \bar{D} = 1 - \bar{D}$ when $D_i = 1$, and $= -\bar{D}$ when $D_i = 0$):

$$\sum_{i=1}^{n}(D_i - \bar{D})^2 = n_1(1 - \bar{D})^2 + n_0\bar{D}^2 = n_1\frac{n_0^2}{n^2} + n_0\frac{n_1^2}{n^2} = \frac{n_0 n_1(n_0 + n_1)}{n^2} = \frac{n_0 n_1}{n}.$$

Numerator of $\hat{\beta}_1$. Since $\sum_{i=1}^{n}(D_i - \bar{D}) = n\bar{D} - n\bar{D} = 0$,

$$\sum_{i=1}^{n}(D_i - \bar{D})(Y_i - \bar{Y}) = \sum_{i=1}^{n}(D_i - \bar{D})Y_i - \bar{Y}\sum_{i=1}^{n}(D_i - \bar{D}) = \sum_{i=1}^{n}(D_i - \bar{D})Y_i.$$

Split by group, using $\sum_{i:D_i=1} Y_i = n_1\bar{Y}_{D=1}$ and $\sum_{i:D_i=0} Y_i = n_0\bar{Y}_{D=0}$:

$$\sum_{i=1}^{n}(D_i - \bar{D})Y_i = \frac{n_0}{n}\, n_1\bar{Y}_{D=1} - \frac{n_1}{n}\, n_0\bar{Y}_{D=0} = \frac{n_0 n_1}{n}\big(\bar{Y}_{D=1} - \bar{Y}_{D=0}\big).$$

Slope. Divide numerator by denominator, which is allowed because $\frac{n_0 n_1}{n} > 0$:

$$\hat{\beta}_1 = \frac{\frac{n_0 n_1}{n}\big(\bar{Y}_{D=1} - \bar{Y}_{D=0}\big)}{\frac{n_0 n_1}{n}} = \bar{Y}_{D=1} - \bar{Y}_{D=0}.$$

Intercept. The overall mean is the share-weighted average of the group means, $\bar{Y} = \frac{n_1\bar{Y}_{D=1} + n_0\bar{Y}_{D=0}}{n} = \bar{D}\,\bar{Y}_{D=1} + (1 - \bar{D})\bar{Y}_{D=0}$, so

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{D} = \bar{D}\,\bar{Y}_{D=1} + (1 - \bar{D})\bar{Y}_{D=0} - \bar{D}\,\bar{Y}_{D=1} + \bar{D}\,\bar{Y}_{D=0} = (1 - \bar{D} + \bar{D})\bar{Y}_{D=0} = \bar{Y}_{D=0}.$$

Match these to the reported estimates $\hat{\beta}_0 = 650.0$ and $\hat{\beta}_1 = 7.4$:

$$\bar{Y}_{D=0} = 650.0, \qquad \bar{Y}_{D=1} = \bar{Y}_{D=0} + \hat{\beta}_1 = 650.0 + 7.4 = 657.4, \qquad \bar{Y}_{D=1} - \bar{Y}_{D=0} = 7.4 \text{ points}.$$

Average score: 650.0 (STR $\geq$ 20) vs. 657.4 (STR $<$ 20);
gap $= \hat{\beta}_1 = 7.4$ points higher for small classes.

5.6 Common Slope, Different Intercepts

A binary regressor can also be added *alongside* a continuous regressor, allowing two groups to share the same slope but have different intercepts.

Theorem 5.6: Population regression lines with a common slope

Consider

$$Y_i = \beta_0 + \delta_0 D_i + \beta_1 X_i + u_i, \qquad i = 1, \dots, n,$$

where $D_i \in \{0, 1\}$ is a dummy variable and X_i is continuous. Then, taking $E[\,\cdot \mid D_i, X_i]$ and using $E[u_i \mid D_i, X_i] = 0$:

$$D_i = 0: \quad E[Y_i \mid D_i = 0, X_i] = \beta_0 + \beta_1 X_i; \qquad D_i = 1: \quad E[Y_i \mid D_i = 1, X_i] = (\beta_0 + \delta_0) + \beta_1 X_i.$$

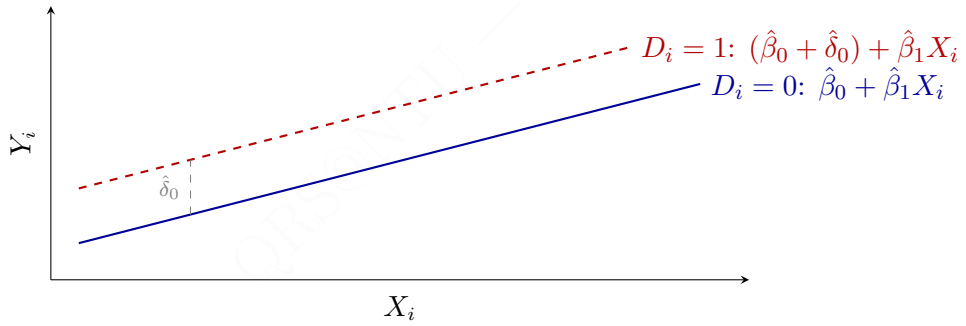
Proof. Substitute $D_i = 0$ into the model and take conditional expectations:

$$E[Y_i \mid D_i = 0, X_i] = \beta_0 + \delta_0(0) + \beta_1 X_i + E[u_i \mid D_i = 0, X_i] = \beta_0 + \beta_1 X_i.$$

Substitute $D_i = 1$ instead:

$$E[Y_i \mid D_i = 1, X_i] = \beta_0 + \delta_0(1) + \beta_1 X_i + E[u_i \mid D_i = 1, X_i] = (\beta_0 + \delta_0) + \beta_1 X_i. \qquad \square$$

Both group-specific lines have identical slope β_1 (the effect of X on Y is assumed the same in both groups) but differ in intercept by δ_0: the two population regression lines are *parallel*, vertically shifted by δ_0. For the California data set, D_i might indicate whether district i lies in Region A ($D_i = 1$) versus Region B ($D_i = 0$); δ_0 then measures the region gap in test scores that is *not* explained by STR.



It is possible to enrich this model further by including additional dummy variables — one per extra category — while keeping a single continuous slope, as the next application shows.

5.7 Multiple Dummy Variables: The Law School Ranking Model

Policy Application: law school rankings, GPA, and starting salary

The data set `lawsch_ranking.dta` records, for each of a sample of U.S. law schools, the average starting salary of its graduates, the school's ranking (1–142), and the average college GPA of its students. Consider

$$\text{salary}_i = \beta_0 + \beta_1\,\text{top10}_i + \beta_2\,\text{r_11_25}_i + \beta_3\,\text{r_26_50}_i + \beta_4\,\text{GPA}_i + u_i,$$

where the three dummy variables are defined as

$$\text{top10}_i = \begin{cases} 1 & \text{school ranked 1–10} \\ 0 & \text{otherwise} \end{cases},$$

$$\text{r_11_25}_i = \begin{cases} 1 & \text{ranked 11–25} \\ 0 & \text{otherwise} \end{cases},$$

$$\text{r_26_50}_i = \begin{cases} 1 & \text{ranked 26–50} \\ 0 & \text{otherwise} \end{cases}.$$

Because at most one dummy is "switched on" for any school, and *none* is switched on for a school ranked below 50, the base group — the omitted category with all three dummies equal to zero — is schools ranked below 50th. Substituting each combination of $(\text{top10}_i, \text{r_11_25}_i, \text{r_26_50}_i)$ into the model (the same substitution technique used for Theorem 5.6) and using $E[u_i \mid \cdot] = 0$ gives the population regression function within each group:

$$\begin{aligned} \text{Top 10:} \quad & E[\text{salary}_i \mid \cdot] = (\beta_0 + \beta_1) + \beta_4\,\text{GPA}_i, \\ \text{11th–25th:} \quad & E[\text{salary}_i \mid \cdot] = (\beta_0 + \beta_2) + \beta_4\,\text{GPA}_i, \\ \text{26th–50th:} \quad & E[\text{salary}_i \mid \cdot] = (\beta_0 + \beta_3) + \beta_4\,\text{GPA}_i, \\ \text{Below 50th (base):} \quad & E[\text{salary}_i \mid \cdot] = \beta_0 + \beta_4\,\text{GPA}_i. \end{aligned}$$

All four lines share the same slope β_4 on GPA; they differ only in intercept. For a Top-10 school, β_1 is precisely the gap in average starting salary relative to a below-50th school *with the same GPA* — the coefficient on each category dummy is always read as "versus the base group, holding the continuous regressor fixed," exactly as β_1 was the male–female wage gap in Theorem 5.4 and δ_0 was the region gap in Theorem 5.6.

Common Mistake: the dummy variable trap

It is tempting to add a fourth dummy, say $\text{below50}_i = 1 - \text{top10}_i - \text{r_11_25}_i - \text{r_26_50}_i$, to "cover" the base group explicitly, *while keeping the intercept β_0 in the model.* This is a mistake: for every observation,

$$\text{top10}_i + \text{r_11_25}_i + \text{r_26_50}_i + \text{below50}_i = 1,$$

which is an exact linear combination of the four dummies that reproduces the constant regressor. Including all four category dummies *and* an intercept creates perfect multicollinearity (the regressor matrix is not full rank), and OLS cannot separate the intercept from the redundant dummy's coefficient — there is no unique solution. The fix is always to omit one category (the *base* or *reference* group) and let the intercept β_0 absorb that group's mean; with k categories, include at most $k-1$ dummies alongside the intercept.

5.8 Setting Up Hypothesis Tests: A First Look

Unbiasedness and efficiency tell us that $\bar{Y}$ (or $\hat{\beta}_1$) is, on average, a good guess of the unknown parameter. The next question — taken up formally in Chapter 6 — is how to use that estimate to make a yes/no

decision about a claim regarding the parameter's true value. This section sets up the vocabulary.

Suppose we want to test whether the population mean μ equals some specific hypothesised value μ_0 (for instance, "college graduates earn \$20/hour on average" translates to $\mu_0 = 20$). We write

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_1 : \mu \neq \mu_0,$$

where H_0 is the *null hypothesis* and H_1 is the (two-sided) *alternative hypothesis* — "two-sided" because it allows μ to be either above or below μ_0, without committing to a direction.

We decide between H_0 and H_1 using the sample $Y_1, \ldots, Y_n$. Because $\bar{Y}$ is a good estimator of μ (unbiased, and BLUE by Theorem 5.2), if H_0 is true then $\bar{Y}$ should typically land close to μ_0; if H_0 is false, $\bar{Y}$ and μ_0 should typically differ by more. This motivates a decision rule of the form

$$\text{reject } H_0 \text{ if } |\bar{Y} - \mu_0| > c,$$

for some critical value c to be chosen. Chapter 1 introduced the resulting classification of outcomes:

	Accept H_0	**Reject** H_0
H_0 **true**	Correct decision	Type I error
H_1 **true**	Type II error	Correct decision

There is a trade-off between the two error types: shrinking the probability of a Type I error (a false rejection) enlarges the probability of a Type II error (a false acceptance), and vice versa — the two cannot be minimised simultaneously by moving c alone. The standard convention is to fix the probability of a Type I error, called the *significance level* of the test, at a prespecified level, typically 10%, 5%, or 1% in economics. In settings where a false rejection is especially costly — a legal test with H_0 : the defendant is not guilty, or a medical test with H_0 : the new drug has no effect — a much stricter significance level such as 0.1% is often used, reflecting a strong reluctance to reject H_0 incorrectly.

Exam Focus: the taxpayer's claim (teaser for Chapter 6)

An angry taxpayer tells the superintendent that cutting class size will not raise test scores, i.e., that the causal effect of STR on test scores is zero: $\beta_1 = 0$. In econometric language this is

$$H_0 : \beta_1 = 0 \qquad \text{versus} \qquad H_1 : \beta_1 \neq 0.$$

The California regression gave $\hat{\beta}_1 = -2.28$: a nonzero *sample* estimate, but on its own that is not enough to reject $H_0 : \beta_1 = 0$, because $\hat{\beta}_1$ is itself a random variable and would rarely equal exactly zero even if the true β_1 were zero. Deciding whether -2.28 is *statistically significantly* different from zero requires comparing it to its own sampling variability — the standard error $\text{SE}(\hat{\beta}_1)$ and the resulting t-statistic, which is exactly where Chapter 6 begins.

5.9 Chapter Summary

Key Formula: Chapter 5 recap

$$\hat{\mu} = \sum_{i=1}^{n} a_i Y_i \text{ unbiased} \iff \sum_{i=1}^{n} a_i = 1, \quad \text{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2 \;\geq\; \frac{\sigma^2}{n} = \text{Var}(\bar{Y}) \text{ } (\bar{Y} \text{ is BLUE}),$$

Gauss–Markov: $\text{Var}(\hat{\beta}_0) \leq \text{Var}(\tilde{\beta}_0)$, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$,
(Assumptions 1–3 + homoskedasticity)

Binary D_i : $\hat{\beta}_0 = \bar{Y}_{D=0}$, $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$.

Exam Focus: Chapter 5 takeaways

- Unbiasedness ($E[\hat{\theta}] = \theta$) and efficiency (smallest variance *among unbiased estimators*) are separate criteria; never compare variances across a biased and an unbiased estimator to declare a "winner."
- Be able to derive, from scratch, both the unbiasedness condition $\sum_{i=1}^{n} a_i = 1$ and the minimisation $\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}(a_i - 1/n)^2 + 1/n$ that makes $\bar{Y}$ BLUE — this is the single most re-usable proof technique in the chapter (it reappears, with X-dependent weights, in the non-examinable OLS BLUE proof).
- State Gauss–Markov precisely: OLS beats only linear, unbiased competitors, and needs homoskedasticity on top of Assumptions 1–3.
- With a binary regressor, $\hat{\beta}_0$ and $\hat{\beta}_0 + \hat{\beta}_1$ are literally the two group sample means; with k categories, use $k - 1$ dummies plus an intercept, never k (the dummy variable trap).
- The hypothesis-testing set-up here (H_0 vs. H_1, Type I/II errors, significance levels) is only the scaffolding; the actual test statistic and decision rule for β_1 are built in Chapter 6.

QRS@NTU — qrs.ntu.org

6 Lecture 6: Inference with a Single Regressor

This chapter answers three questions: (1) *how uncertain is $\hat{\beta}_1$, and how do we measure that uncertainty*; (2) *how do we test a hypothesis about the population slope β_1*; and (3) *how do we summarise the plausible range of β_1 with a confidence interval, and how does that range relate to the p-value reported alongside it.* Chapter 5 showed that $\hat{\beta}_1$ is unbiased and, under the Gauss–Markov conditions, the most efficient linear unbiased estimator of β_1; it also introduced the large-sample normal approximation to the sampling distribution of $\hat{\beta}_1$. This chapter turns that distributional result into the actual machinery of statistical inference – standard errors, t-statistics, confidence intervals, and p-values – and closes Part I of the course before the Week 7 quiz (Chapter 7).

6.1 From the Sampling Distribution of $\hat{\beta}_1$ to Its Standard Error

Under Least Squares Assumptions 1–3, and when the sample size n is large, the OLS slope estimator is approximately normally distributed around the true slope:

$$\hat{\beta}_1 \approx N\big(\beta_1,\ \sigma^2_{\hat{\beta}_1}\big).$$

Subtracting the mean and dividing by the standard deviation standardises this to the familiar bell curve:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \approx N(0,1) \qquad \text{(large samples)}.$$

This is the large-sample counterpart of the exact normality result you would get if the errors u_i were themselves normally distributed; here it follows instead from a Central-Limit-Theorem argument applied to $\hat{\beta}_1$, which is itself (asymptotically) a weighted average of the data.

Extension: the population variance of $\hat{\beta}_1$ (need not recite)

The population variance of $\hat{\beta}_1$ that drives the approximation above is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \cdot \frac{\mathrm{Var}\big[(X_i - \mu_X)u_i\big]}{\big[\mathrm{Var}(X_i)\big]^2},$$

where $\mu_X = E[X_i]$. This object is unobservable because it depends on the population moments of (X_i, u_i), which is exactly why an *estimator* of $\sigma^2_{\hat{\beta}_1}$ is needed in practice.

Because u_i is unobservable, $\sigma^2_{\hat{\beta}_1}$ must be estimated from the data by replacing the population moments with their sample analogues and u_i with the OLS residual $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.

Extension: the (heteroskedasticity-robust) estimator of $\sigma^2_{\hat{\beta}_1}$ (need not recite)

$$\hat{\sigma}^2_{\hat{\beta}_1} \;=\; \frac{1}{n} \cdot \frac{\dfrac{1}{n-2}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2\,\hat{u}_i^2}{\left[\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}. \qquad (*)$$

This formula is not required for recitation; what matters for the exam is the object it defines and how to use it, not its derivation. Two features are worth noting for intuition. First, $(*)$ does not assume the spread of u_i around the regression line is the same at every value of X_i (it is "heteroskedasticity-robust" in this sense) – unlike the simplest textbook formula, which assumes constant error variance (homoskedasticity). Second, the factor $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{u}_i^2$ in the numerator is a variance-type average of the squared residuals, weighted by how far X_i sits from $\bar{X}$.

Key Formula: the standard error of $\hat{\beta}_1$

$$\text{SE}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \hat{\sigma}_{\hat{\beta}_1}.$$

$\text{SE}(\hat{\beta}_1)$ estimates the spread (dispersion) of the sampling distribution of $\hat{\beta}_1$ around β_1. It is computed automatically by regression software (e.g. Stata) and reported directly under the coefficient.

Intuition: why does $\text{SE}(\hat{\beta}_1)$ shrink? Two comparative-statics facts fall directly out of $(*)$:

- **Larger n.** The leading $1/n$ factor means $\hat{\sigma}^2_{\hat{\beta}_1}$ shrinks roughly in proportion to $1/n$, so $\text{SE}(\hat{\beta}_1)$ shrinks roughly like $1/\sqrt{n}$. More observations pin down the slope more precisely – the same logic that makes $\text{Var}(\bar{Y}) = \sigma^2/n$ fall as n grows (Chapter 4).
- **More variation in X.** The denominator $\left[\text{Var}(X_i)\right]^2$ (equivalently $\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2$ in the sample analogue) grows when the regressor X_i is more spread out. A regressor that barely varies across the sample gives the data very little leverage to pin down how Y moves with X, so $\text{SE}(\hat{\beta}_1)$ is large; a widely spread regressor gives many sharply different X_i against which to trace out the slope, so $\text{SE}(\hat{\beta}_1)$ is small.

As a rough check on the first point: since $\hat{\sigma}^2_{\hat{\beta}_1}$ scales like $1/n$, quadrupling the sample size (holding the spread of X fixed) is expected to roughly *halve* $\text{SE}(\hat{\beta}_1)$, because $\sqrt{1/(4n)} = \frac{1}{2}\sqrt{1/n}$. This is the same $\sqrt{n}$-logic behind $\text{Var}(\bar{Y}) = \sigma^2/n$ in Chapter 4, and it is also why Chapter 5's efficiency ranking of estimators matters for inference: a less efficient (higher-variance) unbiased estimator of β_1 would produce a wider $\text{SE}(\hat{\beta}_1)$, a smaller $|t^{\text{act}}|$ for the same underlying effect, and hence a weaker-looking test even though nothing about the true β_1 has changed. Because $\hat{\sigma}^2_{\hat{\beta}_1}$ is itself consistent for $\sigma^2_{\hat{\beta}_1}$, in large samples $\text{SE}(\hat{\beta}_1)$ is very close to the true $\sigma_{\hat{\beta}_1}$, so the standardised statistic still converges to a standard normal when the *estimated* standard error is used in place of the unknown $\sigma_{\hat{\beta}_1}$:

$$\frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \approx \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \approx N(0,1) \qquad \text{(large samples)}.$$

This last display is the entire engine behind everything in this chapter.

Worked Example (Lecture 6): standardising $\hat{\beta}_1$

Suppose $\beta_1 = -2$ and $\sigma_{\hat{\beta}_1} = 0.5$, so $\hat{\beta}_1 \approx N(-2,\, 0.5^2)$.

(a) $\Pr(\hat{\beta}_1 > -1.2)$**.**

Method 1: Standardisation and the upper tail of Z. Let $Z = (\hat{\beta}_1 - \beta_1)/\sigma_{\hat{\beta}_1} = (\hat{\beta}_1 + 2)/0.5$, so $Z \approx N(0,1)$. Subtracting $\beta_1 = -2$ from both sides of the event and dividing by $\sigma_{\hat{\beta}_1} = 0.5 > 0$ keeps the direction of the inequality:

$$\begin{aligned}\Pr(\hat{\beta}_1 > -1.2) &= \Pr\left(\frac{\hat{\beta}_1 - (-2)}{0.5} > \frac{-1.2 - (-2)}{0.5}\right) \\ &= \Pr\left(Z > \frac{0.8}{0.5}\right) = \Pr(Z > 1.6).\end{aligned}$$

By the complement rule and the standard normal table entry $\Phi(1.60) = 0.9452$,

$$\Pr(Z > 1.6) = 1 - \Pr(Z \le 1.6) = 1 - \Phi(1.6) = 1 - 0.9452 = 0.0548.$$

$$\boxed{\Pr(\hat{\beta}_1 > -1.2) = 1 - \Phi(1.6) \approx 0.0548}$$

Method 2: Reflection about the mean $\beta_1 = -2$**.** The normal density of $\hat{\beta}_1$ is symmetric about its mean -2. The point -1.2 lies $-1.2 - (-2) = 0.8$ above the mean, so its mirror image lies 0.8 below the mean, at $-2 - 0.8 = -2.8$. By symmetry the upper tail beyond -1.2 carries the same probability as the lower tail below -2.8:

$$\Pr(\hat{\beta}_1 > -1.2) = \Pr(\hat{\beta}_1 < -2.8).$$

Standardising the lower tail,

$$\Pr(\hat{\beta}_1 < -2.8) = \Pr\left(Z < \frac{-2.8 - (-2)}{0.5}\right) = \Pr\left(Z < \frac{-0.8}{0.5}\right) = \Pr(Z < -1.6) = \Phi(-1.6).$$

The lower-tail table entry is $\Phi(-1.60) = 0.0548$, which equals $1 - \Phi(1.6)$.

$$\boxed{\Pr(\hat{\beta}_1 > -1.2) = 1 - \Phi(1.6) \approx 0.0548}$$

(b) $\Pr(|\hat{\beta}_1 - \beta_1| < 1)$**.**

Method 1: Standardisation and the two symmetric tails. Dividing both sides of $|\hat{\beta}_1 - \beta_1| < 1$ by $\sigma_{\hat{\beta}_1} = 0.5 > 0$,

$$\Pr(|\hat{\beta}_1 - \beta_1| < 1) = \Pr\left(\frac{|\hat{\beta}_1 - \beta_1|}{0.5} < \frac{1}{0.5}\right) = \Pr(|Z| < 2).$$

The complement of $\{|Z| < 2\}$ is the two tails $\{Z \geq 2\} \cup \{Z \leq -2\}$, which have equal probability by the symmetry of $N(0, 1)$:

$$\begin{aligned}\Pr(|Z| < 2) &= 1 - \big[\Pr(Z \geq 2) + \Pr(Z \leq -2)\big] = 1 - 2\Pr(Z \geq 2) \\ &= 1 - 2\big[1 - \Phi(2)\big] = 2\Phi(2) - 1.\end{aligned}$$

With the table entry $\Phi(2.00) = 0.97725$,

$$2\Phi(2) - 1 = 2(0.97725) - 1 = 1.9545 - 1 = 0.9545.$$

$$\boxed{\Pr(|\hat{\beta}_1 - \beta_1| < 1) = 2\Phi(2) - 1 \approx 0.9545}$$

Method 2: Interval on the original scale, endpoint by endpoint. With $\beta_1 = -2$, unpack the absolute value as a double inequality:

$$|\hat{\beta}_1 + 2| < 1 \iff -1 < \hat{\beta}_1 + 2 < 1 \iff -3 < \hat{\beta}_1 < -1.$$

Standardise each endpoint separately:

$$z_{\text{low}} = \frac{-3 - (-2)}{0.5} = \frac{-1}{0.5} = -2, \qquad z_{\text{high}} = \frac{-1 - (-2)}{0.5} = \frac{1}{0.5} = 2.$$

The probability of the interval is the difference of the CDF at its endpoints; the table gives $\Phi(2.00) = 0.97725$ and $\Phi(-2.00) = 0.02275$:

$$\Pr(-3 < \hat{\beta}_1 < -1) = \Phi(2) - \Phi(-2) = 0.97725 - 0.02275 = 0.9545.$$

Since $\Phi(-2) = 1 - \Phi(2)$, the difference $\Phi(2) - \Phi(-2)$ is the same number as $2\Phi(2) - 1$.

$$\boxed{\Pr(|\hat{\beta}_1 - \beta_1| < 1) = 2\Phi(2) - 1 \approx 0.9545}$$

Exam Focus: Quiz scope ends here

The Week 7 Quiz covers this chapter only up to this point – the large-sample distribution of $\hat{\beta}_1$, the (need-not-recite) variance estimator, and $\mathrm{SE}(\hat{\beta}_1)$, together with the standardisation $(\hat{\beta}_1 - \beta_1)/\mathrm{SE}(\hat{\beta}_1) \approx N(0,1)$ – corresponding to Lecture 6 slides 1–3. Everything from the t-statistic onward (the t-test, confidence intervals, and p-values, starting at slide 4) is **examined only in the Final**, not the Quiz. The Quiz also sets no Stata-output questions.

6.2 The t-Statistic and Hypothesis Testing

Definition 6.1: The t-statistic for $\hat{\beta}_1$

For a hypothesised value $\beta_{1,0}$, the t-statistic for $\hat{\beta}_1$ is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\text{estimator} - \text{hypothesised value}}{\text{standard error of the estimator}}.$$

When the null hypothesis is $H_0 : \beta_1 = 0$ (the regressor has *no* effect), this collapses to

$$t = \frac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}.$$

The realised value of t, computed from the actual sample, is written t^{act}.

Because $t \to N(0,1)$ in large samples when H_0 is true, rejecting H_0 whenever t^{act} is "too far" from 0 is a hypothesis test with a controlled false-rejection rate. This is the same Type I/Type II error language introduced in Chapter 1 (there illustrated with a COVID antibody test): rejecting a true H_0 is a Type I error, and the significance level α (conventionally 10%, 5%, or 1% in economics) is exactly the probability of that error that we are willing to tolerate. The t-test built below is simply the class-size/test-score version of the abstract testing framework from Chapter 1, now made concrete with an explicit estimator and standard error.

Theorem 6.1: Asymptotic size of the two-sided t-test

Suppose Least Squares Assumptions 1–3 hold and $H_0 : \beta_1 = \beta_{1,0}$ is true. Consider the decision rule

$$\text{reject } H_0 \iff |t^{\text{act}}| > 1.96.$$

Then, as $n \to \infty$, the probability of rejecting H_0 (a Type I error, since H_0 is true) converges to 0.05:

$$\Pr\big(\text{reject } H_0 \mid H_0 \text{ true}\big) \longrightarrow 0.05.$$

Proof. Under H_0, $\beta_1 = \beta_{1,0}$, so the t-statistic is

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)}.$$

By the large-sample result of the previous subsection, applied with β_1 replaced by the true value $\beta_{1,0}$,

$$t \xrightarrow{d} Z, \qquad Z \sim N(0,1),$$

where $\xrightarrow{d}$ denotes convergence in distribution. The rejection probability is therefore

$$\Pr\big(|t^{\text{act}}| > 1.96 \mid H_0\big) \longrightarrow \Pr\big(|Z| > 1.96\big).$$

Since Z is symmetric about 0,

$$\Pr\big(|Z| > 1.96\big) = \Pr(Z > 1.96) + \Pr(Z < -1.96) = 2\big[1 - \Phi(1.96)\big],$$

where $\Phi(\cdot)$ is the standard normal CDF. Using the standard normal table, $\Phi(1.96) = 0.975$, so

$$2\big[1 - \Phi(1.96)\big] = 2(0.025) = 0.05.$$

Combining the last three displays gives $\Pr(\text{reject } H_0 \mid H_0 \text{ true}) \to 0.05$, as claimed. □

Key Formula: two-sided critical values

Significance level	Two-sided critical value c
10%	1.645
5%	1.96
1%	2.58

Reject $H_0 : \beta_1 = \beta_{1,0}$ at significance level α against $H_1 : \beta_1 \neq \beta_{1,0}$ iff $|t^{\text{act}}| > c_\alpha$.

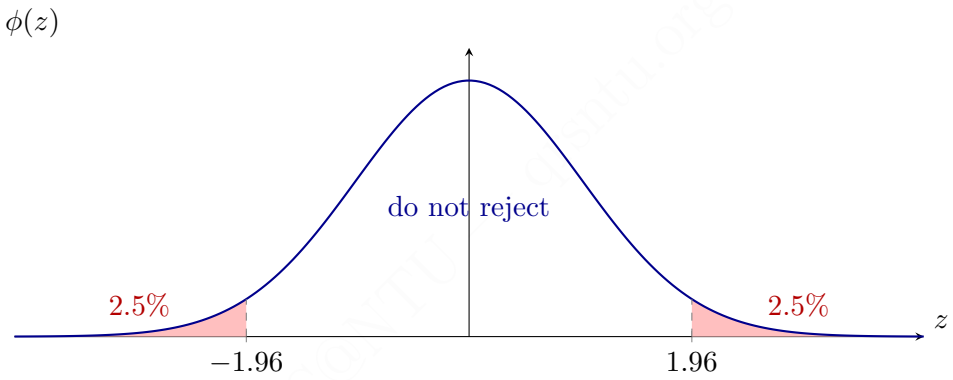


Figure: the $N(0,1)$ density with the two-sided 5% rejection region shaded. Each tail beyond ± 1.96 holds probability 2.5%, for a combined 5% chance of rejecting a true H_0.

Common Mistake: forgetting the absolute value

The decision rule compares $|t^{\text{act}}|$ to the critical value, *not* t^{act} itself. A large negative t-statistic (e.g. $t^{\text{act}} = -4.38$) is just as strong evidence against H_0 as a large positive one. Writing "reject if $t^{\text{act}} > 1.96$" silently throws away every negative rejection and is a common source of lost exam marks.

Exam Focus: the Big Picture (5 steps)

The whole chapter is one five-step recipe for turning a population question into a testable statistic:

1. State the model and the population object of interest (e.g. β_1).
2. Provide an estimator of that object (e.g. the OLS estimator $\hat{\beta}_1$).
3. Study the distribution of the estimator: in large samples, and under Least Squares Assumptions 1–3, $\hat{\beta}_1 \approx N(\beta_1, \sigma^2_{\hat{\beta}_1})$.
4. Compute the standard error $\text{SE}(\hat{\beta}_1)$.
5. Construct the t-statistic $t = (\hat{\beta}_1 - \beta_{1,0})/\,\text{SE}(\hat{\beta}_1)$ for hypothesis testing.

This recipe generalises directly to $\hat{\beta}_0$ (replace $\hat{\beta}_1$ by $\hat{\beta}_0$ throughout) and, in Part II, to every coefficient in a multiple regression.

Extension: one-sided tests

The lecture works only with two-sided alternatives $H_1 : \beta_1 \neq \beta_{1,0}$, but the same t-statistic supports a one-sided test. For $H_1 : \beta_1 > \beta_{1,0}$, reject at the 5% level iff $t^{\text{act}} > 1.645$; for $H_1 : \beta_1 < \beta_{1,0}$, reject iff $t^{\text{act}} < -1.645$. The one-sided critical value 1.645 is smaller in magnitude than the two-sided 1.96 because all 5% of the rejection probability sits in a single tail instead of being split 2.5%/2.5% across both tails.

Method Pattern: the four-beat hypothesis test

Every hypothesis test in this course – and every one you should write out in the exam – follows the same four beats:

1. **Hypotheses.** State H_0 and H_1 in terms of the population parameter.
2. **Test statistic.** Compute $t^{\text{act}} = (\text{estimate} - \text{hypothesised value})/\,\text{SE}(\text{estimate})$.
3. **Decision rule.** State the critical value for the chosen significance level and whether the test is one- or two-sided.
4. **Conclusion in context.** Say whether H_0 is rejected, at which levels, and what that means for the economic question at hand.

Worked Example (Lecture 6): testing class size effects on test scores

The California test score regression (first estimated in Chapter 2) is

$$\widehat{\text{TestScore}} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28}\,\text{STR}, \qquad n = 420,$$

so $\text{SE}(\hat{\beta}_0) = 10.4$ and $\text{SE}(\hat{\beta}_1) = 0.52$.

(a) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Method 1: Two-sided t-test.

Hypotheses. $H_0 : \beta_1 = 0$ (the population slope of TestScore on STR is zero) against the two-sided alternative $H_1 : \beta_1 \neq 0$.

Test statistic. With $\hat{\beta}_1 = -2.28$, $\beta_{1,0} = 0$ and $\text{SE}(\hat{\beta}_1) = 0.52$,

$$t^{\text{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)} = \frac{-2.28 - 0}{0.52} = -4.3846 \approx -4.38, \qquad |t^{\text{act}}| = 4.38.$$

Decision rule. Under H_0 and Least Squares Assumptions 1–3, $t^{\text{act}} \approx N(0,1)$ because $n = 420$ is large. The two-sided critical value c_α solves $2\big[1-\Phi(c_\alpha)\big] = \alpha$; the normal table gives $\Phi(1.96) = 0.975$, so $c_{0.05} = 1.96$, and $\Phi(2.58) = 0.995$, so $c_{0.01} = 2.58$. Reject H_0 at level α iff $|t^{\text{act}}| > c_\alpha$. Here

$$|t^{\text{act}}| = 4.38 > 2.58 > 1.96,$$

so H_0 is rejected at both the 5% and the 1% level.
Conclusion. The estimated slope of -2.28 test-score points per additional student per teacher is statistically significant at the 5% and 1% levels, so the population slope on STR differs from zero.

$$\boxed{t^{\text{act}} = -4.38, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 2: Confidence-interval duality. By Theorem 6.2, $H_0 : \beta_1 = 0$ is rejected at 5% iff 0 lies outside $\hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$; the same argument with 2.58 in place of 1.96 gives the 99% interval and the 1% test.

$$CI_{95\%}(\beta_1) = -2.28 \pm 1.96(0.52) = -2.28 \pm 1.0192 = [-3.2992,\ -1.2608],$$
$$CI_{99\%}(\beta_1) = -2.28 \pm 2.58(0.52) = -2.28 \pm 1.3416 = [-3.6216,\ -0.9384].$$

Both upper endpoints are negative ($-1.2608 < 0$ and $-0.9384 < 0$), so 0 lies outside both intervals and H_0 is rejected at 5% and at 1%. Measured in standard errors, 0 sits $(0 - (-2.28))/0.52 = 4.38$ standard errors above $\hat{\beta}_1$, beyond both half-widths (1.96 and 2.58 standard errors); equivalently $t^{\text{act}} = -4.38$. Hence the slope on STR is statistically significant at both levels and differs from zero.

$$\boxed{t^{\text{act}} = -4.38, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 3: p-value. From $t^{\text{act}} = -4.38$, the two-sided p-value (computed in full in part (c)) is

$$p = 2\Phi(-|t^{\text{act}}|) = 2\Phi(-4.38) = 2(5.93 \times 10^{-6}) = 1.19 \times 10^{-5}.$$

By Theorem 6.3, reject at level α iff $p < \alpha$. Since $1.19 \times 10^{-5} < 0.01 < 0.05$, H_0 is rejected at 1% and at 5%: the slope on STR is statistically significant and differs from zero.

$$\boxed{t^{\text{act}} = -4.38, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

(b) The 95% confidence interval for β_1.

Method 1: Estimate plus or minus 1.96 standard errors. The half-width is $1.96\ \text{SE}(\hat{\beta}_1) = 1.96(0.52) = 1.0192$. Hence

$$\begin{aligned} CI_{95\%}(\beta_1) &= \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1) = -2.28 \pm 1.0192, \\ \text{lower} &= -2.28 - 1.0192 = -3.2992 \approx -3.30, \\ \text{upper} &= -2.28 + 1.0192 = -1.2608 \approx -1.26. \end{aligned}$$

In repeated samples, 95% of intervals built this way contain β_1; every slope from -3.30 to -1.26 test-score points per student per teacher is consistent with the data at the 5% level.

$$\boxed{CI_{95\%}(\beta_1) = [-3.30,\ -1.26]}$$

Method 2: Inverting the 5% t-test. By Definition 6.2, the interval is the set of $\beta_{1,0}$ with $\big|(-2.28 - \beta_{1,0})/0.52\big| \le 1.96$. Multiplying by $0.52 > 0$ and unpacking the absolute value,

$$\begin{aligned} |-2.28 - \beta_{1,0}| &\le 1.96(0.52) = 1.0192, \\ -1.0192 \le -2.28 - \beta_{1,0} &\le 1.0192, \\ -1.0192 + 2.28 \le -\beta_{1,0} &\le 1.0192 + 2.28, \\ 1.2608 \le -\beta_{1,0} &\le 3.2992, \\ -3.2992 \le \beta_{1,0} &\le -1.2608, \end{aligned}$$

where the last line multiplies by -1 and reverses both inequalities. Rounding to two decimals gives $[-3.30,\ -1.26]$.

$$\boxed{CI_{95\%}(\beta_1) = [-3.30,\ -1.26]}$$

Worked Example (Lecture 6): testing class size effects on test scores (cont.)

(c) The p-value.

Method 1: Two tails of the standard normal CDF. By Definition 6.3, with $t^{\text{act}} = -4.38$ and $Z \sim N(0,1)$ under H_0,

$$p = \Pr(|Z| > 4.38) = \Pr(Z > 4.38) + \Pr(Z < -4.38).$$

The two tails are equal by the symmetry of $N(0,1)$, so

$$p = 2\Pr(Z < -4.38) = 2\Phi(-|t^{\text{act}}|) = 2\Phi(-4.38).$$

The printed normal table stops near $z = -3.49$, so $\Phi(-4.38)$ is read from software (Stata: `normal(-4.38)`), which returns $\Phi(-4.38) = 5.93 \times 10^{-6}$. Hence

$$p = 2(5.93 \times 10^{-6}) = 1.186 \times 10^{-5} \approx 1.19 \times 10^{-5}.$$

If β_1 were 0, a t-statistic at least 4.38 away from 0 would arise in about 12 of every million samples.

$$\boxed{p \approx 1.19 \times 10^{-5} \approx 0.0000119}$$

Method 2: Mills-ratio expansion of the normal tail. This method should be used only if the course allows this interpretation. For large $x > 0$ the standard normal lower tail satisfies

$$\Phi(-x) \approx \frac{\phi(x)}{x}\left(1 - \frac{1}{x^2} + \frac{3}{x^4} - \frac{15}{x^6}\right), \qquad \phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

At $x = 4.38$: $x^2 = 19.1844$, so $x^2/2 = 9.5922$ and $e^{-9.5922} = 6.826 \times 10^{-5}$. Dividing by $\sqrt{2\pi} = 2.5066$,

$$\phi(4.38) = \frac{6.826 \times 10^{-5}}{2.5066} = 2.723 \times 10^{-5}, \qquad \frac{\phi(4.38)}{4.38} = 6.217 \times 10^{-6}.$$

The correction terms are $1/x^2 = 0.05213$, $3/x^4 = 0.00815$ and $15/x^6 = 0.00212$, so the bracket is

$$1 - 0.05213 + 0.00815 - 0.00212 = 0.95390.$$

Therefore

$$\Phi(-4.38) \approx (6.217 \times 10^{-6})(0.95390) = 5.931 \times 10^{-6},$$
$$p = 2(5.931 \times 10^{-6}) = 1.186 \times 10^{-5} \approx 1.19 \times 10^{-5}.$$

$$\boxed{p \approx 1.19 \times 10^{-5} \approx 0.0000119}$$

(d) Effect of reducing STR by 2 students.

Method 1: Scaling $\hat{\beta}_1$ and its standard error by ΔX. A reduction of 2 students per teacher is $\Delta X = -2$. The predicted change in test score is

$$\hat{\beta}_1 \, \Delta X = (-2.28)(-2) = 4.56 \text{ points}.$$

Because ΔX is a fixed number, $\text{Var}(\hat{\beta}_1 \Delta X) = (\Delta X)^2 \, \text{Var}(\hat{\beta}_1)$, so the standard error scales by $|\Delta X|$:

$$\text{SE}(\hat{\beta}_1 \Delta X) = |\Delta X| \, \text{SE}(\hat{\beta}_1) = 2(0.52) = 1.04.$$

The 95% interval for $\beta_1 \Delta X$ (Key Formula above) is

$$CI_{95\%}(\beta_1 \Delta X) = 4.56 \pm 1.96(1.04) = 4.56 \pm 2.0384,$$
$$\text{lower} = 4.56 - 2.0384 = 2.5216 \approx 2.52,$$
$$\text{upper} = 4.56 + 2.0384 = 6.5984 \approx 6.60.$$

A class-size reduction of 2 students per teacher is associated with a predicted test score 4.56 points higher, with a 95% interval from 2.52 to 6.60 points.

$$\boxed{\text{predicted gain} = 4.56 \text{ points}, \quad CI_{95\%} = [2.52, 6.60]}$$

Method 2: Rescaling the interval for β_1 endpoint by endpoint. From part (b), before rounding, $-3.2992 \le \beta_1 \le -1.2608$. Multiplying every term by $\Delta X = -2 < 0$ reverses both inequalities:

$$(-2)(-1.2608) \le -2\beta_1 \le (-2)(-3.2992) \quad \Longrightarrow \quad 2.5216 \le \beta_1 \Delta X \le 6.5984.$$

The predicted gain is the midpoint of this interval,

$$\frac{2.5216 + 6.5984}{2} = \frac{9.12}{2} = 4.56 \text{ points},$$

which equals $(-2.28)(-2)$, and the endpoints round to 2.52 and 6.60 points.

$$\boxed{\text{predicted gain} = 4.56 \text{ points}, \quad CI_{95\%} = [2.52, 6.60]}$$

Policy Application: class size and statistical vs. economic significance

Rejecting $H_0 : \beta_1 = 0$ tells a school board only that the data are inconsistent with *zero* effect; it says nothing about whether a 2.28-point drop in test scores per additional student is large enough to justify the cost of hiring more teachers. That second question is about *economic* significance and requires comparing the estimated effect size (and its confidence interval) to a policy-relevant benchmark, not just checking whether $p < 0.05$.

Worked Example (Lecture 6): CAPM market-beta test

Consider an illustrative monthly CAPM regression for Apple Inc. (AAPL):

$$R_{\text{Apple}} - R_f = \alpha + \beta(R_m - R_f) + u,$$

where $R_{\text{Apple}} - R_f$ is Apple's monthly excess return, $R_m - R_f$ is the market (S&P 500) excess return, and R_f is the one-month U.S. Treasury bill rate. Suppose the fitted regression is

$$\widehat{R_{\text{Apple}} - R_f} = 0.15 + 1.25\,(R_m - R_f), \qquad \text{SE}(\hat{\beta}) = 0.10.$$

Test whether Apple has the same market exposure as the market portfolio, i.e. $H_0 : \beta = 1$ vs. $H_1 : \beta \neq 1$.

Method 1: Two-sided t-test against $\beta_0 = 1$.

Hypotheses. $H_0 : \beta = 1$ (Apple's excess return moves one-for-one with the market excess return, the exposure of the market portfolio itself) against the two-sided alternative $H_1 : \beta \neq 1$.

Test statistic. The hypothesised value is $\beta_0 = 1$, not 0, so the numerator is $\hat{\beta} - 1$:

$$t^{\text{act}} = \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})} = \frac{1.25 - 1}{0.10} = \frac{0.25}{0.10} = 2.50.$$

Decision rule. Under H_0 and Least Squares Assumptions 1–3, $t^{\text{act}} \approx N(0,1)$ in large samples (the sample of monthly returns is taken to be large). The two-sided critical value c_α solves $2\big[1 - \Phi(c_\alpha)\big] = \alpha$; from the normal table, $\Phi(1.96) = 0.975$ gives $c_{0.05} = 1.96$ and $\Phi(2.58) = 0.995$ gives $c_{0.01} = 2.58$. Reject at level α iff $|t^{\text{act}}| > c_\alpha$. Here

$$1.96 < |t^{\text{act}}| = 2.50 < 2.58,$$

so H_0 is rejected at 5% but not at 1%.
Conclusion. At the 5% level Apple's market beta differs from 1: the estimate 1.25 lies 0.25 above the market portfolio's exposure and the gap is statistically significant; at the 1% level the data do not reject $\beta = 1$.

$$\boxed{t^{\text{act}} = 2.50, \quad \text{reject } H_0 : \beta = 1 \text{ at 5\% but not at 1\%.}}$$

Method 2: Confidence-interval duality at 95% **and** 99%**.** By Theorem 6.2, $H_0 : \beta = 1$ is rejected at 5% iff $1 \notin \hat{\beta} \pm 1.96\ \text{SE}(\hat{\beta})$, and at 1% iff $1 \notin \hat{\beta} \pm 2.58\ \text{SE}(\hat{\beta})$.

$$CI_{95\%}(\beta) = 1.25 \pm 1.96(0.10) = 1.25 \pm 0.196 = [1.054,\ 1.446],$$
$$CI_{99\%}(\beta) = 1.25 \pm 2.58(0.10) = 1.25 \pm 0.258 = [0.992,\ 1.508].$$

Since $1 < 1.054$, the value 1 lies outside the 95% interval, so H_0 is rejected at 5%. Since $0.992 < 1 < 1.508$, the value 1 lies inside the 99% interval, so H_0 is not rejected at 1%. In standard-error units, 1 sits $(1.25 - 1)/0.10 = 2.50$ standard errors below $\hat{\beta}$, between the half-widths 1.96 and 2.58; equivalently $t^{\text{act}} = 2.50$. Apple's beta therefore differs from 1 at the 5% level but not at the 1% level.

$$\boxed{t^{\text{act}} = 2.50, \quad \text{reject } H_0 : \beta = 1 \text{ at 5\% but not at 1\%.}}$$

Method 3: p**-value.** With $t^{\text{act}} = 2.50$, the two-sided p-value (computed in full below) is

$$p = 2\Phi(-2.50) = 2(0.0062) = 0.0124.$$

By Theorem 6.3, reject at level α iff $p < \alpha$. Since $0.01 < 0.0124 < 0.05$, H_0 is rejected at 5% but not at 1%: Apple's beta is statistically different from 1 at the 5% level only.

$$\boxed{t^{\text{act}} = 2.50, \quad \text{reject } H_0 : \beta = 1 \text{ at 5\% but not at 1\%.}}$$

95% confidence interval:
Method 1: Estimate plus or minus 1.96 **standard errors.** The half-width is $1.96\ \text{SE}(\hat{\beta}) = 1.96(0.10) = 0.196$. Hence

$$\begin{aligned} CI_{95\%}(\beta) &= \hat{\beta} \pm 1.96\ \text{SE}(\hat{\beta}) = 1.25 \pm 0.196, \\ \text{lower} &= 1.25 - 0.196 = 1.054, \\ \text{upper} &= 1.25 + 0.196 = 1.446. \end{aligned}$$

Every beta from 1.054 to 1.446 is consistent with the data at the 5% level; all of them exceed 1.

$$\boxed{CI_{95\%}(\beta) = [1.054,\ 1.446]}$$

Method 2: Inverting the 5% t**-test.** The interval is the set of β_0 with $\left|(1.25 - \beta_0)/0.10\right| \le 1.96$. Multiplying by $0.10 > 0$ and unpacking the absolute value,

$$\begin{aligned} |1.25 - \beta_0| &\le 1.96(0.10) = 0.196, \\ -0.196 \le 1.25 - \beta_0 &\le 0.196, \\ -1.446 \le -\beta_0 &\le -1.054, \\ 1.054 \le \beta_0 &\le 1.446, \end{aligned}$$

where the third line subtracts 1.25 throughout and the last line multiplies by -1, reversing both inequalities.

$$\boxed{CI_{95\%}(\beta) = [1.054,\ 1.446]}$$

p-value:

Method 1: Two tails of the standard normal CDF. By Definition 6.3, with $Z \sim N(0,1)$ under H_0,

$$p = \Pr(|Z| > 2.50) = \Pr(Z > 2.50) + \Pr(Z < -2.50) = 2\big[1 - \Phi(2.50)\big],$$

using the symmetry $\Pr(Z < -2.50) = \Pr(Z > 2.50)$. The table gives $\Phi(2.50) = 0.9938$, so

$$p = 2(1 - 0.9938) = 2(0.0062) = 0.0124.$$

If Apple's beta were exactly 1, a t-statistic at least 2.50 away from 0 would occur in about 1.24% of samples.

$$\boxed{p \approx 0.0124}$$

Policy Application: interpreting Apple's beta

$\hat{\beta} = 1.25 > 1$ with a confidence interval entirely above 1 means the data are consistent with Apple's stock being systematically *riskier* than the market portfolio: a 1 percentage-point move in the market excess return is associated with an estimated 1.054– to 1.446-percentage-point move in Apple's excess return. An investor holding Apple should expect larger swings than holding a market index fund, in both directions.

Common Mistake: testing $\beta_1 = c \neq 0$ with the wrong ratio

When the null hypothesis is not $\beta_1 = 0$, the numerator of the t-statistic must be $\hat{\beta}_1 - \beta_{1,0}$, not $\hat{\beta}_1$ alone. In the CAPM example, computing $\hat{\beta}/\operatorname{SE}(\hat{\beta}) = 1.25/0.10 = 12.5$ instead of $(\hat{\beta} - 1)/\operatorname{SE}(\hat{\beta}) = 2.50$ answers a completely different question ("is β different from 0?" rather than the intended "is β different from 1?") and would lead to the wrong economic conclusion.

6.3 Confidence Intervals for β_1

Definition 6.2: 95% confidence interval for β_1

A 95% (two-sided) confidence interval for β_1 is a random interval, computed from the sample, that contains the true β_1 with probability approximately 95% across repeated samples. Equivalently, it is the set of hypothesised values $\beta_{1,0}$ that a two-sided 5%-level t-test would *not* reject:

$$CI_{95\%}(\beta_1) = \left\{\beta_{1,0} \;:\; \left|\tfrac{\hat{\beta}_1 - \beta_{1,0}}{\operatorname{SE}(\hat{\beta}_1)}\right| \leq 1.96\right\}.$$

Theorem 6.2: Duality between the confidence interval and the two-sided t-test

$$CI_{95\%}(\beta_1) = \big[\hat{\beta}_1 - 1.96\ \operatorname{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\ \operatorname{SE}(\hat{\beta}_1)\big].$$

Proof. By definition, $\beta_{1,0} \in CI_{95\%}(\beta_1)$ iff

$$\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{\operatorname{SE}(\hat{\beta}_1)}\right| \leq 1.96.$$

Since $\operatorname{SE}(\hat{\beta}_1) > 0$, multiplying both sides by $\operatorname{SE}(\hat{\beta}_1)$ preserves the inequality:

$$|\hat{\beta}_1 - \beta_{1,0}| \leq 1.96\ \operatorname{SE}(\hat{\beta}_1).$$

Writing out the absolute-value inequality as a double inequality,

$$-1.96\ \mathrm{SE}(\hat{\beta}_1) \le \hat{\beta}_1 - \beta_{1,0} \le 1.96\ \mathrm{SE}(\hat{\beta}_1).$$

Subtracting $\hat{\beta}_1$ throughout,

$$-\hat{\beta}_1 - 1.96\ \mathrm{SE}(\hat{\beta}_1) \le -\beta_{1,0} \le -\hat{\beta}_1 + 1.96\ \mathrm{SE}(\hat{\beta}_1).$$

Multiplying through by -1 reverses both inequalities,

$$\hat{\beta}_1 - 1.96\ \mathrm{SE}(\hat{\beta}_1) \le \beta_{1,0} \le \hat{\beta}_1 + 1.96\ \mathrm{SE}(\hat{\beta}_1).$$

This is exactly the statement $\beta_{1,0} \in \left[\hat{\beta}_1 - 1.96\ \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\ \mathrm{SE}(\hat{\beta}_1)\right]$, so the two sets are equal. □

Key Formula: confidence interval for β_1 and for a change ΔX

$$CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96\ \mathrm{SE}(\hat{\beta}_1), \qquad CI_{95\%}(\beta_1\,\Delta X) = \hat{\beta}_1\,\Delta X \pm 1.96\ \mathrm{SE}(\hat{\beta}_1)\,|\Delta X|.$$

A confidence interval for β_0 is constructed the same way, using $\hat{\beta}_0$ and $\mathrm{SE}(\hat{\beta}_0)$.

Common Mistake: "95% probability that β_1 is in this interval"

Once the sample is drawn and the numbers $[-3.30, -1.26]$ are computed, β_1 either is or is not in that particular fixed interval – there is nothing left to be random. The correct statement is about the *procedure*: if we repeated the sampling and re-computed the interval many times, about 95% of the resulting intervals would contain the fixed, unknown β_1. "95% probability that β_1 lies in *this* interval" is loose shorthand for that repeated-sampling statement, not a statement about β_1 itself being random.

Common Mistake: "fail to reject" is not "accept as true"

Failing to reject H_0 (e.g. $|t^{\mathrm{act}}| \le 1.96$) means the data are *not sufficiently inconsistent* with H_0 to rule it out at the chosen significance level – it does not mean the data have proven H_0 true, nor that β_1 equals the hypothesised value exactly. A test with low power (e.g. a small sample, or a regressor with little variation) can easily fail to reject a false H_0.

Worked Example (own): testing $H_0 : \beta_1 = c \neq 0$ via the CI duality

Suppose an estimated relationship between weekly study hours (X) and an exam score out of 100 (Y) is

$$\widehat{\mathrm{Score}} = 50 + 3.5\,\mathrm{Hours}, \qquad \mathrm{SE}(\hat{\beta}_1) = 1.2, \qquad n = 150.$$

A widely repeated rule of thumb claims that each additional hour of study raises the exam score by exactly 5 points. Test $H_0 : \beta_1 = 5$ against $H_1 : \beta_1 \neq 5$ at the 5% level, and cross-check the conclusion against the 95% confidence interval.

Method 1: Two-sided t-test against $\beta_{1,0} = 5$.

Hypotheses. $H_0 : \beta_1 = 5$ (in the population, one more weekly study hour is associated with exactly 5 more exam points) against the two-sided alternative $H_1 : \beta_1 \neq 5$.

Test statistic. With $\hat{\beta}_1 = 3.5$, $\beta_{1,0} = 5$ and $\mathrm{SE}(\hat{\beta}_1) = 1.2$,

$$t^{\mathrm{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)} = \frac{3.5 - 5}{1.2} = \frac{-1.5}{1.2} = -1.25, \qquad |t^{\mathrm{act}}| = 1.25.$$

Decision rule. Under H_0 and Least Squares Assumptions 1–3, $t^{\mathrm{act}} \approx N(0,1)$ because $n = 150$ is large. The 5% two-sided critical value is 1.96, since $\Phi(1.96) = 0.975$ gives $2(1 - 0.975) = 0.05$.

Reject H_0 iff $|t^{\text{act}}| > 1.96$. Here

$$|t^{\text{act}}| = 1.25 < 1.96,$$

so H_0 is not rejected at 5%.

Conclusion. The estimated slope of 3.5 points per weekly study hour is not statistically different from 5 points per hour at the 5% level, so the data do not contradict the rule of thumb (which is not the same as proving it true).

$$\boxed{t^{\text{act}} = -1.25, \quad \text{fail to reject } H_0 : \beta_1 = 5 \text{ at } 5\%.}$$

Method 2: Confidence-interval duality (Theorem 6.2). $H_0 : \beta_1 = 5$ is rejected at 5% iff 5 lies outside $\hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$. The half-width is $1.96(1.2) = 2.352$, so

$$CI_{95\%}(\beta_1) = 3.5 \pm 2.352 = [3.5 - 2.352,\ 3.5 + 2.352] = [1.148,\ 5.852].$$

Since $1.148 < 5 < 5.852$, the value 5 lies inside the interval, so H_0 is not rejected at 5%. In standard-error units, 5 sits $(5 - 3.5)/1.2 = 1.25$ standard errors above $\hat{\beta}_1$, inside the half-width of 1.96 standard errors; equivalently $t^{\text{act}} = -1.25$. The slope is therefore not statistically different from 5 points per hour.

$$\boxed{t^{\text{act}} = -1.25, \quad \text{fail to reject } H_0 : \beta_1 = 5 \text{ at } 5\%.}$$

Method 3: p-value. By Definition 6.3 and the symmetry of $N(0,1)$,

$$p = \Pr(|Z| > 1.25) = 2\big[1 - \Phi(1.25)\big] = 2(1 - 0.8944) = 2(0.1056) = 0.2112,$$

using the table entry $\Phi(1.25) = 0.8944$. By Theorem 6.3, reject at 5% iff $p < 0.05$; here $0.2112 > 0.05$, so H_0 is not rejected and the slope is not statistically different from 5 points per hour.

$$\boxed{t^{\text{act}} = -1.25, \quad \text{fail to reject } H_0 : \beta_1 = 5 \text{ at } 5\%.}$$

The 95% confidence interval.

Method 1: Estimate plus or minus 1.96 standard errors. The half-width is $1.96\ \text{SE}(\hat{\beta}_1) = 1.96(1.2) = 2.352$. Hence

$$\begin{aligned} CI_{95\%}(\beta_1) &= \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1) = 3.5 \pm 2.352, \\ \text{lower} &= 3.5 - 2.352 = 1.148, \\ \text{upper} &= 3.5 + 2.352 = 5.852. \end{aligned}$$

Every slope from 1.148 to 5.852 exam points per weekly study hour is consistent with the data at the 5% level, and 5 is one of them.

$$\boxed{CI_{95\%}(\beta_1) = [1.148,\ 5.852]}$$

Method 2: Inverting the 5% t-test. The interval is the set of $\beta_{1,0}$ with $\big|(3.5 - \beta_{1,0})/1.2\big| \le 1.96$. Multiplying by $1.2 > 0$ and unpacking the absolute value,

$$\begin{aligned} |3.5 - \beta_{1,0}| &\le 1.96(1.2) = 2.352, \\ -2.352 \le 3.5 - \beta_{1,0} &\le 2.352, \\ -5.852 \le -\beta_{1,0} &\le -1.148, \\ 1.148 \le \beta_{1,0} &\le 5.852, \end{aligned}$$

where the third line subtracts 3.5 throughout and the last line multiplies by -1, reversing both inequalities.

$$\boxed{CI_{95\%}(\beta_1) = [1.148,\ 5.852]}$$

6.4 The p-Value

Definition 6.3: p-value for the two-sided t-test

The p-value is the probability, computed *as if* H_0 *were true*, of observing a t-statistic at least as extreme (in absolute value) as the one actually obtained:

$$p\text{-value} = \Pr_{H_0}\big(|N(0,1)| > |t^{\text{act}}|\big) = 2\,\Phi(-|t^{\text{act}}|).$$

A small p-value is evidence against H_0: it says that, if H_0 really held, data this extreme would be rare.

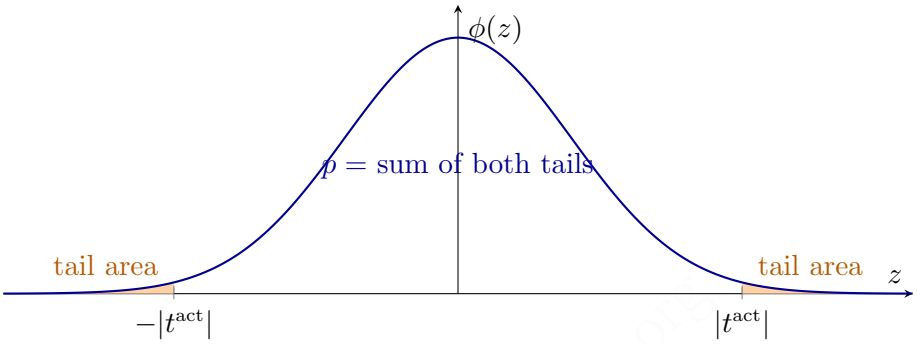


Figure: the two shaded tail areas beyond $\pm|t^{\text{act}}|$ *together give the p-value. Here* $|t^{\text{act}}| = 2.5$ *is drawn for illustration (as in the CAPM example); the smaller the shaded area, the stronger the evidence against* H_0.

Theorem 6.3: Equivalence of the p-value rule and the t-statistic rule

For any significance level $\alpha \in (0,1)$, let c_α denote the two-sided critical value satisfying $2\big[1-\Phi(c_\alpha)\big] = \alpha$ (so $c_{0.05} = 1.96$ and $c_{0.01} = 2.58$). Then

$$p\text{-value} < \alpha \iff |t^{\text{act}}| > c_\alpha.$$

That is, rejecting H_0 because the p-value is below α is exactly the same decision as rejecting H_0 because $|t^{\text{act}}|$ exceeds the critical value c_α.

Proof. By definition, $p = 2\Phi(-|t^{\text{act}}|)$. The condition $p < \alpha$ is therefore

$$2\Phi(-|t^{\text{act}}|) < \alpha \iff \Phi(-|t^{\text{act}}|) < \frac{\alpha}{2}.$$

Because Φ is strictly increasing, it has a well-defined inverse, and applying Φ^{-1} to both sides preserves the direction of the inequality:

$$-|t^{\text{act}}| < \Phi^{-1}\left(\frac{\alpha}{2}\right).$$

Multiplying by -1 reverses the inequality:

$$|t^{\text{act}}| > -\Phi^{-1}\left(\frac{\alpha}{2}\right).$$

By the symmetry of the standard normal distribution, $-\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2)$, and by construction of c_α (the value with $\Phi(c_\alpha) = 1-\alpha/2$), this equals c_α. Hence

$$p < \alpha \iff |t^{\text{act}}| > c_\alpha,$$

as claimed. Taking $\alpha = 0.05$ recovers $c_{0.05} = 1.96$; taking $\alpha = 0.01$ recovers $c_{0.01} = 2.58$. □

Key Formula: reading the p-value

$$p\text{-value} < 5\% \iff \text{reject } H_0 \text{ at } 5\%, \qquad p\text{-value} < 1\% \iff \text{reject } H_0 \text{ at } 1\%.$$

Larger $|t^{\text{act}}|$ always corresponds to a smaller p-value, and vice versa.

Worked Example (Lecture 6): CEO salary and return on equity

Recall the CEO salary regression from Chapter 3, $n = 209$:

$$\widehat{\text{salary}} = 963.191 + 18.501\,\text{roe}, \qquad \text{SE}(\hat{\beta}_1) = 6.829,$$

where salary is in thousands of US dollars and roe (return on equity) is in percent.

(a) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

Method 1: Two-sided t-test.

Hypotheses. $H_0 : \beta_1 = 0$ (the population slope of salary on roe is zero) against the two-sided alternative $H_1 : \beta_1 \neq 0$.

Test statistic. With $\hat{\beta}_1 = 18.501$, $\beta_{1,0} = 0$ and $\text{SE}(\hat{\beta}_1) = 6.829$,

$$t^{\text{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)} = \frac{18.501 - 0}{6.829} = 2.7092 \approx 2.71.$$

Decision rule. Under H_0 and Least Squares Assumptions 1–3, $t^{\text{act}} \approx N(0,1)$ because $n = 209$ is large. The two-sided critical value c_α solves $2\big[1-\Phi(c_\alpha)\big] = \alpha$; the normal table gives $\Phi(1.96) = 0.975$, so $c_{0.05} = 1.96$, and $\Phi(2.58) = 0.995$, so $c_{0.01} = 2.58$. Reject at level α iff $|t^{\text{act}}| > c_\alpha$. Here

$$|t^{\text{act}}| = 2.71 > 2.58 > 1.96,$$

so H_0 is rejected at both the 5% and the 1% level.

Conclusion. The estimated slope of 18.501 thousand dollars of salary per percentage point of roe is statistically significant at the 5% and 1% levels, so the population slope on roe differs from zero.

$$\boxed{t^{\text{act}} = 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 2: Confidence-interval duality at 95% and 99%. By Theorem 6.2, $H_0 : \beta_1 = 0$ is rejected at 5% iff $0 \notin \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$, and at 1% iff $0 \notin \hat{\beta}_1 \pm 2.58\ \text{SE}(\hat{\beta}_1)$.

$$CI_{95\%}(\beta_1) = 18.501 \pm 1.96(6.829) = 18.501 \pm 13.385 = [5.116,\ 31.886],$$
$$CI_{99\%}(\beta_1) = 18.501 \pm 2.58(6.829) = 18.501 \pm 17.619 = [0.882,\ 36.120].$$

Both lower endpoints are positive ($5.116 > 0$ and $0.882 > 0$), so 0 lies outside both intervals and H_0 is rejected at 5% and at 1%. In standard-error units, 0 sits $18.501/6.829 = 2.71$ standard errors below $\hat{\beta}_1$, beyond both half-widths (1.96 and 2.58 standard errors); equivalently $t^{\text{act}} = 2.71$. The slope on roe is statistically significant at both levels.

$$\boxed{t^{\text{act}} = 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 3: p-value. With $t^{\text{act}} = 2.71$, the two-sided p-value (computed in full in part (b)) is

$$p = 2\big[1 - \Phi(2.71)\big] = 2(0.0034) = 0.0068 \approx 0.007.$$

By Theorem 6.3, reject at level α iff $p < \alpha$. Since $0.007 < 0.01 < 0.05$, H_0 is rejected at 1% and at 5%: the slope on roe is statistically significant and differs from zero.

$$\boxed{t^{\text{act}} = 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

(b) p-value.

Method 1: Two tails of the standard normal CDF. By Definition 6.3, with $Z \sim N(0,1)$ under H_0,

$$p = \Pr(|Z| > 2.71) = \Pr(Z > 2.71) + \Pr(Z < -2.71) = 2\big[1 - \Phi(2.71)\big] = 2\Phi(-2.71),$$

using the symmetry $\Pr(Z < -2.71) = \Pr(Z > 2.71)$. The table gives $\Phi(2.71) = 0.9966$, so

$$p = 2(1 - 0.9966) = 2(0.0034) = 0.0068 \approx 0.007.$$

If roe had no linear association with salary in the population ($\beta_1 = 0$), a t-statistic at least 2.71 away from 0 would occur in about 0.7% of samples.

$$\boxed{p \approx 0.007}$$

(c) 95% confidence interval.

Method 1: Estimate plus or minus 1.96 **standard errors.** The half-width is $1.96\ \mathrm{SE}(\hat{\beta}_1) = 1.96(6.829) = 13.38484 \approx 13.385$. Hence

$$\begin{aligned} CI_{95\%}(\beta_1) &= \hat{\beta}_1 \pm 1.96\ \mathrm{SE}(\hat{\beta}_1) = 18.501 \pm 13.385, \\ \text{lower} &= 18.501 - 13.385 = 5.116 \approx 5.12, \\ \text{upper} &= 18.501 + 13.385 = 31.886 \approx 31.89. \end{aligned}$$

Every slope from about 5.12 to 31.89 thousand dollars of salary per percentage point of roe is consistent with the data at the 5% level.

$$\boxed{CI_{95\%}(\beta_1) \approx [5.12,\ 31.89]}$$

Method 2: Inverting the 5% t**-test.** The interval is the set of $\beta_{1,0}$ with $\big|(18.501 - \beta_{1,0})/6.829\big| \leq 1.96$. Multiplying by $6.829 > 0$ and unpacking the absolute value,

$$\begin{aligned} |18.501 - \beta_{1,0}| &\leq 1.96(6.829) = 13.38484 \approx 13.385, \\ -13.385 \leq 18.501 - \beta_{1,0} &\leq 13.385, \\ -31.886 \leq -\beta_{1,0} &\leq -5.116, \\ 5.116 \leq \beta_{1,0} &\leq 31.886, \end{aligned}$$

where the third line subtracts 18.501 throughout and the last line multiplies by -1, reversing both inequalities. Rounding to two decimals gives [5.12, 31.89].

$$\boxed{CI_{95\%}(\beta_1) \approx [5.12,\ 31.89]}$$

Extension: why Stata reports $[5.04, 31.97]$ instead

Statistical software such as Stata often builds the confidence interval from the t_{n-2} distribution rather than the $N(0,1)$ approximation used throughout this chapter, because the exact (finite-sample) distribution of t^{act} under classical normal-error assumptions is Student's t with $n-2$ degrees of freedom, not $N(0,1)$. With $n = 209$, the relevant critical value is $t_{0.025,207} \approx 1.971$, only marginally above 1.96:

$$18.501 \pm 1.971(6.829) = [5.04,\ 31.96],$$

which matches Stata's reported interval up to the last digit (Stata works with the unrounded coefficient and standard error). As $n-2 \to \infty$, the t_{n-2} distribution converges to $N(0,1)$ and $1.971 \to 1.96$, so the two intervals coincide in the large-sample limit; the small gap here (1.971 vs. 1.96) is purely a finite-sample refinement and does not change any conclusion in this chapter.

Common Mistake: statistical significance is not economic significance

A tiny p-value (as in the test-score example, $p \approx 10^{-5}$) shows that β_1 is precisely estimated to be different from 0; it says nothing about whether the estimated *size* of the effect is economically important. Conversely, an effect that is large in magnitude can still fail to be statistically significant if the standard error is large (e.g. a small sample or little variation in X). Always report and interpret both the point estimate (or a confidence interval) and the test result – never the p-value alone.

6.5 Chapter Summary

Key Formula: Chapter 6 recap

$$\text{SE}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}, \qquad t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)},$$

$$\text{reject } H_0 \text{ at } 5\% \iff |t^{\text{act}}| > 1.96, \qquad CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1),$$

$$p\text{-value} = 2\Phi(-|t^{\text{act}}|), \qquad p < \alpha \iff |t^{\text{act}}| > c_\alpha.$$

Exam Focus: Chapter 6 takeaways

- Reproduce the five-step "big picture" recipe from memory, and be ready to apply the four-beat hypothesis-test layout (Hypotheses → Test statistic → Decision rule → Conclusion in context) to *any* coefficient, not just $\hat{\beta}_1$ against 0.
- The t-statistic, the confidence interval, and the p-value are three equivalent ways of asking the same question – a rejection at 5%, an excluded value from the 95% CI, and a p-value below 0.05 always agree (Theorems 6.2 and 6.3). If your numbers disagree, you have made an arithmetic error.
- Never drop the absolute value around t^{act}, never confuse "fail to reject" with "proved true," and never report a p-value or significance star without also reporting the point estimate and its economic magnitude.

7 Part I Summary: Simple Linear Regression (Lectures 1–6)

This chapter compresses Lectures 1–6 (Chapters 1–6 of these notes) for the **Week 7 Quiz**. It introduces no new material; every formula, theorem, and worked number below was already derived in Chapters 1–6 and is only being compared, connected, and drilled here. The order follows the house style for part-summary chapters: a comparison table, then Method Pattern boxes for every recurring archetype, then a bank of common mistakes, then integrated worked examples that force two or more chapters together, then a revision checklist, then a single master memory sheet.

How to use this chapter. First pass: read the comparison table and confirm you recognise every row without going back to Chapters 1–6. Second pass: cover the steps in each Method Pattern box and reproduce them from the one-line prompt alone. Third pass (the day before the quiz): the Common Mistakes bank and the five Integrated Worked Review Examples, timed.

The Week 7 Quiz draws on all six lectures roughly evenly, and the digest of exam-heavy themes (Reference/HE2003_digest.md, §6) flags five topics as tested from the widest range of angles: the closed-form OLS derivation, the residual-identity proof chain (feeding directly into $R^2 \leq 1$), Assumption 1 (diagrams, the RCT benchmark, return-to-education, selection bias), unbiasedness/efficiency/Gauss–Markov built from $\bar{Y}$, and the t-test/CI/p-value triad. Every one of those five is the subject of a dedicated Method Pattern below.

Quiz scope. The Week 7 Quiz does *not* draw on all of Chapter 6: it stops at the large-sample distribution of $\hat{\beta}_1$ and $\text{SE}(\hat{\beta}_1)$ (Lecture 6 slides 1–3), and sets no Stata-output questions. The t-test, confidence interval, and p-value (Lecture 6 slide 4 onward) are examined only in the Final. Boxes and checklist items below that fall past this boundary are labelled "(Final only)".

Exam Focus: Quiz scope

For the Quiz, be ready to standardise $\hat{\beta}_1$ using β_1 and $\sigma_{\hat{\beta}_1}$ (or $\text{SE}(\hat{\beta}_1)$) and to read off a probability from Φ – *not* to run a t-test, build a confidence interval, or compute a p-value. Those three remain examinable, but only in the Final.

Notation and symbol map. Quiz questions rarely lose marks on the formulas themselves – they lose marks on which of three near-identical symbols is meant at a given step. Two tables fix the mapping permanently before anything else in this chapter is used.

Symbol	What it is	Random or fixed?
(β_0, β_1)	true population parameters of $Y_i = \beta_0 + \beta_1 X_i + u_i$	fixed, unknown constants
(b_0, b_1)	generic candidate values in $\min_{b_0,b_1} \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$	dummy arguments, not attached to any sample
$(\hat{\beta}_0, \hat{\beta}_1)$	OLS estimators – the minimizers, for one particular sample	random: functions of the sample $\{(X_i, Y_i)\}$
u_i	true, unobservable error, $u_i = Y_i - (\beta_0 + \beta_1 X_i)$	random, never directly observed
$\hat{u}_i$	OLS residual, $\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$	random, observable, computed from the fitted sample
$\beta_{1,0}$	a hypothesised value of β_1 under H_0 (e.g. 0, or any constant c)	fixed number the analyst chooses; never estimated

Population (true, unknown)	**Sample (OLS, computable)**
Population regression line $\beta_0 + \beta_1 X$	OLS regression line $\hat{\beta}_0 + \hat{\beta}_1 X$
Error term u_i	Residual $\hat{u}_i$
Systematic part $\beta_0 + \beta_1 X_i$ (unobservable)	Predicted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
Population variance $\sigma^2_{\hat{\beta}_1}$ of the sampling distribution of $\hat{\beta}_1$	Estimated standard error $\mathrm{SE}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$

Reminder: never mix a population-column symbol with a sample-column symbol

Writing $\sum_{i=1}^{n} u_i = 0$ (population column) instead of $\sum_{i=1}^{n} \hat{u}_i = 0$ (sample column), or treating $\hat{\beta}_1$ as if it were β_1 itself midway through a proof, is the single most common notational slip across Chapters 2–6. Keep one eye on which column of both tables you are standing in at every line of working.

QRS@NTU — qrsntu.org

7.1 Comparison of Central Results

Topic	Result / formula	Assumptions needed	Where taught
OLS estimator	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$	none (algebra; needs $\sum_{i=1}^n (X_i - \bar{X})^2 > 0$)	L2; Q2/Tut2 Q3
Residual identities	$\sum_{i=1}^n \hat{u}_i = 0$, $\sum_{i=1}^n \hat{u}_i X_i = 0$, $\sum_{i=1}^n \hat{u}_i \hat{Y}_i = 0$	none (pure algebra from the two OLS FOCs)	L2–L3; Q3/Tut3 Q2
R^2 decomposition	$\text{TSS} = \text{ESS} + \text{SSR}$; $R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} \in [0,1]$	algebra, built on the residual identities above	L4; Q4–5/Tut4–5
Least Squares Assumptions 1–3	$E[u_i \mid X_i] = 0$; (X_i, Y_i) i.i.d.; large outliers unlikely	population conditions on the data-generating process (not automatic)	L3; Q3/Tut3
Unbiasedness / BLUE of $\bar{Y}$	$E[\bar{Y}] = \mu$, $\text{Var}(\bar{Y}) = \sigma^2/n \le \text{Var}(\hat{\mu})$ for any linear unbiased $\hat{\mu} = \sum_{i=1}^n a_i Y_i$	i.i.d. sampling; unbiasedness needs $\sum_{i=1}^n a_i = 1$	L4–L5; Q5/Tut5 Q1
Gauss–Markov (OLS is BLUE)	$\text{Var}(\hat{\beta}_0) \le \text{Var}(\tilde{\beta}_0)$, $\text{Var}(\hat{\beta}_1) \le \text{Var}(\tilde{\beta}_1)$ for any linear unbiased $\tilde{\beta}_0, \tilde{\beta}_1$	Assumptions 1–3 **+ homoskedasticity**	L5; Q5/Tut5 Q1(b)
Binary regressor	$\hat{\beta}_0 = \bar{Y}_{D=0}$, $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0} = E[Y \mid D{=}1] - E[Y \mid D{=}0]$	Assumption 1 for the *causal* reading; pure algebra for the OLS values	L3, L5; Q3/Tut3 Q4, Q5/Tut5
t-test	$t^{\text{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)}$; reject iff $\lvert t^{\text{act}} \rvert > 1.96$ (5%) or 2.58 (1%)	Assumptions 1–3 + large n (normal approximation)	L6 (Final)
Confidence interval	$CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$	Assumptions 1–3 + large n	L6 (Final)
p-value	$p = 2\Phi(-\lvert t^{\text{act}} \rvert)$; $p < \alpha \iff \lvert t^{\text{act}} \rvert > c_\alpha$	Assumptions 1–3 + large n	L6 (Final)

Exam Focus: reading the table

Two columns matter more than they look. The "assumptions needed" column tells you *what to check first* on a quiz question: is it pure algebra (residual identities, R^2 decomposition – true in *every* sample), does it need Assumptions 1–3 (unbiasedness, the shape of the sampling distribution), or does it need the extra homoskedasticity condition on top (only Gauss–Markov)? The "where taught" column is your citation key when an examiner asks you to "derive from first principles" – know which lecture's slide-for-slide derivation you are being asked to reproduce.
Each row also maps onto exactly one of the ten Method Pattern boxes below, so a quiz question that names a row is really asking you to run the matching pattern:

- OLS estimator → Method Pattern 1; residual identities and R^2 decomposition → Method Pattern 2;
- Unbiasedness/BLUE of $\bar{Y}$ → Method Pattern 3; binary regressor → Method Pattern 4; Assumptions 1–3 → Method Pattern 5;
- t-test → Method Pattern 6; confidence interval → Method Pattern 7; p-value → Method Pattern 8;
- Gauss–Markov is the one row with no dedicated pattern of its own – it is Method Pattern 3's variance-ranking argument, re-run with X-dependent weights instead of $1/n$ weights (the non-examinable extension in Chapter 5).

How the six lectures chain together. Nothing in Part I is a self-contained island; every lecture either supplies a tool the next one needs, or discharges a promise an earlier lecture made without proof. Seeing this chain explicitly is what lets you answer a quiz question that mixes, say, a residual-identity proof with an R^2 computation without treating them as two unrelated topics.

Exam Focus: the six-lecture dependency chain

- **L1 (Introduction)** supplies the probability toolkit – linearity of E, bilinearity of Cov, the Law of Iterated Expectations, independence ⇒ zero covariance – that every later unbiasedness proof cites by name; it also supplies the causal vocabulary (the three sources of correlation, RCTs, Type I/II errors) that frames Assumption 1 and every hypothesis test that follows.
- **L2 (The OLS Estimator)** derives $\hat{\beta}_0, \hat{\beta}_1$ from the FOCs of $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$ and defines $\hat{Y}_i, \hat{u}_i$ – the four objects every later chapter manipulates but never re-derives.
- **L3 (Least Squares Assumptions)** proves the residual identities purely from L2's two FOCs (no new assumption needed), then states Assumptions 1–3, the conditions under which $\hat{\beta}_0, \hat{\beta}_1$ are more than just "a best-fitting line through whatever data happened to be sampled."
- **L4 (Sampling, Fit, Distribution)** reuses L3's residual identities to prove TSS = ESS + SSR and $0 \leq R^2 \leq 1$; restates *why* Assumptions 1–3 matter (unbiasedness, efficiency, large-sample normality – all promised here, none yet proved); and derives $E[\bar{Y}], \mathrm{Var}(\bar{Y})$ as a rehearsal, in the simplest possible setting, for what L5 does to $\hat{\beta}_1$ itself.
- **L5 (Unbiasedness, Efficiency, Binary Regressors)** proves unbiasedness and the BLUE property for $\bar{Y}$ in full generality, states Gauss–Markov for $\hat{\beta}_0, \hat{\beta}_1$ by direct analogy, introduces the binary regressor as conditional means, and opens the hypothesis-testing vocabulary (H_0, H_1, Type I/II errors) that L6 turns into an operational test.
- **L6 (Inference with a Single Regressor)** converts L4–L5's large-sample normal approximation for $\hat{\beta}_1$ into a usable $\mathrm{SE}(\hat{\beta}_1)$, the t-statistic, the confidence interval, and the p-value – three equivalent ways of answering the one question every earlier chapter was building toward: is β_1 different from some hypothesised value, and how confident can we be?

7.2 Standard Problem Types and Solution Patterns

Method Pattern 1: derive the OLS estimator from the FOCs

When you see it. "Derive $\hat{\beta}_0, \hat{\beta}_1$" or "show that $\hat{\beta}_1 = s_{XY}/s_X^2$" with no numbers attached – a from-scratch algebra question, not a plug-in one.

1. Write the objective $S(b_0, b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$.
2. Differentiate with respect to b_0 and b_1; set both partial derivatives to zero (drop the harmless -2) to get the two normal equations.
3. Solve the first normal equation for $\hat{\beta}_0$ in terms of $\hat{\beta}_1$: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.
4. Substitute into the second normal equation; collect the terms in $\hat{\beta}_1$ and use $\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}$ (and the analogous X-only identity) to convert raw sums into deviation form.
5. Read off $\hat{\beta}_1 = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) / \sum_{i=1}^n (X_i - \bar{X})^2 = s_{XY}/s_X^2$, then back out $\hat{\beta}_0$.

Cite: L2; Question 2/Tutorial 2 Q3.

Method Pattern 2: the residual-identity proof chain, through $0 \le R^2 \le 1$

When you see it. "Show $\sum_{i=1}^n \hat{u}_i = 0$" or "show $R^2 \le 1$" – these are never independent facts to memorise separately; they are one five-step chain starting from the same two FOCs.

1. From the first FOC, $\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$, i.e. $\sum_{i=1}^n \hat{u}_i = 0$.
2. From the second FOC, $\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0$, i.e. $\sum_{i=1}^n \hat{u}_i X_i = 0$.
3. Expand $\sum_{i=1}^n \hat{u}_i \hat{Y}_i = \hat{\beta}_0 \sum_{i=1}^n \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^n \hat{u}_i X_i = 0$ using Steps 1–2.
4. Write $Y_i - \bar{Y} = \hat{u}_i + (\hat{Y}_i - \bar{Y})$, square, and sum; the cross term $2[\sum_{i=1}^n \hat{u}_i \hat{Y}_i - \bar{Y} \sum_{i=1}^n \hat{u}_i]$ vanishes by Steps 1 and 3.
5. Conclude TSS = ESS + SSR; divide by TSS > 0 to get $1 = R^2 + \text{SSR/TSS}$, hence $R^2 = 1 - \text{SSR/TSS} \in [0, 1]$, with $R^2 = 1$ iff every $\hat{u}_i = 0$.

Cite: L2–L3, L4; Question 3/Tutorial 3 Q2; Question 5/Tutorial 5 Q2.

Method Pattern 3: show a linear estimator is unbiased, or compare two variances

1. Write the candidate as $\hat{\mu} = \sum_{i=1}^n a_i Y_i$ for fixed (non-random) weights a_i.
2. Take $E[\cdot]$: $E[\hat{\mu}] = \sum_{i=1}^n a_i E[Y_i] = \mu \sum_{i=1}^n a_i$ (uses only that every Y_i has the same mean μ, i.e. identical distribution).
3. $\hat{\mu}$ is unbiased iff $\sum_{i=1}^n a_i = 1$.
4. The variance, using independence to kill every cross term, is $\text{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2$.
5. To rank two unbiased candidates, compute $\sum_{i=1}^n a_i^2$ for each; the smaller value wins. The trick $a_i = (a_i - \frac{1}{n}) + \frac{1}{n}$ shows $\sum_{i=1}^n a_i^2 = \sum_{i=1}^n (a_i - \frac{1}{n})^2 + \frac{1}{n} \ge \frac{1}{n}$, so $\bar{Y}$ (all $a_i = 1/n$) beats every other linear unbiased competitor.

Cite: L4–L5; Question 5/Tutorial 5 Q1.

Method Pattern 4: interpret a binary-regressor coefficient as conditional means

When you see it. X_i (or D_i) is explicitly 0/1-valued – gender, treatment, region, union status – and the question asks for an interpretation or the two group means.

1. Confirm $X_i \in \{0, 1\}$; never call $\hat{\beta}_1$ a "slope" – there are only two points on the X-axis.
2. Under Assumption 1, take $E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$ at $X_i = 0$ and $X_i = 1$: $E[Y \mid X{=}0] = \beta_0$, $E[Y \mid X{=}1] = \beta_0 + \beta_1$.
3. Read β_0 as the base-group mean and β_1 as the *difference* between the $X{=}1$ group mean and the base-group mean (the average treatment effect if X is randomly assigned).
4. On sample data this is exact, not approximate: $\hat{\beta}_0 = \bar{Y}_{D=0}$ and $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$ – no separate regression is needed once the two group means are known.

Cite: L3, L5; Question 3/Tutorial 3 Q4, Question 5/Tutorial 5.

Method Pattern 5: judge whether Assumption 1 holds in a new economic example

When you see it. A short-answer or essay prompt describing a new regression in an unfamiliar applied setting and asking whether the estimate can be read causally.

1. Write the model $Y = \beta_0 + \beta_1 X + u$ and name, concretely, what is plausibly bundled inside u for this context.
2. Ask: does the value of X (or inclusion in the sample) get *chosen* by units in a way correlated with something in u?
3. If yes, state the selection mechanism explicitly – who selects in (or is observed), and how they systematically differ on u from who does not.
4. Sign the bias: does the selection push $\mathrm{Cov}(X, u)$ positive or negative, and does that make $\hat{\beta}_1$ overstate or understate the true causal β_1?
5. Contrast with the RCT benchmark: only *random* assignment of X guarantees $E[u \mid X] = 0$ by construction; observational X always needs a case-by-case argument.

Cite: L1, L3, L4; Question 1/3/4 and the corresponding tutorials.

Method Pattern: probabilities from the sampling distribution of $\hat{\beta}_1$

When you see it. A question gives (or implies) β_1 and $\sigma_{\hat{\beta}_1}$ (or $\text{SE}(\hat{\beta}_1)$) and asks for $\Pr(\hat{\beta}_1 > c)$, $\Pr(\hat{\beta}_1 < c)$, or $\Pr(|\hat{\beta}_1 - \beta_1| < c)$ – no hypothesis test, CI, or p-value is requested.

1. **Identify.** Note the mean β_1 and the standard deviation $\sigma_{\hat{\beta}_1}$ (population) or $\text{SE}(\hat{\beta}_1)$ (estimated) of the large-sample distribution $\hat{\beta}_1 \approx N(\beta_1, \sigma^2_{\hat{\beta}_1})$.
2. **Standardise.** Convert the event on $\hat{\beta}_1$ into an event on $Z = (\hat{\beta}_1 - \beta_1)/\sigma_{\hat{\beta}_1}$ (or $/\,\text{SE}(\hat{\beta}_1)$), which is approximately $N(0,1)$.
3. **Read Φ.** Look up (or compute) the corresponding standard normal probability, e.g. $1 - \Phi(z)$ for a one-sided upper tail, or $2\Phi(z) - 1$ for $\Pr(|Z| < z)$.
4. **State the answer in words.** Report the probability in context (e.g. "there is about a 5.5% chance that $\hat{\beta}_1$ exceeds -1.2").

Cite: L6 (slides 1–3; Quiz scope).

Method Pattern 6: run a two-sided t-test (four beats) (Final only)

When you see it. A reported coefficient and its standard error, and a claim to test – "the effect is zero," "the effect is exactly c," "the gap is significant."

1. **Hypotheses.** $H_0 : \beta_1 = \beta_{1,0}$ vs. $H_1 : \beta_1 \neq \beta_{1,0}$.
2. **Test statistic.** $t^{\text{act}} = (\hat{\beta}_1 - \beta_{1,0})/\,\text{SE}(\hat{\beta}_1)$.
3. **Decision rule.** Reject at 5% iff $|t^{\text{act}}| > 1.96$; at 1% iff $|t^{\text{act}}| > 2.58$.
4. **Conclusion in context.** State the decision *and* what it means for the underlying economic question, not just "reject/do not reject."

Cite: L6.

Method Pattern 7: build and use a confidence interval (Final only)

When you see it. "Construct a 95% CI" or "is c a plausible value of β_1?" – both are answered by the same interval, built once.

1. Start from the duality: $CI_{95\%}(\beta_1) = \{\beta_{1,0} : |\hat{\beta}_1 - \beta_{1,0}|/\,\text{SE}(\hat{\beta}_1) \leq 1.96\}$.
2. Simplify to $\hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$.
3. To test any $H_0 : \beta_1 = c$ at 5%, just check whether c falls inside the interval – this is exactly equivalent to the t-test and needs no separate calculation.
4. For a scaled quantity such as $\hat{\beta}_1 \Delta X$, scale *both* the point estimate and the margin by $|\Delta X|$: $\hat{\beta}_1 \Delta X \pm 1.96\ \text{SE}(\hat{\beta}_1)\,|\Delta X|$.

Cite: L6.

Method Pattern 8: read and use a p-value (Final only)

When you see it. Software output that reports a p-value directly instead of (or alongside) t^{act} and asks you to state a conclusion at a given significance level.

1. Compute $p = 2\Phi(-|t^{\text{act}}|)$.
2. Compare p to the significance level α: $p < \alpha \iff |t^{\text{act}}| > c_\alpha$, for *any* α, not just 5% or 1%.
3. Never report p alone; pair it with the point estimate (or CI) to speak to economic, not just statistical, significance.
4. Use monotonicity as a sanity check: a larger $|t^{\text{act}}|$ always gives a smaller p, and vice versa – if your numbers disagree, recheck the arithmetic.

Cite: L6.

Method Pattern 9: compute a prediction, or a change in prediction, with units

When you see it. A fitted equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ and a specific X value, or a proposed change ΔX, with no test or interval requested.

1. Plug the given X into $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$; keep the units of Y attached throughout.
2. For a *change*, use $\Delta\hat{Y} = \hat{\beta}_1 \, \Delta X$ directly – do not compute two separate predictions and subtract unless the question explicitly asks you to cross-check both routes.
3. If a confidence interval is wanted for the change, scale the margin by $|\Delta X|$: $\hat{\beta}_1 \Delta X \pm 1.96 \, \text{SE}(\hat{\beta}_1) \, |\Delta X|$.
4. Flag, rather than silently perform, any extrapolation to an X value far outside the range of the data used to fit the line.

Cite: L2, L4; Question 2/4.

Method Pattern 10: mean and variance of a difference of independent sample means

When you see it. Two separate populations (two countries, treatment vs. control, before vs. after with independent samples) and a question about the gap between their means.

1. Write $\hat{d} = \bar{Y}_1 - \bar{Y}_2$ from two *independently* drawn samples of sizes N_1, N_2.
2. $E[\hat{d}] = \mu_1 - \mu_2$ by linearity of expectation alone – independence is not needed for this step.
3. Independence of the two samples gives $\text{Cov}(\bar{Y}_1, \bar{Y}_2) = 0$.
4. Hence $\text{Var}(\hat{d}) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2) = \sigma_1^2/N_1 + \sigma_2^2/N_2$, with no cross term.

Cite: L5; Question 5/Tutorial 5 Q3.

7.3 Common Mistakes and Exam Traps

Trap 1: u_i is not $\hat{u}_i$

u_i is the true, unobservable population error; $\hat{u}_i$ is the observable OLS residual computed from one particular sample. Only $\sum_{i=1}^{n} \hat{u}_i = 0$ holds, and it holds *automatically* by construction of the OLS FOCs – never write $\sum_{i=1}^{n} u_i = 0$, since u_i is a random population deviation with no such guarantee. *Quick fix.* Before writing any summation identity, ask "is this symbol hatted?" – if the claim would need u_i (no hat) to sum to zero, it is false in general.

Trap 2: conditional vs. unconditional, and zero correlation vs. independence

$E[u] = 0$ (unconditional) is automatic by convention; Assumption 1 requires the much stronger $E[u \mid X] = 0$ at *every* value of X. Likewise, $E[u \mid X] = 0 \Rightarrow \text{Cov}(X, u) = 0$, but zero correlation does *not* imply $E[u \mid X] = 0$ (Chapter 3's $u = X^2 - 1$ counterexample), and zero correlation never implies full independence (Chapter 1). All three are progressively stronger statements; do not treat any of them as interchangeable. *Quick fix.* Rank the three claims from weakest to strongest before answering: $E[u] = 0 \Leftarrow E[u \mid X] = 0 \Rightarrow \text{Cov}(X, u) = 0 \Leftarrow$ independence – and never reverse an arrow.

Trap 3: reading a binary-regressor $\hat{\beta}_1$ as a slope

When $X_i \in \{0, 1\}$, $\hat{\beta}_1$ is a *difference in group means*, not a rate of change – there is no meaningful "one-unit increase in X" to speak of. Report it as "the D=1 group's mean exceeds the D=0 group's mean by $\hat{\beta}_1$," never as "Y rises by $\hat{\beta}_1$ per unit of X." *Quick fix.* Before interpreting any coefficient, check whether its regressor can take values other than 0 and 1; if not, switch immediately from "slope" language to "group-mean gap" language.

Trap 4: dropping the absolute value on t^{act}

The decision rule compares $|t^{\text{act}}|$ to the critical value, never t^{act} itself. A large *negative* t-statistic is exactly as strong evidence against H_0 as a large positive one; writing "reject if $t^{\text{act}} > 1.96$" silently discards every negative rejection. *Quick fix.* Compute t^{act}, then immediately write $|t^{\text{act}}| = \ldots$ on the next line before comparing to 1.96 or 2.58 – make the absolute value a physical step, not a mental one.

Trap 5: testing $\beta_1 = c \neq 0$ with $\hat{\beta}_1 / \text{SE}(\hat{\beta}_1)$

When the null is not $\beta_1 = 0$, the numerator of the t-statistic must be $\hat{\beta}_1 - c$, not $\hat{\beta}_1$ alone. Computing $\hat{\beta}_1 / \text{SE}(\hat{\beta}_1)$ answers "is β_1 different from 0?", a different question from the one asked, and leads to the wrong conclusion whenever $c \neq 0$ (Chapter 6's CAPM $H_0 : \beta = 1$ example). *Quick fix.* Write out $H_0 : \beta_1 =$ ___ explicitly before touching a calculator; whatever number fills the blank is what gets subtracted from $\hat{\beta}_1$.

Trap 6: "95% probability that β_1 is in this interval"

Once the sample is drawn and $[\hat{\beta}_1 - 1.96\,\text{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 1.96\,\text{SE}(\hat{\beta}_1)]$ is a fixed pair of numbers, β_1 either is or is not in it – nothing is left random. The correct statement is about the *procedure*: repeated over many samples, about 95% of such intervals would contain the fixed, unknown β_1. *Quick fix.* Say "95% of intervals built this way would contain β_1," never "there is a 95% chance β_1 is in *this* interval."

Trap 7: "fail to reject" is not "accept as true"

Failing to reject H_0 (e.g. $|t^{\text{act}}| \leq 1.96$) means the data are not sufficiently inconsistent with H_0 to rule it out at that significance level – it does not prove H_0, and a low-power test (small n, or little variation in X) can easily fail to reject a false H_0. *Quick fix.* Write "fail to reject H_0" or "the data are consistent with H_0," and never write "we accept H_0" or "H_0 is proven true."

Trap 8: a low R^2 is not a "bad" regression

R^2 only measures how much of Y's sample variation X explains; it says nothing about whether $\hat{\beta}_1$ is an accurate, causally valid estimate. A regression can have a tiny R^2 and a precisely estimated, causally valid $\hat{\beta}_1$, or a large R^2 (e.g. from an omitted-variable-driven correlation) and a badly biased $\hat{\beta}_1$ – whether $\hat{\beta}_1$ is trustworthy is an Assumption-1 question, never an R^2 question. *Quick fix.* Answer "how well does X fit Y" with R^2, and "is $\hat{\beta}_1$ believable as a causal effect" with Assumption 1 – never let one question answer the other.

Trap 9: causal versus predictive interpretation of $\hat{\beta}_1$

A statistically significant $\hat{\beta}_1$ (tiny p-value, or a CI excluding 0) shows only that $\hat{\beta}_1$ is precisely estimated to differ from 0 – *as a description of the sample*. Reading $\hat{\beta}_1$ as the causal effect of X on Y additionally requires Assumption 1 to hold; without it, $\hat{\beta}_1$ remains a valid predictor of Y from X but not a valid estimate of a causal effect (the return-to-education and class-size examples of Chapters 3–4). *Quick fix.* After any significance test, ask separately: "would I trust this number as a *causal* effect?" – statistical significance and Assumption 1 are independent hurdles, and a question can fail either one alone.

Trap 10: a missing or misplaced $1/n$ in the summation identities

Question 2 Q4's printed identity $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_iY_i - \bar{X}\bar{Y}$ has a $\frac{1}{n}$ on the right but not the left; it is only correct with $\frac{1}{n}$ on *both* sides, or on *neither* side (the deviation-sum form actually used inside the OLS derivation). For $n = 420$ (the California data), using the wrong version is a two-orders-of-magnitude arithmetic error – always check which of the two internally consistent forms you are using before substituting into $\hat{\beta}_1$. *Quick fix.* Re-derive the identity from $(X_i - \bar{X})(Y_i - \bar{Y}) = X_iY_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y}$ on the spot if unsure, rather than trusting a half-remembered formula sheet.

7.4 Integrated Worked Review Examples

Integrated Review Example 1: OLS by hand, residual check, and R^2 (Lectures 2–4)

A tutor records weekly one-on-one tutoring hours (X) and the resulting improvement in a student's quiz score (Y, points) for $n = 5$ students:

i	1	2	3	4	5
X_i	2	4	6	8	10
Y_i	5	6	8	7	10

(a) $\hat{\beta}_0, \hat{\beta}_1$.
Method 1: Closed-form deviation formula. The sample means are

$$\bar{X} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6,$$
$$\bar{Y} = \frac{5+6+8+7+10}{5} = \frac{36}{5} = 7.2.$$

Subtract the means from each observation:

$$X_i - \bar{X} = (2-6,\ 4-6,\ 6-6,\ 8-6,\ 10-6) = (-4,-2,0,2,4),$$
$$Y_i - \bar{Y} = (5-7.2,\ 6-7.2,\ 8-7.2,\ 7-7.2,\ 10-7.2) = (-2.2,-1.2,0.8,-0.2,2.8).$$

Multiply the deviations pairwise and square the X-deviations:

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = (-4)(-2.2)+(-2)(-1.2)+(0)(0.8)+(2)(-0.2)+(4)(2.8)$$
$$= 8.8+2.4+0-0.4+11.2 = 22.0,$$
$$\sum_{i=1}^{n}(X_i-\bar{X})^2 = (-4)^2+(-2)^2+0^2+2^2+4^2 = 16+4+0+4+16 = 40.$$

Hence

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{22.0}{40} = 0.55,$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 7.2 - 0.55(6) = 7.2 - 3.3 = 3.9.$$

$$\boxed{\hat{\beta}_0 = 3.9, \qquad \hat{\beta}_1 = 0.55, \qquad \hat{Y}_i = 3.9 + 0.55X_i.}$$

Method 2: Normal equations (FOCs) in raw sums. Setting both partial derivatives of $S(b_0,b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ to zero and dividing by -2 gives

$$\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i) = 0 \implies \sum_{i=1}^{n}Y_i = n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}X_i,$$
$$\sum_{i=1}^{n}(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)X_i = 0 \implies \sum_{i=1}^{n}X_iY_i = \hat{\beta}_0\sum_{i=1}^{n}X_i + \hat{\beta}_1\sum_{i=1}^{n}X_i^2.$$

The four raw sums are

$$\sum_{i=1}^{n}X_i = 2+4+6+8+10 = 30, \qquad \sum_{i=1}^{n}Y_i = 5+6+8+7+10 = 36,$$
$$\sum_{i=1}^{n}X_iY_i = 2(5)+4(6)+6(8)+8(7)+10(10) = 10+24+48+56+100 = 238,$$
$$\sum_{i=1}^{n}X_i^2 = 4+16+36+64+100 = 220.$$

With $n=5$ the normal equations read

$$36 = 5\hat{\beta}_0 + 30\hat{\beta}_1,$$
$$238 = 30\hat{\beta}_0 + 220\hat{\beta}_1.$$

Divide the first by 5 to get $\hat{\beta}_0 = 7.2 - 6\hat{\beta}_1$, and substitute into the second:

$$238 = 30(7.2-6\hat{\beta}_1) + 220\hat{\beta}_1$$
$$= 216 - 180\hat{\beta}_1 + 220\hat{\beta}_1$$
$$= 216 + 40\hat{\beta}_1$$
$$\implies 40\hat{\beta}_1 = 22 \implies \hat{\beta}_1 = 0.55,$$
$$\hat{\beta}_0 = 7.2 - 6(0.55) = 7.2 - 3.3 = 3.9.$$

$$\boxed{\hat{\beta}_0 = 3.9, \qquad \hat{\beta}_1 = 0.55, \qquad \hat{Y}_i = 3.9 + 0.55X_i.}$$

(b) Fitted values and residuals.
Method 1: Observation-by-observation residuals. Substitute each X_i into $\hat{Y}_i = 3.9 + 0.55X_i$ and subtract from Y_i:

$$\begin{aligned}
\hat{Y}_1 &= 3.9 + 0.55(2) = 3.9 + 1.1 = 5.0, & \hat{u}_1 &= 5 - 5.0 = 0,\\
\hat{Y}_2 &= 3.9 + 0.55(4) = 3.9 + 2.2 = 6.1, & \hat{u}_2 &= 6 - 6.1 = -0.1,\\
\hat{Y}_3 &= 3.9 + 0.55(6) = 3.9 + 3.3 = 7.2, & \hat{u}_3 &= 8 - 7.2 = 0.8,\\
\hat{Y}_4 &= 3.9 + 0.55(8) = 3.9 + 4.4 = 8.3, & \hat{u}_4 &= 7 - 8.3 = -1.3,\\
\hat{Y}_5 &= 3.9 + 0.55(10) = 3.9 + 5.5 = 9.4, & \hat{u}_5 &= 10 - 9.4 = 0.6.
\end{aligned}$$

Adding the five residuals:

$$\boxed{\sum_{i=1}^{n} \hat{u}_i = 0 - 0.1 + 0.8 - 1.3 + 0.6 = 0.}$$

Method 2: First normal equation (orthogonality to the constant). Expand the sum of residuals and use $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i &= \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i \\
&= n\big(\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1\bar{X}\big) = n\big(\bar{Y} - (\bar{Y} - \hat{\beta}_1\bar{X}) - \hat{\beta}_1\bar{X}\big) = 0.
\end{aligned}$$

With the numbers of part (a):

$$\sum_{i=1}^{n} \hat{u}_i = 36 - 5(3.9) - 0.55(30) = 36 - 19.5 - 16.5 = 0.$$

The five residuals $Y_i - 3.9 - 0.55X_i$ that make up this total are $5 - 3.9 - 1.1 = 0$, $6 - 3.9 - 2.2 = -0.1$, $8 - 3.9 - 3.3 = 0.8$, $7 - 3.9 - 4.4 = -1.3$, $10 - 3.9 - 5.5 = 0.6$, so

$$\boxed{\sum_{i=1}^{n} \hat{u}_i = 0 - 0.1 + 0.8 - 1.3 + 0.6 = 0.}$$

(c) R^2.
Method 1: R^2 = ESS/TSS from the three sums of squares. Using $Y_i - \bar{Y} = (-2.2, -1.2, 0.8, -0.2, 2.8)$ from part (a),

$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (-2.2)^2 + (-1.2)^2 + 0.8^2 + (-0.2)^2 + 2.8^2 \\
&= 4.84 + 1.44 + 0.64 + 0.04 + 7.84 = 14.8.
\end{aligned}$$

The fitted values minus the mean are $\hat{Y}_i - \bar{Y} = (5.0 - 7.2,\ 6.1 - 7.2,\ 7.2 - 7.2,\ 8.3 - 7.2,\ 9.4 - 7.2) = (-2.2, -1.1, 0, 1.1, 2.2)$, so

$$\begin{aligned}
\text{ESS} &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = (-2.2)^2 + (-1.1)^2 + 0^2 + 1.1^2 + 2.2^2 \\
&= 4.84 + 1.21 + 0 + 1.21 + 4.84 = 12.1.
\end{aligned}$$

Using the residuals of part (b),

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^{n} \hat{u}_i^2 = 0^2 + (-0.1)^2 + 0.8^2 + (-1.3)^2 + 0.6^2 \\ &= 0 + 0.01 + 0.64 + 1.69 + 0.36 = 2.7.\end{aligned}$$

Therefore

$$\boxed{\text{TSS} = \text{ESS} + \text{SSR} = 12.1 + 2.7 = 14.8\ (\checkmark), \qquad R^2 = \frac{12.1}{14.8} \approx 0.8176.}$$

Method 2: $R^2 = 1 - \text{SSR}/\text{TSS}$, with ESS from the slope. Since $\hat{Y}_i - \bar{Y} = (\hat{\beta}_0 + \hat{\beta}_1 X_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) = \hat{\beta}_1 (X_i - \bar{X})$,

$$\text{ESS} = \sum_{i=1}^{n} \hat{\beta}_1^2 (X_i - \bar{X})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0.55^2(40) = 0.3025(40) = 12.1.$$

With TSS = 14.8 and SSR = 2.7 as in Method 1,

$$\begin{aligned}R^2 &= 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{2.7}{14.8} = \frac{14.8 - 2.7}{14.8} = \frac{12.1}{14.8} \approx 0.8176, \\ \text{ESS} + \text{SSR} &= 12.1 + 2.7 = 14.8 = \text{TSS}.\end{aligned}$$

$$\boxed{\text{TSS} = \text{ESS} + \text{SSR} = 12.1 + 2.7 = 14.8\ (\checkmark), \qquad R^2 = \frac{12.1}{14.8} \approx 0.8176.}$$

Method 3: Squared sample correlation. The sample correlation uses the three deviation sums already computed:

$$r_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{22.0}{\sqrt{40(14.8)}} = \frac{22.0}{\sqrt{592}}.$$

Hence

$$\begin{aligned}r_{XY}^2 &= \frac{22.0^2}{592} = \frac{484}{592} = \frac{12.1}{14.8} \approx 0.8176, \\ \text{ESS} &= r_{XY}^2\, \text{TSS} = \frac{484}{592}(14.8) = \frac{484}{40} = 12.1, \qquad \text{SSR} = \text{TSS} - \text{ESS} = 14.8 - 12.1 = 2.7.\end{aligned}$$

$$\boxed{\text{TSS} = \text{ESS} + \text{SSR} = 12.1 + 2.7 = 14.8\ (\checkmark), \qquad R^2 = \frac{12.1}{14.8} \approx 0.8176.}$$

(d) Mean-point check at $(\bar{X}, \bar{Y}) = (6, 7.2)$.
Method 1: Substitution of $\bar{X}$ into the OLS line. Evaluate the fitted line at $X = \bar{X} = 6$:

$$\boxed{\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = 3.9 + 0.55(6) = 3.9 + 3.3 = 7.2 = \bar{Y}\ (\checkmark)}$$

Method 2: Average of the fitted values. Average $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ over i, and use $\hat{Y}_i = Y_i - \hat{u}_i$ with $\sum_{i=1}^{n} \hat{u}_i = 0$ from part (b):

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X}, \\ \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i &= \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = \bar{Y} - 0 = \bar{Y}.\end{aligned}$$

Numerically, $\frac{1}{5}(5.0 + 6.1 + 7.2 + 8.3 + 9.4) = \frac{36}{5} = 7.2$, so

$$\boxed{\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = 3.9 + 0.55(6) = 3.9 + 3.3 = 7.2 = \bar{Y}\ (\checkmark)}$$

Integrated Review Example 2: full inference from a reported regression (Lecture 6) (Final only)

A campus dining survey estimates, for $n = 85$ students, the relationship between weekly cups of coffee (X) and a self-reported productivity index (Y, 0–100 scale):

$$\widehat{\text{Productivity}} = \underset{(3.1)}{45.2} + \underset{(1.4)}{3.8}\ \text{Coffee}.$$

(a) t-test on β_1.
Method 1: Two-sided t-test against the critical values.
Hypotheses. Let β_1 be the population difference in mean productivity index (points) per additional weekly cup of coffee. Test

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 \neq 0.$$

Test statistic. With $\hat{\beta}_1 = 3.8$, $\text{SE}(\hat{\beta}_1) = 1.4$ and $\beta_{1,0} = 0$,

$$t^{\text{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)} = \frac{3.8 - 0}{1.4} = 2.714,$$
$$|t^{\text{act}}| = 2.714.$$

Decision rule. Under Least Squares Assumptions 1–3 and $n = 85$ (large), t^{act} is approximately $N(0, 1)$ under H_0. The two-sided 5% critical value solves $\Pr(|Z| > c) = 0.05$, i.e. $\Phi(c) = 0.975$, so $c = 1.96$; the 1% value solves $\Phi(c) = 0.995$, so $c = 2.58$. Reject H_0 at 5% iff $|t^{\text{act}}| > 1.96$, and at 1% iff $|t^{\text{act}}| > 2.58$.
Conclusion. Since $2.714 > 1.96$ and $2.714 > 2.58$, H_0 is rejected at both levels: the Coffee coefficient of 3.8 index points per weekly cup is statistically significantly different from zero at the 1% level (and hence also at 5%).

$$\boxed{t^{\text{act}} \approx 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 2: Confidence-interval duality. $H_0 : \beta_1 = 0$ is rejected at level α iff 0 lies outside the $(1 - \alpha)$ confidence interval $\hat{\beta}_1 \pm c_\alpha \,\text{SE}(\hat{\beta}_1)$. The two intervals are

$$CI_{95\%} = 3.8 \pm 1.96(1.4) = 3.8 \pm 2.744 = [1.056,\ 6.544],$$
$$CI_{99\%} = 3.8 \pm 2.58(1.4) = 3.8 \pm 3.612 = [0.188,\ 7.412].$$

Both lower endpoints are positive, so 0 lies outside both intervals. The distance from the centre 3.8 to 0 in standard-error units is $(3.8 - 0)/1.4 = 2.714$, larger than both half-widths 1.96 and 2.58 (in standard-error units).

$$\boxed{t^{\text{act}} \approx 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

Method 3: p-value comparison. With $t^{\text{act}} = (3.8 - 0)/1.4 = 2.714$, part (c) gives

$$p = 2\Phi(-2.714) \approx 0.0066.$$

Reject at level α iff $p < \alpha$: here $0.0066 < 0.01$ and $0.0066 < 0.05$.

$$\boxed{t^{\text{act}} \approx 2.71, \quad \text{reject } H_0 : \beta_1 = 0 \text{ at 5\% and 1\%.}}$$

(b) 95% confidence interval.

Method 1: Estimate plus or minus 1.96 standard errors. The margin is $1.96\ \mathrm{SE}(\hat{\beta}_1) = 1.96(1.4) = 2.744$, so

$$\hat{\beta}_1 - 1.96\ \mathrm{SE}(\hat{\beta}_1) = 3.8 - 2.744 = 1.056,$$
$$\hat{\beta}_1 + 1.96\ \mathrm{SE}(\hat{\beta}_1) = 3.8 + 2.744 = 6.544.$$

$$\boxed{CI_{95\%}(\beta_1) \approx [1.06,\ 6.54]}$$

Method 2: Inverting the 5% t-test. The 95% interval is the set of hypothesised values $\beta_{1,0}$ that a 5% two-sided test does not reject:

$$\begin{aligned}\frac{|3.8-\beta_{1,0}|}{1.4} \le 1.96 &\iff -1.96 \le \frac{3.8-\beta_{1,0}}{1.4} \le 1.96\\ &\iff -2.744 \le 3.8-\beta_{1,0} \le 2.744\\ &\iff 3.8-2.744 \le \beta_{1,0} \le 3.8+2.744\\ &\iff 1.056 \le \beta_{1,0} \le 6.544.\end{aligned}$$

$$\boxed{CI_{95\%}(\beta_1) \approx [1.06,\ 6.54]}$$

(c) p-value.
Method 1: Lower-tail doubling. The two-sided p-value is $p = 2\Phi(-|t^{\mathrm{act}}|) = 2\Phi(-2.714)$. From the normal table, $\Phi(-2.71) = 0.00336$ and $\Phi(-2.72) = 0.00326$; interpolating 0.4 of the way,

$$\Phi(-2.714) \approx 0.00336 - 0.4(0.00010) = 0.00332,$$
$$p \approx 2(0.00332) = 0.00664.$$

$$\boxed{p \approx 0.0066 < 0.01}$$

Method 2: Complement of the central probability. p is the probability, under H_0, of a $|Z|$ at least as large as 2.714:

$$\begin{aligned}p &= 1 - \Pr(-2.714 \le Z \le 2.714) = 1 - \big[\Phi(2.714) - \Phi(-2.714)\big]\\ &= 1 - \big[2\Phi(2.714) - 1\big] = 2 - 2\Phi(2.714).\end{aligned}$$

From the table, $\Phi(2.71) = 0.99664$ and $\Phi(2.72) = 0.99674$, so $\Phi(2.714) \approx 0.99664 + 0.4(0.00010) = 0.99668$ and

$$p \approx 2 - 2(0.99668) = 2 - 1.99336 = 0.00664.$$

$$\boxed{p \approx 0.0066 < 0.01}$$

(d) Predicted change for 3 extra cups a week.
Method 1: Scaled coefficient and scaled standard error. For $\Delta X = 3$ cups per week the predicted change is $\hat{\beta}_1 \Delta X$. Since $\mathrm{Var}(3\hat{\beta}_1) = 3^2\,\mathrm{Var}(\hat{\beta}_1)$, its standard error is $3\ \mathrm{SE}(\hat{\beta}_1)$:

$$\begin{aligned}\widehat{\Delta \mathrm{Productivity}} &= \hat{\beta}_1(3) = 3.8(3) = 11.4,\\ \mathrm{SE}(3\hat{\beta}_1) &= 3(1.4) = 4.2,\\ CI_{95\%} &= 11.4 \pm 1.96(4.2) = 11.4 \pm 8.232 = [3.168,\ 19.632].\end{aligned}$$

$$\boxed{\text{predicted gain} = 11.4 \text{ points}, \quad CI_{95\%} = [3.17,\ 19.63]}$$

Method 2: Difference of two predictions and a rescaled interval for β_1. For any coffee level X,

$$\begin{aligned}\widehat{\text{Productivity}}(X+3) - \widehat{\text{Productivity}}(X) &= \big[45.2 + 3.8(X+3)\big] - \big[45.2 + 3.8X\big] \\ &= 3.8(3) = 11.4.\end{aligned}$$

The population change is $3\beta_1$, and multiplying by $3 > 0$ preserves order, so from part (b)

$$\begin{aligned}1.056 \le \beta_1 \le 6.544 &\iff 3(1.056) \le 3\beta_1 \le 3(6.544) \\ &\iff 3.168 \le 3\beta_1 \le 19.632.\end{aligned}$$

$$\boxed{\text{predicted gain} = 11.4 \text{ points}, \quad CI_{95\%} = [3.17,\ 19.63]}$$

(e) Point prediction at Coffee $= 10$.
Method 1: Substitution into the OLS line.

$$\widehat{\text{Productivity}}\big|_{\text{Coffee}=10} = 45.2 + 3.8(10) = 45.2 + 38 = 83.2$$

The value 83.2 lies inside the 0–100 range of the index.

$$\boxed{\widehat{\text{Productivity}}\big|_{\text{Coffee}=10} = 83.2 \text{ points (out of 100).}}$$

Integrated Review Example 3: binary regressor – group means and the test of the gap (Lectures 5–6) (Final only)

A labour economist regresses monthly wage (Y, SGD hundreds) on $\text{Union}_i \in \{0, 1\}$ (union membership) for $n = 300$ workers:

$$\widehat{\text{Wage}} = \underset{(0.9)}{18.5} + \underset{(1.1)}{2.7}\ \text{Union}.$$

(a) Group means (Quiz scope).
Method 1: Fitted values at Union $= 0$ and Union $= 1$. For a binary regressor the OLS fitted value in each group equals that group's sample mean: $\hat\beta_0 = \bar Y_{\text{Union}=0}$ and $\hat\beta_0 + \hat\beta_1 = \bar Y_{\text{Union}=1}$ (Lecture 5). Evaluate the fitted line at the two values of the regressor:

$$\begin{aligned}\widehat{\text{Wage}}\big|_{\text{Union}=0} &= 18.5 + 2.7(0) = 18.5, \\ \widehat{\text{Wage}}\big|_{\text{Union}=1} &= 18.5 + 2.7(1) = 21.2.\end{aligned}$$

In SGD this is 1,850 per month for non-members and 2,120 per month for members.

$$\boxed{\text{non-union mean wage} = 18.5, \quad \text{union mean wage} = 18.5 + 2.7 = 21.2\ (\times 100 \text{ SGD}).}$$

Method 2: OLS slope formula with a 0/1 regressor. Write $D_i = \text{Union}_i$, $n_1 = \sum_{i=1}^n D_i$ (members), $n_0 = n - n_1$, $\bar D = n_1/n$, and $\bar Y_1, \bar Y_0$ for the two group means, so that $\sum_{i=1}^n D_i Y_i = n_1 \bar Y_1$ and $n\bar Y = n_1 \bar Y_1 + n_0 \bar Y_0$. Both groups are non-empty ($n_1 \ge 1$ and $n_0 \ge 1$), so the denominator $\sum_{i=1}^n (D_i - \bar D)^2$ computed next is positive. Since $D_i^2 = D_i$,

$$\sum_{i=1}^n (D_i - \bar D)^2 = \sum_{i=1}^n D_i^2 - n\bar D^2 = n_1 - \frac{n_1^2}{n} = \frac{n_1(n - n_1)}{n} = \frac{n_1 n_0}{n}.$$

Since $\sum_{i=1}^{n}(D_i-\bar{D})=n_1-n\bar{D}=0$, the term $\bar{Y}\sum_{i=1}^{n}(D_i-\bar{D})$ drops out of the cross-product sum:

$$\begin{aligned}\sum_{i=1}^{n}(D_i-\bar{D})(Y_i-\bar{Y}) &= \sum_{i=1}^{n}(D_i-\bar{D})Y_i-\bar{Y}\sum_{i=1}^{n}(D_i-\bar{D})=\sum_{i=1}^{n}D_iY_i-\bar{D}\sum_{i=1}^{n}Y_i\\ &= n_1\bar{Y}_1-\frac{n_1}{n}\big(n_1\bar{Y}_1+n_0\bar{Y}_0\big)=n_1\bar{Y}_1\Big(1-\frac{n_1}{n}\Big)-\frac{n_1n_0}{n}\bar{Y}_0\\ &= \frac{n_1n_0}{n}\big(\bar{Y}_1-\bar{Y}_0\big).\end{aligned}$$

Hence

$$\begin{aligned}\hat{\beta}_1 &= \frac{(n_1n_0/n)(\bar{Y}_1-\bar{Y}_0)}{n_1n_0/n}=\bar{Y}_1-\bar{Y}_0,\\ \hat{\beta}_0 &= \bar{Y}-\hat{\beta}_1\bar{D}=\frac{n_1\bar{Y}_1+n_0\bar{Y}_0}{n}-\frac{n_1}{n}\big(\bar{Y}_1-\bar{Y}_0\big)=\frac{(n_0+n_1)\bar{Y}_0}{n}=\bar{Y}_0.\end{aligned}$$

Inverting with the reported estimates, $\bar{Y}_0=\hat{\beta}_0=18.5$ and $\bar{Y}_1=\hat{\beta}_0+\hat{\beta}_1=18.5+2.7=21.2$.

$$\boxed{\text{non-union mean wage} = 18.5, \quad \text{union mean wage} = 18.5+2.7 = 21.2\ (\times 100\ \text{SGD}).}$$

(b) Significance of the gap.
Method 1: Two-sided t-test against the critical values.
Hypotheses. β_1 is the population difference in mean monthly wage (SGD hundreds) between union and non-union workers. Test

$$H_0:\beta_1=0 \qquad \text{vs.} \qquad H_1:\beta_1\neq 0.$$

Test statistic. With $\hat{\beta}_1=2.7$, $\mathrm{SE}(\hat{\beta}_1)=1.1$ and $\beta_{1,0}=0$,

$$\begin{aligned}t^{\text{act}} &= \frac{\hat{\beta}_1-\beta_{1,0}}{\mathrm{SE}(\hat{\beta}_1)}=\frac{2.7-0}{1.1}=2.455,\\ |t^{\text{act}}| &= 2.455.\end{aligned}$$

Decision rule. Under Least Squares Assumptions 1–3 and $n=300$ (large), t^{act} is approximately $N(0,1)$ under H_0, so the two-sided critical values are 1.96 ($\Phi(1.96)=0.975$) at 5% and 2.58 ($\Phi(2.58)=0.995$) at 1%. Reject at 5% iff $|t^{\text{act}}|>1.96$, and at 1% iff $|t^{\text{act}}|>2.58$.
Conclusion. Since $1.96<2.455<2.58$, the union–non-union difference in mean monthly wage of 2.7 hundred SGD (SGD 270) is statistically significant at the 5% level but not at the 1% level.

$$\boxed{t^{\text{act}}\approx 2.45, \quad \text{reject at 5\%, not at 1\%.}}$$

Method 2: Confidence-interval duality. Reject $H_0:\beta_1=0$ at level α iff 0 lies outside $\hat{\beta}_1\pm c_\alpha\,\mathrm{SE}(\hat{\beta}_1)$:

$$\begin{aligned}CI_{95\%} &= 2.7\pm 1.96(1.1)=2.7\pm 2.156=[0.544,\ 4.856],\\ CI_{99\%} &= 2.7\pm 2.58(1.1)=2.7\pm 2.838=[-0.138,\ 5.538].\end{aligned}$$

The 95% interval excludes 0 (reject at 5%); the 99% interval contains 0 (do not reject at 1%). In standard-error units, 0 lies $(2.7-0)/1.1=2.455$ from the centre, between 1.96 and 2.58.

$$\boxed{t^{\text{act}}\approx 2.45, \quad \text{reject at 5\%, not at 1\%.}}$$

Method 3: p-value comparison. With $t^{\text{act}} = 2.455$, part (c) gives $p = 2\Phi(-2.455) \approx 0.0141$. Reject at level α iff $p < \alpha$: $0.0141 < 0.05$, but $0.0141 > 0.01$.

$$\boxed{t^{\text{act}} \approx 2.45, \quad \text{reject at 5\%, not at 1\%.}}$$

(c) 95% CI and p-value.
Method 1: Estimate ± 1.96 standard errors; lower-tail doubling. The margin is $1.96(1.1) = 2.156$, so the interval is $2.7 \pm 2.156 = [0.544, 4.856]$. For the p-value, $\Phi(-2.45) = 0.00714$ and $\Phi(-2.46) = 0.00695$; interpolating at 2.4545,

$$\Phi(-2.455) \approx 0.00714 - 0.45(0.00019) = 0.00705,$$
$$p = 2\Phi(-2.455) \approx 2(0.00705) = 0.0141.$$

$$\boxed{CI_{95\%}(\beta_1) \approx [0.54,\ 4.86], \quad p \approx 0.0141}$$

Method 2: Test inversion; complement of the central probability. The interval collects every $\beta_{1,0}$ not rejected at 5%:

$$\frac{|2.7 - \beta_{1,0}|}{1.1} \le 1.96 \iff -2.156 \le 2.7 - \beta_{1,0} \le 2.156$$
$$\iff 0.544 \le \beta_{1,0} \le 4.856.$$

For the p-value, $p = 1 - \Pr(|Z| \le 2.455) = 1 - [2\Phi(2.455) - 1] = 2 - 2\Phi(2.455)$. With $\Phi(2.45) = 0.99286$ and $\Phi(2.46) = 0.99305$, $\Phi(2.455) \approx 0.99286 + 0.45(0.00019) = 0.99295$, so

$$p \approx 2 - 2(0.99295) = 2 - 1.98590 = 0.0141.$$

$$\boxed{CI_{95\%}(\beta_1) \approx [0.54,\ 4.86], \quad p \approx 0.0141}$$

(d) $\text{SE}(\hat{\beta}_1)$ as a difference of independent group means (s_j^2: within-group sample variance).
Method 1: Variance of a difference of independent sample means. By part (a), $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$. Hence

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_0) - 2\,\text{Cov}(\bar{Y}_1, \bar{Y}_0).$$

The two groups consist of different, independently sampled workers, so $\text{Cov}(\bar{Y}_1, \bar{Y}_0) = 0$. Within group j the n_j wages are i.i.d. with variance σ_j^2, so $\text{Var}(\bar{Y}_j) = \sigma_j^2/n_j$. Therefore

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}.$$

Replacing each unknown σ_j^2 by the within-group sample variance $s_j^2 = \frac{1}{n_j - 1}\sum_{i:D_i=j}(Y_i - \bar{Y}_j)^2$ and taking the square root gives the standard error:

$$\boxed{\text{SE}(\hat{\beta}_1) = \sqrt{\text{Var}(\bar{Y}_{\text{Union}=1}) + \text{Var}(\bar{Y}_{\text{Union}=0})} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}}$$

Integrated Review Example 4: comparing linear estimators of μ (Lecture 5)

Let Y_1, Y_2, Y_3 be i.i.d. with $E[Y_i] = \mu$, $\text{Var}(Y_i) = \sigma^2$. Compare $\bar{Y} = \frac{1}{3}(Y_1 + Y_2 + Y_3)$ against the

candidate

$$W = \frac{Y_1 + 2Y_2 + Y_3}{4}.$$

(a) Unbiasedness of W.
Method 1: Linearity of expectation applied directly. Each Y_i has mean μ (identical distribution), so

$$\begin{aligned} E[W] &= E\Big[\frac{Y_1 + 2Y_2 + Y_3}{4}\Big] = \frac{1}{4}\big(E[Y_1] + 2E[Y_2] + E[Y_3]\big) \\ &= \frac{1}{4}(\mu + 2\mu + \mu) = \frac{1}{4}(4\mu) = \mu, \\ E[\bar{Y}] &= \frac{1}{3}\big(E[Y_1] + E[Y_2] + E[Y_3]\big) = \frac{1}{3}(3\mu) = \mu. \end{aligned}$$

$$\boxed{E[W] = \mu, \text{ same as } E[\bar{Y}] = \mu.}$$

Method 2: Weight-sum criterion for a linear estimator. Write $W = \sum_{i=1}^n a_i Y_i$ with fixed weights:

$$\begin{aligned} W &= \sum_{i=1}^n a_i Y_i, \quad (a_1, a_2, a_3) = \left(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{4}\right), \\ \sum_{i=1}^n a_i &= \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4} = 1, \\ E[W] &= \sum_{i=1}^n a_i E[Y_i] = \mu \sum_{i=1}^n a_i = \mu \cdot 1 = \mu \qquad \text{(linearity of } E\text{)}. \end{aligned}$$

$\bar{Y}$ has weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ with $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$, so $E[\bar{Y}] = \mu$ by the same criterion.

$$\boxed{E[W] = \mu, \text{ same as } E[\bar{Y}] = \mu.}$$

(b) Variance comparison.
Method 1: Variance of a sum of independent terms. Independence makes every covariance between distinct Y_i zero, and $\mathrm{Var}(cY_i) = c^2\sigma^2$:

$$\begin{aligned} \mathrm{Var}(W) &= \frac{1}{16}\mathrm{Var}(Y_1 + 2Y_2 + Y_3) = \frac{1}{16}\big[\mathrm{Var}(Y_1) + 2^2\,\mathrm{Var}(Y_2) + \mathrm{Var}(Y_3)\big] \\ &= \frac{1}{16}\big(\sigma^2 + 4\sigma^2 + \sigma^2\big) = \frac{6}{16}\sigma^2 = 0.375\,\sigma^2, \\ \mathrm{Var}(\bar{Y}) &= \frac{1}{9}\mathrm{Var}(Y_1 + Y_2 + Y_3) = \frac{1}{9}\big(3\sigma^2\big) = \frac{\sigma^2}{3} \approx 0.333\,\sigma^2. \end{aligned}$$

$$\boxed{\mathrm{Var}(\bar{Y}) \approx 0.333\,\sigma^2 \;<\; \mathrm{Var}(W) = 0.375\,\sigma^2.}$$

Method 2: Decomposition $a_i = (a_i - \frac{1}{n}) + \frac{1}{n}$. For any linear estimator, $\mathrm{Var}(\sum_{i=1}^n a_i Y_i) = \sigma^2 \sum_{i=1}^n a_i^2$. Expand the square and use $\sum_{i=1}^n (a_i - \frac{1}{n}) = \sum_{i=1}^n a_i - 1 = 0$ for unbiased weights:

$$\sum_{i=1}^n a_i^2 = \sum_{i=1}^n \Big(a_i - \tfrac{1}{n}\Big)^2 + \frac{2}{n}\sum_{i=1}^n \Big(a_i - \tfrac{1}{n}\Big) + n\Big(\tfrac{1}{n}\Big)^2 = \sum_{i=1}^n \Big(a_i - \tfrac{1}{n}\Big)^2 + \frac{1}{n}.$$

With $n = 3$, the deviations of W's weights from $\frac{1}{3}$ are $\frac{1}{4} - \frac{1}{3} = -\frac{1}{12}$, $\frac{1}{2} - \frac{1}{3} = \frac{1}{6}$, $\frac{1}{4} - \frac{1}{3} = -\frac{1}{12}$, so

$$\sum_{i=1}^{n}\Big(a_i - \tfrac{1}{3}\Big)^2 = \tfrac{1}{144} + \tfrac{4}{144} + \tfrac{1}{144} = \tfrac{6}{144} = \tfrac{1}{24},$$

$$\mathrm{Var}(W) = \sigma^2\Big(\tfrac{1}{24} + \tfrac{1}{3}\Big) = \sigma^2\Big(\tfrac{1}{24} + \tfrac{8}{24}\Big) = \tfrac{9}{24}\sigma^2 = 0.375\,\sigma^2,$$

$$\mathrm{Var}(\bar{Y}) = \sigma^2\Big(0 + \tfrac{1}{3}\Big) \approx 0.333\,\sigma^2.$$

The gap is $\mathrm{Var}(W) - \mathrm{Var}(\bar{Y}) = \sigma^2/24 > 0$.

$$\boxed{\mathrm{Var}(\bar{Y}) \approx 0.333\,\sigma^2 \;<\; \mathrm{Var}(W) = 0.375\,\sigma^2.}$$

(c) Third comparator Y_1.
Method 1: Weights $(1, 0, 0)$ in the general formulas.

$$(a_1, a_2, a_3) = (1, 0, 0), \quad \sum_{i=1}^{n} a_i = 1 + 0 + 0 = 1 \quad \text{(unbiased)},$$

$$\mathrm{Var}(Y_1) = \sigma^2(1^2 + 0^2 + 0^2) = \sigma^2.$$

Combining with part (b), $0.333\,\sigma^2 < 0.375\,\sigma^2 < \sigma^2$:

$$\boxed{\mathrm{Var}(\bar{Y}) \approx 0.333\,\sigma^2 \;<\; \mathrm{Var}(W) = 0.375\,\sigma^2 \;<\; \mathrm{Var}(Y_1) = \sigma^2.}$$

Method 2: Ranking by squared deviation from equal weights. By the decomposition of part (b), $\mathrm{Var}(\hat{\mu}) = \sigma^2\big[\sum_{i=1}^{n}(a_i - \frac{1}{3})^2 + \frac{1}{3}\big]$, so the three estimators are ranked by $\sum_{i=1}^{n}(a_i - \frac{1}{3})^2$. For Y_1 the deviations are $1 - \frac{1}{3} = \frac{2}{3}$, $0 - \frac{1}{3} = -\frac{1}{3}$, $0 - \frac{1}{3} = -\frac{1}{3}$:

$$\sum_{i=1}^{n}\Big(a_i - \tfrac{1}{3}\Big)^2 = \tfrac{4}{9} + \tfrac{1}{9} + \tfrac{1}{9} = \tfrac{6}{9} = \tfrac{2}{3},$$

$$\mathrm{Var}(Y_1) = \sigma^2\Big(\tfrac{2}{3} + \tfrac{1}{3}\Big) = \sigma^2.$$

The deviation sums are 0 $(\bar{Y}) < \frac{1}{24}$ $(W) < \frac{2}{3}$ (Y_1).

$$\boxed{\mathrm{Var}(\bar{Y}) \approx 0.333\,\sigma^2 \;<\; \mathrm{Var}(W) = 0.375\,\sigma^2 \;<\; \mathrm{Var}(Y_1) = \sigma^2.}$$

Integrated Review Example 5: Assumption 1 mini-essay – internships and starting salary (Lectures 1, 3–4)

Question. A university placement office estimates $\text{Salary}_i = \beta_0 + \beta_1\,\text{Internship}_i + u_i$, where $\text{Internship}_i = 1$ if the graduate completed a structured internship during their degree and 0 otherwise, and wants to interpret $\hat{\beta}_1$ as the causal salary boost from doing an internship. Assess whether Assumption 1 plausibly holds, state the likely direction of any bias, and support your reasoning with a numerical illustration of $E[u] = 0$ coexisting with $E[u \mid \text{Internship}] \neq 0$.
Answer.

- The error u collects every determinant of starting salary other than the internship, measured in the same salary units as Salary_i: networking ability, motivation, family connections and GPA are all inside u.

- Internships are not randomly assigned: students choose to apply and firms choose whom to accept, so motivated, high-GPA, well-connected students are more likely to have $\text{Internship}_i = 1$. Then $E[u \,|\, \text{Internship} = 1] > E[u \,|\, \text{Internship} = 0]$, so $E[u \,|\, \text{Internship}] \neq 0$ and $\text{Cov}(\text{Internship}, u) > 0$; Least Squares Assumption 1 plausibly fails.

- Writing $D = \text{Internship}$, the binary-regressor $\hat{\beta}_1$ estimates the population gap in mean salary, $E[\text{Salary} \,|\, D{=}1] - E[\text{Salary} \,|\, D{=}0] = \beta_1 + \big(E[u \,|\, D{=}1] - E[u \,|\, D{=}0]\big)$. The bracketed selection term is positive, so $\hat{\beta}_1$ likely overstates the causal salary effect β_1.

- $\hat{\beta}_1$ remains a valid sample description of how mean starting salary differs between graduates with and without an internship; it is not the causal boost. Only random assignment of internships (the RCT benchmark) would guarantee $E[u \,|\, \text{Internship}] = 0$ by construction.

Numerical illustration (hypothetical values). $P(\text{Internship} = 0) = 0.4$, $E[u \,|\, \text{Internship} = 0] = -3$; $P(\text{Internship} = 1) = 0.6$, $E[u \,|\, \text{Internship} = 1] = 2$.
Method 1: Law of Total Expectation. Weight each conditional mean by the probability of its group:

$$\begin{aligned} E[u] &= P(\text{Internship} = 0)\, E[u \,|\, \text{Internship} = 0] + P(\text{Internship} = 1)\, E[u \,|\, \text{Internship} = 1] \\ &= 0.4(-3) + 0.6(2) = -1.2 + 1.2 = 0. \end{aligned}$$

Both conditional means differ from 0, so $E[u \,|\, \text{Internship}] \neq 0$ even though the unconditional mean is exactly 0.

$$\boxed{E[u] = 0 \text{ exactly, while } E[u \,|\, \text{Internship} = 0] = -3 \neq 0 \text{ and } E[u \,|\, \text{Internship} = 1] = 2 \neq 0.}$$

Method 2: Linear conditional mean of u on the binary regressor. Because Internship takes only the values 0 and 1, its conditional mean function is exactly linear, $E[u \,|\, \text{Internship}] = \gamma_0 + \gamma_1 \,\text{Internship}$, with

$$\begin{aligned} \gamma_0 &= E[u \,|\, \text{Internship} = 0] = -3, \\ \gamma_1 &= E[u \,|\, \text{Internship} = 1] - E[u \,|\, \text{Internship} = 0] = 2 - (-3) = 5. \end{aligned}$$

Taking expectations of both sides (Law of Iterated Expectations), with $E[\text{Internship}] = 0.6$,

$$E[u] = E\big[E[u \,|\, \text{Internship}]\big] = \gamma_0 + \gamma_1 E[\text{Internship}] = -3 + 5(0.6) = -3 + 3 = 0.$$

The slope $\gamma_1 = 5 \neq 0$ gives the covariance:

$$\begin{aligned} \text{Cov}(\text{Internship}, u) &= E[\text{Internship} \cdot u] - E[\text{Internship}]\, E[u] = 0.6(2) - 0.6(0) = 1.2 > 0, \\ \text{Var}(\text{Internship}) &= 0.6(1 - 0.6) = 0.24, \qquad \frac{\text{Cov}(\text{Internship}, u)}{\text{Var}(\text{Internship})} = \frac{1.2}{0.24} = 5 = \gamma_1. \end{aligned}$$

In this illustration the intern/non-intern gap in mean salary exceeds β_1 by 5 salary units, the positive selection term of the third bullet.

$$\boxed{E[u] = 0 \text{ exactly, while } E[u \,|\, \text{Internship} = 0] = -3 \neq 0 \text{ and } E[u \,|\, \text{Internship} = 1] = 2 \neq 0.}$$

7.5 Part I Revision Checklist

Checklist: can you, without notes, . . .

- ☐ derive $\hat{\beta}_0, \hat{\beta}_1$ from the FOCs of $\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$, ending at $\hat{\beta}_1 = s_{XY}/s_X^2$;
- ☐ state, from memory, which symbol is fixed/unknown (β_0, β_1), which is a dummy candidate (b_0, b_1), and which is the random estimator $(\hat{\beta}_0, \hat{\beta}_1)$;
- ☐ prove $\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = 0$, $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$ from the two OLS FOCs, then chain them into $\text{TSS} = \text{ESS} + \text{SSR}$ and $0 \le R^2 \le 1$;
- ☐ state Least Squares Assumptions 1–3 *by number* (never "SLR.1"–"SLR.5") and identify, for a new example, which assumption is under threat;
- ☐ construct a two-value numerical example with $E[u] = 0$ but $E[u \mid X] \neq 0$;
- ☐ name at least four distinct selection-bias contexts (hospital survey, return to education, Wald's bombers, investor survivorship) and explain the mechanism in each;
- ☐ derive $E[\bar{Y}] = \mu$ and $\text{Var}(\bar{Y}) = \sigma^2/n$ from i.i.d. sampling, stating exactly which step used identical distribution and which used independence;
- ☐ derive the unbiasedness condition $\sum_{i=1}^{n} a_i = 1$ and the variance formula $\text{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2$ for a general linear estimator $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$, and show $\bar{Y}$ minimises it;
- ☐ state the Gauss–Markov theorem precisely, including the extra homoskedasticity condition beyond Assumptions 1–3;
- ☐ translate a binary-regressor equation into the two group means, both ways ($\beta_0, \beta_1 \to$ means; means $\to \hat{\beta}_0, \hat{\beta}_1$);
- ☐ standardise $\hat{\beta}_1$ using β_1 and $\sigma_{\hat{\beta}_1}$ (or $\text{SE}(\hat{\beta}_1)$) and read off a probability from Φ (e.g. $\Pr(\hat{\beta}_1 > c)$ or $\Pr(|\hat{\beta}_1 - \beta_1| < c)$);
- ☐ run the four-beat t-test (hypotheses, statistic, decision rule, conclusion in context) for *any* hypothesised value, not just $\beta_1 = 0$ (Final only);
- ☐ build a 95% CI and use the CI/t-test/p-value duality to answer any one of the three from the other two (Final only);
- ☐ compute a prediction and a change in prediction, with correct units, and attach a CI to the change when asked;
- ☐ compute the mean and variance of a difference of two independent sample means;
- ☐ give a fresh example of each of the three sources of correlation (causality, reverse causality, common factor) and say which one a proposed regression is implicitly assuming away;
- ☐ state the p-value's definition precisely, as $\Pr_{H_0}(|N(0,1)| > |t^{\text{act}}|)$, and explain in one sentence why $p < \alpha \iff |t^{\text{act}}| > c_\alpha$ (Final only);
- ☐ explain, in one sentence each, why a small R^2 does not invalidate a causally valid $\hat{\beta}_1$, and why a large R^2 does not certify one;
- ☐ list all ten Method Patterns above by name from the one-line "when you see it" cue alone, with no notes.

Master Memory Sheet: Part I (Lectures 1–6)

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad \text{(population model)}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \qquad \hat{u}_i = Y_i - \hat{Y}_i$$

$$\sum_{i=1}^{n} \hat{u}_i = 0, \qquad \sum_{i=1}^{n} \hat{u}_i X_i = 0, \qquad \sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$$

$$\text{TSS} = \text{ESS} + \text{SSR}, \qquad R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} = [\text{Corr}(X, Y)]^2$$

Assumption 1: $E[u_i \mid X_i] = 0$, Assumption 2: (X_i, Y_i) i.i.d.

Assumption 3: large outliers unlikely

$$E[\bar{Y}] = \mu, \qquad \text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$$

$$\hat{\mu} = \sum_{i=1}^{n} a_i Y_i \text{ unbiased} \iff \sum_{i=1}^{n} a_i = 1, \qquad \text{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2 \geq \frac{\sigma^2}{n} = \text{Var}(\bar{Y})$$

Gauss–Markov: $\text{Var}(\hat{\beta}_0) \leq \text{Var}(\tilde{\beta}_0)$, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$ (+ homoskedasticity)

Binary D_i : $\hat{\beta}_0 = \bar{Y}_{D=0}, \qquad \hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0} = E[Y \mid D{=}1] - E[Y \mid D{=}0]$

$$E[\hat{d}] = \mu_1 - \mu_2, \qquad \text{Var}(\hat{d}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} \qquad (\hat{d} = \bar{Y}_1 - \bar{Y}_2, \text{ independent samples})$$

$$\hat{\beta}_1 \approx N(\beta_1, \sigma^2_{\hat{\beta}_1}), \qquad \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \approx \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \approx N(0,1) \quad \text{(large samples; Quiz scope)}$$

$$\text{SE}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}, \qquad t^{\text{act}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)} \quad \text{(Final only)}$$

reject H_0 at 5% $\iff |t^{\text{act}}| > 1.96$ (1%: 2.58; 10%: 1.645)

$$CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96 \text{ SE}(\hat{\beta}_1)$$

$$CI_{95\%}(\beta_1 \Delta X) = \hat{\beta}_1 \Delta X \pm 1.96 \text{ SE}(\hat{\beta}_1) |\Delta X|$$

$$p\text{-value} = 2\Phi(-|t^{\text{act}}|), \qquad p < \alpha \iff |t^{\text{act}}| > c_\alpha$$

A Formula & Key Results Reference

Chapter 1: Introduction and Causality

Definitions

- **Three sources of correlation.** X, Y correlated $\Rightarrow$ (i) causality $X \to Y$, (ii) reverse causality $Y \to X$, or (iii) common factor $C \to X,\ C \to Y$. A raw correlation cannot tell these apart.
- **Conditional mean.** $E(Y \mid X = x)$: average of Y in the sub-population with $X = x$.
- **Population vs. sample.** Population = all units of interest; sample = the units actually drawn (usually by random selection).
- **Null/alternative hypothesis, Type I/II error.** H_0 = default claim, H_1 = claim entertained if H_0 is contradicted. Type I = reject H_0 when true; Type II = accept H_0 when false. Significance level $= P(\text{Type I error})$.
- **RCT.** Random assignment to treatment/control makes the two groups identical on average in every other respect, so causal effect $\approx E(Y \mid \text{treat}) - E(Y \mid \text{control})$.

Results

- $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$. [definition]
- $E[aX + b] = aE[X] + b;\ \ \text{Var}(aX + b) = a^2 \,\text{Var}(X)$. [linearity of E; always true]
- $\text{Cov}(aX + b, Y) = a\,\text{Cov}(X, Y);\ \ \text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$. [bilinearity; always true]
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X, Y)$. [always true]
- $X \perp Y \Rightarrow \text{Cov}(X, Y) = 0$ and $E[XY] = E[X]E[Y]$. [independence; converse false – zero covariance does *not* imply independence]
- Law of Iterated Expectations: $E\big[E(Y \mid X)\big] = E[Y]$. [always true]
- $X \sim N(0, 1)$: $E[X] = 0,\ E[X^2] = 1,\ E[X^3] = 0$. [normality; all odd moments vanish by symmetry]

Procedures / Templates

- **Classify a correlation.** Name the plausible mechanism (which variable could move first; is there a third driver?) and label it causality / reverse causality / common factor – a bare correlation coefficient cannot do this alone.
- **LIE two-route trick.** To find $E[Y]$ or $E[g(X)Y]$ when Y is built from X, Z: (1) compute $E(Y \mid X)$ first using $E[g(X) \mid X] = g(X)$ and $Z \perp X \Rightarrow E(Z \mid X) = E[Z]$; (2) apply the LIE, $E[Y] = E[E(Y \mid X)]$; (3) cross-check against the direct (unconditional) expansion.

Chapter 2: The OLS Estimator

Definitions

- **Simple linear regression model.** $Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \ldots, n$.
- **Three kinds of beta.** (β_0, β_1) = true, fixed, unknown population parameters; (b_0, b_1) = generic dummy candidates in the minimisation; $(\hat{\beta}_0, \hat{\beta}_1)$ = OLS estimators, the minimisers for one sample (random).
- **OLS predicted value / residual.** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i;\ \ \hat{u}_i = Y_i - \hat{Y}_i$.

Results

- Summation identities: $\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^n X_iY_i - n\bar{X}\bar{Y}$; $\sum_{i=1}^n (X_i-\bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2$. [pure algebra, always true]
- OLS closed form:
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^n (X_i-\bar{X})^2} = \frac{s_{XY}}{s_X^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}.$$
[requires $\sum_{i=1}^n (X_i-\bar{X})^2 > 0$, i.e. X varies in the sample]
- The OLS line passes through $(\bar{X},\bar{Y})$: $\hat{\beta}_0 + \hat{\beta}_1\bar{X} = \bar{Y}$. [always true, from the intercept formula]
- $S(b_0,b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1X_i)^2$ is strictly convex whenever $\sum_{i=1}^n (X_i-\bar{X})^2 > 0$, so the FOC stationary point is the unique global minimum. [same non-degeneracy condition]

Procedures / Templates

- **OLS by hand.** (1) Compute $\bar{X},\bar{Y}$. (2) Tabulate deviations $X_i-\bar{X}$, $Y_i-\bar{Y}$. (3) Sum cross-products $\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})$ and squares $\sum_{i=1}^n (X_i-\bar{X})^2$. (4) $\hat{\beta}_1$ = ratio of the two sums. (5) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$. (6) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_i$, $\hat{u}_i = Y_i - \hat{Y}_i$; check $\sum_{i=1}^n \hat{u}_i = 0$ as an arithmetic check.
- **Prediction and change.** Predict at new X: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1X$. Change in predicted Y for ΔX: either difference two predictions, or $\hat{\beta}_1 \times \Delta X$ directly – both routes must agree (line is linear).
- **Recover $\bar{Y}$ from $\bar{X}$.** $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1\bar{X}$ – no new data needed, direct application of the mean-point fact.

Chapter 3: The Least Squares Assumptions

Definitions

- **Five objects.** Population line $\beta_0+\beta_1X$ vs. OLS line $\hat{\beta}_0+\hat{\beta}_1X$; error $u_i = Y_i - (\beta_0+\beta_1X_i)$ vs. residual $\hat{u}_i = Y_i - (\hat{\beta}_0+\hat{\beta}_1X_i)$; unobservable systematic part $\beta_0+\beta_1X_i$ vs. predicted value $\hat{Y}_i$.
- **Least Squares Assumptions.** Assumption 1: $E[u_i \mid X_i] = 0$ (zero conditional mean). Assumption 2: (X_i,Y_i) i.i.d. Assumption 3: large outliers unlikely (finite 4th moments).
- **Selection / survivorship bias.** Sample inclusion (or the value of X_i) correlated with u_i, so Assumption 1 fails in the observed sample even if it would hold in the population.

Results

- Residual identities: $\sum_{i=1}^n \hat{u}_i = 0$, $\sum_{i=1}^n \hat{u}_iX_i = 0$, $\sum_{i=1}^n \hat{u}_i\hat{Y}_i = 0$. [always true – pure algebraic consequence of the OLS FOCs, no assumption needed]
- Assumption 1 $\Rightarrow E[Y_i \mid X_i] = \beta_0 + \beta_1X_i$. [Assumption 1]
- Assumption 1 $\Rightarrow \mathrm{Cov}(X,u) = 0$ (via LIE). [Assumption 1; converse false – zero correlation does not imply Assumption 1]
- Binary $X_i \in \{0,1\}$: $\beta_0 = E(Y \mid X{=}0)$, $\beta_1 = E(Y \mid X{=}1) - E(Y \mid X{=}0)$. [Assumption 1; β_1 is a group-mean difference, not a slope]

Procedures / Templates

- **Assumption 1 judgement.** (1) Identify everything bundled into u. (2) Ask whether the average of u shifts systematically across values/groups of X (not just whether $E[u]=0$ overall). (3) If it fails, name the mechanism (self-selection, survivorship, omitted common factor). (4) State the plausible direction of bias if askable.
- **Residual-identity derivation.** Divide each OLS FOC by -2; substitute $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1X_i$; the third identity ($\sum_{i=1}^n \hat{u}_i\hat{Y}_i = 0$) follows by substituting $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_i$ and using the first two.

Chapter 4: Sampling, Fit and Distribution

Definitions

- **Simple random sampling.** Every population member equally likely to be drawn ⇒ independence; every draw from the same population ⇒ identical distribution. Together: Assumption 2.
- $\text{TSS}, \text{ESS}, \text{SSR}, R^2$**.** $\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, $\text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, $\text{SSR} = \sum_{i=1}^{n}\hat{u}_i^2$, $R^2 = \text{ESS}/\text{TSS}$.
- **Sampling distribution.** The probability distribution of an estimator's value across repeated random samples of size n from the same population.
- **CAPM regression.** $R - R_f = \alpha + \beta(R_m - R_f) + u$; β = market beta, α = abnormal return (0 under strict CAPM).

Results

- $Y_i = \hat{Y}_i + \hat{u}_i$; $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$. [from $\sum_{i=1}^{n}\hat{u}_i = 0$, always true]
- $\hat{\beta}_1 = 0 \Rightarrow \hat{Y}_i = \bar{Y}\ \forall i \Rightarrow \text{ESS} = 0,\ R^2 = 0$. [always true]
- $\text{TSS} = \text{ESS} + \text{SSR}$. [always true – cross term vanishes via $\sum_{i=1}^{n}\hat{u}_i = 0, \sum_{i=1}^{n}\hat{u}_i X_i = 0$]
- $0 \le R^2 \le 1$; $R^2 = 1 \iff \hat{u}_i = 0\ \forall i$. [always true, given $\text{TSS} > 0$]
- $R^2 = [\text{Corr}(X, Y)]^2$. [simple/one-regressor case only]
- $E[\bar{Y}] = \mu$, $\text{Var}[\bar{Y}] = \sigma^2/n$. [i.i.d. sample, Assumption 2]
- $\bar{Y}$ approximately $N(\mu, \sigma^2/n)$ for $n > 100$. [CLT, large sample]

Procedures / Templates

- **Compute** R^2**.** From residuals: $\text{TSS}, \text{ESS}, \text{SSR}$ then $R^2 = \text{ESS}/\text{TSS} = 1 - \text{SSR}/\text{TSS}$; or, in simple regression, square the sample correlation of X, Y.
- **Diagnose Assumption 2.** For: name the random-selection mechanism explicitly (e.g. "addresses drawn at random"). Against: identify either a non-random filter (who is likely sampled) or sequential dependence (time series).
- **Diagnose selection bias.** Name the unobserved factor that drives both sample inclusion (or the value of X) and the outcome Y.

Chapter 5: Unbiasedness, Efficiency and Binary Regressors

Definitions

- **Estimator vs. estimate.** Estimator $\hat{\theta}$ = a rule (function of the sample), random; estimate = its realised numerical value for one sample.
- **Unbiasedness.** $E[\hat{\theta}] = \theta$. **Efficiency.** Among unbiased estimators, $\hat{\theta}$ is more efficient than $\tilde{\beta}$ if $\text{Var}(\hat{\theta}) < \text{Var}(\tilde{\beta})$.
- **Linear estimator of** μ**.** $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$, weights a_i fixed before the data are observed.
- **Gauss–Markov / BLUE.** Best Linear Unbiased Estimator: smallest variance among all linear, unbiased competitors.
- **Binary regressor (dummy).** $D_i \in \{0, 1\}$. **Dummy variable trap.** k categories need $k - 1$ dummies plus an intercept, never k.

Results

- $\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$ unbiased for $\mu \iff \sum_{i=1}^{n} a_i = 1$. [i.i.d. Y_i]
- $\text{Var}(\hat{\mu}) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \left[\sum_{i=1}^{n} (a_i - \frac{1}{n})^2 + \frac{1}{n}\right] \geq \sigma^2/n$, minimised uniquely at $a_i = 1/n \ \forall i$. So $\bar{Y}$ is BLUE for μ. [i.i.d., $\sum_{i=1}^{n} a_i = 1$]
- Gauss–Markov: $\text{Var}(\hat{\beta}_0) \leq \text{Var}(\tilde{\beta}_0)$, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$ for any linear unbiased $\tilde{\beta}_0, \tilde{\beta}_1$. [Assumptions 1–3 + homoskedasticity $\text{Var}(u_i \mid X_i) = \sigma_u^2$]
- Binary D_i: $\hat{\beta}_0 = \bar{Y}_{D=0}$, $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$. [OLS FOCs split by group]
- Common slope, different intercepts: $Y_i = \beta_0 + \delta_0 D_i + \beta_1 X_i + u_i \Rightarrow$ group lines $\beta_0 + \beta_1 X_i$ (D=0) and $(\beta_0 + \delta_0) + \beta_1 X_i$ (D=1), parallel, shifted by δ_0. [$E[u_i \mid D_i, X_i] = 0$]

Procedures / Templates

- **Compare two linear estimators.** (1) Check unbiasedness via $\sum_{i=1}^{n} a_i = 1$. (2) Compute Var $= \sigma^2 \sum_{i=1}^{n} a_i^2$. (3) Prefer the smaller variance among the unbiased candidates – never compare variance across a biased vs. unbiased pair.
- **Binary-regressor reading.** $\hat{\beta}_0$ = base-group (D=0) sample mean; $\hat{\beta}_0 + \hat{\beta}_1$ = other-group (D=1) sample mean; $\hat{\beta}_1$ = difference in group means – state explicitly that it is *not* a slope (only two points on the X-axis).
- **Hypothesis-test skeleton (previewed here, built in Ch. 6).** State $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$; decision rule reject if $|\bar{Y} - \mu_0| > c$; significance level = $P(\text{reject} \mid H_0 \text{ true})$ fixed first, then evaluate power.

Chapter 6: Inference with a Single Regressor

Definitions

- **Standard error.** $\text{SE}(\hat{\beta}_1) = \hat{\sigma}_{\hat{\beta}_1}$, the estimated spread of $\hat{\beta}_1$'s sampling distribution.
- **t-statistic.** $t = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{\text{SE}(\hat{\beta}_1)}$; realised value written t^{act}. Reporting convention: $\underset{(\text{SE}(\hat{\beta}_1))}{\hat{\beta}_1}$.
- **p-value.** $p = 2\Phi(-|t^{\text{act}}|) = \Pr_{H_0}(|N(0,1)| > |t^{\text{act}}|)$.
- **95% CI for β_1.** The set of $\beta_{1,0}$ not rejected by a two-sided 5% test.

Results

- $\dfrac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \approx N(0,1)$. [Assumptions 1–3, n large]
- Two-sided test with $|t^{\text{act}}| > 1.96$: $\Pr(\text{reject} \mid H_0 \text{ true}) \to 0.05$ as $n \to \infty$. [Assumptions 1–3, n large]
- CI duality: $CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96 \ \text{SE}(\hat{\beta}_1) = \{\beta_{1,0} : |t^{\text{act}}| \leq 1.96\}$. [same conditions]
- $p < \alpha \iff |t^{\text{act}}| > c_\alpha$. [$c_\alpha$ = two-sided critical value at level α]
- Standard normal two-sided critical values: $c_{10\%} = 1.645$, $c_{5\%} = 1.96$, $c_{1\%} = 2.58$; one-sided 5% uses 1.645.

Procedures / Templates

- **t-test, four beats.** (1) *Hypotheses*: $H_0 : \beta_1 = \beta_{1,0}$ vs. H_1 (two- or one-sided). (2) *Test statistic*: $t^{\text{act}} = (\hat{\beta}_1 - \beta_{1,0})/\,\text{SE}(\hat{\beta}_1)$. (3) *Decision rule*: reject iff $|t^{\text{act}}| > c_\alpha$ (two-sided: 1.645/1.96/2.58 at 10%/5%/1%; one-sided at 5%: $t^{\text{act}} > 1.645$ or $t^{\text{act}} < -1.645$). (4) *Conclusion in context*: state rejection/non-rejection and the economic meaning.
- **CI construction.** $CI_{95\%}(\beta_1) = \hat{\beta}_1 \pm 1.96\ \text{SE}(\hat{\beta}_1)$; for a scaled effect, $\hat{\beta}_1 \Delta X \pm 1.96\ \text{SE}(\hat{\beta}_1)|\Delta X|$. Check: $\beta_{1,0}$ outside the CI $\iff$ rejected at that level.
- **p-value.** Compute $p = 2\Phi(-|t^{\text{act}}|)$; reject H_0 at level α iff $p < \alpha$; never report p without the point estimate and its economic magnitude.
- **Binary-regressor t-test.** Same four beats with $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$ and its SE – tests whether the group-mean gap is statistically significant.

B Notation

Probability and Sampling

Symbol	Meaning	Context
$\mu,\ \sigma^2$	population mean, population variance	generic population parameters (e.g. of Y_i)
$E(\cdot)$, $E(\cdot \mid \cdot)$	(unconditional) expectation; conditional expectation	Ch. 1 toolkit; $E(Y \mid X{=}x)$ is the conditional mean
$\text{Var}(\cdot)$, $\text{Cov}(\cdot,\cdot)$, $\text{Corr}(\cdot,\cdot)$	variance; covariance; correlation	Ch. 1 toolkit, used throughout
$X_i,\ Y_i$	the regressor and outcome for observation i	population random variables / one sampled draw
i.i.d.	independently and identically distributed	property of (X_i, Y_i) under Assumption 2 (simple random sampling)
$H_0,\ H_1$	null hypothesis; alternative hypothesis	hypothesis testing (Ch. 1, formalised in Ch. 6)
Type I / Type II error	false rejection of a true H_0 / false acceptance of a false H_0	significance level $= P(\text{Type I error})$

Regression Model and OLS

Symbol	Meaning	Context
β_0, β_1	population (true, unknown) intercept and slope	regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$
b_0, b_1	generic candidate values in the least-squares problem	dummy arguments of $\min_{b_0,b_1} \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$
$\hat{\beta}_0, \hat{\beta}_1$	OLS estimators (functions of the sample)	estimation; random – vary from sample to sample
$\tilde{\beta}$ (e.g. $\tilde{\beta}_1$)	a generic alternative linear estimator	used to compare against OLS in efficiency / Gauss–Markov arguments
u_i / $\hat{u}_i$	population error (unobserved) / OLS residual	model vs. fit; only $\sum_{i=1}^{n} \hat{u}_i = 0$ holds, never $\sum_{i=1}^{n} u_i = 0$
$\hat{Y}_i$	OLS predicted (fitted) value $\hat{\beta}_0 + \hat{\beta}_1 X_i$	fit
$\bar{X}$, $\bar{Y}$	sample means of X and Y	the OLS line always passes through $(\bar{X}, \bar{Y})$
s_{XY}, s_X^2	sample covariance of X, Y; sample variance of X	$\hat{\beta}_1 = s_{XY}/s_X^2$
D_i	binary (dummy/indicator) regressor, $D_i \in \{0, 1\}$	binary-regressor models; $\hat{\beta}_0 = \bar{Y}_{D=0}$, $\hat{\beta}_1 = \bar{Y}_{D=1} - \bar{Y}_{D=0}$
δ_0	intercept shift for the D_i=1 group	common-slope, different-intercept model $Y_i = \beta_0 + \delta_0 D_i + \beta_1 X_i + u_i$
$\underset{\cdot}{\cdot}$	coefficient with its standard error set underneath	reporting convention, e.g. $\underset{(0.52)}{-2.28}$

Measures of Fit

Symbol	Meaning	Context
TSS, ESS, SSR	total / explained / residual sum of squares	measure of fit; TSS = ESS + SSR
R^2	coefficient of determination, ESS/TSS = 1−SSR/TSS	fraction of Y's variation explained by X; = $[\text{Corr}(X, Y)]^2$ in simple regression

Inference

Symbol	Meaning	Context
$\sigma^2_{\hat{\beta}_1}$	population variance of the sampling distribution of $\hat{\beta}_1$	unobservable; drives the large-sample normal approximation
$\hat{\sigma}^2_{\hat{\beta}_1}$	sample (heteroskedasticity-robust) estimator of $\sigma^2_{\hat{\beta}_1}$	computed from the OLS residuals $\hat{u}_i$
$\mathrm{SE}(\hat{\beta}_1)$	standard error of $\hat{\beta}_1$, $= \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$	inference; reported under $\hat{\beta}_1$ in $\underset{(\mathrm{SE}(\hat{\beta}_1))}{\hat{\beta}_1}$
t^{act}	realised t-statistic, $(\hat{\beta}_1 - \beta_{1,0})/\,\mathrm{SE}(\hat{\beta}_1)$	hypothesis testing
α	significance level, $P(\text{Type I error})$	conventionally $10\%, 5\%, 1\%$; critical values $1.645/1.96/2.58$
$\Phi(\cdot)$	standard normal CDF	used in the p-value and in critical-value lookups
p-value	$2\Phi(-\lvert t^{\mathrm{act}}\rvert)$	reject H_0 at level α iff $p < \alpha$
$CI_{95\%}(\beta_1)$	95% confidence interval for β_1, $\hat{\beta}_1 \pm 1.96\ \mathrm{SE}(\hat{\beta}_1)$	duality: excludes $\beta_{1,0}$ iff a 5% two-sided test rejects it

C Exam & Admin Information

Assessment Breakdown

Component	Weight	Timing / Coverage / Format
Quiz	30%	Week 7 (after Recess Week); simple linear regression, Lectures 1–5 and Lecture 6 slides 1–3 (up to the standard error of $\hat{\beta}_1$); no Stata questions; no tutorial that day.
Tutorial questions and presentations	20%	Weekly Tutorial Question Sets; answers presented in tutorials.
Final Examination	50%	Comprehensive; Lectures 1–12.

The outline does not state the allowed materials (help sheet, calculator) or the duration of the Quiz or the Final; confirm with the instructor.

Weekly Schedule and Supplementary Reading

Week	Topic	Stock & Watson
1	Introduction	Ch. 1 (1.1–1.2), Ch. 2 (2.1–2.3)
2–6	Simple linear regression model 1–5	Ch. 4 (4.1–4.4), Ch. 2 (2.5), Ch. 3 (3.1), Ch. 5
7	Quiz	—
8–10	Multiple linear regression model 1–3	Ch. 3 (3.2–3.3), Ch. 5, Ch. 6
11	Nonlinear regression model 1	Ch. 7 / Ch. 8 (by edition)
12	Nonlinear regression model 2; course review	Ch. 8

Chapter numbers past Week 6 are approximate (the outline's table layout is ambiguous); the lecture slides are the main reference.

Resources

- Main reference: lecture slides and notes (Assoc. Prof. Wang Wenjie).
- Supplementary textbook: J. H. Stock and M. W. Watson, *Introduction to Econometrics*, 3rd or 4th edition.
- Software: Stata (computer labs, HSS building B1).
- Prerequisite knowledge: MH1820 Introduction to Probability and Statistical Methods (expectation, variance, covariance, normal distribution).
- QRS@NTU materials: tutorial solutions, cheatsheets and practice papers at qrsntu.org.

Revision Checklist Before the Quiz

☐ Derive $\hat{\beta}_0, \hat{\beta}_1$ from the least-squares FOCs without notes.

☐ Prove $\sum_{i=1}^{n} \hat{u}_i = 0$, $\sum_{i=1}^{n} \hat{u}_i X_i = 0$, $\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = 0$, then $\text{TSS} = \text{ESS} + \text{SSR}$ and $0 \leq R^2 \leq 1$.

☐ State Least Squares Assumptions 1–3 and judge Assumption 1 in a new economic example.

☐ Interpret β_0, β_1 with a binary regressor as conditional means.

☐ Show that $\sum_i a_i Y_i$ is unbiased for μ iff $\sum_i a_i = 1$; compare variances.

☐ Standardise $\hat{\beta}_1$ using β_1 and $\sigma_{\hat{\beta}_1}$ (or $\text{SE}(\hat{\beta}_1)$) and compute a probability via $\Phi(\cdot)$ (e.g. $\Pr(\hat{\beta}_1 > c)$); the t-test, 95% CI, and p-value are Final-only, not Quiz material.

QRS@NTU – Revision Notes