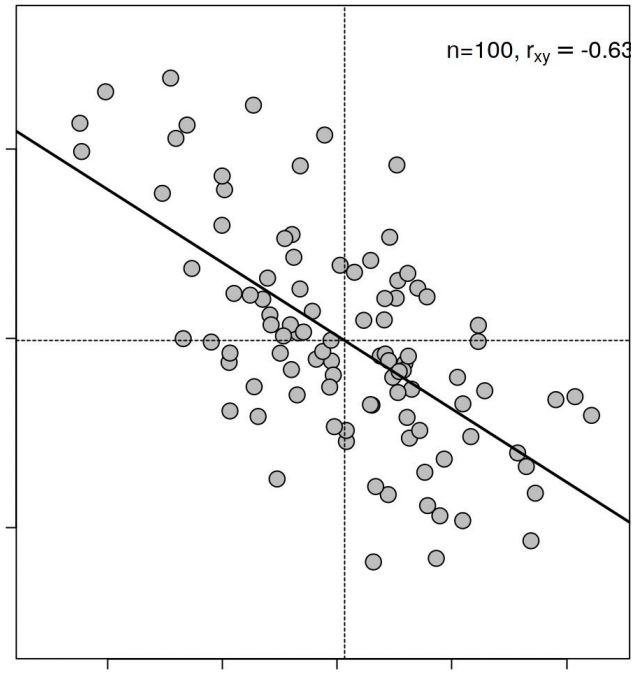
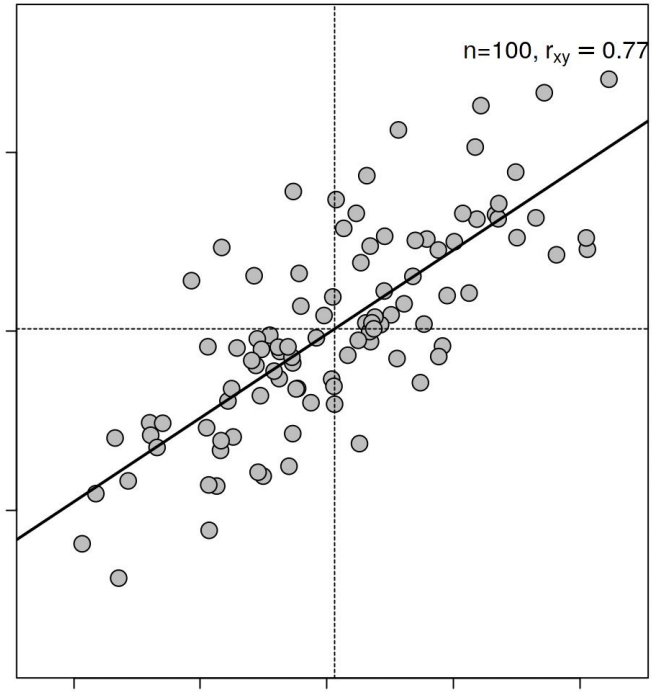
Fundamentals of Econometrics Models



Vicenç Soler
v.soler@tbs-education.org
~~vincent.soler@tbs-education.org~~



Quick reminder

What are linear regressions useful for?

They allow to **explain how a variable changes in relation to another variable**... And from there, derive **predictive models**.

Quick reminder

What is a dependent variable?

It is the variable that we try to **explain or predict**. Usually, denoted by the letter Y.

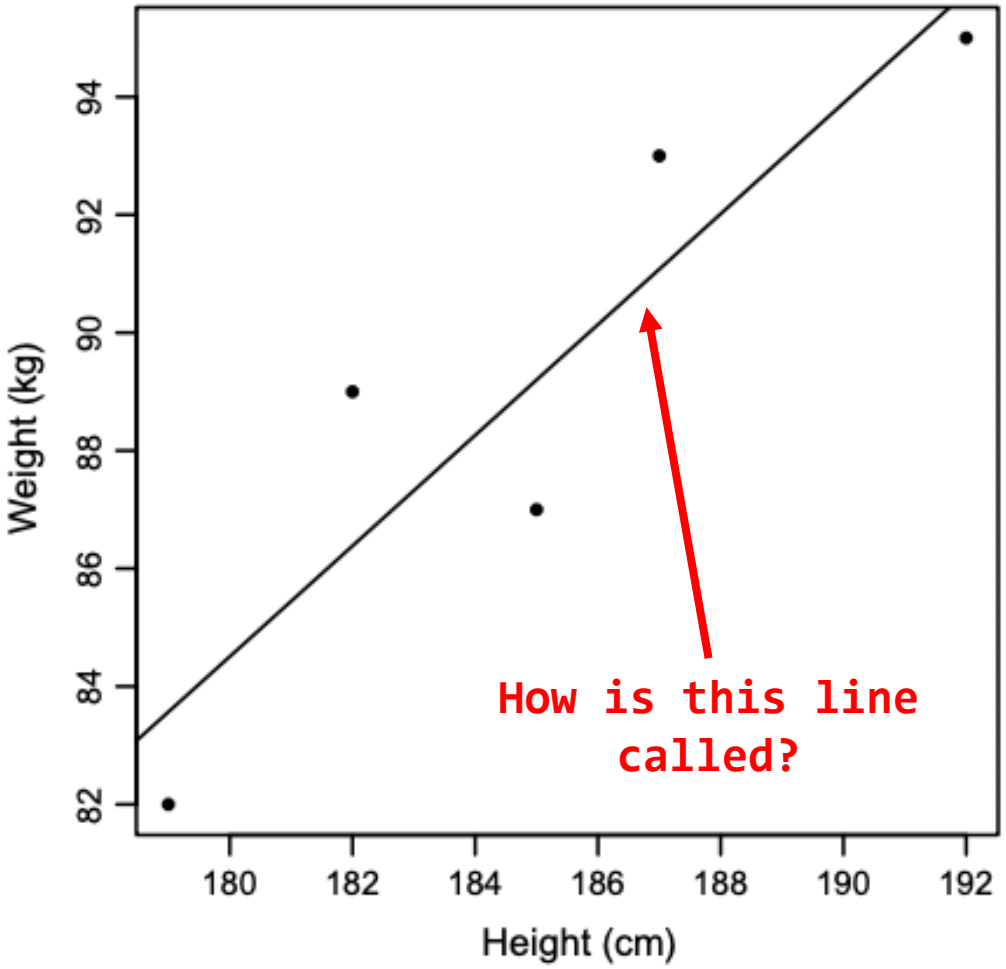
Quick reminder

What is an independent variable?

It is the variable that we use to **explain or predict** our dependent variable. Usually, denoted by the letter X.

Quick reminder

Regression line



Quick reminder

In this linear equation…

$$y = a + \beta x$$

$$Weight = -83.47 + 0.94 * Height$$

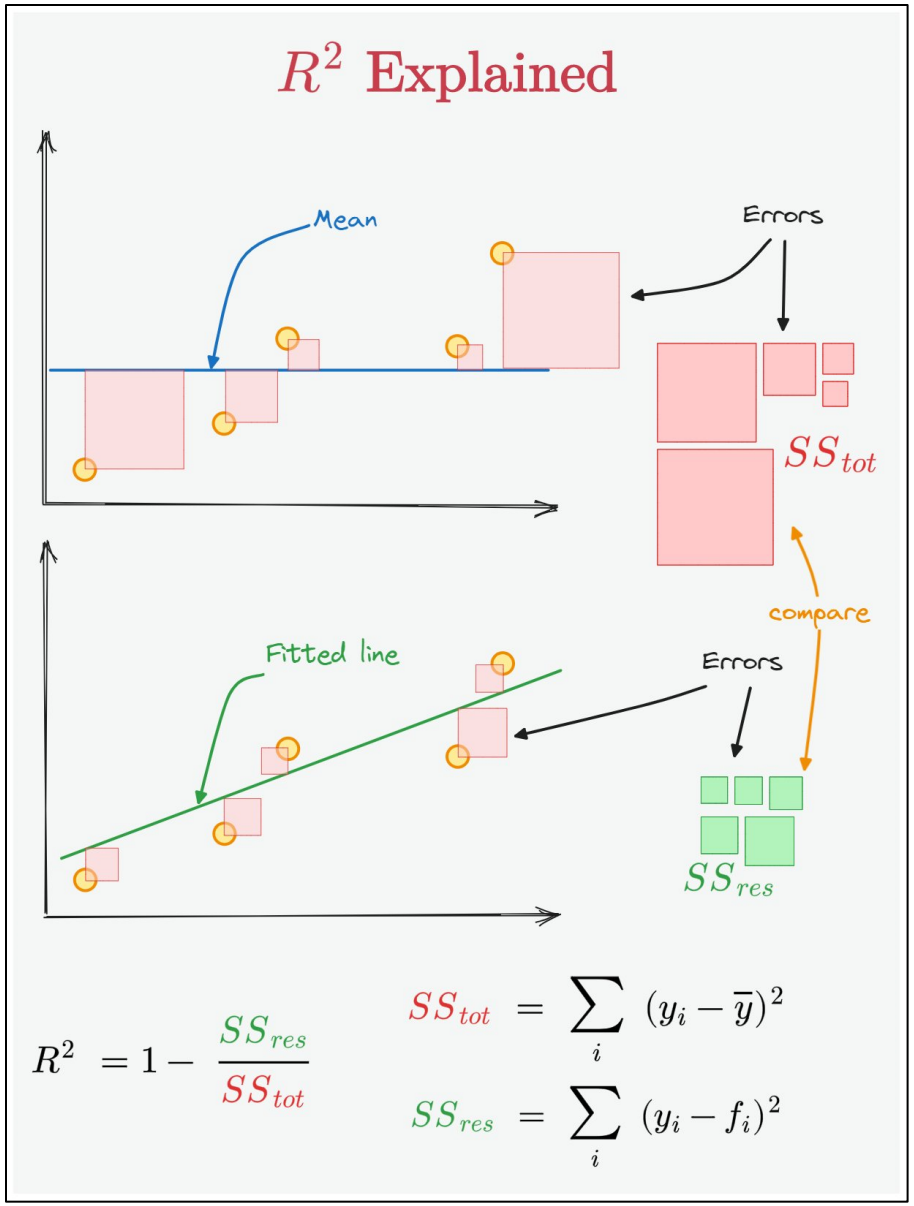
What is the slope? And the intercept?

Slope = 0.93 (or β) and Intercept = -83.47 (or a)

The essentials of regression

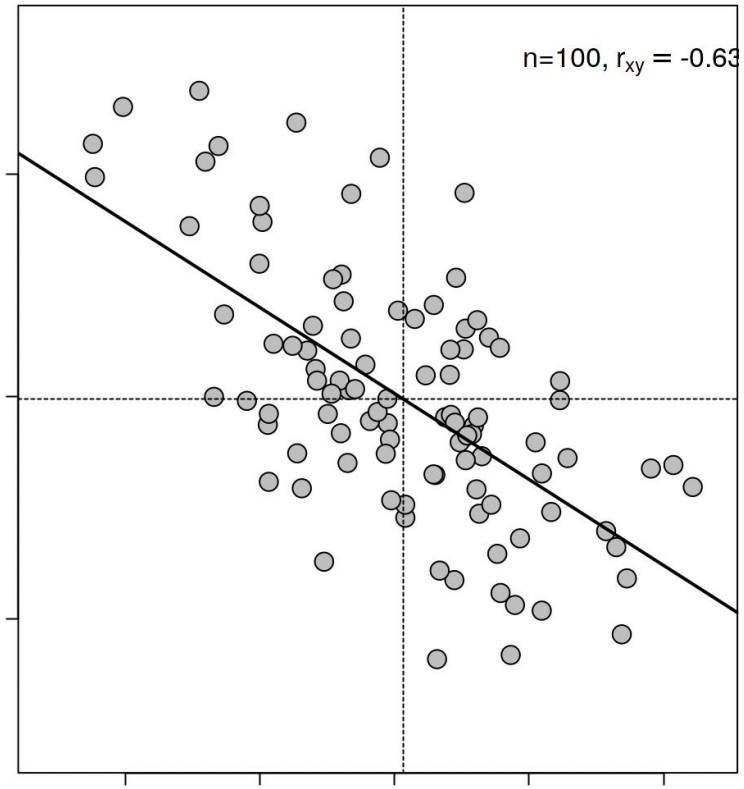
To measure the fitness between our predicted and the observed data we use

CORRELATION (explained visually)

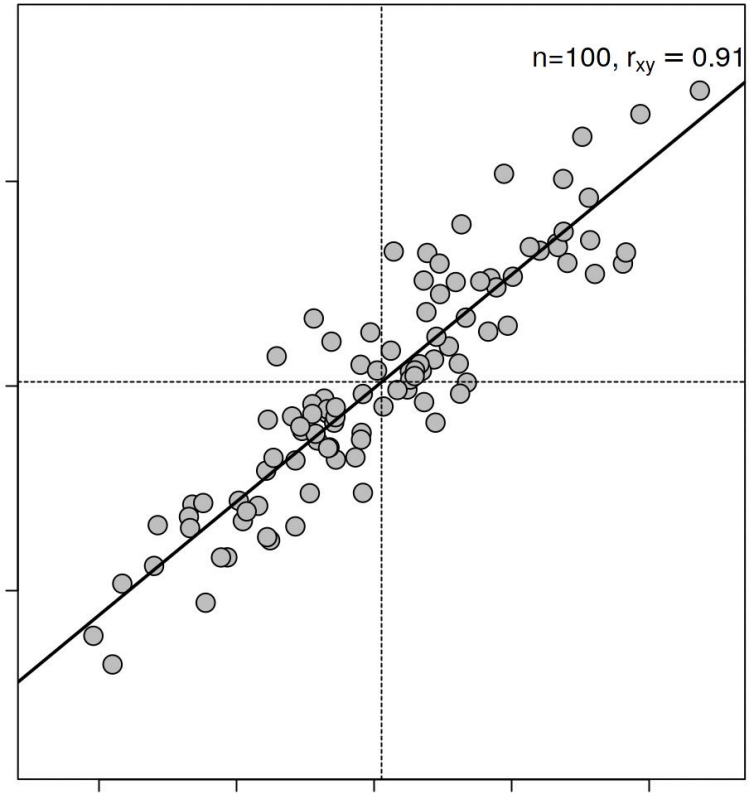


Quick reminder

Which linear regression has a smaller error?

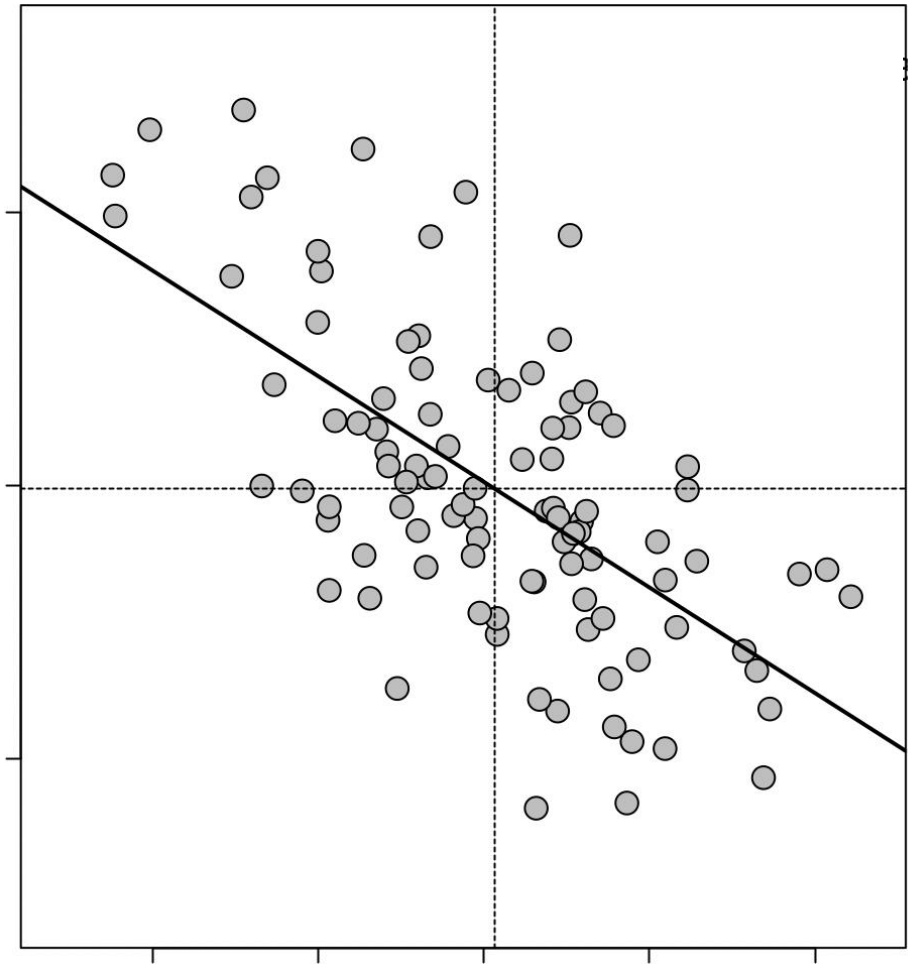


Bigger error



Smaller error

Quick reminder



Is there a correlation?

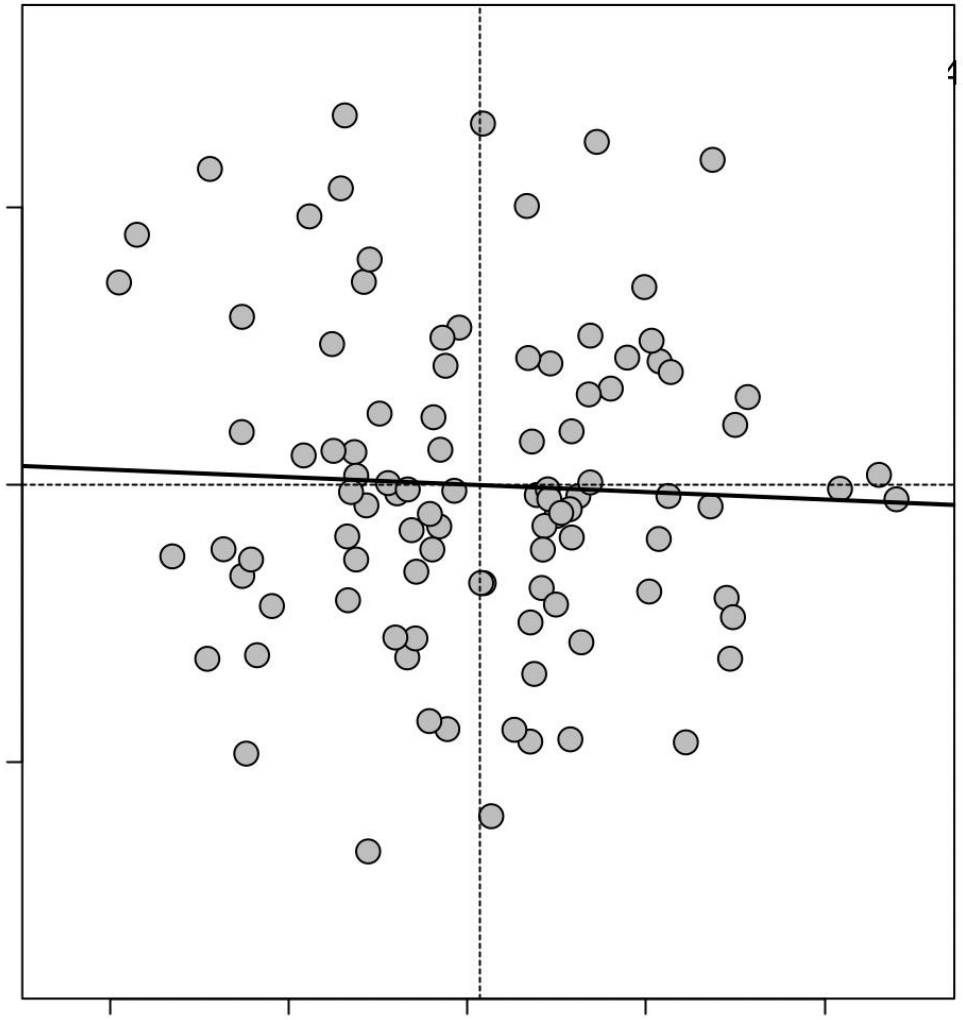
If so, positive or negative?

Weak or strong?

Can you guess the correlation coefficient (between -1 and 1)?

R = -0.63

Quick reminder



Is there a correlation?

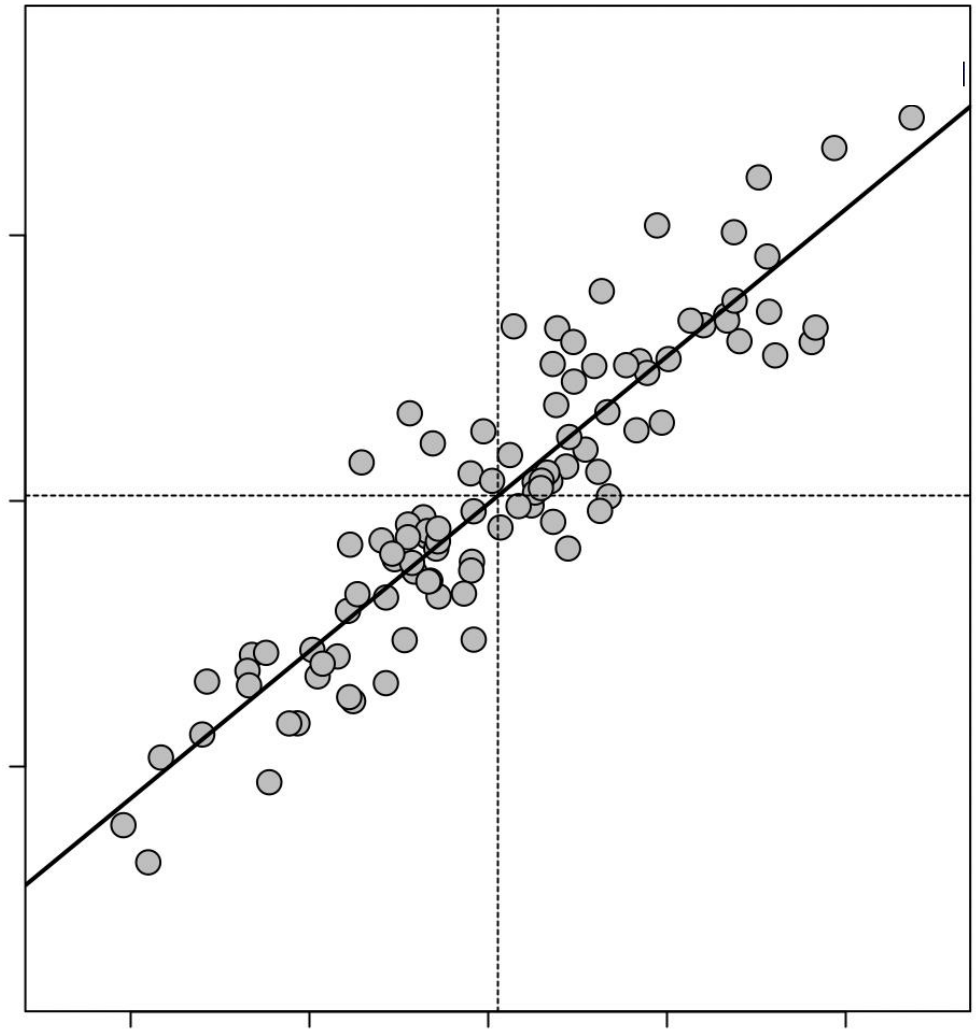
If so, positive or negative?

Weak or strong?

Can you guess the correlation coefficient (between -1 and 1)?

R = -0.04

Quick reminder



Is there a correlation?

If so, positive or negative?

Weak or strong?

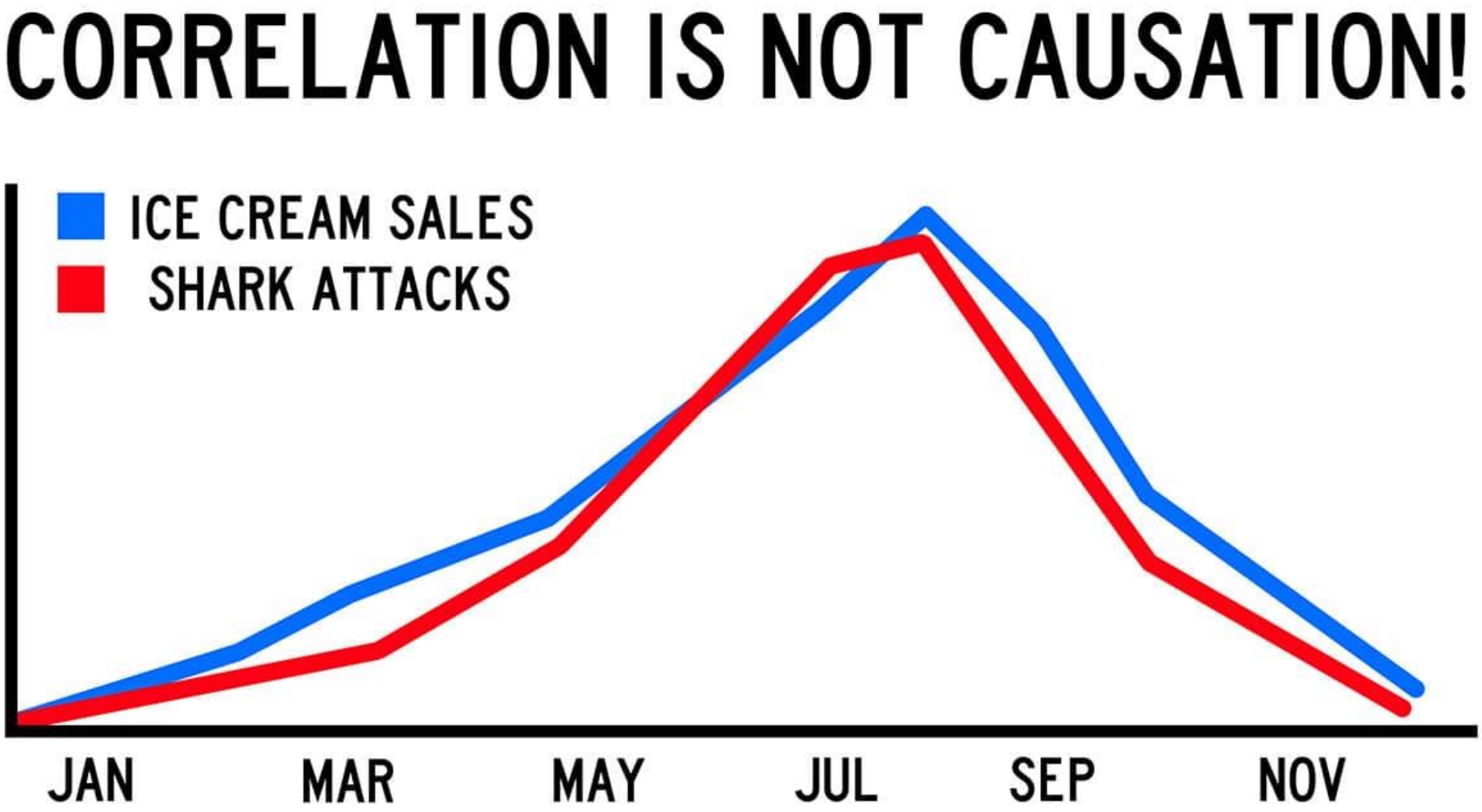
Can you guess the correlation coefficient (between -1 and 1)?

R = 0.91

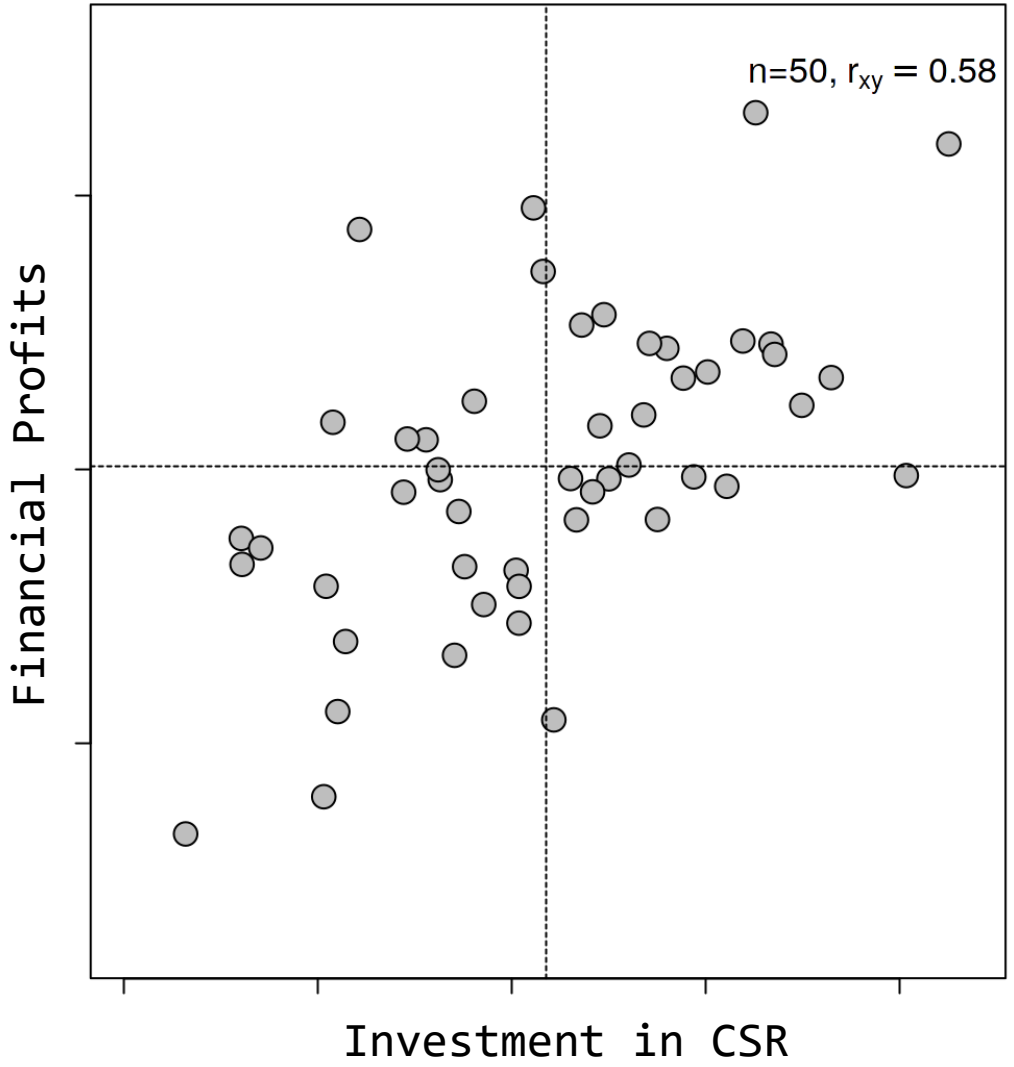
SESSION 2

-MULTIPLE LINEAR REGRESSION-

The essentials of regression



The essentials of regression



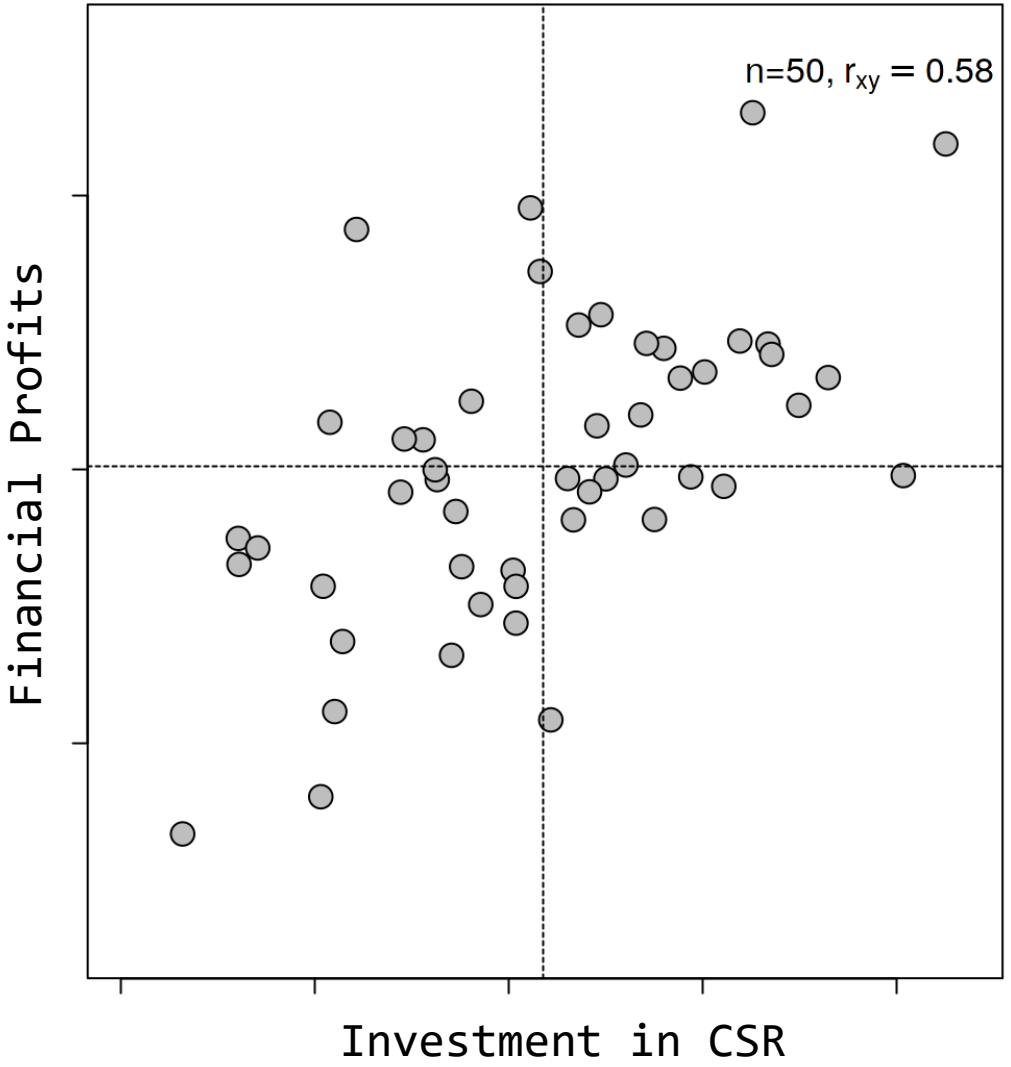
Strong correlation between companies that invest in CSR and financial profits.

Does it mean that investing in CSR is a good idea for companies?

NOT NECESSARILY

Could be the other way around: companies that already have high financial benefits can invest in CSR

The essentials of regression

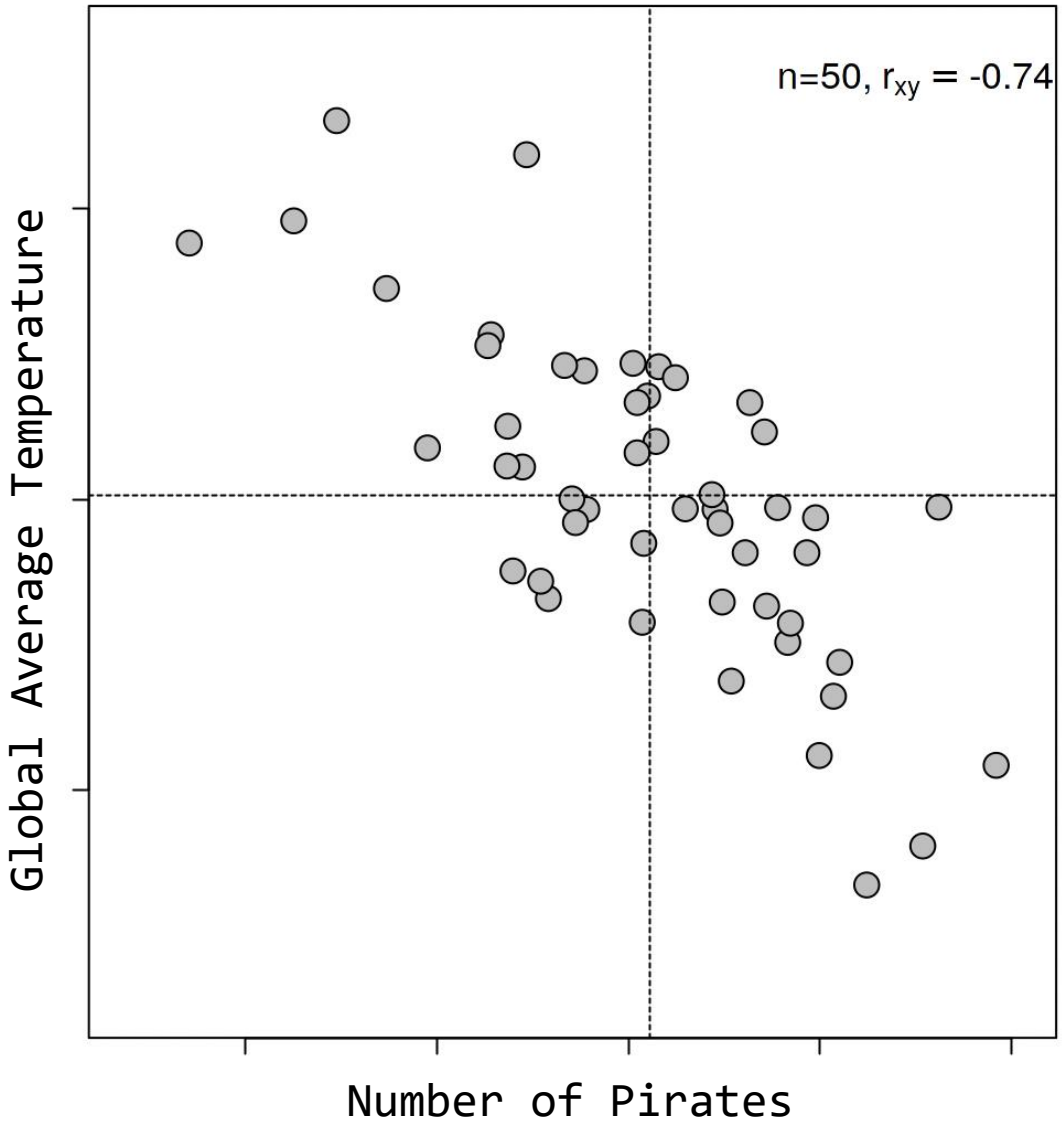


This problem is called

REVERSE CAUSALITY

Could be the other way around: companies that already have high financial benefits can invest in CSR

The essentials of regression

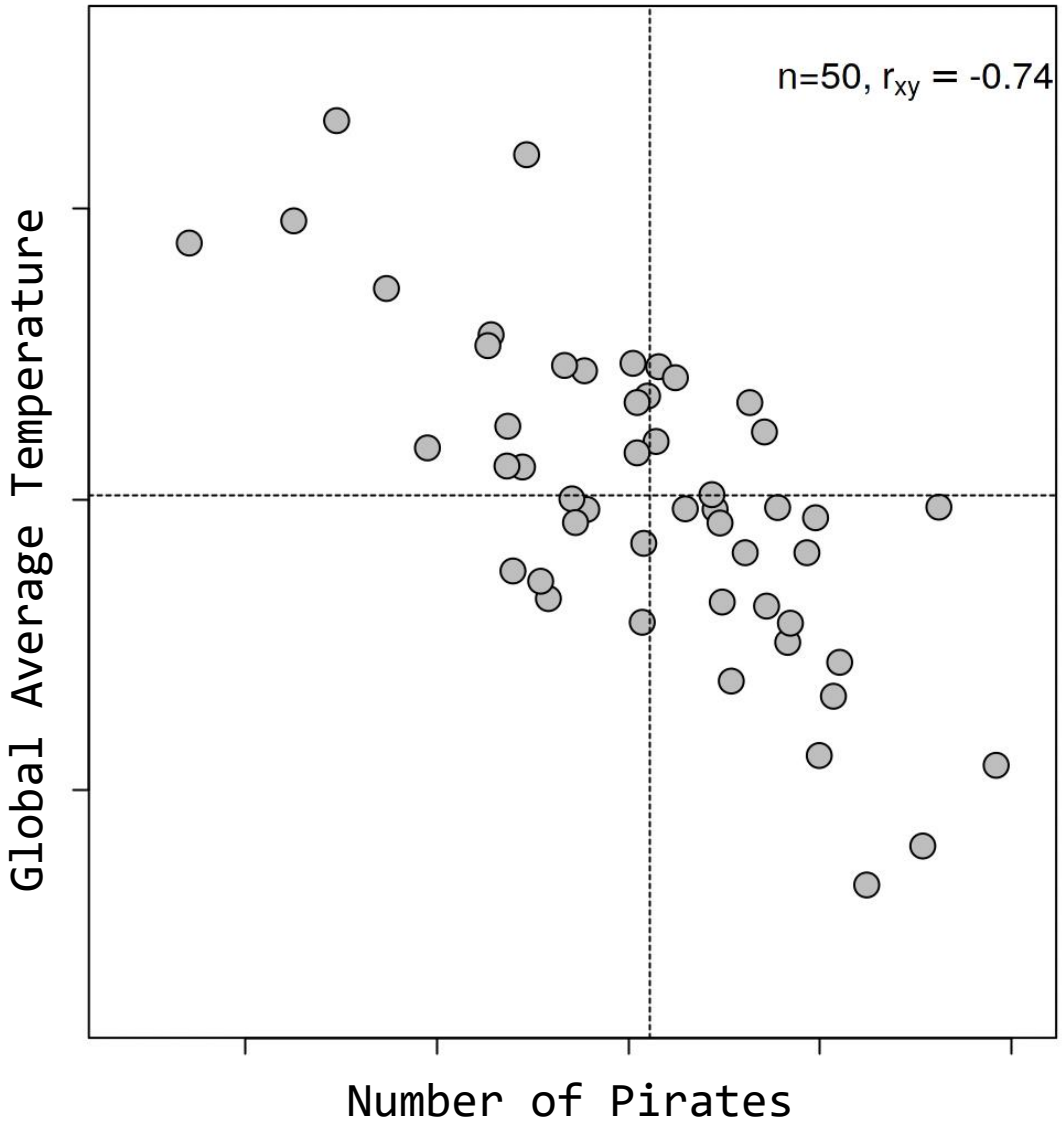


Are pirates preventing climate change?

NOT NECESSARILY

Could be for another reason: Time goes by. Around 1860, temperatures started to grow due to industrialization. At the same time, pirates started to decline due to UK's Royal Navy.

The essentials of regression



This problem is called

OMITTED VARIABLE

Two things happened at the same time, but independently of each other!

The essentials of regression

We can solve these issues with

MULTIPLE LINEAR REGRESSION

The essentials of regression

In general, **multiple linear regression is the same as a linear regression**, but **using more than one variable** to explain variation in our dependent variable.

Multiple Linear Regression

As we have seen, a linear regression looks like this:

$$y = a + \beta x$$

While a multiple linear regression looks like this:

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

When we only have a single independent variable (k = 1), we have a simple linear regression.

Multiple Linear Regression

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Like in simple linear regression, a value of *Y* which has been calculated with a regression equation is called a **predicted value**. The difference of the actual value minus the predicted value is called **residual**:

$$\text{Residual} = \text{Actual value} - \text{Predicted value}$$

The residuals are taken as prediction errors. So, the smaller the residuals, the better the model.

In practice, the **regression algorithm** chooses the coefficients so that the sum of the squares of the residuals is minimum. This is called the **method of least squares**. With more than one variable on the right side of the equation, implementing the least squares method involves a set of formulas which are usually packed into a matrix formula. The mathematics are hard to understand without some familiarity with matrix algebra.

Multiple Linear Regression

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

Again, the interpretation of results is somehow similar to the interpretation in simple linear regression: The coefficient of a particular **X** variable in a regression equation is frequently interpreted as its effect on **Y**.

However, this is the interesting part of multiple linear regression:

In multiple linear regression, increasing one unit the value of a particular variable, **holds constant the other variables**, therefore the change in the predicted value of Y is closer to reality.

Multiple Linear Regression

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k$$

But, be careful!

This interpretation is bounded to situations in which it makes sense to increase or decrease a particular *X* variable while holding constant the rest.

This is not realistic when this variable is strongly correlated among the independent variables.

This problem is known as **multicollinearity**.

TIME TO PRACTICE!

Multiple Linear Regression in Excel

Clarice Pereira is a production manager at CEMEX, one of the top concrete producers in the world.

She is in charge of supervising the quality of the production of concrete from several plants.

Multiple Linear Regression in Excel

Concrete is a main resource for construction, since it is easy to work with and is highly resistant for a variety of applications.

To produce concrete, it is required to use a mixture of cement, additives, water and other components such as rock, gravel and sand.

Multiple Linear Regression in Excel

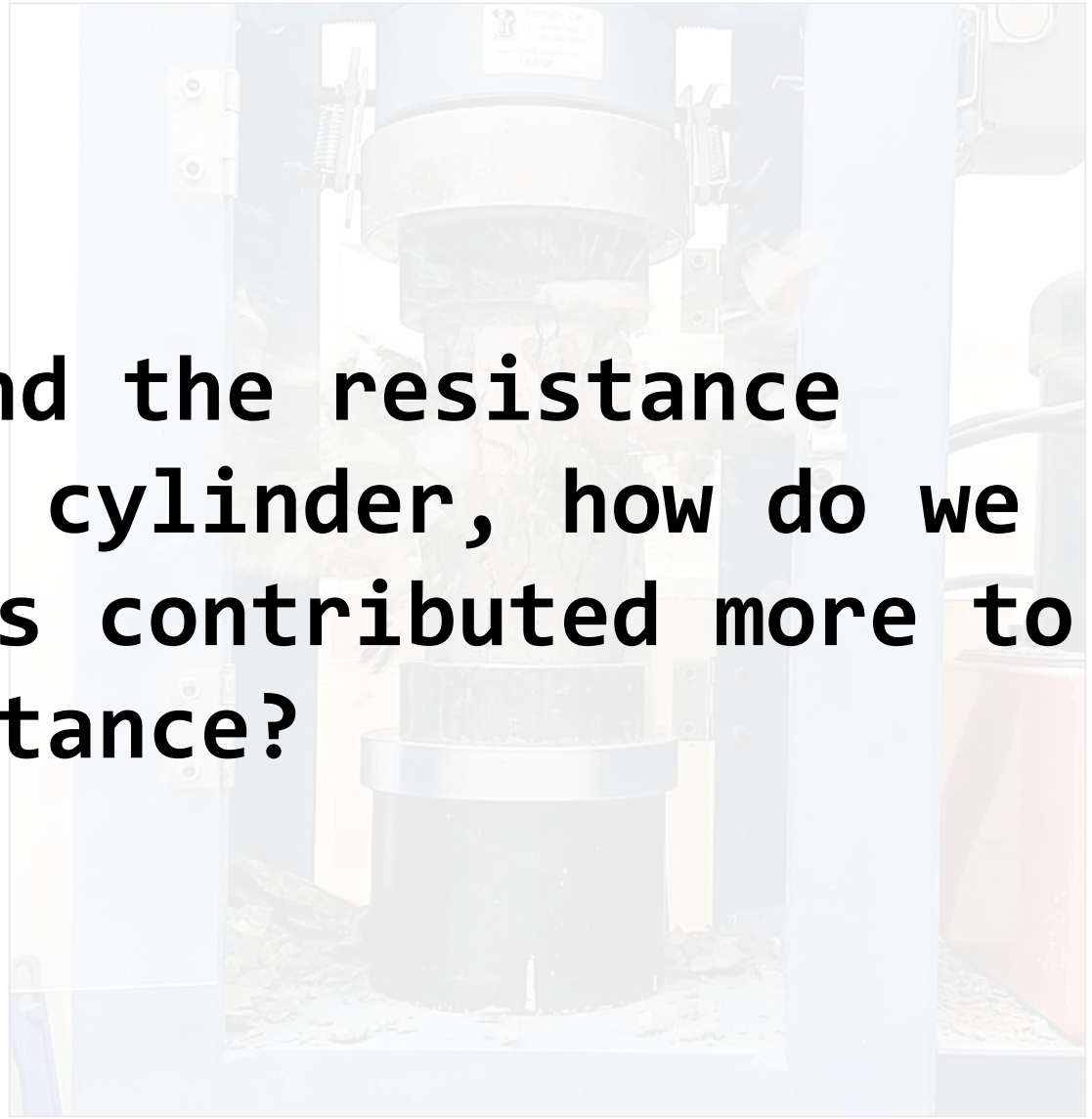
One of the most basic ways to perform a quality control of concrete is to test its resistance 28 days after it was produced.

To do this, a test is made to a concrete cylinder in which pressure is applied until the concrete breaks.

Multiple Linear Regression in Excel

One of the most basic ways to perform a quality control of concrete is to test its resistance 28 days after it was produced.

To do this, a test is made to a concrete cylinder in which pressure is applied until the concrete breaks.



However, once we find the resistance (quality) of a concrete cylinder, how do we know which ingredient has contributed more to its resistance?

Multiple Linear Regression in Excel

Clarice plans to analyze how the ingredients of that mix impact the resistance of the concrete after 28 days.

To this purpose, she gathers data on 804 production samples. The data set used for the analysis (file concrete.xls, sheet Data) contains the following variables:

- Resistance (kg/cm2).
- Cement (kg/m3).
- Additives (kg/m3).
- Water (kg/m3).

Fitting the Regression Line

We start by fitting a regression line (Resistance on Cement) to the data. The equation obtained is

Resistance = ? + ? Cement

Try to calculate it!
(Check the previous session slides if needed)

Fitting the Regression Line

We start by fitting a regression line (Resistance on Cement) to the data. The equation obtained is

Resistance = 0.74 + 0.99 Cement

How much is the correlation between Resistance and Cement?

R = 0.786

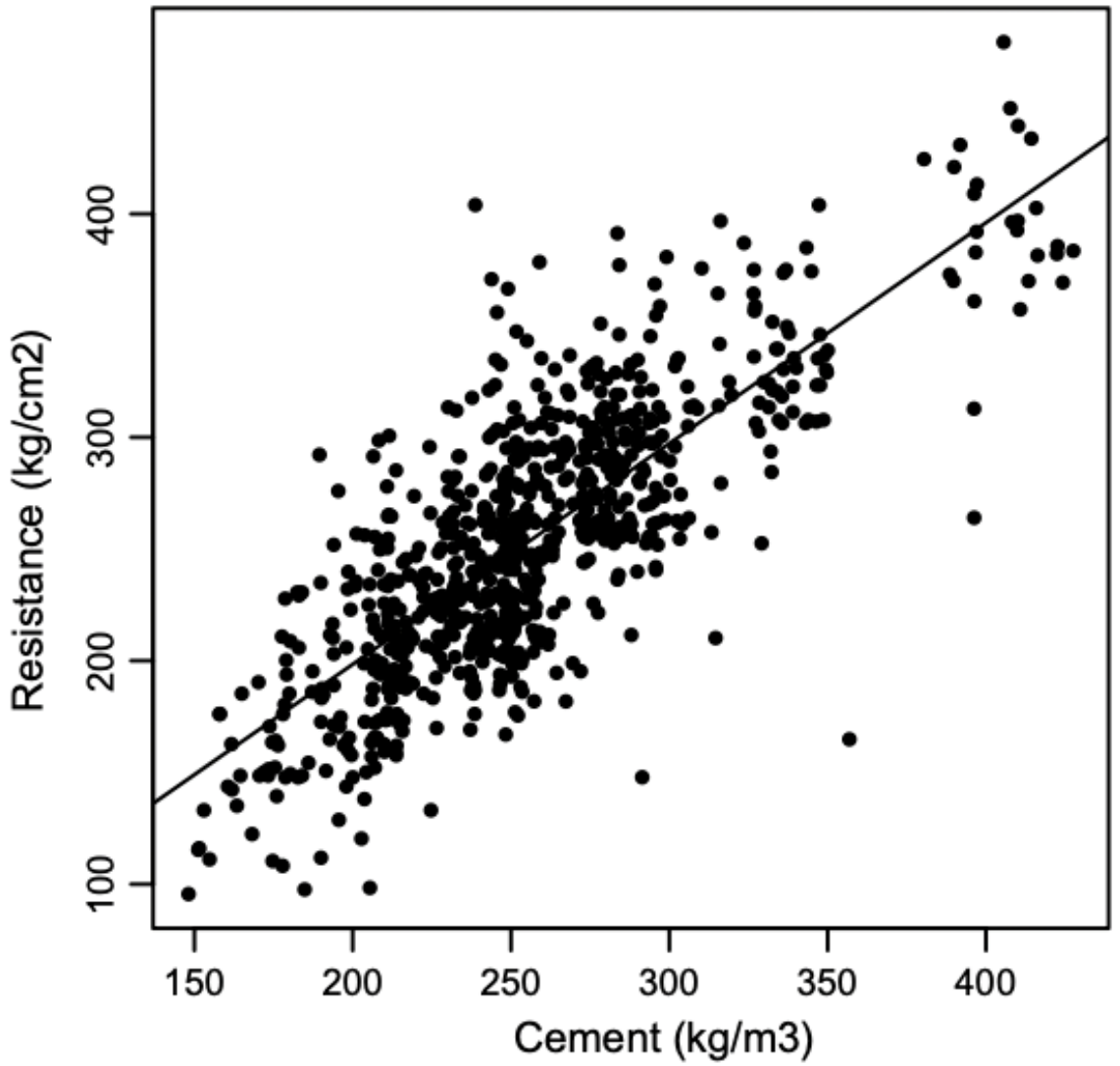
Fitting the Regression Line

Create a scatter plot!

How do you interpret the previous correlation and the graph?

Is it a positive/negative relationship?

Is it a strong correlation?



Fitting the Regression Line

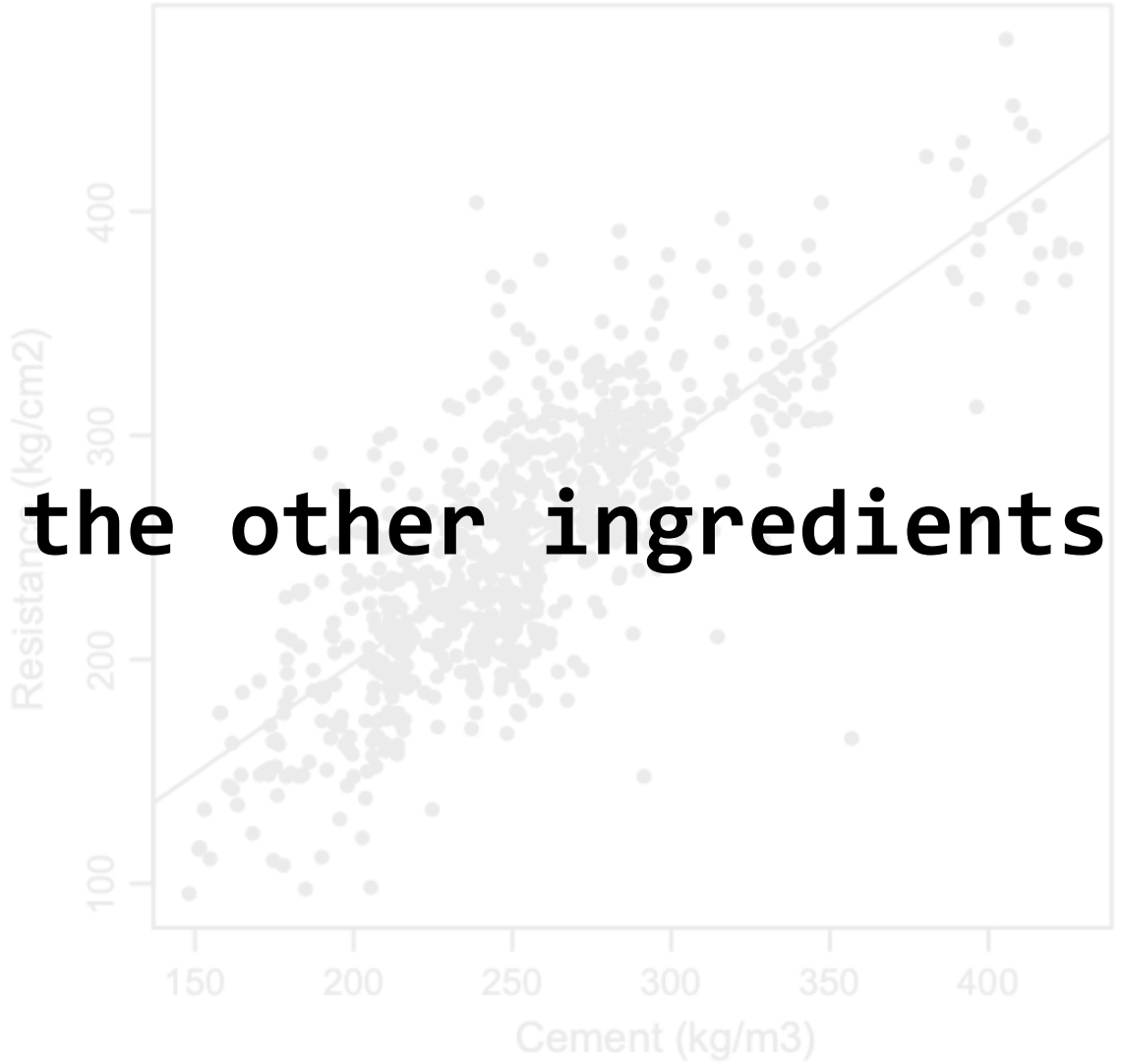
Create a scatter plot!

How do you interpret the previous correlation and the graph?

Let’s do the same with the other ingredients!

Is it a positive/negative relationship?

Is it a strong correlation?



Fitting the Regression Line

We start by fitting a regression line (Resistance on Additives) to the data. The equation obtained is

Resistance = ? + ? Additives

Try to calculate it!
(Check the previous session slides if needed)

Fitting the Regression Line

We start by fitting a regression line (Resistance on Additives) to the data. The equation obtained is

Resistance = 97.83 + 64.29 Additives

How much is the correlation between Resistance and Additives?

R = 0.646

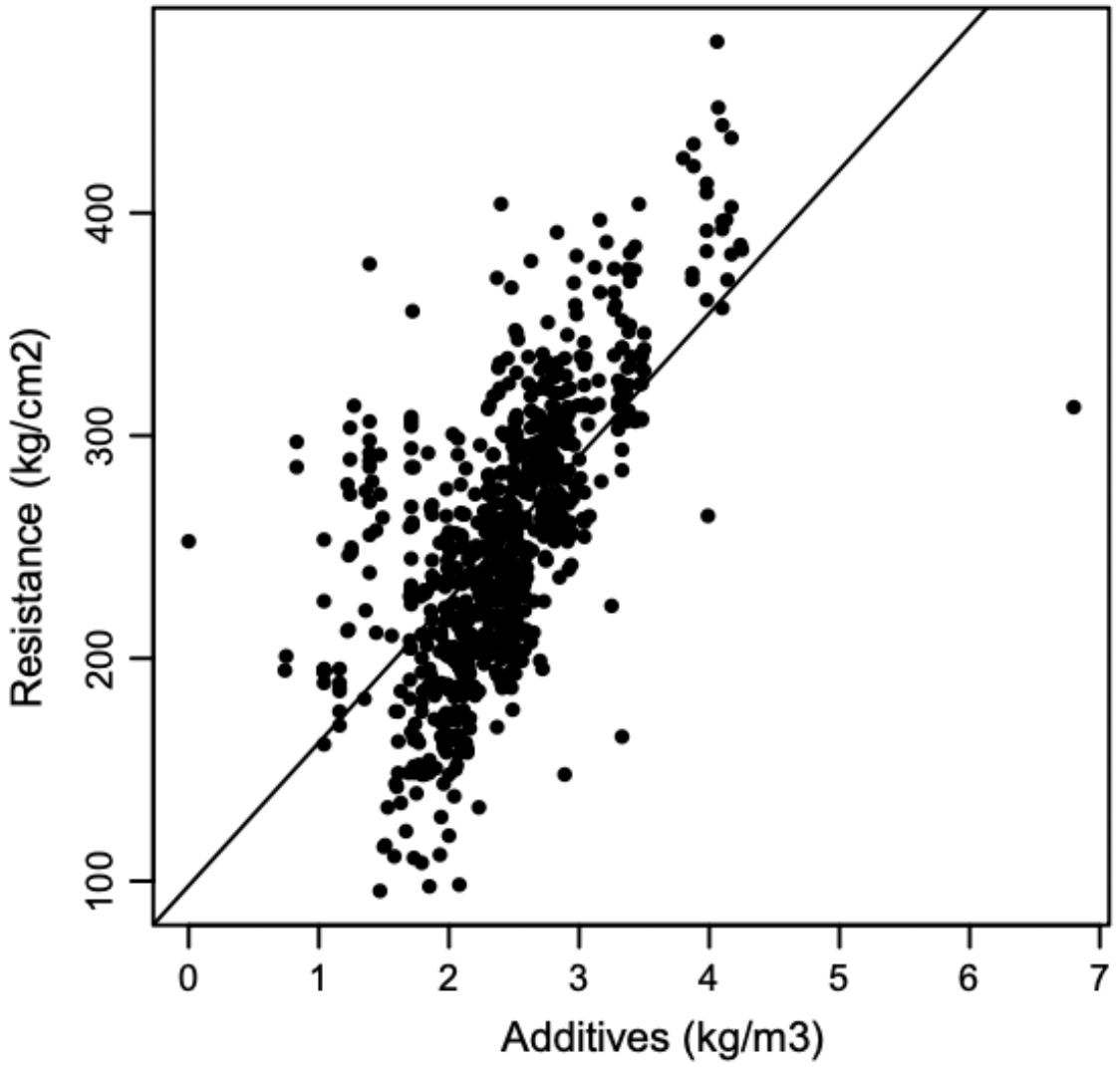
Fitting the Regression Line

Create a scatter plot!

How do you interpret the previous correlation and the graph?

Is it a positive/negative relationship?

Is it a strong correlation?



Fitting the Regression Line

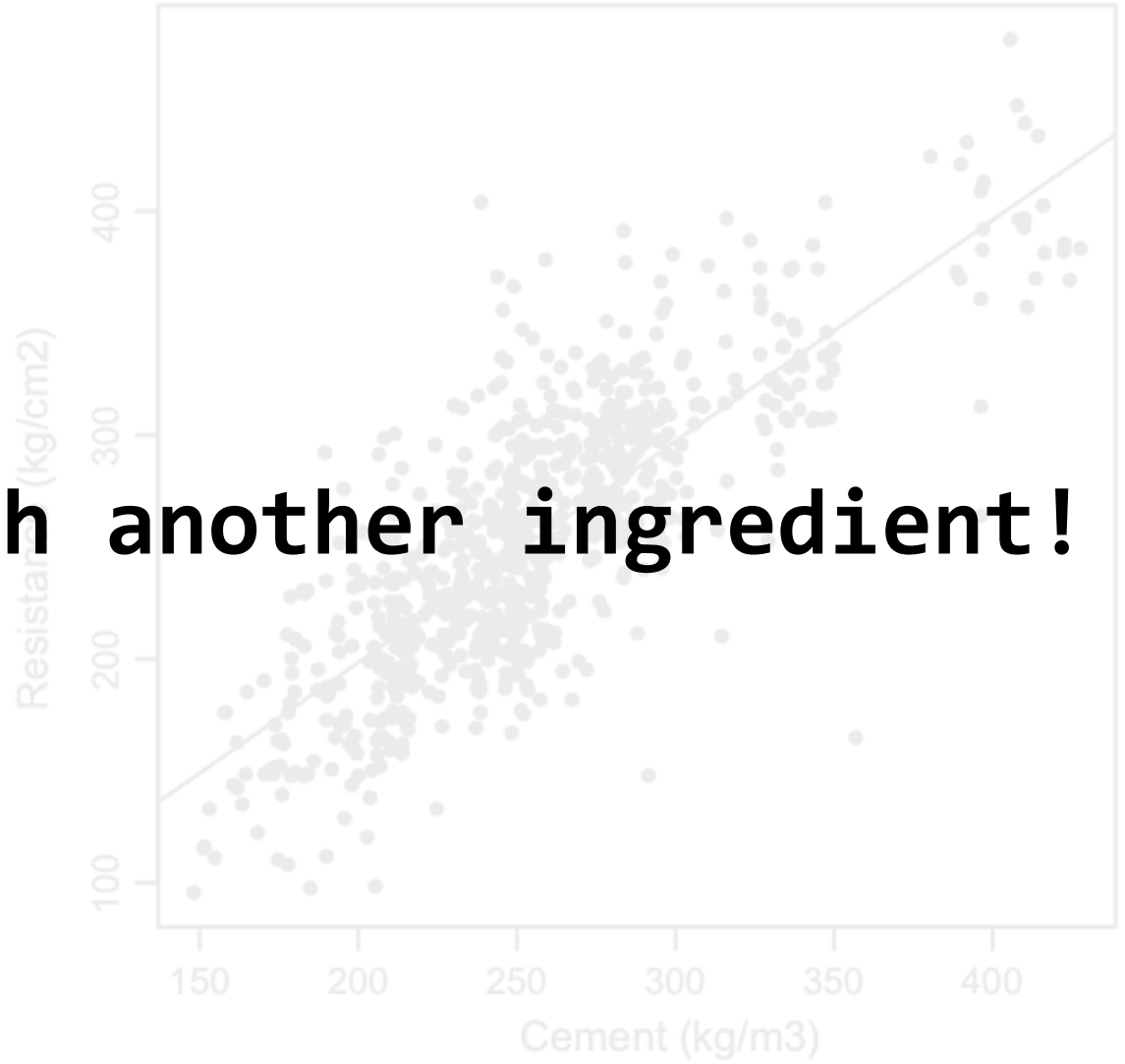
Create a scatter plot!

How do you interpret the previous correlation and the graph?

Let’s do the same with another ingredient!

Is it a positive/negative relationship?

Is it a strong correlation?



Fitting the Regression Line

We start by fitting a regression line (Resistance on Water) to the data. The equation obtained is

Resistance = ? + ? Water

Try to calculate it!
(Check the previous session slides if needed)

Fitting the Regression Line

We start by fitting a regression line (Resistance on Water) to the data. The equation obtained is

Resistance = 211.88 + 0.196 Water

How much is the correlation between Resistance and Water?

R = 0.104

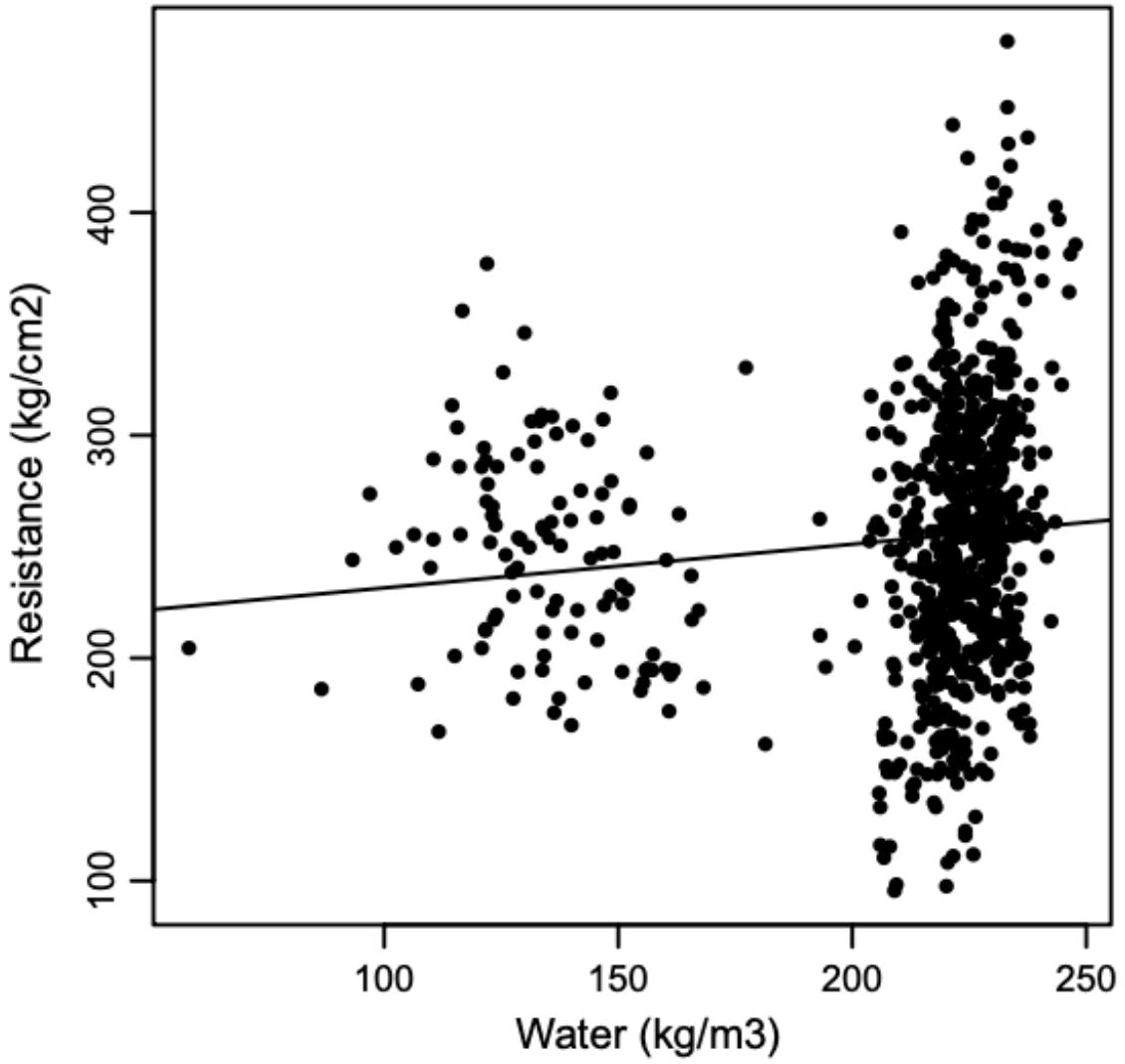
Fitting the Regression Line

Create a scatter plot!

How do you interpret the previous correlation and the graph?

Is it a positive/negative relationship?

Is it a strong correlation?



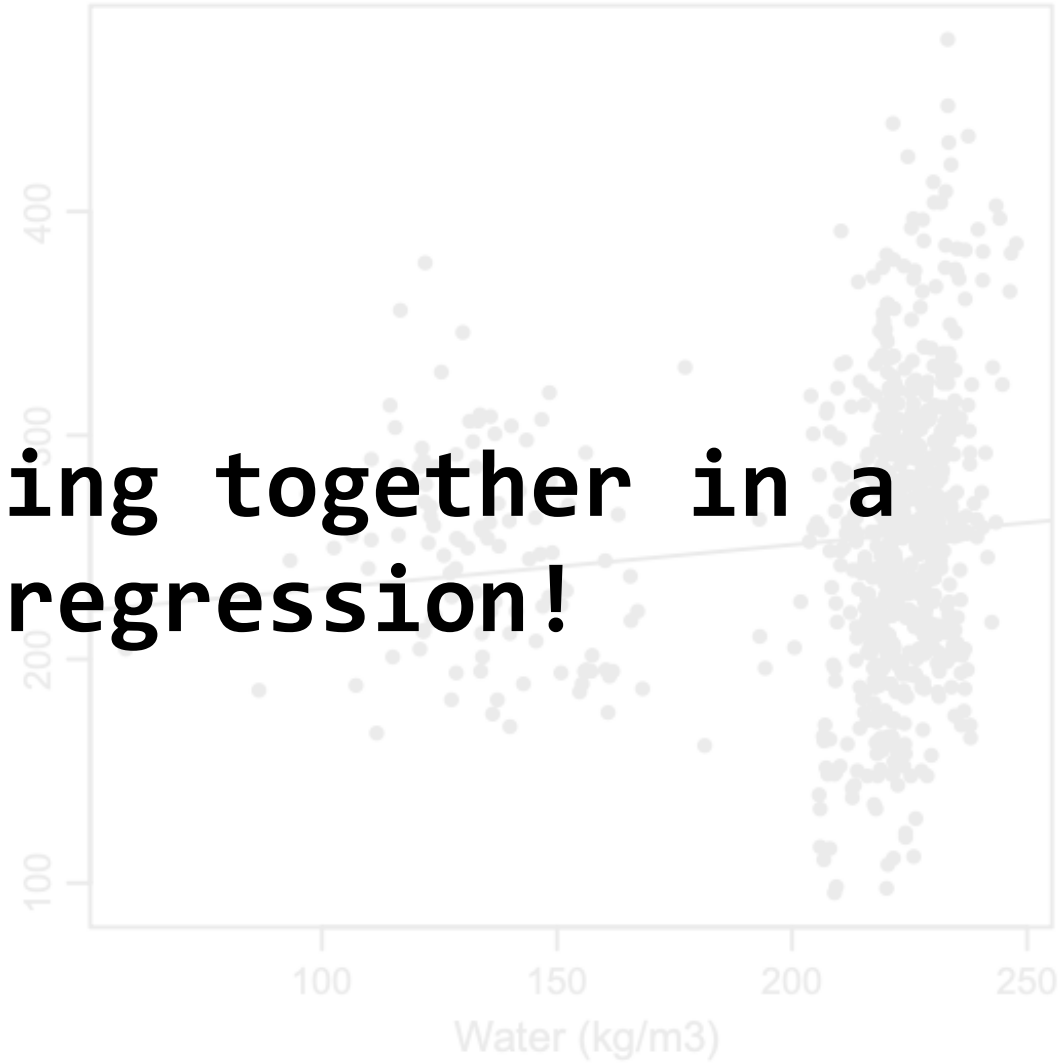
Fitting the Regression Line

Create a scatter plot!

How do you interpret the previous correlation and the graph?

Is it a positive/negative relationship?

Is it a strong correlation?

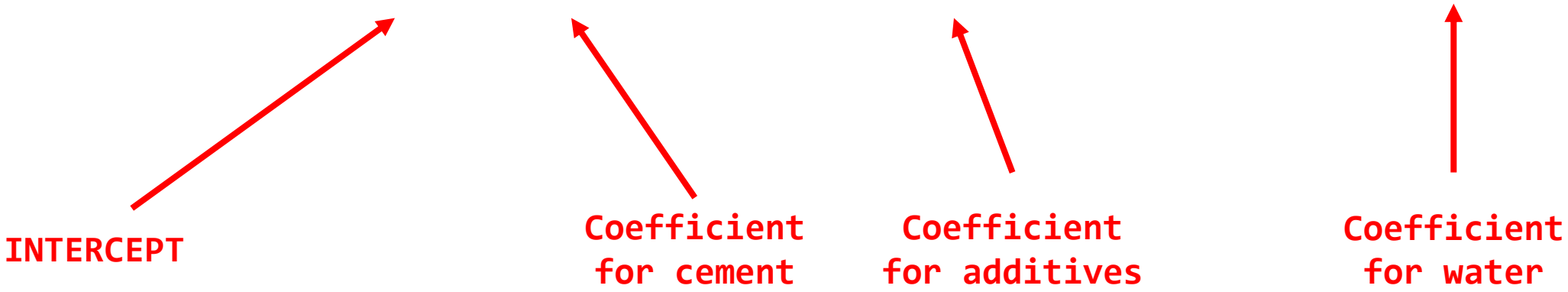


Now, let’s put everything together in a single multiple regression!

Multiple Regression Analysis

What is the obtained equation?

Resistance = ? + ? Cement + ? Additives + ? Water



Multiple Regression Analysis

What is the obtained equation?

Resistance = 15.83 + 0.85 Cement + 15.12 Additives - 0.07 Water

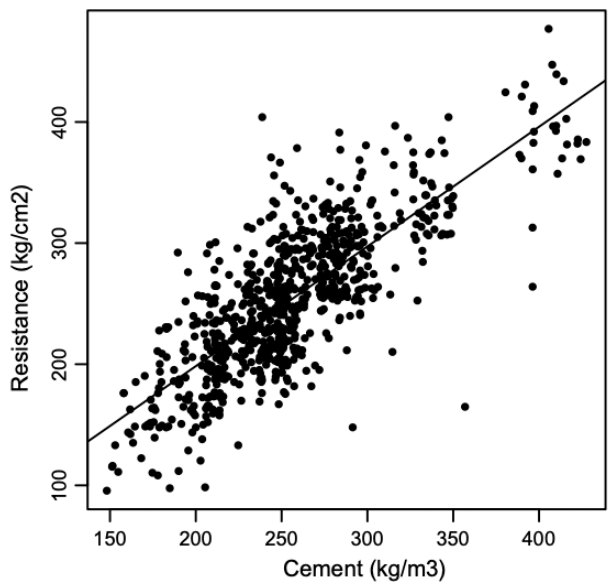
Multiple Regression Analysis

The “Adjusted R2” is the correlation between our predicted values and the observed (real) values. In other words, how good our model is!

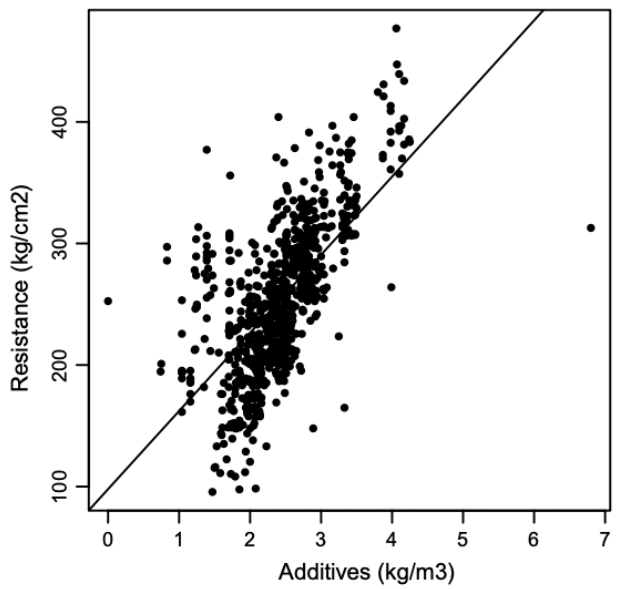
	Dependent variable:			
	Resistance	Resistance		Resistance
	(1)	(2)	(3)	(4)
Cement	0.988***			0.846***
	(0.027)			(0.056)
Additives		64.266***		
		(2.682)		
Water			0.196***	
			(0.067)	
Additives				15.160***
				(5.284)
Water				-0.073
				(0.065)
Constant	0.738	97.890***	211.884***	15.925
	(7.150)	(6.694)	(14.240)	(14.846)
Observations	804	804	804	804
R2	0.617	0.417	0.011	0.623
Adjusted R2	0.617	0.417	0.009	0.621
Residual Std. Error	37.778 (df = 802)	46.616 (df = 802)	60.740 (df = 802)	37.563 (df = 800)
F Statistic	1,293.782*** (df = 1; 802)	574.388*** (df = 1; 802)	8.700*** (df = 1; 802)	439.929*** (df = 3; 800)
Note:				*p<0.1; **p<0.05; ***p<0.01

Multiple Regression Analysis

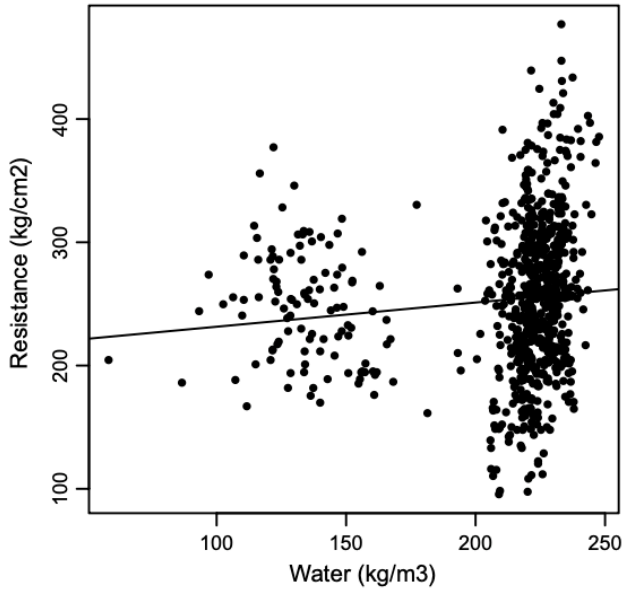
Compare this result with the simple linear regression results.



R = 0.786



R = 0.646

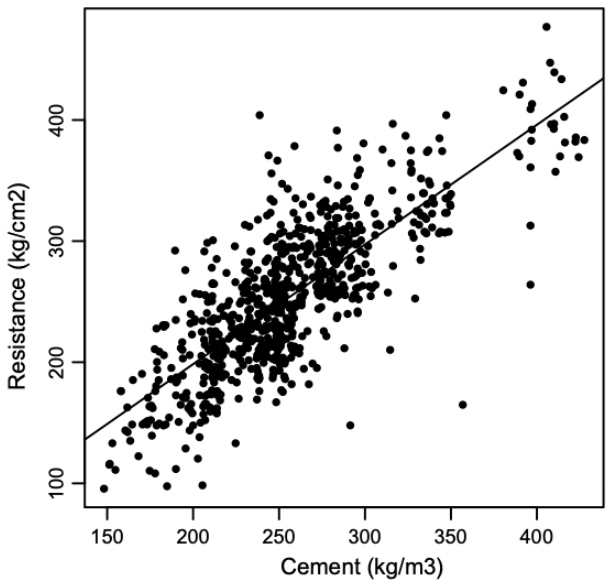


R = 0.104

Can you see something weird in this results?

Multiple Regression Analysis

Compare this result with the simple linear regression results.



R = 0.786

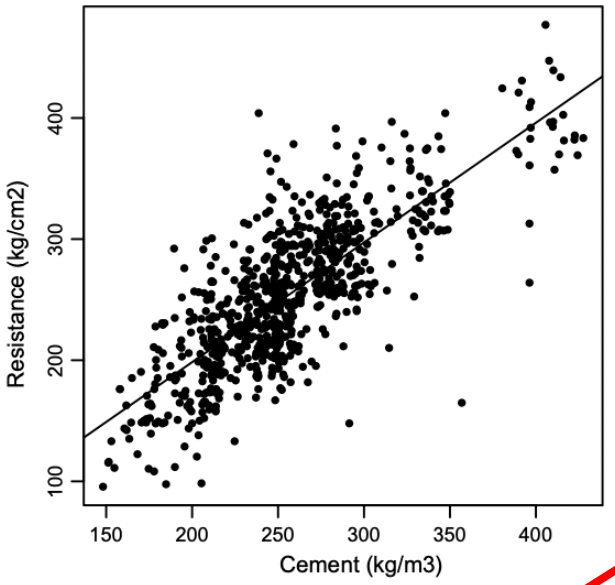
R = 0.646

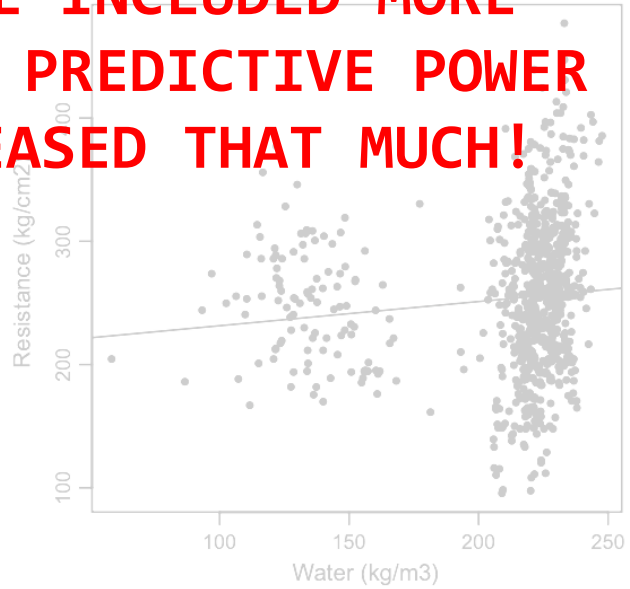
R = 0.104

Can you see something weird in this results?

Multiple Regression Analysis

Compare this result with the simple linear regression results.





EVENTHOUGH WE INCLUDED MORE VARIABLES, OUR PREDICTIVE POWER HAS NOT INCREASED THAT MUCH!

R = 0.786

R = 0.646

R = 0.104

Can you see something weird in this results?

Multiple Regression Analysis

Compare this result with the simple linear regression results.

R = 0.786

EVENTHOUGH WE INCLUDED MORE VARIABLES, OUR PREDICTIVE POWER HAS NOT INCREASED THAT MUCH!

THIS MEANS THAT CONCRETE'S STRENGHT IS LARGELY DEPENDENT ON THE AMOUNT OF CEMENT IT HAS

Multiple Regression Analysis

Compare this result with the simple linear regression results.

EVENTHOUGH WE INCLUDED MORE VARIABLES, OUR PREDICTIVE POWER HAS NOT INCREASED THAT MUCH!

THIS MEANS THAT CONCRETE'S STRENGHT IS LARGELY DEPENDENT ON THE AMOUNT OF CEMENT IT HAS

Which is something we can also see in the table. Cement has the lowest residual standard error.

Multiple Regression Analysis

Compare this result with the simple linear regression results.

EVENTHOUGH WE INCLUDED MORE VARIABLES, OUR PREDICTIVE POWER HAS NOT INCREASED THAT MUCH!

THIS MEANS THAT CONCRETE'S STRENGHT IS LARGELY DEPENDENT ON THE AMOUNT OF CEMENT IT HAS

Regression statistics	
Multiple R	0.89
R square	0.625
Adjusted R square	0.621
Standard error	37.6
Observations	1030

	Coefficients	P-value
Intercept	15.83	0.286
Cement	0.846	0.000
Additives	0.015	0.004
Water	−0.072	0.266

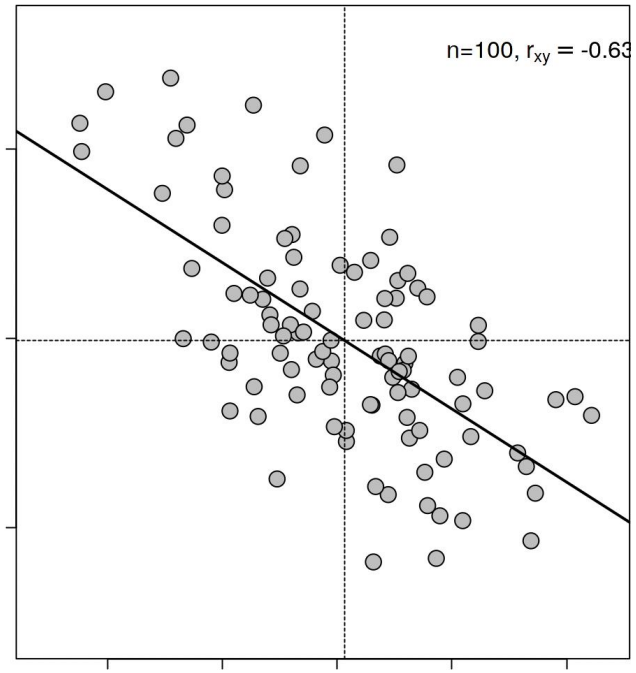
This is something we can also see in the table. Cement has the biggest coefficient (meaning biggest effect on Resistance) and is highly significant (low p-value)

QUESTIONS?

Remember:

There are no dumb questions, nobody is born knowing!

Fundamentals of Econometrics Models



Vicenç Soler
v.soler@tbs-education.org
~~vincent.soler@tbs-education.org~~