SciencesPo
DEPARTMENT OF ECONOMICS

Applied Data Analysis for Public Policy Studies

Simple Linear Regression

Michele Fioretti
SciencesPo Paris
2020-09-11

Recap from past weeks

- `R` basics, importing data
- Exploratory data analysis:
  - Summary statistics: *mean*, *median*, *variance*, *standard deviation*
  - Data visualization: base `R` and `ggplot2`
  - Data wrangling: `dplyr`

Recap from past weeks

- `R` basics, importing data

- Exploratory data analysis:

    - Summary statistics: *mean*, *median*, *variance*, *standard deviation*
    - Data visualization: base `R` and `ggplot2`
    - Data wrangling: `dplyr`

Today - Real 'metrics finally ✌️

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

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

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

Class size and student performance

- What policies *lead* to improved student learning?
- Class size reduction has been at the heart of policy debates for *decades*.

Class size and 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 (7 minutes)

1. Load the data from here. You need to find the function that enables importing *.dta* files. (FYI: *.dta* is the extension for data files used in *Stata*)

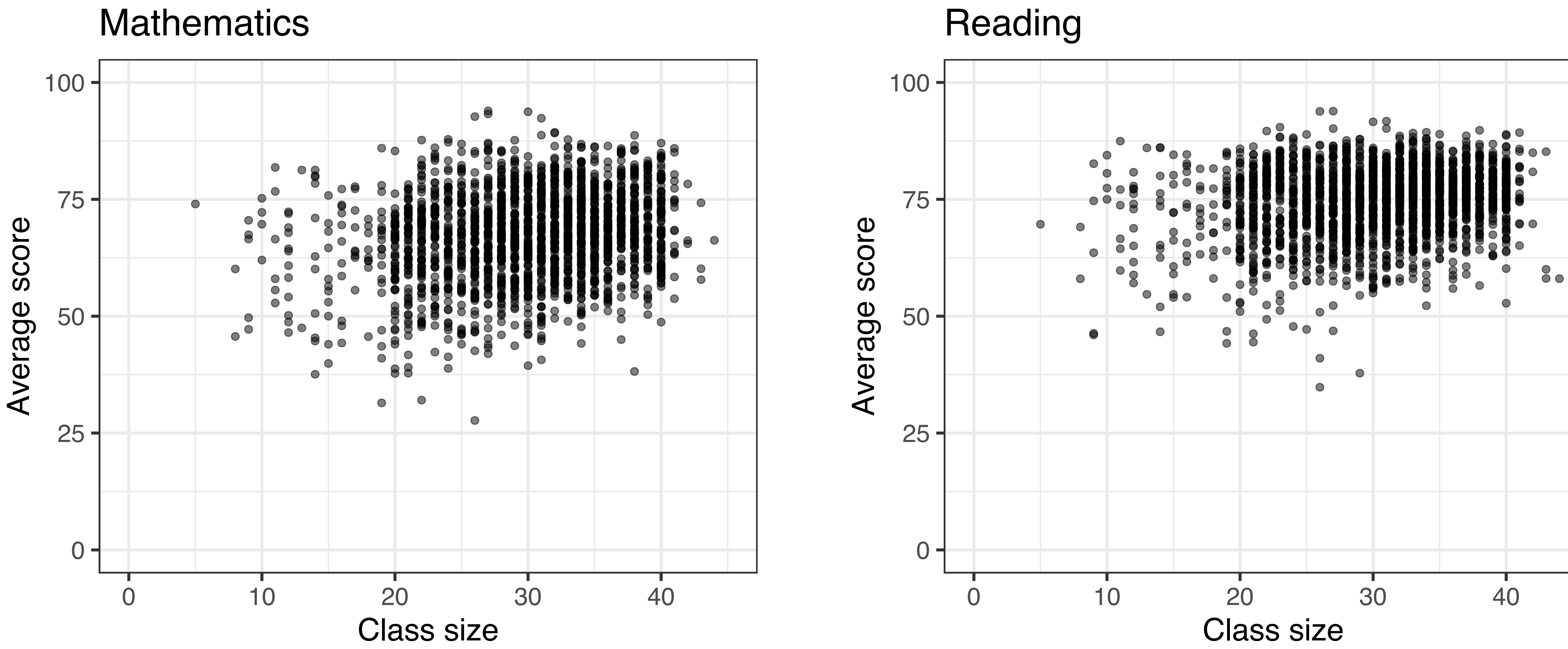
2. Describe the dataset:

   - What is the unit of observations, i.e. what does each row correspond to?
   - How many observations are there?
   - What variables do we have? `View` the dataset to see what the variables correspond to.
   - 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()`.

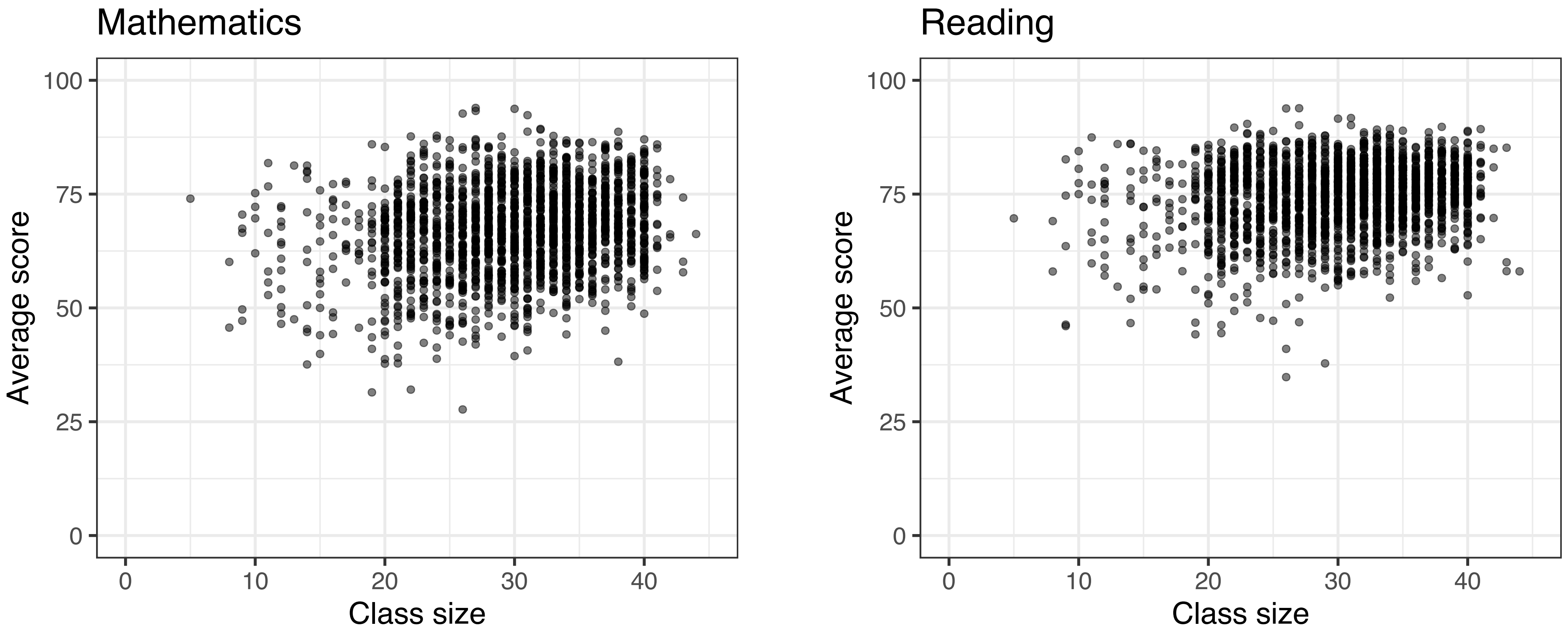
3. Do you have any priors about the actual (linear) relationship between class size and student achievement? What would you do to get a first insight?

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

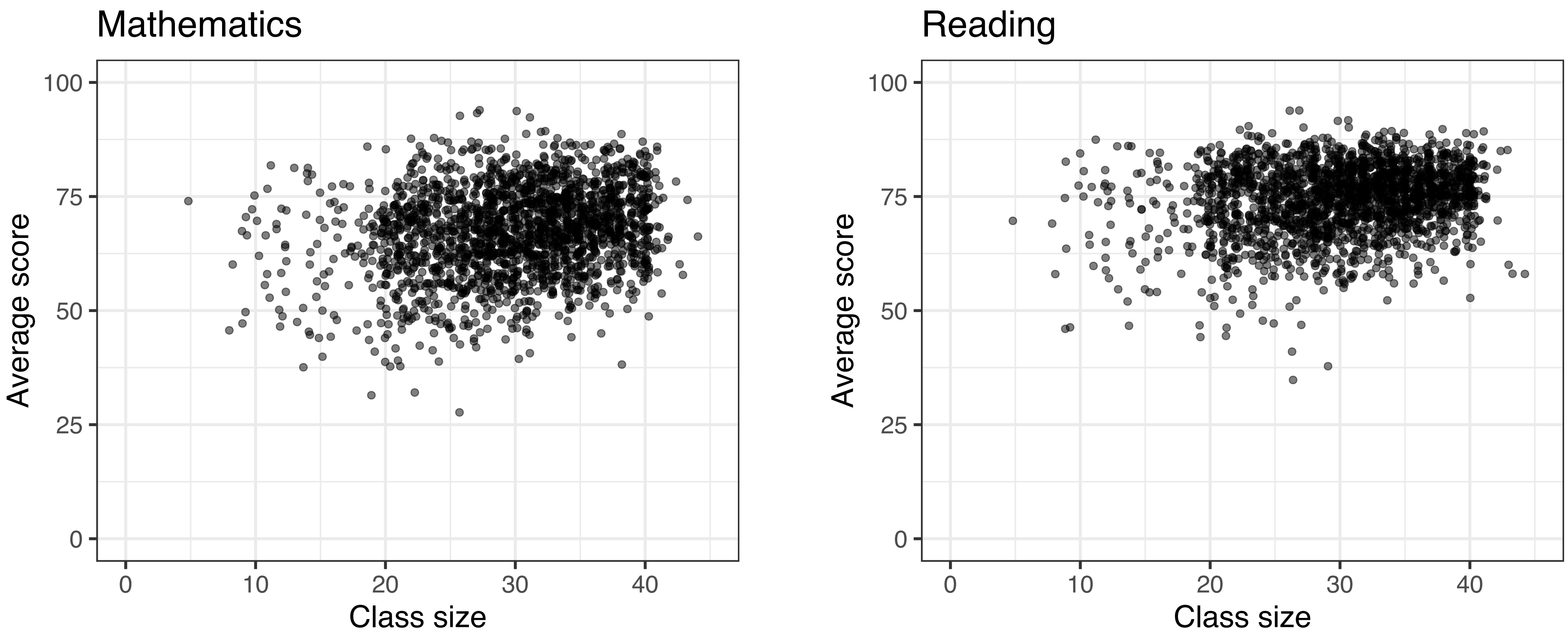


Class size and student performance: Scatter plot

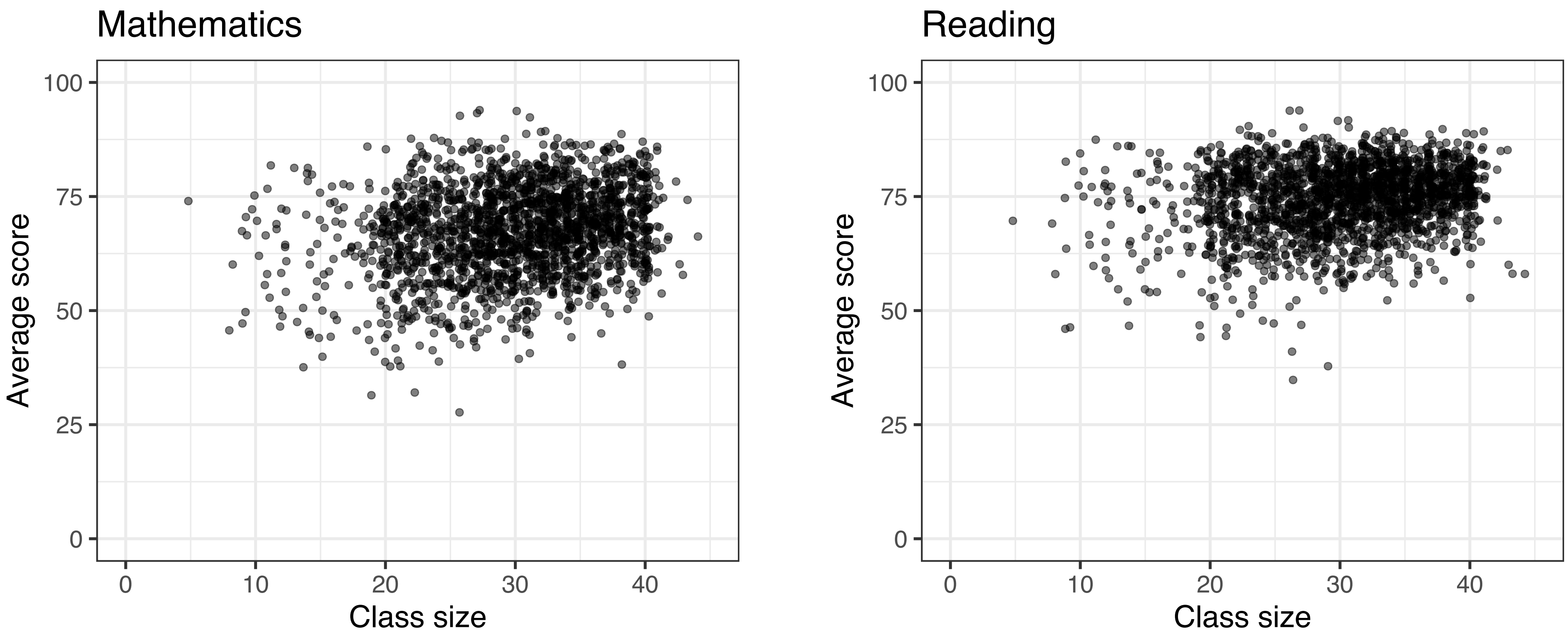


- Hard to see much because all the data points are aligned vertically. Let's add a bit of `jitter` to disperse the data slightly.

Class size and student performance: `jitter` scatter plot



Class size and student performance: `jitter` scatter plot



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

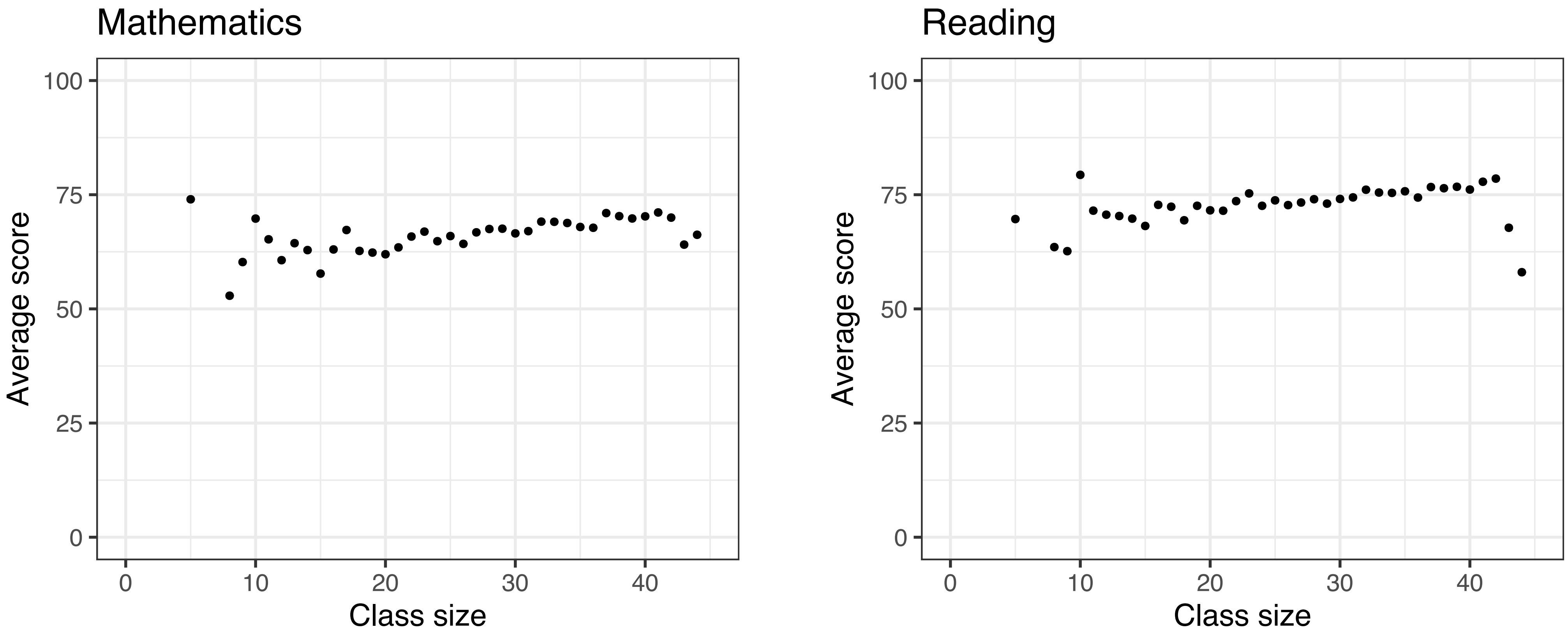
Task 2: Binned scatter plot (7 minutes)

1. Create a new dataset (`grades_avg_cs`) where math and verbal scores are averaged by class size. Let's call these new average scores `avgmath_cs` and `avgverb_cs`.
   N.B.: the "raw" scores are already averages at the class level. Here we average these averages by class size.

2. Redo the same plots as before. Is the sign of the relationship more apparent?

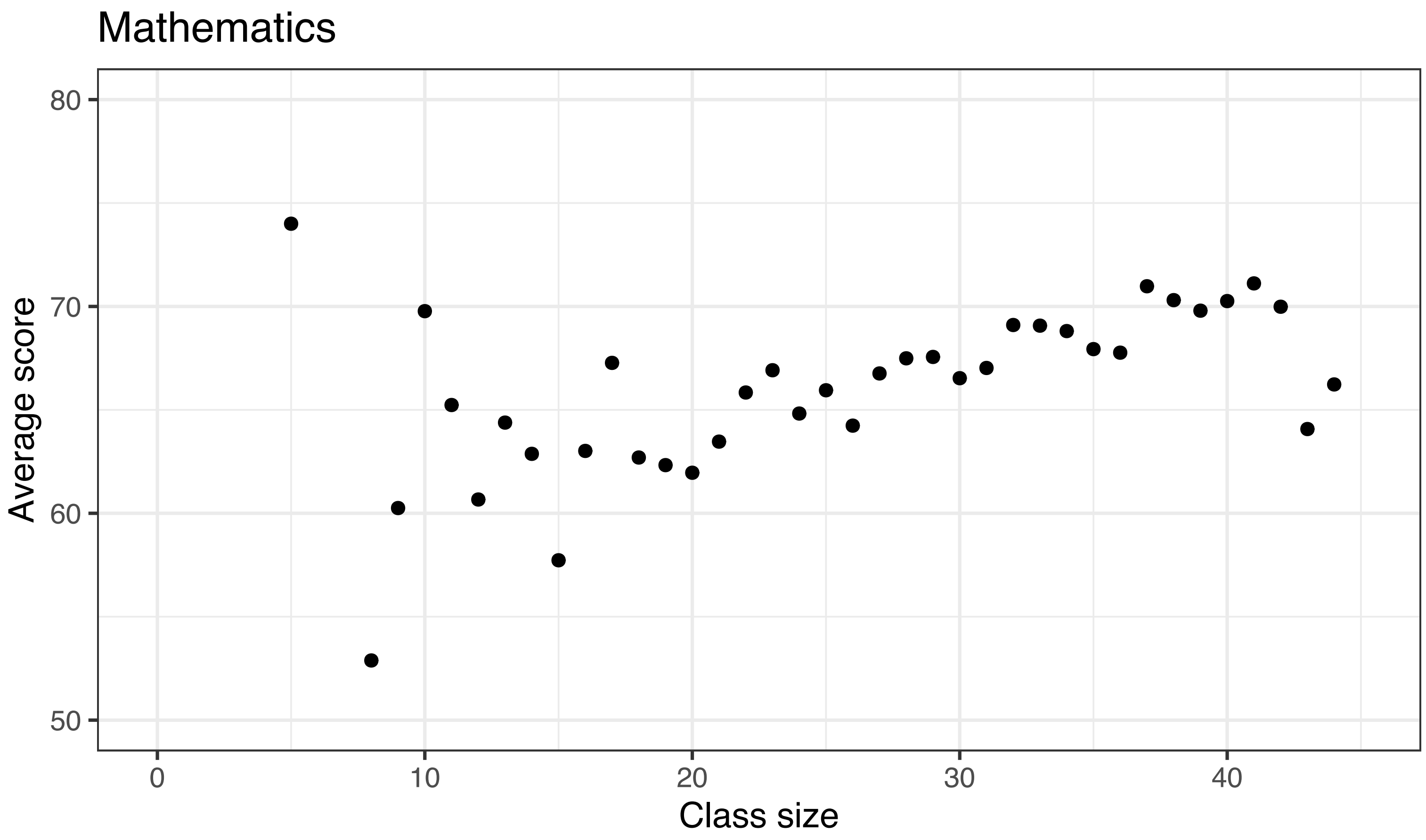
3. Compute the correlation between class size and the new aggreagated math and verbal scores variables. Why is the (linear) association so much stronger?

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

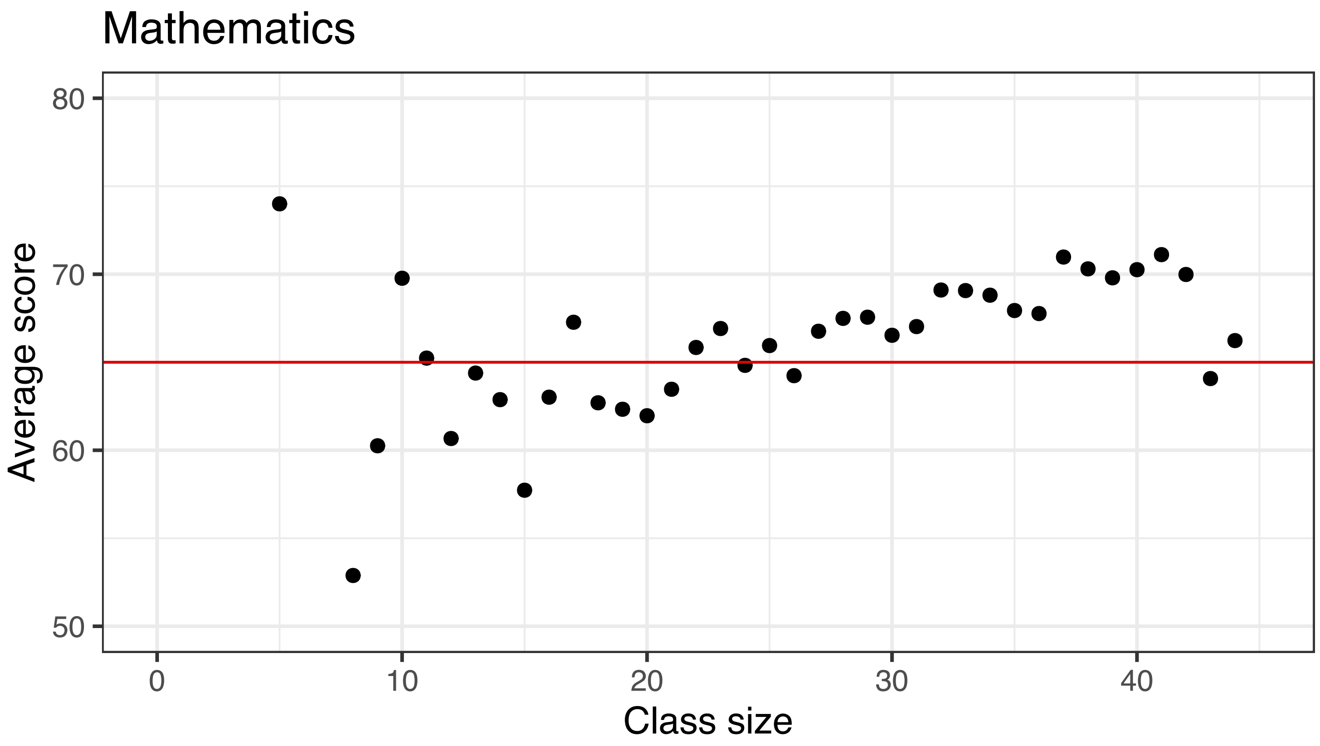


Class size and student performance: Regression line

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

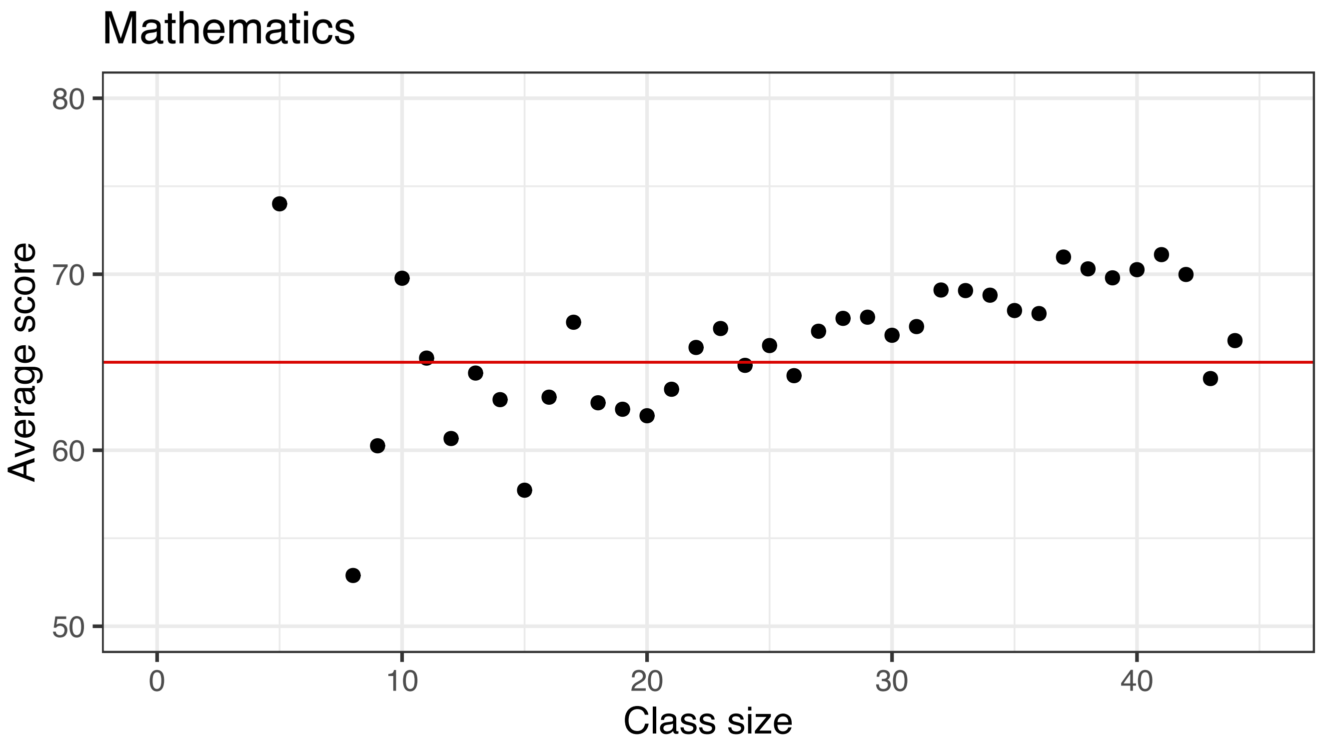
Class size and student performance: Regression line

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



Class size and student performance: Regression line

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

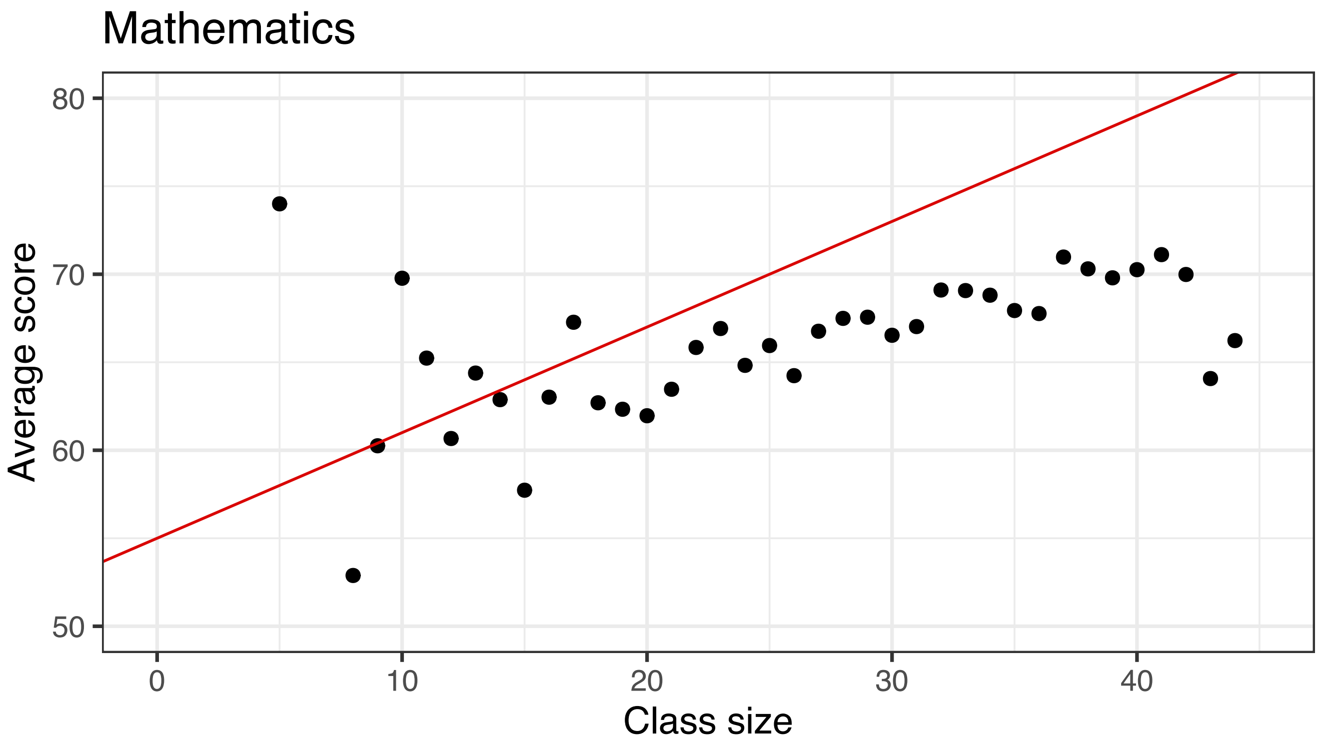


- A *line*! Great. But **which** line? This one?

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

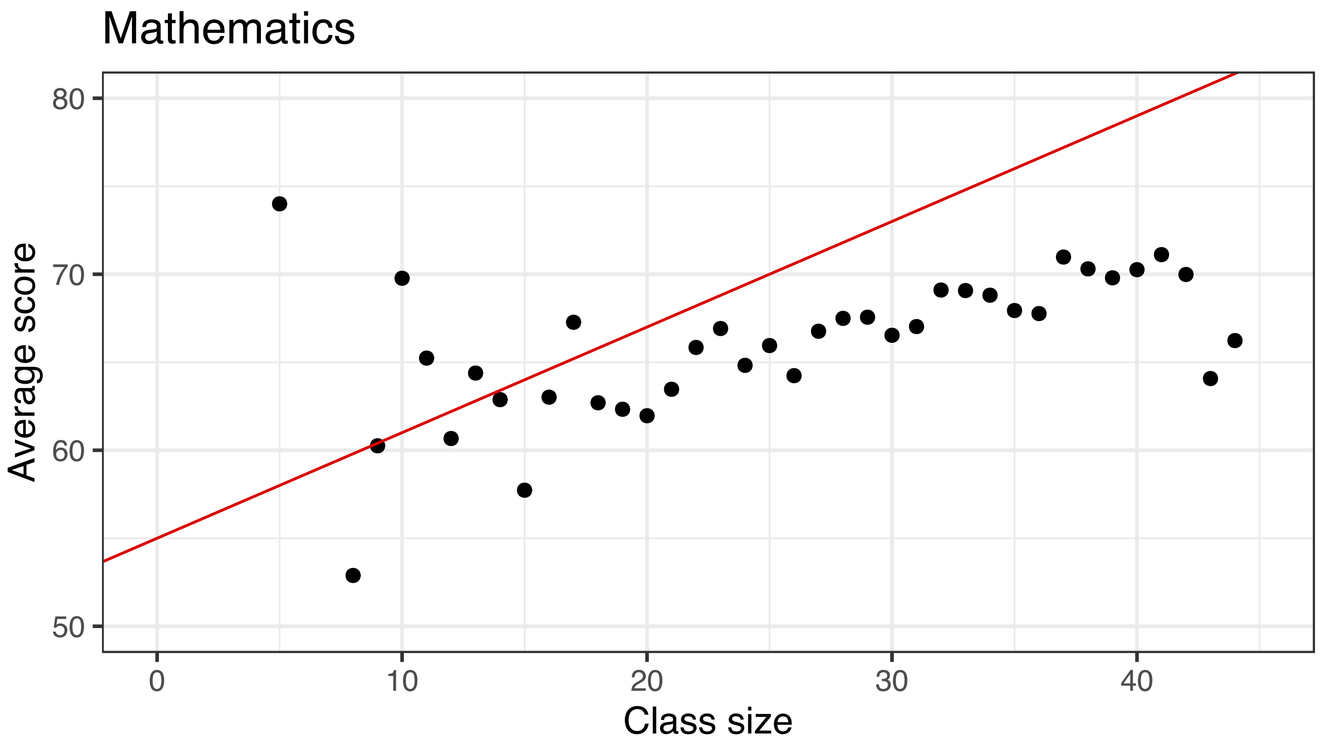
Class size and student performance: Regression line

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



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:

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)

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)

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.

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:

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

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:

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

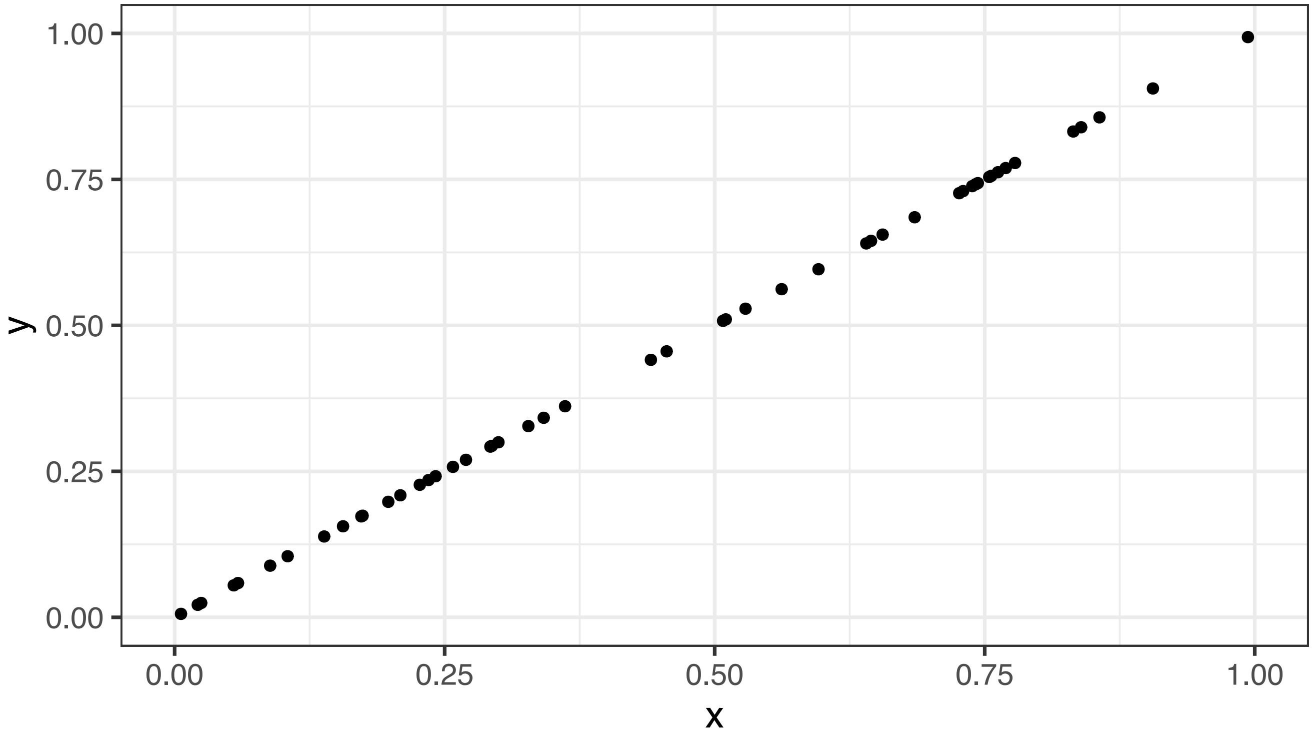
- $\hat{y}_i$ is our *prediction* for y at observation i (y_i) given our model (i.e. the line).

Simple Linear Regression: Error term

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

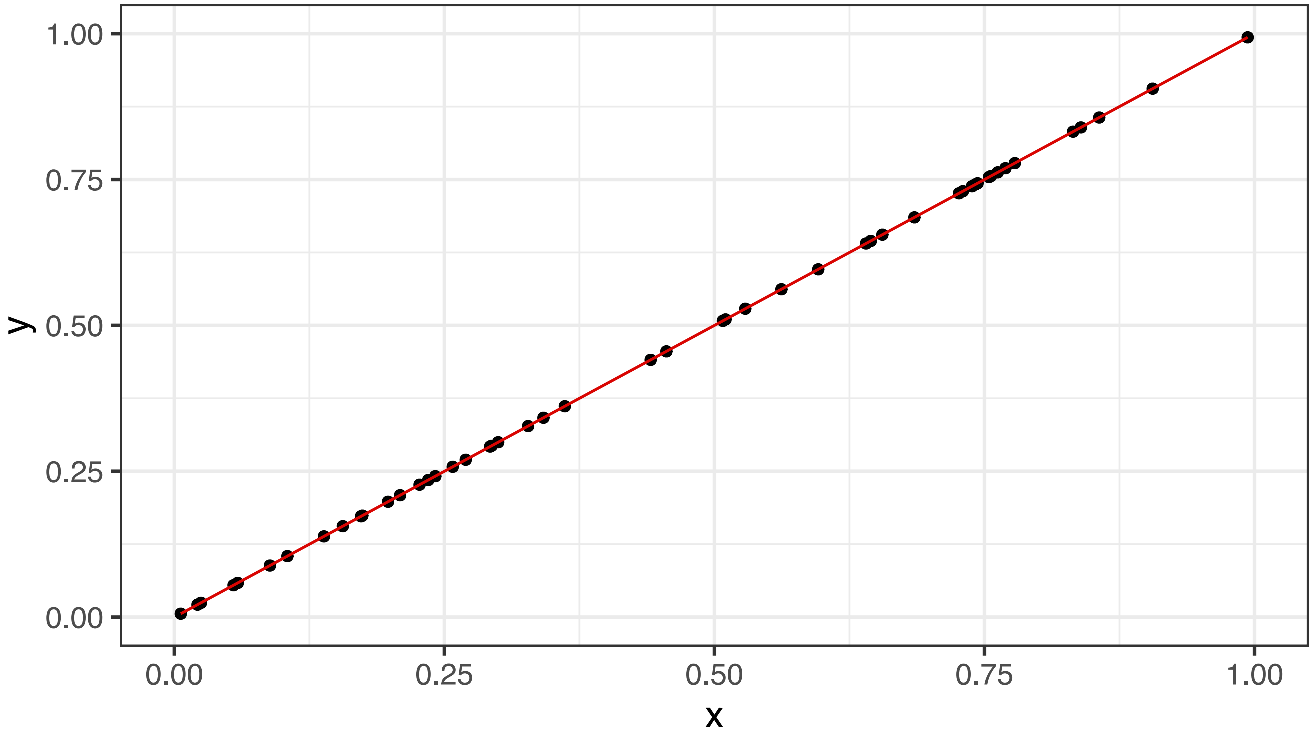
Simple Linear Regression: Error term

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



Simple Linear Regression: Error term

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



Simple Linear Regression: Error term

- 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 **error term**.

Simple Linear Regression: Error term

- 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 **error term**.
- At point (x_i, y_i), we note this error e_i.

Simple Linear Regression: Error term

- 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 **error term**.
- At point (x_i, y_i), we note this error e_i.
- The *actual data* (x_i, y_i) can thus be written as *prediction* + *error*:

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

Simple Linear Regression: Error term

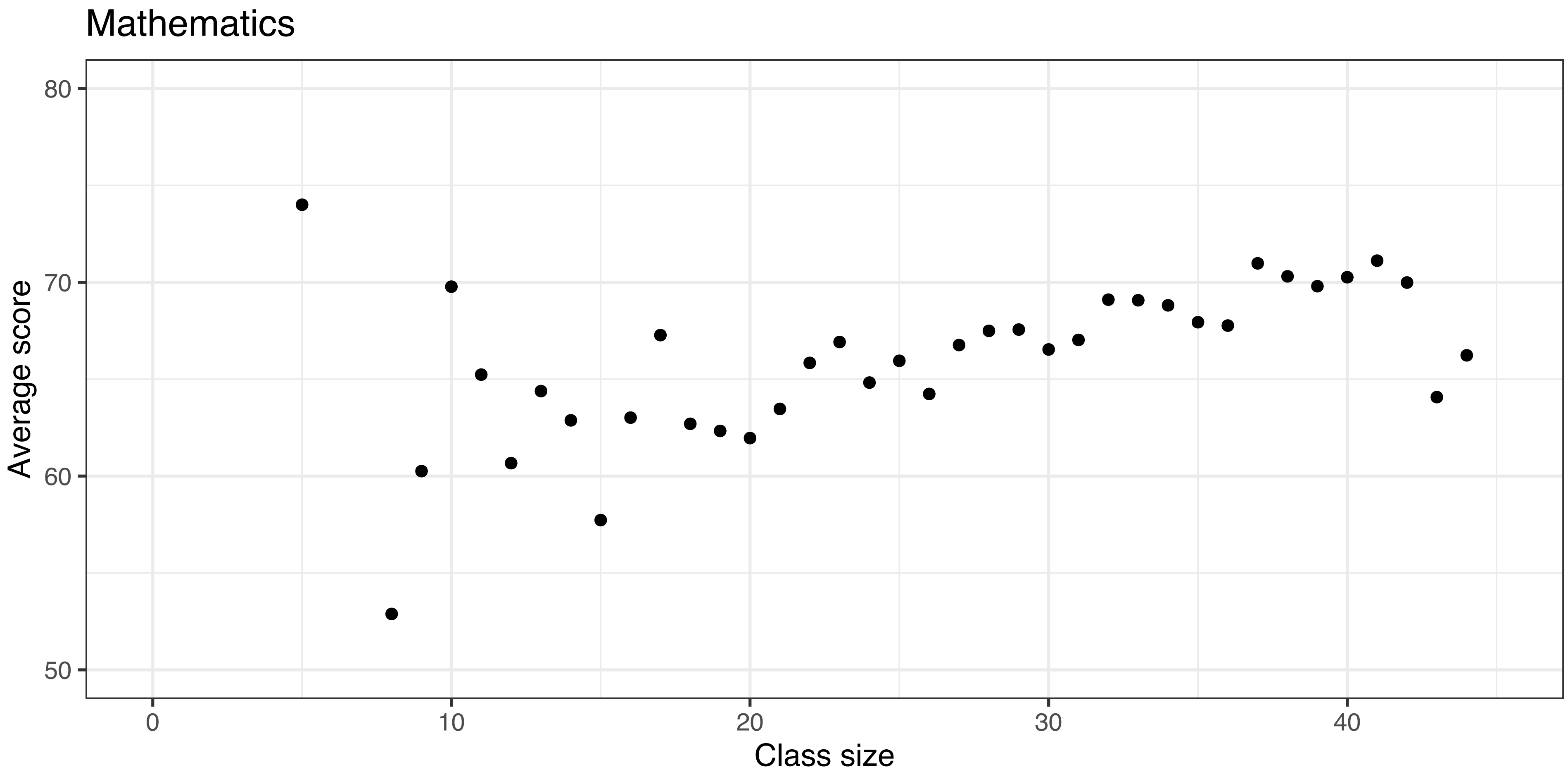
- If all the data points were **on** the line then $\hat{y}_i = y_i$.
- However, since in most cases the *dependent variable* (y) is not ***only*** explained by the chosen *independent variable* (x), $\hat{y}_i \neq y_i$, i.e. we make an **error**.
  This **error** is called the **error term**.
- At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

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

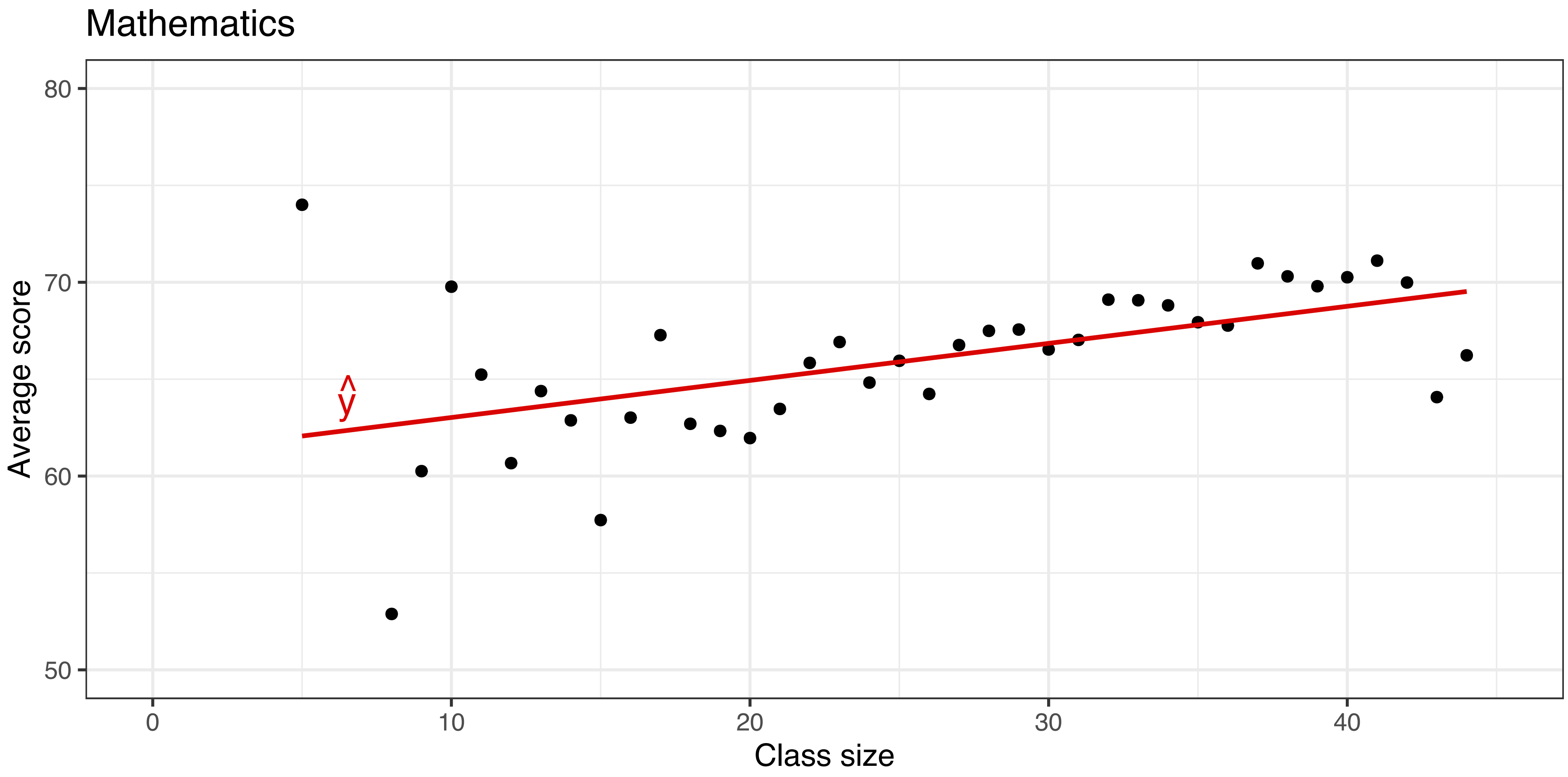
- **Goals**

  1. Find the values for b_0 and b_1 that **make the errors as small as possible**,
  2. Check whether these values **give a reasonable description of the data**.

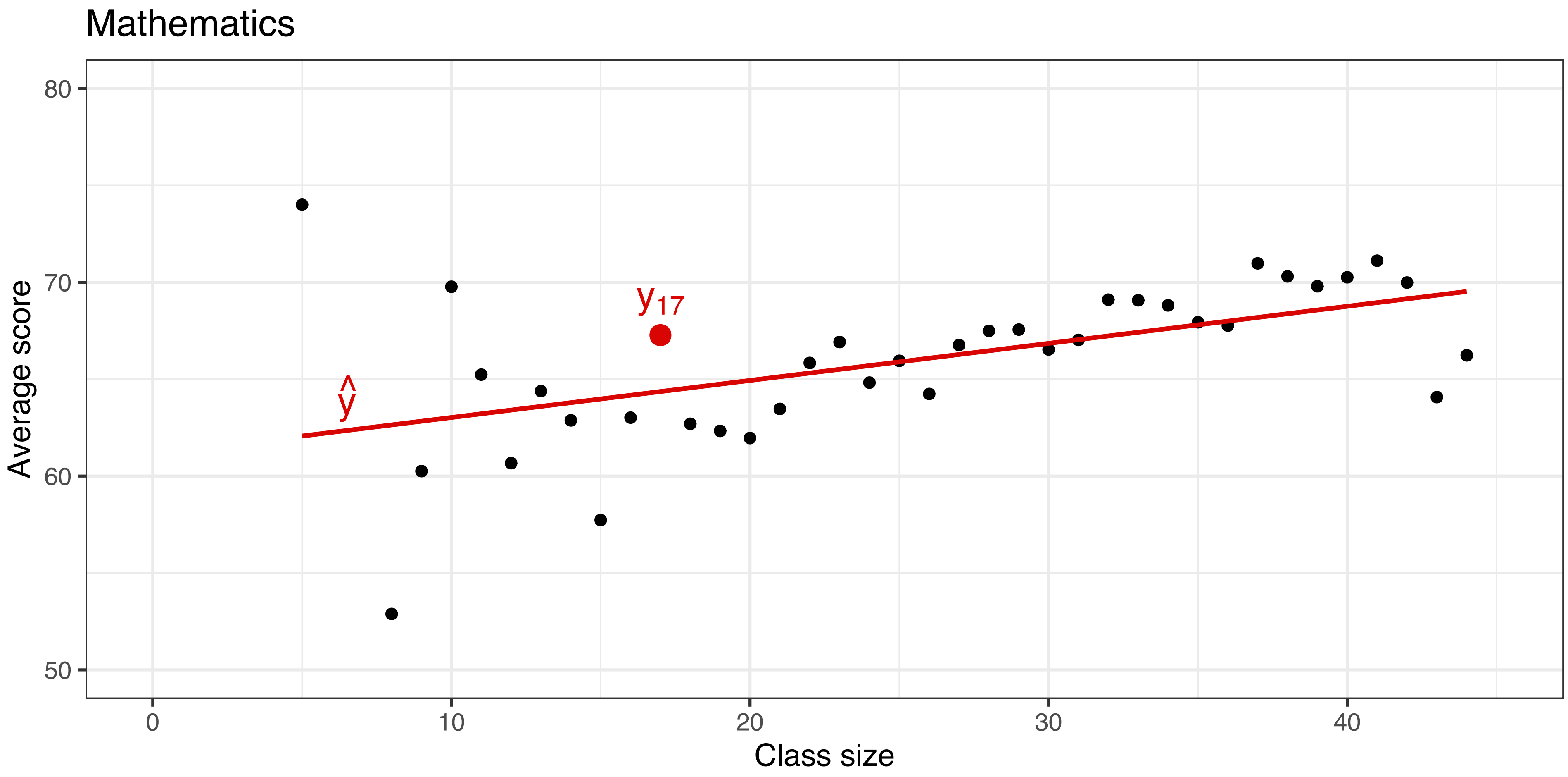
Simple Linear Regression: Graphically



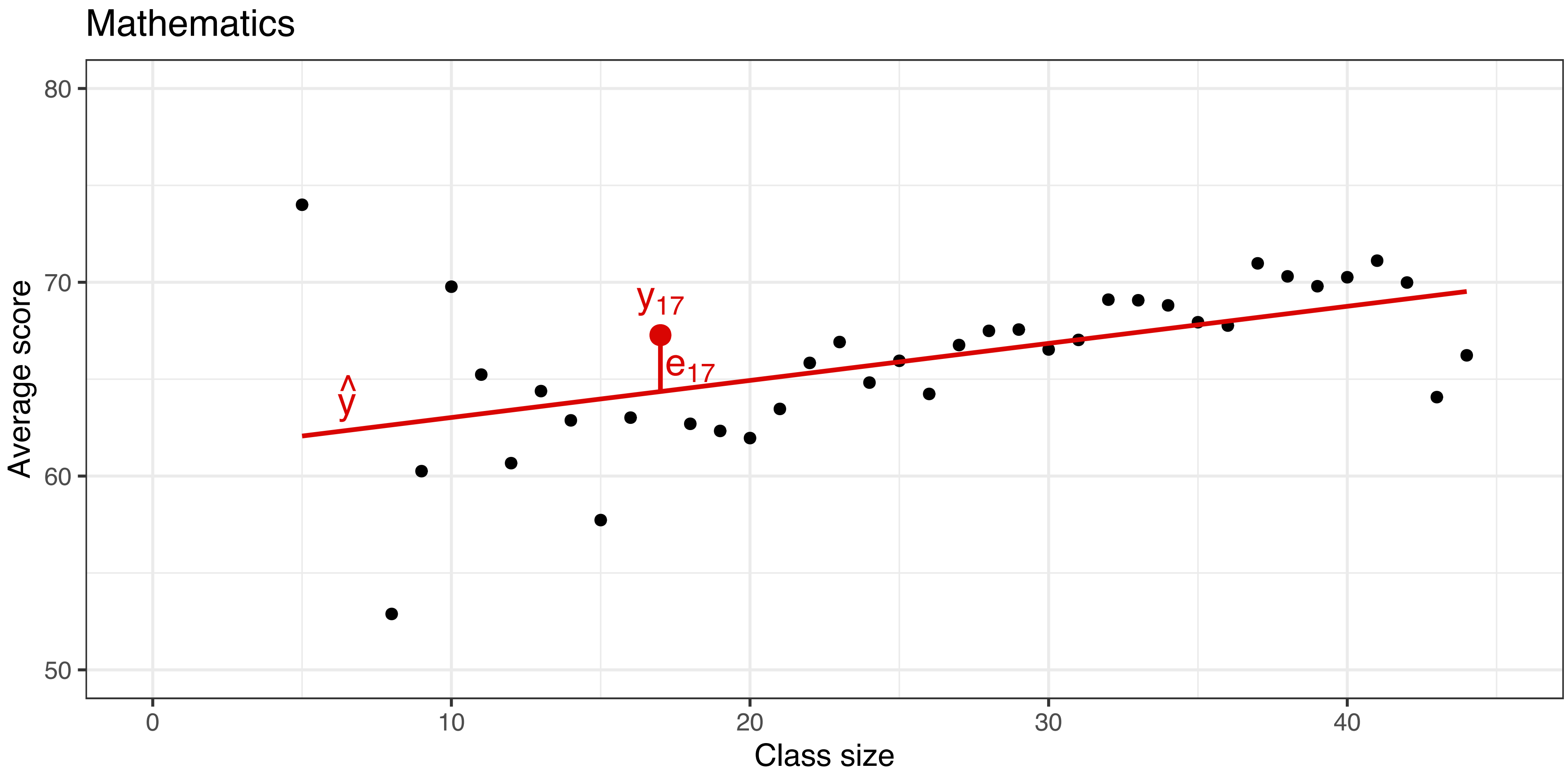
Simple Linear Regression: Graphically



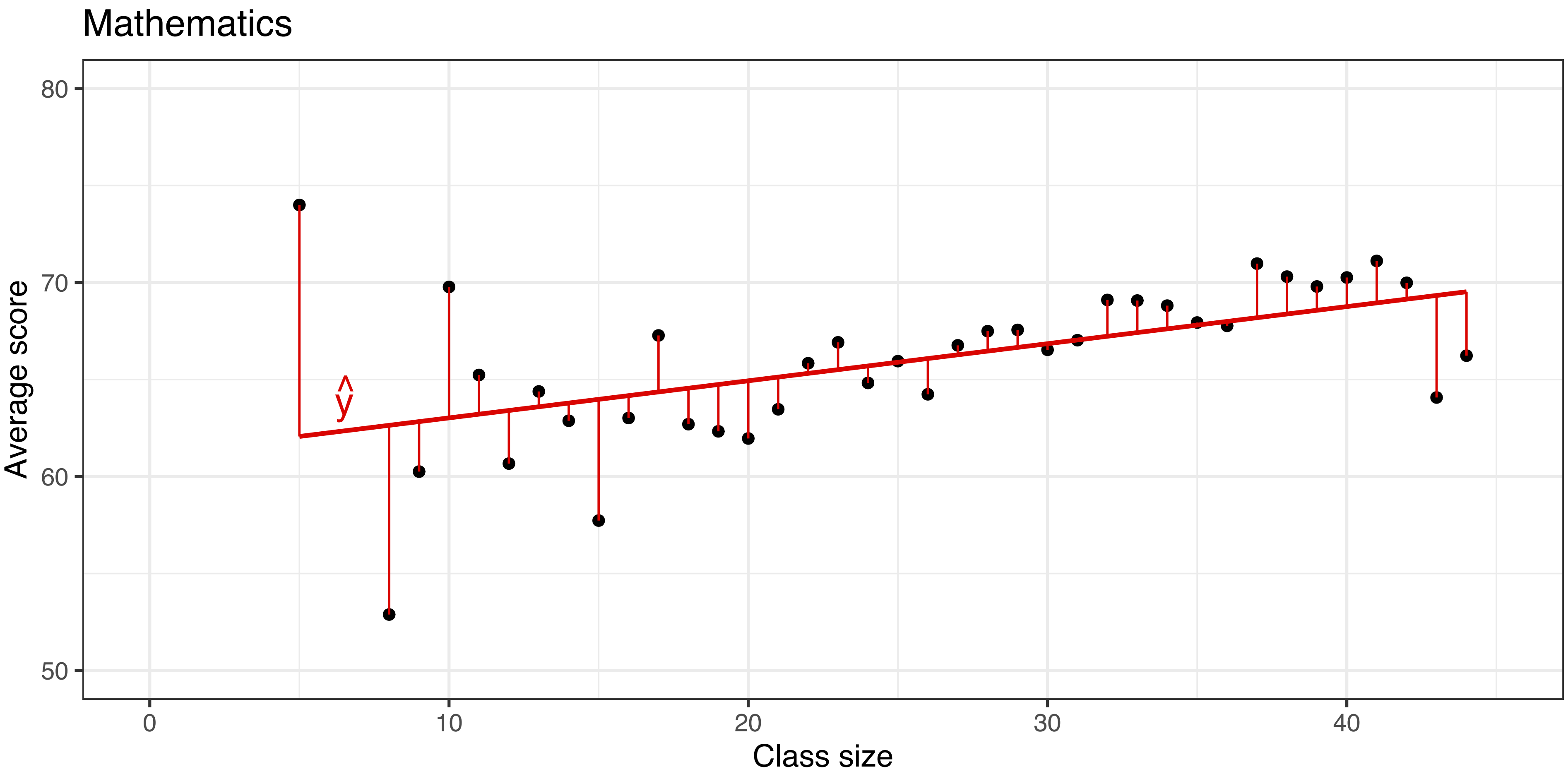
Simple Linear Regression: Graphically



Simple Linear Regression: Graphically



Simple Linear Regression: Graphically



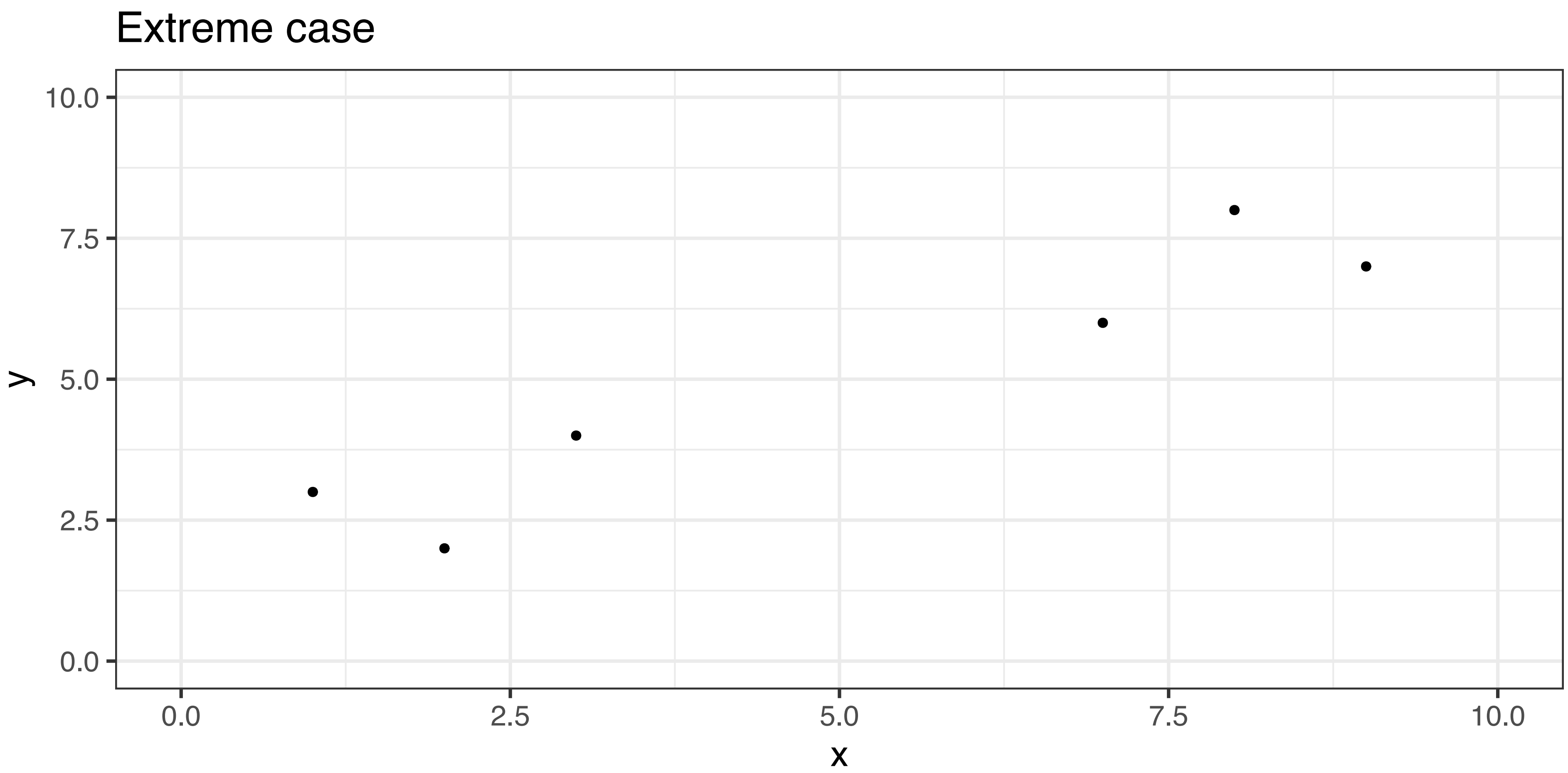
App Time! (5 minutes)

Intuitively one might want to simply minimize the absolute value of the sum of all the errors $|\sum_{i=1}^n e_i|$, that is one might want the sum of errors as close to 0 as possible.

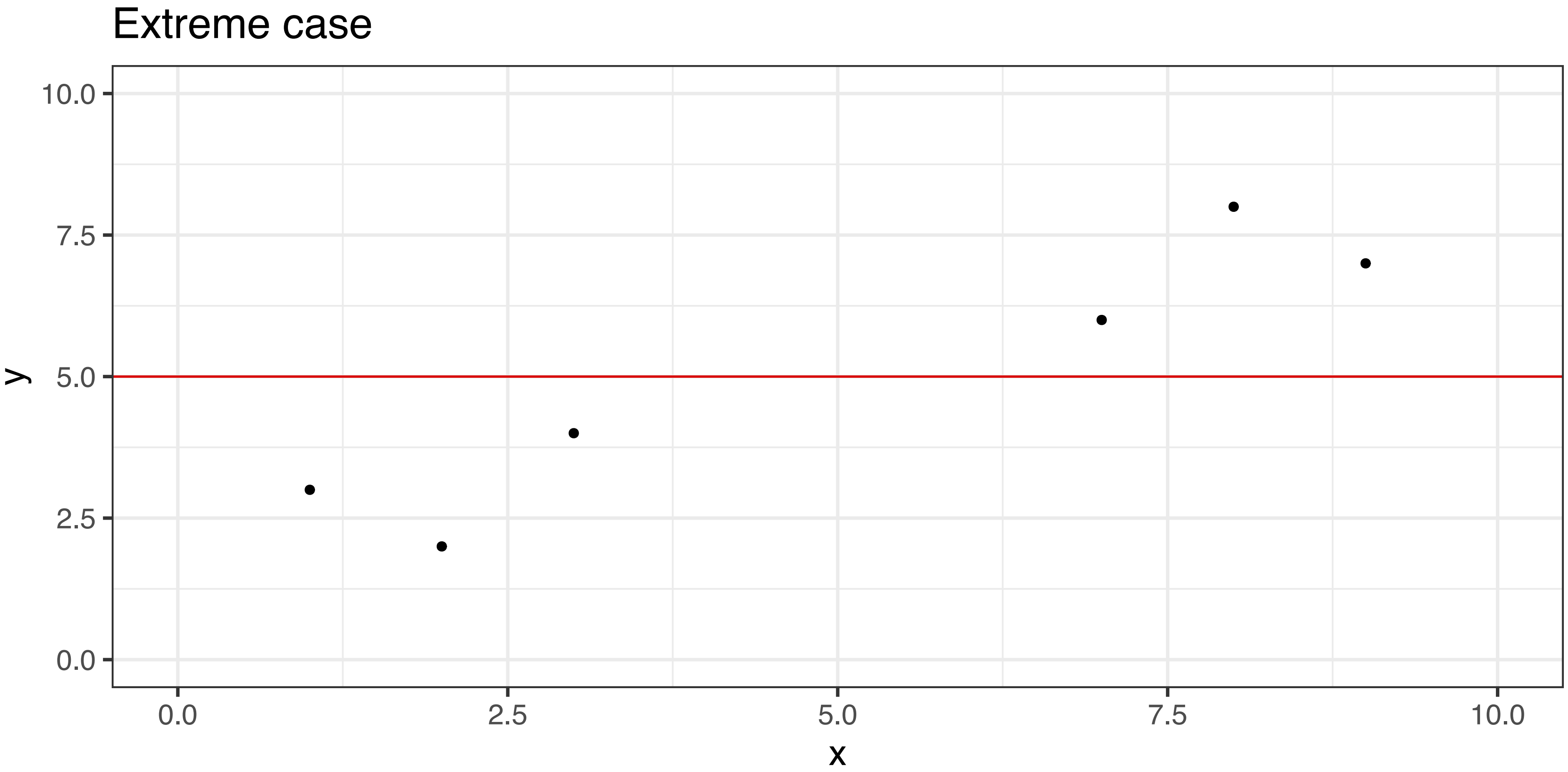
Let's try to find the best line by minimizing the absolute value of the sum of errors. The line *won't* turn green so don't spend all day waiting for it.

```
library(ScPoApps) # load our library
launchApp('reg_simple_arrows')
aboutApp('reg_simple_arrows') # explainer about app
```

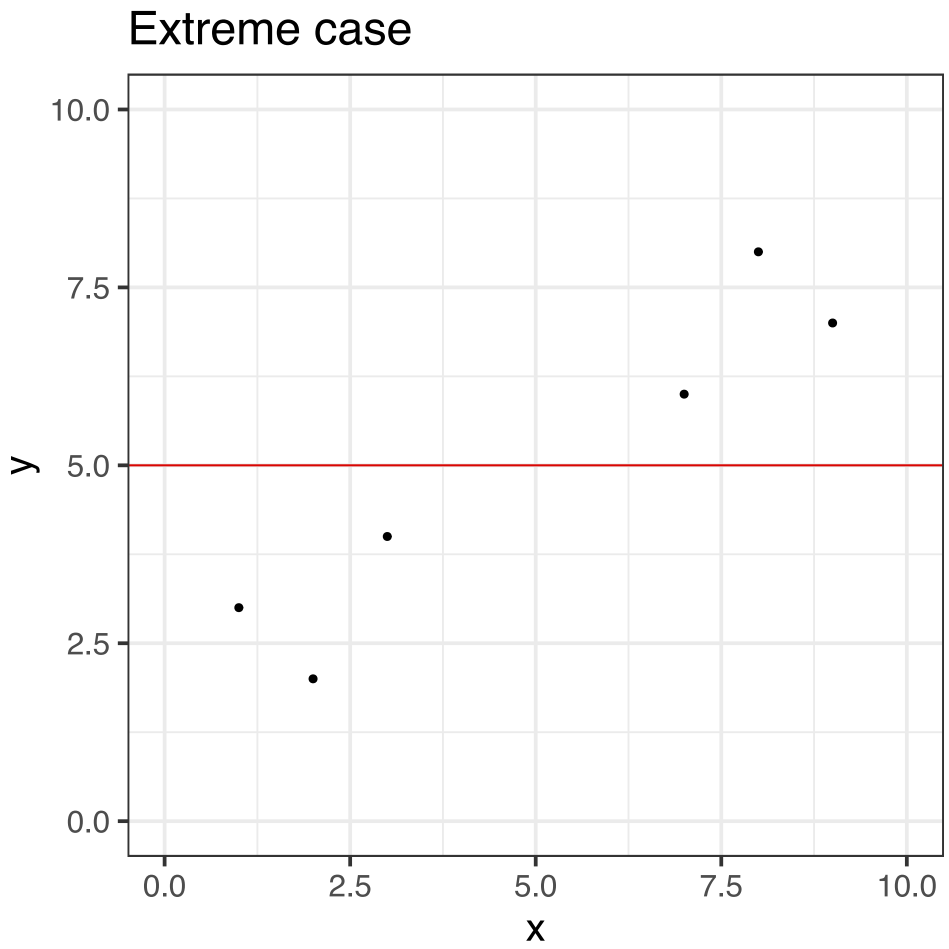
Minimizing the Absolute Value of the Sum of Errors



Minimizing the Absolute Value of the Sum of Errors



Minimizing the Absolute Value of the Sum of Errors



- This line minimizes the absolute value of the sum of errors since the data points are symmetric around $y = 5$.
- Yet it clearly does not fit the data well!
- Note also that many other lines would also yield a sum of errors of 0 since the data are symmetric. A unique solution is not guaranteed!

Ordinary Least Squares (OLS) Estimation

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

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

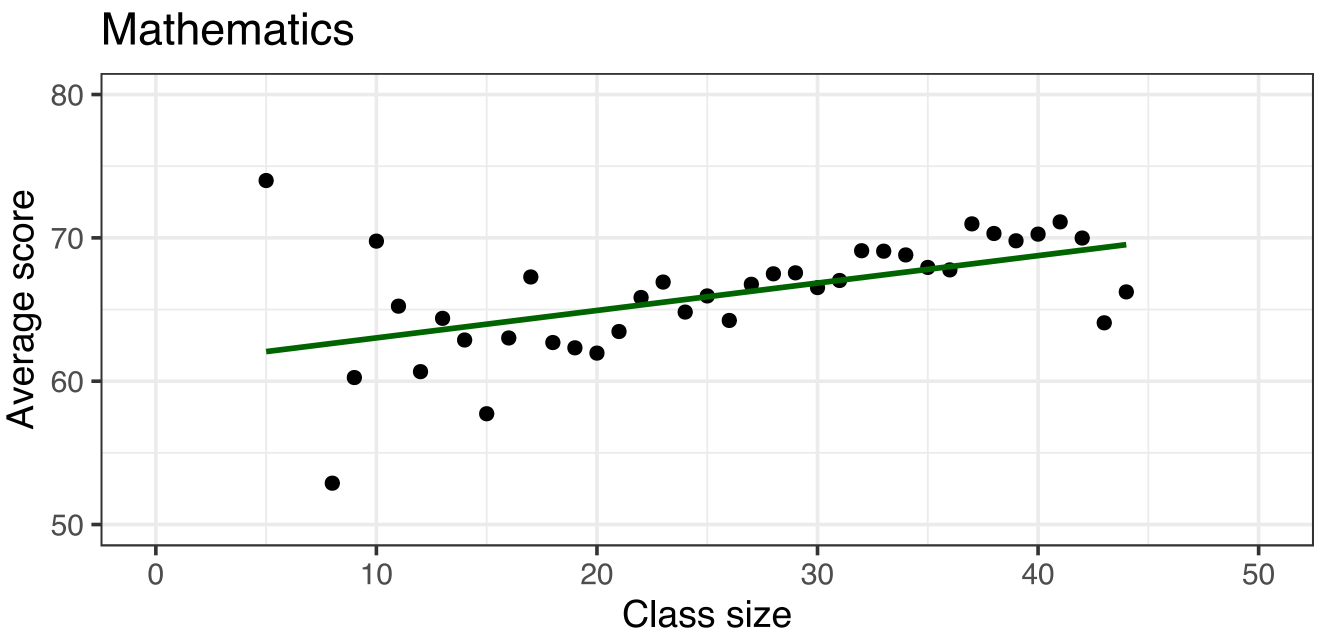
- Choose (b_0, b_1) such that $\sum_{i=1}^N e_1^2 + \cdots + e_N^2$ is **as small as possible**.

Ordinary Least Squares (OLS) Estimation

- Errors of different sign $(+/-)$ cancel out, so let's 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**.

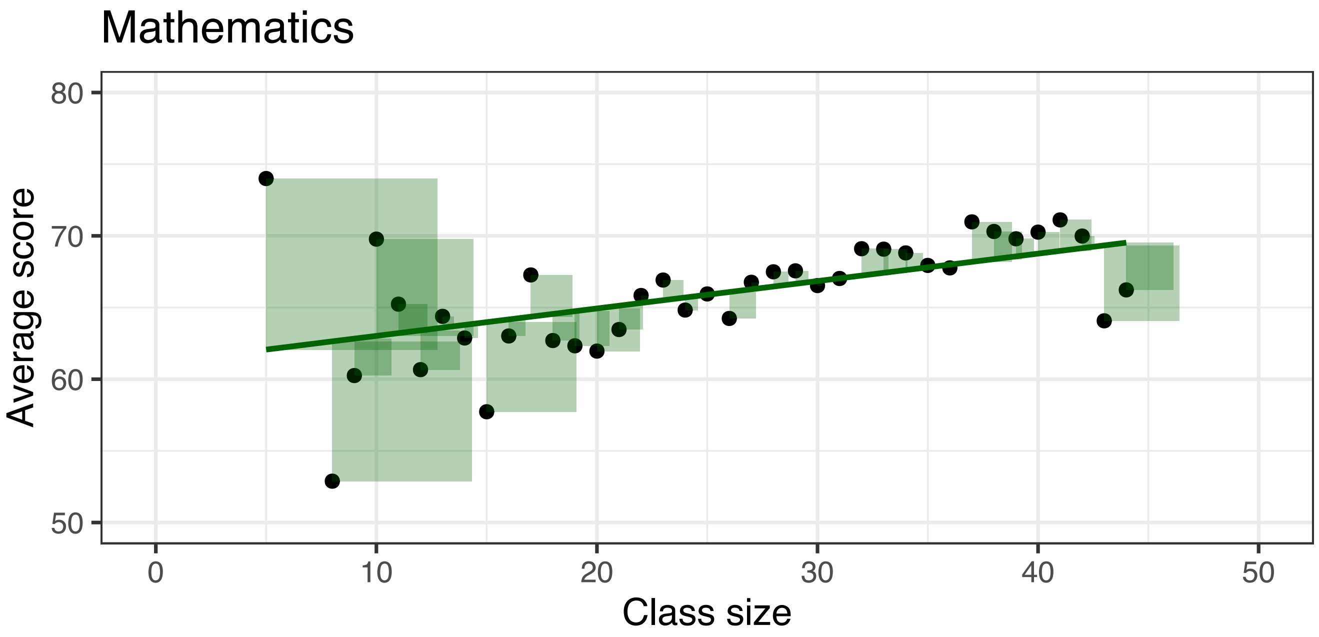


Ordinary Least Squares (OLS) Estimation

- Errors of different sign $(+/-)$ cancel out, so let's 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**.



App Time! #2 (3 minutes)

Let's minimize some squared errors!

```
launchApp('reg_simple')
aboutApp('reg_simple')
```

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)?

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:** $b_0^{OLS} = \bar{y} - b_1\bar{x}$

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:** $b_0^{OLS} = \bar{y} - b_1\bar{x}$

- ❗ You should know these formulas, especially the one for b_1^{OLS} ❗

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:** $b_0^{OLS} = \bar{y} - b_1\bar{x}$

- **!** You should know these formulas, especially the one for b_1^{OLS} **!**
- 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.

App Time! #3 (3 minutes)

How does OLS actually perform the minimization problem? Some intuition without maths.

```
launchApp('SSR_cone')
aboutApp('SSR_cone')  # after
```

Ordinary Least Squares (OLS): Interpretation

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

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 $x = 0$.**

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.**

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!)

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.

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`

- 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)
```

Class size and student performance

Let's estimate the following model by OLS: $\text{avgmath\_cs}_i = b_0 + b_1 \text{classsize}_i + e_i$

```
# 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
```

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):

$$\hat{y} = b_0 + b_1 x$$
$$\widehat{\text{avgmath\_cs}} = b_0 + b_1 \cdot \text{classize}$$
$$\widehat{\text{avgmath\_cs}} = 61.11 + 0.19 \cdot \text{classize}$$

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):

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

$$\widehat{\text{avgmath\_cs}} = b_0 + b_1 \cdot \text{classize}$$

$$\widehat{\text{avgmath\_cs}} = 61.11 + 0.19 \cdot \text{classize}$$

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

$$\widehat{\text{avgmath\_cs}} = 61.11 + 0.19 \cdot 26$$

$$\widehat{\text{avgmath\_cs}} = 66.08$$

Task 3: OLS Regression (7 minutes)

1. Compute the OLS coefficients using the formulas on slide 28.
2. Regress class size (independant variable) on average verbal score (dependent variable).
3. Is the slope coefficient similar to the one found for average math score? Was it expected based on the graphical evidence?
4. What is the predicted average verbal score when class size is equal to 0? (Does that even make sense?!)
5. What is the predicted average verbal score when the class size is equal to 30 students?

OLS variations / restrictions

- All are described in the book. Optional 🤓.
- There is an app for each of them:

type	App
No Intercept, No regressors	`launchApp('reg_constrained')`
Centered Regression	`launchApp('demeaned_reg')`
Standardized Regression	`launchApp('reg_standardized')`

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 errors is 0.**

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

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 errors 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 errors are uncorelated (by definition).**

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

- **Prediction and errors 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!

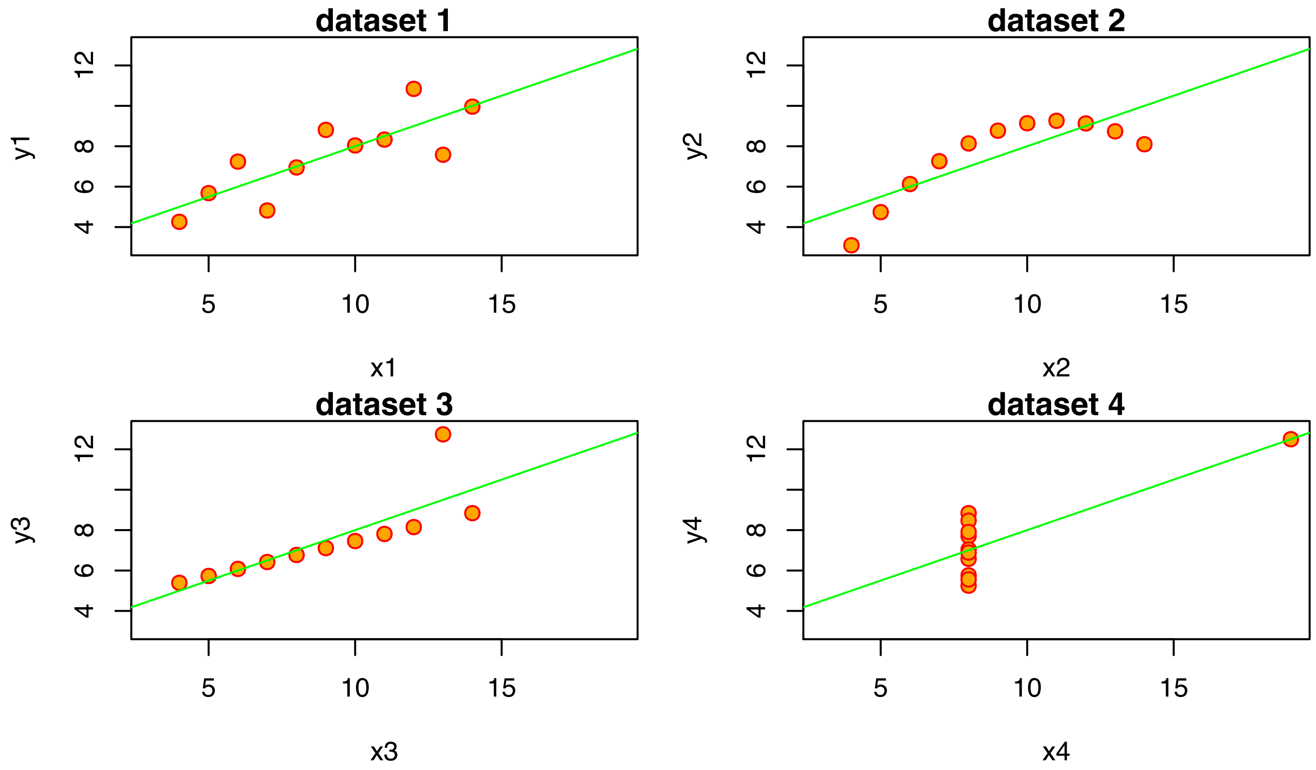
- It's important to keep in mind that covariance, correlation and simple OLS regression only measure **linear relationships** between two variables.
- Two datasets with *identical* correlations and regression lines could look *vastly* different.

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?

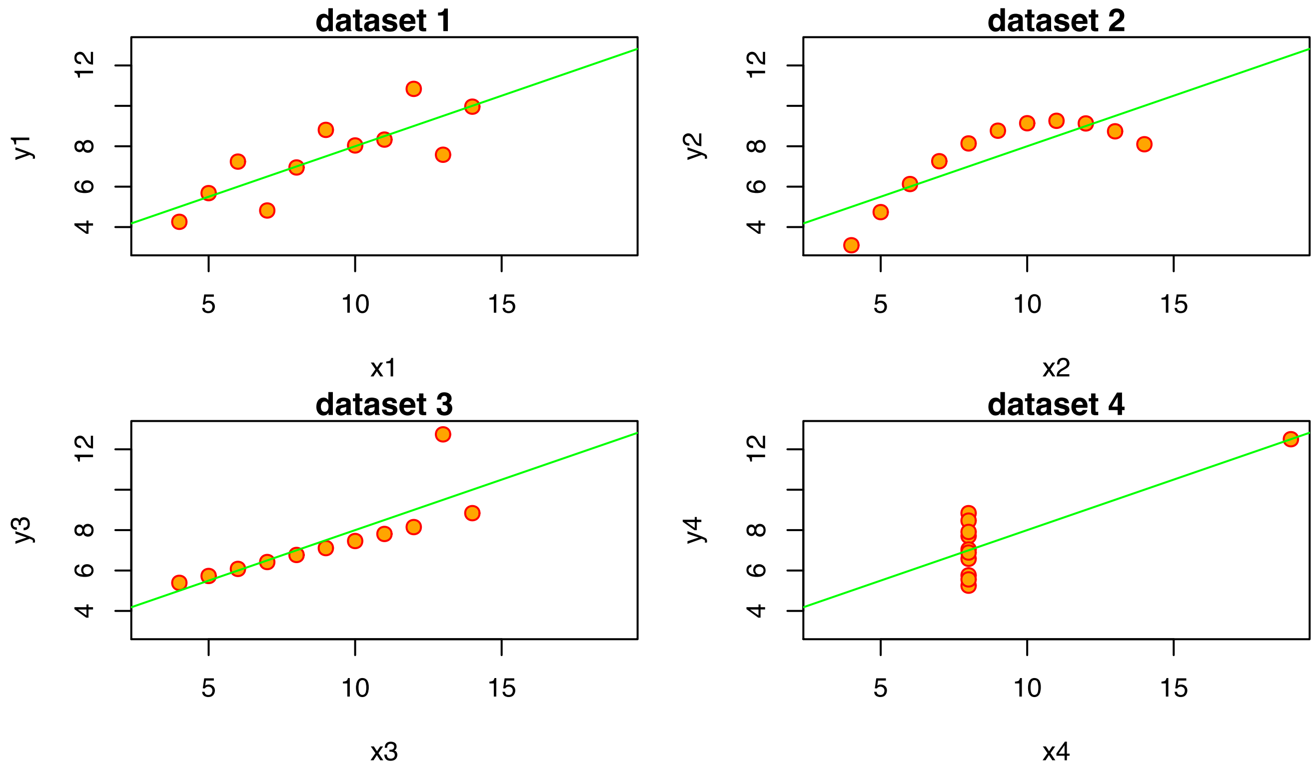
Linearity Assumption: Anscombe

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



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!)

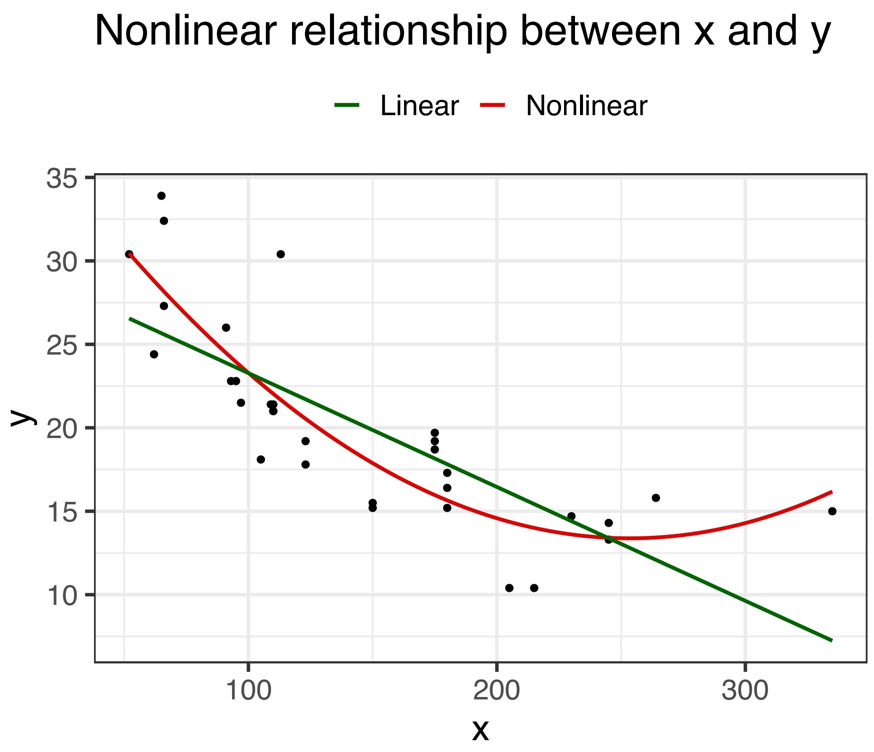
Nonlinear Relationships in Data?

- 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:



Analysis of Variance (ANOVA)

- 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)**
  - SST = sum of squares total
  - SSE = sum of squares error
  - SSE = sum of squares regresion

Goodness of Fit

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

Goodness of Fit

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

$$R^2 = \frac{\text{variance explained}}{\text{total variance}} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \in [0, 1]$$

Goodness of Fit

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

$$R^2 = \frac{\text{variance explained}}{\text{total variance}} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \in [0, 1]$$

- 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.

Goodness of Fit

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

$$R^2 = \frac{\text{variance explained}}{\text{total variance}} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \in [0, 1]$$

- 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.

Goodness of Fit

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

$$R^2 = \frac{\text{variance explained}}{\text{total variance}} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \in [0, 1]$$

- 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.
- ⚠️ Low R^2 does **NOT** mean it's a useless model! Remember that econometrics is interested in causal mechanisms, not prediction!

Task 4: R^2 and goodness of fit (10 minutes)

1. Regress `classize` on `avgmath_cs`. 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 of the relationship between R^2 and the correlation in a regression with only one regressor?

4. Install and load the `broom` package. Pass `math_reg` in the `broom::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 in the previous slide.

5. 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?

SciencesPo
DEPARTMENT OF ECONOMICS

SEE YOU NEXT WEEK!

	michele.fioretti@sciencespo.fr
	Slides
	Book
	@ScPoEcon
	@ScPoEcon