

# Ec140 - Regression as Line Fitting and Conditional Expectation (Part I)

---

Fernando Hoces la Guardia  
07/14/2022

# Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix + [SoPo Econometrics](#). (Part I today)
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79.
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

# Regression Journey

- Regression as Matching on Groups. Ch2 of MM up to page 68 (not included).
- **Regression as Line Fitting and Conditional Expectation. Ch2 of MM, Appendix + SoPo Econometrics. (Part I today)**
- Multiple Regression and Omitted Variable Bias. Ch2 of MM pages 68-79.
- Regression Inference, Binary Variables and Logarithms. Ch2 of MM, Appendix + others.

# Regression as Line Fitting and Conditional Expectation

# Regression as Line Fitting: Today's Goal

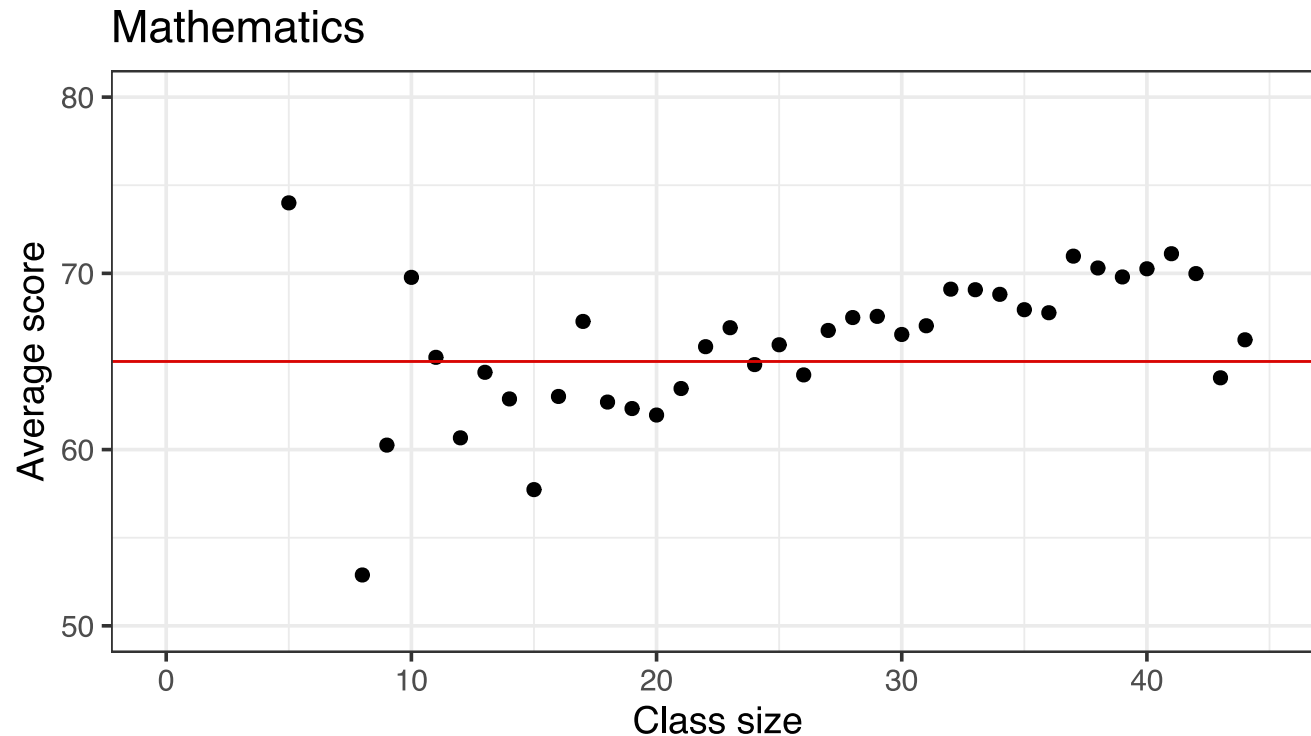
- The goals of today's class are two:
  1. Provide an explanation to what regression does when "it generate fitted values" (or "it fits a line"), and
  2. Provide some insight to a useful formula that represents the main coefficient of interest ( $\beta$ ).
- Today's class will be a bit more technical than previous classes.
- For this reason it is important to always keep in mind what the goal is.
- Even if you end up completely lost about today's material, these explanations are not essential for you to do well in class.

# Regression as Line Fitting

- Example: Class size and student performance (Slides adapted from [SciencePo Econometrics](#) course, and [data from Raj Chetty and Greg Bruich's course](#))

# Class size and student performance: Regression line

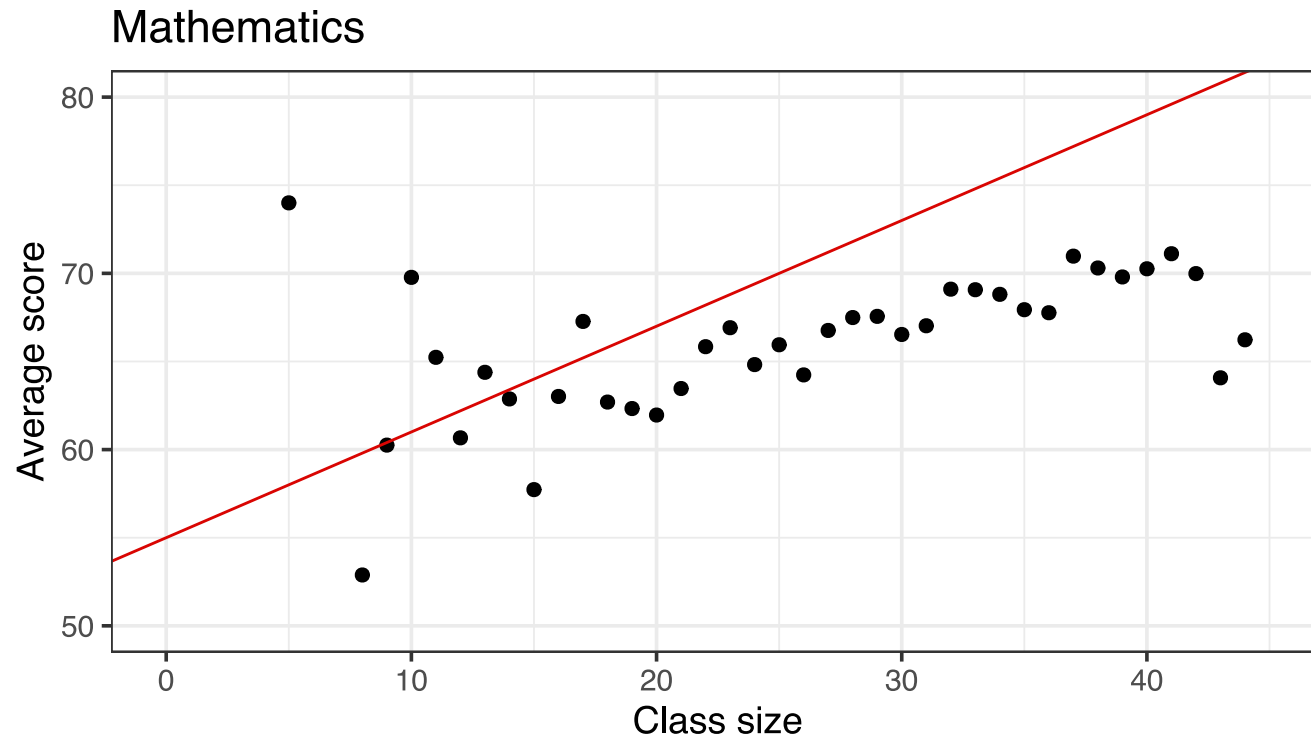
How to visually summarize the relationship: **a line through the scatter plot**



- 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!



# It's All About the Residuals

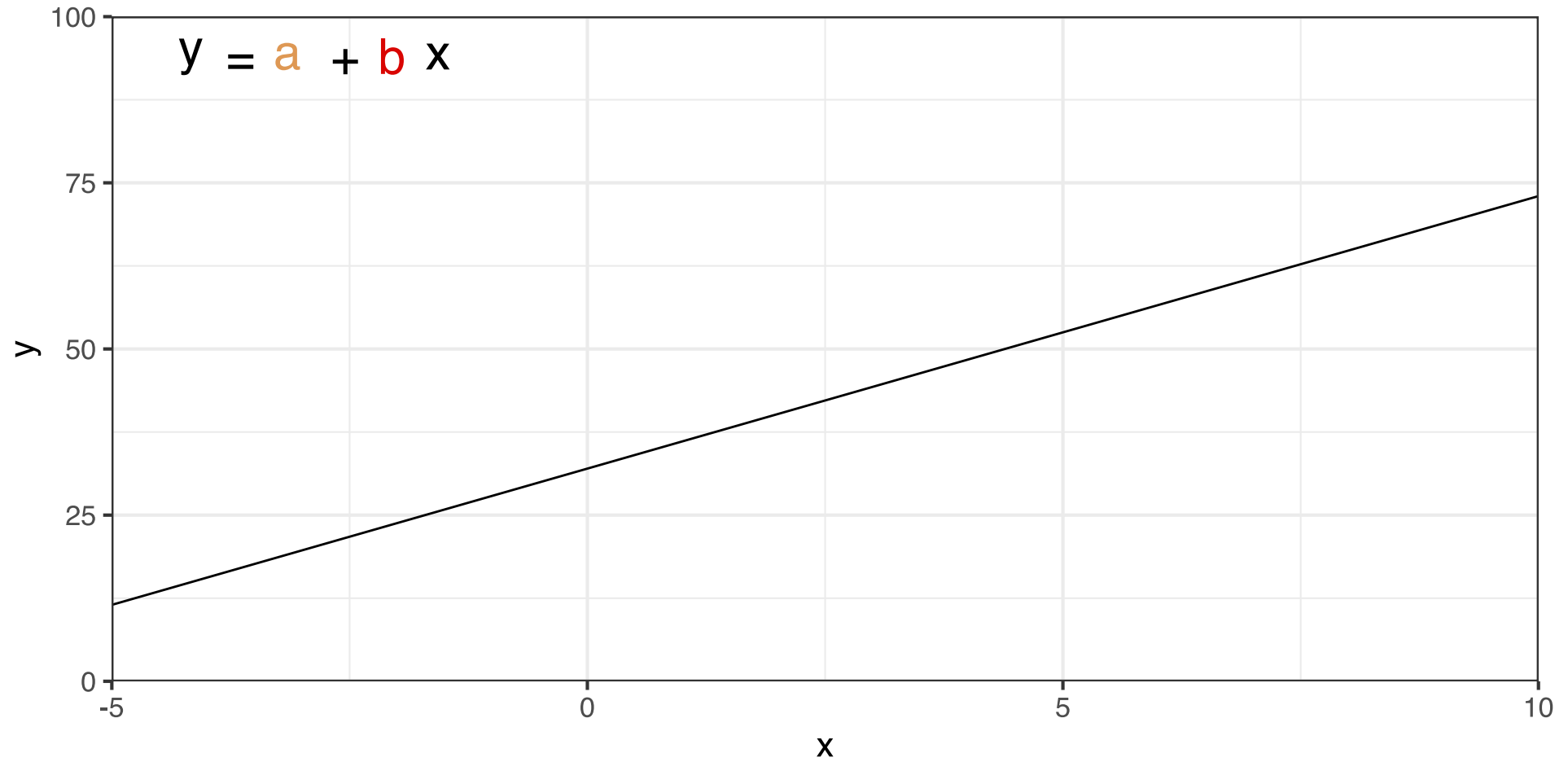
- In *Regression as Matching* we define the residuals,  $e_i$ , as the difference between the observed ( $Y_i$ ) and fitted values ( $\hat{Y}_i$ ).

$$e_i \equiv Y_i - \hat{Y}_i$$

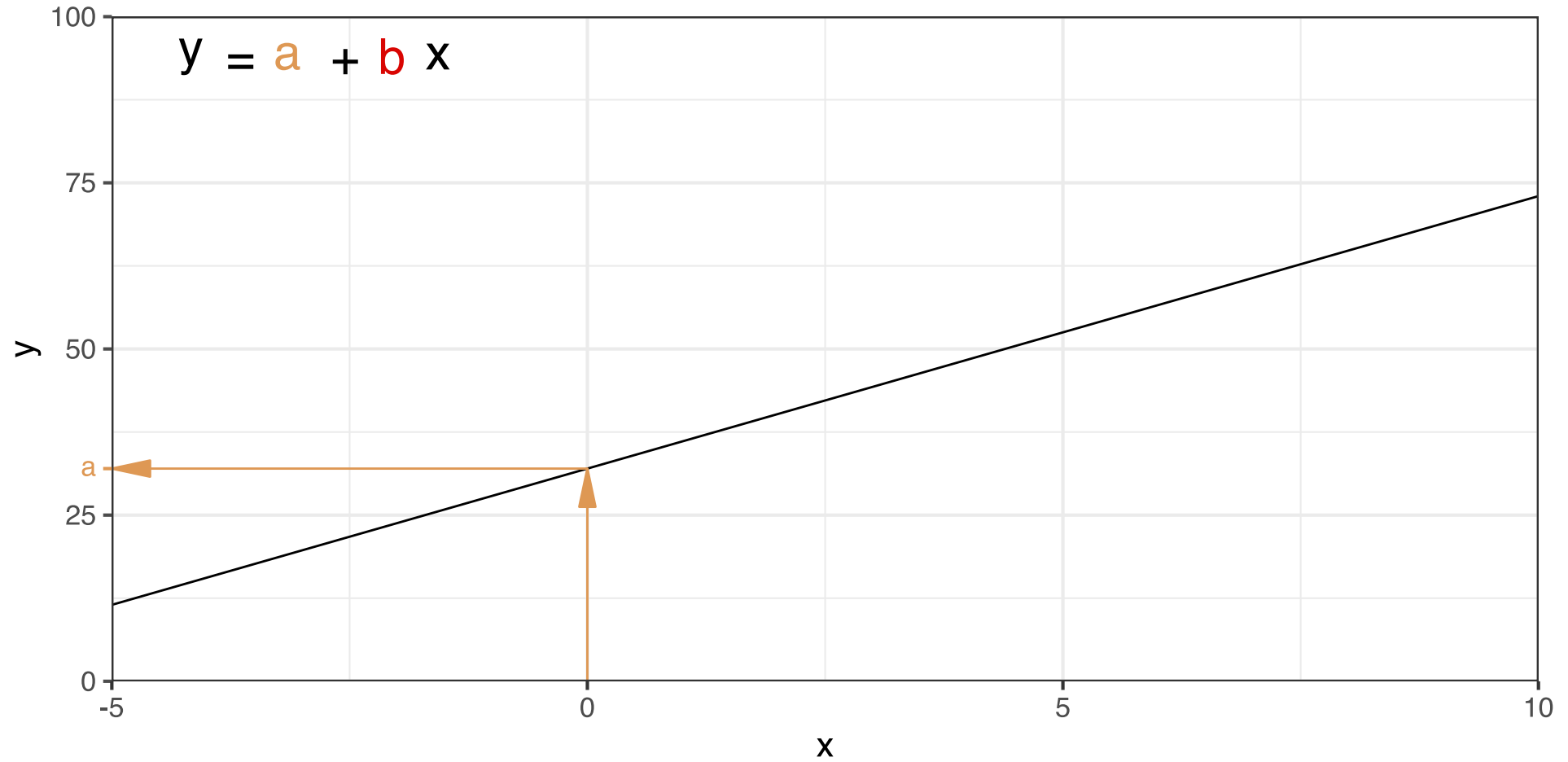
- By fitted values, we mean a line (for now) that summarizes the relationship between  $X$  and  $Y$ .
- The equation for such a line with an intercept  $a$  and a slope  $b$  is:

$$\hat{Y}_i = a + bX_i$$

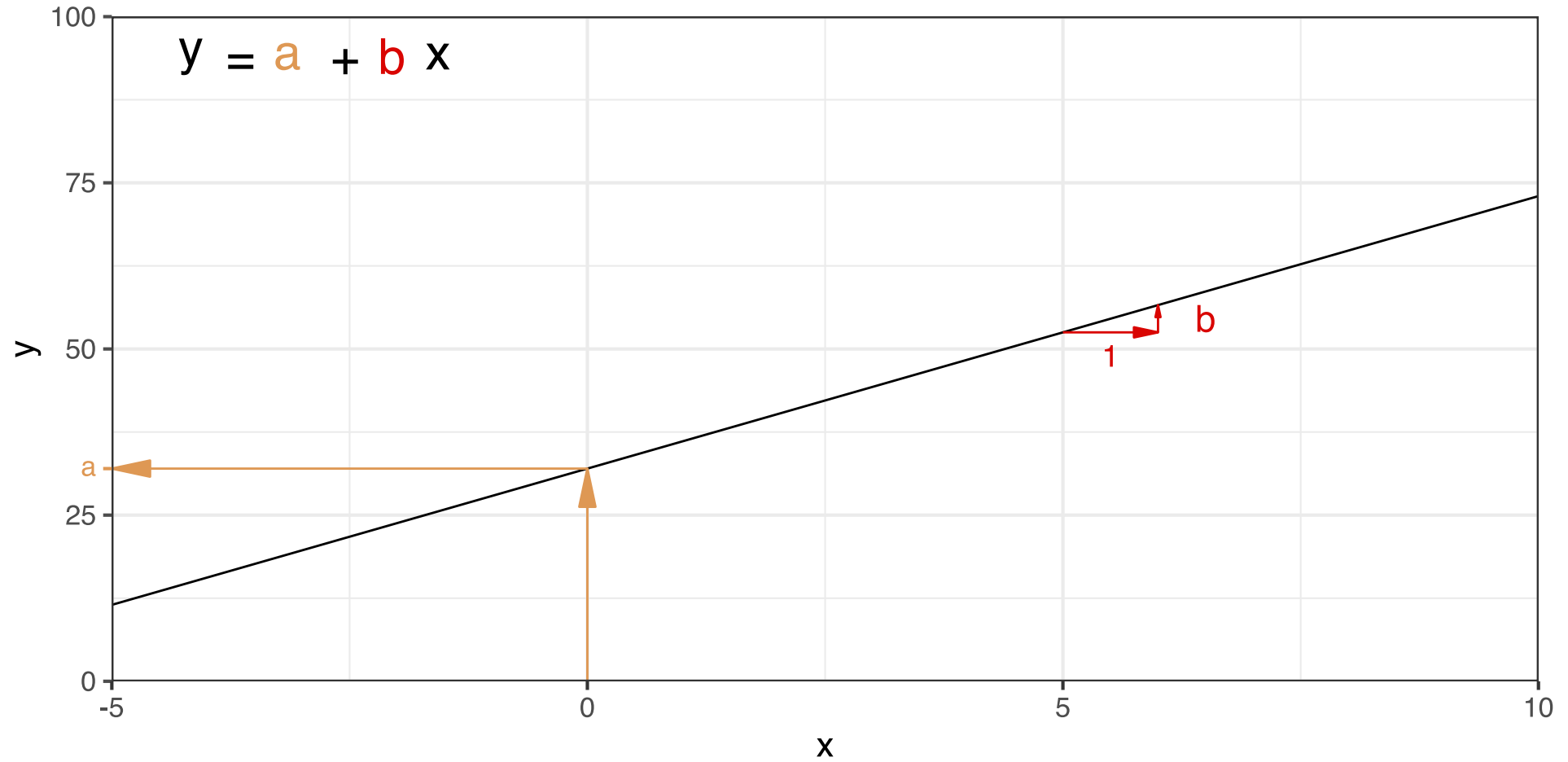
# What's A Line: A Refresher



# What's A Line: A Refresher

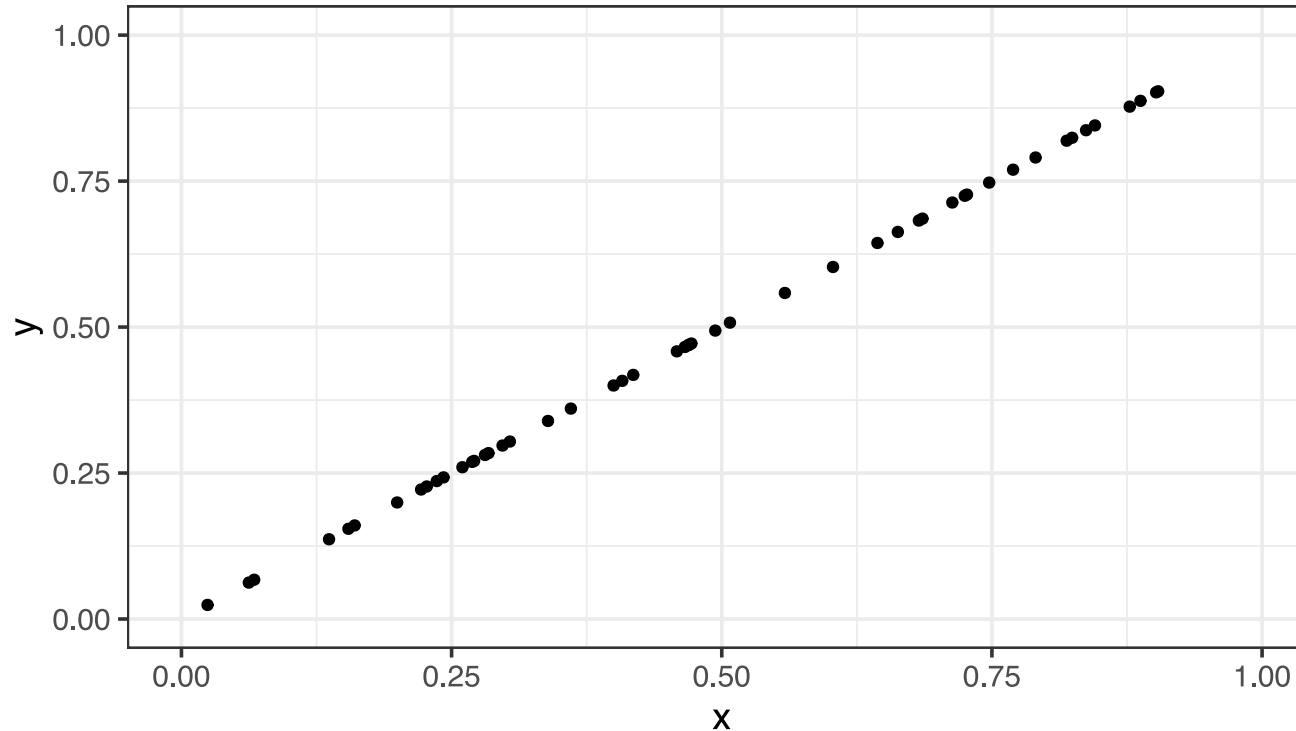


# What's A Line: A Refresher



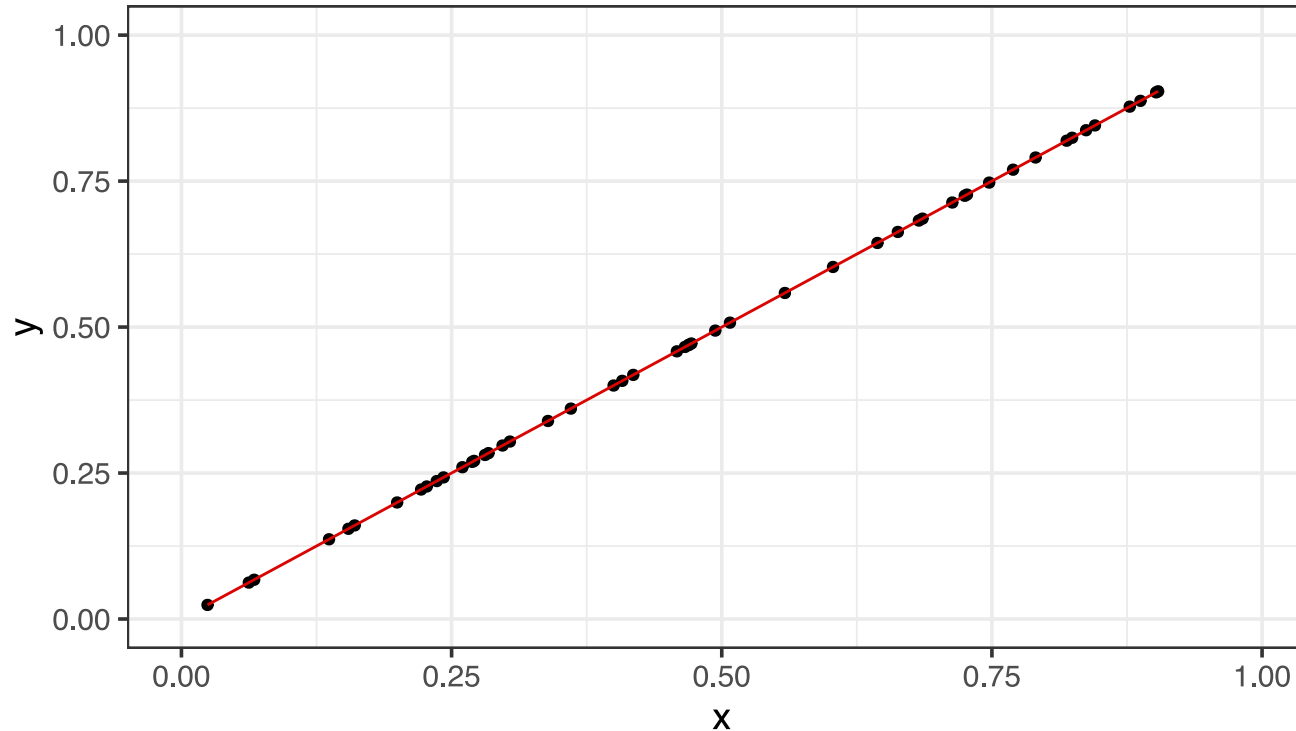
# Simple Linear Regression: Residual

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

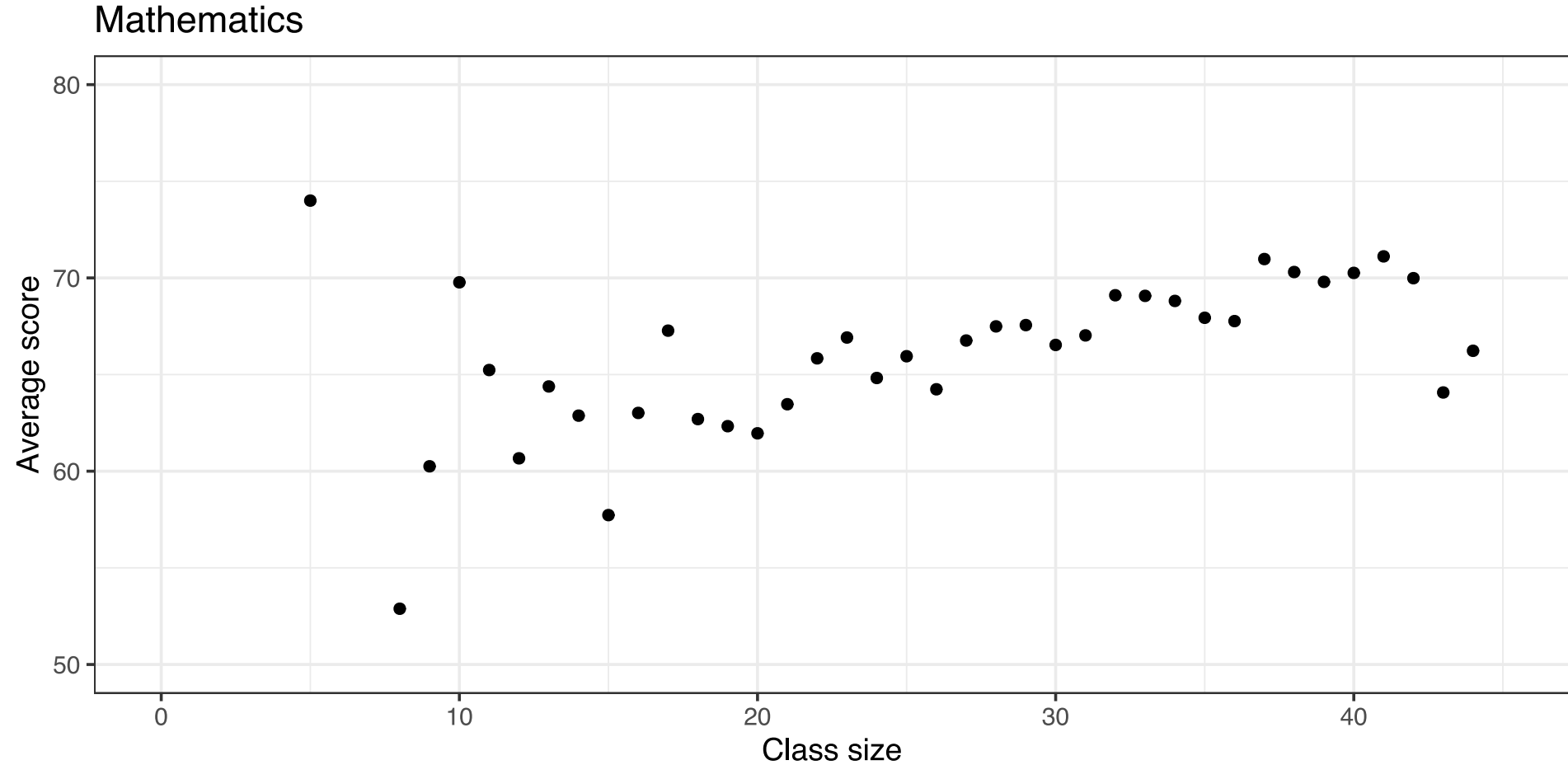


# Simple Linear Regression: Residual

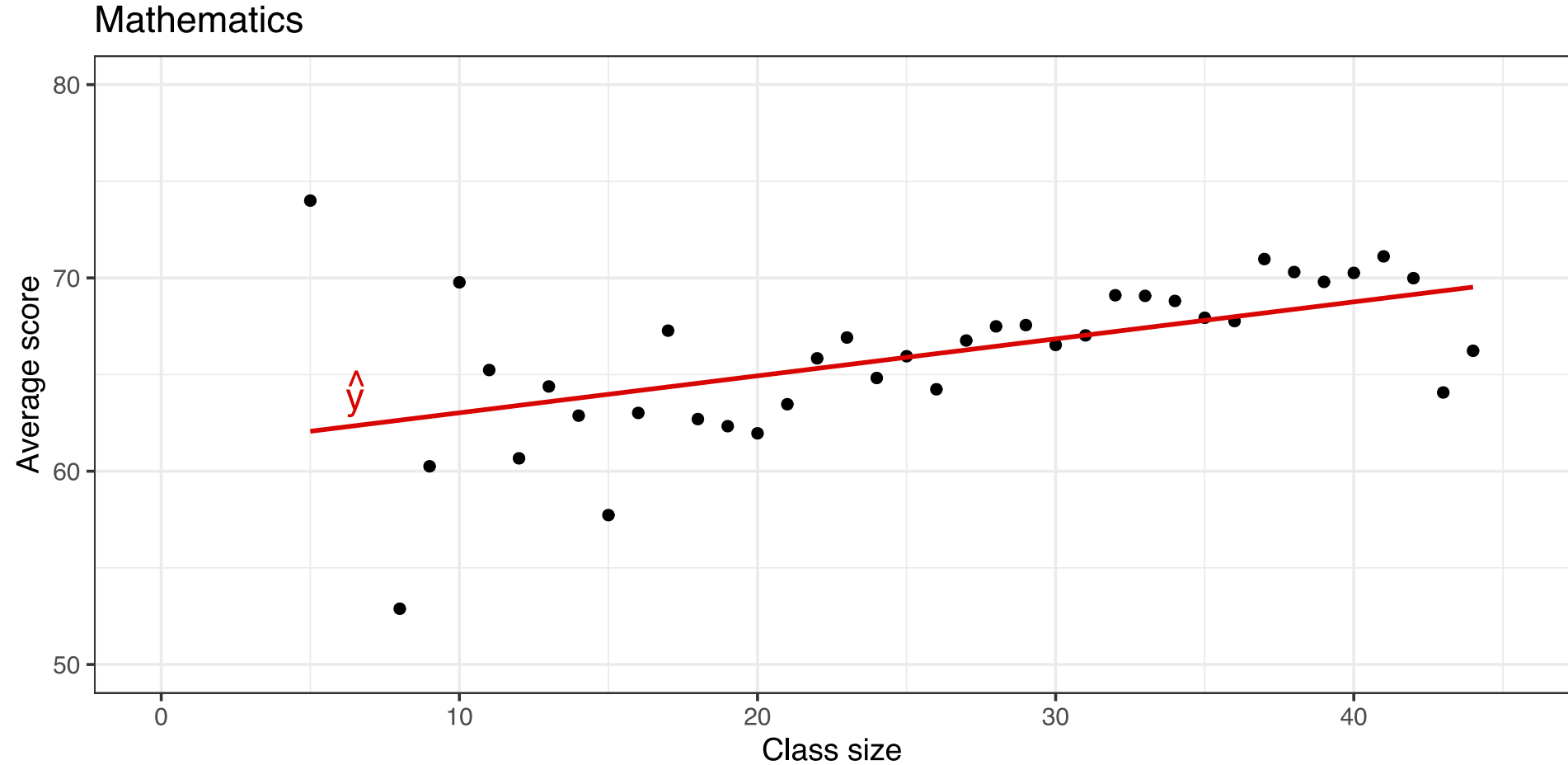
- If all the data points were **on** the line then  $\hat{Y}_i = Y_i$ .



# Simple Linear Regression: Graphically

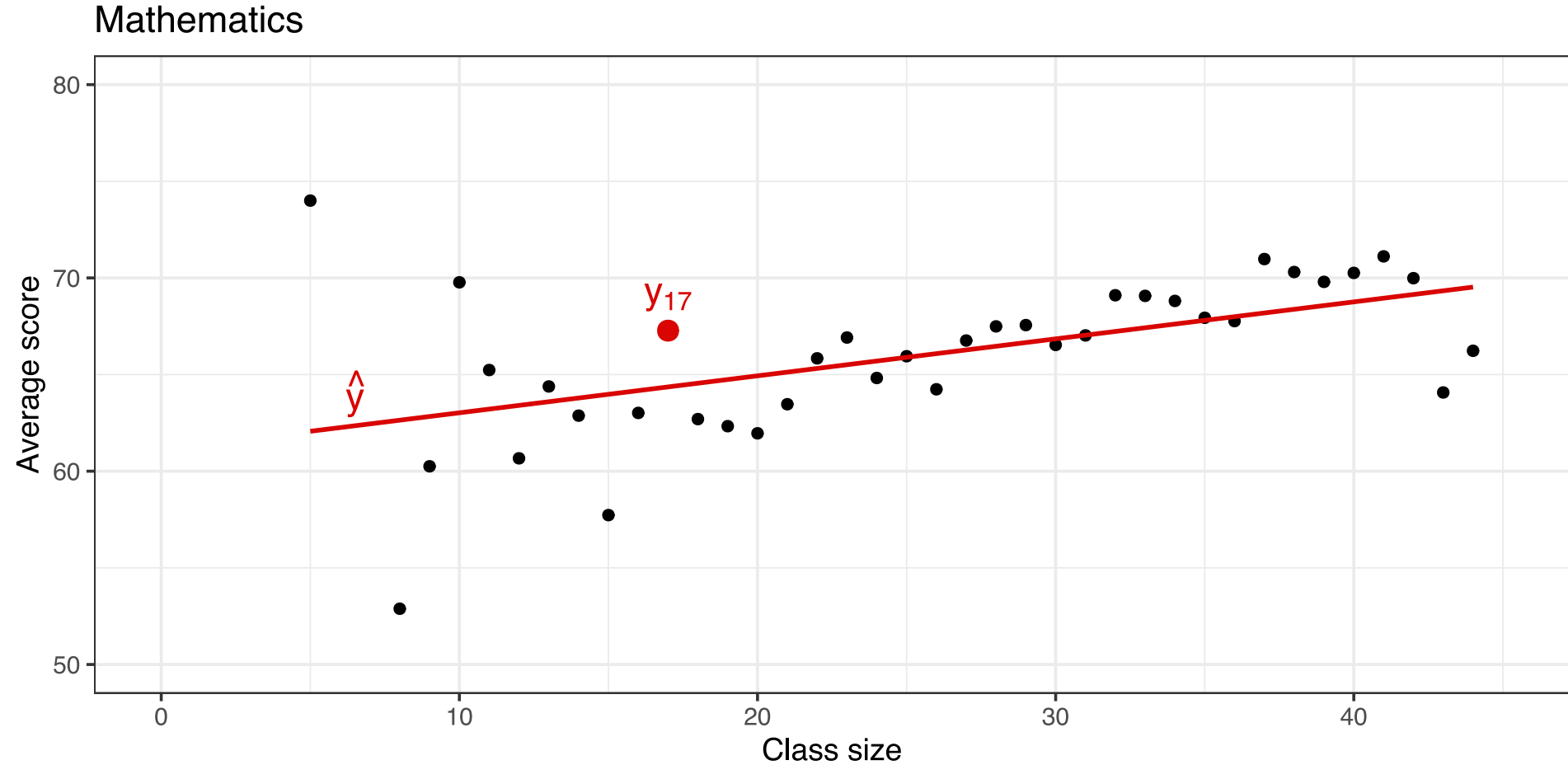


# Simple Linear Regression: Graphically

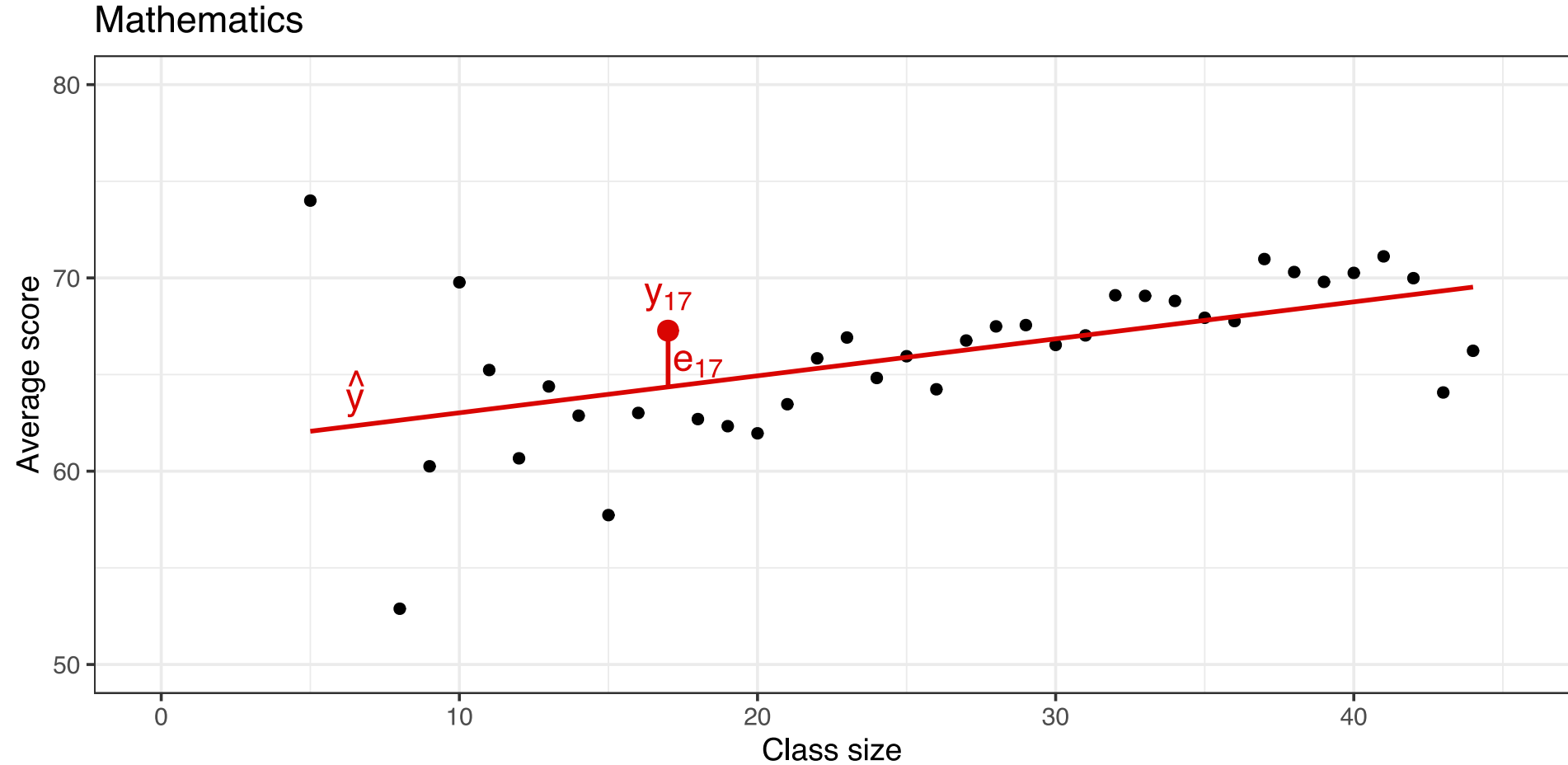




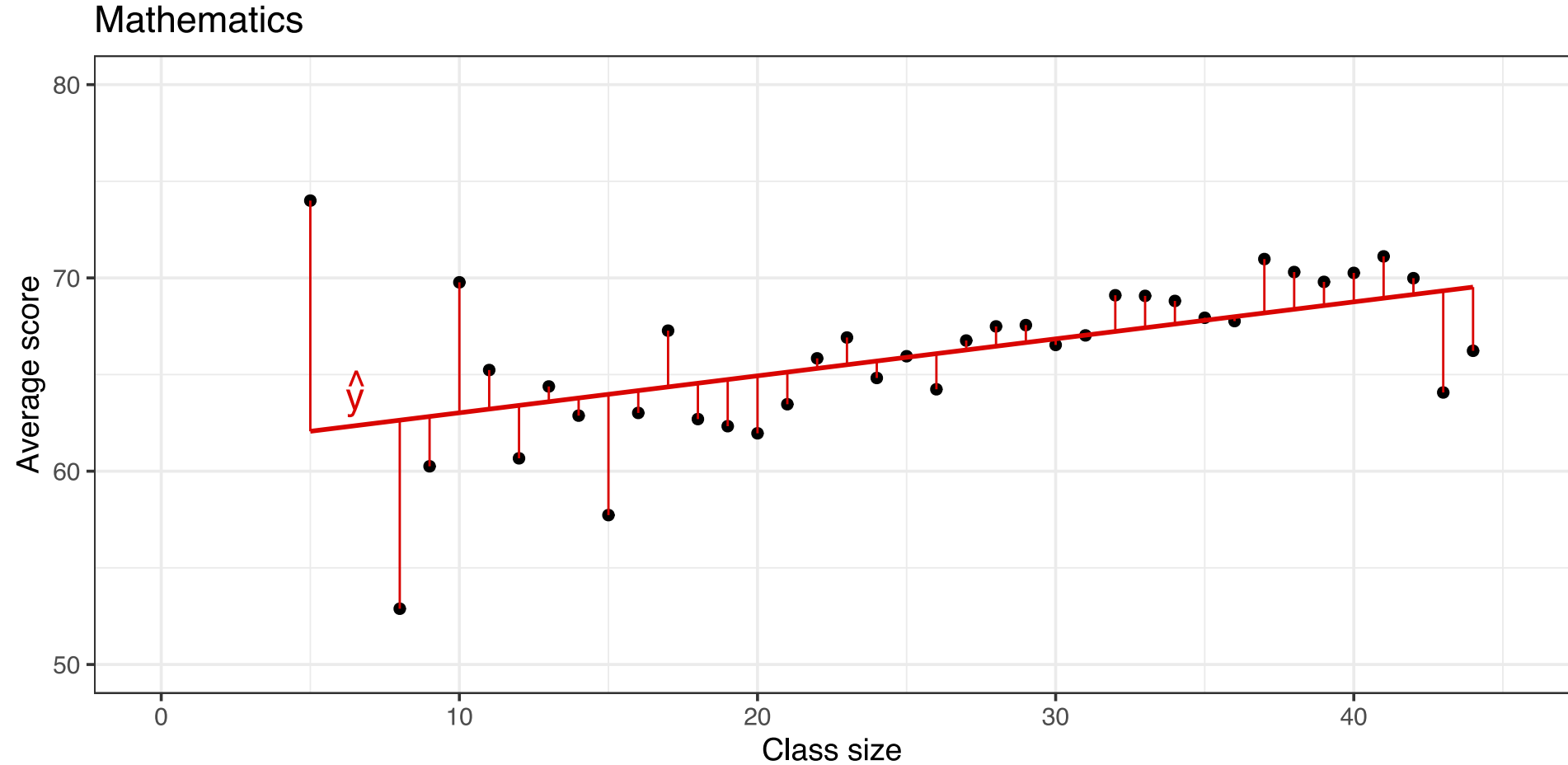
# Simple Linear Regression: Graphically



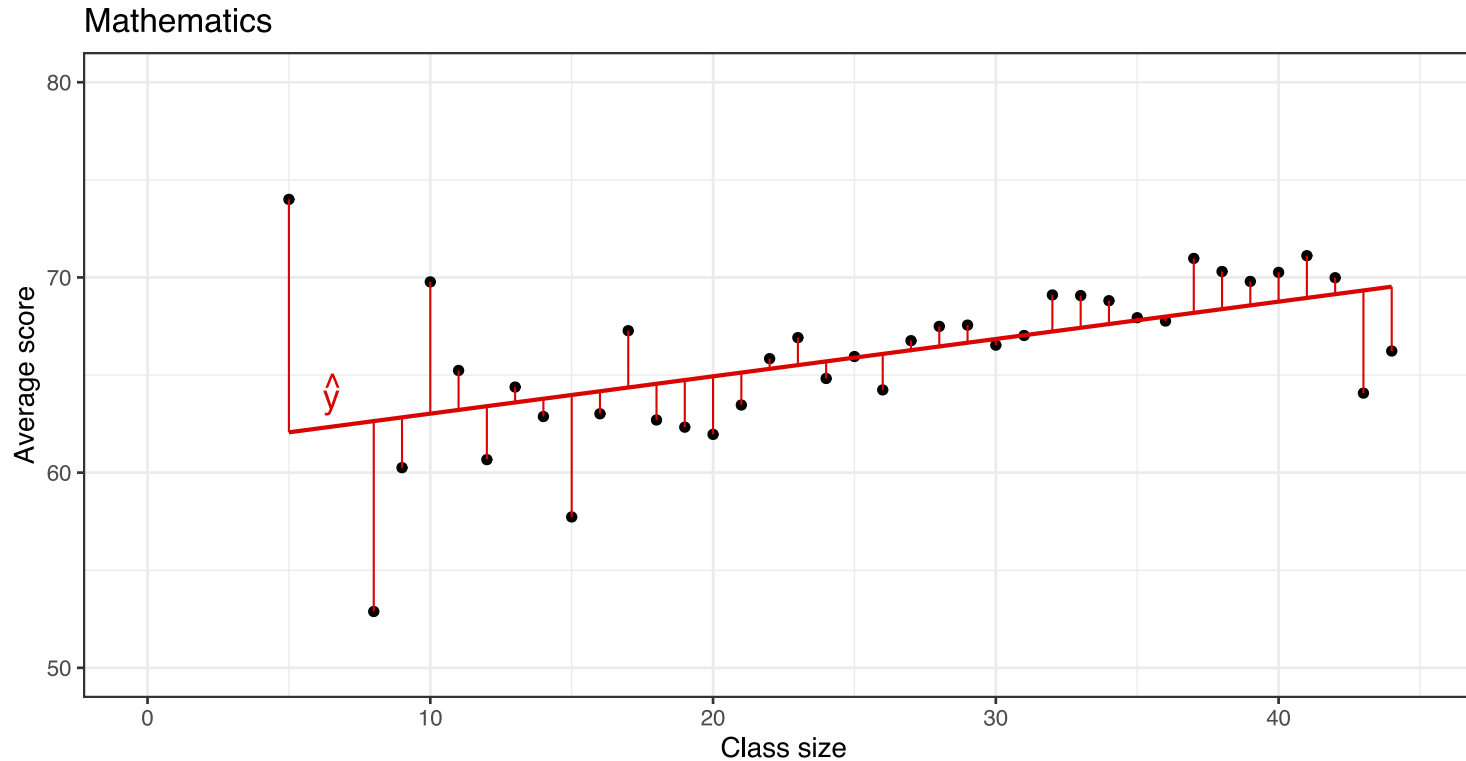
# Simple Linear Regression: Graphically



# Simple Linear Regression: Graphically



# Simple Linear Regression: Graphically



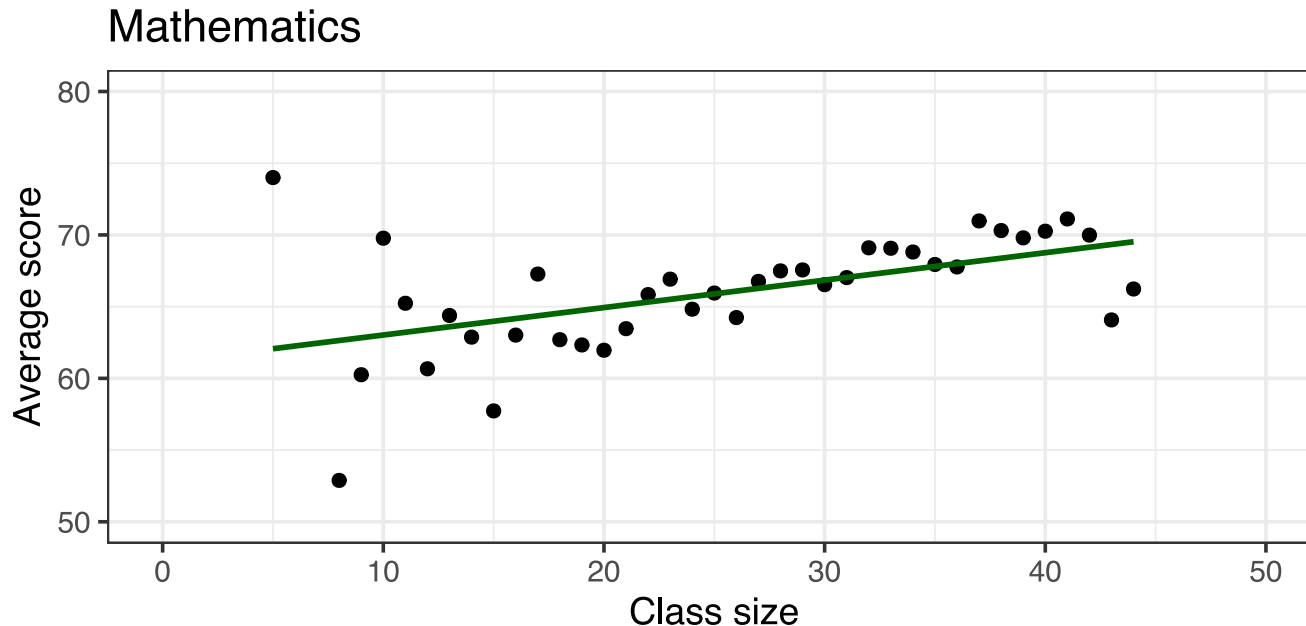
**Which "minimisation" criterion should (can) be used?**

# Ordinary Least Squares (OLS) Estimation

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

$$e_i^2 = (Y_i - \hat{Y}_i)^2 = (Y_i - a - bX_i)^2$$

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

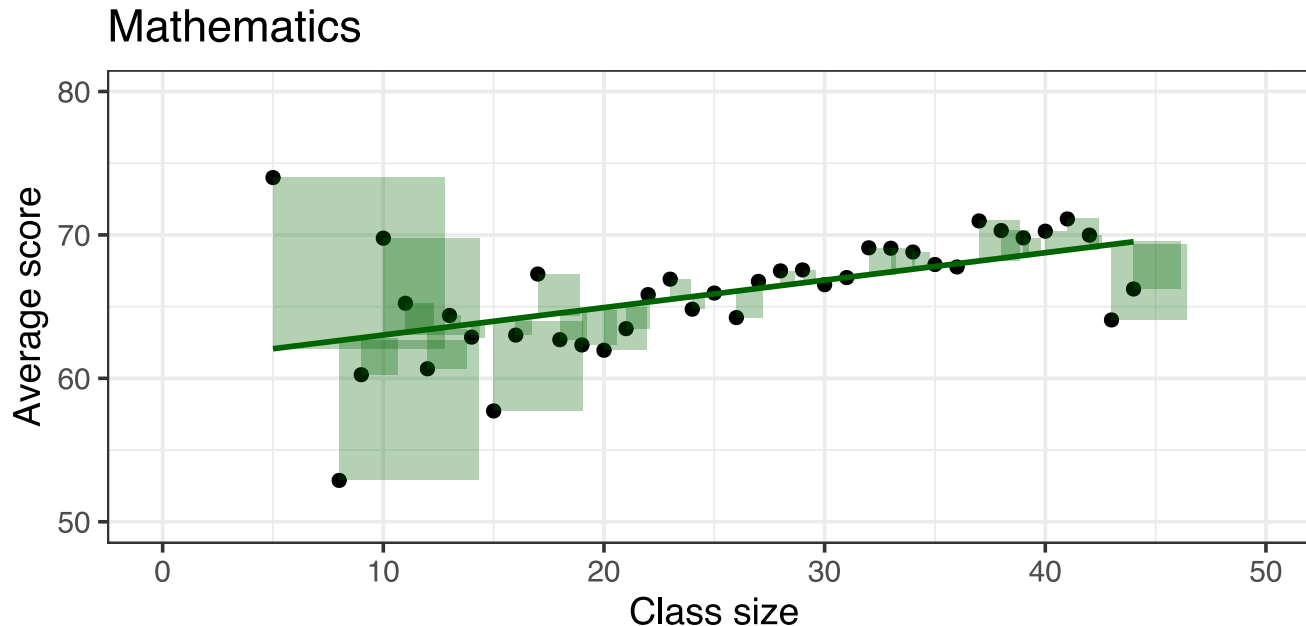


# Ordinary Least Squares (OLS) Estimation

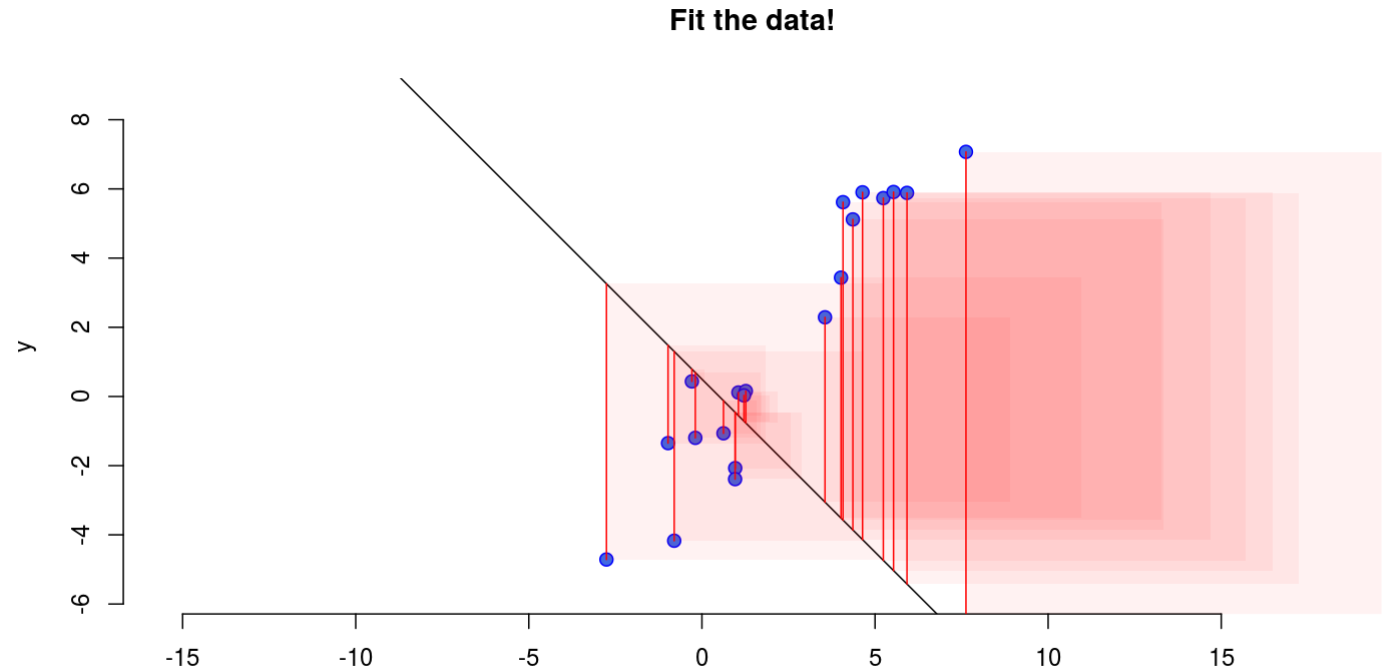
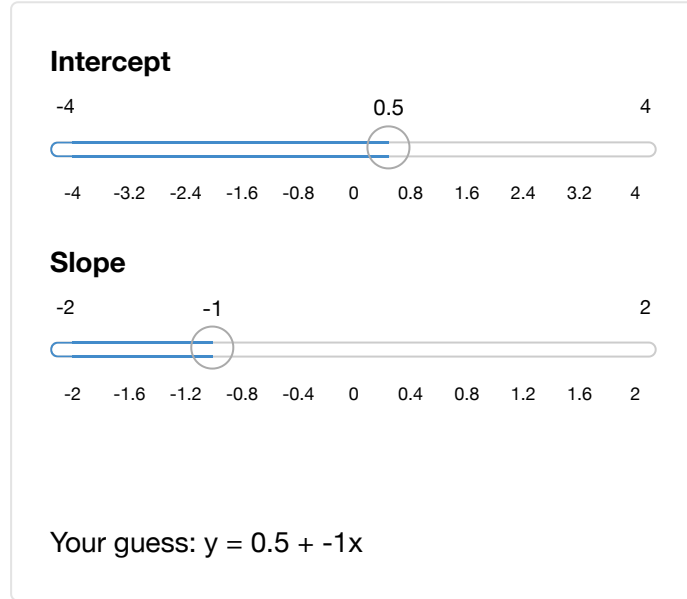
- Errors of different sign (+/−) cancel out, so we consider **squared residuals**

$$e_i^2 = (Y_i - \hat{Y}_i)^2 = (Y_i - a - bX_i)^2$$

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



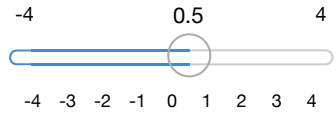
# Ordinary Least Squares (OLS) Estimation



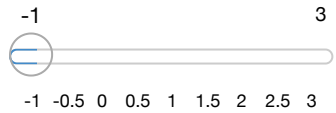
Link

# Ordinary Least Squares (OLS) Estimation

**Intercept**



**Slope**



[Link](#)



# Covariance: Brief Explainer 1/2

- The covariance is a measure of co-movement between two random variables  $(X_i, Y_i)$ :

$$\text{Cov}(X_i, Y_i) = \sigma_{XY} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])]$$

- With its sample counterpart (for the case of equally likely observations):

$$\hat{\sigma}_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- If either formula looks weird, think of the variance, as the covariance between  $X_i$  and itself ( $X_i$ ) and the above should look more familiar:

$$\sigma_{XX} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_i - \mathbb{E}[X_i])] = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = \sigma_X^2$$

# Covariance: Brief Explainer 2/2

In addition to  $\sigma_{XX} = \sigma_X^2$ , we might use two other properties of the covariance:

- If the expectation of either  $X_i$  or  $Y_i$  is 0, the covariance between them is the expectation of their product:  $Cov(X_i, Y_i) = E(X_i Y_i)$
- The covariance linear functions of variables  $X_i$  and  $Y_i$  -- written as  $W_i = c_1 + c_2 X_i$  and  $Z_i = c_3 + c_4 Y_i$  for constants  $c_1, c_2, c_3, c_4$  -- is given by:

$$Cov(W_i, Z_i) = c_2 c_4 Cov(X_i, Y_i)$$

- You are not asked to memorize any of these formulas. Just used them to understand many concepts in regression.

# Ordinary Least Squares (OLS): Coefficient Formulas 1/4

- **OLS**: *estimation* method consisting in choosing  $a$  and  $b$  to minimize the sum of squared residuals.
- In the case of one regressor (and a constant), the result of this minimization generates the following formulas: (derivation [in this video](#) and [these slides](#)).
- So what are the formulas for  $a$  (intercept) and  $b$  (slope)?
- We can solve this problem for the population or for random sample.
- Warning: the next 3 slides are heavy on notation. If you lose track, the main takeaway is that we want an intuitive formula for the solution to this problem.

# Ordinary Least Squares (OLS): Coefficient Formulas 2/4

## Population

Problem to solve:

$$\arg \min_{a,b} \left\{ \mathbb{E}[(Y_i - a - bX_i)^2] \right\}$$

Solution:

$$b = \beta = \frac{\mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])]}{\mathbb{E}[(X_i - \mathbb{E}[X_i])^2]}$$

$$a = \alpha = \mathbb{E}[Y_i] - b \mathbb{E}[X_i]$$

## Sample

Problem to solve:

$$\arg \min_{a,b} \left\{ \sum (Y_i - a - bX_i)^2 \right\}$$

Solution:

$$b = \hat{\beta} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$

$$a = \hat{\alpha} = \bar{Y} - b\bar{X}$$

- Let's bring the concept of Covariance to make this formulas more intuitive

# Ordinary Least Squares (OLS): Coefficient Formulas 3/4

## Population

$$b = \beta = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

$$a = \alpha = \mathbb{E}[Y_i] - b \mathbb{E}[X_i]$$

## Sample

$$b = \hat{\beta} = \frac{\frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{n}}{\frac{\sum (X_i - \bar{X})^2}{n}}$$

$$a = \hat{\alpha} = \bar{Y} - b\bar{X}$$

# Ordinary Least Squares (OLS): Coefficient Formulas 3/4

## Population

$$b = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} = \frac{\sigma_{XY}}{\sigma_X^2}$$

$$a = \alpha = \mathbb{E}[Y_i] - b \mathbb{E}[X_i]$$

## Sample

$$b = \hat{\beta} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

$$a = \hat{\alpha} = \bar{Y} - b\bar{X}$$

# Ordinary Least Squares (OLS): Coefficient Formulas 4/4

- The main takeaway:

$$b = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

# Properties of Residuals 1/2

- As we saw at the beginning of this class, in a regression the observed outcome ( $Y_i$ ) can be separated into a component "explained" by the regression equation (aka model) and a residual component:

$$Y_i = \underbrace{\hat{Y}_i}_{\text{fitted values (explained)}} + \underbrace{e_i}_{\text{residuals}}$$

- Two important properties of the residuals:
  - They have expectation 0.  $E(e_i) = 0$
  - They are uncorrelated with all the regressors that made them and with the corresponding fitted values. For each regressor  $X_{ki}$ :  
 $E[X_{ki}e_i] = 0$  and  $E[\hat{Y}_ie_i] = 0$



# Properties of Residuals 2/2

- We take these properties as given in this course (they come from the calculus of the minimization problem).
- One important point is that these properties are true always (regardless of biased coefficients).
- This does not imply however that we have solved the problem of selection bias.
- In the traditional way of teaching econometrics these two concepts are mixed (hence required a distinction between residuals ( $e_i$ ) and unobservables ( $u_i$ )).

# (OLS with R)

- In R, OLS regressions are estimated using the `lm` function.
- This is how it works:

```
lm(formula = dependent variable ~ independent variable)
```

Let's estimate the following model by OLS:  $\text{average math score}_i = a + b\text{class size}_i + e_i$

```
# OLS regression of class size on average maths  
lm(avgmth_cs ~ classsize, grades_avg_cs)
```

```
#>  
#> Call:  
#> lm(formula = avgmth_cs ~ classsize, data = grades_avg_cs)  
#>  
#> Coefficients:  
#> (Intercept)      classsize  
#>    61.1092         0.1913
```

# Acknowledgments

- Ed Rubin's Undergraduate Econometrics II
- ScPoEconometrics
- MM