LECTURE 1.1

Data and Variation

ECON30130 Econometrics I

Sam Deegan
University College Dublin

Autumn 2027

Where We Are

What a Year of Schooling Is Worth

Write down a euro figure for one extra year of schooling. Then say whose earnings you would need to see:

> **Question**
> How much more does a person earn per hour, on average, for each extra year they stayed in school?

Lecture Plan

How the module runs takes the first seventeen minutes; then today's four sections:

1. **Populations and Samples.**
   The question is about a population; the data is a sample.
2. **Centre and Spread.**
   Where a variable sits, and how far people scatter from its mean.
3. **Moments and Shape.**
   Which tail is longer, and how much sits far out.
4. **Describing Two Variables.**
   Covariance, correlation, and what one number cannot show.

SECTION 1

How This Module Works

What the Module Covers: the Linear Regression Model

More advanced models solve specific problems but rest on linear regression. This module teaches linear regression in four blocks:

- **Foundations:** we describe samples and how estimates vary between them.
- **Simple Regression:** a straight line relates one outcome to one predictor.
- **Multiple Regression:** we hold other variables fixed, and see the bias from omitting them.
- **Limits:** we ask what breaks a regression, and how you would know.

> **Math Is Unavoidable**
> No equation here is copied off a textbook page. Each one is built in front of you. Intuition comes first, then a picture where one exists.

Wooldridge's Introductory Econometrics Is the Set Text

Our textbook is the 7th edition of Jeffrey Wooldridge's *Introductory Econometrics*:

- **Chapters:** we cover chapter 1 and Part 1, chapters 2 to 9.
- **Math Refreshers:** A to C cover the notation, if it slows you down.
- **Cost:** libraries hold copies, and older editions cost less.

Regression Studies How an Outcome Varies with Predictors

Regression studies how an outcome (y_i) varies with one or more predictors (x_i):

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

- β_0: the intercept
- β_1: the slope
- u_i: the error, or unexplained outcome

> **Question**
> Can you think of situations where this method could be useful in research, policy, business or everyday life?

What to Expect from Lectures

The module runs in four blocks, each lecture in three or four sections. Each section runs in this order, and exercises are optional:

- **Intuition:** why the topic matters, before any symbol.
- **Notation:** the symbols that describe the idea.
- **Picture:** a figure of the idea, where one exists.

> **The Math Is Highly Repetitive**
> The math looks scary, but it repeats. One or two elements tell you what the econometricians are up to.

Timetable

Week	No.	Lecture	Due	Weight
1	1.1	Data and Variation	—	—
2	1.2	Conditional Means	—	—
3	1.3	Sampling Variation	—	—
4	2.1	Least Squares	Set 1	10%
5	2.2	Units and Fit	—	—
6	2.3	Bias and Precision	—	—
7	3.1	Estimation	Set 2 + **Midterm** (lab)	10% + **20%**
8	3.2	Inference	—	—
9	3.3	Functional Form	—	—
10	3.4	Dummies	—	—
11	4.1	Heteroskedasticity	Set 3	10%
12	4.2	Specification	—	—
Final exam			—	**50%**

The Problem Sets Are Exam Practice

The problem sets prepare you for the exam, and happen to carry marks:

> **AI Will Not Sit the Exam**
> AI will get you full marks on the problem sets. AI will not sit the exam for you. A set handed to a model is a week’s practice you did not do.

Expected Workload

Most of your own hours go on the problem sets and their reading:

24
hours of lectures, two hours a week

12
hours of labs

75
hours of your own, a little over six a week in term

SECTION 2

Populations and Samples

The Gap Between the Population and the Sample

The schooling question needs everybody's earnings, and we only ever hold some:

- **Population:** every person the question is about.
- **Sample:** the people from the population whose numbers we hold.
- **Statistical Inference:** arguing from the sample back to the population.

> **Later Techniques Measure the Gap, Not Remove It**
> Nothing later in the term removes the gap. Later techniques only measure how wide the gap might be.

Notation for a Single Observation

Our whole sample is written as one set:

$$\left\{ \overbrace{(x_i, y_i)}^{\text{an observation}} \right\}_{i=1}^{n} \tag{2}$$

- x_i: years of schooling completed by person i
- y_i: hourly earnings of person i, in euro
- i: which person, running from 1 to n
- n: how many people are in the sample

Cross-Section, Time Series and Panel Data

What you can estimate depends on how the data is indexed:

- **Cross-Section Data:** many people, each seen once, indexed by i.
- **Time-Series Data:** one unit, seen over many periods, indexed by t.
- **Panel Data:** many people, each over many periods, indexed by i and t.

Every Dataset Here Is Cross-Section
Our 120 people are each observed once. The assumptions change when the subscript does.

Roman Letters for the Sample, Greek for the Population

Roman here means the everyday alphabet, the one English is written in. Our sample holds n people; the population holds N.

> **Every Summary Comes as a Roman and a Greek Pair**
> Every summary from here comes as a pair. The Roman letter is computed from our 120. The Greek letter is the fact about all 200,000 that nobody sees.

The Population We Will Use All Term

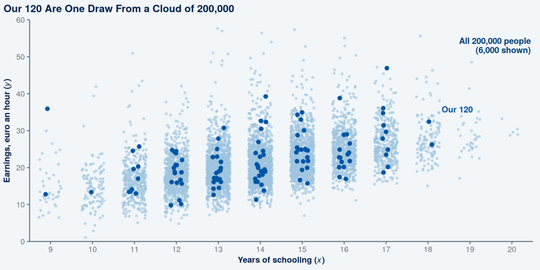

Figure 1: Schooling and hourly earnings: the population pale, our 120 in solid blue.

> **We Use the Solid Blue to Speak About the Pale Cloud**
> Our questions are about the pale cloud, all 200,000 people. The solid blue 120 are one random draw, and our numbers come from them.

Why a Simulated Population Rather Than Real Data

Real earnings data gives an estimate, never the population's true return.

> **1.30 Is a Setting We Chose**
> We wrote the generator, so 1.30 euro an hour a year is a setting. Nobody estimated 1.30.

Every estimate this term can be marked against that 1.30.

Each Flaw in the Generator Starts Switched Off

Real data has every flaw at once; later lectures switch the generator's flaws on:

- **Lecture 3.1:** ability, left out of the regression.
- **Lecture 4.1:** an error whose size grows with schooling.
- **Lecture 4.2:** schooling, measured with error.

SECTION 3

Centre and Spread

The Sample Mean and the Balance Property

At the sample mean, the deviations above and below sum to zero:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

- $\bar{x}$: the mean of x in our sample

> **Reading the Summation Sign**
> The $i = 1$ underneath says start at the first person. The n on top says stop at the last one.

Why the Bar Marks a Sample Quantity

A different 120 would give a different bar, but the same population mean.

> **Roman Bar, Greek Mu**
> A bar means the average of that letter in our sample. The population mean is Greek μ_X, because we do not know it.

Our 120 earn 12.25 euro an hour on average; all 200,000 earn 11.52.

Expectation: the Population Average of an Expression

The population average of anything built from x gets its own notation:

E Takes a Population Average

$E(\cdot)$ takes whatever expression sits inside it and returns its population average. $E(x)$ is μ_x, and $E(x^2)$ is the population average of x^2.

Mean, Median and Mode

A symmetric, single-peaked distribution gives all three the same value:

- **Mean:** the balance point, and the one a long tail pulls.
- **Med**i**an:** the m**i**ddle value, with half the people either side.
- **M**o**de:** the m**o**st common value, at the peak.

> **The I of Middle, the O of Most**
> Median and mode are the pair students swap. The marked letters tell them apart.

A Long Right Tail Puts the Mean Above the Median

A town's earnings, with a long right tail, pull the three centres apart:

- **Median:** 26,000, with half the town earning less.
- **Mean:** 34,000, pulled up by the highest earners.
- **Mode:** lower still, at the peak.

> **The Mean Can Describe Nobody**
> Report the mean and you describe a person who may not exist. A headline on skewed earnings needs the median beside it.

The Three Centres on Two Distributions

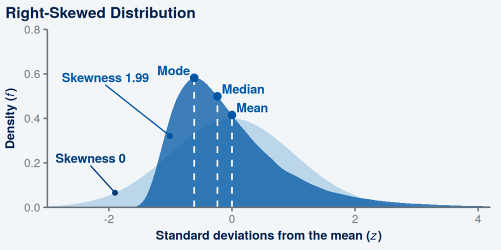

Figure 2: A long right tail: the three separate.

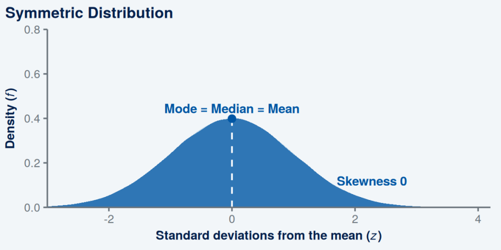

Figure 3: No tail either way: the three coincide.

A Symmetric Distribution Has One Centre

Dashed leaders mark the mode, the median and the mean. On a symmetric distribution all three sit at the same point.

Why a Mean Need Not Be a Possible Value

Roll a die long enough and the rolls average 3.5. No face carries 3.5, so no single roll ever lands on the average.

> **Every Roll Lands 0.5 to 2.5 Away**
> The mean tells you where the rolls sit. Every roll lands 0.5, 1.5 or 2.5 from the average.

The Variance: Average Squared Distance from the Mean

The sample variance measures spread by squared distance from the mean:

$$s_x^2 = \overbrace{\frac{1}{n-1}}^{\text{Bessel's correction}} \sum_{i=1}^{n} \underbrace{(x_i - \bar{x})^2}_{\text{squared distance from the mean}} \tag{4}$$

- s_x^2: the variance of x in the sample

> **Why We Square the Distance**
> We care how far a person sits from the mean, not which side. Unsquared, the distances cancel and the total comes to exactly zero.

Bessel's Correction: Only N Minus One Deviations Are Free

Deviations from $\bar{x}$ sum to zero, so any $n - 1$ of them fix the last:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0 \tag{5}$$

Our 120 deviations hold 119 free numbers, so the divisor is 119.

> **Dividing by N Understates the Spread**
> The sum of squared deviations from $\bar{x}$ is never larger than from μ_X. Dividing by n therefore gives too small an answer on average. The $n - 1$ divisor lifts every sample's answer. Averaged over many samples, the answers come out at the population variance σ_X^2.

Bessel's Correction Matters Only in Small Samples

The $n - 1$ divisor matters little at 120 and a great deal at two:

- $n = 120$: 1/119 against 1/120, a difference under one per cent.
- $n = 2$: the divisor is 1, so the sum of squares is the estimate.
- $n = 1$: the divisor is 0, so there is no estimate at all.

The Standard Deviation Is Measured in Euro

Taking the positive square root returns a variance to the data's own units:

$$s_x = \sqrt{s_x^2} \qquad \sigma_x = \sqrt{\sigma_x^2} \tag{6}$$

- s_x: the standard deviation of x in the sample
- σ_x: the standard deviation of x in the population

Our 120's hourly earnings have a standard deviation of 5.91 euro.

> **Squaring Cost the Units**
> Squaring stopped the distances cancelling, and it put earnings in squared euro. Every square is positive, so the root exists and returns euro.

Comparing Centre and Spread Across Distributions

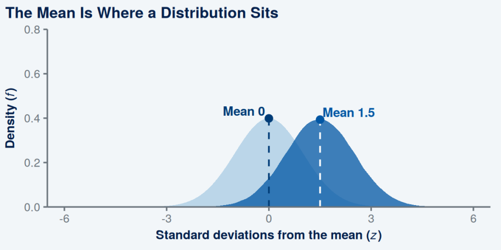

Figure 4: The mean: same variance, different mean.

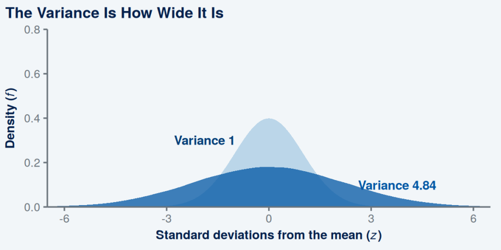

Figure 5: The variance: same mean, different variance.

> **Changing the Mean Leaves the Variance Alone**
> Light blue is the same reference curve in both panels. Shifting the mean leaves the variance unchanged, and widening the variance leaves the mean.

SECTION 4

Moments and Shape

The Moments of a Distribution

Mean, variance, skewness and kurtosis are the first four moments. The last three raise one bracket to a power:

$$(x_i - \bar{x})^2\,, \qquad (x_i - \bar{x})^3\,, \qquad (x_i - \bar{x})^4 \tag{7}$$

Today's Moments Are Sample Moments

We average these brackets over our 120, so today's moments are sample moments. The population versions put μ_X where $\bar{x}$ stands.

A Higher Power Weights Far-Out People More

Double a person's distance from the mean, and each power multiplies their contribution by its own factor:

- **Second Power:** their contribution grows fourfold.
- **Third Power:** their contribution grows eightfold and keeps its sign.
- **Fourth Power:** their contribution grows sixteenfold.

> **The Tails Decide the Fourth Power**
> At the fourth power, the people furthest from the mean mostly decide the sum. The cube keeps the sign, so it says which tail is longer.

Skewness Cubes the Standardised Deviation

Skewness asks which side of the mean has the longer tail:

$$\text{skew}(x) = \frac{1}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{x_i - \bar{x}}{s_x}\right)^3}_{\text{the standardised deviation, cubed}} \tag{8}$$

- skew(x): the skewness of x in the sample

> **Dividing by the Spread Removes the Units**
> A deviation and s_x share their units, so the ratio has none. Skewness in cents equals skewness in euro.

Bulmer’s Bands for the Size of Skewness

Bulmer’s bands classify skewness by its size, sign stripped off (Bulmer, 1979):

Size of the skewness	**Bulmer’s reading**
Below 0.5	approximately symmetric
0.5 to 1	moderately skewed
Above 1	highly skewed

> **The Sign Gives the Direction**
> Positive means a long right tail, negative a long left one.

Skewness: Which Tail Is Longer

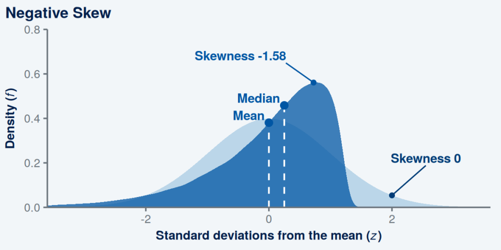

Figure 6: Negative skew: the long tail runs left.

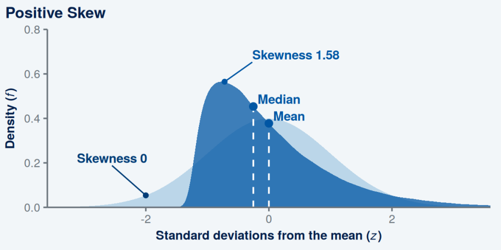

Figure 7: Positive skew: the long tail runs right.

> **The Mean Sits Further into the Tail Than the Median**
> Both curves share the mean and variance; only the third power separates them. Dashed leaders mark the median and the mean, out in the tail.

Kurtosis: the Standardised Deviation to the Fourth

Kurtosis asks how much of the distribution sits far from the mean:

$$\text{kurt}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 \tag{9}$$

- $\text{kurt}(x)$: the kurtosis of x in the sample

Kurtosis Is Read Against the Normal Curve's Three

The normal curve has a kurtosis of three, and every reading starts there:

Kurtosis	Tails, against the normal curve's
Below 3	thinner
Exactly 3	the same
Above 3	thicker, with more weight in them

No Agreed Bands, Only the Benchmark

Nobody has settled bands for kurtosis the way Bulmer did for skewness. Daily stock returns sit well above three. Finance cares about kurtosis because tail thickness sets the probability of a large loss.

Kurtosis: How Much Sits Far from the Middle

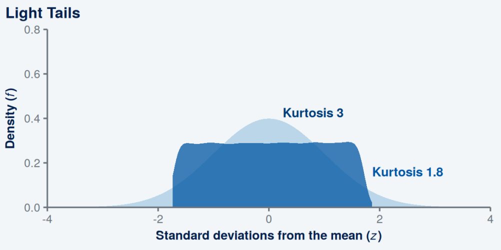

Figure 8: Light tails: almost nothing lands far out.

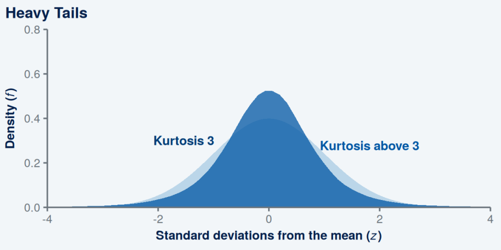

Figure 9: Heavy tails: much more lands a long way out.

> **Thicker Tails, Same Variance**
> Both panels share the normal curve's mean and variance; only the fourth power separates them. The right panel produces far-out values more often.

The Range: Spread from Only Two Observations

Unlike the moments, the range is built from just two observations:

$$\text{range} = x_{\max} - x_{\min} \tag{10}$$

- $x_{\max}, x_{\min}$: the largest and smallest values of x

> **The Range Uses Only the Largest and Smallest Values**
> Every other observation could change without changing the range. A bigger sample can only widen the range, never narrow it.

SECTION 5

Describing Two Variables

Covariance: Measuring How Two Variables Move Together

The covariance asks whether schooling and earnings rise and fall together:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{(x_i - \bar{x})(y_i - \bar{y})}_{\text{the cross-product}} \tag{11}$$

- s_{xy}: the covariance of x and y in the sample
- $\bar{y}$: the mean of y in our sample

> **Same Side Positive, Opposite Sides Negative**
> Above on both, or below on both, and the product is positive. Above on one and below on the other, and the product is negative.

Only the Sign of a Covariance Is Interpretable

Our 120 cross-products sum to a positive number, and only the sign can be read:

- **Positive:** people with above-average schooling mostly earn above average.
- **Negative:** people with above-average schooling mostly earn below average.
- **Near Zero:** the positive products and the negative ones roughly cancel.

> **Zero Is Cancellation, Not a Definition**
> A covariance near zero is not a rule about unrelated variables. Zero is what happens when the agreements and the disagreements cancel out.

Why the Size of a Covariance Is Not Interpretable

Like the variance, the covariance carries units, so rescaling changes its size:

- **Earnings in Euro an Hour:** the covariance is 8.11 euro-years.
- **Earnings in Cents an Hour:** the covariance is 811.16 cent-years.
- **Schooling in Months Instead:** the covariance is 97.34 euro-months.

> **Why We Normalise the Covariance**
> Only the units changed, yet the covariance ran from 8.11 to 811.16. Dividing the units out leaves one number all three measurements share.

Correlation Is Covariance with the Units Divided Out

Dividing the covariance by both standard deviations cancels its euro-years:

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y} \tag{12}$$

- r_{xy}: the correlation of x and y in the sample
- s_y: the standard deviation of y in the sample

All Three Covariances Give 0.594

Every one of the three covariances, 8.11, 811.16 and 97.34, divides down to 0.594. Both standard deviations are positive, so the covariance's sign survives.

Correlation Runs from Minus One to One

A correlation is bounded, and each bound is an exact straight line:

$$-1 \leq r_{xy} \leq 1 \tag{13}$$

Our correlation is 0.594.

At Either Bound the Points Lie on a Line
At 1 the points lie on an upward straight line. At -1 the line slopes down. Values nearer either end put the points closer to a straight line.

Cohen's Bands: Small, Medium and Large Correlations

Cohen's bands for a correlation's size are the ones you will meet everywhere (Cohen, 1988):

- $r = 0.1$: Cohen calls a correlation this size small.
- $r = 0.3$: Cohen calls a correlation this size medium.
- $r = 0.5$: Cohen calls a correlation this size large.

> **Cohen Wrote His Bands for Psychology**
> Wooldridge gives no bands at all. Our 0.594 is large by Cohen. Squared, 0.594 says schooling and earnings share a third of their variation.

Correlations of 0.10, 0.30 And 0.50

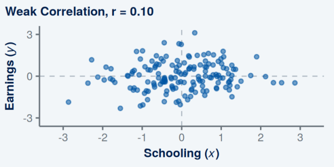

Figure 10: Cohen's small, constructed, not our sample.

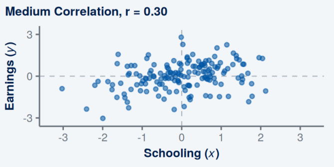

Figure 11: Cohen's medium, constructed, not our sample.

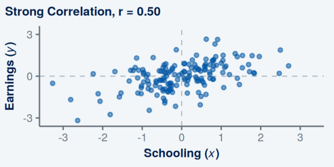

Figure 12: Cohen's large, constructed, not our sample.

> **Earnings Depend on More Than Schooling**
> These clouds are built to hit each value exactly; none is our 120. Even the one Cohen calls large is a wide cloud, and block 3 adds the other variables earnings depend on.

Correlation Measures Straight-Line Association Only

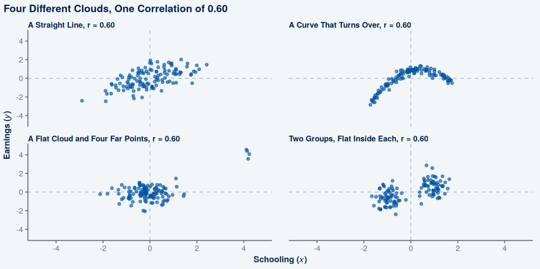

Figure 13: Four constructed samples, not from our population, each with a correlation of 0.60.

> **Same Number, Four Clouds**
> A curve, four outliers and two separate groups all return 0.60. Only closeness to a straight line enters the number (Anscombe, 1973; Matejka & Fitzmaurice, 2017).

Why a Correlation Is Not a Causal Claim

We wrote a return of 1.30 euro an hour a year into the generator.

> **Chance and Mechanism Look Alike**
> The same 0.594 could be the product of chance, or of a mechanism running from schooling to earnings. A correlation measures co-movement and says nothing about which of the two produced it.

Outside a simulation nobody hands you the mechanism, only the number.

Three Ways a Correlation Misleads

All three correlations are real. A causal reading of each fails for a different reason:

- **Coincidence:** films starring Nicolas Cage track pool drownings, $r = 0.66$.
- **A Confounder:** sunshine raises both ice cream sales and swimming.
- **A Confounder Beside a Mechanism:** schooling opens better careers, and those who stay differ in other ways.

> **Ice Cream Does Not Cause Drowning**
> Nothing connects Cage to drownings. Sunshine causes the ice cream and the swimming. Ability raises schooling and earnings alike.

What Would Have to Be True?

Today’s number is the correlation between schooling and earnings in our sample.

> **Question**
> What would have to be true for this number to mean what we just said it means? And what is the plausible story where it isn’t?

Build a story in which this number means what we said. Then build the story that makes it false.

Two Conditions Behind Our 0.594

Our 0.594 means what we said only if two conditions hold:

- **A Straight-Line Cloud:** the 120 points lie close to a straight line.
- **A Representative Sample:** our 120 represent the whole population.

Neither condition can be read off 0.594 itself.

Summary: Samples, Centre, Shape and Two Variables

Today was four ideas, each answering a different question:

1. **Populations and Samples.**
   The question is about a population; the gap to the sample never closes.
2. **Centre and Spread.**
   The mean says where; the variance and the standard deviation say how far.
3. **Moments and Shape.**
   Higher powers read the tails; the range reads only two people.
4. **Describing Two Variables.**
   Covariance and correlation measure straight-line co-movement, and nothing else.

Before Lecture 1.2

1. **Play:**
   load the 120 in R and compute its mean, spread and correlation.
2. **Apply:**
   problem set 1, which opens with the appendix derivation of Bessel's correction.
3. **Review:**
   Wooldridge (2020), Math Refresher B-3 and B-4, then C-1 and C-2.

> **Diagnostic in This Week's Lab**
> Ten anonymous, ungraded items on expectation, sampling, logs and proportions.

Next Lecture

Instead of one average for everybody, an average for each level of schooling.

> **An Average at Each Year**
> The line a regression draws joins one average per year of schooling.

Thank you

sam.deegan@ucdconnect.ie
sam-deegan.com

BACKUP

Appendix

Notation Used in This Lecture

Symbol	What it is	First met
y_i	the outcome; here, hourly earnings of person i, in euro	Regression Studies How an Outcome Varies With Predictors
x_i	the predictor; here, years of schooling completed by person i	Regression Studies How an Outcome Varies With Predictors
β_0	the intercept	Regression Studies How an Outcome Varies With Predictors
β_1	the slope	Regression Studies How an Outcome Varies With Predictors
u_i	the error, or unexplained outcome	Regression Studies How an Outcome Varies With Predictors
i	which person, running from 1 to n	Notation for a Single Observation
n	how many people are in the sample	Notation for a Single Observation
t	which period, when one unit is seen over many	Cross-Section, Time Series and Panel Data
N	how many people the population holds	Roman Letters for the Sample, Greek for the Population

Notation Used in This Lecture

Symbol	What it is	First met
$\sum_{i=1}^{n}$	add up what follows, from the first person to the last	The Sample Mean and the Balance Property
$\bar{x}$	the mean of x in our sample	The Sample Mean and the Balance Property
μ_x	the population mean of x	Why the Bar Marks a Sample Quantity
$E(\cdot)$	the population average of the expression inside it	Expectation: The Population Average of an Expression
s_x^2	the variance of x in the sample	The Variance: Average Squared Distance From the Mean
σ_x^2	the population variance of x	Bessel's Correction: Only n Minus One Deviations Are Free
s_x	the standard deviation of x in the sample	The Standard Deviation Is Measured in Euro
σ_x	the standard deviation of x in the population	The Standard Deviation Is Measured in Euro
$\text{skew}(x)$	the skewness of x in the sample	Skewness Cubes the Standardised Deviation

Notation Used in This Lecture

Symbol	What it is	First met
kurt(x)	the kurtosis of x in the sample	Kurtosis: The Standardised Deviation to the Fourth
range	the largest value of x less the smallest	The Range: Spread From Only Two Observations
x_{max}	the largest value of x in the sample	The Range: Spread From Only Two Observations
x_{min}	the smallest value of x in the sample	The Range: Spread From Only Two Observations
s_{xy}	the covariance of x and y in the sample	Covariance: Measuring How Two Variables Move Together

Notation Used in This Lecture

Symbol	What it is	First met
$\bar{y}$	the mean of y in our sample	Covariance: Measuring How Two Variables Move Together
r_{xy}	the correlation of x and y in the sample	Correlation Is Covariance With the Units Divided Out
s_y	the standard deviation of y in the sample	Correlation Is Covariance With the Units Divided Out
r	a correlation, written without its subscripts	Cohen's Bands: Small, Medium and Large Correlations

Derivation: Bessel's Correction

No other number makes the squared deviations smaller than $\bar{x}$ does. Their expected sum comes to $(n-1)\sigma_x^2$, not $n\sigma_x^2$.

> **The Cross Term Becomes Part of the Squared Gap**
> Write each deviation from $\bar{x}$ as a deviation from μ_x minus the gap. Square and sum, and the cross term merges into the squared gap. The gap has variance σ_x^2/n, so n copies of it remove one σ_x^2.

Dividing by n is too small on average by the factor $(n-1)/n$.

[Full derivation. Set as homework; worked solution published with the set.]

The Greek Alphabet: Alpha to Mu

Say them out loud. A symbol you cannot pronounce is one you will never ask about:

Lower	Upper	Name	Said as	Used here for
α	A	alpha	AL-fa	
β	B	beta	BEE-ta	the regression coefficients
γ	Γ	gamma	GAM-a	
δ	Δ	delta	DEL-ta	a shift; Δ for a change in
ϵ	E	epsilon	EP-si-lon	
ζ	Z	zeta	ZEE-ta	
η	H	eta	EE-ta	
θ	Θ	theta	THEE-ta	
ι	I	iota	eye-OH-ta	
κ	K	kappa	KAP-a	
λ	Λ	lambda	LAM-da	
μ	M	mu	MEW	a population mean

The Greek Alphabet: Nu to Omega

Lower	Upper	Name	Said as	Used here for
ν	N	nu	NEW	
ξ	Ξ	xi	KSY	
o	O	omicron	OH-mi-kron	
π	Π	pi	PY	
ρ	P	rho	ROH	
σ	Σ	sigma	SIG-ma	a population spread; Σ means add up
τ	T	tau	TAW	
υ	Υ	upsilon	UP-si-lon	
ϕ	Φ	phi	FY	
χ		chi	KY	
ψ	Ψ	psi	SY	
ω	Ω	omega	OH-mig-a	

Thirteen capitals are identical to ours, so nobody uses them as symbols. A capital alpha is just an A.

How to Study This Subject

Four habits decide what your own hours are worth:

- **Re-derive:** the lecture’s central result, from a blank page, before the tutorial.
- **Attempt:** the problem set before reading the answers, not after.
- **Space:** the work across the week, since four short days beat one long.
- **Avoid:** rereading and highlighting, which give fluency without retrieval.

> **Nothing on This Frame Is Examinable**
> Ten minutes today, and no time after that. The four habits above are worth more than any revision guide.

Notes and Tools

- **Wooldridge:** chapter 1 for the subject, Math Refreshers B and C for today (Wooldridge, 2020).
- **Angrist and Pischke:** section 6.1, on why the schooling question fills a term (Angrist & Pischke, 2014).
- **The Train-Test Split:** chapter 2 of James et al. (2021), stated properly.
- **Pictures First:** Healy (2018) on the plot before the statistic, Wickham et al. (2023) for the R.
- **Software:** R and RStudio, both free. [Install links and the module page.]

References

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering 'Metrics: The path from cause to effect*. Princeton University Press.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21. https://doi.org/10.1080/00031305.1973.10478966

Bulmer, M. G. (1979). *Principles of statistics*. Dover Publications.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Healy, K. (2018). *Data visualization: A practical introduction*. Princeton University Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025912

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for data science* (2nd ed.). O'Reilly Media.

References

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.