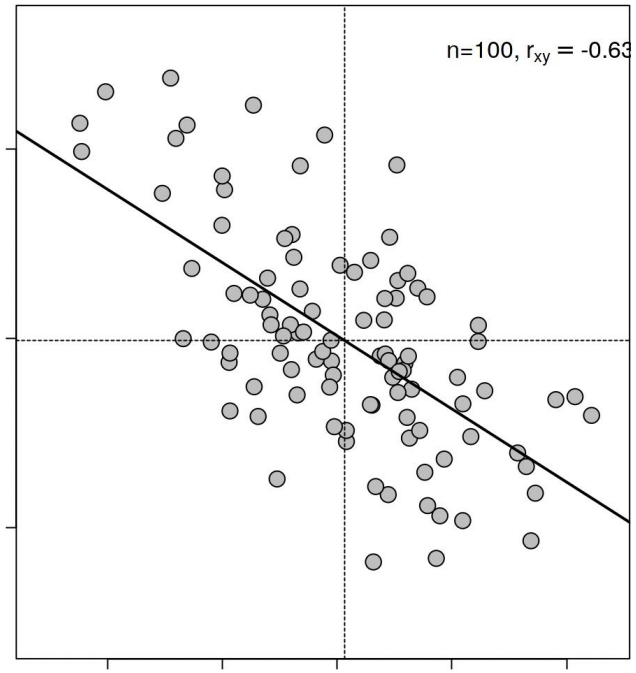
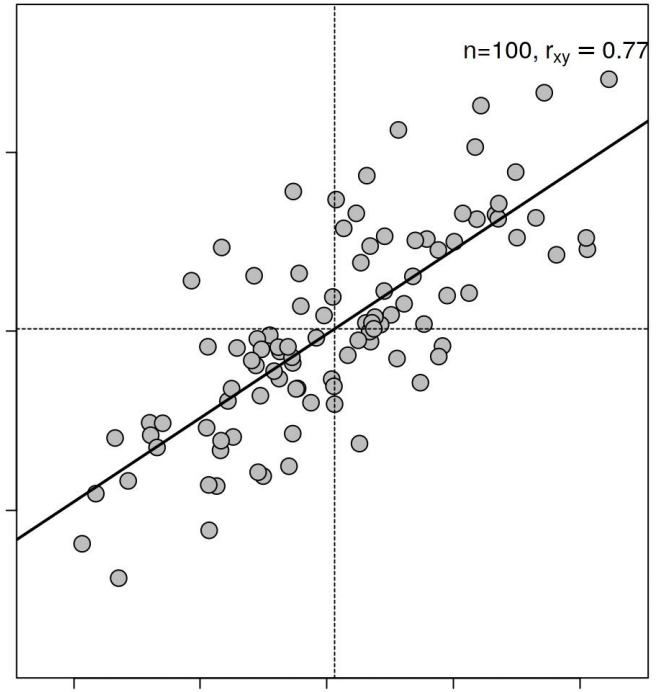
# Fundamentals of Econometrics Models




**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.

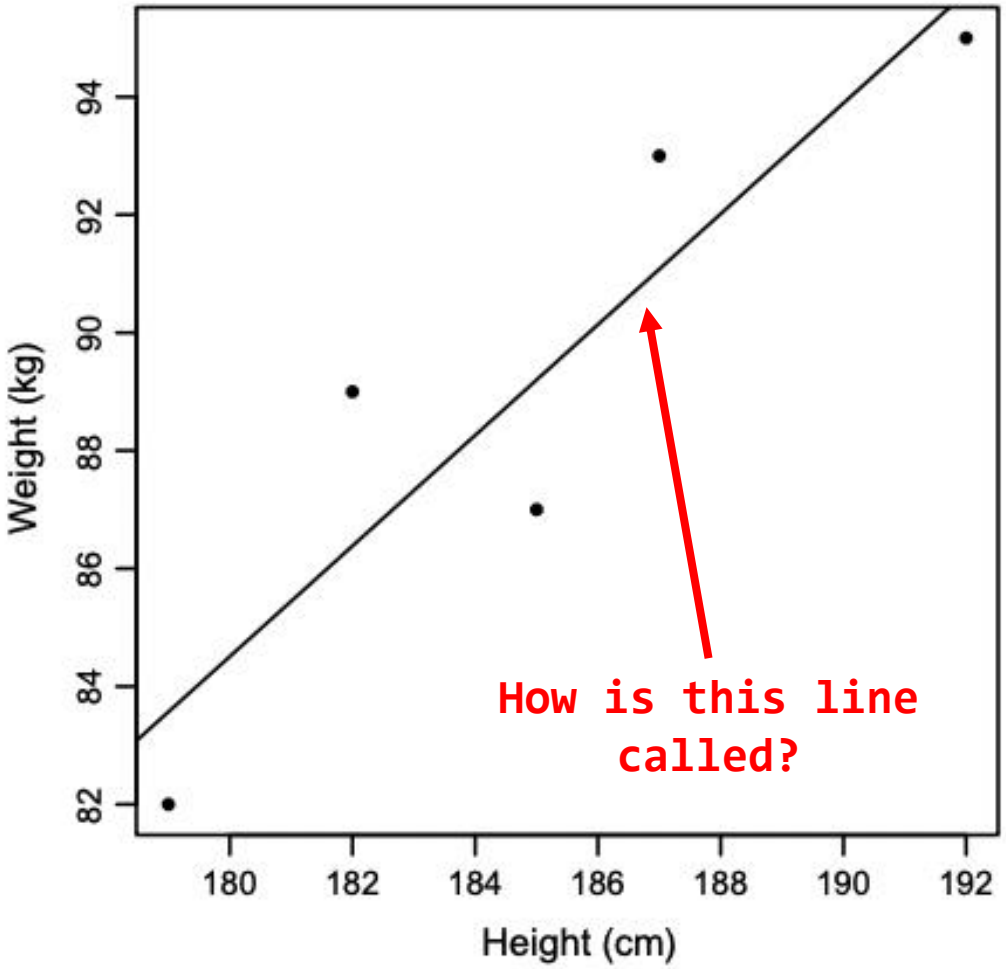
# Quick reminder

**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…**

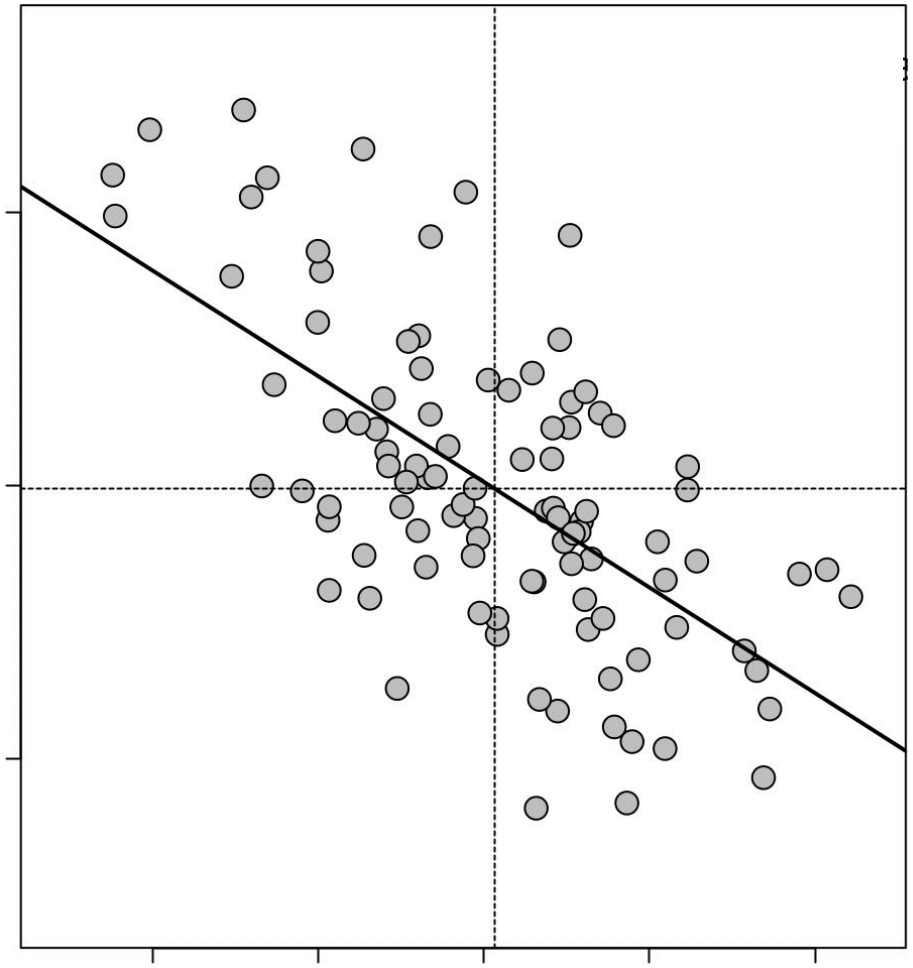
$$y = a + \beta x$$

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

**What is the slope? And the intercept?**

**Slope = 0.93 (or $\beta$) and Intercept = -83.47 (or a)**

# Quick reminder



Is there a correlation?

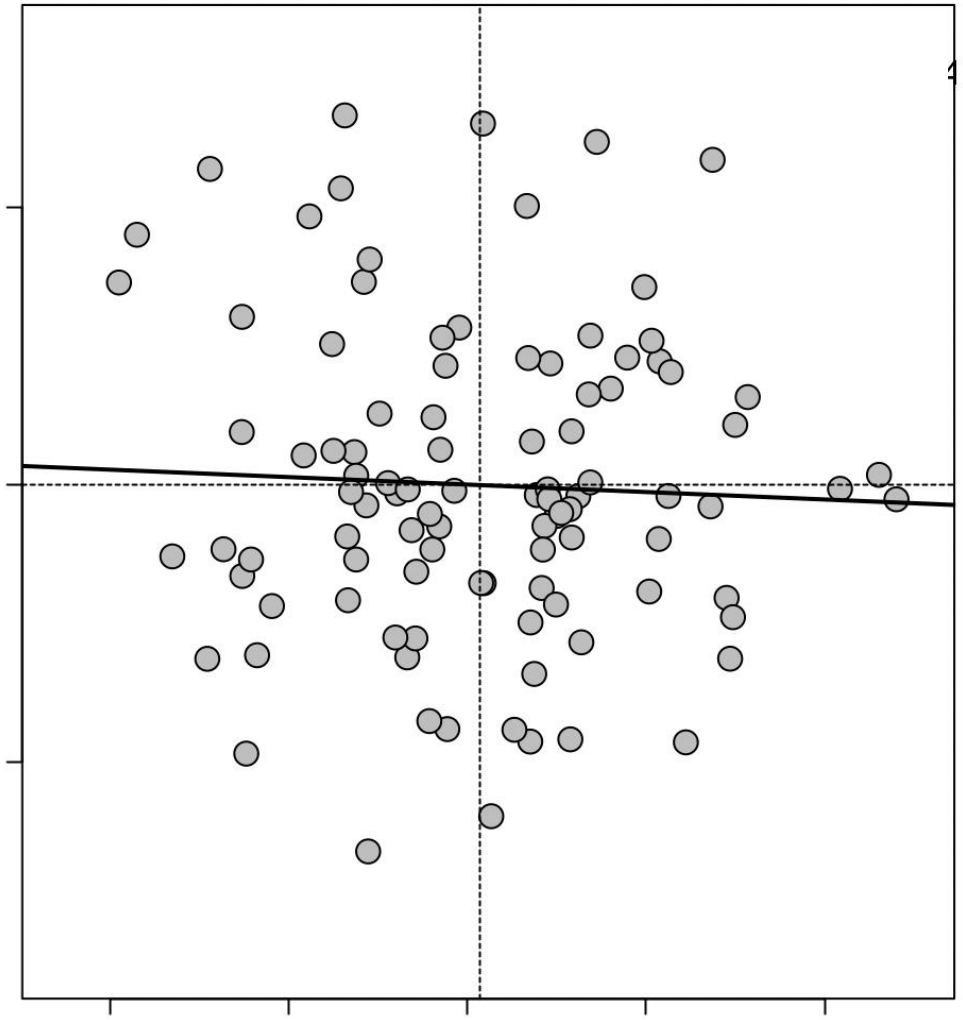
If so, positive or negative?

Weak or strong?

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

**R = -0.63**

# Quick reminder



Is there a correlation?

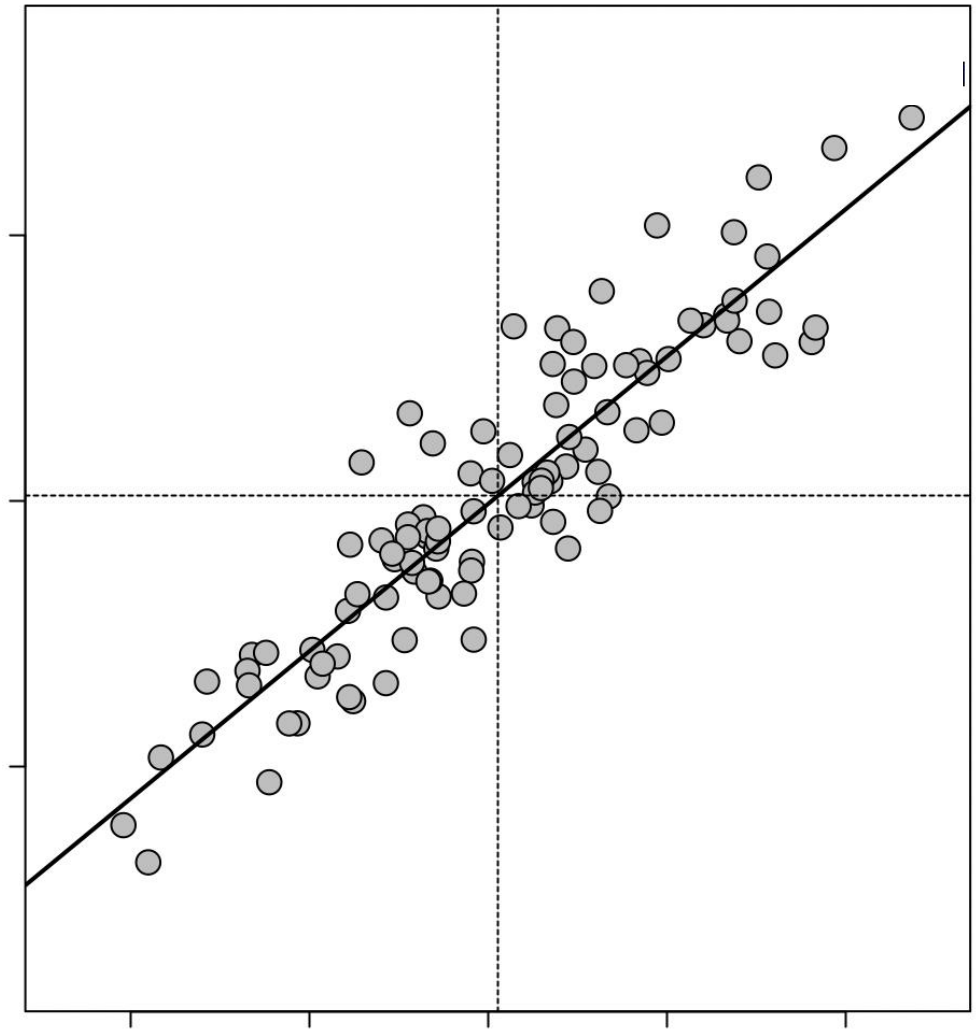
If so, positive or negative?

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**

# Quick reminder

What is this?

| | Dependent variable: | | | |
|---|---|---|---|---|
| | Resist | 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 |

# Quick reminder

And this?

| | Dependent variable: | | | |
|---|---|---|---|---|
| | Resist | 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 |

# SESSION 3

## REGRESSION SIGNIFICANCE TEST

# Statistical Inference

We have seen that the regression coefficients are reported with extra information, which we have not discussed yet.

The goal of today: **How can we use this information?**

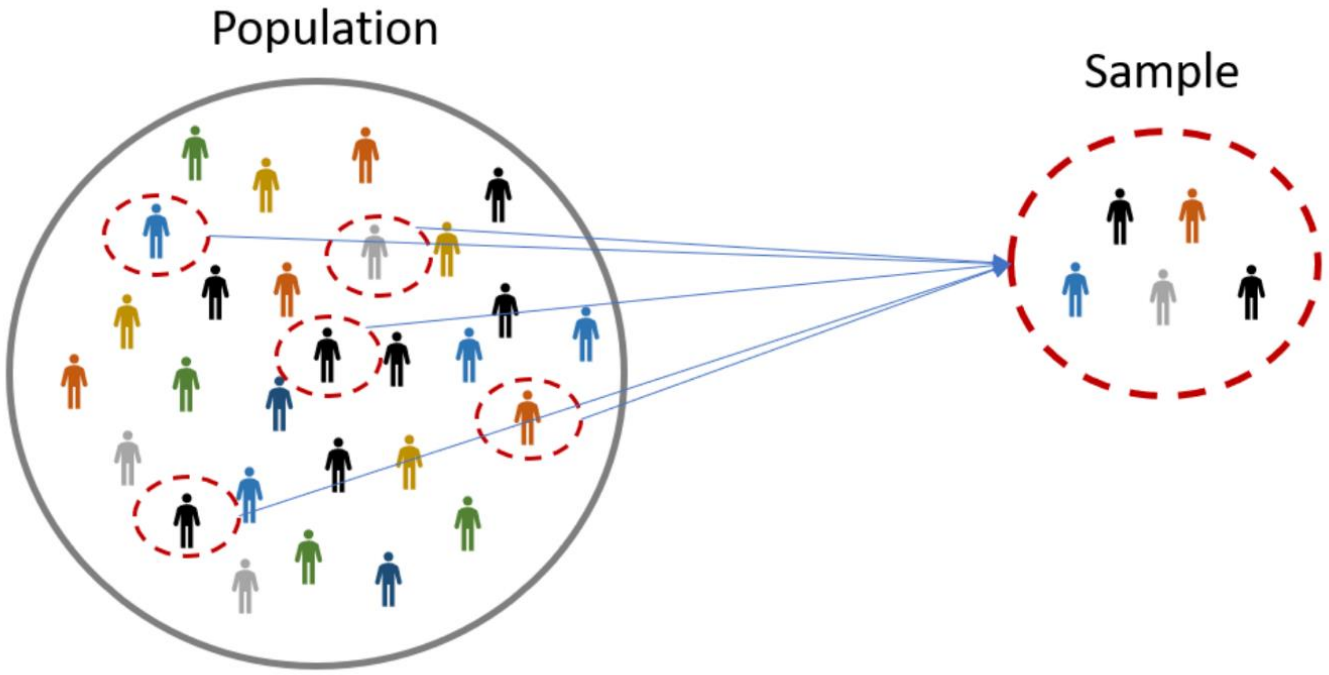
| | Dependent variable: | | | |
|---|---|---|---|---|
| | Resistance (1) | Resistance (2) | (3) | Resistance (4) |
| Cement | 0.988*** (0.027) | | | 0.846*** (0.056) |
| Additives | | 64.266*** (2.682) | | |
| Water | | | 0.196*** (0.067) | |
| Additives | | | | 15.160*** (5.284) |
| Water | | | | -0.073 (0.065) |
| Constant | 0.738 (7.150) | 97.890*** (6.694) | 211.884*** (14.240) | 15.925 (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 |

```
Call:
lm(formula = dataset$Resistance ~ dataset$Water)

Residuals:
     Min       1Q   Median       3Q      Max
-157.467  -43.549   -0.175   39.885  219.183

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   211.88390   14.23989  14.880  < 2e-16 ***
dataset$Water   0.19617    0.06651   2.949  0.00328 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.74 on 802 degrees of freedom
Multiple R-squared:  0.01073,   Adjusted R-squared:  0.009497
F-statistic:   8.7 on 1 and 802 DF,  p-value: 0.003275
```

# Statistical Inference

Before explaining it, we need to discuss briefly the context.

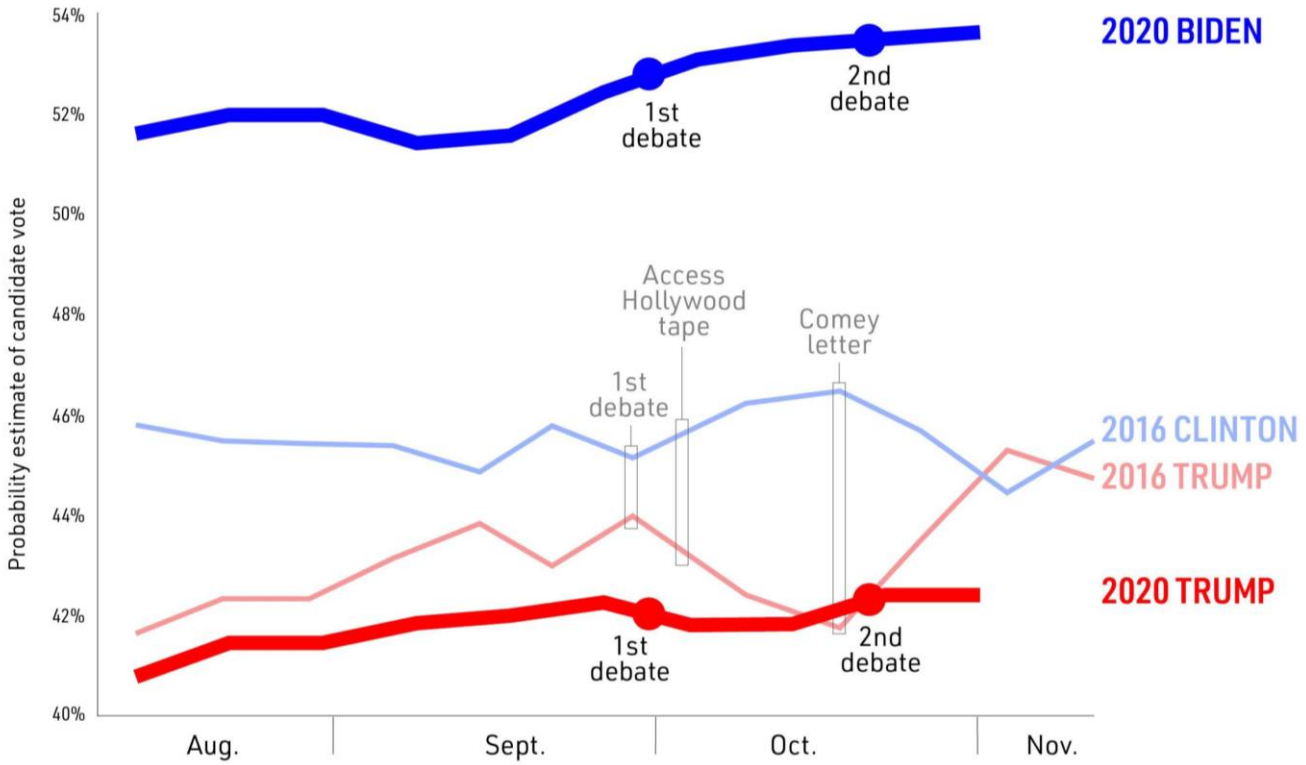
Populations are usually too big to be observed completely. What we do, in practice, is to **draw a sample from the population**, perform our statistical analysis on the sample, and extend the results to the population.




This is called
**Statistical Inference**

# Statistical Inference

For instance, if in a survey on voting intention in the presidential elections in the US, a 41.7% of the interviewees declare their intention to vote for the Republican candidate, what can we say about the population?

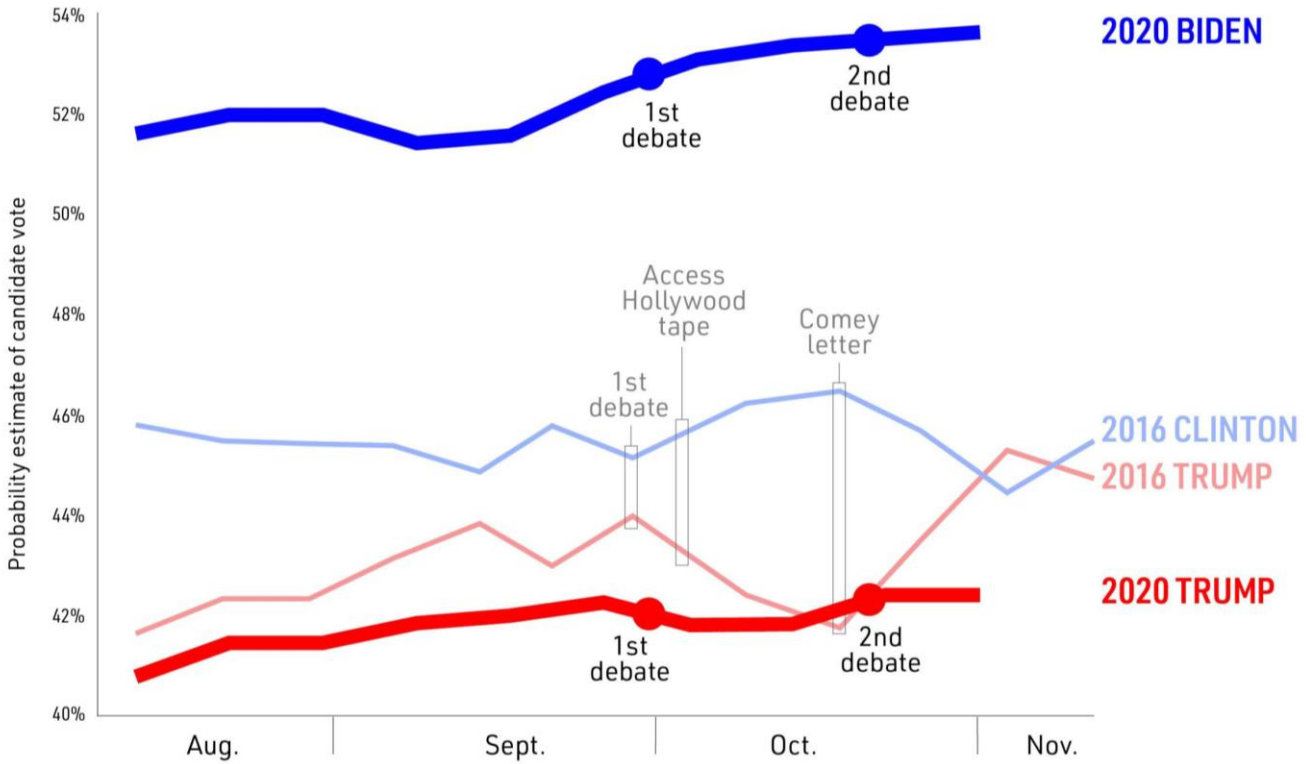



How **sure** are we about these numbers?

# Statistical Inference

Sample 1: 41.7% for Republicans
Sample 2: 42.3% for Republicans
Population: ?% for Republicans

How **big the error** could be?




How **sure** are we about these numbers?

# Statistical Inference

Be careful with this:

**1.** Reliability of an estimate based on a sample **depends on sample size**. The bigger the sample, the more we trust the results.

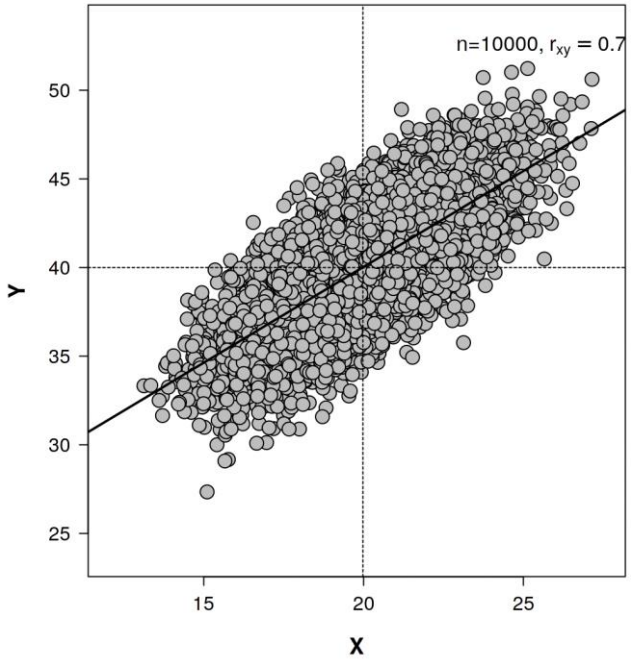
**2.** For the inference from sample to population to be valid, **the sample has to be extracted by means of a random sampling procedure**. In practice, sampling is accepted as random when no part of the population is favoured.

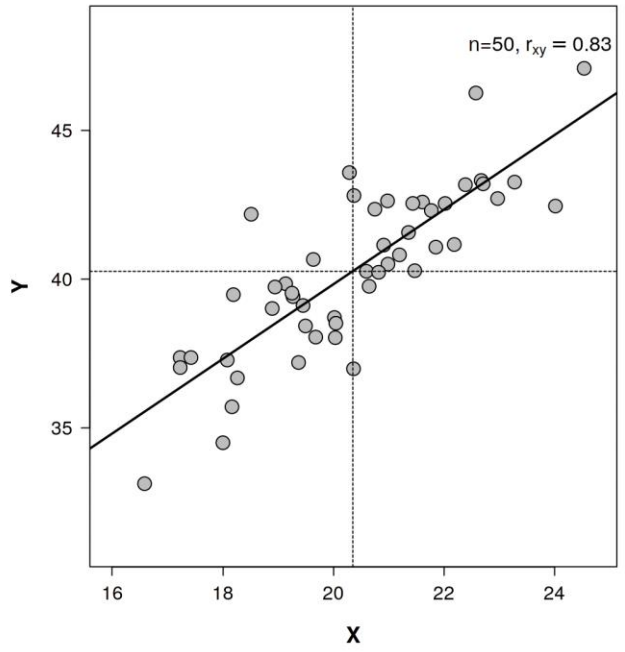
**3.** When the sampling procedure favours a certain subpopulation, it is said to be **biased**. A sample extracted from Twitter would probably be taken as biased (toward young people), because old electors would be less represented there than in the population.

# Statistical Inference

Let us suppose a population on whose individuals we can observe two variables *X* and *Y* , for which there is a linear equation *Y=a+bX*. We call this equation the **population equation**. Unless we measure the values of *X* and *Y* on every individual of the population, we cannot know exactly the population equation (coefficients *a* and *b)*.

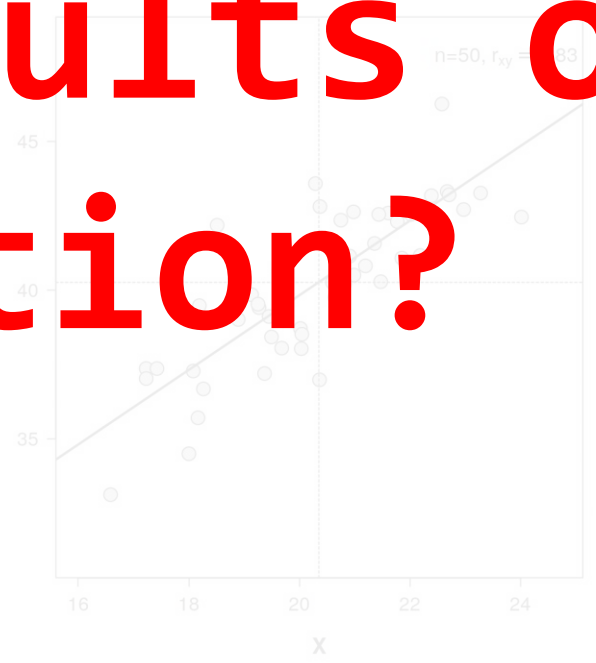
We extract a random sample from that population, measuring *X* and *Y*. Then, we fit a regression line to these data, using R software. The equation of this line, derived from the sample data, will not be the same as the true population equation. We call it the **sample equation**.



# Statistical Inference

**How do we know if the results of the sample match the results of the population?**

# Statistical Inference

**Collecting more samples!**

Our sample estimate has an error, which can be put in evidence by collecting data from other samples and comparing, across samples, the different coefficients obtained.

Nevertheless, we never do this in real-world statistical analysis.

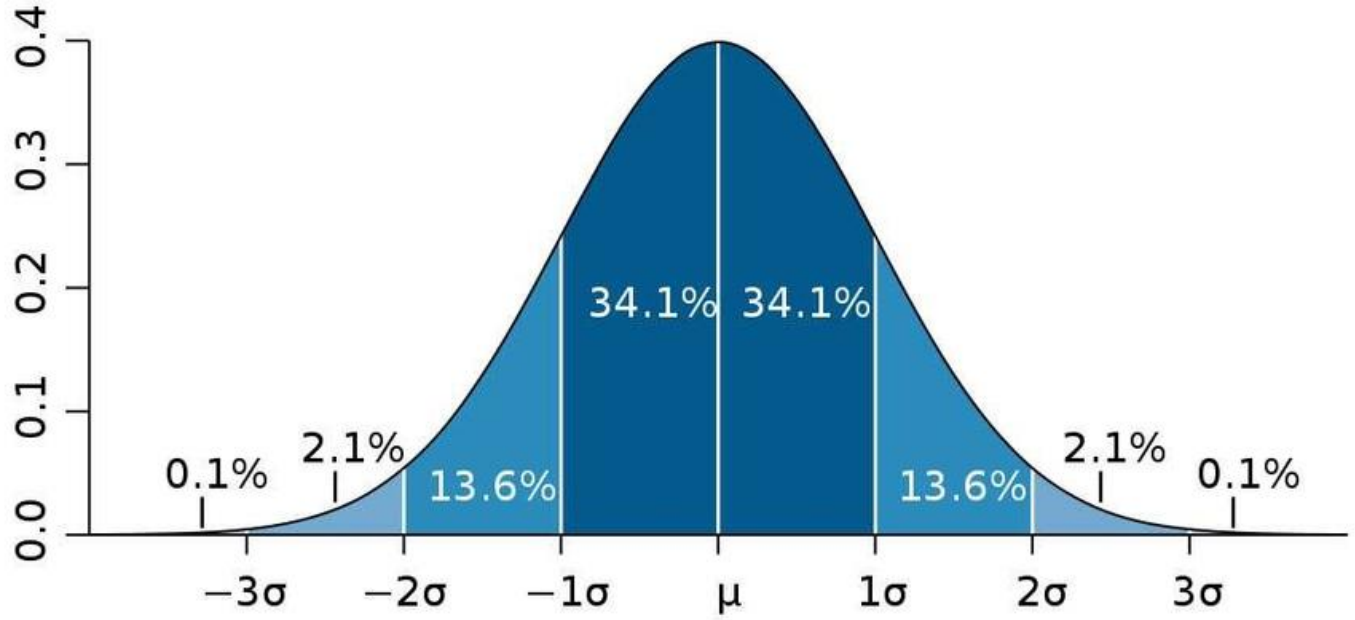
**Guess the distribution!**

When the sample size is big enough (like n > 100), the sample coefficients (a and b) are approximately normally distributed, with a mean which is equal to the population coefficient and a standard deviation equal to the standard error reported in the coefficients table of the regression output. In practice, we take advantage of this fact by using the formula:

$$\text{Coefficient} \pm (2 \times \text{Standard Error})$$

# Error in coefficient

## Coefficient ± (2 × Standard Error)




lower confidence interval

upper confidence interval

# Statistical Inference

## Coefficient ± (2 × Standard Error)

This formula will give us two numbers: **lower confidence interval** and **upper confidence interval**, for which we are 95% confident, that the true coefficient falls between these two numbers.

In other words, if we collect more samples, just 5% of the time the coefficient will be out of the lower and upper confidence intervals.

# Statistical Inference

**OK but, what do we do with these confidence intervals?**

Coefficient ± (2 × Standard Error)

This formula will give us two numbers, lower confidence interval and upper confidence interval, for which we are 95% confident, that the true coefficient falls between these two numbers.

In other words, if we collect more samples, just 5% of the time the coefficient will be out of the lower and upper confidence intervals.

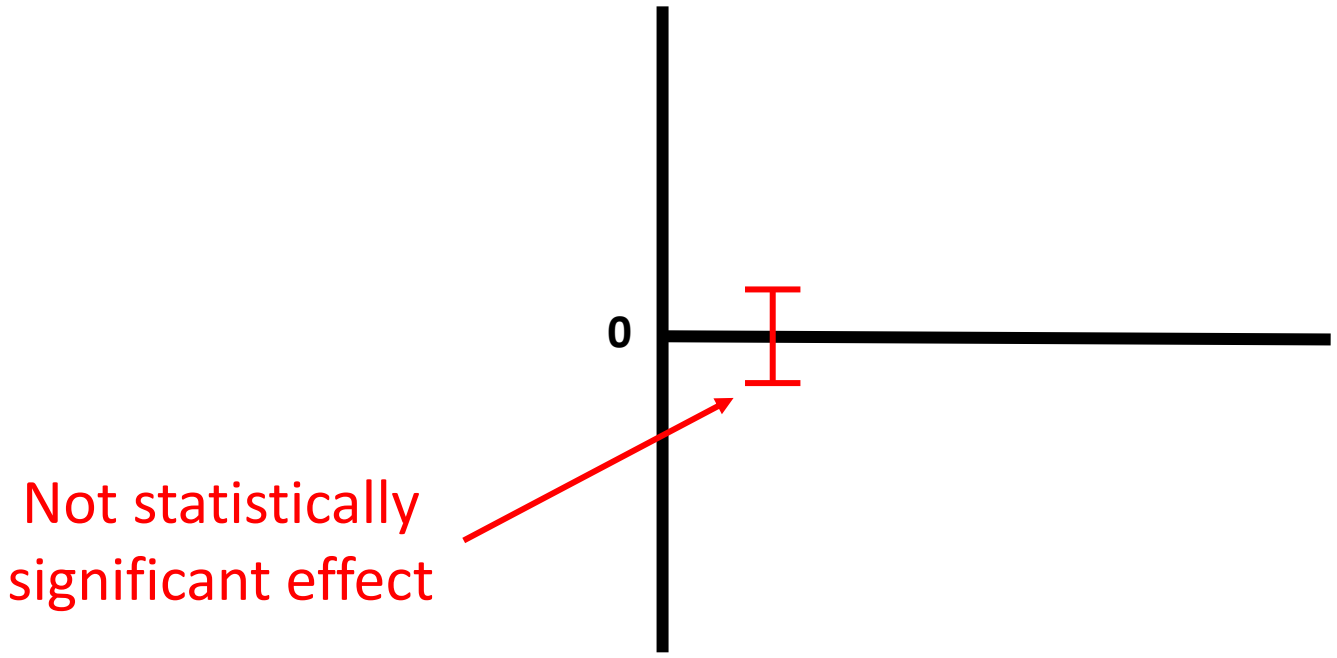
# Significance

When applied to a regression coefficient derived from a sample (randomly) extracted from a population, the term significant means that we can conclude, with a certain **confidence level (usually 95%)**, that, in the population equation, **the coefficient is nonzero**.

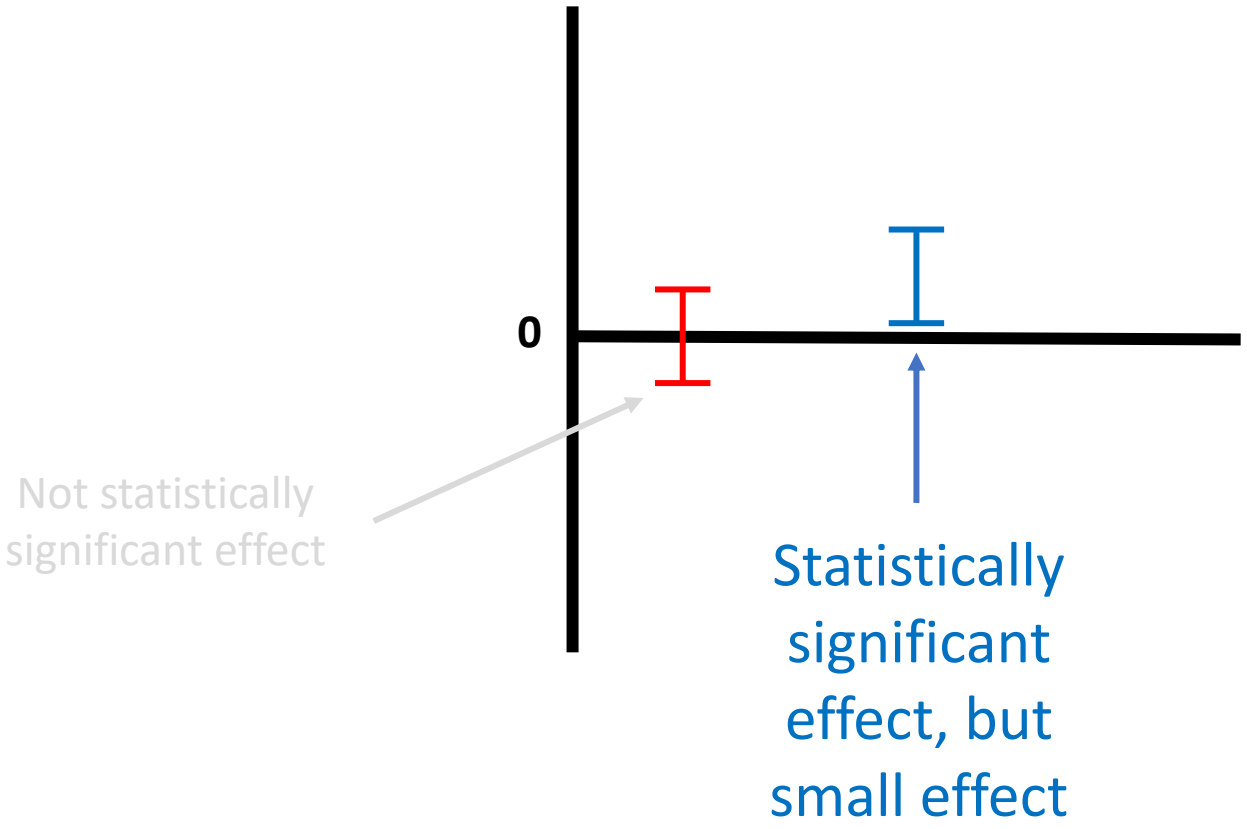
In other words,

**THAT THIS VARIABLE HAS A REAL EFFECT**
**(this variable matters!)**

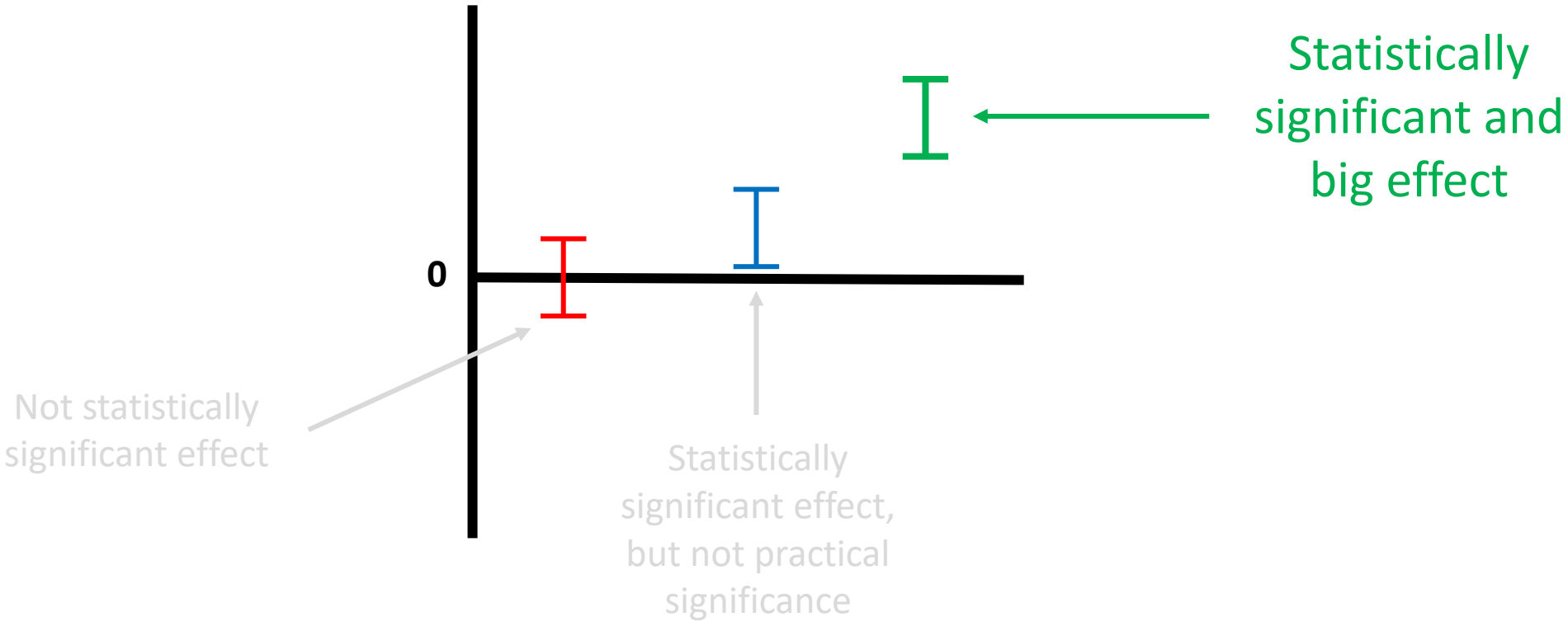
# Significance



# Significance



# Significance



# P-Values

Instead of the confidence limits, we can use the ***p*-value** to assess the significance of a variable.

The *p*-value is a probability (a number between 0 and 1) which is read as follows: **the lower *p*, the more significant the coefficient**.

By consensus, **a coefficient is considered to be significant when $p < 0.05$** (this is equivalent to setting 95% as the confidence level). source: U.S. Census Bureau

# P-value = Pr(>|t|) in R

```
Call:
lm(formula = dataset$Resistance ~ dataset$Water)

Residuals:
     Min       1Q   Median       3Q      Max
-157.467  -43.549   -0.175   39.885  219.183

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   211.88390   14.23989  14.880  < 2e-16 ***
dataset$Water   0.19617    0.06651   2.949  0.00328 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.74 on 802 degrees of freedom
Multiple R-squared:  0.01073,   Adjusted R-squared:  0.009497
F-statistic:   8.7 on 1 and 802 DF,  p-value: 0.003275
```

# P-Values

Using *p*-values is very easy,

**we just check whether they are small enough.**

*But understanding the mathematics behind them is more demanding, since calculating p-values involves integral calculus.*

# TIME TO PRACTICE!

# Orange Juice Pricing


**Minute Maid** is one of the world's most famous orange juice brands. It is now produced by the Coca-Cola company, the world's leading marketer of fruit juices and drinks.

# Orange Juice Pricing


**Jorge Luzio**, Minute Maid sales director, is worried about Aldi, a discount supermarket that is growing fast in the Paris area. Specially because he thinks that a real danger to Minute Maid comes from the **Aldi's branded orange juice**.

# Orange Juice Pricing


However, another manager at Minute Maid believes that Tropicana is their main competitor and, therefore, they consider that special attention should be paid to **Tropicana's pricing strategy**.

# Orange Juice Pricing


OK, let’s take a closer look at the Minute Maid, Aldi, and Tropicana prices to learn how they affect Minute Maid’s market share.



# Orange Juice Pricing


**Tropicana** produces two kinds of orange juice, the regular juice and the premium juice, whose prices are **$3.50** and **$4.45** per package.


**Aldi**'s current prices are **$2.20** per package


**Minute Maid** prices are **$3.99**.

In this setting, **Minute Maid's market share is 13.71%.**

# Orange Juice Pricing


However, when Jorge's team is collecting the data, they receive a memo warning that Tropicana plans to reduce substantially its prices.

While the price of the regular Tropicana juice would be **lowered at $3.25**, two rumours circulate about the premium brand: one is that the Tropicana Premium's price is going to be set to **$3.75**, and the other rumour says that it will be set at **$4.25**.

# Orange Juice Pricing – The Data


The data set (file orange.csv, sheet Data) contains information on **Minute Maid's market share** (percentage scale) and **prices** (dollars per package) of Minute Maid and its competitors, covering 121 weeks.

**CALCULATE THE MEAN AND STANDARD DEVIATION OF EACH VARIABLE**

| | MShare | TropPremium | Trop | MMaid | Aldi |
|---|---|---|---|---|---|
| Mean | 17.27 | 4.39 | 3.39 | 3.29 | 2.74 |
| St. deviation | 6.72 | 0.54 | 0.54 | 0.58 | 0.57 |

**Who is the main competitor of Minute Maid?**

# Orange Juice Pricing


**How does Minute Maid's price affect its market share?**

Take a simple linear regression approach to this question. How Mshare changes when (dependent variable) when we change MMaid price (independent variable)?

$$MShare = ? - ? \, MMaid$$

$$R = ?$$

# Orange Juice Pricing


**How does Minute Maid's price affect its market share?**

Take a simple linear regression approach to this question. How Mshare changes when (dependent variable) when we change MMaid price (independent variable)?

This means that an increase of 1 cent in the price leads, on average, to a loss of 0.07% in the market share of Minute Maid.

$$MShare = 40.01 - 6.91\ MMaid$$

$$R = 0.597$$

# Orange Juice Pricing

**How does Minute Maid's market share is affected by its price and the price of its competitors?**

Take a multiple linear regression approach to this question.

$$\text{MShare} = ? + ? \text{ TropPremium} - ? \text{ Trop} - ? \text{ MMaid} + ? \text{ Aldi}$$

$$R = ?$$

# Orange Juice Pricing

## How does Minute Maid's market share is affected by its price and the price of its competitors?

Take a multiple linear regression approach to this question.

MShare = **11.41** + **8.40** TropPremium – **4.15** Trop – **8.57** Mmaid + **4.13** Aldi

$R$ = **0.750**

It went up from 0.597 to 0.750, so we increased our predictive power!

# Orange Juice Pricing

## How do you interpret the results?

```
Call:
lm(formula = Mshare ~ TropPr                          ldi, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9771 -2.9296 -0.6031  2.3847 16.6914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4101     8.2955   1.375    0.172
TropPremium   8.3962     6.9659   1.205    0.231
Trop         -4.1488     6.9982  -0.593    0.554
MMaid        -8.5738     0.7558 -11.343  < 2e-16 ***
Aldi          4.1263     0.7390   5.584 1.57e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.519 on 116 degrees of freedom
Multiple R-squared:  0.5627,    Adjusted R-squared:  0.5476
F-statistic: 37.32 on 4 and 116 DF,  p-value: < 2.2e-16
```

What about the signs?
And the effect size?

**What about the signs?**
When TropPremium and Aldi increase their price, MMaid gains market share (positive sign). The other direction for Trop and MMaid.

**And the effect size?**
Both MMaid own price and TropPremium have the biggest effect on MMaid market share.

# Orange Juice Pricing

## How do you interpret the results?

```
Call:
lm(formula = Mshare ~ TropPremium + Trop + MMaid + Aldi, data = dataset)

Residuals:
    Min      1Q  Median      3Q
-8.9771 -2.9296 -0.6031  2.3847 16.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4101     8.2955   1.375    0.172
TropPremium   8.3962     6.9659   1.205    0.231
Trop         -4.1488     6.9982  -0.593    0.554
MMaid        -8.5738     0.7558 -11.343  < 2e-16 ***
Aldi          4.1263     0.7390   5.584 1.57e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.519 on 116 degrees of freedom
Multiple R-squared:  0.5627,    Adjusted R-squared:  0.5476
F-statistic: 37.32 on 4 and 116 DF,  p-value: < 2.2e-16
```

**What about the significance?**
TropPremium and Trop are not significant (in other words, it does not matter what TropPremium and Trop do, they do not affect MMaid market share).

What about the
significance of variables?

# Lower confidence interval and upper confidence interval

```
> model_parameters(z1, summary = TRUE)
Parameter    | Coefficient |   SE |         95% CI | t(116) |      p
-----------------------------------------------------------------------
(Intercept)  |       11.41 | 8.30 | [ -5.02, 27.84] |   1.38 | 0.172
TropPremium  |        8.40 | 6.97 | [ -5.40, 22.19] |   1.21 | 0.231
Trop         |       -4.15 | 7.00 | [-18.01,  9.71] |  -0.59 | 0.554
MMaid        |       -8.57 | 0.76 | [-10.07, -7.08] | -11.34 | < .001
Aldi         |        4.13 | 0.74 | [  2.66,  5.59] |   5.58 | < .001

Model: Mshare ~ TropPremium + Trop + MMaid + Aldi (121 Observations)
Residual standard deviation: 4.519 (df = 116)
R2: 0.563; adjusted R2: 0.548
```

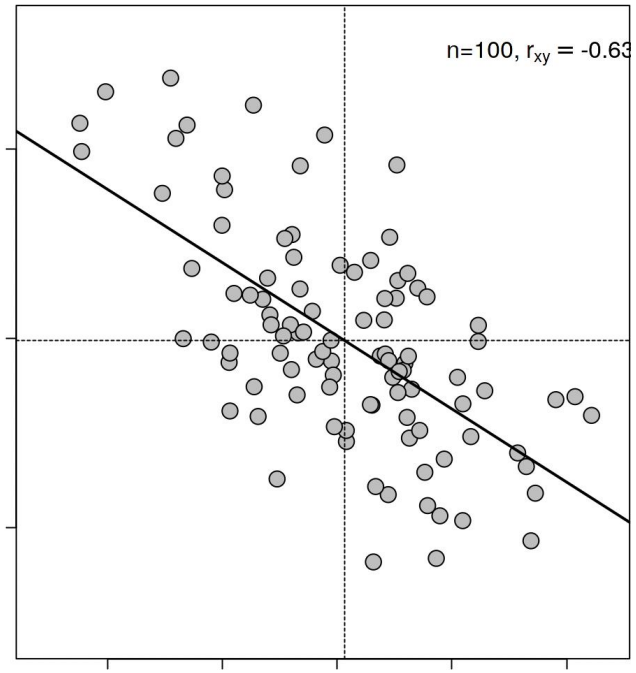
# Multiple Regression Analysis

# 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~~