LECTURE 2.1

The Simple Regression Model

ECON30130 Econometrics I

Sam Deegan
University College Dublin

Autumn 2027

Where We Are

1. Foundations	2. Simple Regression	3. Multiple Regression	4. Broken Assumptions
1.1 Data and Variation	**2.1 Least Squares**	3.1 Estimation	4.1 Heteroskedasticity
1.2 Conditional Means	2.2 Units and Fit	3.2 Inference	4.2 Specification
1.3 Sampling Variation	2.3 Bias and Precision	3.3 Functional Form	
		3.4 Dummies	

Block 2 Is Three Lectures on One Estimator

Today's lecture fits the line and measures its fit. Lecture 2.2 reads the slope in euro and tests the fit on new rows, and Lecture 2.3 asks when the slope is unbiased.

Last Week's Lecture

Lecture 1.3 asked how far an estimate sits from the number it estimates:

1. **The Sampling Distribution.**
   The standard error measures how far a sample mean typically sits from the truth.

2. **The Central Limit Theorem and Confidence Intervals.**
   Sample means are close to normal, so one estimate can carry a 95 per cent interval.

3. **Hypothesis Tests.**
   A test says what a claimed mean would produce, not whether the claim is true.

Guessing Earnings from Schooling Alone

Each of our 120 people comes with years of schooling and an hourly wage.

> **Question**
> Given one person's years of schooling and nothing else, what would you guess they earn? What would you need to know to make that guess?

What Lectures 1.2 and 1.3 Left Open

The average earnings of everyone with that schooling answers the question, and we hold only 120 people:

- **Lecture 1.2:** that average is the **conditional expectation** of earnings given schooling.
- **Lecture 1.3:** a number computed from 120 people is one draw.

> **Today's Slope Comes from One Draw**
> Our sample mean differed from the population mean by 0.49 euro. Today's lecture chooses one shape and fits it to the same 120 people. Our slope comes from one draw, so it differs from the population slope.

Lecture Plan

1. **The Line as a Conditional Expectation.**
   A line is a claim about the average earnings of people with a given schooling.

2. **Least Squares.**
   Minimising the squared gaps picks one line and gives its slope formula.

3. **Fitted Values, Residuals and Sums of Squares.**
   The chosen line splits every person's earnings into a fitted value and a residual.

4. **Goodness of Fit.**
   R-squared measures the share of earnings variation the line accounts for.

SECTION 1

The Line as a Conditional Expectation

Schooling and Earnings in Our 120 People

Figure 1: Our 120 people: years of schooling across, hourly earnings up the side.

> **Earnings Rise with Schooling and Still Scatter**
> Earnings rise with schooling across the picture. At any one level of schooling people still earn very different amounts.

A Line Is the Simplest Conditional Expectation

The simplest form of a conditional expectation function is a straight line. A straight line commits to one claim:

> *One more year of schooling changes expected earnings by the same amount. The ninth year and the nineteenth are worth exactly the same.*

A Line Has Two Numbers Where the CEF Has Many

The conditional expectation function (CEF) can bend, jump or flatten at any level of schooling. A line has two numbers; the CEF has one at every level.

> **A Straight Line Is Imposed, Not Found**
> Nothing in the definition of a conditional expectation makes it straight. We impose the line; the population does not supply it.

The constant slope is a restriction, not a discovery (Angrist & Pischke, 2015).

The Simple Regression Model and Its Notation

The model writes down what one line claims about one person:

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

- β_0: expected earnings at zero years of schooling
- β_1: the change in expected earnings for one more year
- u_i: every factor in person i's earnings other than schooling

> **Both Betas Describe the Population**
> β_0 and β_1 are Greek, so no sample changes either one.

Centring Schooling at a Reference Level

Nobody in this population left school at zero years, where β_0 sits. Subtracting a reference level makes β_0 the expected earnings at that level:

$$y_i = \beta_0 + \beta_1 \underbrace{(x_i - x_0)}_{\text{the centred regressor}} + u_i \qquad (2)$$

- x_0: the reference level of schooling, fixed before we see any data

> **Our Reference Level Is the Leaving Certificate**
> We set x_0 at thirteen years, eight of primary and five of second level. Centring renames β_0 and leaves the slope β_1 untouched.

Taking the Expectation of the Model at Fixed Schooling

At a fixed level of schooling (x), the two β terms are constants and pass through unchanged:

$$E(y \mid x) = \beta_0 + \beta_1(x - x_0) + E(u \mid x) \tag{3}$$

- $E(u \mid x)$: the average of everything else among people with that schooling

When the Line Is the Conditional Expectation

If everything else averages zero at every level of schooling, only the line is left (Wooldridge, 2020):

$$E(u \mid x) = 0 \quad \Longrightarrow \quad E(y \mid x) = \beta_0 + \beta_1(x - x_0) \tag{4}$$

> **Everything Else Must Not Differ with Schooling**
> Suppose people who stayed longer in school also have higher ability. Then $E(u \mid x)$ rises with schooling, and the line is not the CEF.

The Hat on the Slope Marks a Sample Estimate

An estimate from our 120 carries a hat, as $\hat{\mu}_y$ did in Lecture 1.3:

- $\hat{\beta}_1$: the slope computed from our 120

A Hat Marks a Sample Quantity
β_1 is Greek with no hat: one number for the population of 200,000. $\hat{\beta}_1$ is that letter estimated from the 120 rows you hold. A different 120 would return a different number.

SECTION 2

Least Squares

Comparing Candidate Lines Through the Same Scatter

Figure 2: Three lines through our 120: least squares and two drawn by hand.

> **The Totals Rank the Three Lines**
> Every line leaves each person a gap, and the squared gaps total differently. Nothing so far says why we square each gap instead of adding up their sizes.

Least Squares Minimises the Total of Squared Gaps

Least squares adds the squared gap a line leaves at every person (Wooldridge, 2020):

$$\mathrm{SSR}(b_0, b_1) = \sum_{i=1}^{n} \underbrace{\big(y_i - b_0 - b_1(x_i - x_0)\big)^2}_{\text{the squared residual}} \tag{5}$$

- b_0: a candidate intercept, not yet the chosen one
- b_1: a candidate slope, not yet the chosen one

> **Least Squares Is a Chosen Criterion**
> Least squares picks the pair minimising this total. The model does not require squares.

Why the Residuals Are Squared

A residual is squared for Lecture 1.1's reason and for two more:

- **Raw Gaps Cancel:** residuals sum to zero for a line through both means.
- **Absolute Gaps Give No Formula:** the absolute gap has no derivative at zero.
- **Squares Are Differentiable:** two derivatives set to zero give two solvable equations.

> **One Person Can Dominate the Total**
> A residual of 10 adds 100, and ten residuals of 1 add ten. Our largest residual, 21.60, supplies eleven per cent of the total.

Squares Rather Than Absolute Values

Minimising the total absolute gap picks a different line, called least absolute deviations. Least absolute deviations gives an extreme earner less weight than least squares does.

> **Question**
> Which of the two would you fit to earnings data? What does that choice assume about the highest earners?

Least Absolute Deviations Fits the Conditional Median

The two criteria estimate different averages of earnings at each level of schooling (Wooldridge, 2020):

- **Least Squares:** estimates the conditional mean, $E(y \mid x)$.
- **Least Absolute Deviations:** estimates the conditional median.

> **Mean and Median Agree Under Symmetry**
> The conditional mean and median agree when earnings are symmetric about the line. Least squares estimates the mean either way.

Deriving the Slope: Two Conditions

At the smallest total, changing either number cannot lower it further. Both derivatives are therefore zero at the least-squares pair $(\hat{\beta}_0, \hat{\beta}_1)$ (Wooldridge, 2020):

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - x_0)\right) &= 0 \\ \sum_{i=1}^{n} (x_i - x_0)\left(y_i - \hat{\beta}_0 - \hat{\beta}_1(x_i - x_0)\right) &= 0 \end{aligned} \tag{6}$$

Zero Mean, Zero Co-Movement with Schooling

The plain sum says the residuals average to zero. The schooling sum says the residuals carry no remaining co-movement with schooling.

Deriving the Slope: Solving for the Intercept

The y_i sum to $n\bar{y}$, and a constant sums to n times itself. Dividing the first condition by n leaves the intercept in terms of the slope:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1(\bar{x} - x_0) \qquad (7)$$

Deriving the Slope: Solving the Pair

Substituting the intercept into the second condition cancels the constants in pairs, leaving deviations from the two means:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \tag{8}$$

The Slope Is a Sample Covariance over a Sample Variance
Dividing both sums by $n-1$ gives Lecture 1.1's s_{xy} and s_x^2. Subtracting x_0 shifts x_i and $\bar{x}$ alike, so the centring cancels.

Worked example: our 120 people

A Slope Needs Schooling to Vary

The slope's denominator adds up squared deviations of schooling from its mean:

> **The Denominator Must Not Be Zero**
> If every person had the same schooling, the denominator would be zero. A line needs schooling to vary. The less schooling varies, the more the slope differs from sample to sample.

Residuals as Vertical Distances from the Fitted Line

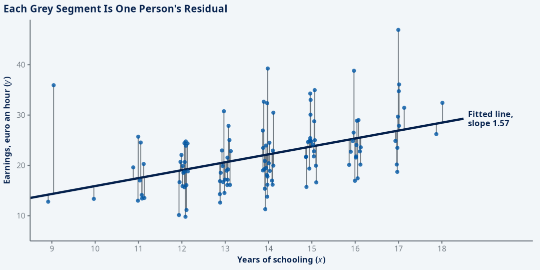

Figure 3: The least-squares line and the 120 residuals it leaves.

> **Every Person Has One Vertical Residual**
> Each segment runs vertically, because the criterion scores earnings alone. No other line makes the squared segments total less than 4,238.3.

Moving the Line and Watching the SSR

In the app you place a line by hand, read its SSR, then reveal the least-squares line.

> **Question**
> Drop the three highest earners: does the fitted slope rise, fall, or hold? Would your answer change if those three had left school at fifteen?

The Earners’ Schooling Decides Whether the Slope Rises

Dropping three earners above the line changes the slope in a direction set by their schooling:

- **High Schooling:** dropping them lowers the fitted slope.
- **Low Schooling:** dropping them raises the fitted slope.

> **Schooling Sets the Direction, Not Earnings**
> The slope weights each person’s earnings by their distance from average schooling, 13.93. People below 13.93 years enter with a negative weight.

SECTION 3

Fitted Values, Residuals and Sums of Squares

Splitting One Person's Earnings into Two Parts

Each person's earnings split into the line's prediction plus a residual (Wooldridge, 2020):

$$y_i = \hat{y}_i + \hat{u}_i \qquad \text{where} \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1(x_i - x_0) \tag{9}$$

- $\hat{y}_i$: the fitted value for person i
- $\hat{u}_i$: the residual for person i

> **A Residual Is Not an Error**
> u_i is the population error, and nobody ever sees it. A hat marks a sample quantity, so $\hat{u}_i$ estimates u_i from our line. A different line changes every residual and leaves every error alone.

The Split Shown on One Person

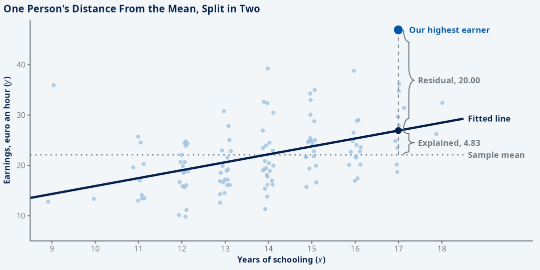

Figure 4: Our highest earner's distance from the average, split into two parts.

> **One Distance, Cut into Two Segments**
> The line accounts for the 4.83 from the average up to the fitted value. The residual, 20.00, is exactly the rest of the distance to this person.

The Two Sums Least Squares Sets to Zero

The two conditions behind the slope are two sums of residuals, both exactly zero (Wooldridge, 2020):

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad \sum_{i=1}^{n} (x_i - x_0)\,\hat{u}_i = 0 \tag{10}$$

The Two Sums Hold on Any Sample

These two sums are not assumptions about the population. Both sums are derivatives of the squared residual total, set to zero. Both therefore hold on any sample at all.

The Mean Point Lies on the Fitted Line

The plain sum alone fixes two facts about the fitted line (Wooldridge, 2020):

- **The Fitted Values:** their average equals $\bar{y}$, the average of earnings.
- **The Mean Point:** the line passes through $(\bar{x}, \bar{y})$, here (13.93, 22.09).

> **A Zero Total Says Nothing About One Person**
> The residuals add to zero on any data set. Their zero total says nothing about any one person's residual.

A Large Residual Is Information About the Model

Our largest residual, 21.60, supplied 466.6 of the total 4,238.3. That residual has three possible sources:

- **A Rare Occupation:** the person may earn far more than schooling predicts.
- **A Mistyped Wage:** the recorded wage may be wrong.
- **A Missing Variable:** the line may be missing a variable, such as ability.

> **A Residual Belongs to the Person and the Line**
> Reading that person as a data error is tempting. The size of a residual alone cannot say whether the person or the line produced it.

No Residual Total Detects a Missing Variable

The two conditions constrain the residuals as a set, not one person at a time.

> **Both Conditions Hold with Ability Left Out**
> Both conditions hold for a line that leaves ability out and for one that does not. No residual total can therefore tell you that a variable is missing, or that the true relationship bends.

A pattern across the residuals is evidence about the line, not about one person.

The Three Sums of Squares and What Each Measures

Adding one person's split over all 120 people gives three totals (Wooldridge, 2020):

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad \text{SSR} = \sum_{i=1}^{n}\hat{u}_i^2 \tag{11}$$

- SST: the total variation in earnings
- SSE: the variation in the fitted values
- SSR: the total of the squared residuals

> **SSE Is the Explained Total, Not an Error Total**
> Other texts and some software call the residual total the error sum of squares. Here SSE is the explained total, following Wooldridge (2020).

The Total Sum of Squares as Vertical Segments

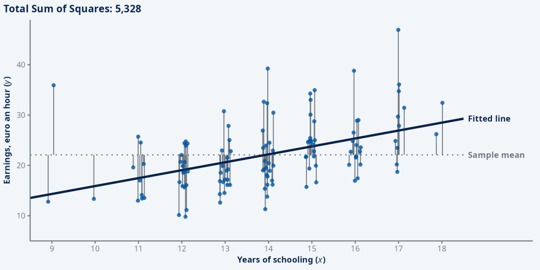

Figure 5: Each segment runs from a person to the sample average of earnings.

> **The Benchmark Is the Sample Average**
> SST is the total we would face without any schooling data. Predict 22.09 for everybody, square what is left, and the total comes to 5,327.6.

The Explained Sum of Squares as Vertical Segments

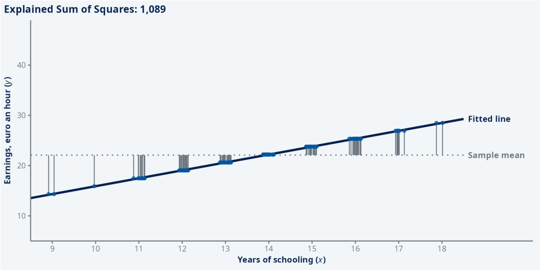

Figure 6: Each segment now runs from the fitted line down to the sample average.

> **A Flat Line Would Contribute Nothing Here**
> If $\hat{\beta}_1$ were zero, every fitted value would equal 22.09. SSE would then be zero, and on our 120 SSE is 1,089.3.

The Residual Sum of Squares as Vertical Segments

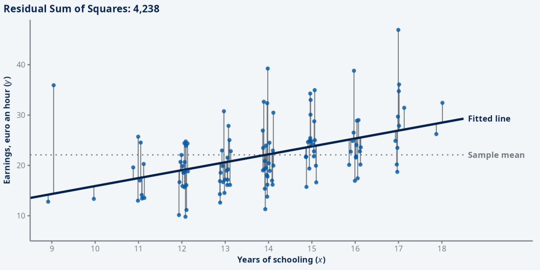

Figure 7: Each segment now runs from a person to the fitted line.

> **The Line Leaves This Total Unexplained**
> SSR adds up each person's squared gap to the fitted line. On our 120 it is 4,238.3, and least squares made it as small as any line can.

SECTION 4

Goodness of Fit

Deriving the Identity: Splitting the Deviation from the Mean

Adding and subtracting the fitted value splits every person's distance from the average in two:

$$y_i - \bar{y} = \hat{u}_i + (\hat{y}_i - \bar{y}) \tag{12}$$

Squaring both sides and summing over the 120 gives three terms:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{13}$$

The Identity Holds Only If the Middle Term Is Zero

Any line at all gives these three terms, not just least squares. The middle term doubles because squaring gives the same cross product twice.

Deriving the Identity: Why the Cross Term Is Zero

Substituting the intercept into the fitted value cancels the centring constant:

$$\hat{y}_i - \bar{y} = \hat{\beta}_0 + \hat{\beta}_1(x_i - x_0) - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) \tag{14}$$

Put that deviation into the cross term and take $\hat{\beta}_1$ outside:

$$\sum_{i=1}^{n} \hat{u}_i(\hat{y}_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})\,\hat{u}_i = \hat{\beta}_1\left[\sum_{i=1}^{n}(x_i - x_0)\,\hat{u}_i - (\bar{x} - x_0)\sum_{i=1}^{n}\hat{u}_i\right] = 0 \tag{15}$$

Both Bracketed Sums Are Least Squares' Conditions

The schooling sum is zero because the residuals retain no co-movement with it. The plain sum is zero because the residuals average to zero. With both zero, SST equals SSE plus SSR exactly.

The Three Totals on Our 120 People

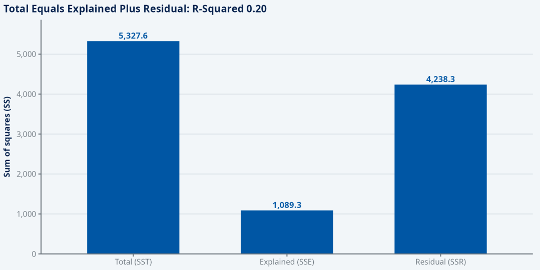

Figure 8: SST, SSE and SSR as bars: SST is the height of the other two stacked.

> **The Arithmetic Closes Exactly**
> SSE of 1,089.3 plus SSR of 4,238.3 gives SST of 5,327.6. The cross term vanished, so the closure is exact rather than approximate.

Defining R-Squared from the Three Totals

The three totals, in squared euro per hour, cannot be compared across datasets. Dividing by SST, the whole the other two split, removes the units and the sample size (Wooldridge, 2020):

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{16}$$

- R^2: the share of the sample variation in earnings the line accounts for

> **R-Squared Is a Share Because the Identity Holds**
> The first ratio is SSE over SST, the second one minus SSR over SST. They agree, and R^2 lies between nought and one, only because SSE and SSR add to SST.

R-Squared Is the Squared Correlation

Dividing top and bottom by $s_x s_y$ writes the slope with Lecture 1.1's correlation (r_{xy}) (Wooldridge, 2020):

$$\hat{\beta}_1 = r_{xy} \cdot \frac{s_y}{s_x} \qquad \Longrightarrow \qquad R^2 = r_{xy}^2 \tag{17}$$

For our 120 the correlation is 0.4522, and squaring it gives 0.2045.

> **Squaring Discards the Sign**
> R-squared never says whether the slope is positive.

Worked example: our 120 people

R-Squared Is a Squared Correlation in Any Regression

In any regression, R-squared is the squared correlation between earnings and the fitted values (Wooldridge, 2020).

Why One Regressor Gives the Schooling Correlation

With one regressor the fitted values are a straight-line function of schooling. Their squared correlation with earnings is therefore r_{xy}^2.

Adding a Variable That Cannot Matter

We run two regressions on the same 120: schooling alone, then schooling plus height in centimetres.

> **Question**
> Which of the two regressions has the higher R-squared? Does your answer depend on whether height matters for earnings?

Adding a Regressor Cannot Lower R-Squared

The regression with height has an R-squared at least as high, whether or not height matters for earnings.

> **Adding a Regressor Can Only Raise R-Squared**
> Adding any regressor to a model can only raise R-squared, never lower it. Least squares could set height's coefficient to zero, so the SSR cannot rise.

R-squared alone cannot decide whether height belongs in the model.

What Would Have to Be True?

The fitted slope, 1.57, reads as euro an hour per extra year of schooling.

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?

Everything Else Must Average Zero at Each Schooling Level

The line is the conditional expectation when everything else averages zero, and ability can break that condition:

- **The Condition:** $E(u \mid x) = 0$ at every level of schooling.
- **The Story Where It Fails:** higher schooling comes with higher ability, so $E(u \mid x)$ rises.

> **R-Squared Cannot Reveal Missing Ability**
> A high R-squared would reveal neither missing ability nor a bend in the true relationship.

Summary: Line, Least Squares, Residuals and Fit

1. **The Line as a Conditional Expectation.**
   A line is the simplest conditional expectation, and it is the CEF when $E(u \mid x) = 0$.

2. **Least Squares.**
   Covariance over variance gave 1.57 euro an hour per year, against a population 1.60.

3. **Fitted Values, Residuals and Sums of Squares.**
   The residuals sum to zero on any sample, whatever is wrong with the line.

4. **Goodness of Fit.**
   R-squared is 0.2045 here, the squared correlation of schooling and earnings.

Before Lecture 2.2

1. **Play:**
   Fit a line by eye in the app, watch the SSR, then reveal the least-squares line.

2. **Apply:**
   Problem set 2 repeats today's derivation with $n = 6$, by hand.

3. **Review:**
   Wooldridge (2020), sections 2-1 to 2-3 on today's model, then 2-4 before Lecture 2.2.

Next Lecture

Lecture 2.2 reads the slope in euro and asks what R-squared cannot report:

1. **Reading the Slope.**
   What 1.57 euro an hour per year says, and how a few high earners change it.
2. **What R-Squared Will Not Tell You.**
   A share of variation reports neither the line's shape nor the slope's size.
3. **Fit on a Hold-Out Sample.**
   Why R-squared on the fitting rows usually exceeds R-squared on new rows.

A line can explain little of the variation and still give an unbiased estimate of β_1. Earnings vary widely at every level of schooling, and no line removes that scatter.

Thank you

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

BACKUP

Appendix

Notation Used in This Lecture, Part 1

Symbol	What it is	First met
y_i	hourly earnings of person i, in euro	Lecture 1.1
i	which person, running from 1 to n	Lecture 1.1
x_i	years of schooling completed by person i	Lecture 1.1
β_0	expected earnings at zero years of schooling	The Simple Regression Model and Its Notation
β_1	the change in expected earnings for one more year	The Simple Regression Model and Its Notation
u_i	every factor in person i's earnings other than schooling	The Simple Regression Model and Its Notation

Notation Used in This Lecture, Part 1

Symbol	**What it is**	**First met**
x_0	the reference level of schooling; ours is thirteen years	Centring Schooling at a Reference Level
$(x_i - x_0)$	the centred regressor	Centring Schooling at a Reference Level
$E(y \mid x)$	the conditional expectation of earnings given schooling	Lecture 1.2
x	a level of schooling, rather than one person's x_i	Lecture 1.2
$E(u \mid x)$	the average of everything else among people with that schooling	Taking the Expectation of the Model at Fixed Schooling

Notation Used in This Lecture, Part 2

Symbol	What it is	First met
$\hat{\mu}_y$	another name for $\bar{y}$: the estimate of μ_y from our 120	Lecture 1.3
$\hat{\beta}_1$	the slope computed from our 120	The Hat on the Slope Marks a Sample Estimate
SSR	the total of the squared residuals	Least Squares Minimises the Total of Squared Gaps
$\sum_{i=1}^{n}$	the sum over the n people in the sample	Lecture 1.1
n	how many people are in the sample, 120 here	Lecture 1.1
b_0	a candidate intercept, not yet the chosen one	Least Squares Minimises the Total of Squared Gaps
b_1	a candidate slope, not yet the chosen one	Least Squares Minimises the Total of Squared Gaps
$\hat{\beta}_0$	the least-squares intercept from our 120	Deriving the Slope: Two Conditions
$\bar{y}$	the average earnings of our 120, 22.09	Lecture 1.1
$\bar{x}$	the average schooling of our 120, 13.93	Lecture 1.1

Notation Used in This Lecture, Part 3

Symbol	What it is	First met
s_{xy}	the sample covariance of schooling and earnings, 691.9 over $n-1$	Lecture 1.1
s_x^2	the sample variance of schooling, 439.5 over $n-1$	Lecture 1.1
$\hat{y}_i$	the fitted value for person i	Splitting One Person's Earnings Into Two Parts
$\hat{u}_i$	the residual for person i	Splitting One Person's Earnings Into Two Parts
SST	the total variation in earnings	The Three Sums of Squares and What Each Measures

Notation Used in This Lecture, Part 3

Symbol	What it is	First met
SSE	the variation in the fitted values	The Three Sums of Squares and What Each Measures
R^2	the share of the sample variation in earnings the line accounts for	Defining R-Squared From the Three Totals
s_x	the standard deviation of schooling in the sample	Lecture 1.1
s_y	the standard deviation of earnings in the sample	Lecture 1.1
r_{xy}	the correlation of schooling and earnings in our 120, 0.4522	Lecture 1.1

Worked Example: the Slope on Our 120 People

Both sums were computed on the 120 people in front of us:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{691.89}{439.47} = 1.5744 \tag{18}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1(\bar{x} - 13) = 22.0936 - 1.5744\,(13.9333 - 13) = 20.6242$$

Back to the slide

Worked Example: R-Squared on Our 120 People

Both routes give the same share from the bar chart's three totals:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{1089.3}{5327.6} = 0.2045$$
$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{4238.3}{5327.6} = 0.2045 \tag{19}$$

What 0.2045 Says About This Sample

Schooling accounts for 20.45 per cent of the variation in earnings here. The other 79.55 per cent is variation the line does not reproduce. That share describes this sample and the spread of earnings in it. R-squared says nothing about how far 1.5744 sits from the population slope.

Back to the slide

Derivation: the Two Derivatives

The body frames start from the two conditions, and this frame derives them. Differentiate the SSR with respect to each unknown in turn:

$$\frac{\partial}{\partial b_0} \underbrace{\sum_{i=1}^{n} \left(y_i - b_0 - b_1(x_i - x_0)\right)^2}_{\text{SSR}} = -2 \sum_{i=1}^{n} \left(y_i - b_0 - b_1(x_i - x_0)\right)$$
$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} \left(y_i - b_0 - b_1(x_i - x_0)\right)^2 = -2 \sum_{i=1}^{n} (x_i - x_0)\left(y_i - b_0 - b_1(x_i - x_0)\right) \tag{20}$$

Setting Both to Zero Gives the Two Conditions

The -2 divides out, leaving the two sums the body frames use. The SSR is a sum of squares in b_0 and b_1. The stationary point is therefore a minimum, not a maximum.

Derivation: R-Squared Is the Squared Correlation

The body frame asserts $R^2 = r_{xy}^2$. Start from the fitted deviation derived in the body, $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$, so the explained total is the slope squared times the schooling total:

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{xy}^2 \, \frac{s_y^2}{s_x^2} \cdot \frac{s_x^2}{s_y^2} = r_{xy}^2 \qquad (21)$$

Dividing Both Totals by N Minus One Turns Them into Variances

Substituting $\hat{\beta}_1 = r_{xy} s_y / s_x$ from p.26 gives the middle expression, and the two variance ratios cancel (Wooldridge, 2020). That cancellation is why R-squared carries no information beyond the correlation when there is one regressor.

Notes and Tools

The app and one line of R reproduce this lecture's line and its R-squared:

- **The App:** the line-fitting app, and the script that draws this deck's exhibits.
- **The Lecture in One Line:** `summary(lm(y ~ I(x - 13)))`.
- **This Lecture's Lab:** `lm()`, `coef()`, `fitted()`, `resid()`, and R-squared by hand.

References

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

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