Econometrics

Simple Linear Regression

Mustapha Douch based on
Florian Oswald's slides
UniTo ESOMAS
2025-10-22

UNIVERSITÀ
DI TORINO

Department of
Economics, Social Studies,
Applied Mathematics
and Statistics

Where We Are Now: Building the Foundation 🏗️

Covered Concepts (Stock & Watson Chapters 1–3)

Up until now, we've focused on the core tools of **descriptive statistics** and **probability** that form the language of econometrics:

- **Descriptive Statistics:**
  Summarizing data using the *mean*, *median*, *variance*, and *standard deviation*.

- **Probability Theory:**
  Understanding *random variables*, their *distributions* (especially the *normal distribution*), and the concept of *covariance* and *correlation* to measure association.

- **Asymptotics:**
  Grasping the crucial role of the *Law of Large Numbers (LLN)* and the *Central Limit Theorem (CLT)* in ensuring our sample estimates are reliable.

Next Steps: From Description to Causation 🚀

This week, we begin our journey into **the Simple Regression Model** — using data to explain **how one variable affects another**.

We'll build on the descriptive and probabilistic foundations from before to estimate relationships and test hypotheses.

Next week, we move further into **causal inference** with **Difference-in-Differences (DiD)** — comparing before-and-after outcomes to identify policy or treatment effects.

Today - Real 'metrics finally ✌️

- Introduction to the ***Simple Linear Regression Model*** and ***Ordinary Least Squares (OLS)*** *estimation*.

- Empirical application: *class size* and *student performance*

- Keep in mind that we are interested in uncovering **causal** relationships

How Does One Variable Affect Another? 🎯

> A state implements tough new penalties on drunk drivers — what is the effect on highway fatalities?
> A school district cuts the size of its elementary school classes — what is the effect on its students’ standardized test scores?
> You successfully complete one more year of college classes — what is the effect on your future earnings?

All three questions are about the **unknown effect of changing one variable**, (X), (on penalties, class size, or years of schooling) on another variable, (Y), (highway deaths, test scores, or earnings).

This week, we introduce the **linear regression model** relating (X) to (Y). It postulates a **linear relationship** between (X) and (Y): the **slope** represents the effect of a one-unit change in (X) on (Y).

Just as the **mean of (Y)** is an unknown population characteristic, the **slope of the line** relating (X) and (Y) is an unknown feature of their **joint distribution**.

How Does One Variable Affect Another? 🎯

Our econometric task:
Estimate this slope — that is, estimate **the effect of a unit change in (X) on (Y)** — using a random sample of data.

Finally, we'll see how this is done using **Ordinary Least Squares (OLS)**, which allows us to test hypotheses and construct confidence intervals for the slope.

Student performance

- What policies *lead* to improved student learning?
- Class size reduction has been at the heart of policy debates for *decades*.
- We will be using data from a famous paper by Joshua Angrist and Victor Lavy (1999), obtained from Raj Chetty and Greg Bruich's course.
- Consists of test scores and class/school characteristics for fifth graders (10-11 years old) in Jewish public elementary schools in Israel in 1991.
- National tests measured *mathematics* and (Hebrew) *reading* skills. The raw scores were scaled from 1-100.

Task 1: Getting to know the data

07:00

1. Load the data from here as `grades`. *Hint: Use the `read_dta` from the `haven` library to import the file, which has a format .dta.* (FYI: *.dta* is the extension for data files used in *Stata*)

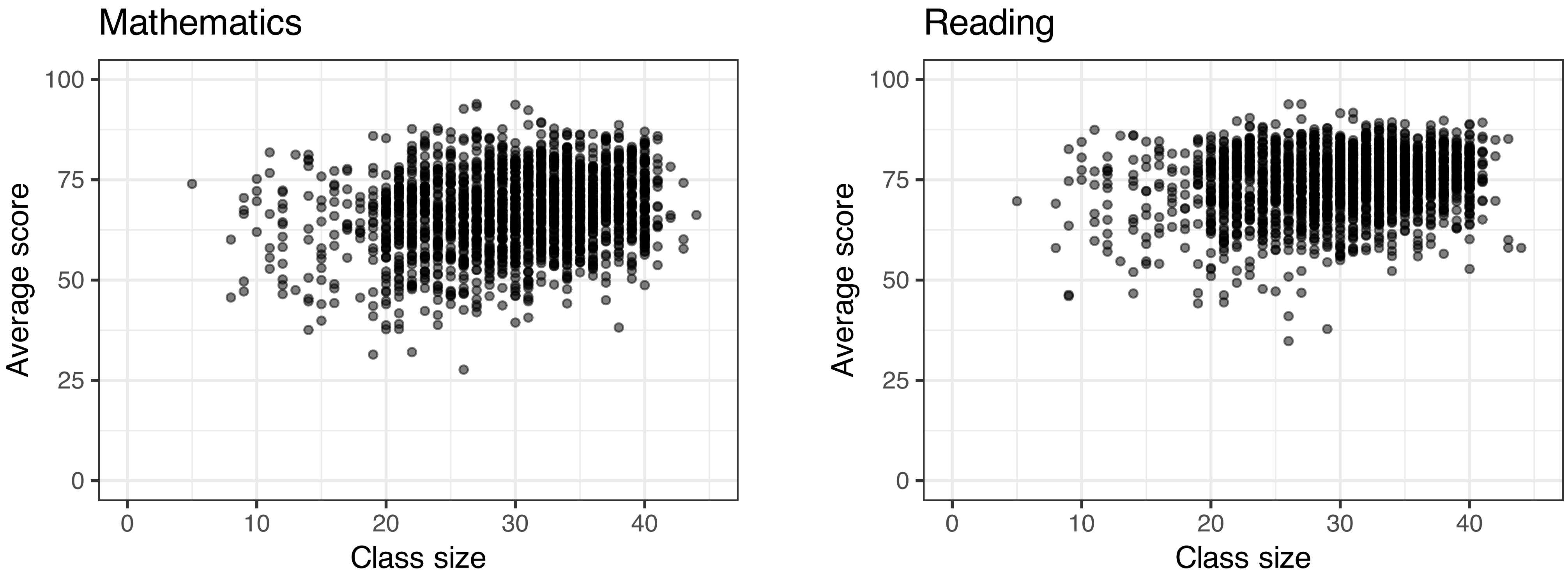
2. Describe the dataset:

    - What is the unit of observations, i.e. what does each row correspond to?
    - How many observations are there?
    - View the dataset. What variables do we have? What do the variables `avgmath` and `avgverb` correspond to?
    - Use the `skim` function from the `skimr` package to obtain common summary statistics for the variables `classize`, `avgmath` and `avgverb`. (*Hint: use `dplyr` to `select` the variables and then simply pipe (`%>%`) `skim()`.*)

3. Do you have any priors about the actual (linear) relationship between class size and student achievement? What would you do to get a first insight?

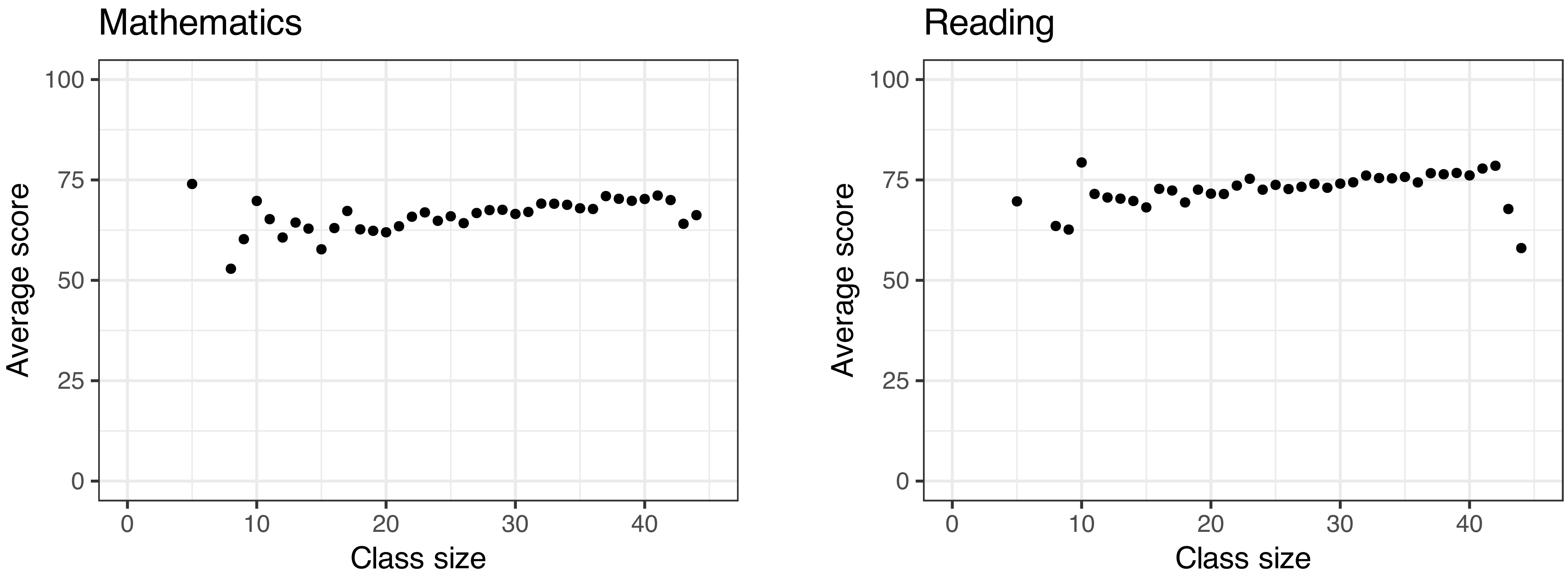
4. Compute the correlation between class size and math and verbal scores. Is the relationship positive/negative, strong/weak?

Class size and student performance: Scatter plot



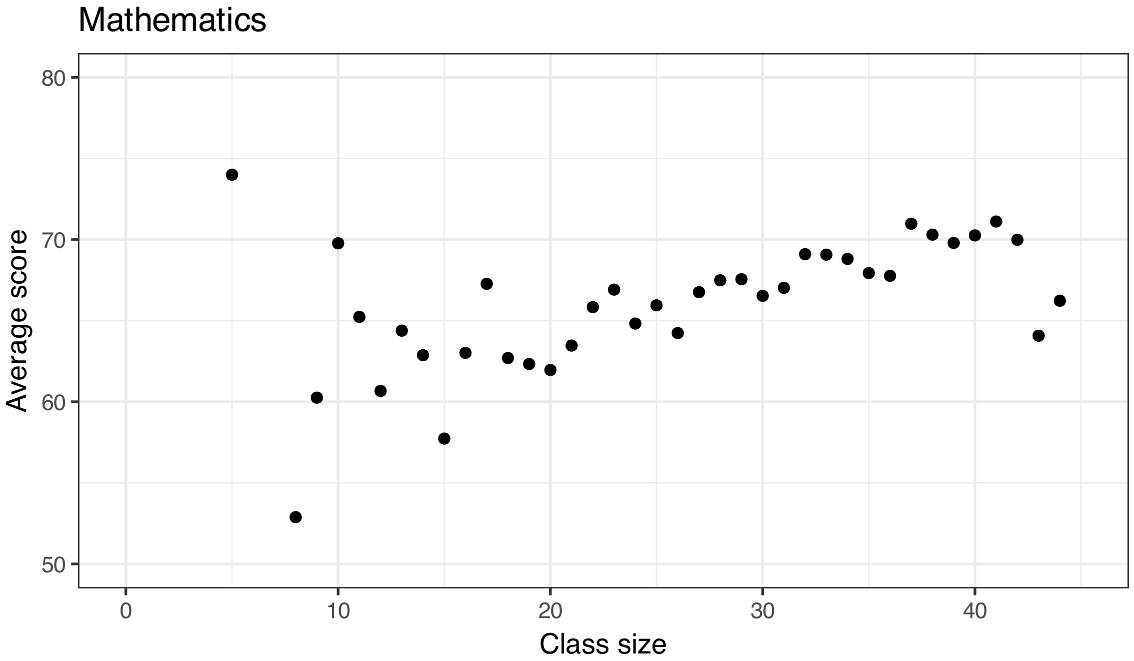
- Somewhat positive association as suggested by the correlations. Let's compute the average score by class size to see things more clearly!

Class size and student performance: Binned scatter plot



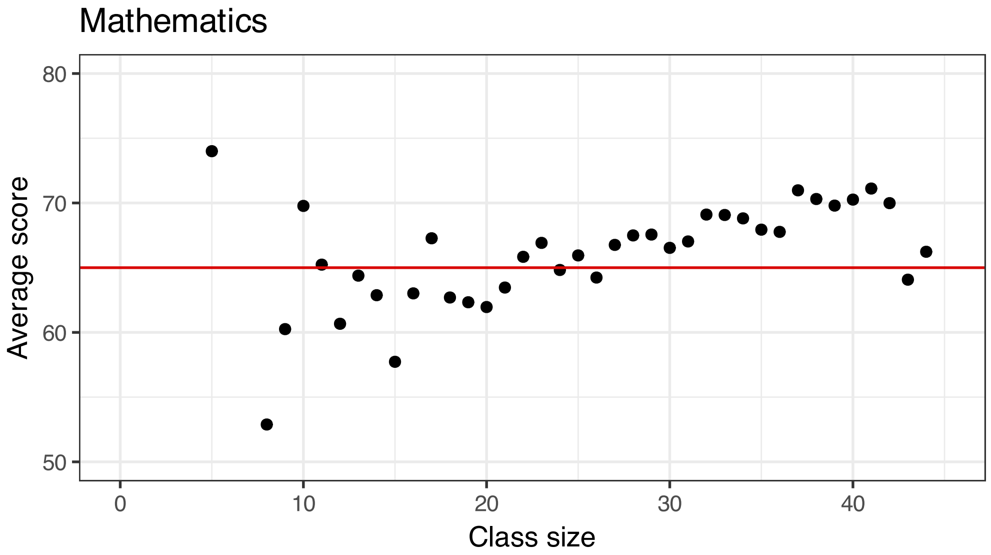
Class size and student performance: Binned scatter plot

- We'll first focus on the mathematics scores and for visual simplicity we'll zoom in



Class size and student performance: Regression line

How to visually summarize the relationship: **a line through the scatter plot**

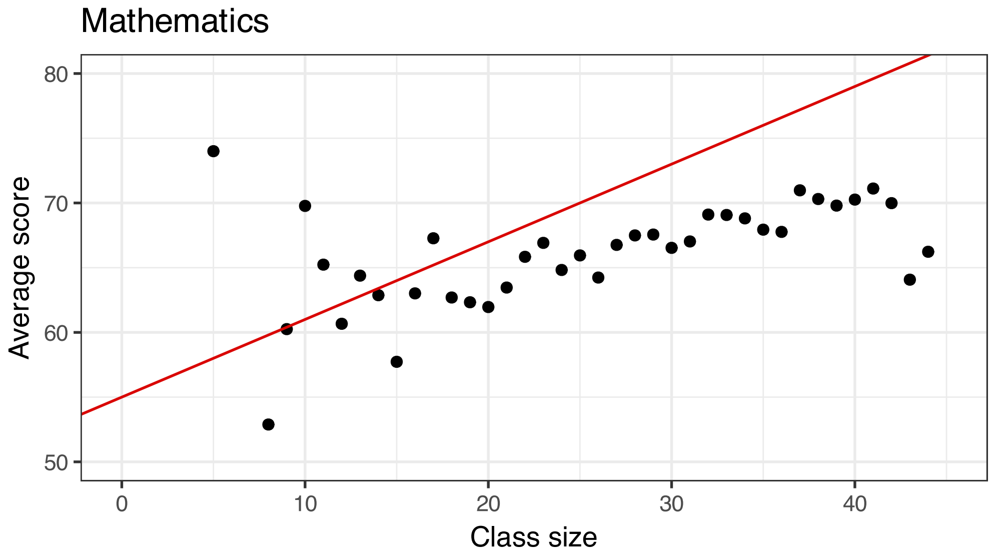


- A *line*! Great. But **which** line? This one?

- That's a *flat* line. But average mathematics score is somewhat *increasing* with class size

Class size and student performance: Regression line

How to visually summarize the relationship: **a line through the scatter plot**



- **That** one?
- Slightly better! Has a **slope** and an **intercept** 😐
- We need a rule to decide!

Simple Linear Regression

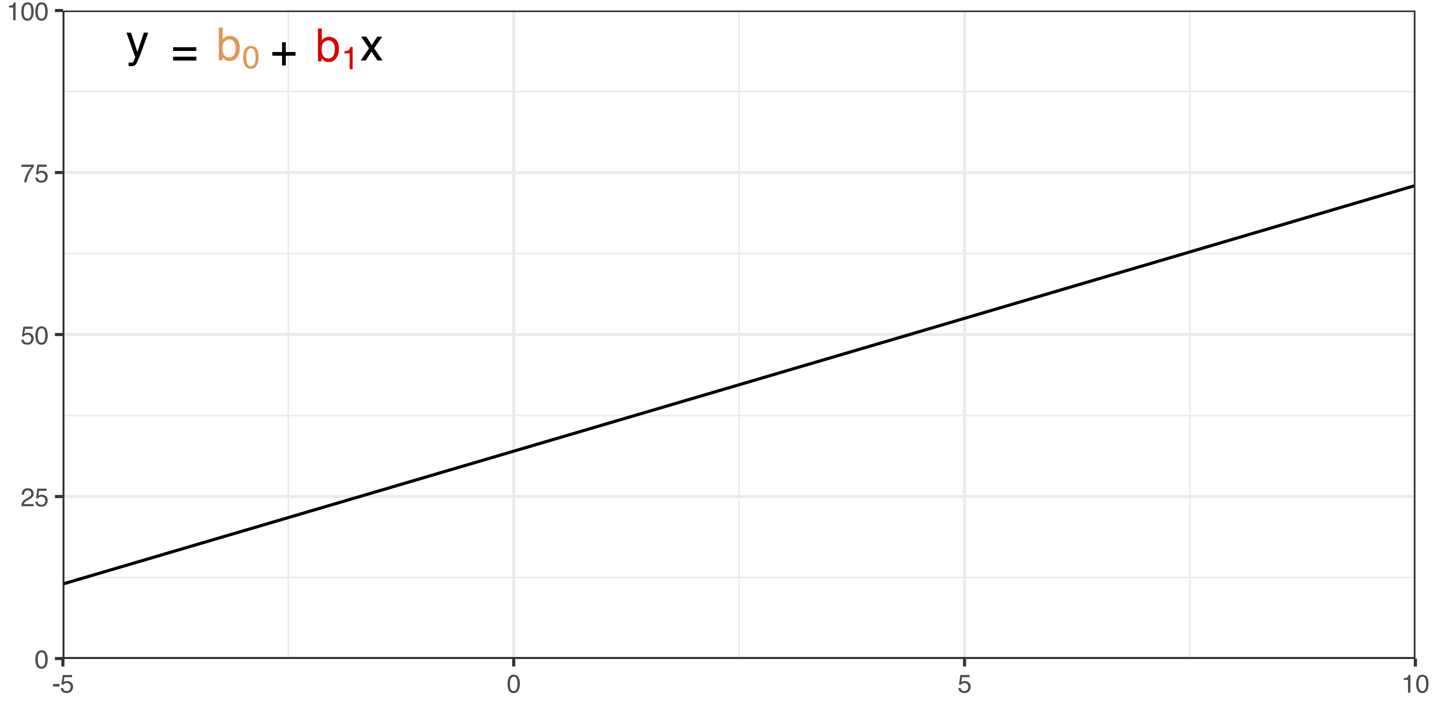
Let's formalise a bit what we are doing so far.

- We are interested in the relationship between two variables:
  - an **outcome variable** (also called **dependent variable**): *average mathematics score* (y)
  - an **explanatory variable** (also called **independent variable** or **regressor**): *class size* (x)
- For each class i we observe both x_i and y_i, and therefore we can plot the *joint distribution* of class size and average mathematics score.
- We summarise this relationship with a line (for now). The equation for such a line with an intercept b_0 and a slope b_1 is:

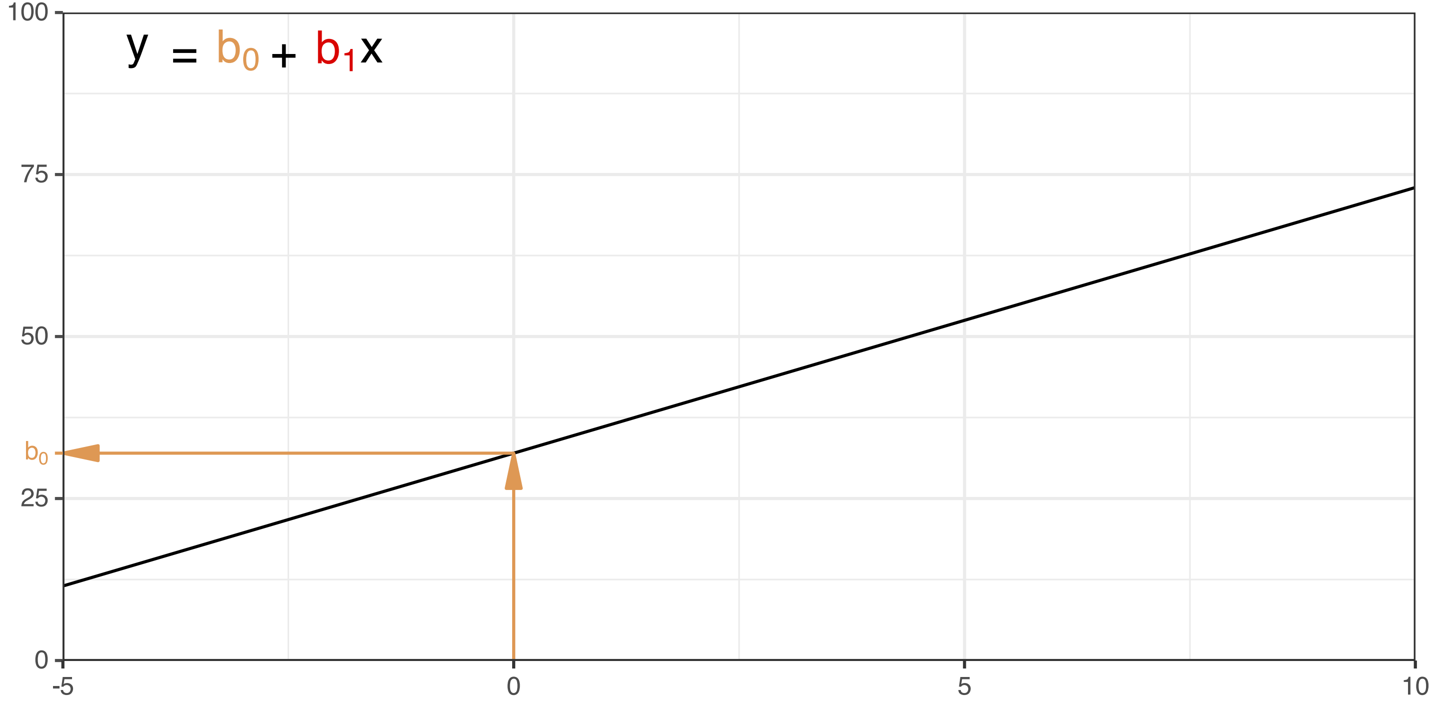
$$\hat{y}_i = b_0 + b_1 x_i$$

- $\hat{y}_i$ is our *prediction* for y at observation i (y_i) given our model (i.e. the line).

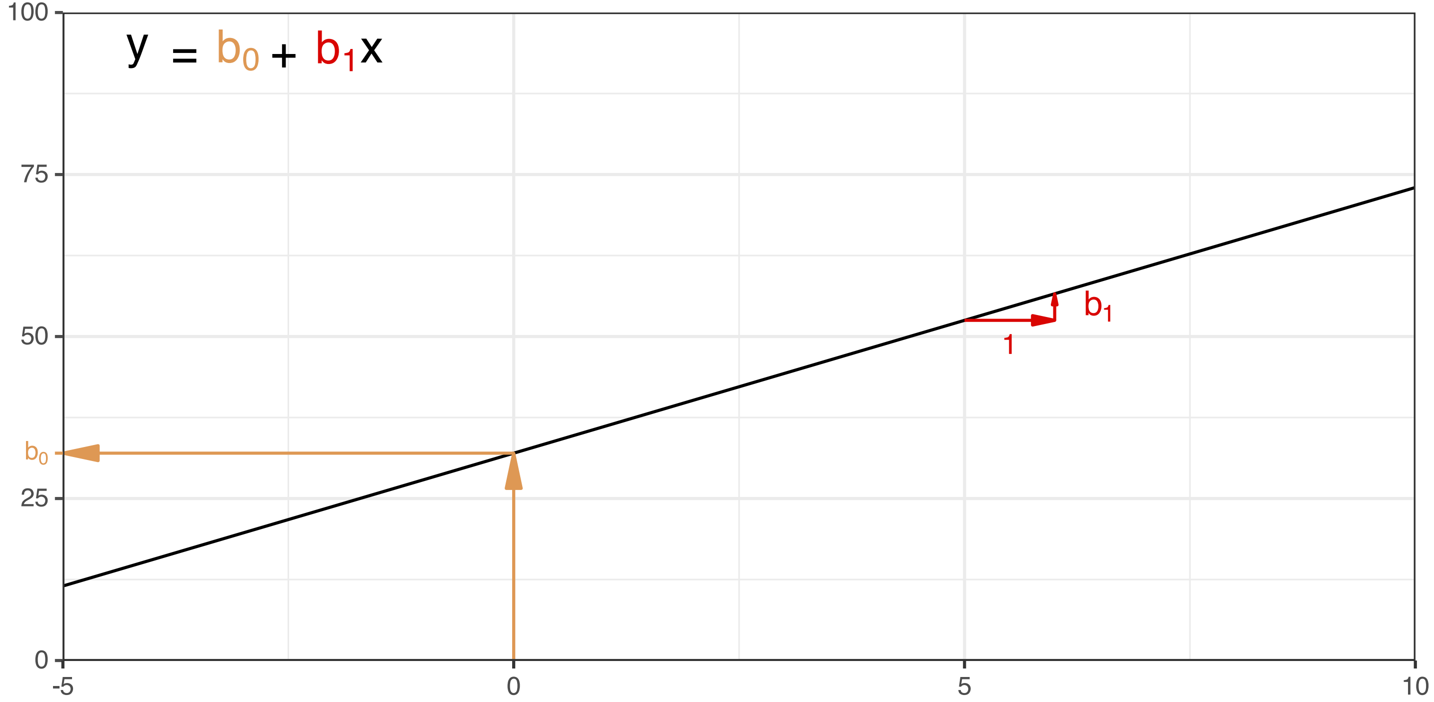
What's A Line: A Refresher



What's A Line: A Refresher

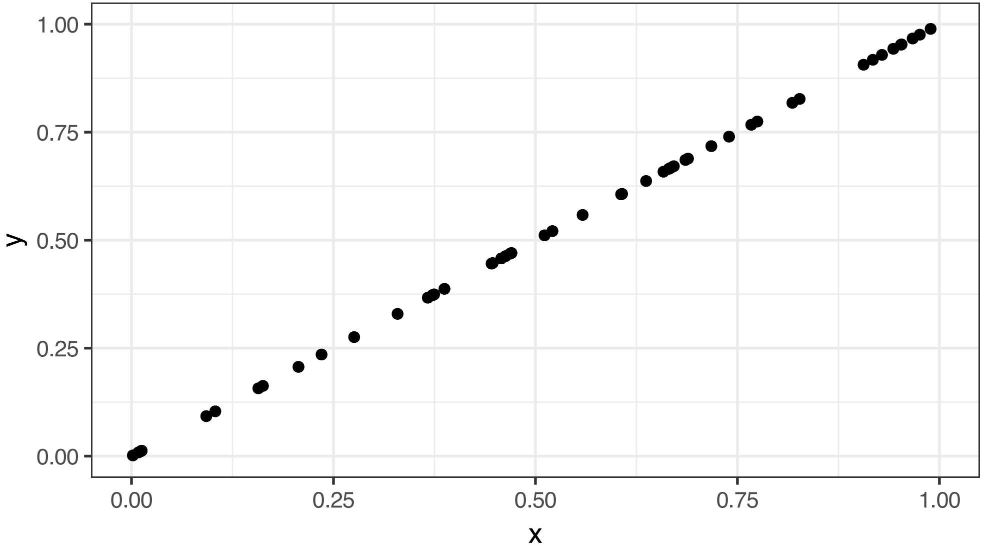


What's A Line: A Refresher



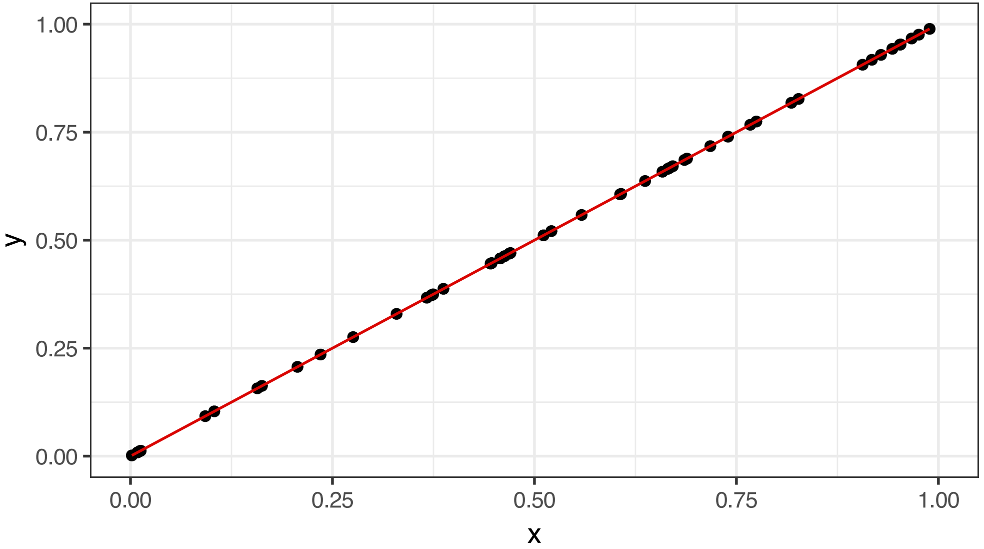
Simple Linear Regression: Residual

- If all the data points were **on** the line then $\hat{y}_i = y_i$.



Simple Linear Regression: Residual

- If all the data points were **on** the line then $\hat{y}_i = y_i$.

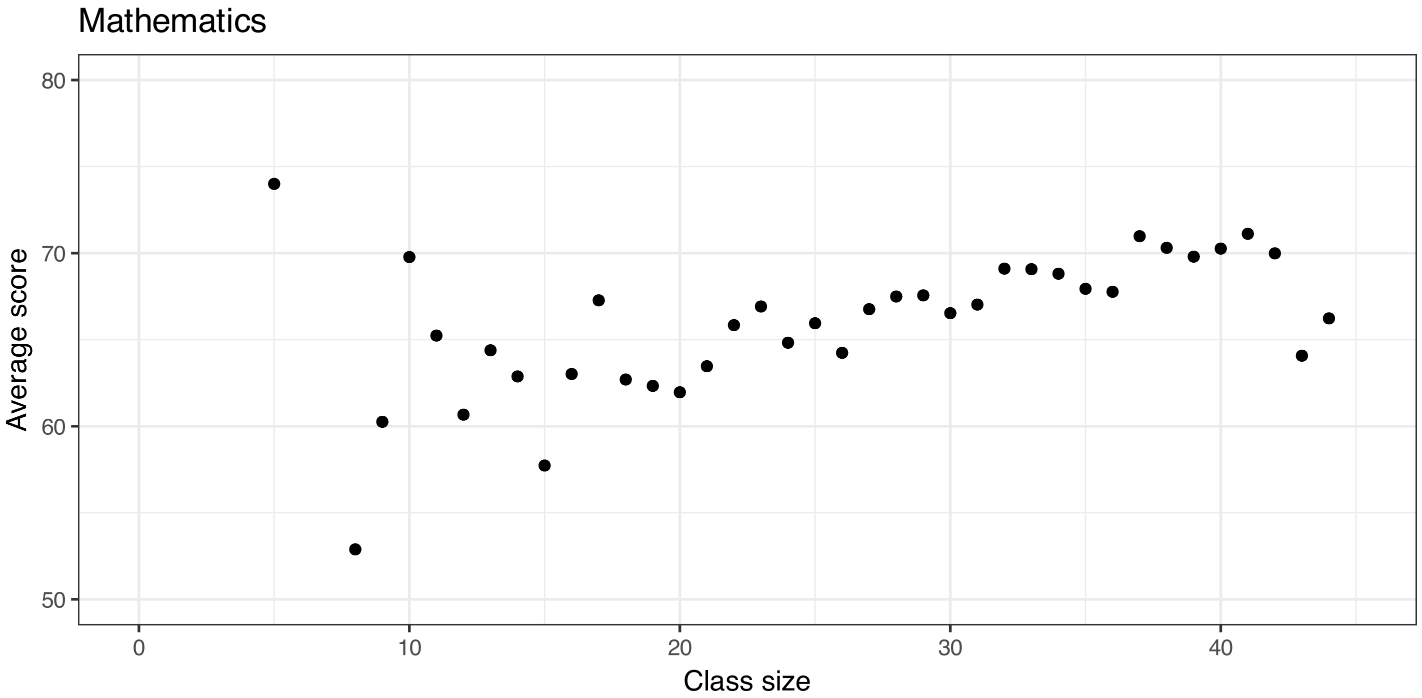


Simple Linear Regression: Residual

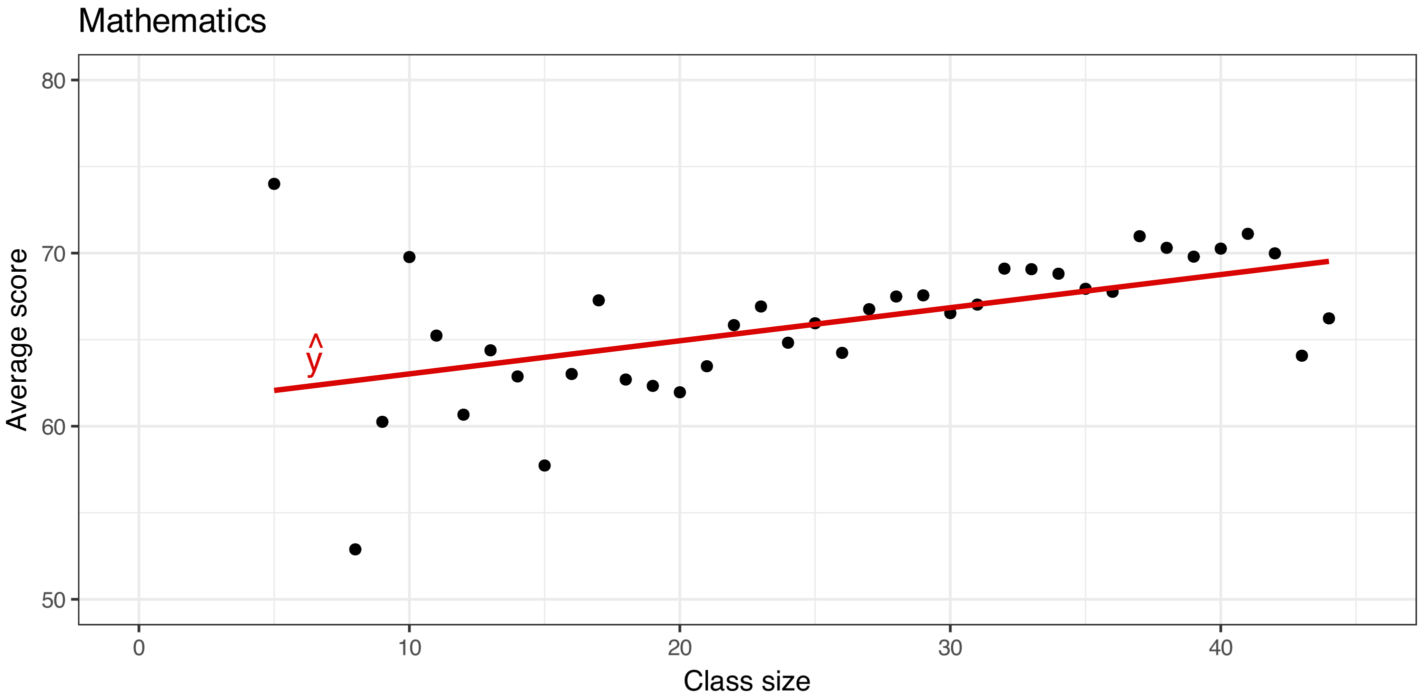
- If all the data points were **on** the line then $\hat{y}_i = y_i$.
- However, since in most cases the *dependent variable* (y) is not *only* explained by the chosen *independent variable* (x), $\hat{y}_i \neq y_i$, i.e. we make an **error**.
  This **error** is called the **residual**.
- At point (x_i, y_i), we note this residual e_i.
- The *actual data* (x_i, y_i) can thus be written as *prediction* + *residual*:

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + e_i$$

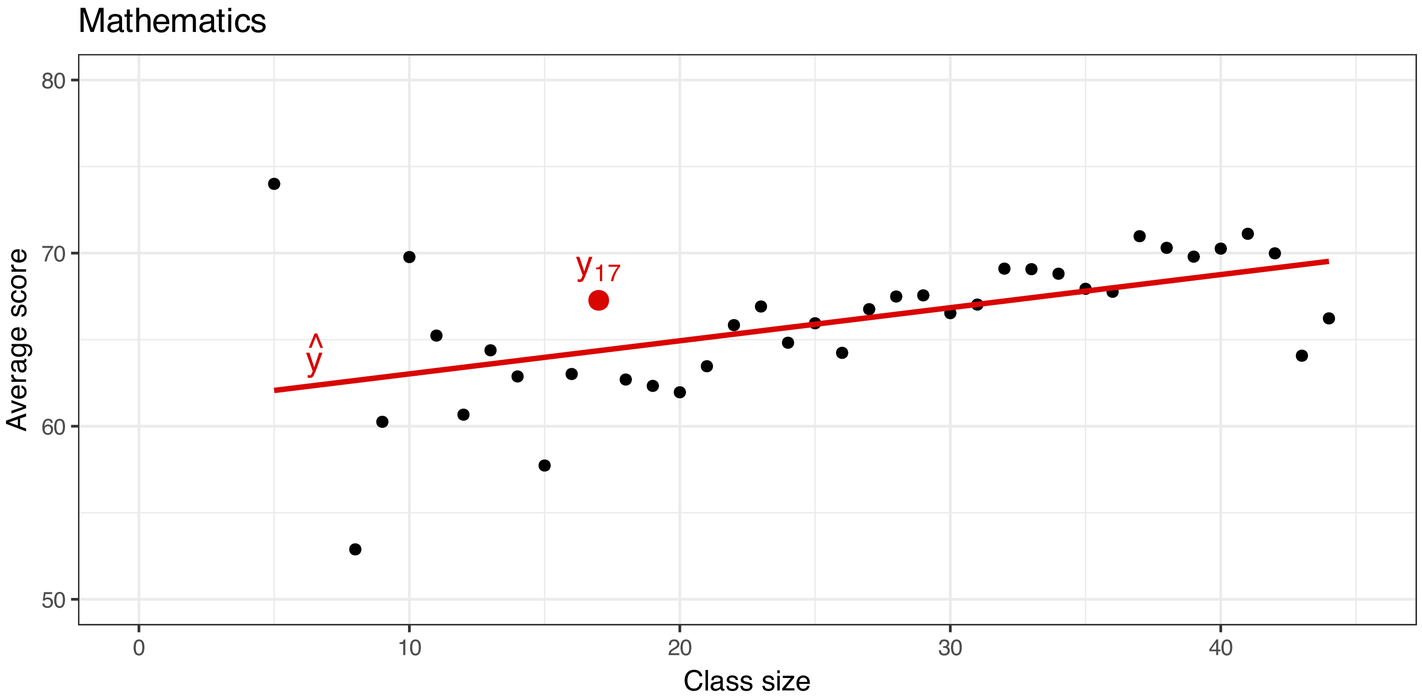
Simple Linear Regression: Graphically



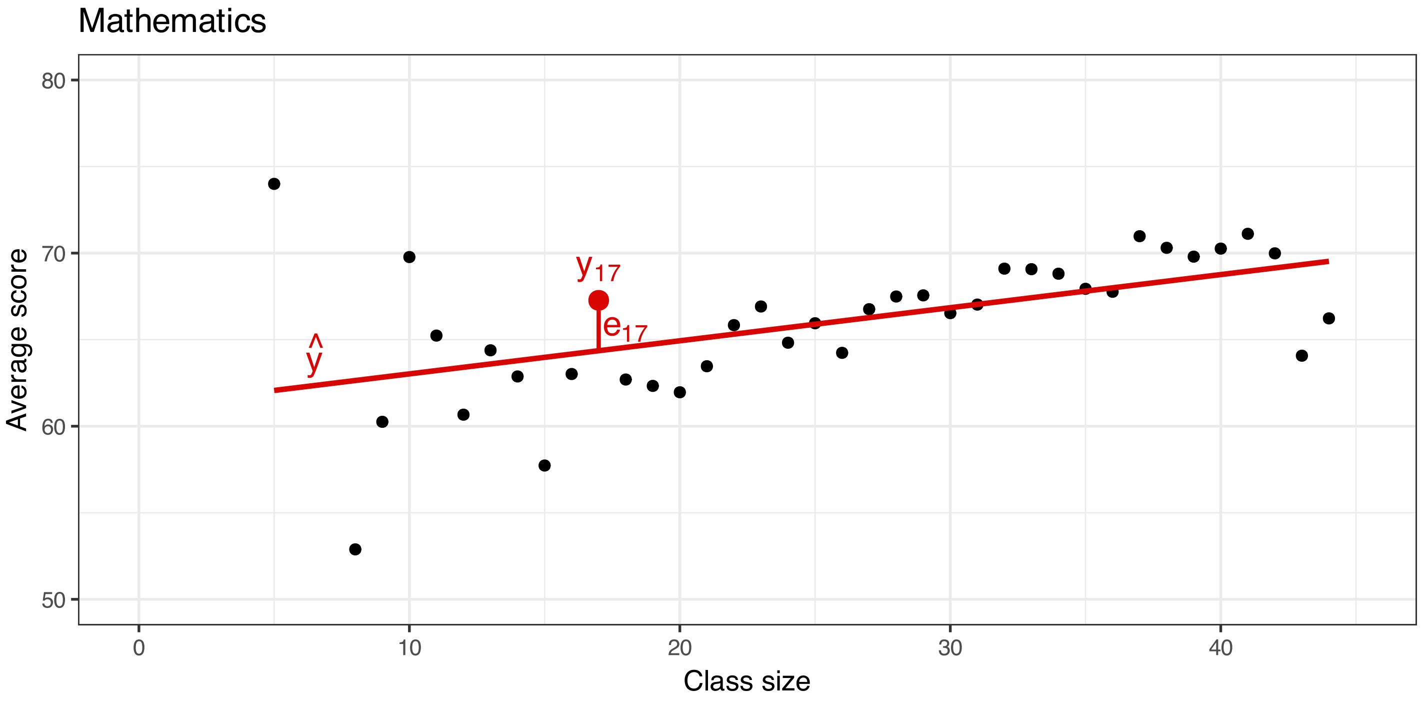
Simple Linear Regression: Graphically



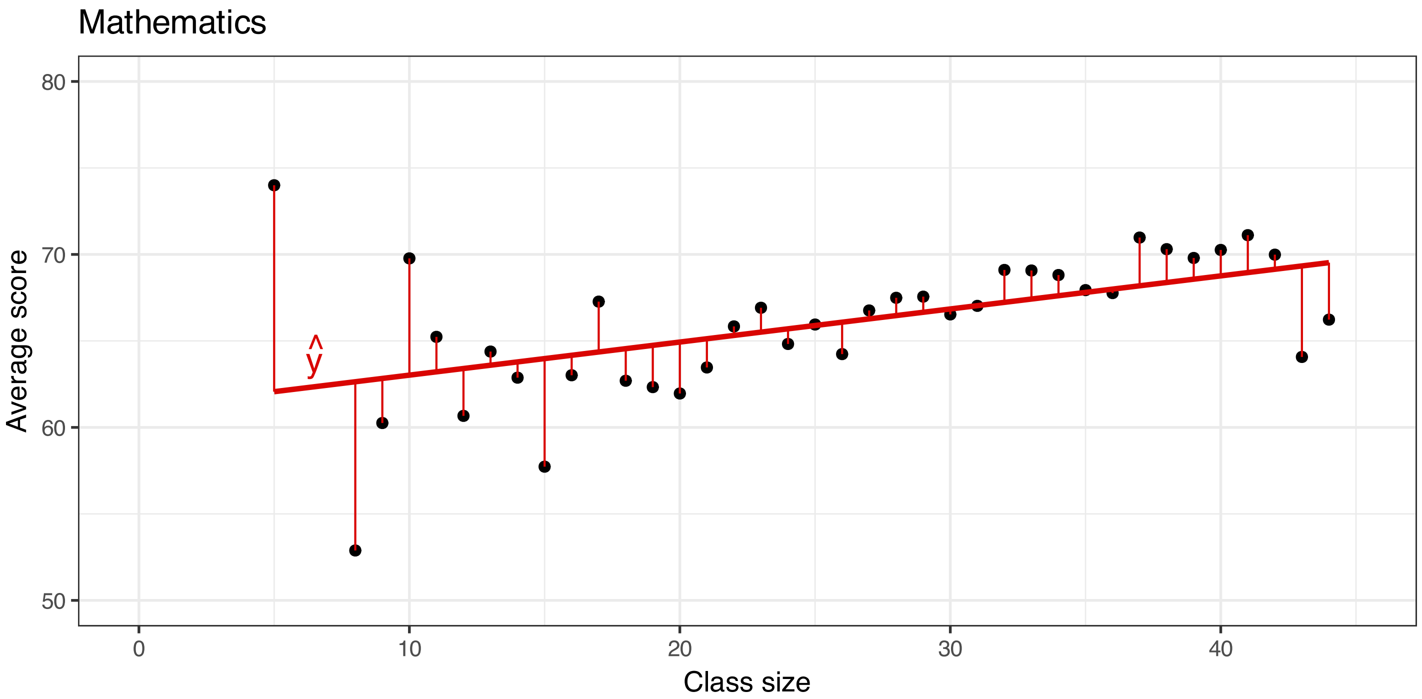
Simple Linear Regression: Graphically



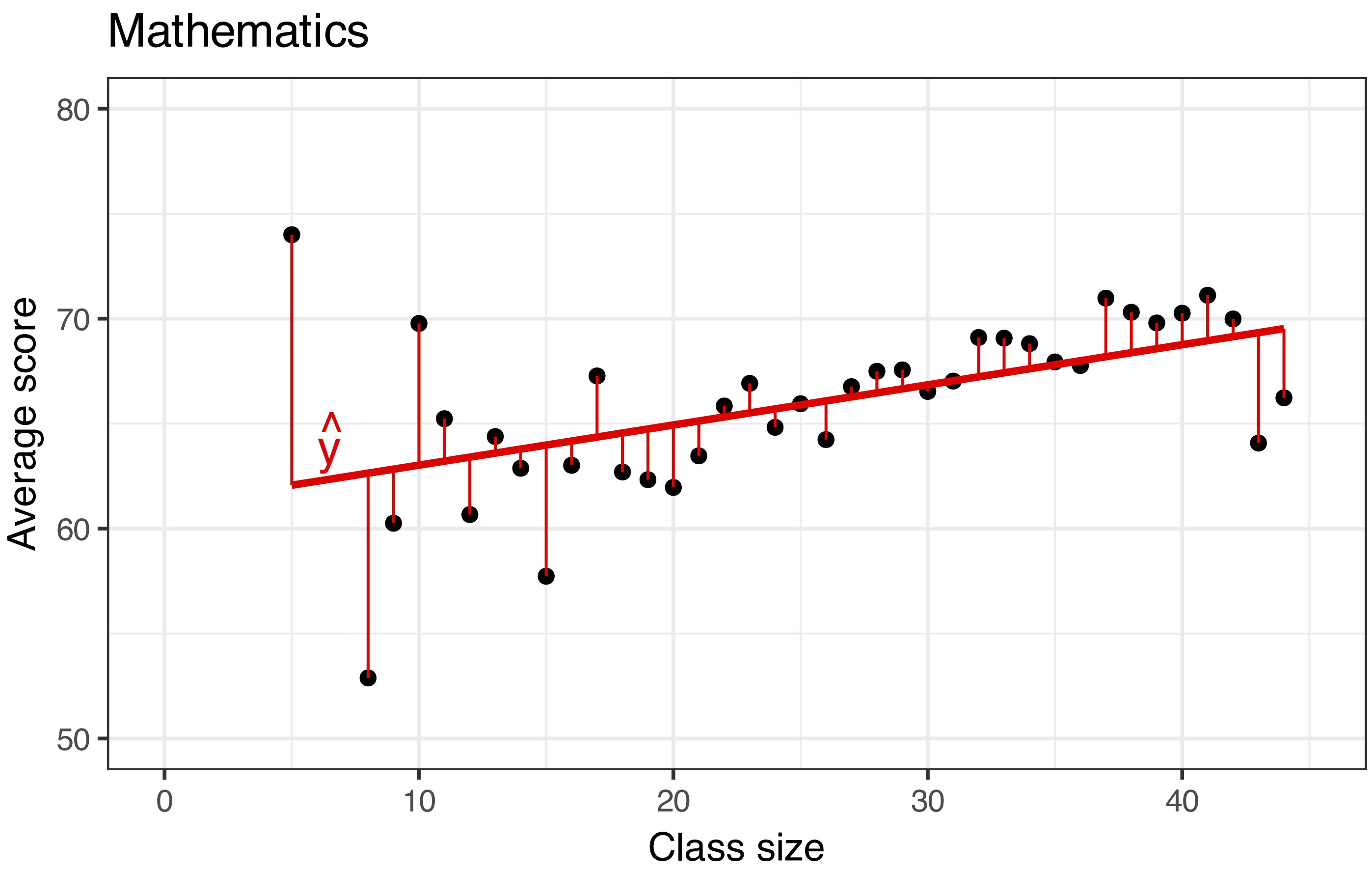
Simple Linear Regression: Graphically



Simple Linear Regression: Graphically



Simple Linear Regression: Graphically



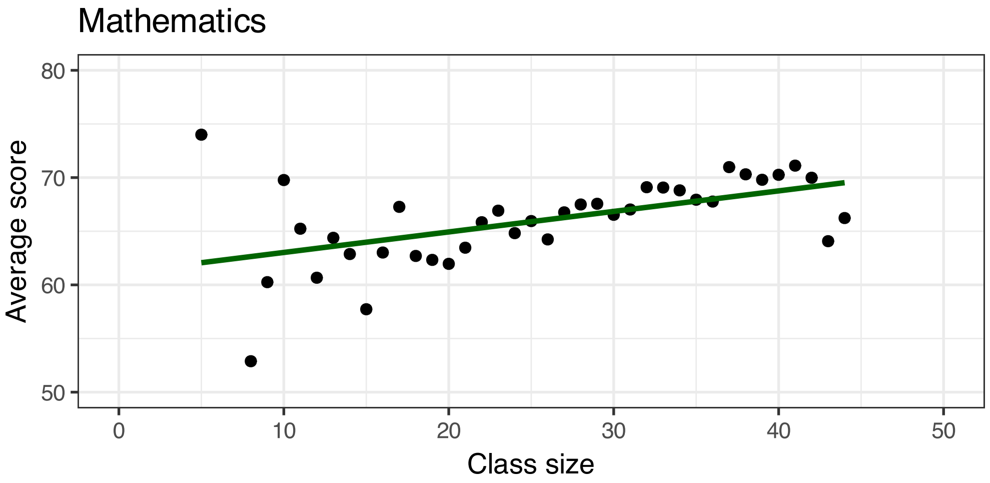
Which "minimisation" criterion should (can) be used?

Ordinary Least Squares (OLS) Estimation

- Errors of different sign $(+/-)$ cancel out, so we consider **squared residuals**

$$\forall i \in [1, N], e_i^2 = (y_i - \hat{y}_i)^2 = (y_i - b_0 - b_1 x_i)^2$$

- Choose (b_0, b_1) such that $\sum_{i=1}^{N} e_1^2 + \cdots + e_N^2$ is **as small as possible**.

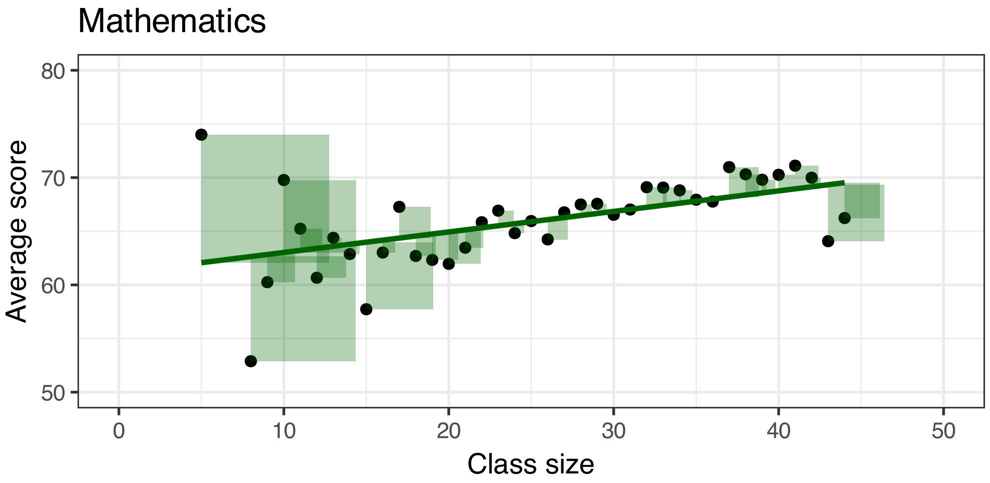


Ordinary Least Squares (OLS) Estimation

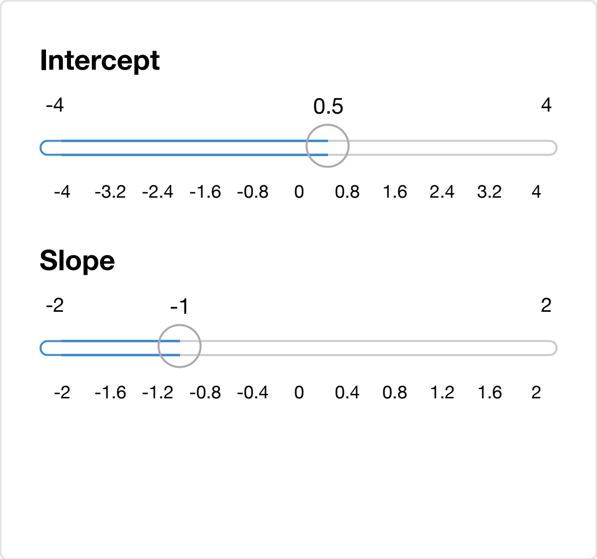
- Errors of different sign $(+/-)$ cancel out, so we consider **squared residuals**

$$\forall i \in [1, N], e_i^2 = (y_i - \hat{y}_i)^2 = (y_i - b_0 - b_1 x_i)^2$$

- Choose (b_0, b_1) such that $\sum_{i=1}^{N} e_1^2 + \cdots + e_N^2$ is **as small as possible**.



Ordinary Least Squares (OLS) Estimation



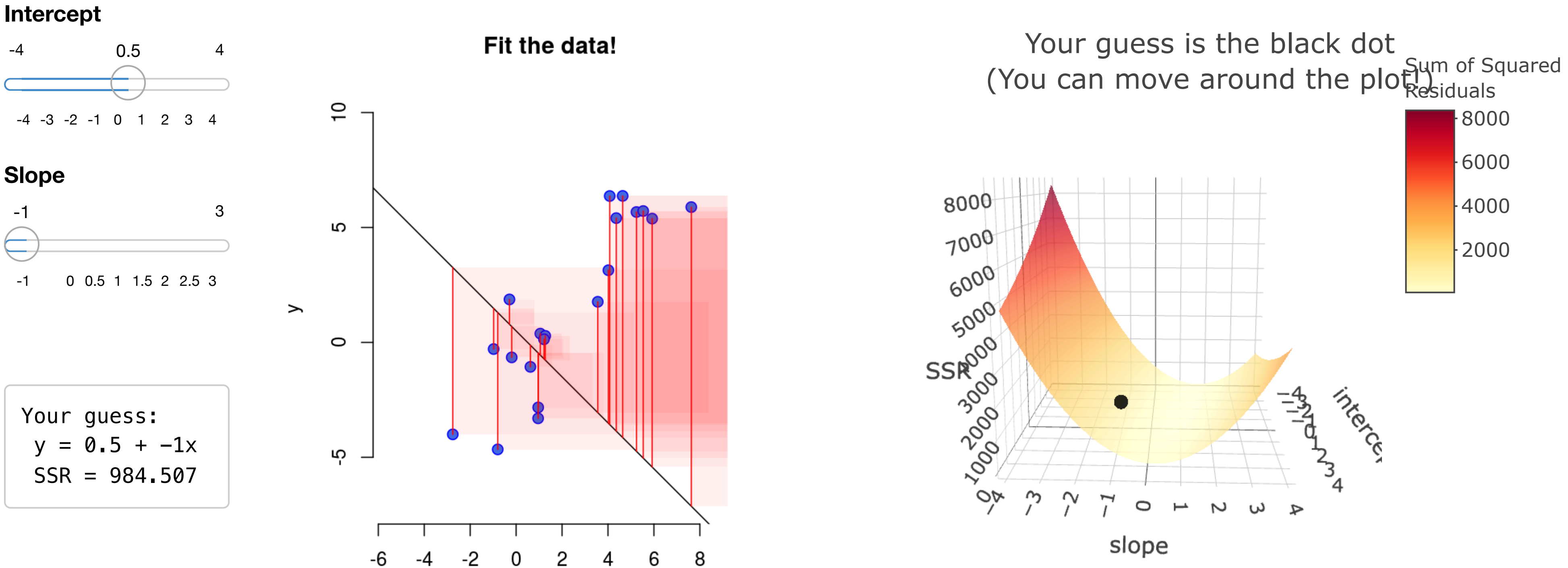
Intercept

0.5

Slope

-1

Ordinary Least Squares (OLS) Estimation



Intercept

0.5

Slope

-1

```
Your guess:
 y = 0.5 + −1x
 SSR = 984.507
```

Ordinary Least Squares (OLS): Coefficient Formulas

- **OLS**: *estimation* method consisting in minimizing the sum of squared residuals.
- Yields **unique** solutions to this minization problem.
- So what are the formulas for b_0 (intercept) and b_1 (slope)?
- In our single independent variable case:

> **Slope:** $b_1^{OLS} = \frac{cov(x,y)}{var(x)}$ **Intercept:** $b_0^{OLS} = \bar{y} - b_1\bar{x}$

- These formulas do not appear from magic. They can be found by solving the minimisation of squared errors. The maths can be found here for those who are interested.

Ordinary Least Squares (OLS): Interpretation

For now assume both the dependent variable (y) and the independent variable (x) are numeric.

> Intercept (b_0): **The predicted value of y $(\hat{y})$ if $x = 0$.**
>
> Slope (b_1): **The predicted change, on average, in the value of y *associated* to a one-unit increase in x.**

- ⚠ Note that we use the term *associated*, **clearly avoiding interpreting b_1 as the causal impact of x on y**. To make such a claim, we need some specific conditions to be met. (Next week!)
- Also notice that the units of x will matter for the interpretation (and magnitude!) of b_1.
- **You need to be explicit about what the unit of x is!**

OLS with `R`

- In `R`, OLS regressions are estimated using the `lm` function.

- This is how it works:

```
lm(formula = dependent variable ~  independent variable, data = data.frame containing the data)
```

OLS with R ## Class size and student performance Let's estimate the following model by OLS: `$\text{average math score}_i = b_0 + b_1 \text{class size}_i + e_i$ r # OLS regression of class size on average maths score lm(avgmath_cs ~ classize, grades_avg_cs) ## ## Call: ## lm(formula = avgmath_cs ~ classize, data = grades_avg_cs) ## ## Coefficients: ## (Intercept) classize ## 61.1092 0.1913 `

Ordinary Least Squares (OLS): Prediction

```
##
## Call:
## lm(formula = avgmath_cs ~ classize, data = grades_avg_cs)
##
## Coefficients:
## (Intercept)     classize
##     61.1092       0.1913
```

This implies (abstracting the $_i$ subscript for simplicity):

$$\hat{y} = b_0 + b_1 x$$

$$\widehat{\text{average math score}} = b_0 + b_1 \cdot \text{class size}$$

$$\widehat{\text{average math score}} = 61.11 + 0.19 \cdot \text{class size}$$

What's the predicted average score for a class of 26 students? (Using the *exact* coefficients.)

$$\widehat{\text{average math score}} = 61.11 + 0.19 \cdot 26$$

$$\widehat{\text{average math score}} = 66.08$$

Task 2: OLS Regression

10:00

Run the following code to aggregate the data at the class size level:

```
grades_avg_cs <- grades %>%
  group_by(classize) %>%
  summarise(avgmath_cs = mean(avgmath),
            avgverb_cs = mean(avgverb))
```

1. Regress average verbal score (dependent variable) on class size (independant variable). Interpret the coefficients.

2. Compute the OLS coefficients b_0 and b_1 of the previous regression using the formulas on slide 25. (*Hint:* you need to use the `cov`, `var`, and `mean` functions.)

3. What is the predicted average verbal score when class size is equal to 0? (Does that even make sense?!)

4. What is the predicted average verbal score when the class size is equal to 30 students?

Predictions and Residuals: Properties

- **The average of $\hat{y}_i$ is equal to $\bar{y}$.**

$$\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i = \frac{1}{N}\sum_{i=1}^{N} b_0 + b_1 x_i$$
$$= b_0 + b_1\bar{x} = \bar{y}$$

- **The average (or sum) of residuals is 0.**

$$\frac{1}{N}\sum_{i=1}^{N} e_i = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$
$$= \bar{y} - \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i$$
$$= 0$$

- **Regressor and residuals are uncorrelated (by definition).**

$$Cov(x_i, e_i) = 0$$

- **Prediction and residuals are uncorrelated.**

$$\begin{aligned} Cov(\hat{y}_i, e_i) &= Cov(b_0 + b_1 x_i, e_i) \\ &= b_1 Cov(x_i, e_i) \\ &= 0 \end{aligned}$$

Since $Cov(a + bx, y) = bCov(x, y)$.

Linearity Assumption: Visualize your Data!

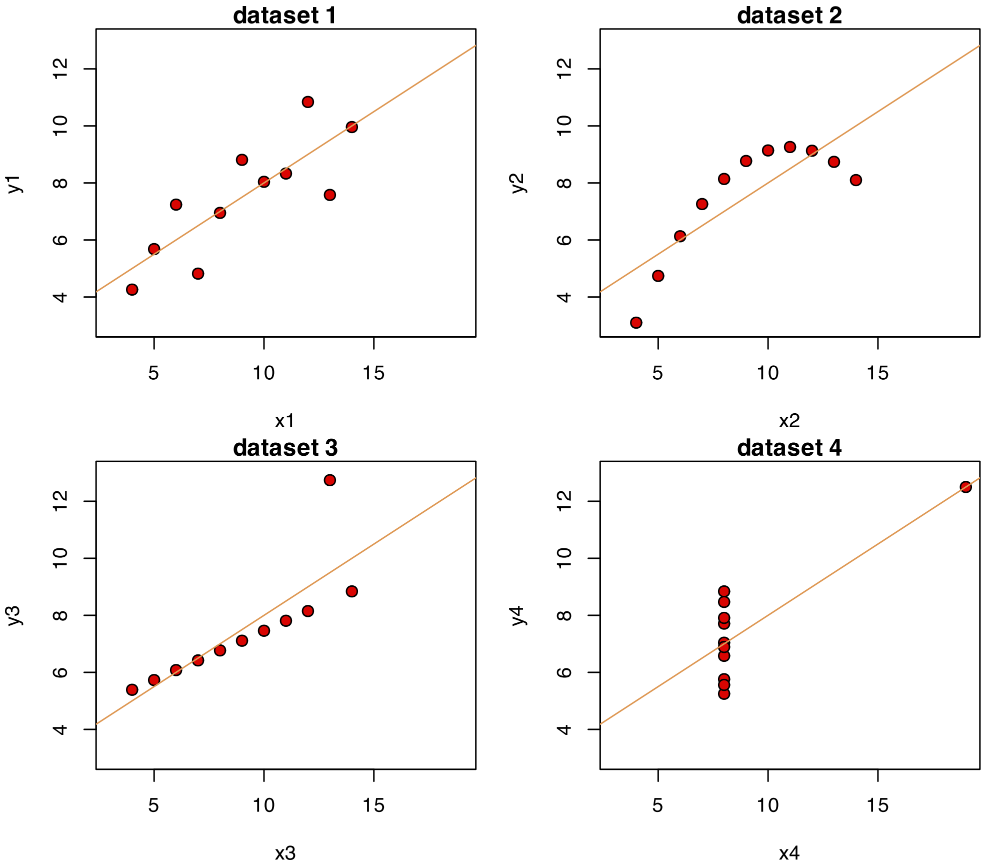
- It's important to keep in mind that covariance, correlation and simple OLS regression only measure **linear relationships** between two variables.

- Two datasets with *identical* correlations and regression lines could look *vastly* different.

- Is that even possible?

Linearity Assumption: Anscombe

- Francis Anscombe (1973) came up with 4 datasets with identical stats. But look!



dataset	cov	var(y)	var(x)
1	5.501	4.127	11
2	5.500	4.128	11
3	5.497	4.123	11
4	5.499	4.123	11

Nonlinear Relationships in Data?

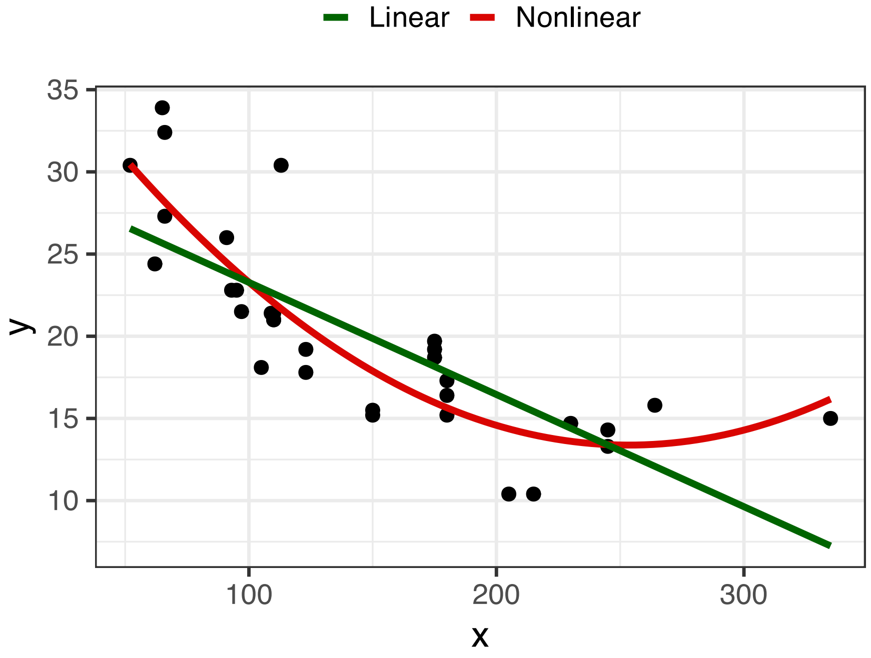
- We can accomodate non-linear relationships in regressions.
- Just add a *higher order* term like this:

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + e_i$$

- This is **multiple regression** (in 2 weeks!)

- For example, suppose we had this data and fit the previous regression model:

Nonlinear relationship between x and y



Analysis of Variance

- Remember that $y_i = \hat{y}_i + e_i$.
- We have the following decomposition:

$$\begin{aligned} Var(y) &= Var(\hat{y} + e) \\ &= Var(\hat{y}) + Var(e) + 2Cov(\hat{y}, e) \\ &= Var(\hat{y}) + Var(e) \end{aligned}$$

- Because:
  - $Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$
  - $Cov(\hat{y}, e) = 0$
- **Total variation (SST) = Model explained (SSE) + Unexplained (SSR)**

Goodness of Fit

- The R^2 measures how well the **model fits the data**.

The formula expresses the R^2 as the ratio of what's **Explained** to the **Total** variation, which is equivalent to 1 minus the ratio of what's **Unexplained** to the **Total** variation.

$$\mathbf{R^2 = \frac{Explained\ Variation}{Total\ Variation} \quad = \quad 1 - \frac{Unexplained\ Variation}{Total\ Variation}}$$

Goodness of Fit

- The R^2 measures how well the **model fits the data**.

Formula using Sums of Squares

In terms of the standard regression sums of squares:

$$\mathbf{R^2} = \frac{\mathbf{ESS}}{\mathbf{TSS}} \quad = \quad \mathbf{1} - \frac{\mathbf{SSR}}{\mathbf{TSS}} \in [0,1]$$

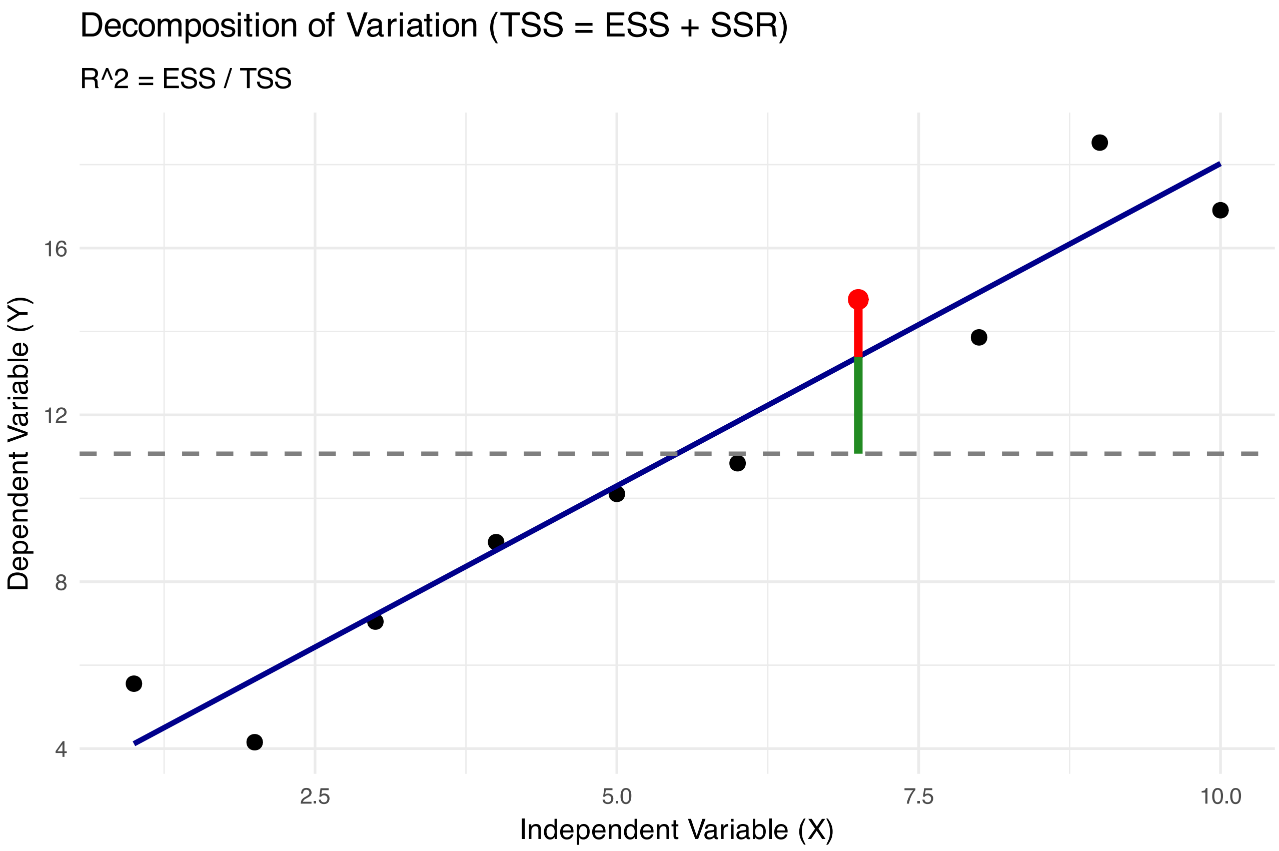
- **ESS (Explained Sum of Squares):** Variation explained by the regression.
- **SSR (Sum of Squared Residuals):** Variation **unexplained** (the error).
- **TSS (Total Sum of Squares):** The total variation in Y.

- R^2 close to 1 indicates a **very *high* explanatory power** of the model.

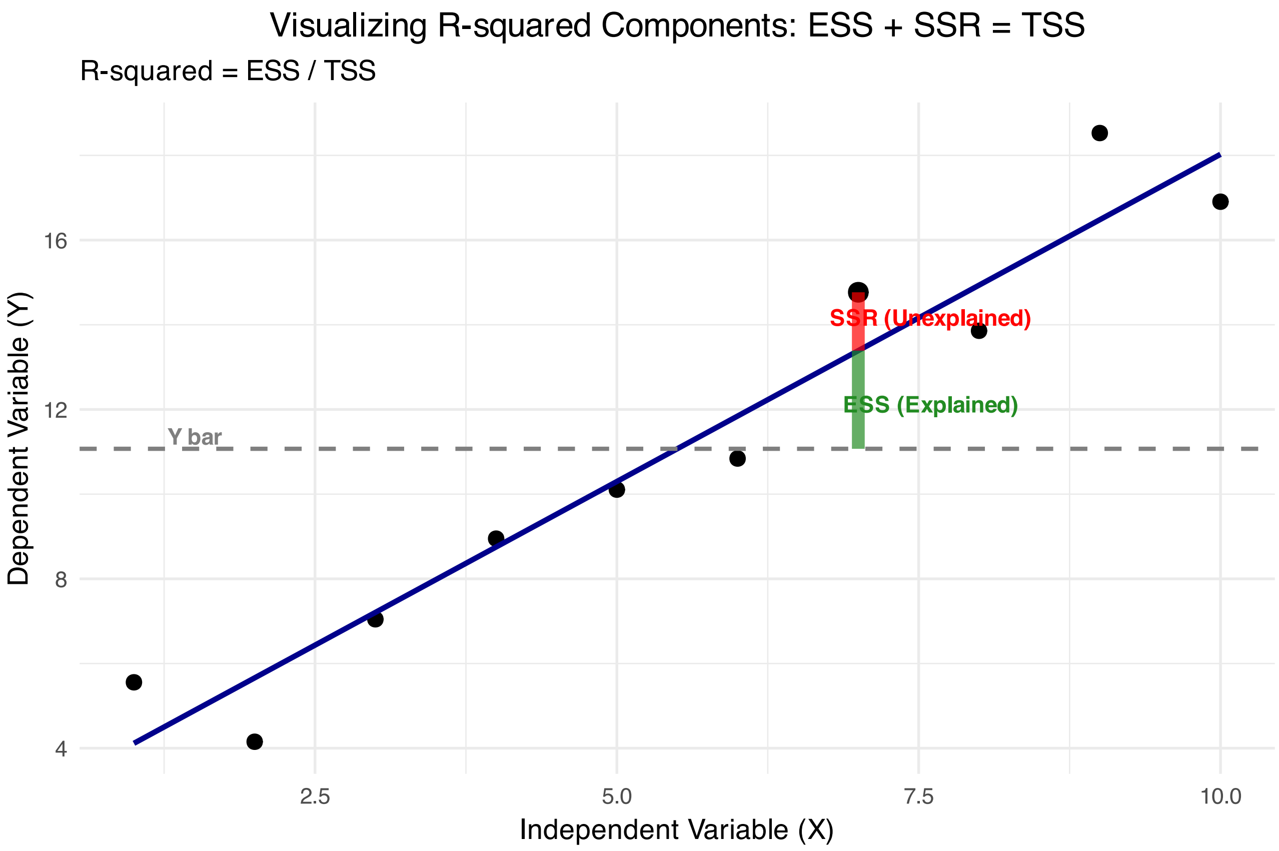
- R^2 close to 0 indicates a **very *low* explanatory power** of the model.

- *Interpretation:* an R^2 of 0.5, for example, means that the variation in x "explains" 50% of the variation in y.

Graphically



Graphically



Visualizing R^2: Components and Sign ⚠️

Recall that the visualization shows the **magnitude** of the vertical distances. Mathematically, the components used in the R^2 formula are **squared** to ensure they are positive and can be summed (TSS = ESS + SSR).

Component	Visual Representation (Distance)	Mathematical Term (Before Squaring)	Sign
TSS (Total)	Y_i to $\bar{Y}$	$(Y_i - \bar{Y})$	Can be positive or negative.
ESS (Explained)	$\widehat{Y}_i$ to $\bar{Y}$	$(\widehat{Y}_i - \bar{Y})$	Can be positive or negative.
SSR (Unexplained)	Y_i to $\widehat{Y}_i$	$(Y_i - \widehat{Y}_i)$	Can be positive or negative.

Key Point

The **Sum of Squares** $(\sum(\cdot)^2)$ is necessary to eliminate the sign and aggregate the variation across **all** data points, resulting in a positive R^2 value between 0 and 1.

Task 3: R^2 and goodness of fit

10:00

1. Regress `avgmath_cs` on `classize`. Assign to an object `math_reg`.

2. Pass `math_reg` in the `summary()` function. What is the (multiple) R^2 for this regression? How can you interpret it?

3. Compute the squared correlation between `classize` and `avgmath_cs`. What does this tell you about the relationship between R^2 and the correlation in a regression with only one regressor?

4. Repeat steps 1 and 2 for `avgverb_cs`. For which exam does the variance in class size explain more of the variance in students' scores?

5. (Optional) Install and load the `broom` package. Pass `math_reg` in the `augment()` function and assign it to a new object. Use the variance in `avgmath_cs` (SST) and the variance in `.fitted` (predicted values; SSE) to find the R^2 using the formula on the previous slide.

Why Hypothesis Testing? 🤔

The Challenge: Sample vs. Population

We're not just interested in the results from our **small sample of data**; we want to make confident **conclusions about the entire population**.

- **Example 1 (Mean):** If the average test score in our sample is 700, can we confidently say the true **population average** (μ) is really 700?
- **Example 2 (Regression):** If our regression shows a coefficient of $-\mathbf{5.82}$ for class size, can we say for sure that this effect is **real** and not just due to random chance in our sample?

The Solution: Statistical Inference

This is where **Hypothesis Testing** comes in.

It's the essential tool we use to move from a **sample finding** to a **population conclusion**—to determine if our results are **statistically significant**.

Our Plan:

1. Testing the **Population Mean (μ)**.
2. Apply that logic to our core econometrics task, e.g. testing the **Regression Coefficients (β_i)**.

Hypothesis Testing for the Population Mean (μ)

The Simple t-Test

1. Hypotheses

$$\begin{aligned} H_0 &: \mu = \mu_0 \quad \text{(Population mean equals hypothesized value)} \\ H_1 &: \mu \neq \mu_0 \quad \text{(Population mean is different from hypothesized value)} \end{aligned}$$

2. Test Statistic The test statistic is: `

$$t^* = \frac{\bar{X} - \mu_0}{SE(\bar{X})} \sim t_{n-1}$$

$where: -\bar{X}: samp \leq mean - \mu_0: hypothesizedpop\underline{a}tionmean(\mathfrak{o}mH_0) -$
$SE(\bar{X}) = \frac{s}{\sqrt{n}}: S\tan dardErr\ \mathrm{or}\ ofthesamp \leq mean - n$
$: \nu mberofobservations(samp \leq size) - n - 1: degreesof\mathfrak{r}edomofthet$`-distribution

3. Decision Rule (Critical Value Approach)

Decision rule: Reject H_0 if the calculated absolute t^*:

$$|t^*| > t_{\alpha/2,\, n-1}$$

The interval $\left[-\, t_{\alpha/2,\, n-1},\ t_{\alpha/2,\, n-1} \right]$ is the **Non-rejection Region**.

4. Decision Rule (p-value approach)

If (p-value < α), we **reject the null hypothesis** at the α significance level. (Commonly α =0.05).

5. Interpretation of Results

If (H_0) is rejected $\rightarrow$ The true population mean (μ) is **statistically different** from (μ_0).

If (H_0) is not rejected $\rightarrow$ There is **insufficient evidence** to conclude that μ is different from (μ_0).

Practical Example

The Scenario: State Mandate

- **Problem:** A parent group claims the true **average test score (μ)** is different from the mandated 650 points.
- H_0: The district meets the mandate ($\mu = 650$)
- H_1: The district fails the mandate ($\mu \neq 650$)
- **Significance Level:** $\alpha = 0.05$

The Data & Test Statistic

We take a sample ($n = 25$) and find strong evidence:

Statistic	**Value**
Sample Mean ($\bar{X}$)	**628 points**
Sample Standard Deviation (s)	50 points
Sample Size (n)	25

The calculated **Test Statistic (t^*)** is:

$$t^* = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{628 - 650}{50/\sqrt{25}} = \frac{-22}{10} = \mathbf{-2.2}$$

The Decision: Reject H_0

1. Critical Value Approach

$|t^*| = |-2.2| = 2.2 \quad \text{is} \quad > 2.064 \quad (t_{0.025,24})$

Conclusion: Reject H_0

2. p-value Approach

$p\text{-value} \approx \mathbf{0.037}$

Conclusion: Since $0.037 < 0.05$, **Reject H_0**

The Interpretation

The sample mean of 628 is **statistically significantly** different from 650. We reject the null hypothesis and conclude there is strong evidence the true average score is below the state mandate.

Hypothesis Testing for Regression Coefficients

Individual Significance Test

1. Hypotheses

$$\begin{aligned} H_0 &: \beta_i = 0 \quad \text{(no effect)} \\ H_1 &: \beta_i \neq 0 \quad \text{(significant effect)} \end{aligned}$$

2. Test statistic The test statistic is: `

$$t^* = \frac{\widehat{\beta}_i - \beta_i}{SE(\widehat{\beta}_i)} \sim t_{n-k-1}$$

$Thety\pi caltestf$ or $H_0 : \beta_i = 0simpl$ if $iesthis \rightarrow :$

$$t^* = \frac{\widehat{\beta}_i - 0}{SE(\widehat{\beta}_i)} \sim t_{n-k-1}$$

` where: - $n$: number of observations - $k$: number of parameters - $n - k - 1$: degrees of freedom of the t-distribution - (`$\alpha) : sign$ if $icance \leq vel(probabilityofreject \in gH_0$ $whenH_0istrue, usually\alpha$`=0.05)

**3. Decision Rule (Critical Value Approach)

Decision rule: If the absolute calculated t^* is within the Non-rejection Region (N.R.):

$$t^* \in \text{N.R.} = \left[-t_{\alpha/2,\, n-k-1},\ t_{\alpha/2,\, n-k-1} \right]$$

we **do not reject the null hypothesis** (H_0).

The critical values in this interval are obtained from the t-distribution tables.

**4. Test p-value

If (p-value < 0.05),
we **reject the null hypothesis** at the 5% significance level.

**5. Interpretation of Results

If (H_0) is rejected → variable (X_i) is **relevant** to explain variable (Y).
If (H_0) is not rejected → variable (X_i) has **no statistically significant effect** on (Y).

On the way to causality

✅ How to manage data? Read it, tidy it, visualise it…

🚧 **How to summarise relationships between variables?** Simple linear regression… to be continued

❌ What is causality?

❌ What if we don't observe an entire population?

❌ Are our findings just due to randomness?

❌ How to find exogeneity in practice?

SEE YOU NEXT WEEK!

| | | | :-- | :------------------------------ | | ✈ | florian.oswald@sciencespo.fr | | 🔗 | Slides | | 🔗 | Book | | 🐦 | @ScPoEcon | | | @ScPoEcon |