LECTURE 2.2

Units, Functional Form and Fit

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
Lecture 2.1 fitted the line and measured its fit. This lecture reads the slope in euro and asks what R-squared cannot report. Lecture 2.3 asks how far the slope typically sits from the truth.

Last Week's Lecture

Lecture 2.1 turned the scatter into one line and one number:

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.

Lecture Plan

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.

SECTION 1

Reading the Slope

The Slope Is in Euro an Hour per Year

The slope divides earnings by schooling, so its unit is euro an hour per year (Wooldridge, 2020).

> **Reading the Estimate Out Loud**
> Two people in this sample differ by one year of schooling. On average the one with more schooling earns **1.57 euro more per hour**. At thirteen years of schooling the line predicts about **20.62 euro an hour**.

Our Estimate Differs from the Population Slope

Both estimates differ from the population values because our 120 people are one draw:

- **Slope:** we estimated 1.57 against a population 1.60.
- **Intercept:** we estimated 20.62 against a population 20.00.

A different 120 would give different numbers, which Lecture 1.3 called sampling variation.

> **A Fitted Line Is Not an Effect**
> Nothing here says a year more school would raise a person's earnings. Simple regression analyses correlation, so causal readings need care (Wooldridge, 2020).

How a Few High Earners Change the Fitted Slope

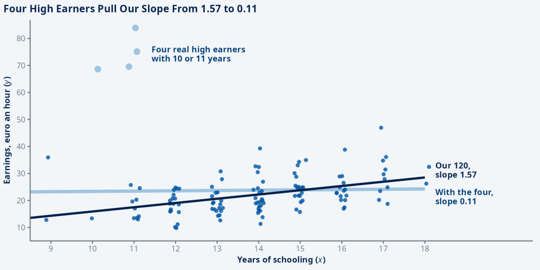

Figure 1: Blue: our 120 and their line. Pale: the same 120 plus four real high earners.

> **1.57 Without Them, 0.11 with Them**
> Squared residuals let a few people far from the line outweigh many near it.

Whether to Drop the Highest Earners

Four very high earners changed the fitted slope from 1.57 to 0.11. One response is to remove them and refit.

> **Question**
> Should those earners be dropped from the sample? What would you need to know before answering?

Whether to Drop Them Depends on Who They Are

The decision turns on what the high earners are:

- **Recording Errors:** fix or drop the rows, and say so.
- **Outside the Population:** directors in an employee survey belong to a different question.
- **Genuine High Earners:** keep them, or the estimate loses the population's top tail.

> **A Rule Chosen After Seeing the Data**
> A trimmed mean fixed before the draw is a method. Dropping whoever changed the slope is a choice made after the data. Report that choice.

SECTION 2

What R-Squared Will Not Tell You

How Small Is a Small Sum of Squared Residuals?

Least squares chose the line with the smallest sum of squared residuals: 4,238.3 on our 120.

> **Question**
> Suppose I tell you a regression's sum of squared residuals is 4,238.3. Have I told you anything about how well the line fits?

A Residual Total Has Units and No Ceiling

The total alone says nothing about fit, for three reasons:

- **Units:** the total is in euro per hour, squared.
- **Sample Size:** doubling the sample roughly doubles the total, with the fit unchanged.
- **Ceiling:** a sum of squares has no upper limit to compare against.

> **A Share Has No Units and a Ceiling of One**
> Lecture 2.1 compared the total with SST, 5,327.6: $R^2 = 1 - 4{,}238.3/5{,}327.6 = 0.2045$. On its own the residual total only ranks lines.

Adding a Regressor Cannot Lower the R-Squared

Adding height in centimetres gives least squares more candidate lines to search:

- **SSR:** cannot rise, because the schooling-only fit is still a candidate.
- **SST:** is unchanged, because it never involved a regressor.

Height Need Not Affect Earnings

So the R-squared, one minus SSR over SST, cannot fall. The inequality holds whether or not height affects earnings.

An Irrelevant Regressor Still Raises the R-Squared

Height is irrelevant in the population but varies with earnings by chance in our 120.

So in practice the R-squared rises rather than staying still: on our 120 it goes from 0.2045 to 0.2106.

> **R-Squared Cannot Compare Specifications**
> A criterion that never falls ranks the largest model first, every time. The F-test of Lecture 3.2 tests between models. Lecture 3.3's adjusted R-squared charges for each variable.

A Straight Line Through a Curved Relationship

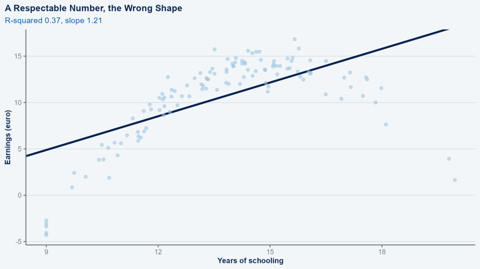

Figure 2: A constructed example, not our sample: earnings rise and then fall.

> **The R-Squared Does Not Report the Curve**
> The R-squared is 0.37 around the wrong shape. SSR adds the squared residuals without asking where in schooling they fall.

A Low R-Squared Does Not Mean a Small Slope

Figure 3: A constructed example, not our sample: a wide scatter around a clearly rising line.

> **Scatter Raises SST and Leaves SSE Alone**
> The share falls however steeply the line rises. Wooldridge's CEO salary regression returns 0.0132 (Wooldridge, 2020).

Errors That Widen with Schooling Lower the R-Squared

Hold the population slope (β_1) at 1.60, and widen the errors with schooling around zero:

- **Errors of One Width:** our 120 give a slope of 1.5744 and an R^2 of 0.2045.
- **Errors That Widen:** the same 120 give 1.7809 and 0.1926.
- **The Whole Population:** the slope stays 1.60 while R^2 falls from 0.1980 to 0.1543.

> **Only the Slope Is Measured in Euro**
> The slope counts extra euro an hour per year of schooling. R-squared is a share with no units, so it cannot say whether a euro figure matters. Wooldridge warns that beginners weight R-squared too heavily (Wooldridge, 2020).

A High R-Squared Does Not Mean a Large Slope

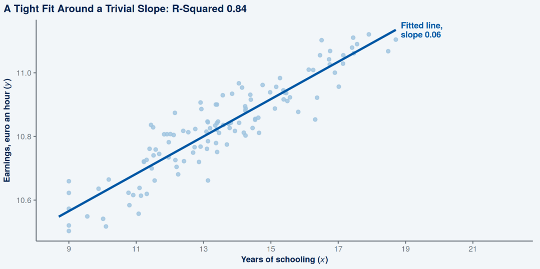

Figure 4: A constructed example, not our sample: R-squared 0.84 with a slope of 0.06.

> **The Slope Depends on Both Standard Deviations**
> R-squared fixes only the size of r_{xy}, and $\hat{\beta}_1 = r_{xy}s_y/s_x$. Any R-squared is compatible with any slope, in either direction.

SECTION 3

Fit on a Hold-Out Sample

Fit Measured Where the Line Was Chosen

R-squared is computed on the very 120 people the coefficients were fitted to:

- **The Choice:** least squares picked the pair with the smallest SSR on those rows.
- **The Consequence:** that pair fits every chance pattern in those rows that a line can.

Fit on a Sample the Line Has Not Seen

The second 120 come from the same population, none of them in our sample, and have not been used. We draw our line through them without refitting it.

> **Question**
> Is the hold-out R-squared above 0.2045, below it, or about the same?

R-Squared on the Hold-Out Sample

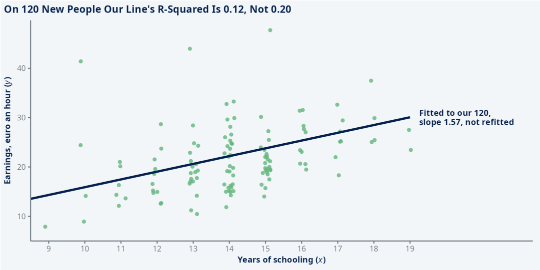

Figure 5: Green: 120 hold-out people. Navy: the line fitted to our 120, not refitted.

> **0.2045 in Sample, 0.1240 on the Hold-Out**
> Only the rows differ: the line, its slope of 1.5744 and the population are unchanged. Report which of the two numbers you quote.

Fresh Rows Usually Give a Lower R-Squared

Fresh rows from the same population carry chance errors our line was never fitted to:

- **SSR:** usually larger on fresh rows, so the R-squared is usually lower.
- **The Drop:** measures how much of our 0.2045 was chance rather than schooling.

> **The Hold-Out Was Set Aside Before Fitting**
> Using a hold-out sample to choose a specification stops that sample being a hold-out. The chosen line has then been suited to those rows too. Its R-squared then overstates the fit on new rows, as the first one does.

What Would Have to Be True?

Read out loud, R-squared says schooling accounts for 20.45 per cent of the variation in earnings.

> **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?

Reading 20.45 per Cent as a Fact About 120 Rows

The 20.45 per cent is arithmetic on 120 rows, so its conditions concern the sample:

- **The Rows:** the 120 must be the people you mean to describe.
- **The Units:** any scale will do, since rescaling earnings rescales SST and SSE together.
- **The Verb:** “accounts for” means a share of squared deviations, not a cause.

> **Ability Could Produce the Same Share**
> Suppose ability causes both schooling and earnings. The share is still 20.45 per cent, but ability put part of that variation there.

Dragging One Person and Watching the R-Squared

Dragging one person in the app changes the slope and the R-squared by different amounts.

> **A Person at Average Schooling Cannot Tilt the Line**
> Raising person i's earnings by one euro changes the slope by $x_i - \bar{x}$ over 439.5, our sum of squared schooling deviations. At average schooling that change is zero.

The lab computes R-squared from the three totals, checks `summary()` and draws the residual plot.

Summary: the Slope, R-Squared and the Hold-Out

1. **Reading the Slope.**
   1.57 euro an hour per year, from one draw of 120 people.

2. **What R-Squared Will Not Tell You.**
   A share of spread reports neither the line's shape nor the slope's size.

3. **Fit on a Hold-Out Sample.**
   0.2045 on our 120 against 0.1240 on rows the line never saw.

Before Lecture 2.3

1. **Play:**
   The app. Drag one person and watch the slope and R-squared change by different amounts.

2. **Apply:**
   Problem set 2, covering 2.1 and 2.2, repeats the sums-of-squares identity with $n = 6$.

3. **Review:**
   Wooldridge (2020), sections 2-3c and 2-4 on R-squared and reading the slope, then 2-5 before Lecture 2.3.

Next Lecture

Lecture 2.3 asks how far the slope typically sits from the true 1.60, in four steps:

1. **The Gauss-Markov Assumptions.**
   Five claims about the population, and the weight each person carries in the slope.
2. **Unbiasedness.**
   Why assumption four makes the fitted slope centre on the truth.
3. **Precision.**
   How far the slope varies across samples, and the three things that narrow it.
4. **Consistency and Asymptotic Efficiency.**
   The slope concentrates on the truth, and least squares stays best in a wider class in the limit.

Thank you

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

BACKUP

Appendix

Notation Used in This Lecture (1 Of 2)

Symbol	What it is	First met
i	which person, running from 1 to n	Lecture 1.1
n	how many people are in the sample; 120 here	Lecture 1.1
x_i	years of schooling completed by person i	Lecture 1.1
y_i	what person i earns per hour	Lecture 1.1
$\bar{x}$	the average schooling of our 120, 13.9333	Lecture 1.1
$\bar{y}$	the average earnings of our 120, 22.0936	Lecture 1.1
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

Notation Used in This Lecture (2 Of 2)

Symbol	What it is	First met
β_1	the population slope; 1.60 here	Lecture 2.1
$\hat{\beta}_1$	the least-squares slope from our 120, 1.5744	Lecture 2.1
b_0	a candidate intercept, not yet the chosen one	Lecture 2.1
b_1	a candidate slope, not yet the chosen one	Lecture 2.1
x_0	the reference level of schooling; ours is thirteen years	Lecture 2.1
$\hat{y}_i$	the fitted value, what the line predicts for person i	Lecture 2.1
SST	how much earnings vary around their average; 5,327.6	Lecture 2.1
SSE	how much the fitted values vary around the same average; 1,089.3	Lecture 2.1
SSR	the sum of squared residuals; 4,238.3	Lecture 2.1
R^2	the share of the sample variation in earnings the line accounts for; 0.2045	Lecture 2.1

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:

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

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.

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:

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

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.

Notes and Tools

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

References

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