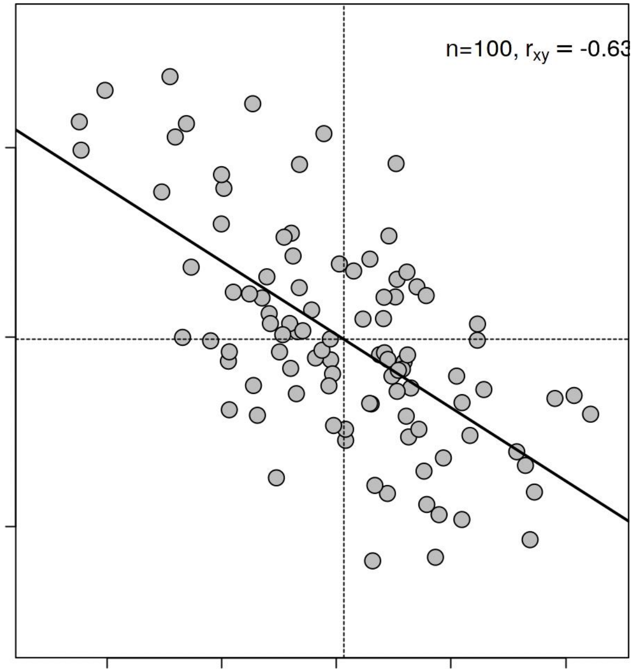
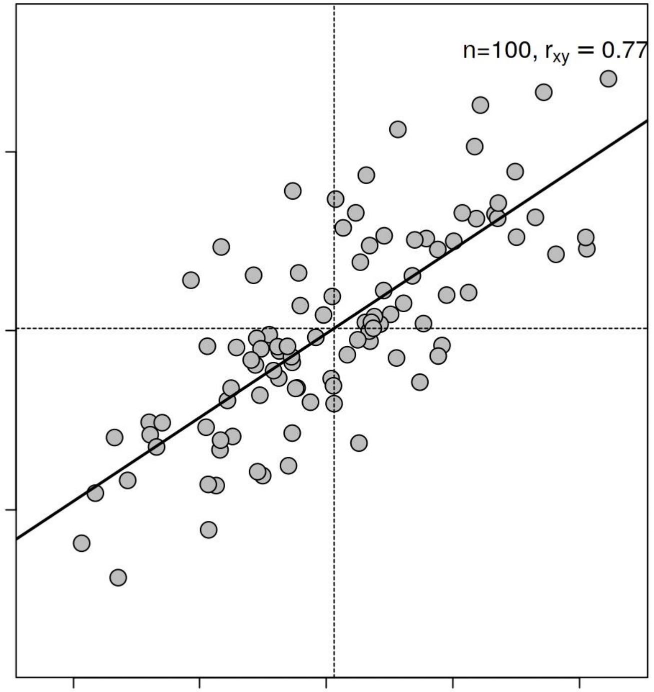
Fundamentals of Econometrics Models



Vicenç Soler

v.soler@tbs-education.org

~~vincent.soler@tbs-education.org~~



The essentials of regression

Do taller people weight more than shorter people?

The essentials of regression

Do taller people weight more than shorter people?

By how much?

The essentials of regression

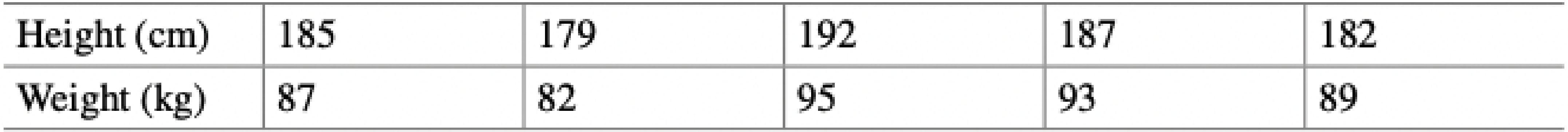
Do taller people weight more than shorter people?

By how much?

If I tell you how tall somebody is, can you guess -approximately- how much he weights?

The essentials of regression

Suppose that we observe the **height** and the **weight** in a sample of five 40-year-old men

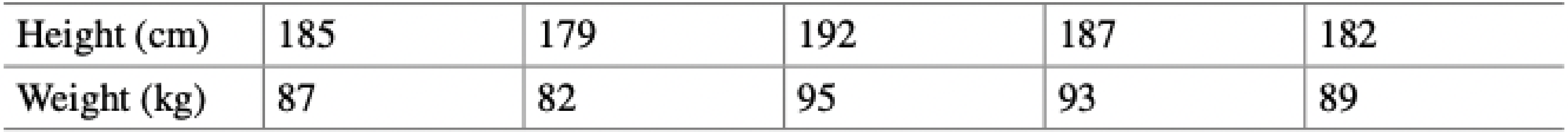


Height (cm)	185	179	192	187	182
Weight (kg)	87	82	95	93	89

The essentials of regression

In regression, we are interested in **explaining how a variable changes in relation to another variable.** The concepts are:

DEPENDENT VARIABLE: The variable we try to explain or predict
INDEPENDENT VARIABLE: The variable we use to explain or predict



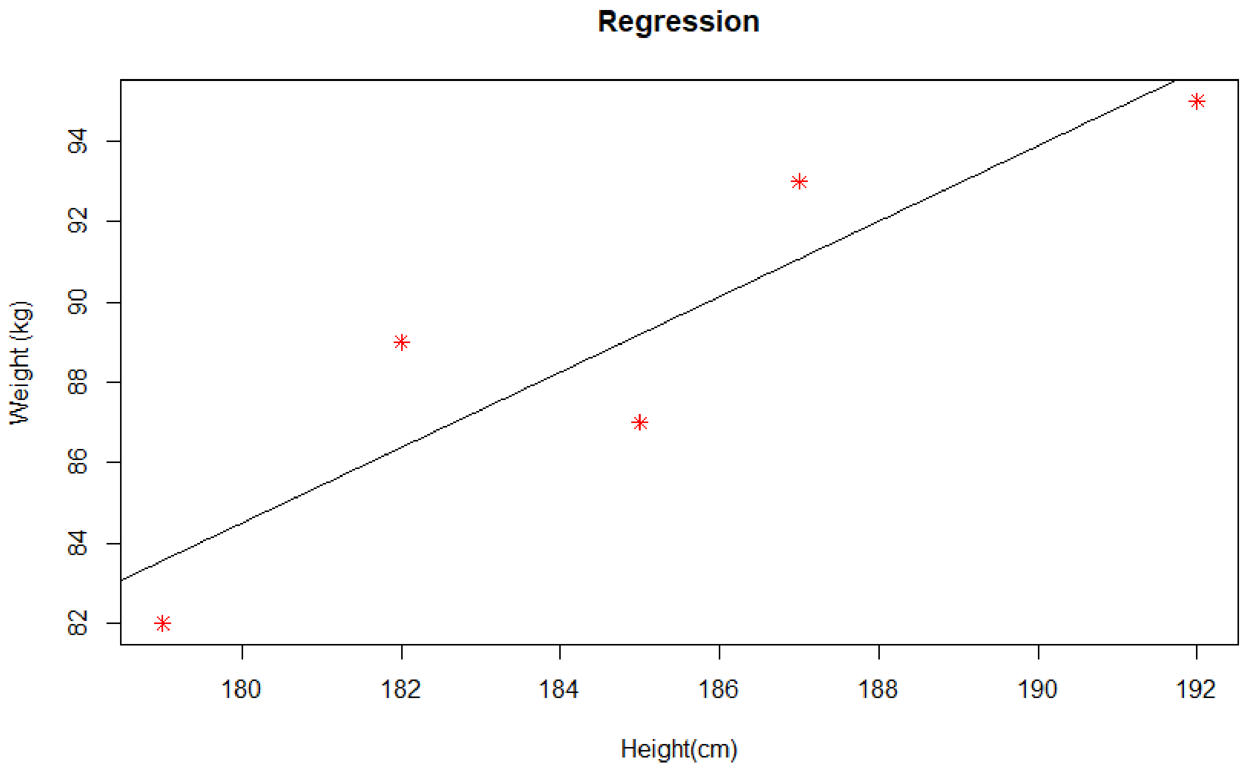
Height (cm)	185	179	192	187	182
Weight (kg)	87	82	95	93	89

In this case, we try to explain WEIGHT based on HEIGHT

The essentials of regression

Suppose that we observe the **height** and the **weight** in a sample of five 40-year-old men

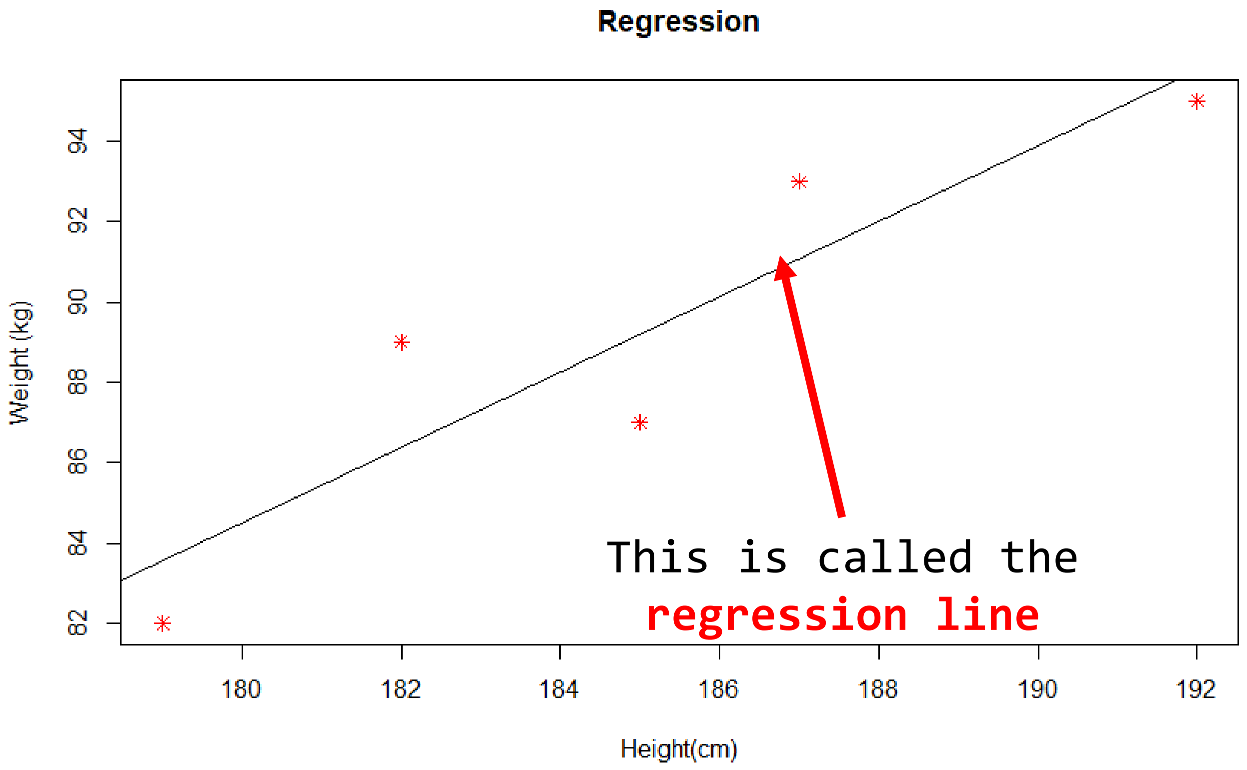
Putting the height in the horizontal axis and the weight in the vertical axis, we can plot these observations as the following five points:



This representation is called a scatter plot. Scatter plots are typically used to explore the association between two variables.

The essentials of regression

The scatter plot suggests a positive relationship between height and weight, meaning that, the more height, the more weight.



This is called the **regression line**

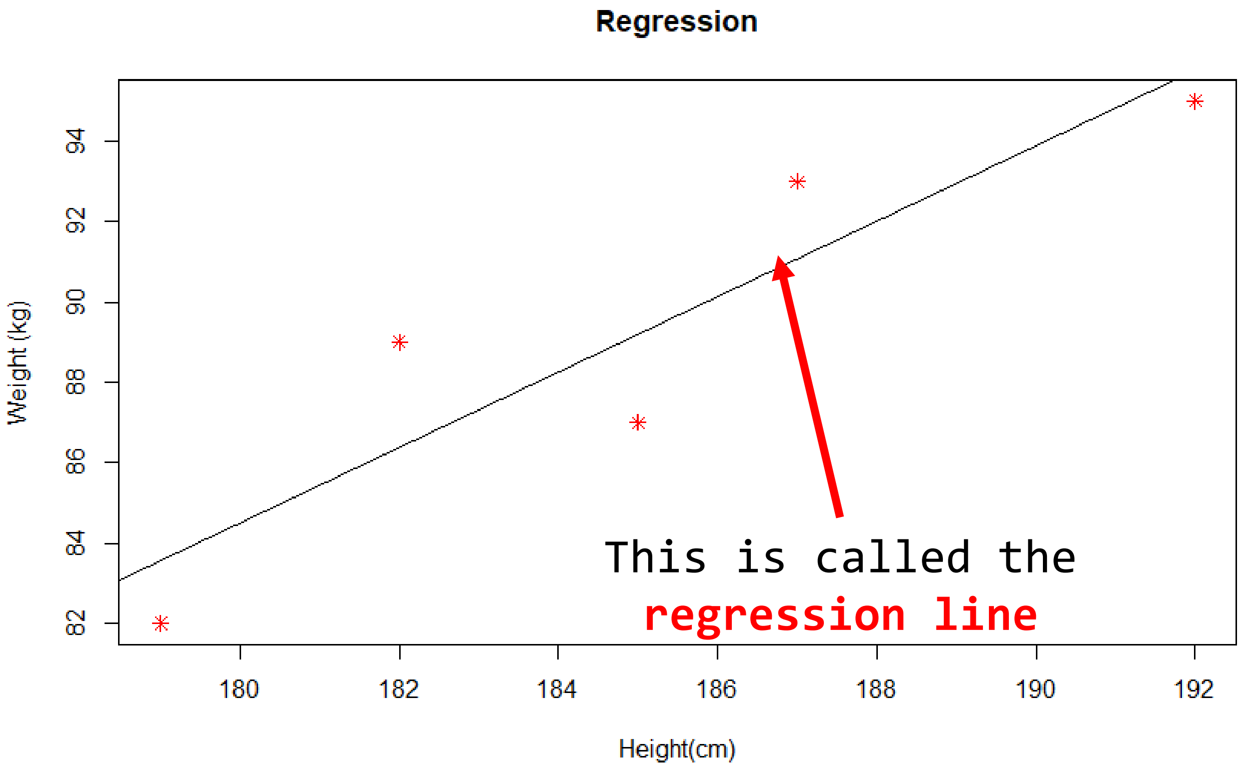
The essentials of regression

The scatter plot suggests a positive relationship between height and weight, meaning that, the more height, the more weight.

The regression line can be expressed, in mathematical terms, as an equation of the form

$$y = a + bx$$

called the **regression equation**.



This is called the **regression line**

The essentials of regression

The scatter plot suggests a positive relationship between height and weight, meaning that, the more height, the more weight.

The regression line can be expressed, in mathematical terms, as an equation of the form

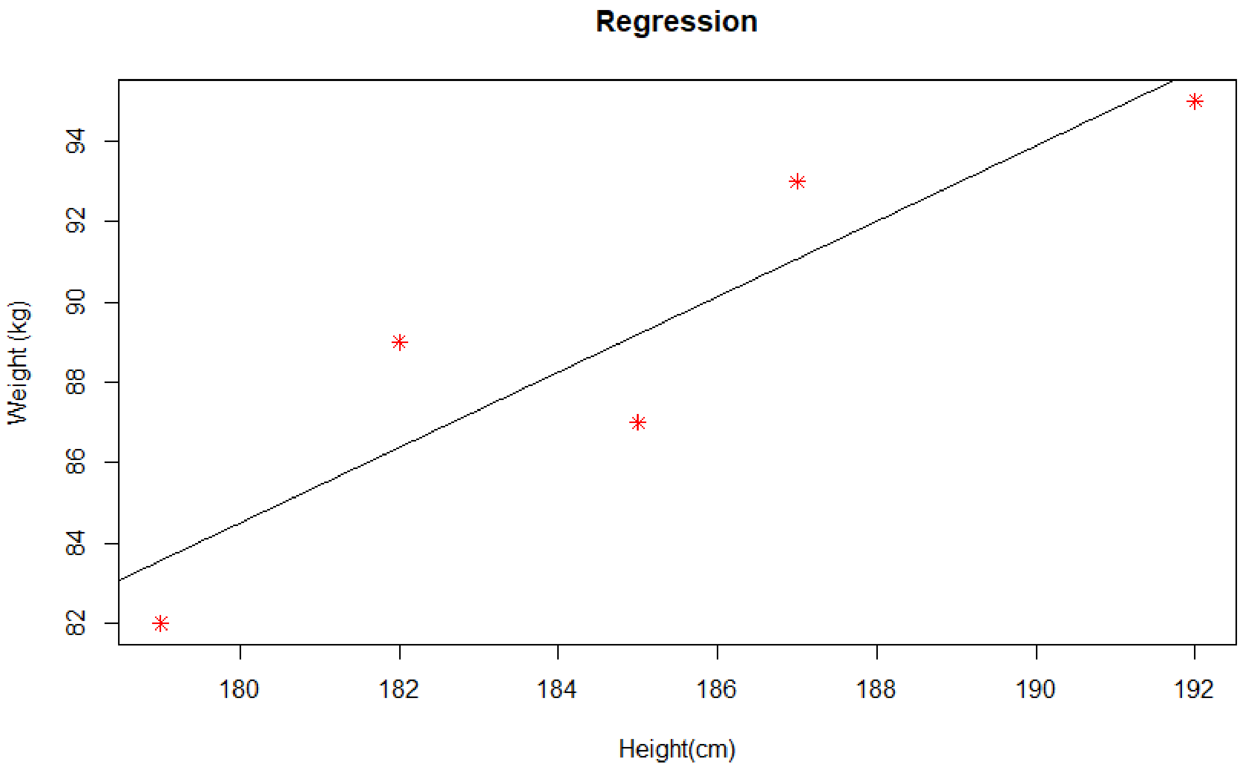
$$y = a + bx$$

called the **regression equation**.

In this figure,

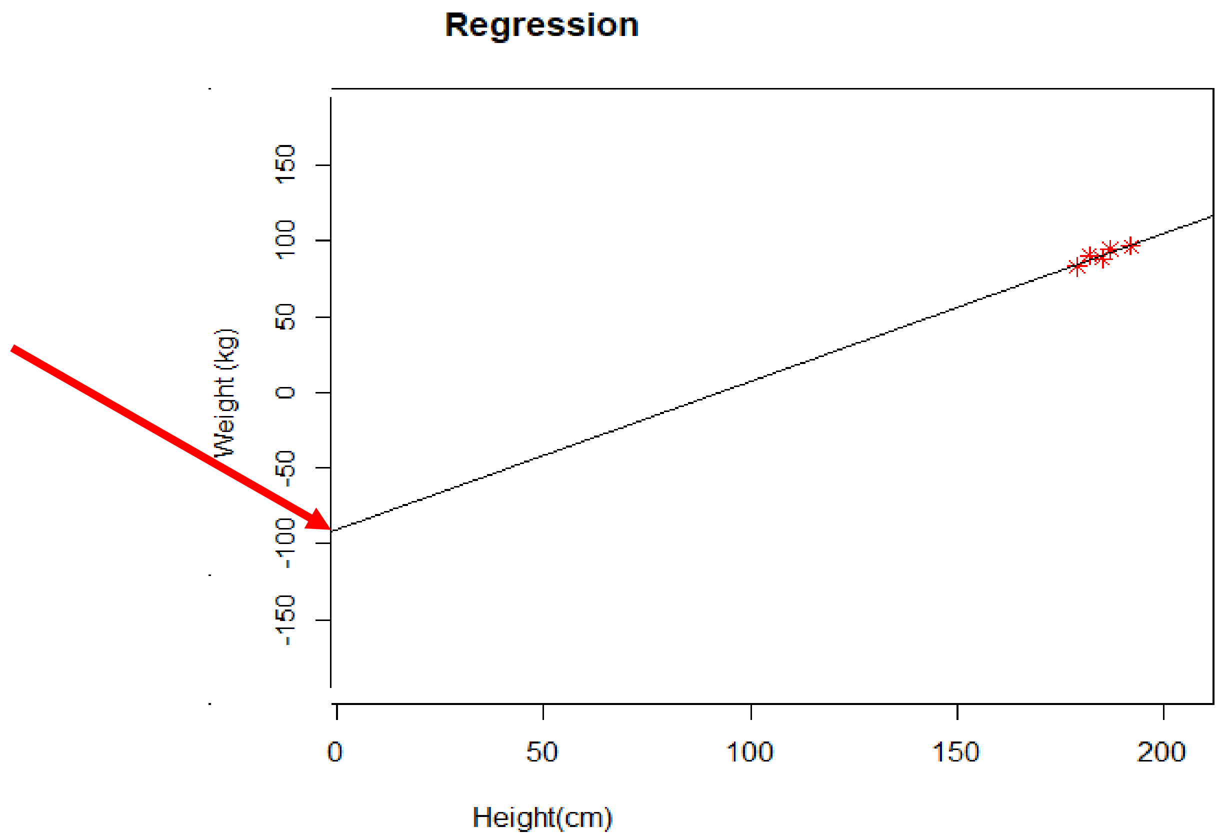
a = -83.47

This number (a) is called the INTERCEPT.



The essentials of regression

It tells you where the line starts.



Not always starts at 0, since having 0 weight/height does not make sense.

The essentials of regression

The scatter plot suggests a positive relationship between height and weight, meaning that, the more height, the more weight.

The regression line can be expressed, in mathematical terms, as an equation of the form

$$y = a + bx$$

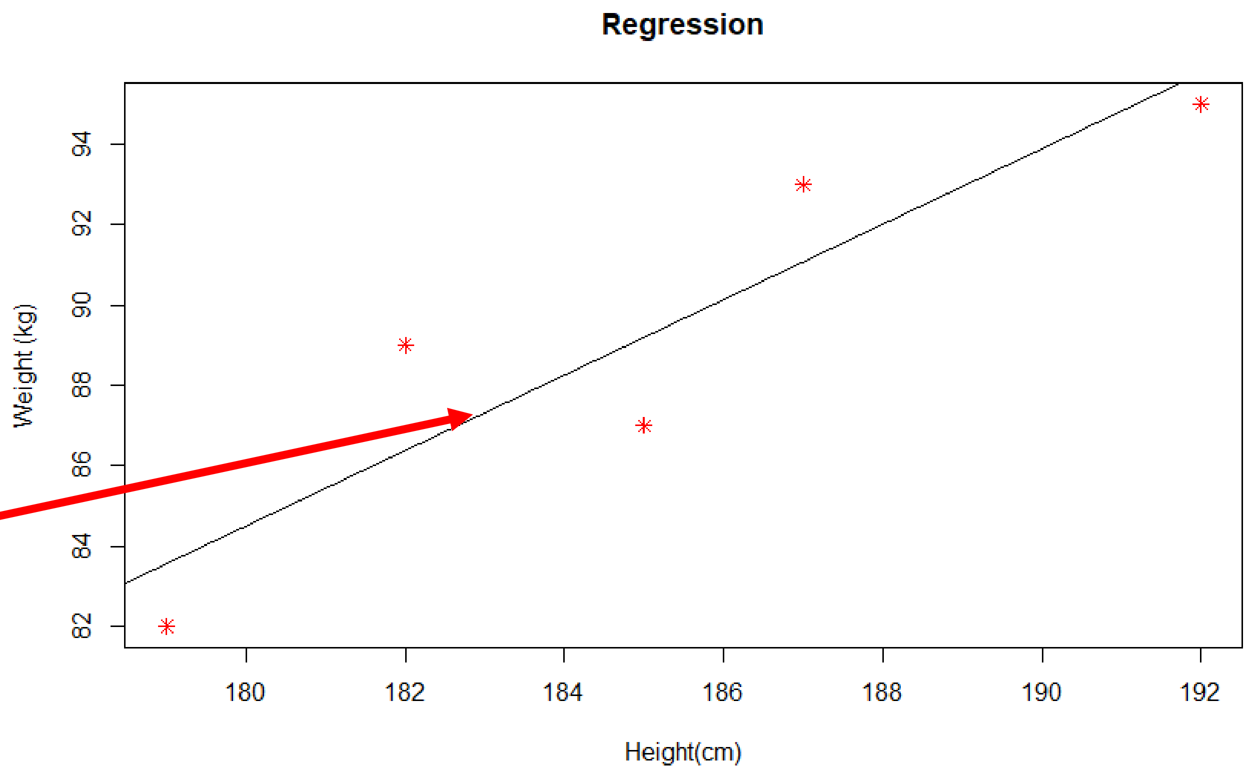
called the **regression equation**.

In this figure,

a = -83.47 and
b = 0.94.

This number (b or β) is called the SLOPE.

It tells you by how much Y will increase if we increase X.



The essentials of regression

The scatter plot suggests a positive relationship between height and weight, meaning that, the more height, the more weight.

The regression line can be expressed, in mathematical terms, as an equation of the form

$$y = a + bx$$

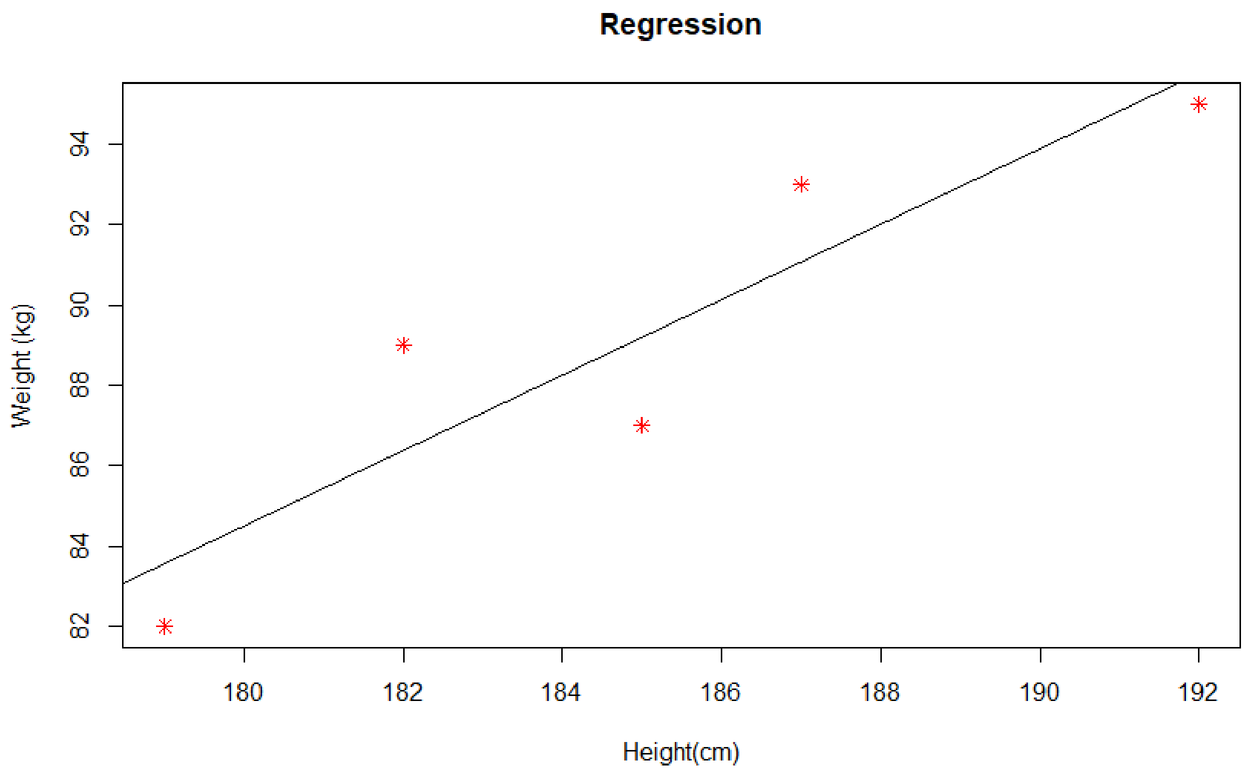
called the **regression equation**.

In this figure,

a = -83.47 and
b = 0.94.

So, the regression equation is

Weight = -83.47 + 0.94 Height.

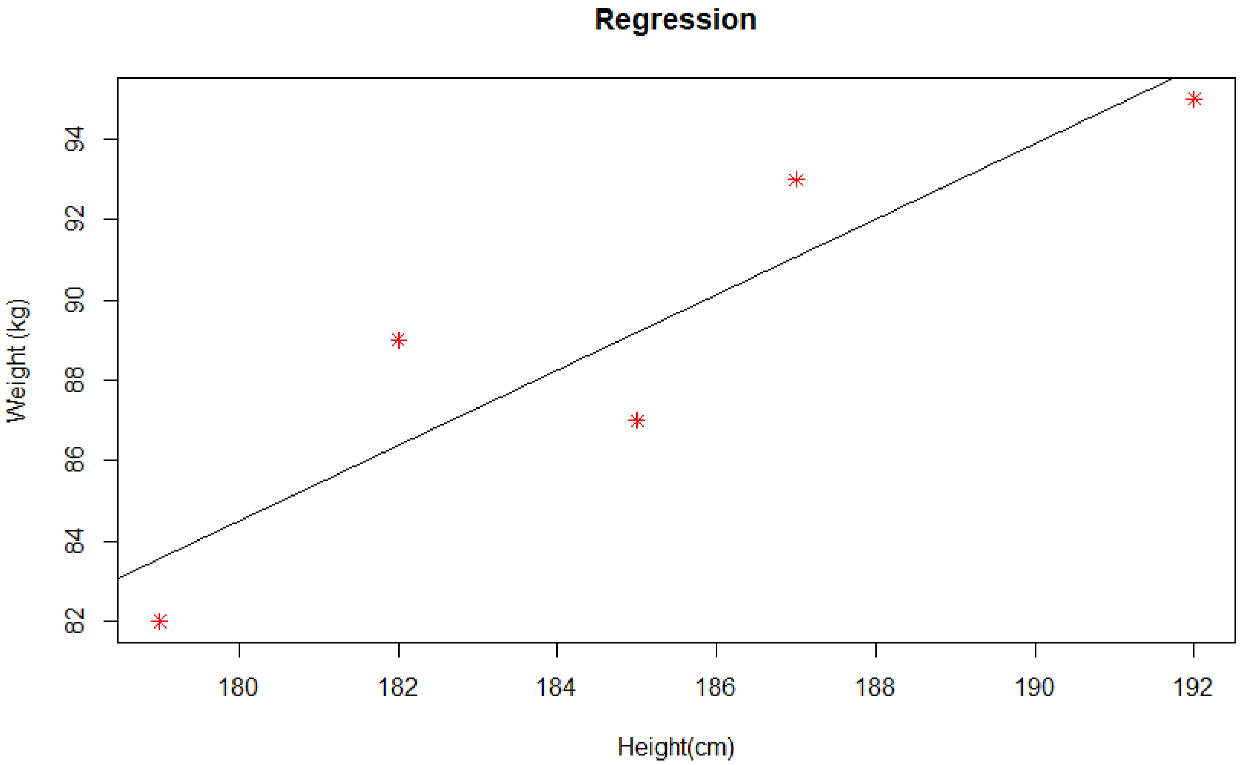


The essentials of regression

So we have

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

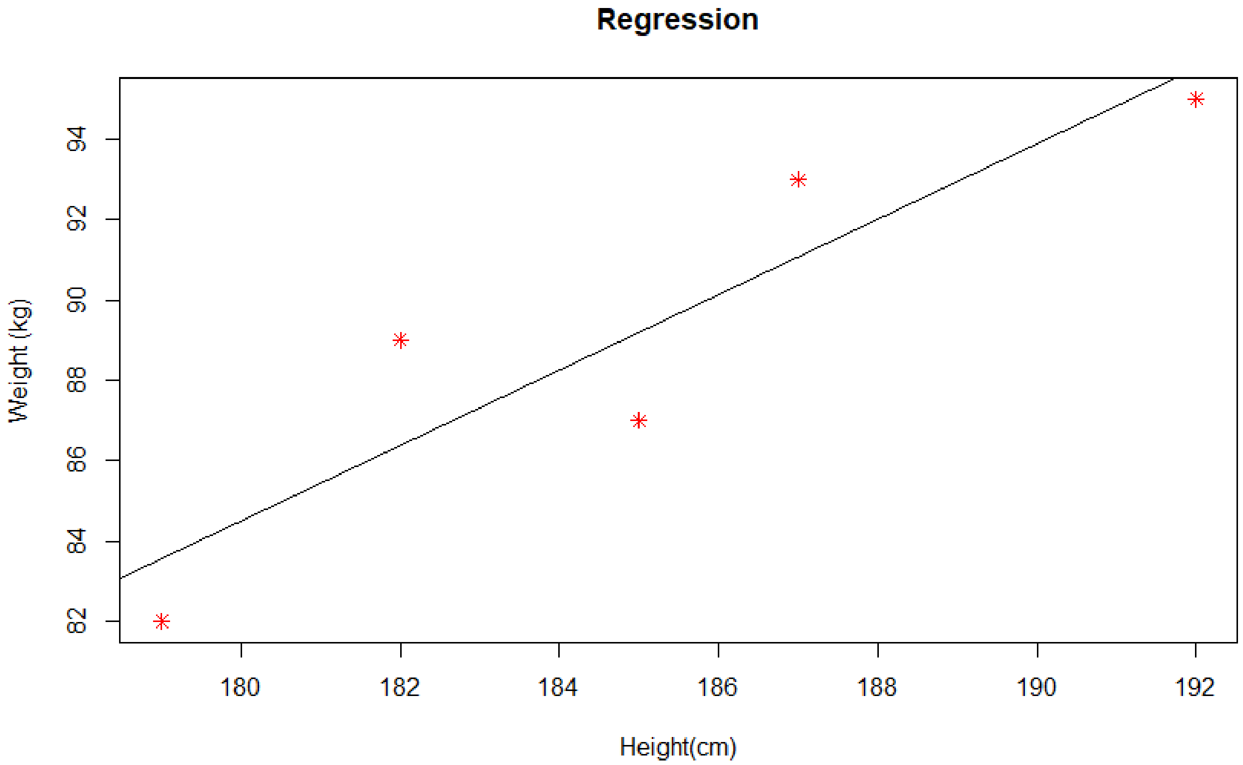
This equation is a **predictive model**!



The essentials of regression

So we have

*Weight = -83.47 + 0.94 * Height*

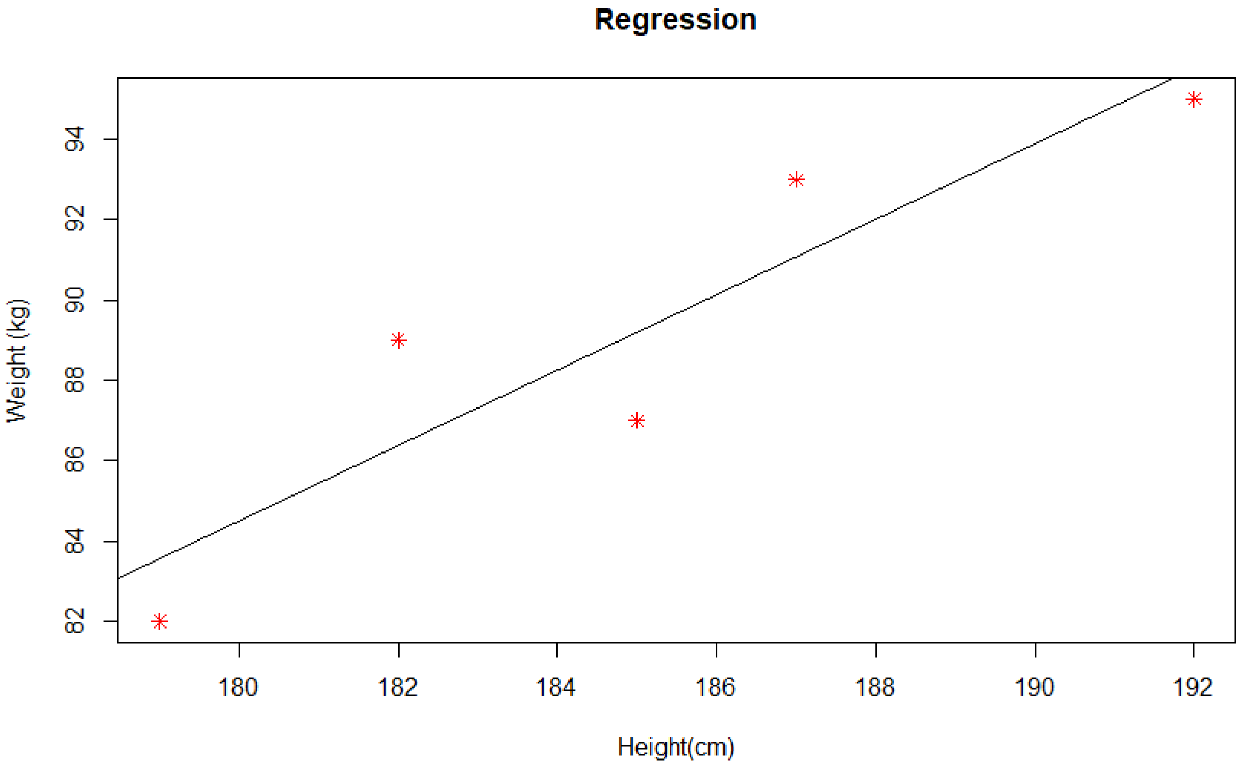


The essentials of regression

So we have

*Weight = -83.47 + 0.94 * Height*

If somebody has a **height of 185cm,** how much would he **weight?**

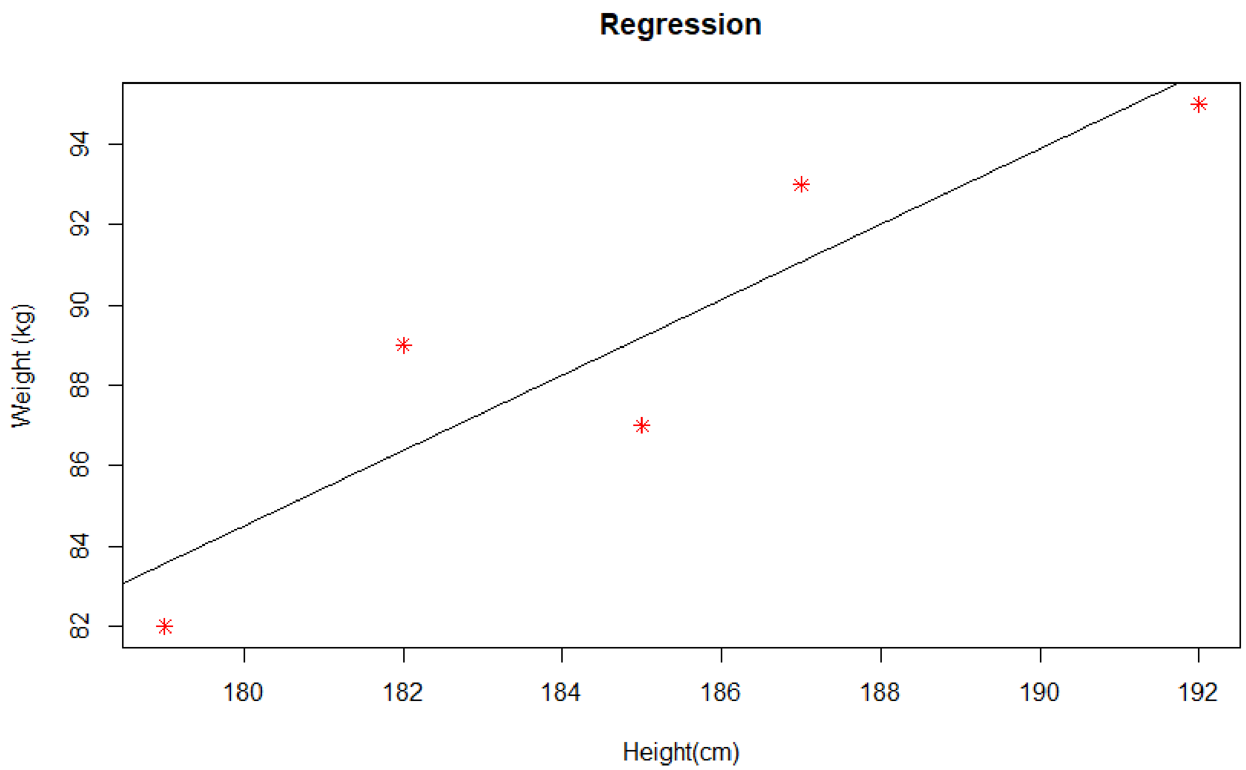


The essentials of regression

If somebody has a **height of 185cm**, how much would he **weight?**

In theory, **89.2kg**

But check the plot. What is wrong?

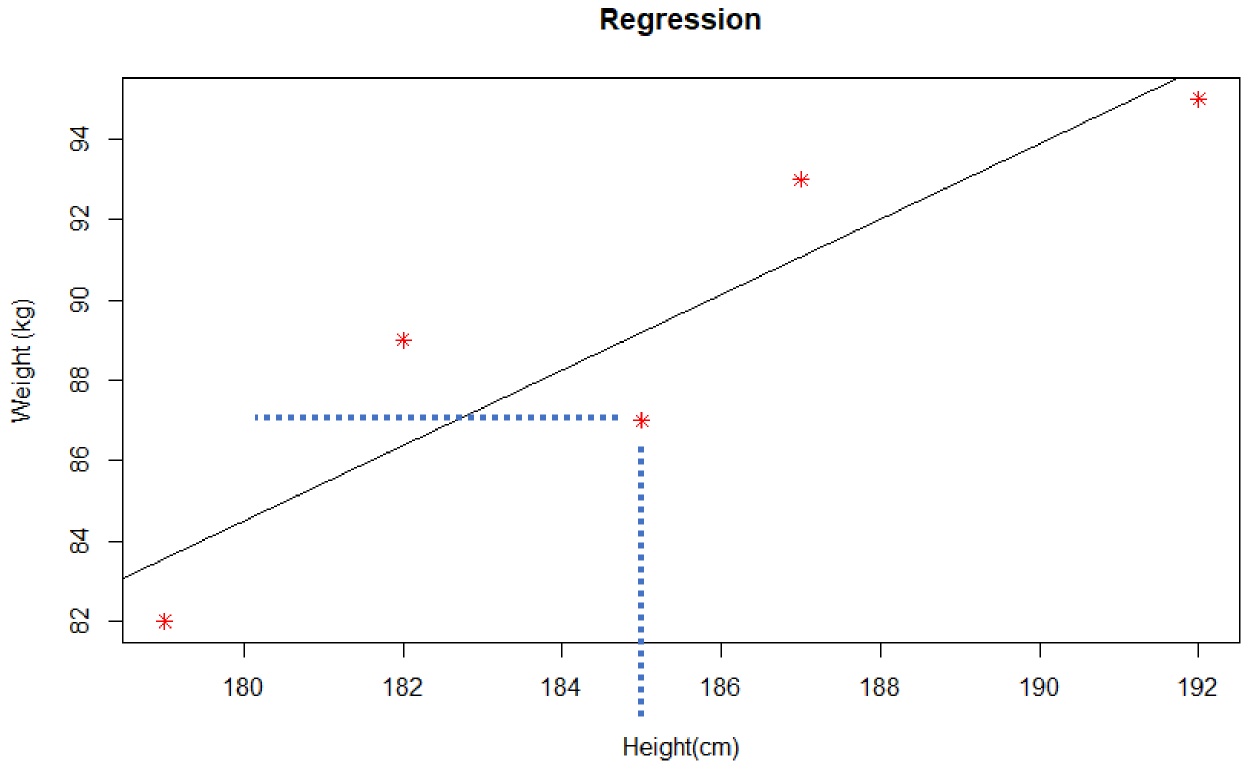


The essentials of regression

We predicted, **89.2kg**

But he actually weights **87kg**

Every prediction has an error!



The essentials of regression

Every prediction has an error!

Prediction error

=

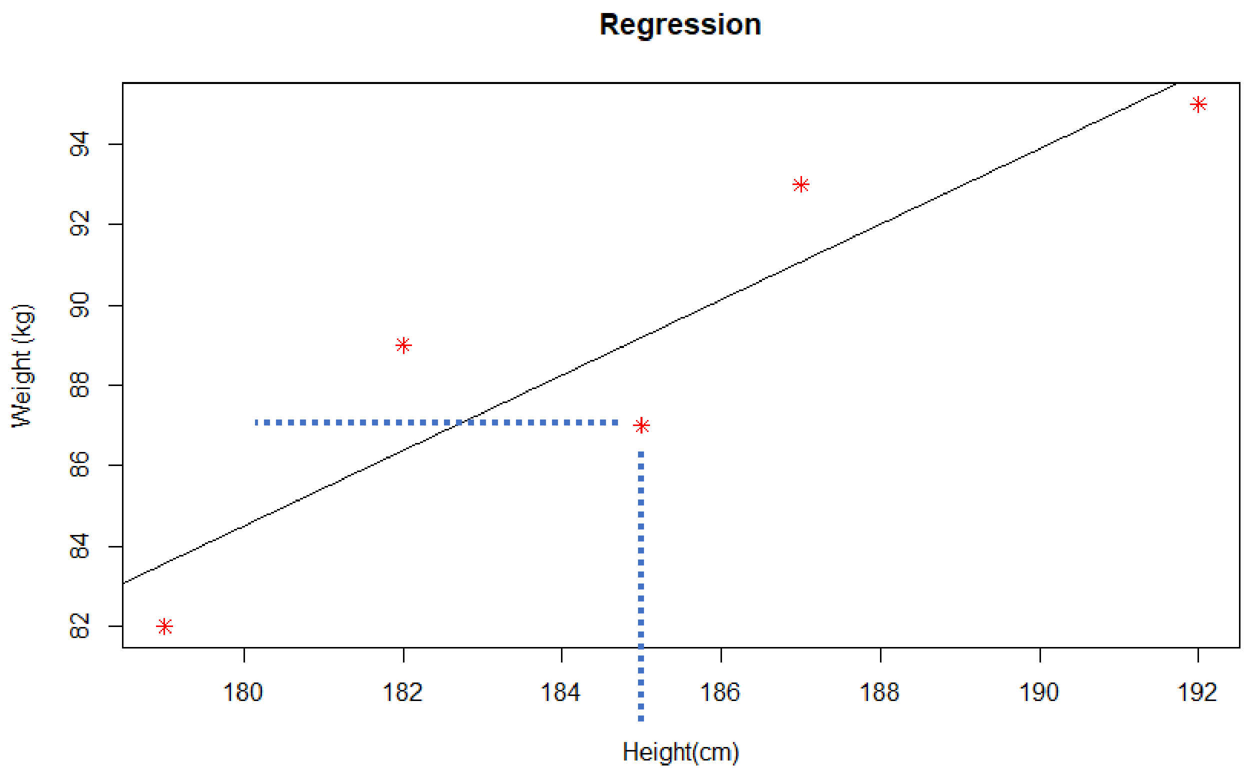
Actual Weight – Predicted Weight

=

87kg – 89.2kg

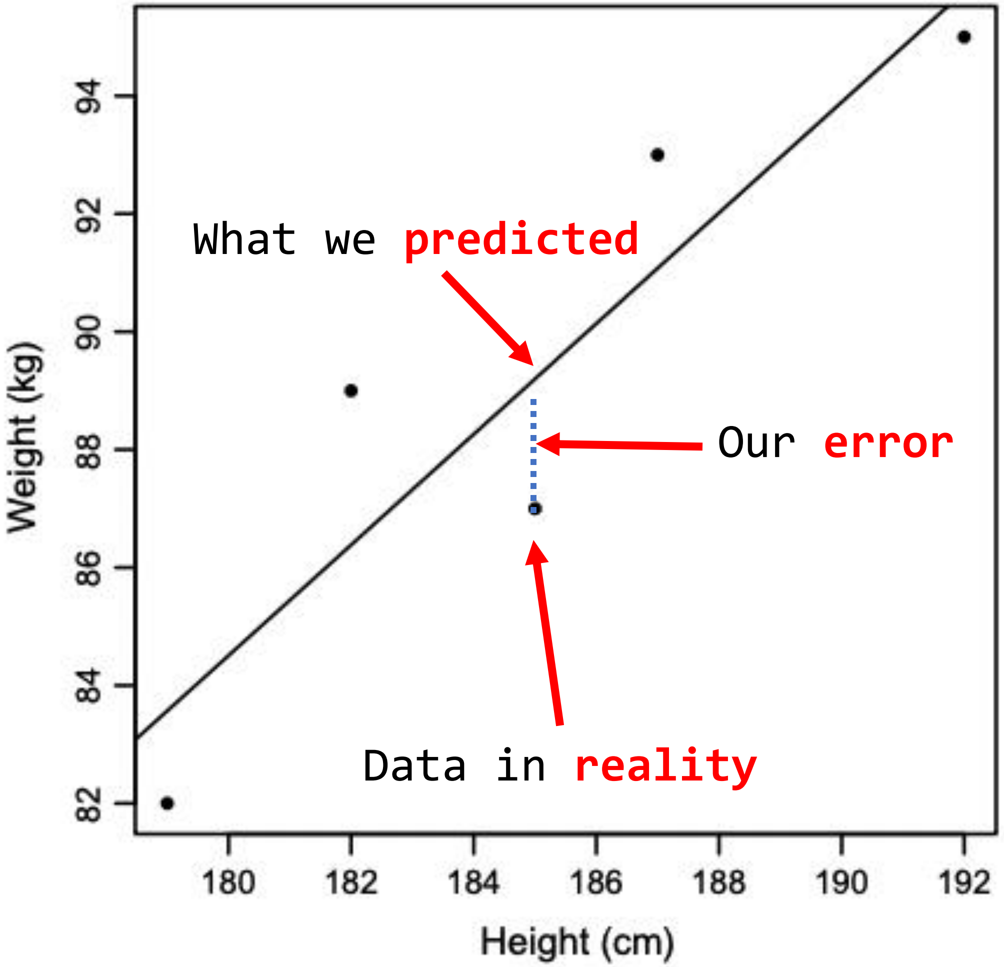
=

-2.2kg

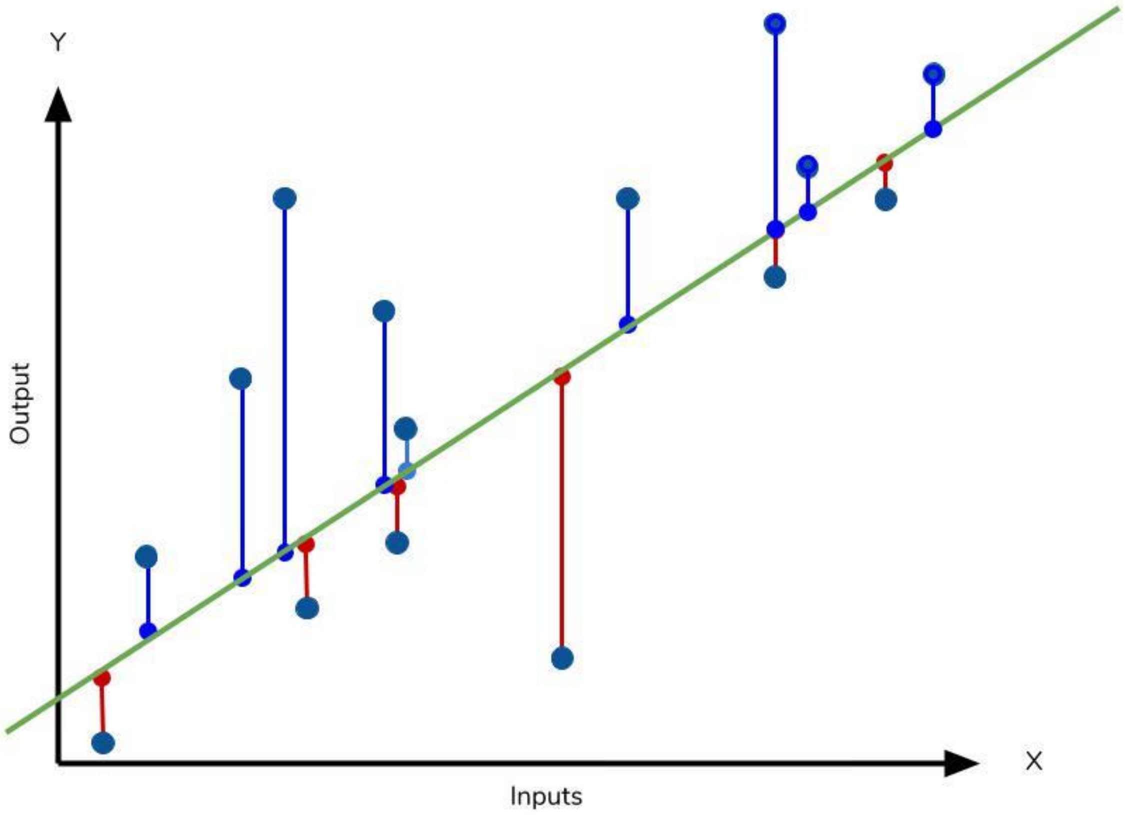


The essentials of regression

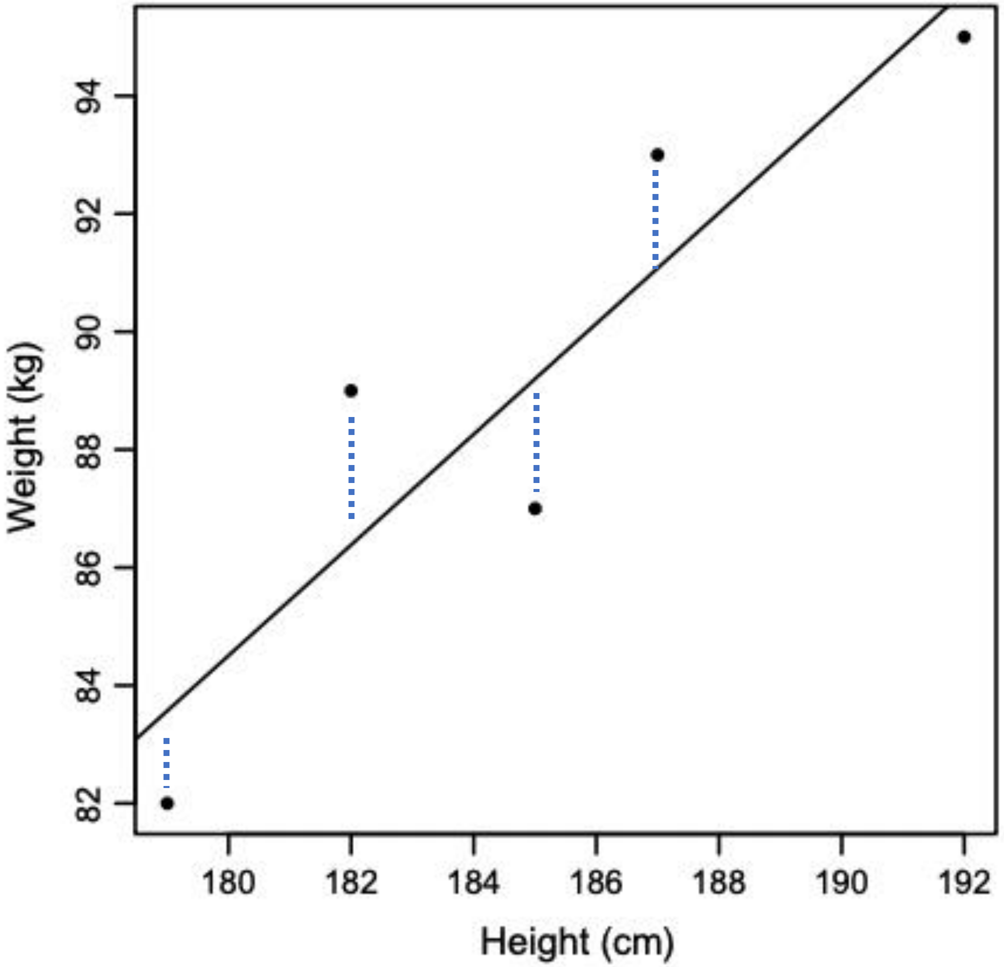
The smaller the error,
the better our
predictive model.



The essentials of regression



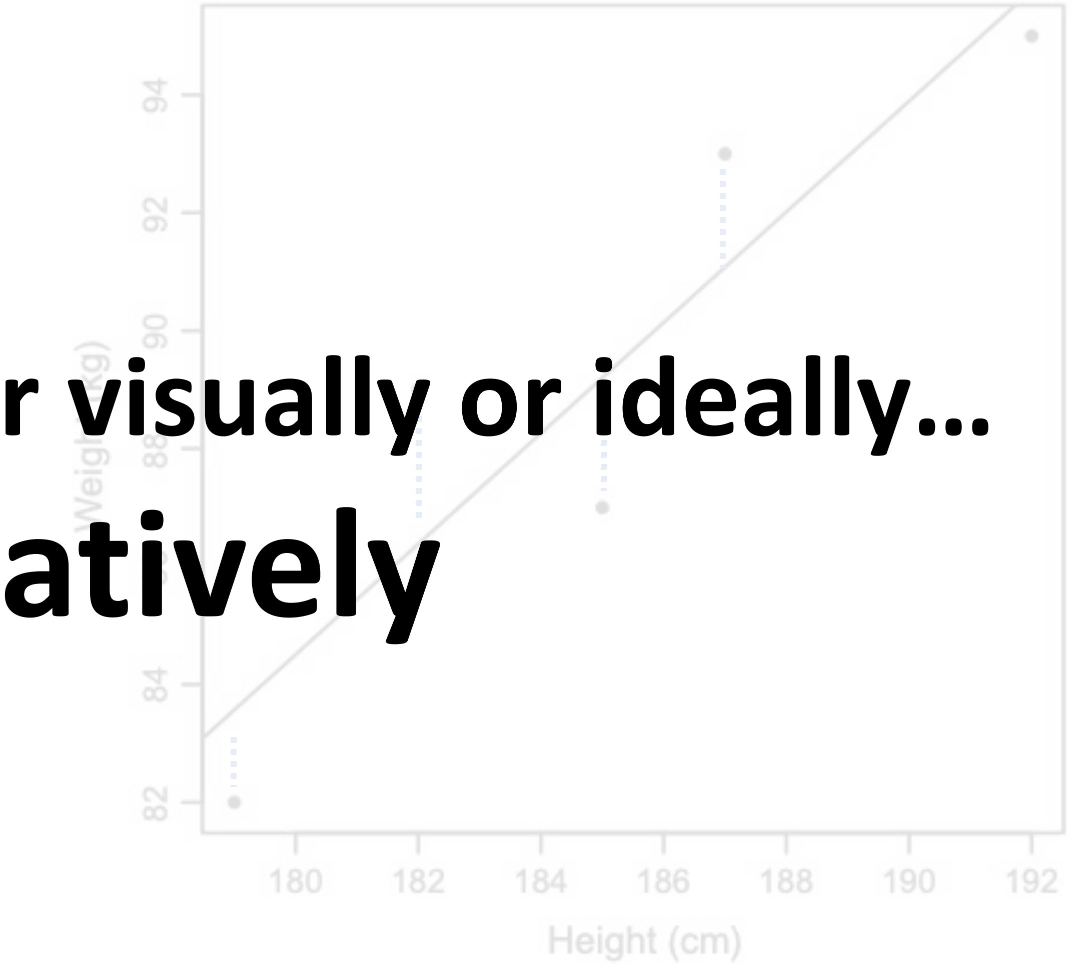
Bigger error



Smaller error

The essentials of regression

Y
Output
X
Inputs
MPE tells us if there's
more positive errors
than negative, or
vice versa



We can measure error visually or ideally…

quantitatively

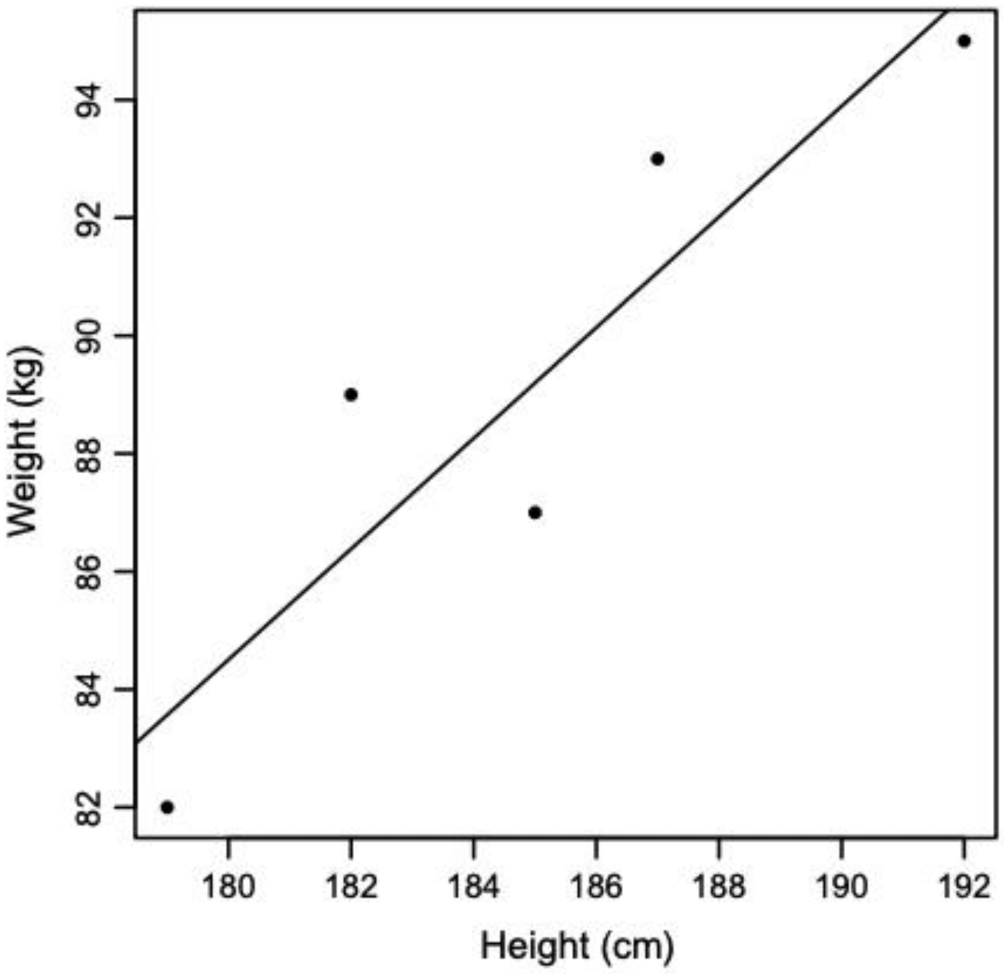
Bigger error

Smaller error

The essentials of regression

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

CORRELATION COEFFICIENT

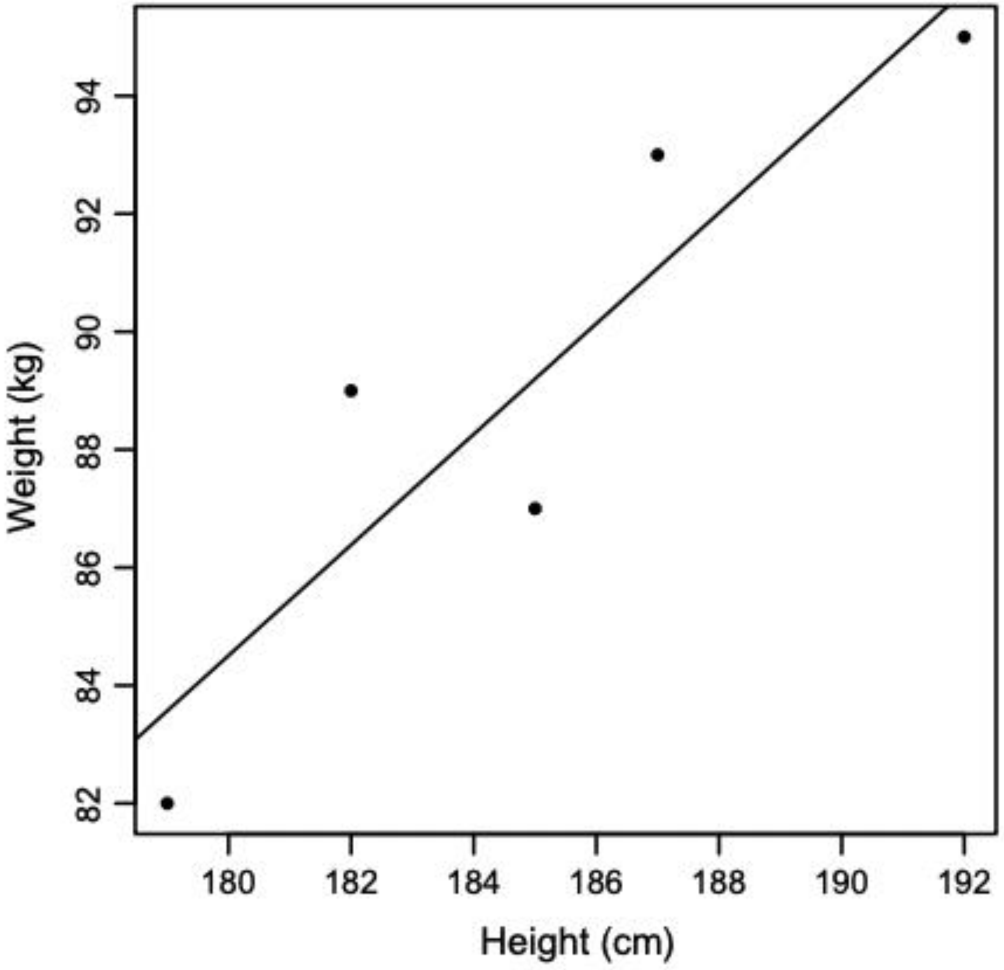


The essentials of regression

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

CORRELATION COEFFICIENT

In this case, the correlation (R) is 0.908

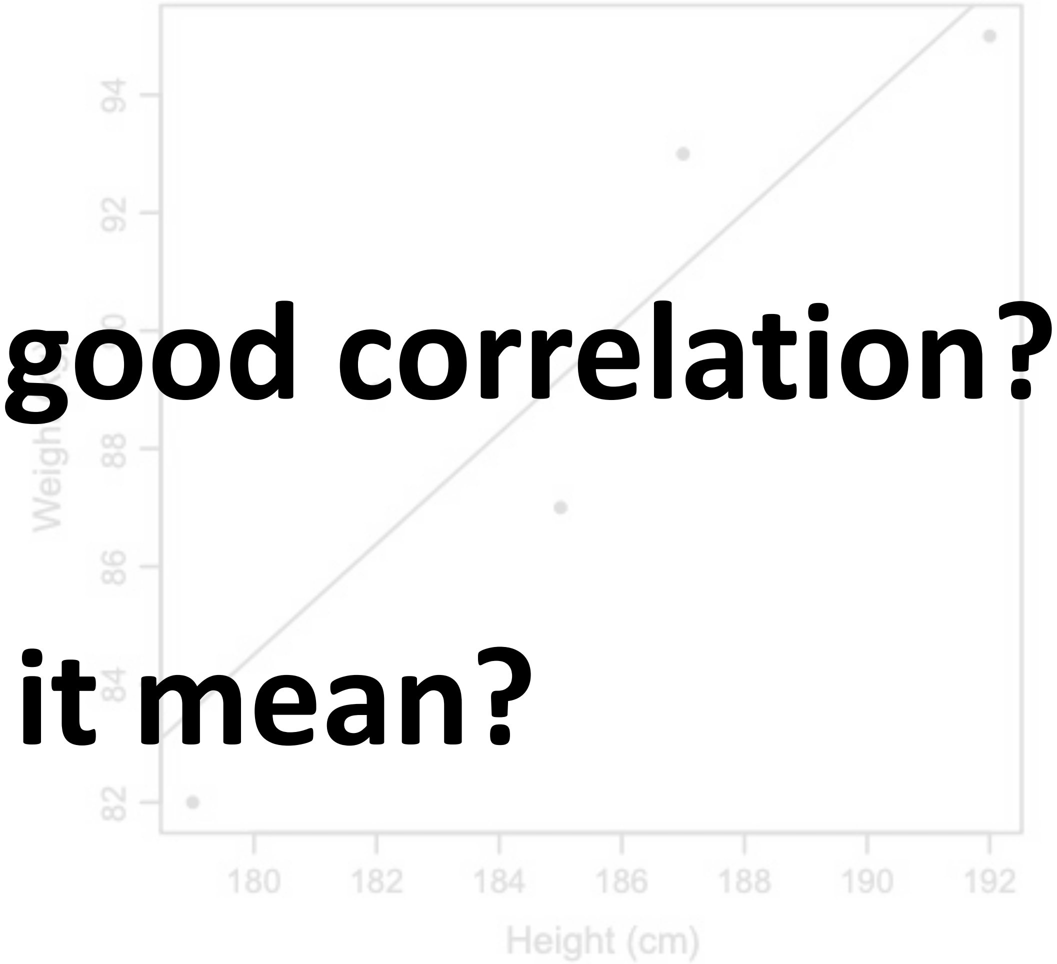


The essentials of regression

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

CORRELATION

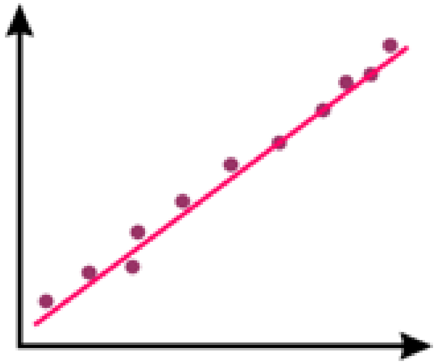
In this case, the correlation (R) is 0.908



However, is 0.908 a good correlation?

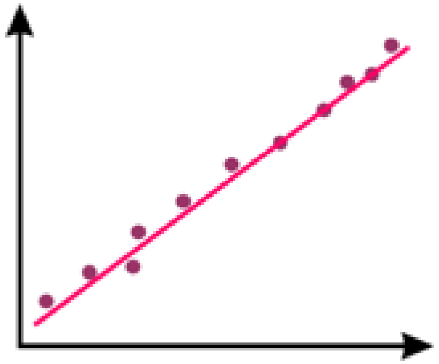
What does it mean?

The essentials of regression

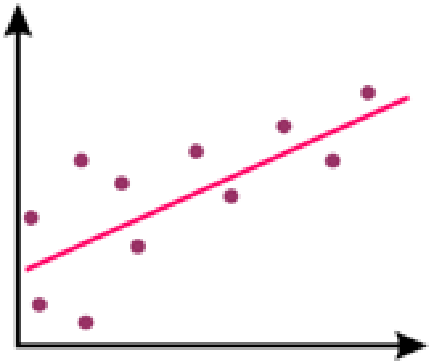


STRONG POSITIVE CORRELATION

The essentials of regression

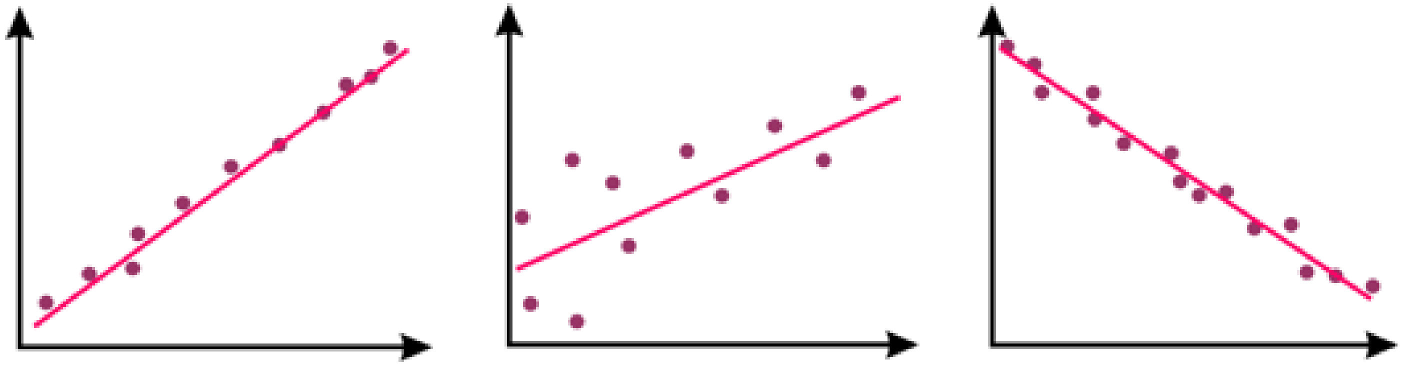


STRONG POSITIVE CORRELATION



WEAK POSITIVE CORRELATION

The essentials of regression

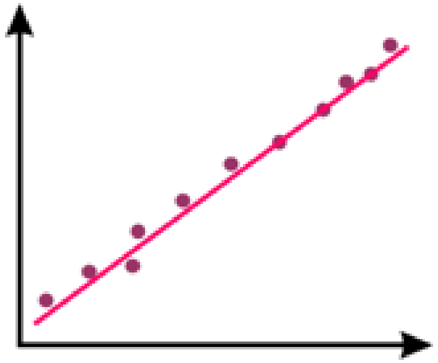


STRONG POSITIVE CORRELATION

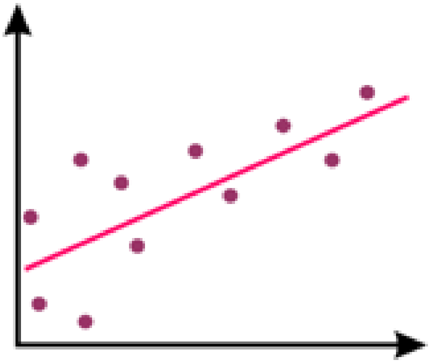
WEAK POSITIVE CORRELATION

STRONG NEGATIVE CORRELATION

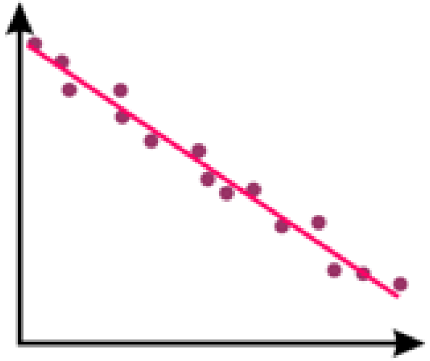
The essentials of regression



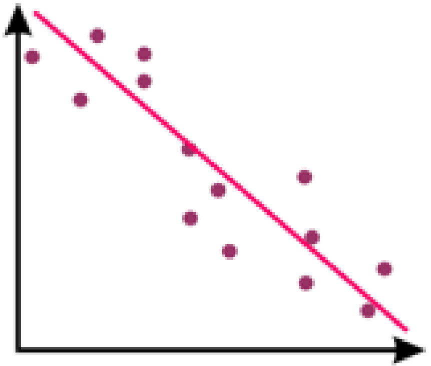
STRONG POSITIVE CORRELATION



WEAK POSITIVE CORRELATION

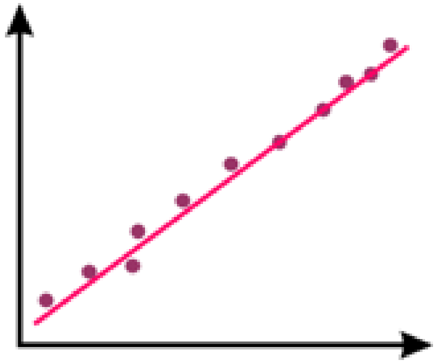


STRONG NEGATIVE CORRELATION

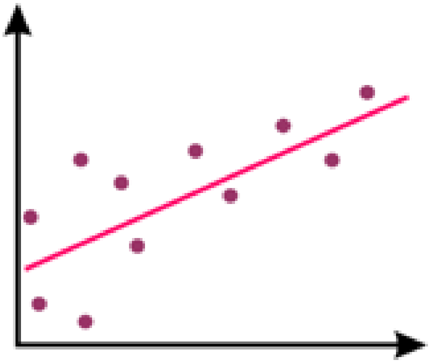


WEAK NEGATIVE CORRELATION

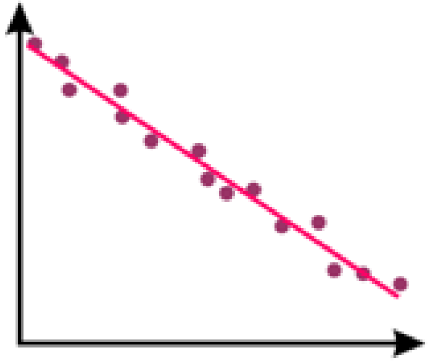
The essentials of regression



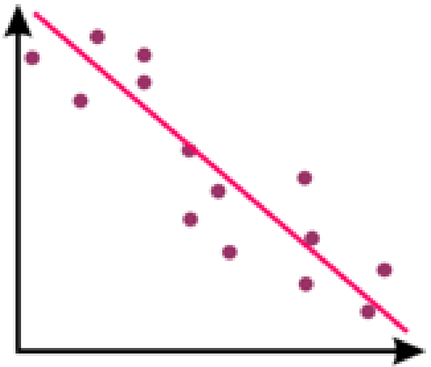
STRONG POSITIVE CORRELATION



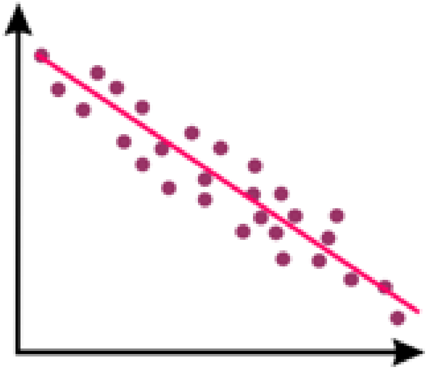
WEAK POSITIVE CORRELATION



STRONG NEGATIVE CORRELATION

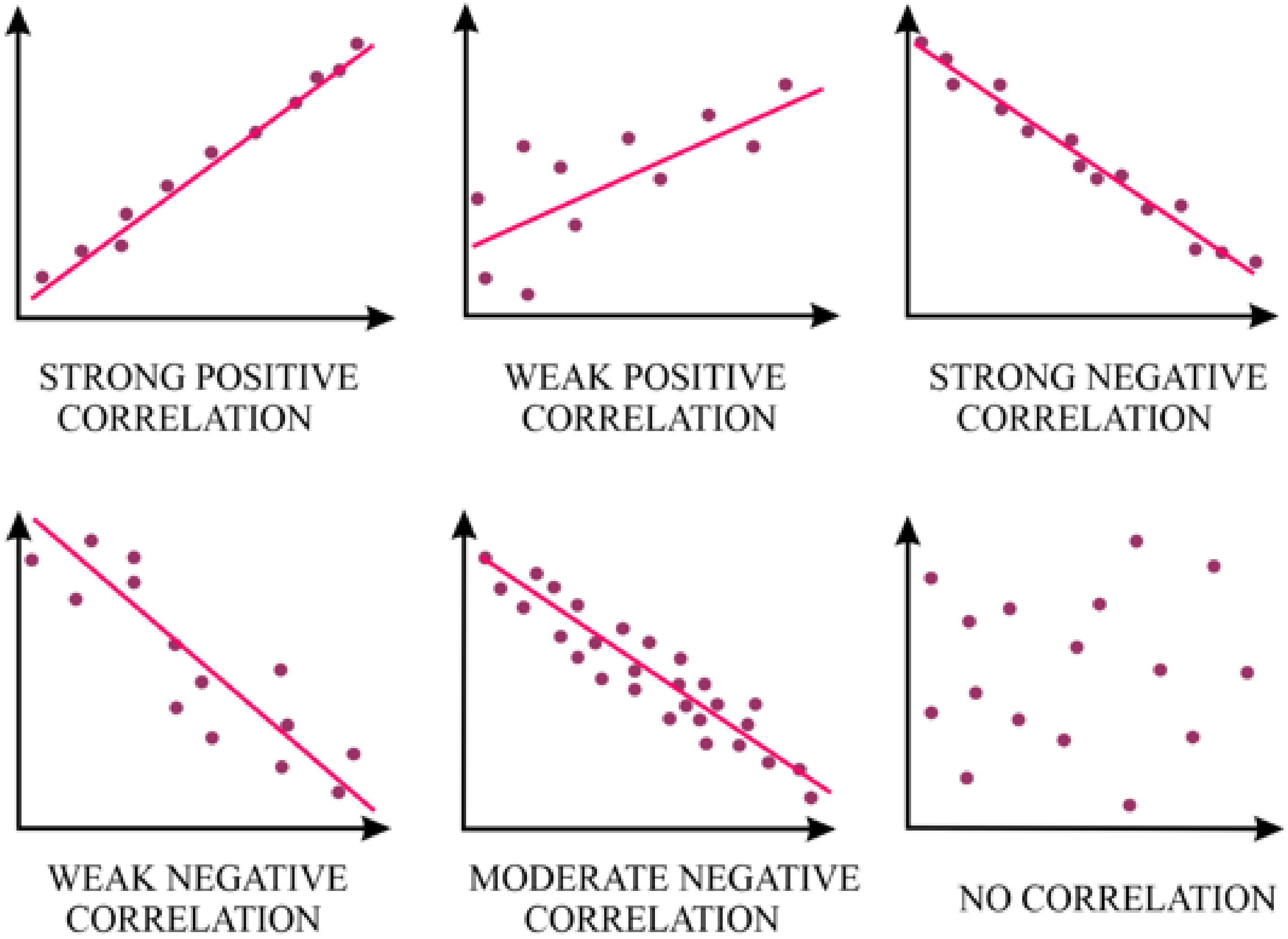


WEAK NEGATIVE CORRELATION



MODERATE NEGATIVE CORRELATION

The essentials of regression



STRONG POSITIVE CORRELATION

WEAK POSITIVE CORRELATION

STRONG NEGATIVE CORRELATION

WEAK NEGATIVE CORRELATION

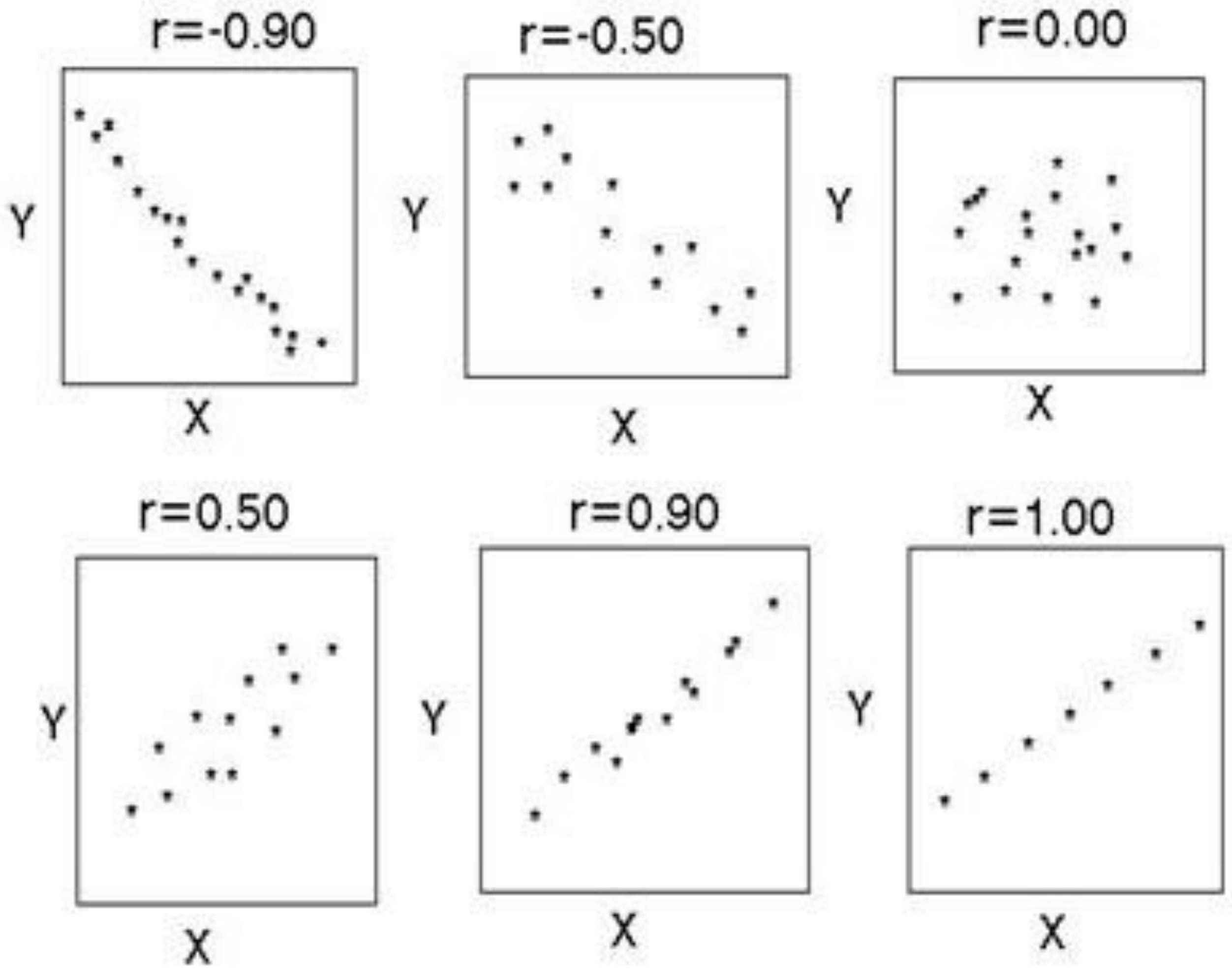
MODERATE NEGATIVE CORRELATION

NO CORRELATION

The essentials of regression

Quantitatively, correlation coefficients range between -1 (perfect negative correlation) to 1 (perfect positive correlation).

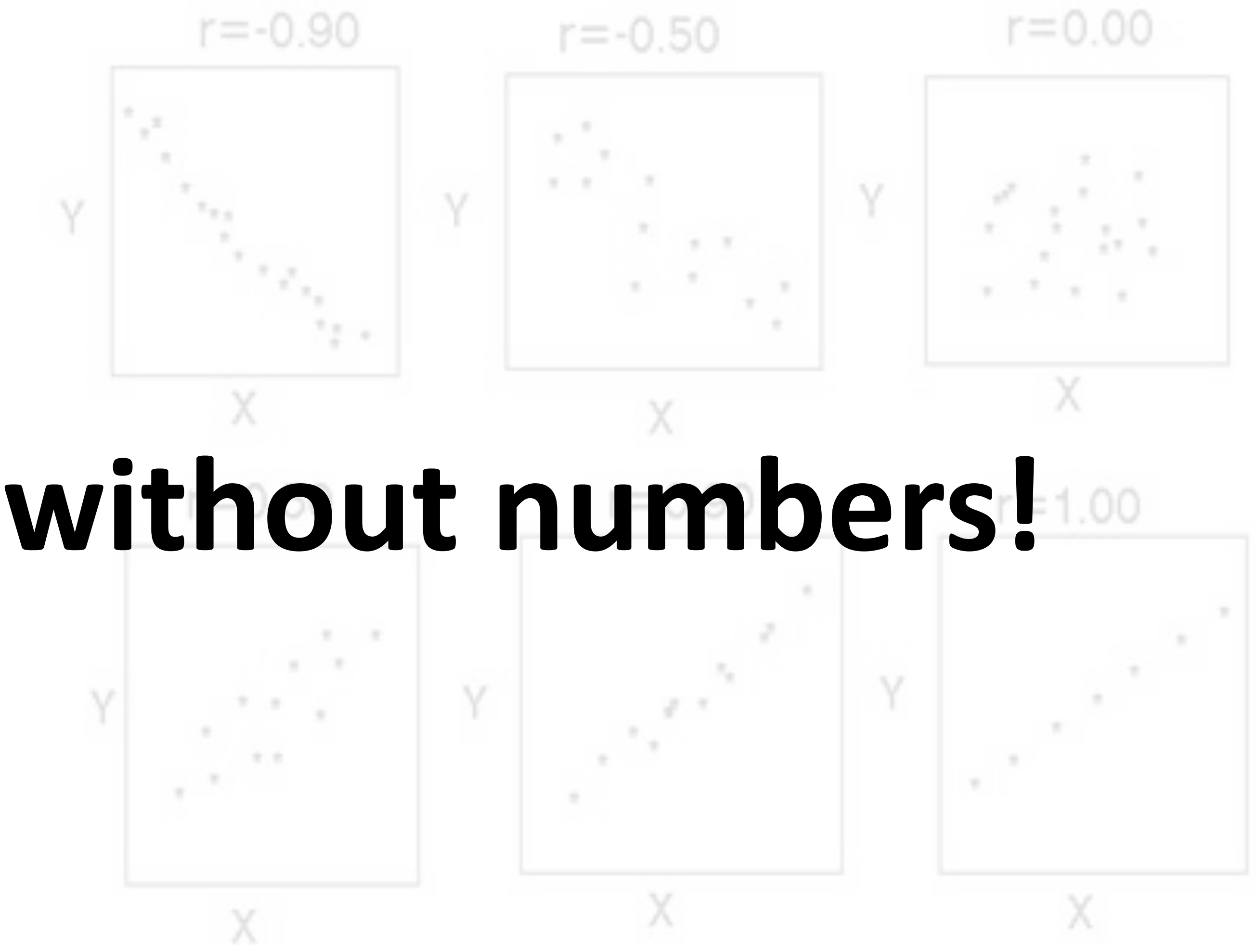
Close to 0 means no correlation.



The essentials of regression

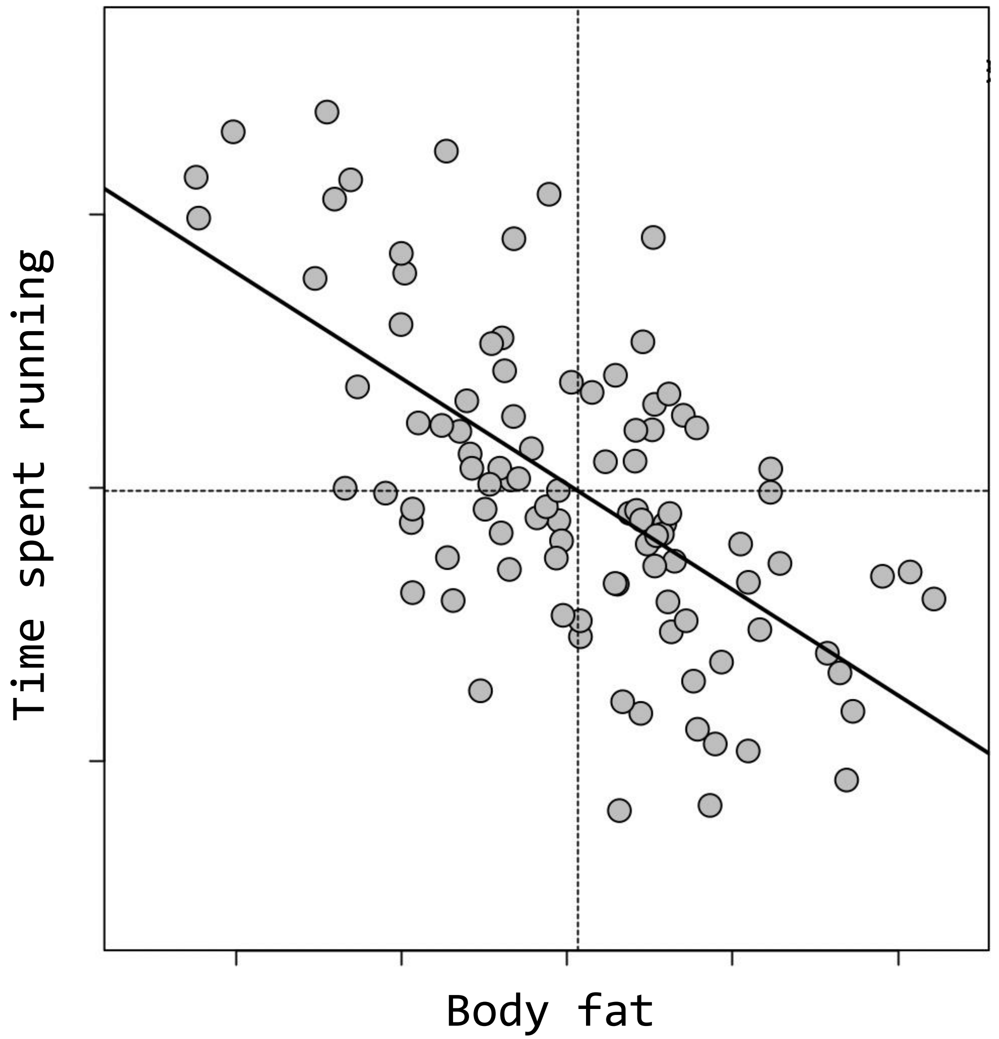
Quantitatively, correlation coefficients range between -1 (perfect negative correlation) to 1 (perfect positive correlation).

Close to 0 means no correlation.



Let’s practice without numbers!

The essentials of regression



Is there a correlation?

If so, positive or negative?

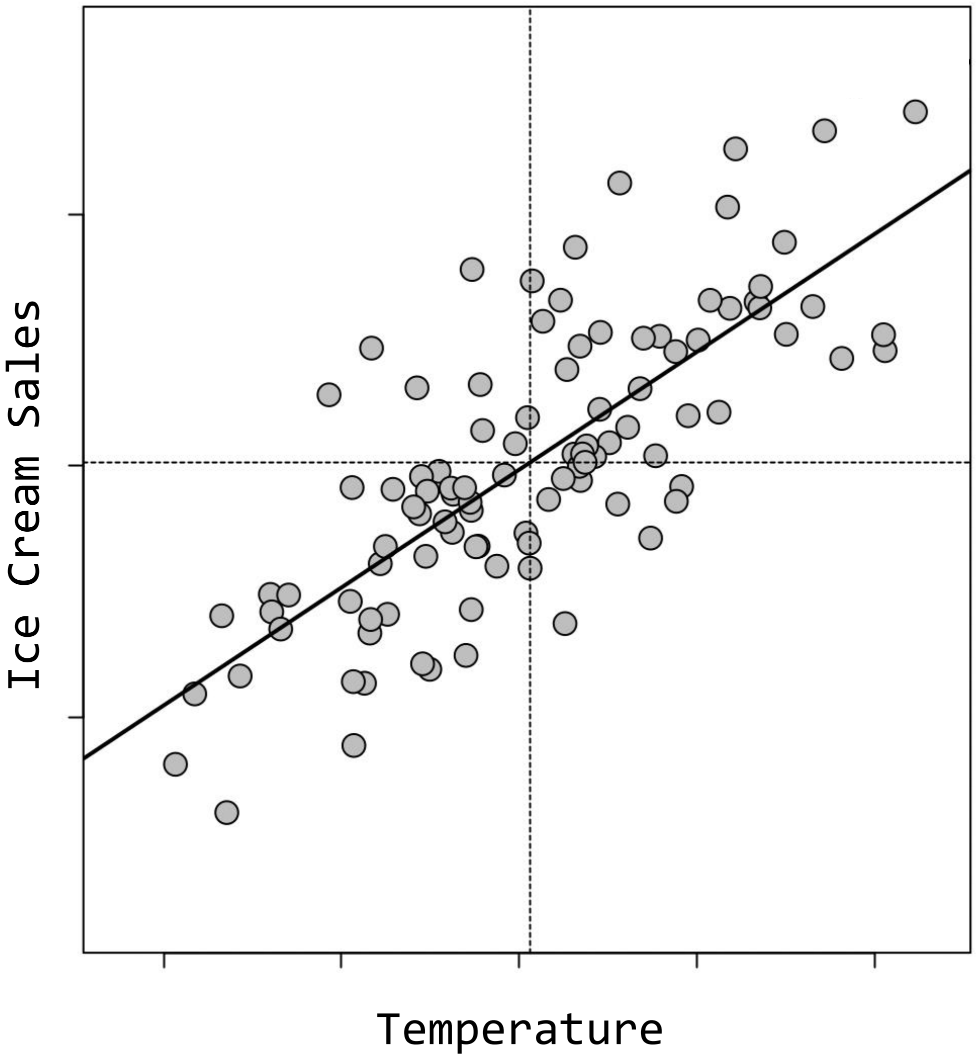
Weak or strong?

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

R = -0.63

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

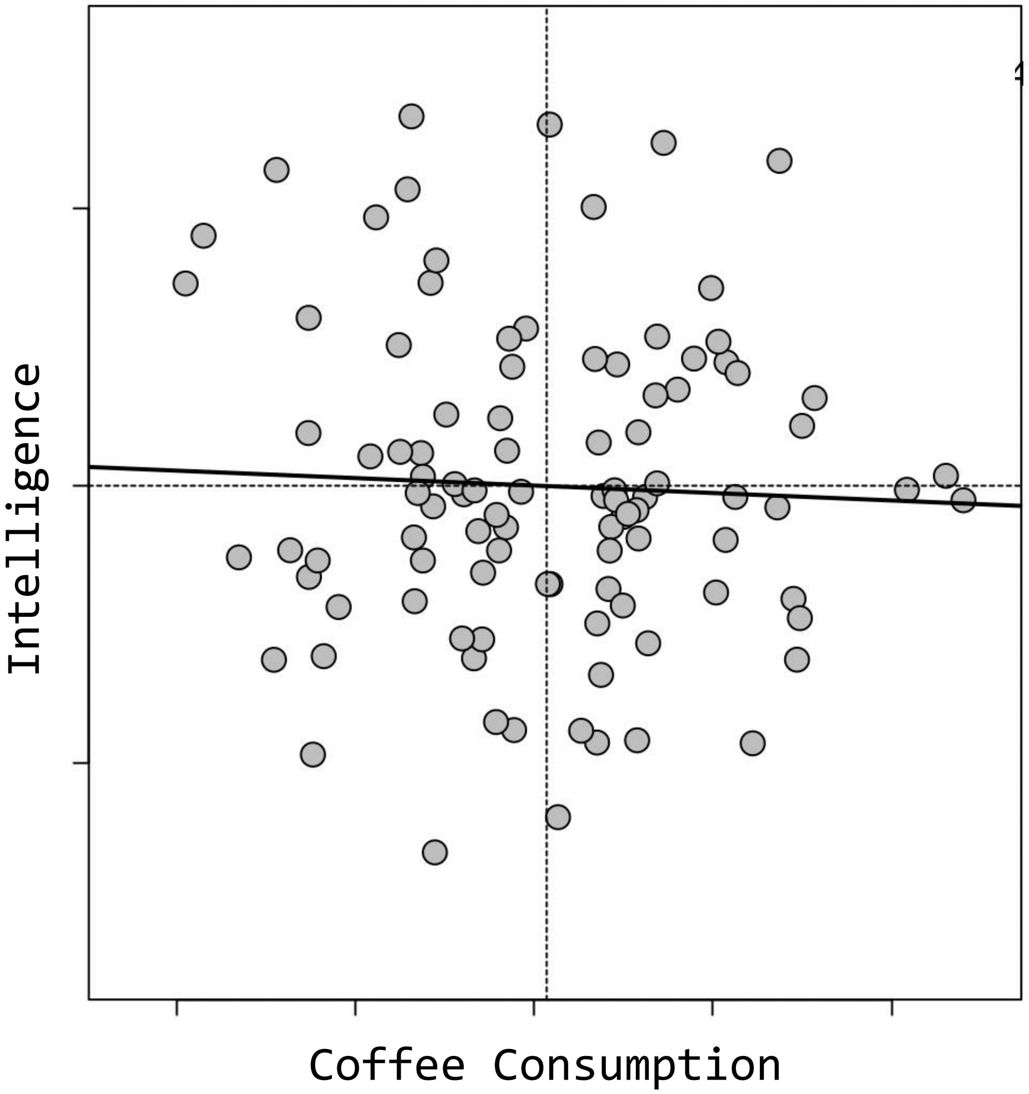
Weak or strong?

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

R = 0.76

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

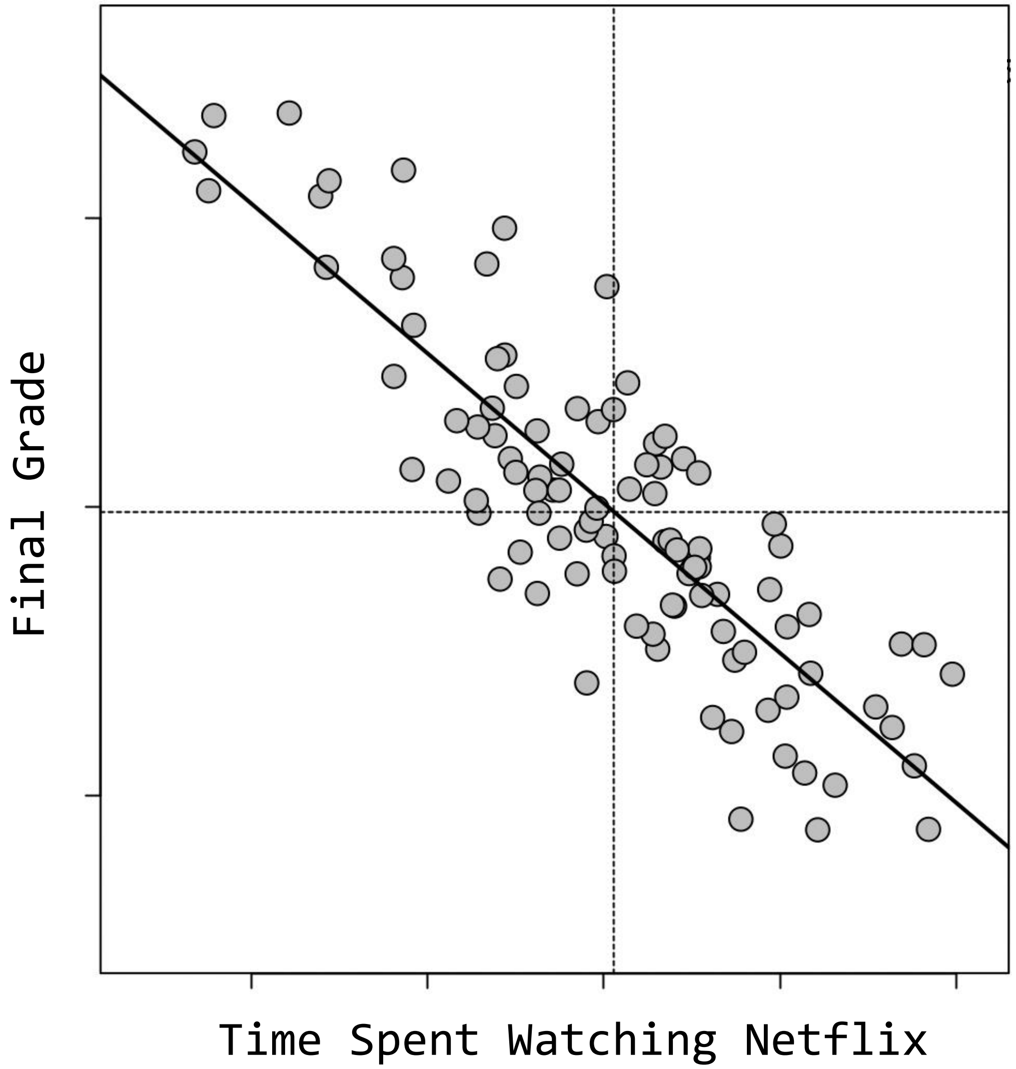
Weak or strong?

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

R = -0.04

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

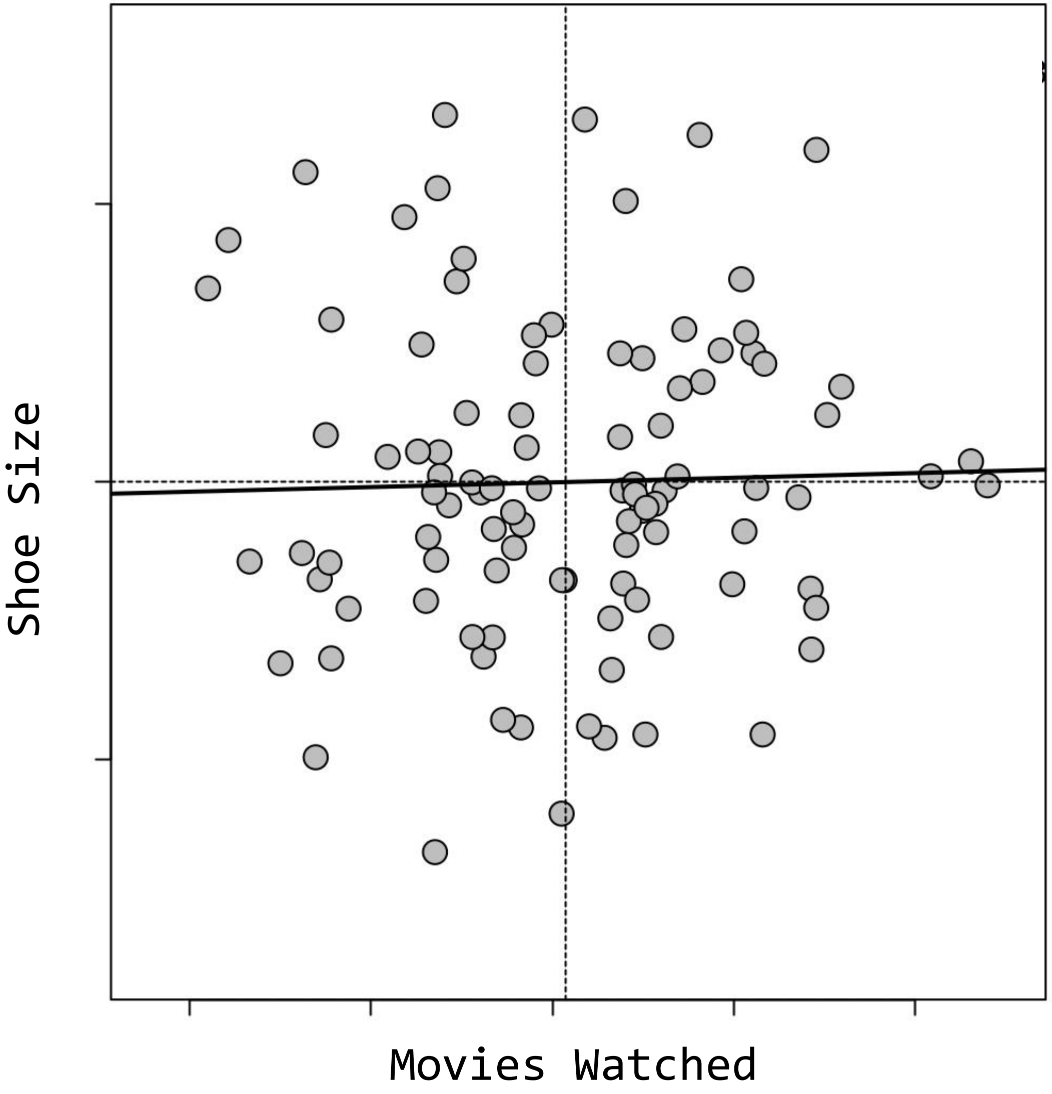
Weak or strong?

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

R = -0.85

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

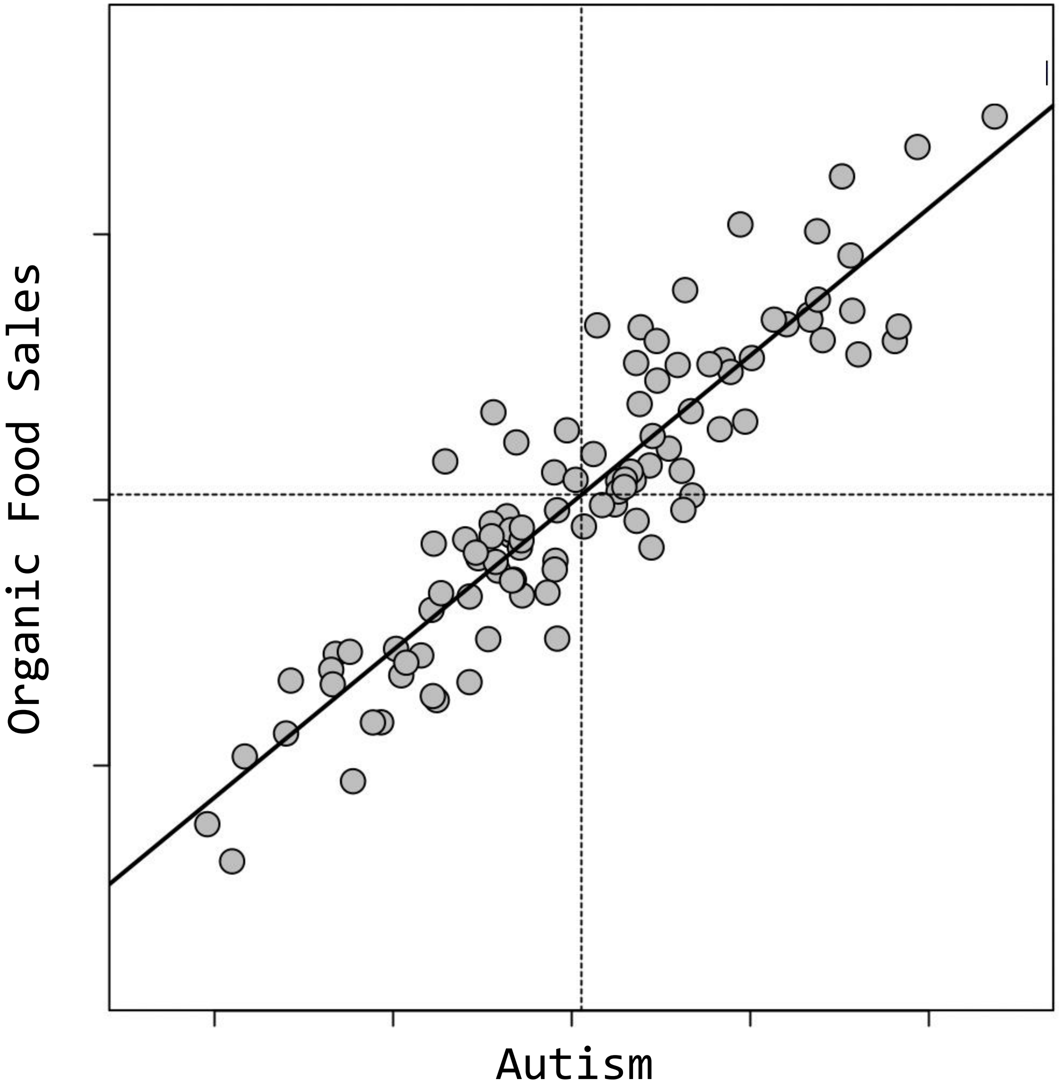
Weak or strong?

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

R = 0.02

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

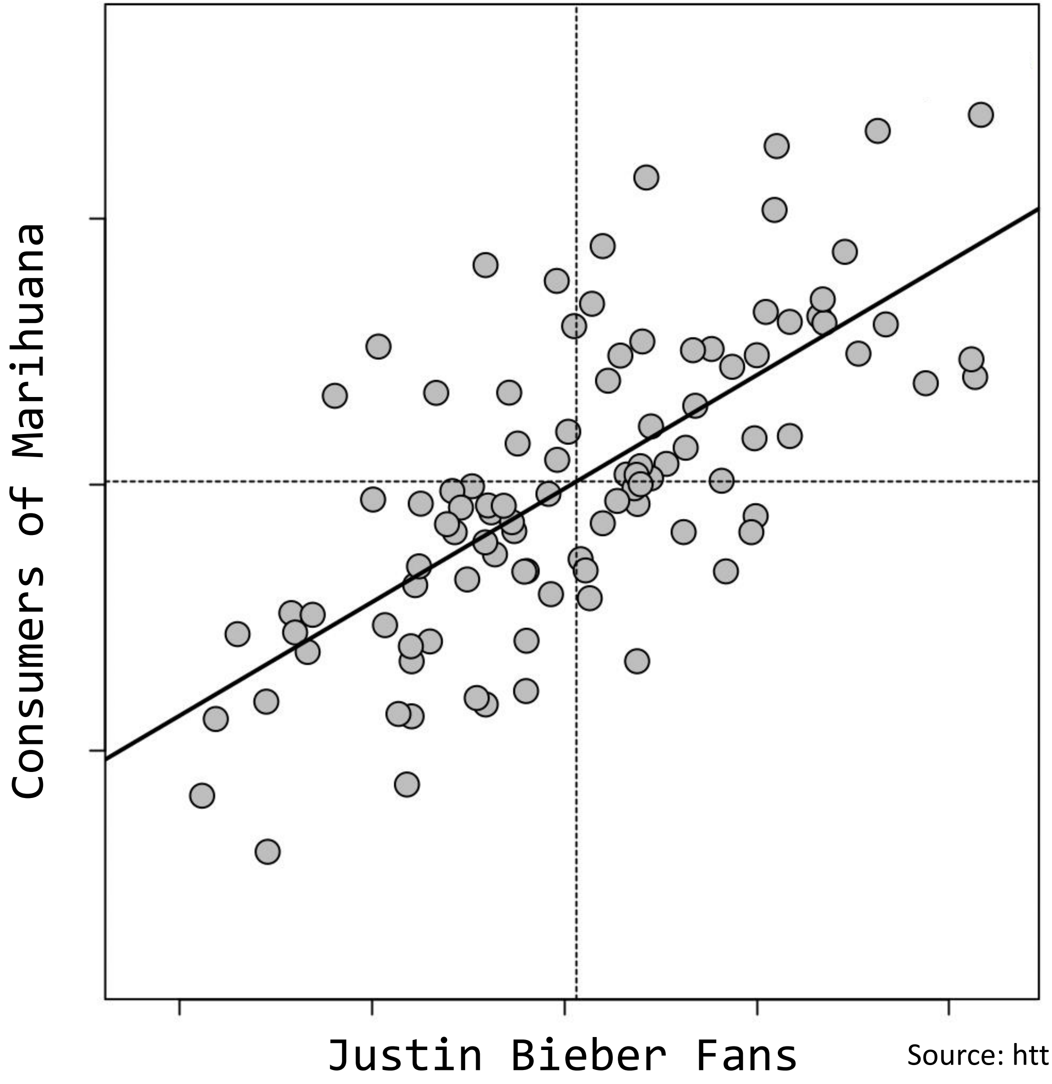
Weak or strong?

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

R = 0.91

How do you explain this result?

The essentials of regression



Is there a correlation?

If so, positive or negative?

Weak or strong?

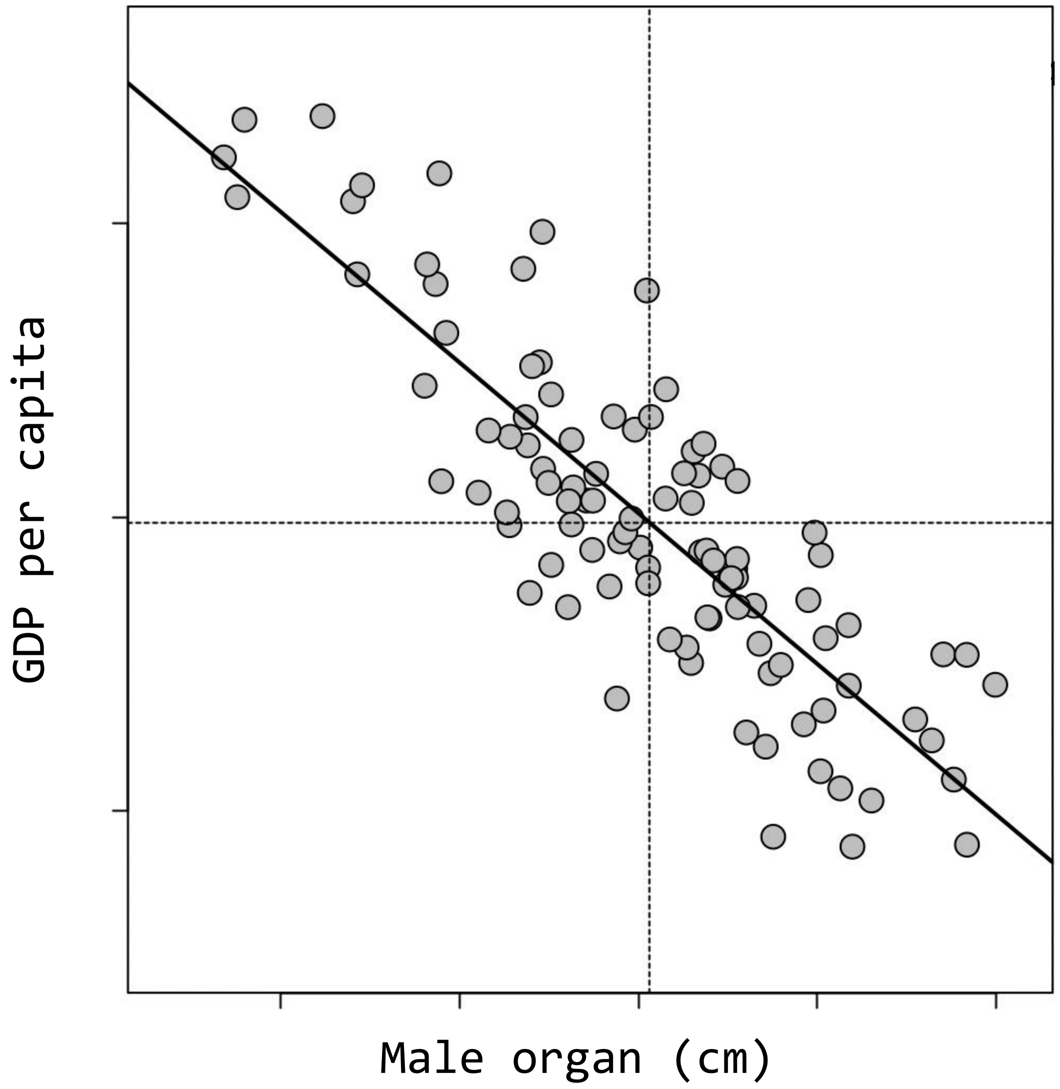
Can you guess the correlation coefficient (between -1 and 1)?

R = 0.71

How do you explain this result?

Source: https://nypost.com/2021/10/04/justin-bieber-breaks-into-cannabis-market/

The essentials of regression



Is there a correlation?

If so, positive or negative?

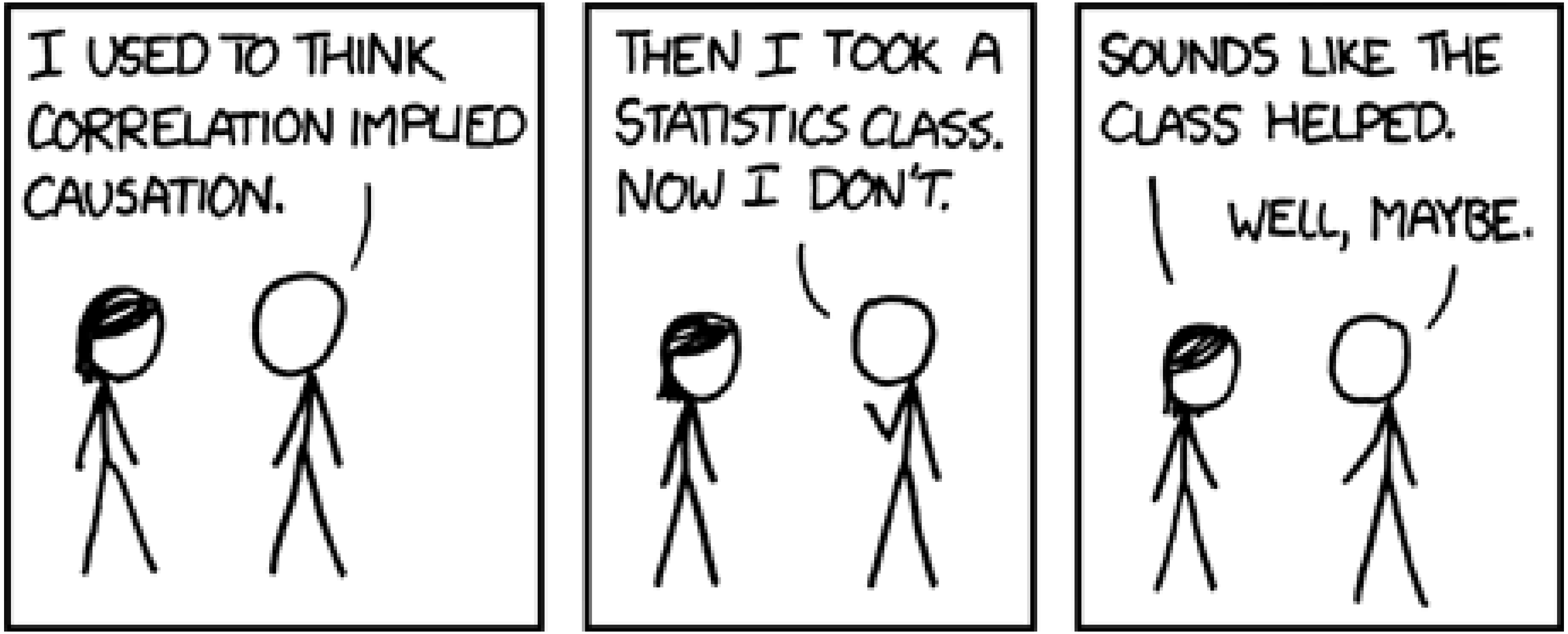
Weak or strong?

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

R = -0.84

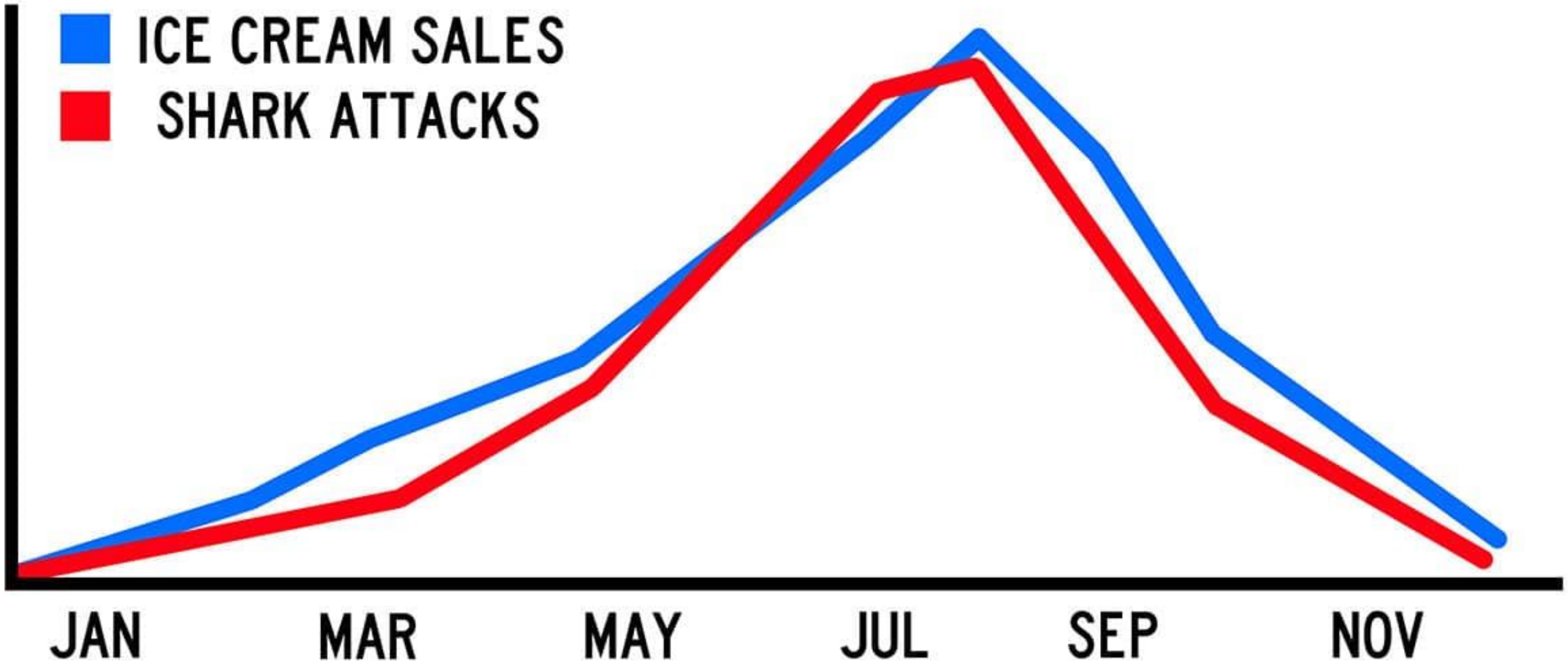
How do you explain this result?

The essentials of regression

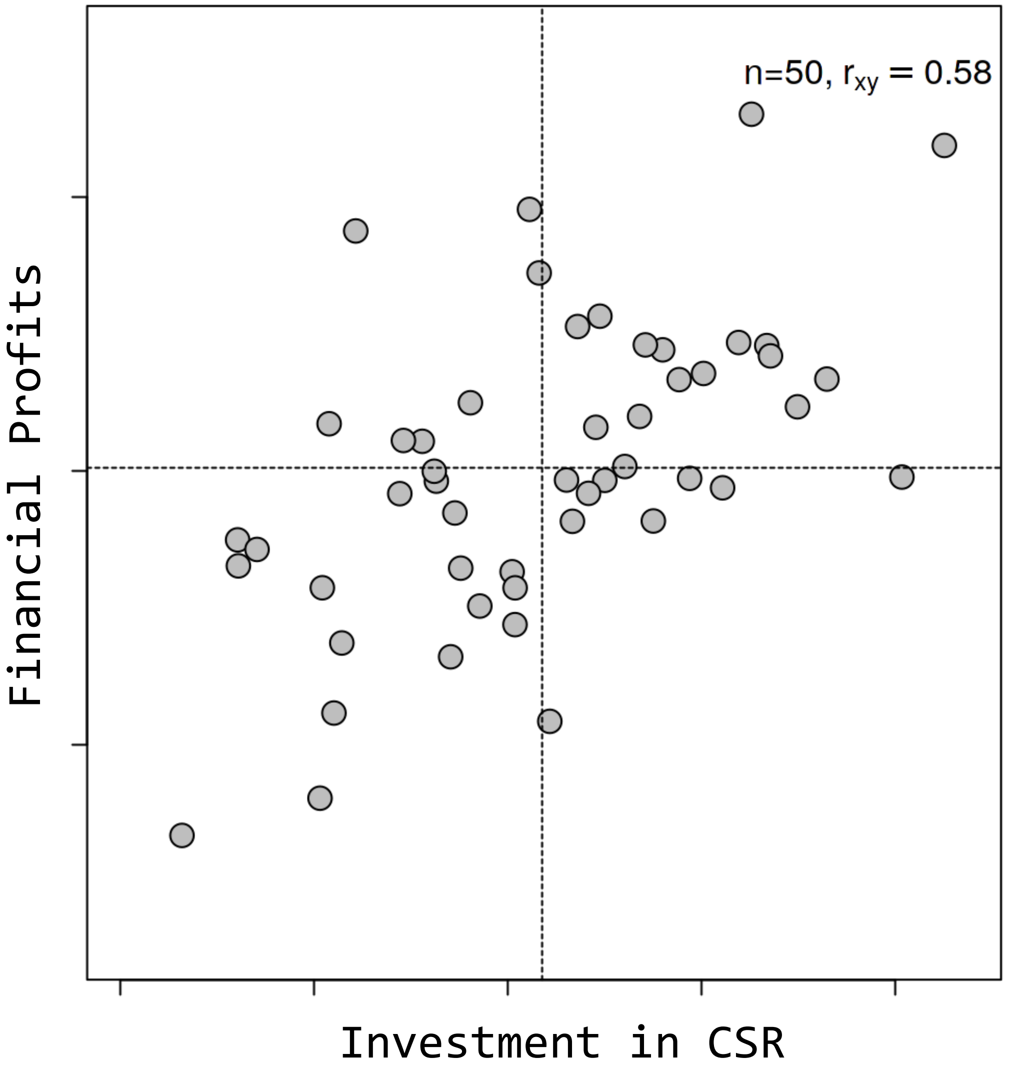

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

The essentials of regression

CORRELATION IS NOT CAUSATION!



The essentials of regression



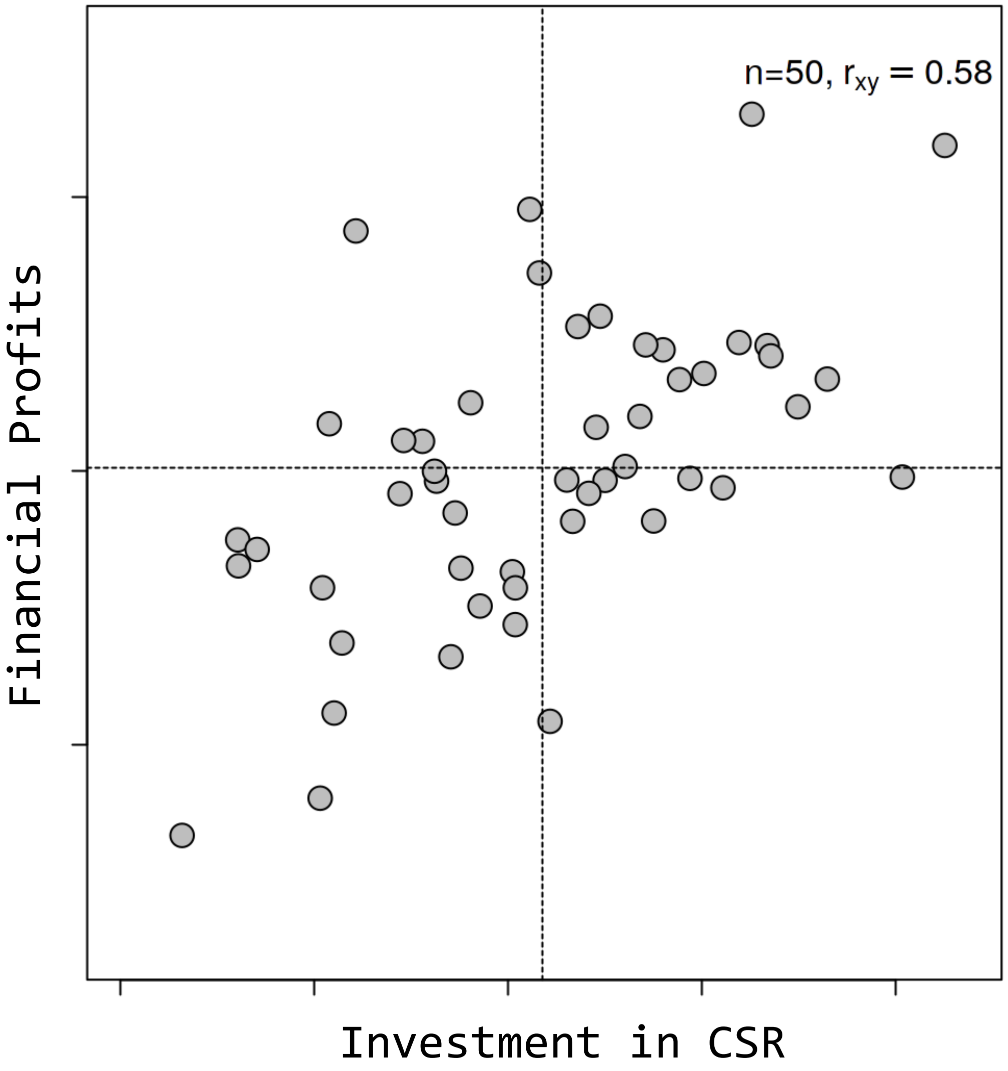
Strong correlation between companies that invest in CSR and financial profits.

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

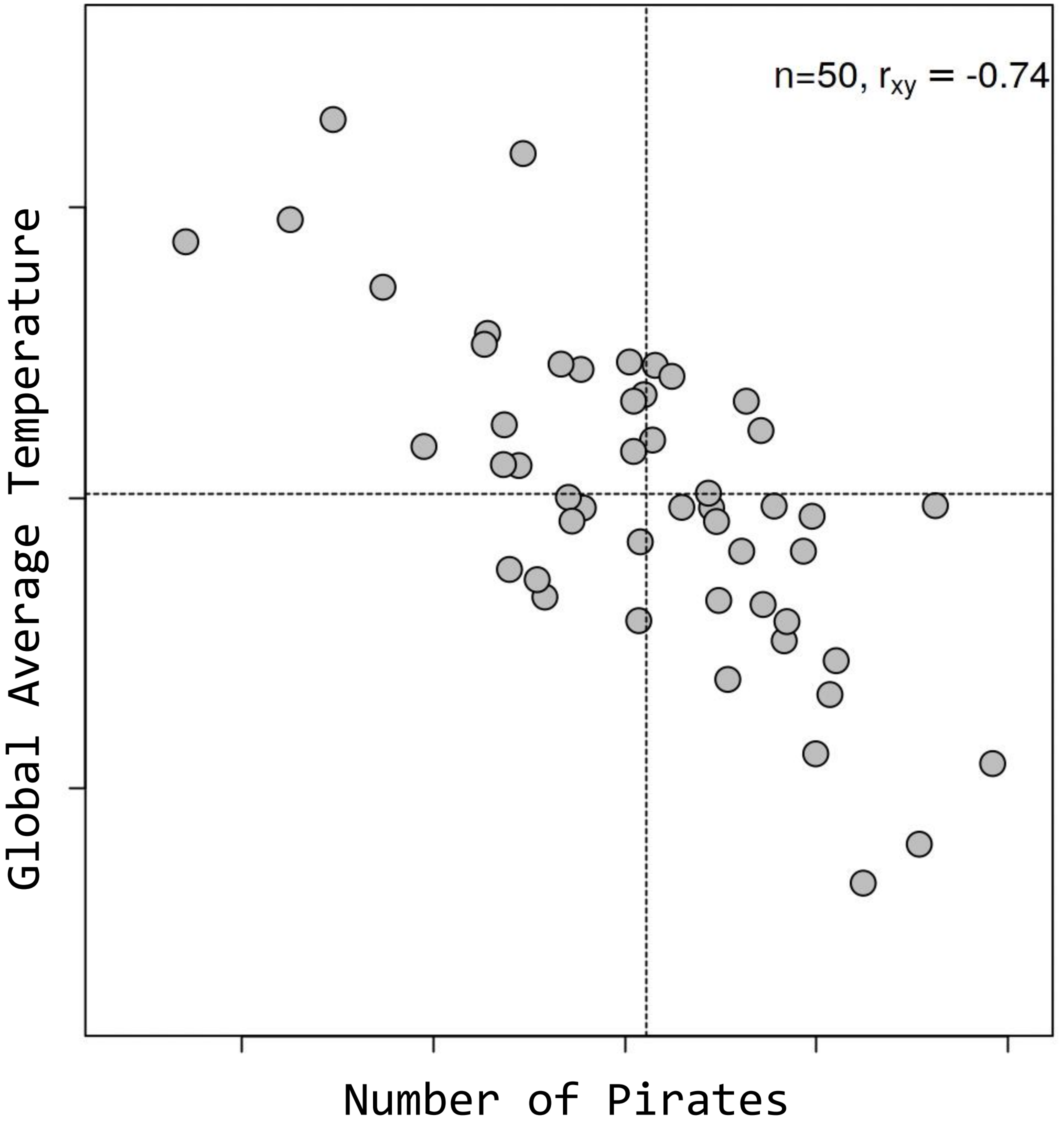


This problem is called

REVERSE CAUSALITY

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

The essentials of regression

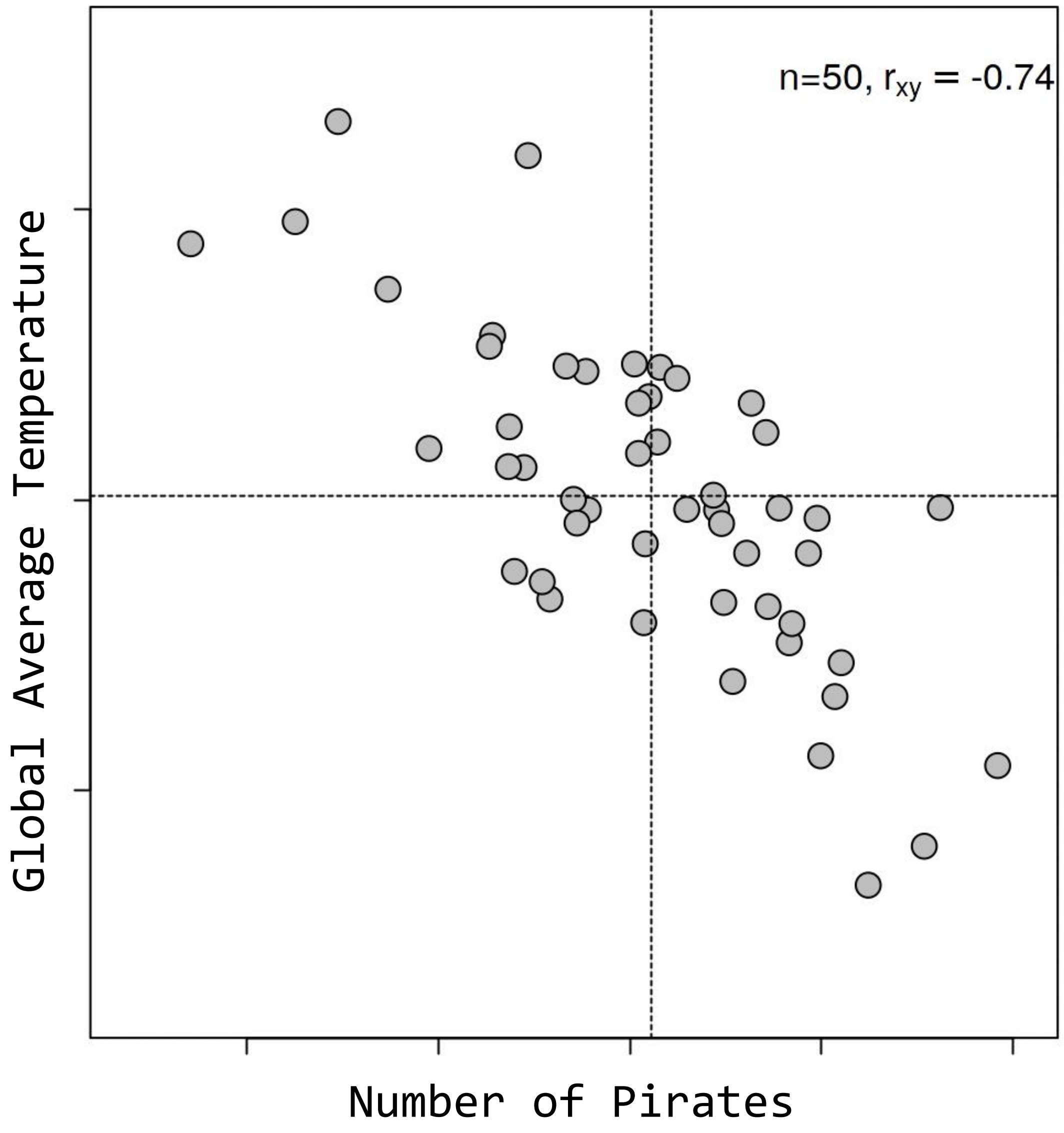


Are pirates preventing climate change?

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

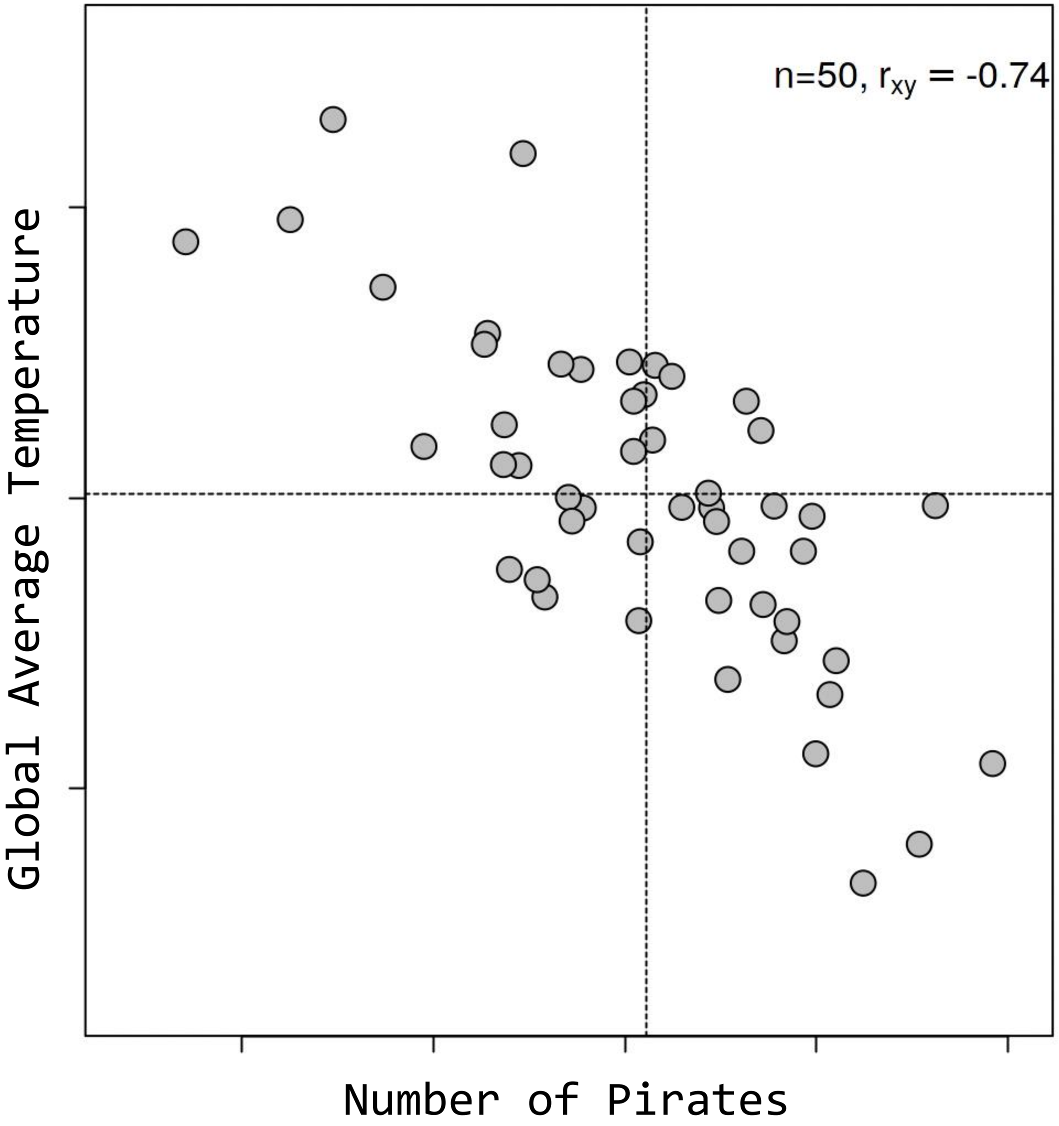


This problem is called

OMITTED VARIABLE

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

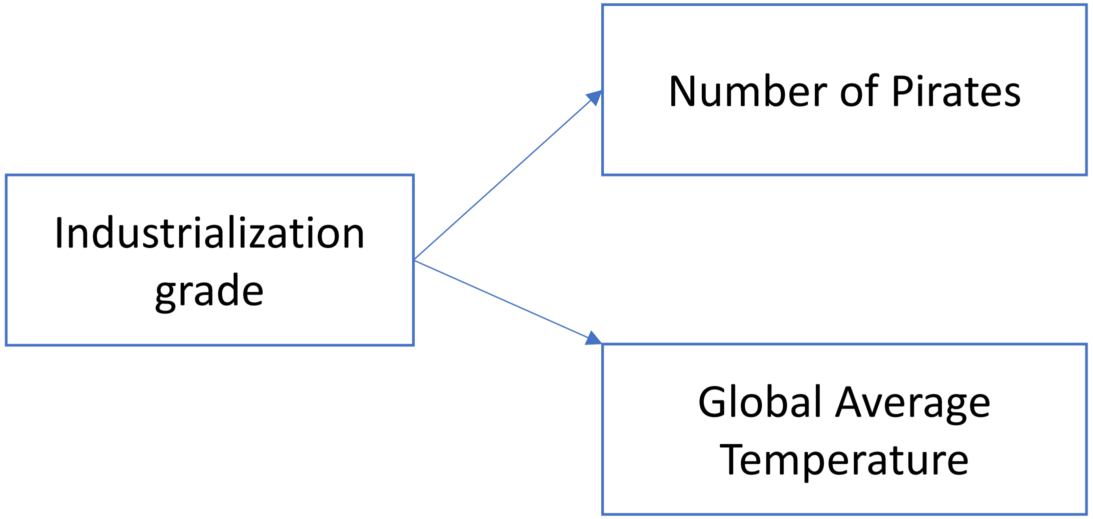
The essentials of regression



This problem is called

OMITTED VARIABLE

Two things are dependent on another (omitted) variable.


Industrialization grade

Number of Pirates

Global Average Temperature

The essentials of regression

We can solve these issues with

MULTIPLE LINEAR REGRESSION

The essentials of regression

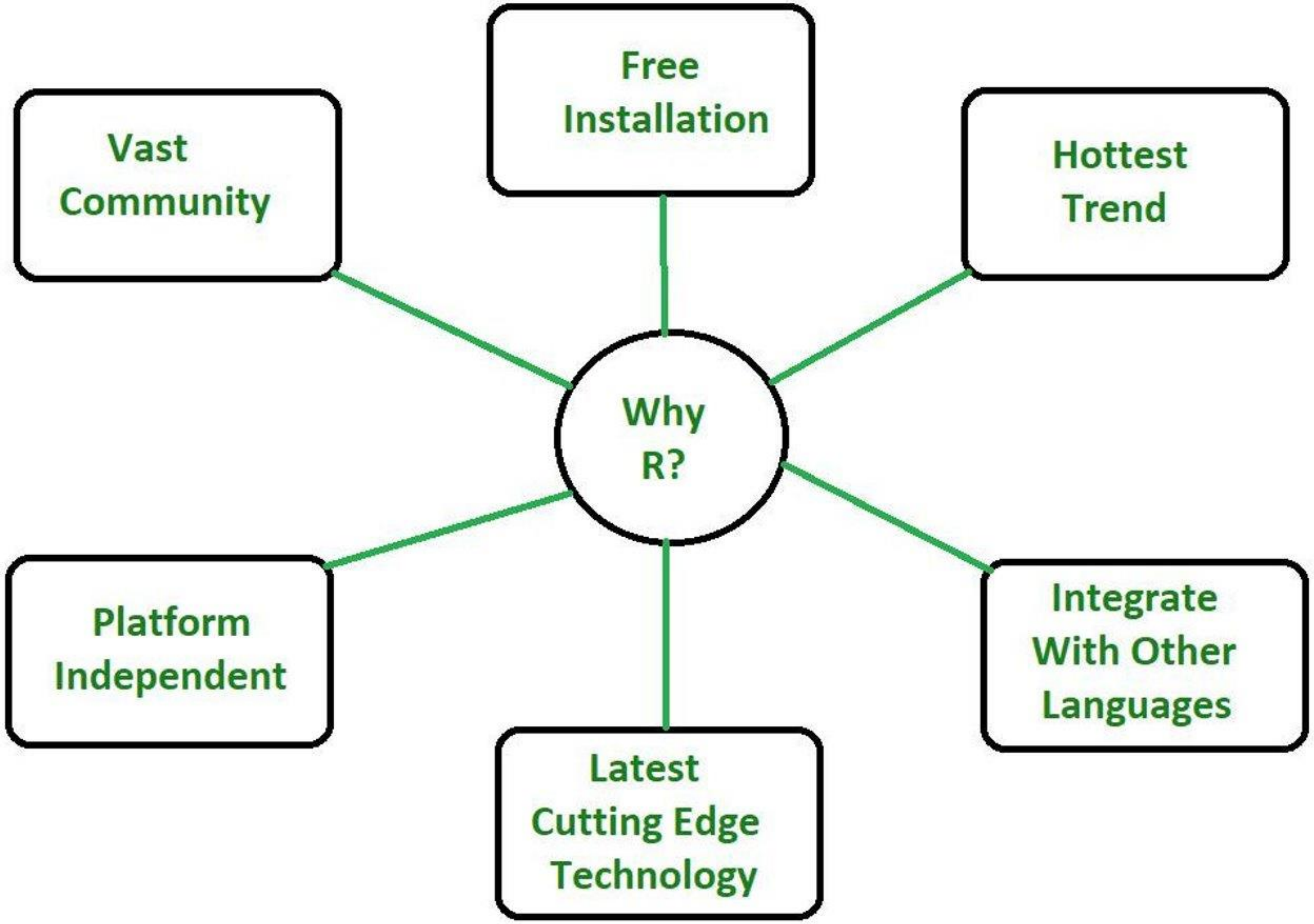
In general, **multiple linear regression is the same as a linear regression**, but **using more than one variable** to explain variation in our dependent variable.

TIME TO PRACTICE!

Agenda

1. Basic Commands of R
2. Reproduce Weight/Height exercise in class
3. Predicting sales from price

Why R?



Why R?

Free Installation

Vast Community

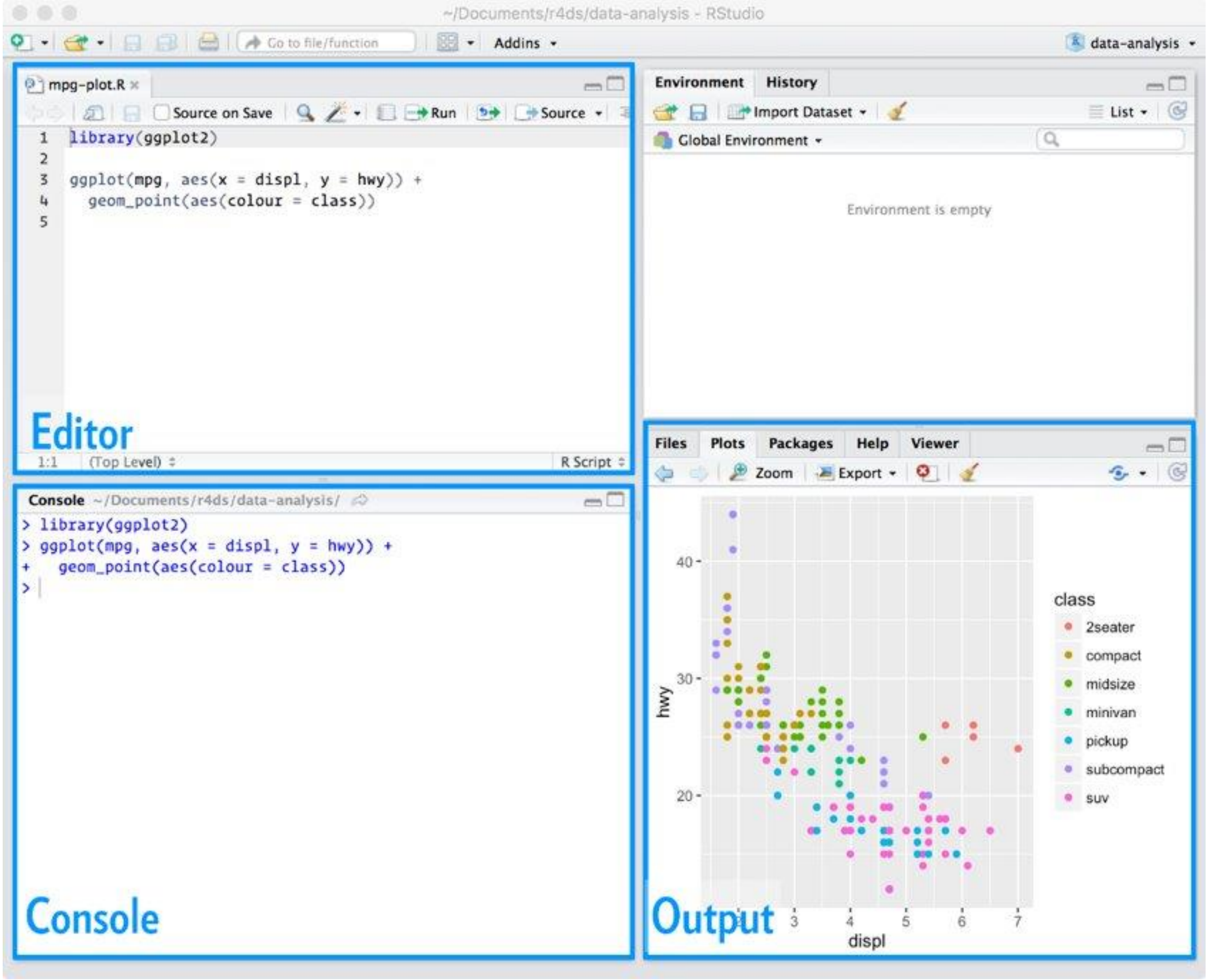
Hottest Trend

Platform Independent

Latest Cutting Edge Technology

Integrate With Other Languages

How does it work?



```
library(ggplot2)

ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(colour = class))
```

```
> library(ggplot2)
> ggplot(mpg, aes(x = displ, y = hwy)) +
+   geom_point(aes(colour = class))
> 
```

How does it work?

- As you start to run R code, you’re likely to run into problems.

- Don’t worry — it happens to everyone. I have been writing R code for years, and every day I still write code that doesn’t work!

- Start by carefully comparing the code that you’re running to the code in the session.

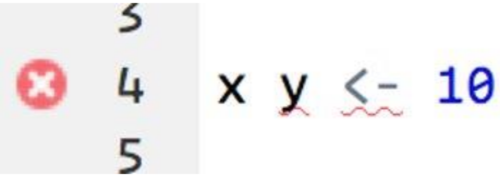
- R is extremely picky, and a misplaced character can make all the difference. For example, if you write “Dataset” instead of “dataset” it will give you an error.

- Sometimes you’ll run the code and nothing happens. Check the left-hand of your console: if there is a + sign, it means that R doesn’t think you’ve typed a complete expression and it’s waiting for you to finish it.

- In this case, it’s usually easy to start from scratch again by pressing ESCAPE to abort processing the current command.

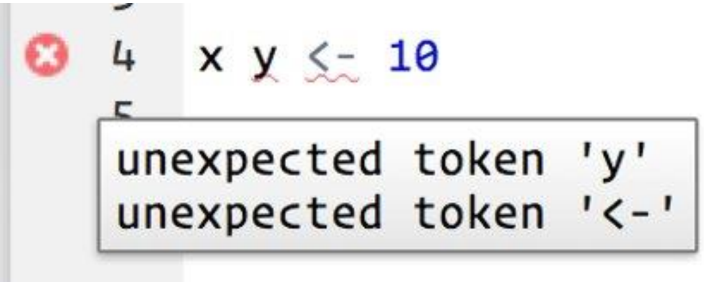
How does it work?

- The script editor will also highlight syntax errors with a red squiggly line and a cross in the sidebar:



```
x y <- 10
```

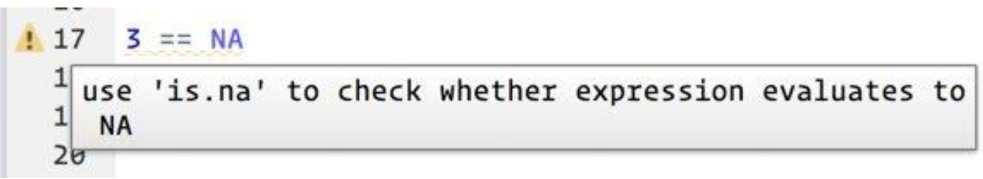
- Fly over the cross to see what the problem is:



```
x y <- 10
```

```
unexpected token 'y'
unexpected token '<-'
```

- RStudio will also let you know about potential problems:



```
3 == NA
```

```
use 'is.na' to check whether expression evaluates to
 NA
```

How does it work?

R works with libraries

To install a given library just type

install.packages(“name_of_the_library”)

It is not enough to install them, you also need to call the library when you want to use it (only once per session). To do so, just type:

library(name_of_the_library)

Looking for help

- As you start to code, you will soon find questions, I will try to answer as little as possible because I want you to learn how to find solutions. Mainly:

- If you get stuck, start with Google. Typically adding “R” to a query is enough to restrict it to relevant results

- Google is particularly useful for error messages.
    - If you get an error message and you have no idea what it means, try googling it!
    - Chances are that someone else has been stuck in the past, and there will be help somewhere on the web.

- If Google doesn’t help, try stackoverflow.
    - Start by spending a little time searching for an existing answer, including [R] to restrict your search to questions and answers that use R.

ALWAYS ANNOTATE YOUR CODE

In R, this is done using #

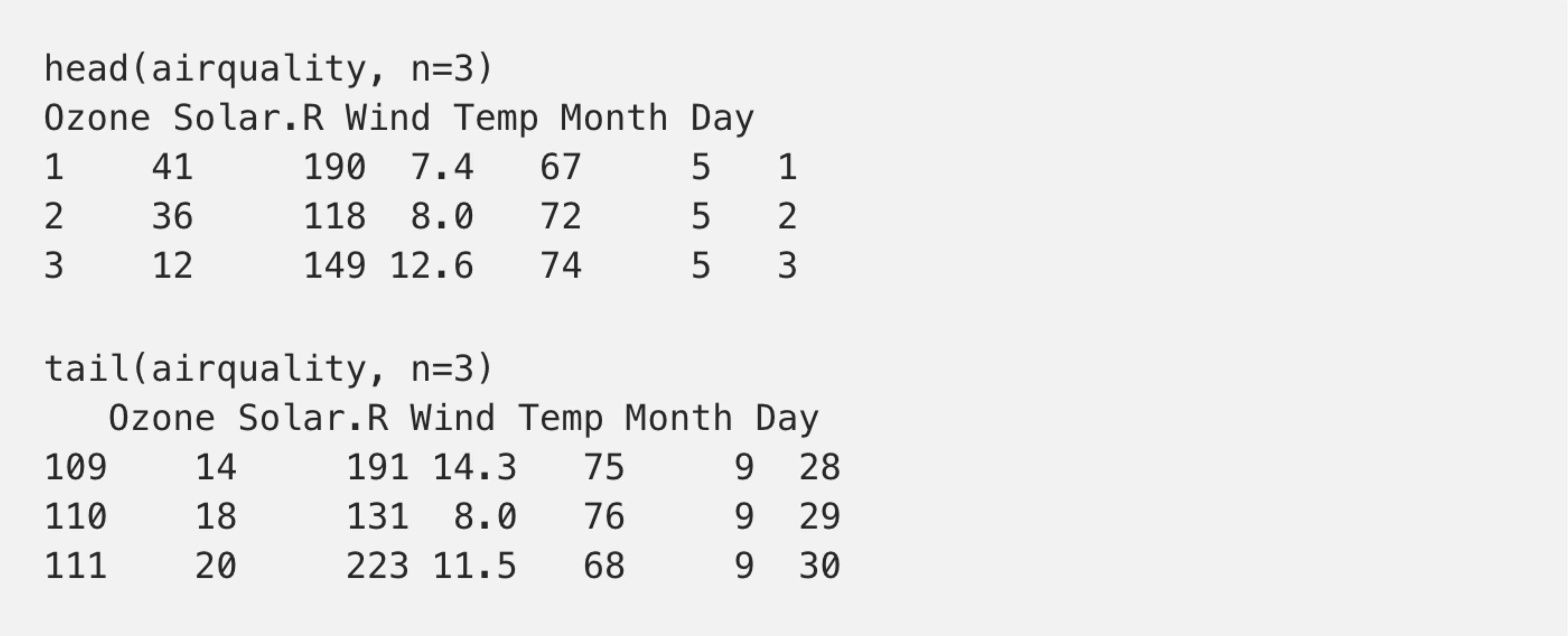
```
# This is just an example
```

Type and run:

```
airquality <- airquality
```

- **head(data,n)** and **tail(data,n)**

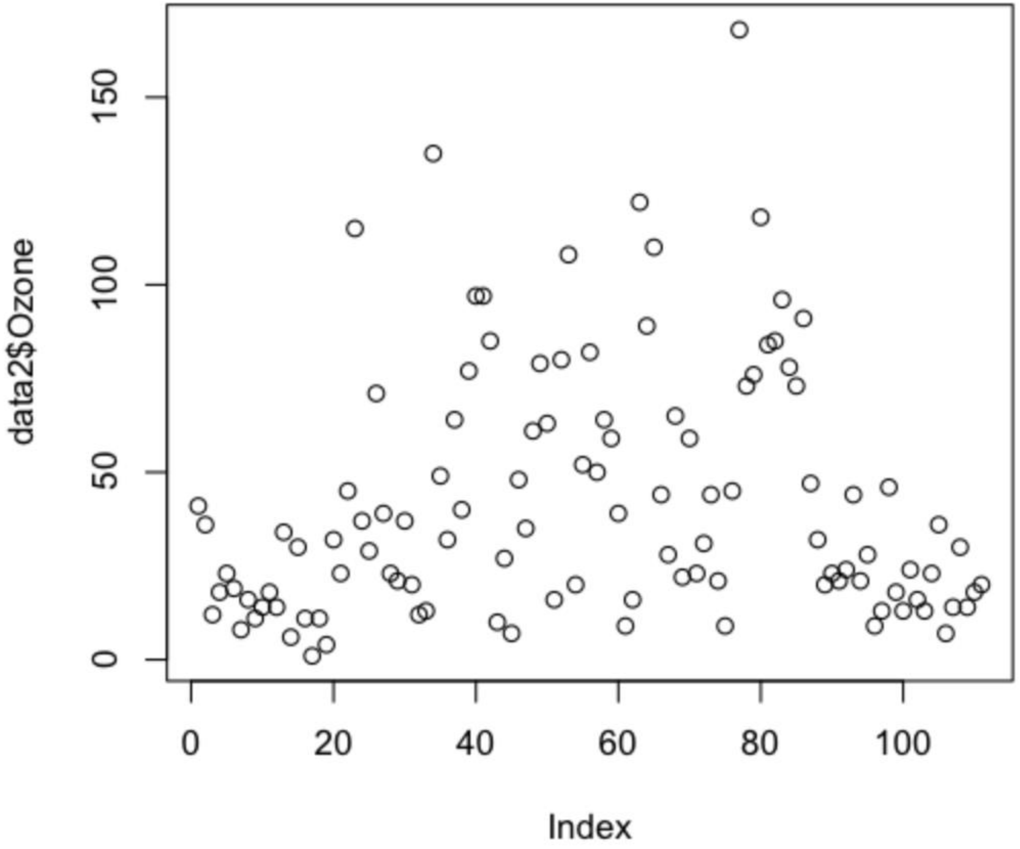
  The head outputs the top **n** elements in the dataset while the tail method outputs the bottom **n**.



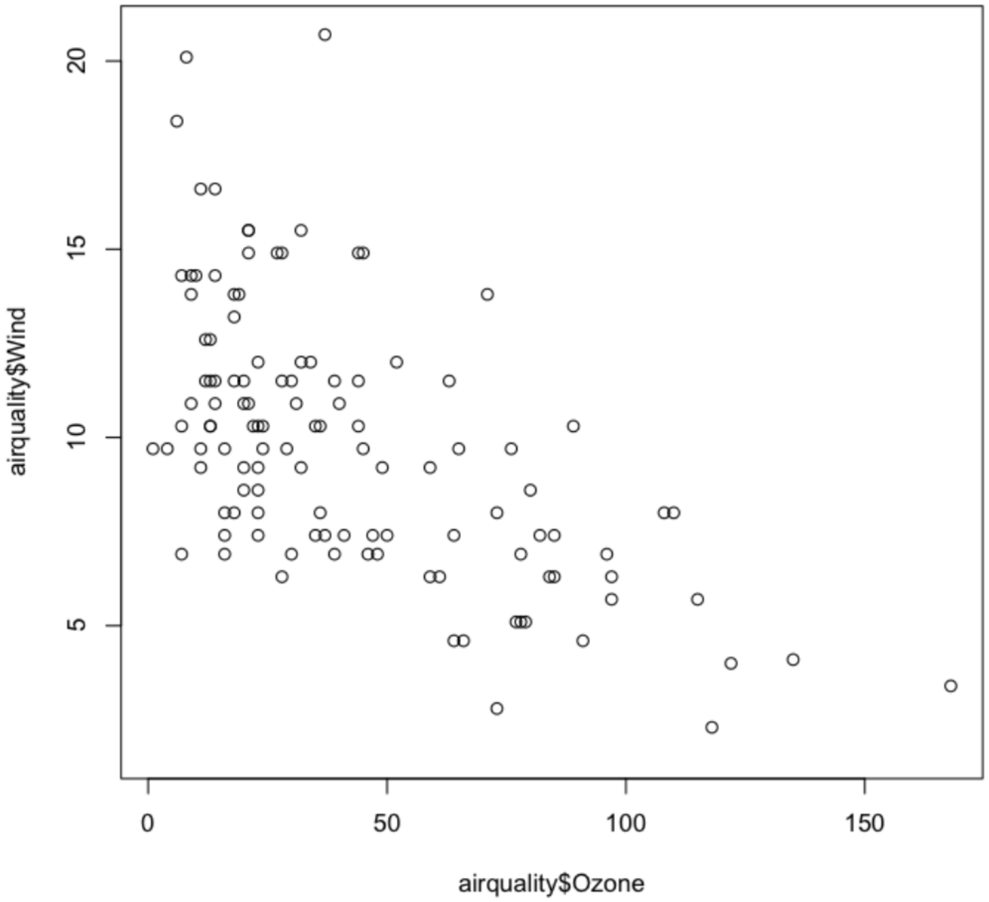
```
head(airquality, n=3)
Ozone Solar.R Wind Temp Month Day
1     41      190  7.4    67      5    1
2     36      118  8.0    72      5    2
3     12      149 12.6    74      5    3

tail(airquality, n=3)
   Ozone Solar.R Wind Temp Month Day
109     14      191 14.3    75      9   28
110     18      131  8.0    76      9   29
111     20      223 11.5    68      9   30
```

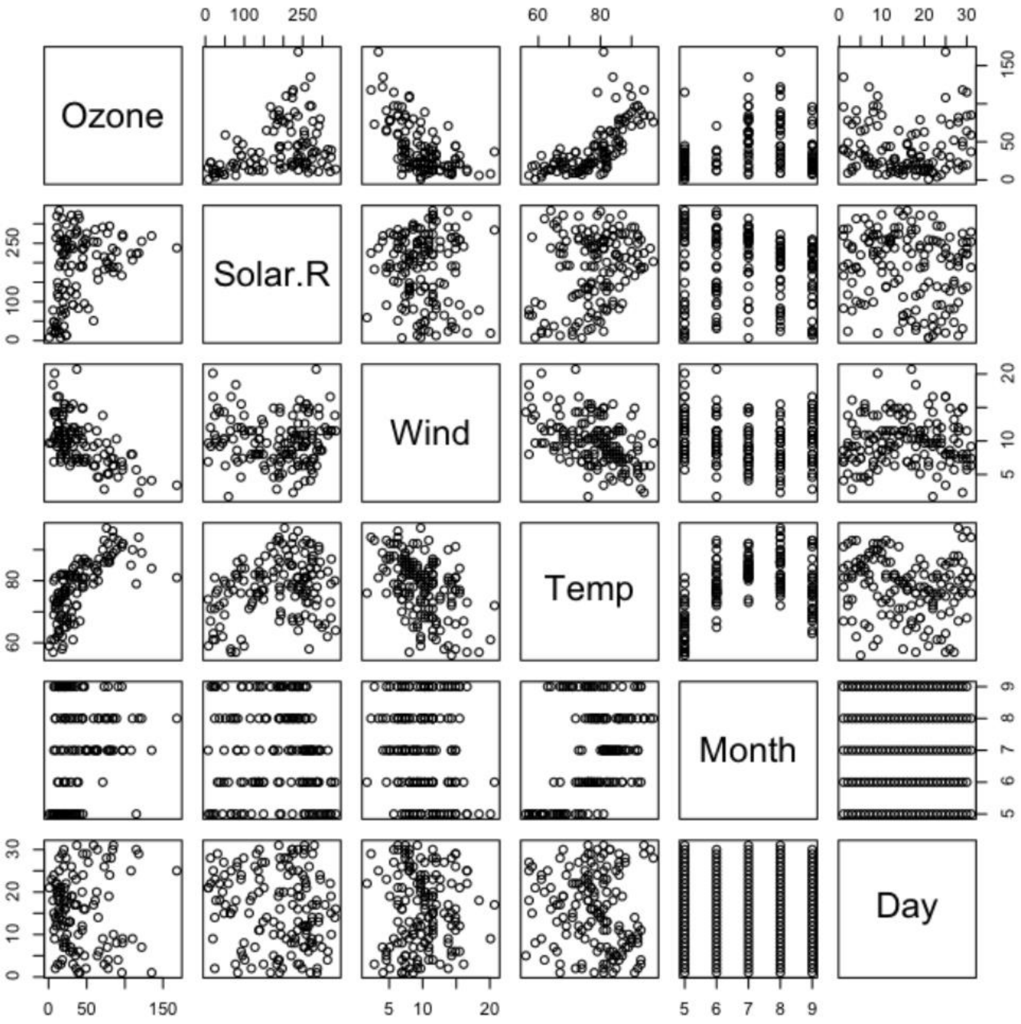
```
plot(airquality$Ozone)
```



```
plot(airquality$Ozone, airquality$Wind)
```



```
plot(airquality)
```



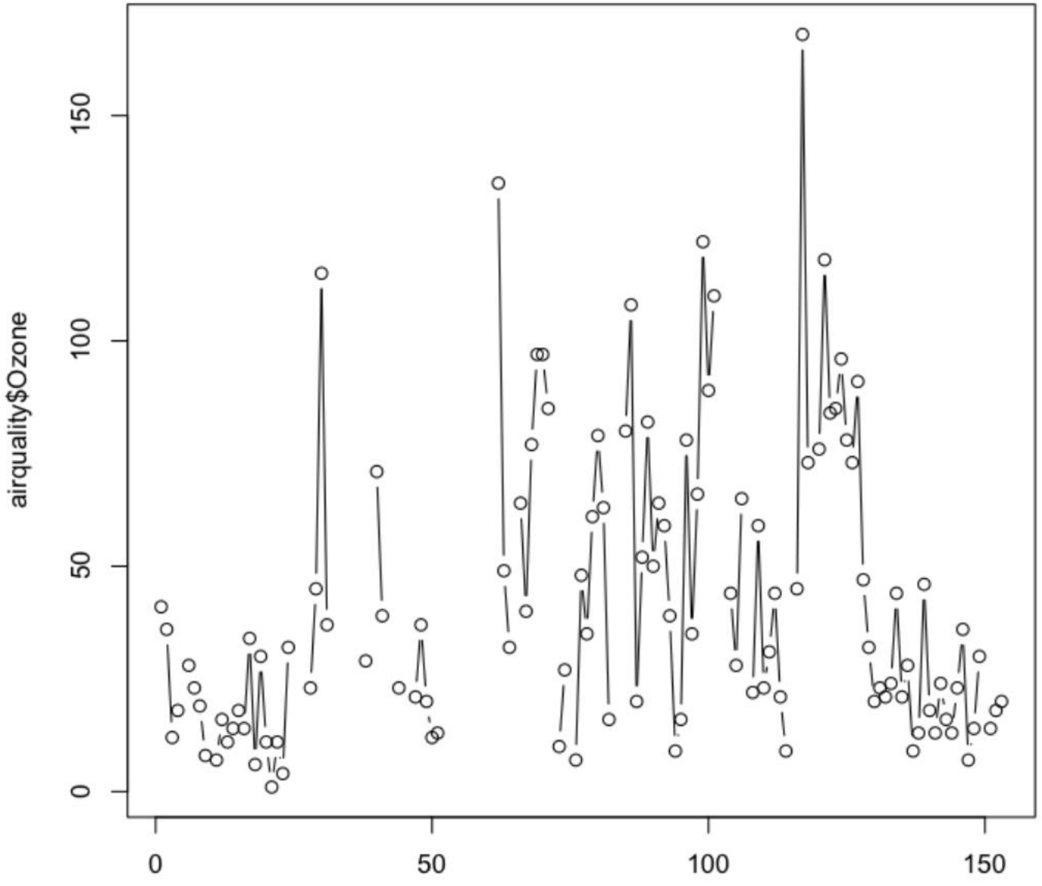
We get a matrix of scatterplots which is a correlation matrix of all the columns. The plot above instantly shows that:

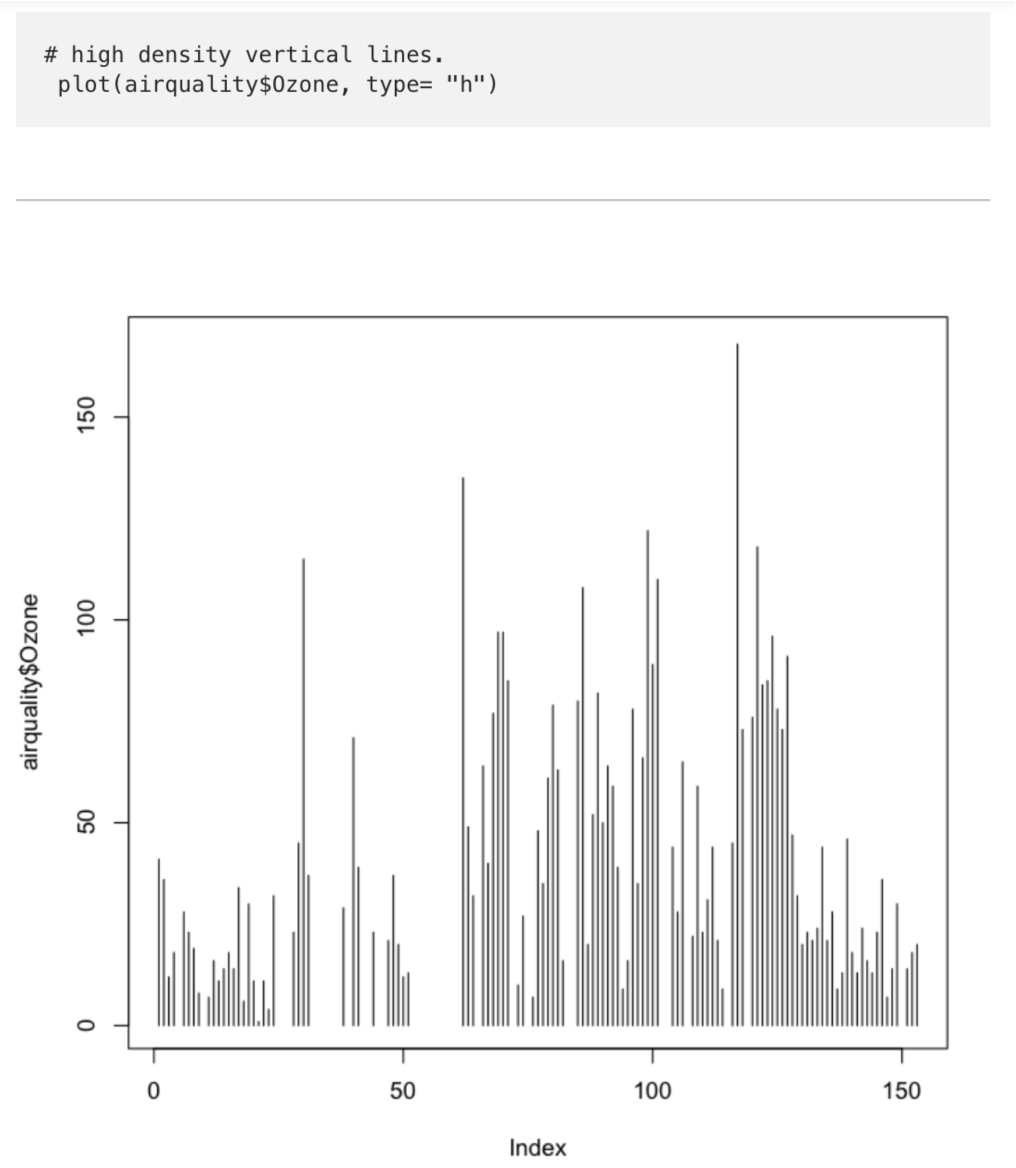
- The level of Ozone and Temperature is correlated positively.

- Wind speed is negatively correlated to both Temperature and Ozone level.

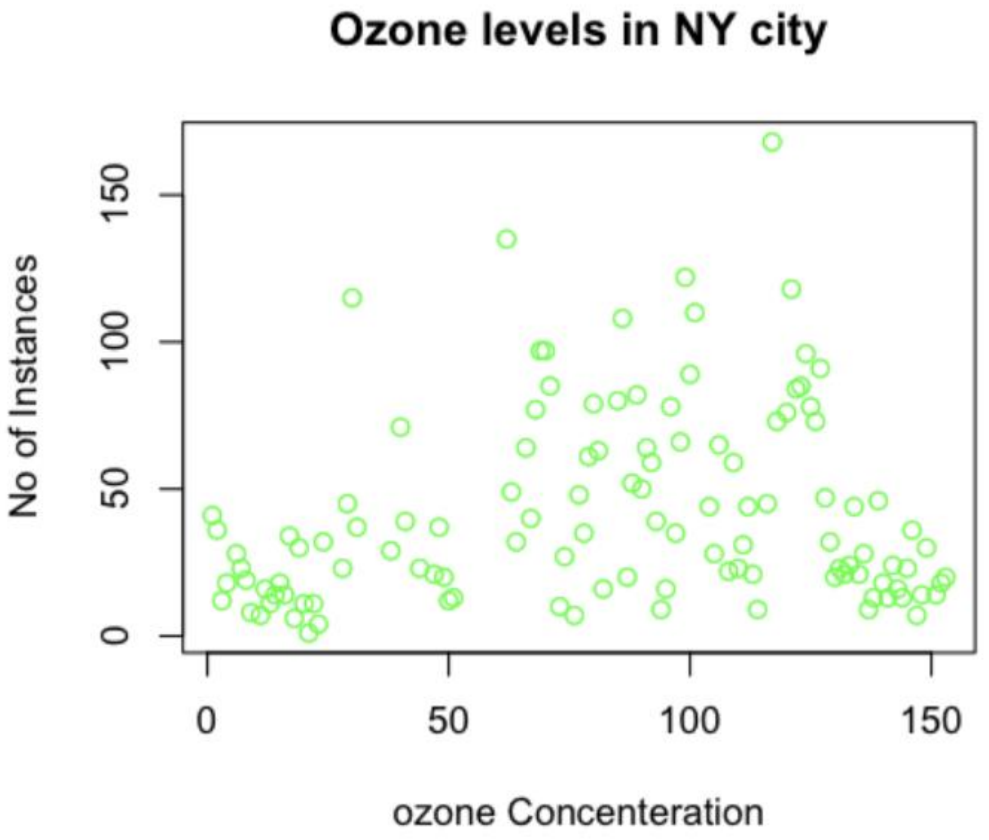
We can quickly discover the relationship between variables by merely looking at the plots drawn between them.

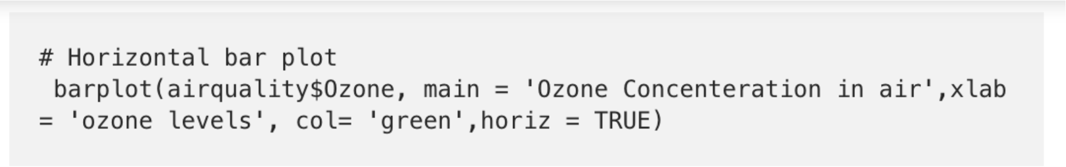
```
# points and lines
 plot(airquality$Ozone, type= "b")
```



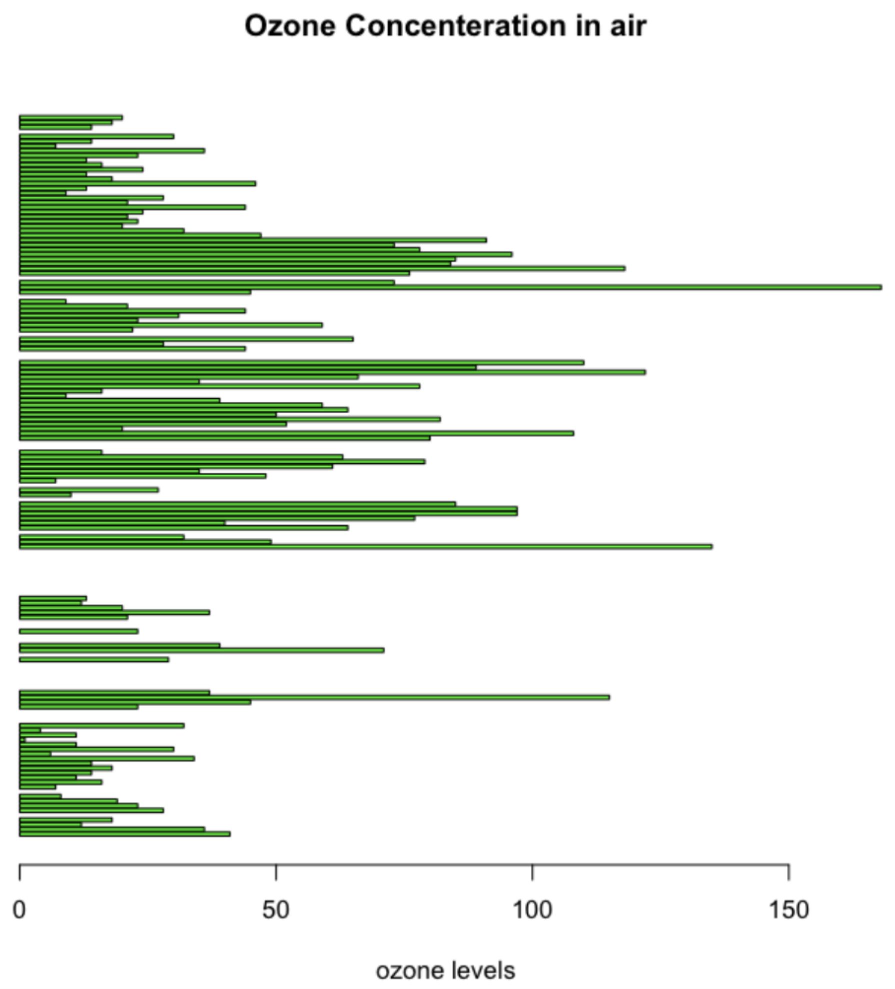


```
# high density vertical lines.
 plot(airquality$Ozone, type= "h")
plot(airquality$Ozone, xlab = 'ozone Concentration', ylab = 'No of
Instances', main = 'Ozone levels in NY city', col = 'green')
```

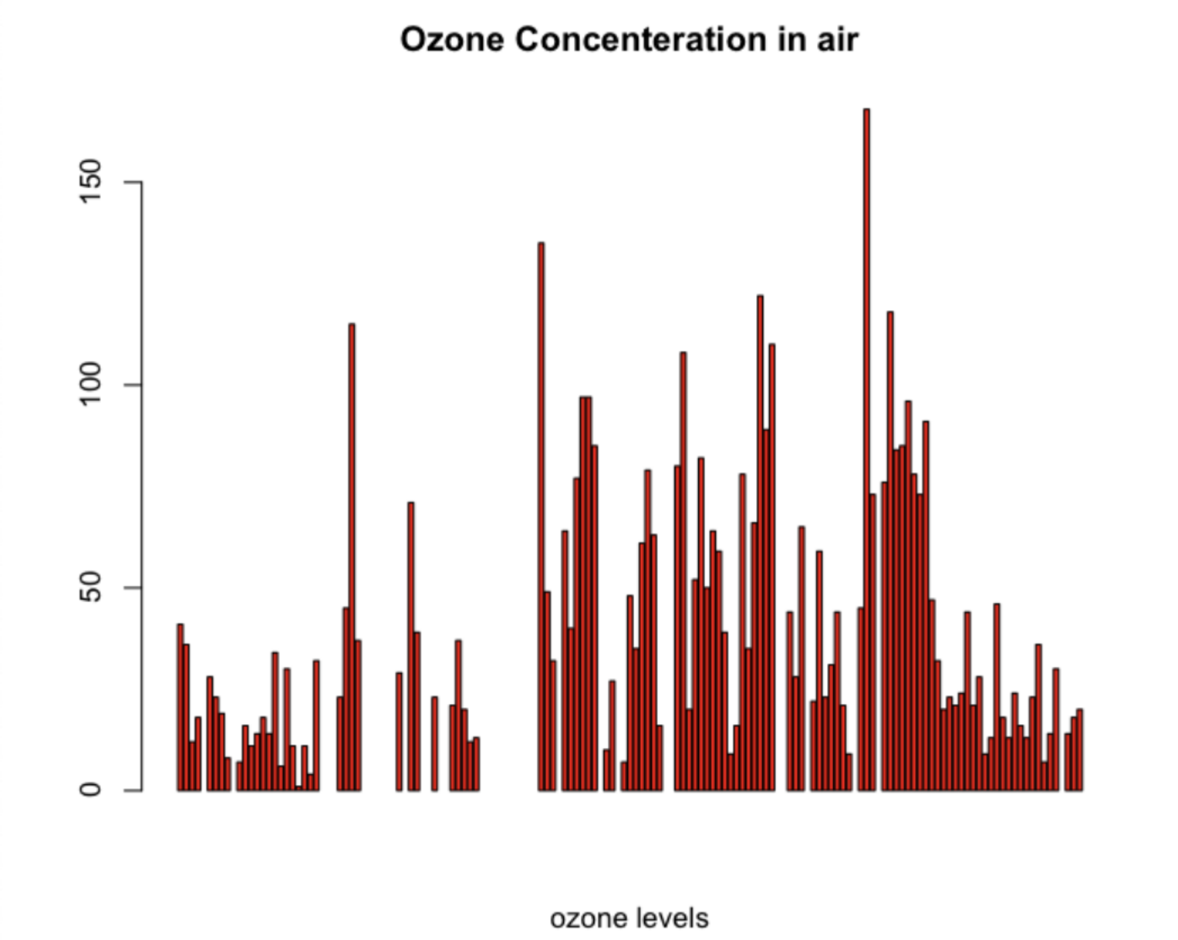




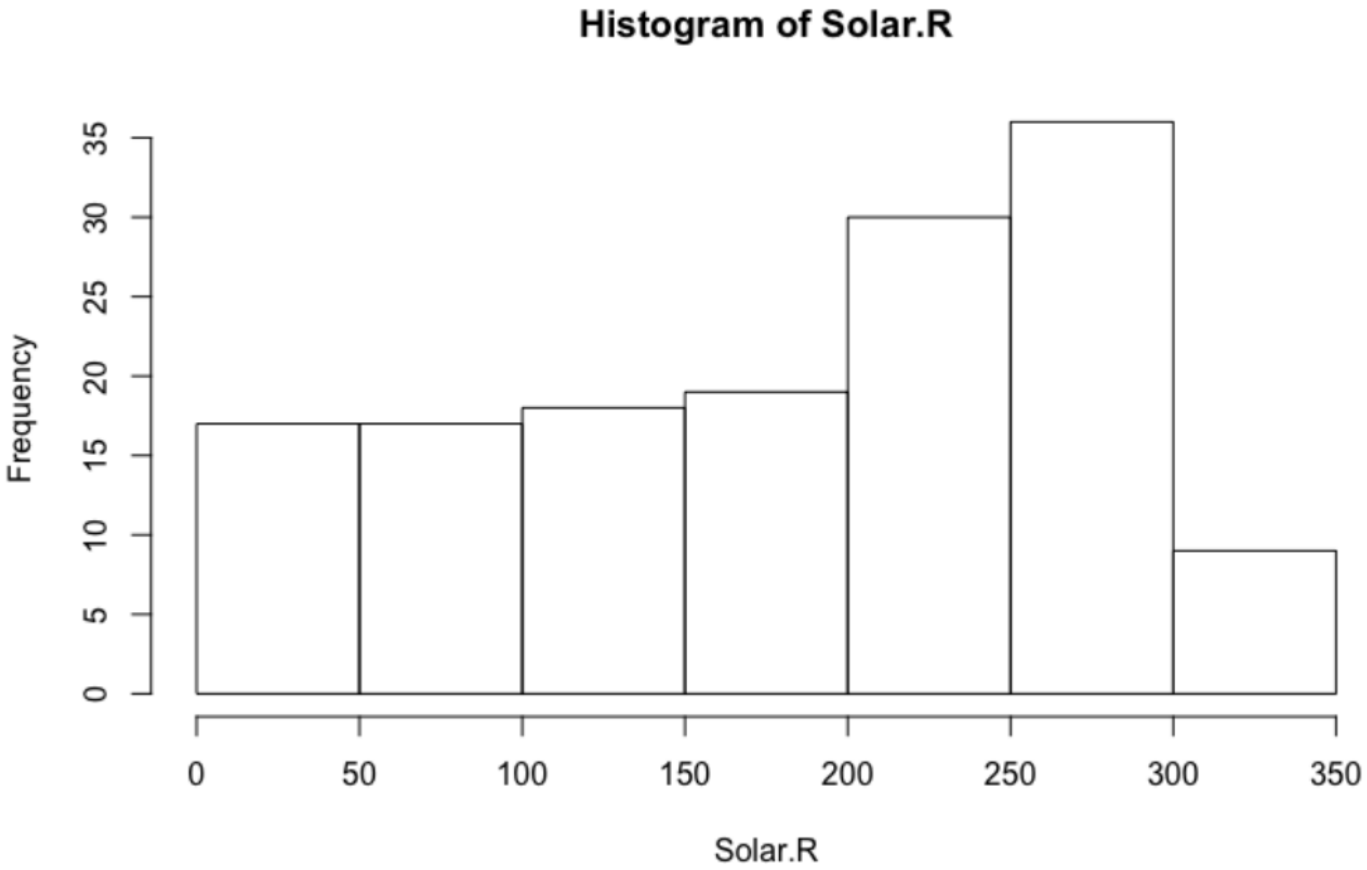
```
# Horizontal bar plot
 barplot(airquality$Ozone, main = 'Ozone Concenteration in air',xlab
= 'ozone levels', col= 'green',horiz = TRUE)
```



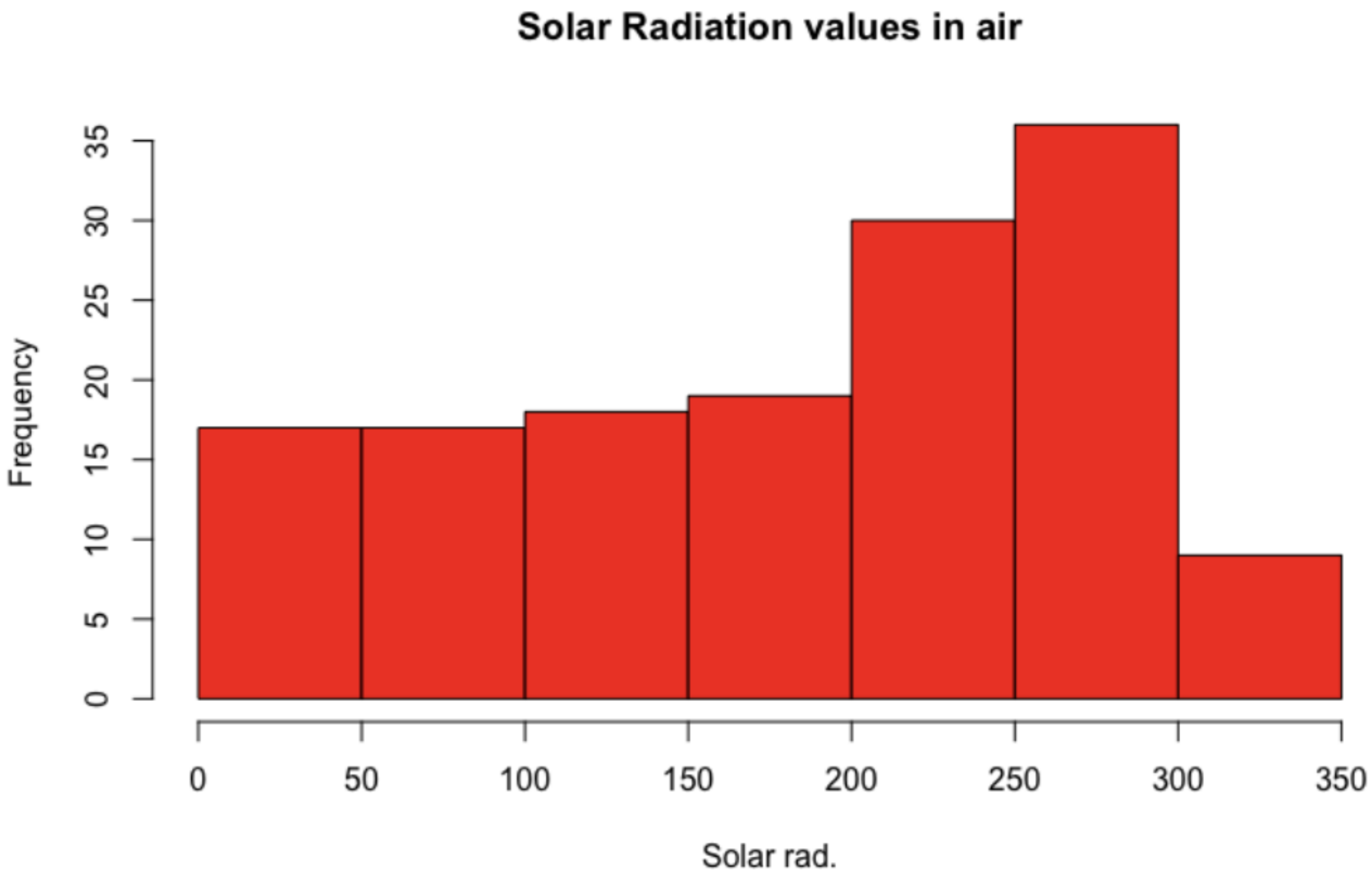
```
# Vertical bar plot
barplot(airquality$Ozone, main = 'Ozone Concenteration in air',xlab
= 'ozone levels', col='red',horiz = FALSE)
```



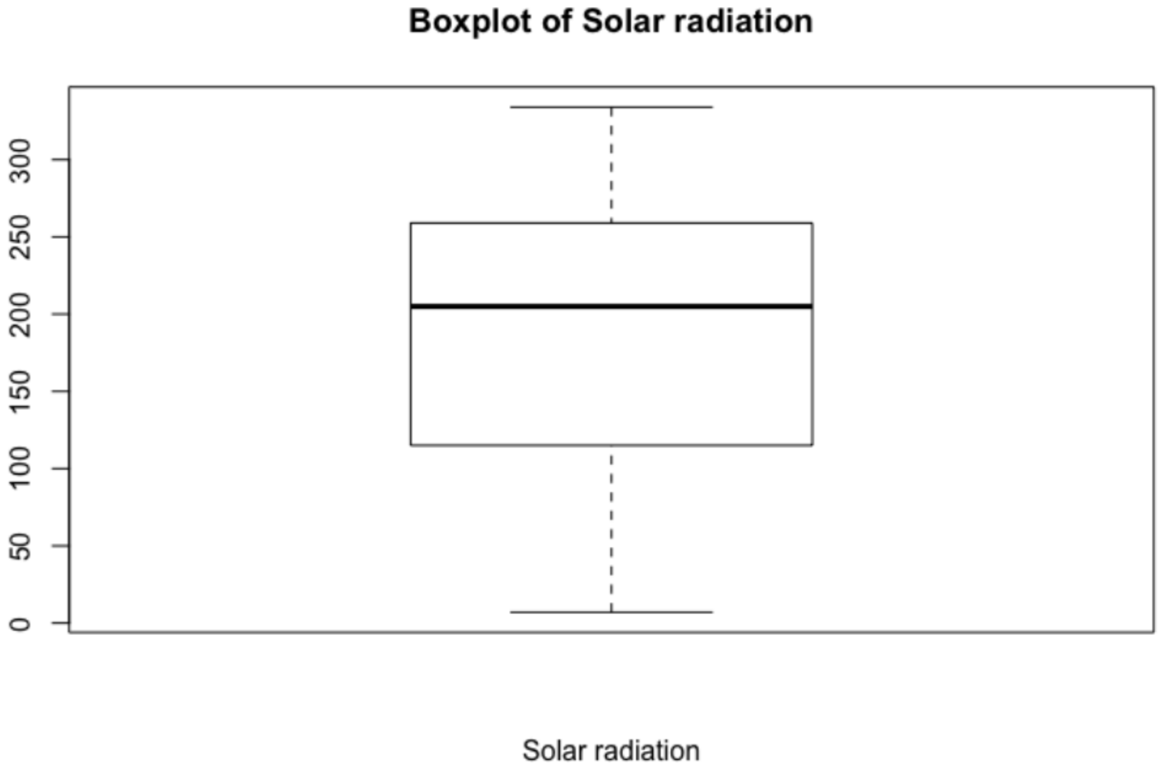
```
hist(airquality$Solar.R)
```



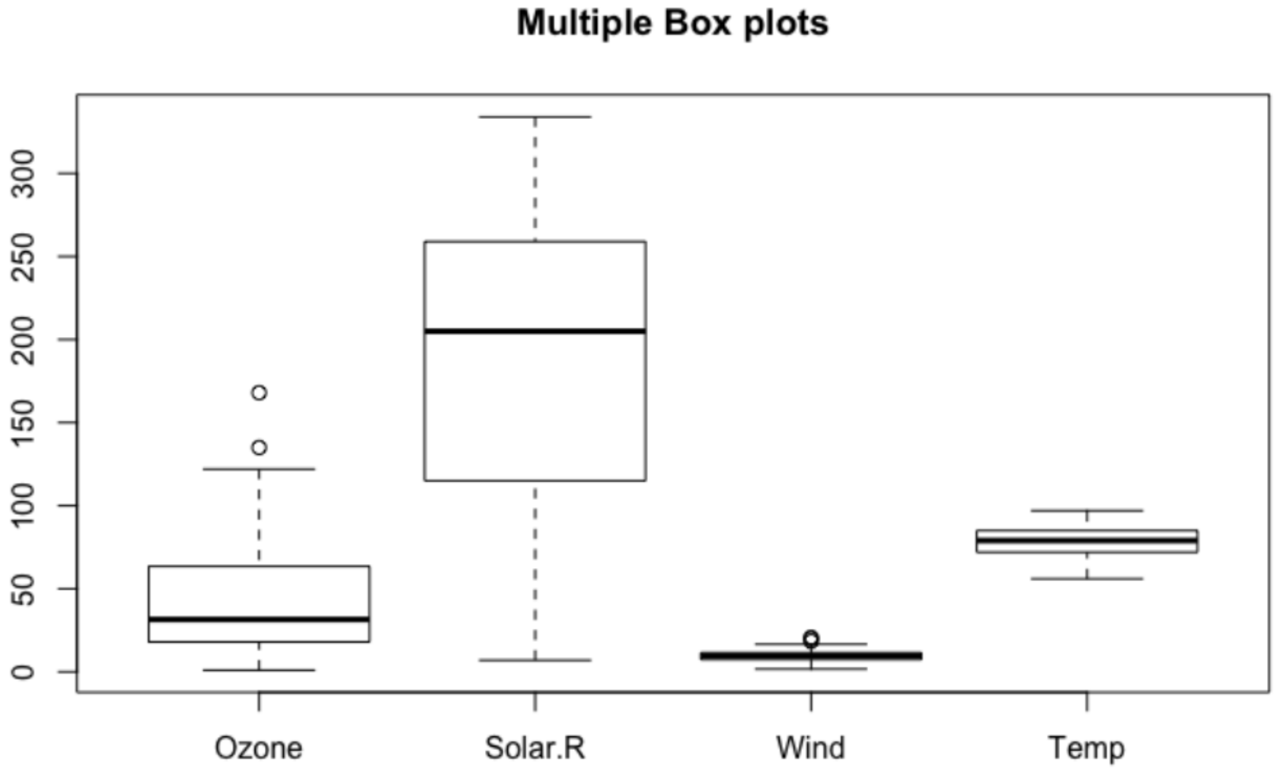
```
hist(airquality$Solar.R, main = 'Solar Radiation values in air',xlab
= 'Solar rad.', col='red')
```

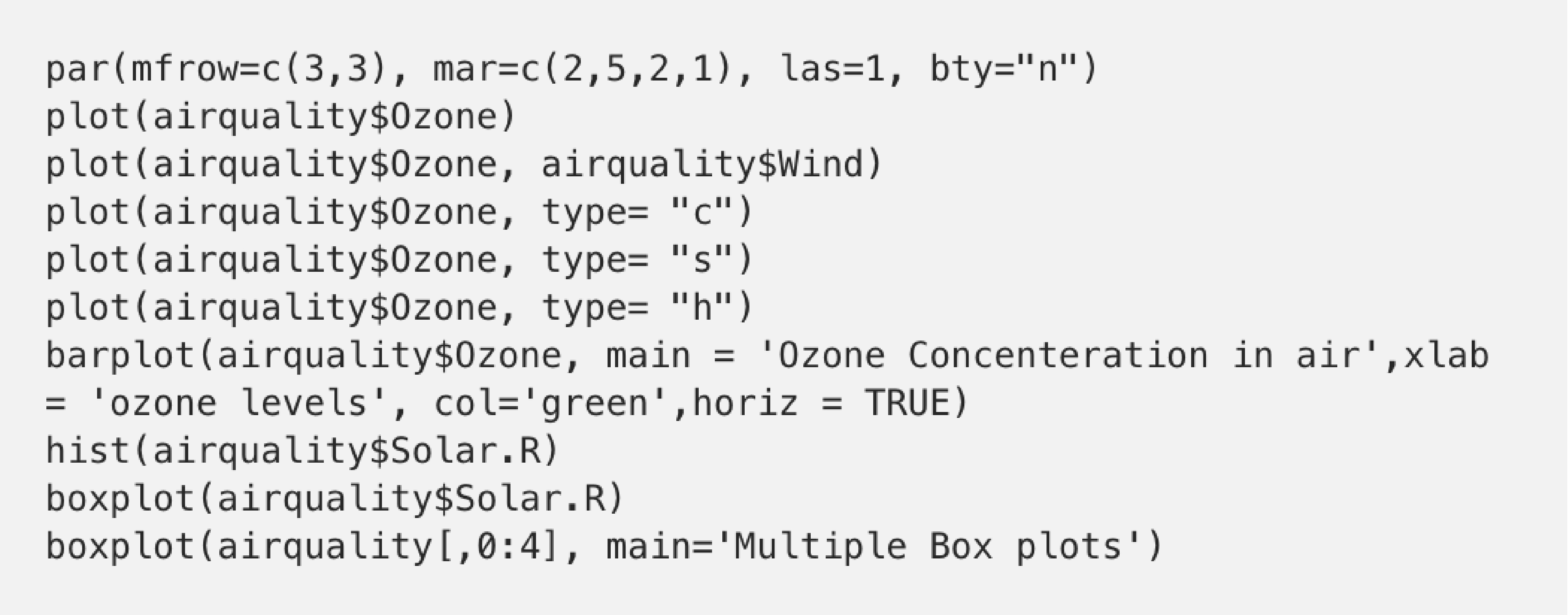


```
#Single box plot
boxplot(airquality$Solar.R)
```



```
# Multiple box plots
boxplot(airquality[,0:4], main='Multiple Box plots')
```





```
par(mfrow=c(3,3), mar=c(2,5,2,1), las=1, bty="n")
plot(airquality$Ozone)
plot(airquality$Ozone, airquality$Wind)
plot(airquality$Ozone, type= "c")
plot(airquality$Ozone, type= "s")
plot(airquality$Ozone, type= "h")
barplot(airquality$Ozone, main = 'Ozone Concenteration in air',xlab
= 'ozone levels', col='green',horiz = TRUE)
hist(airquality$Solar.R)
boxplot(airquality$Solar.R)
boxplot(airquality[,0:4], main='Multiple Box plots')
```

Agenda

1. Basic Commands of R
2. Reproduce Weight/Height exercise in class
3. Predicting sales from price

Agenda

1. Basic Commands of R
2. Reproduce Weight/Height exercise in class
3. Predicting sales from price

Predicting sales from price

BASIC ECONOMICS

What happens when you increase the price of a product?
Do sales increase or decrease?

Why?

But when deciding for the price of a product we need to be more specific…

By how much do sales increase or decrease?

Predicting sales from price

Greenchips is a brand of snacks.

Greenchips snacks are made of dehydrated fruits or vegetables.

They are packaged in 40g bags, as if they were potato chips, but advertised as a much healthier option.

Predicting sales from price

Greenchips produces snacks of dehydrated apple, pineapple and strawberry as well as chips made out of green peas or chickpeas.

Their products are vegan, gluten free, with no palm oil, made of natural ingredients and oven baked instead of fried.

Predicting sales from price

This example uses a sales and price data set of the Greenchips dehydrated apple snack.

The data consist of weekly unit sales (thousands) of the standard 100g package and the weekly average price (in euros) over a period of 104 weeks.

Our objective:

Develop a simple model, based on a linear equation, to **predict the sales from the price**.

Predicting sales from price

Thus, the regression equation is

SALES = 626.6 – 305.6 PRICE

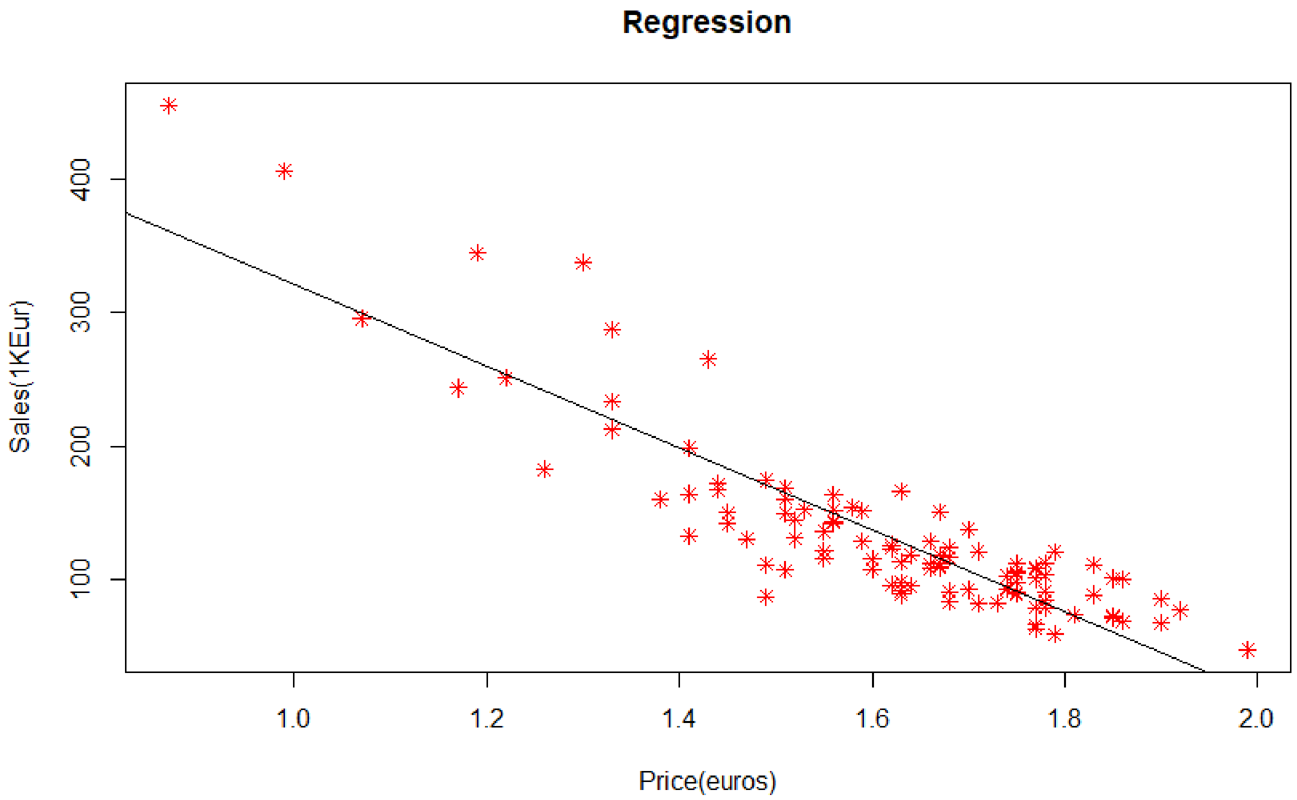
Now, to visualize the data with a scatter plot:

Predicting sales from price

You should get something like this:

Now:

- Is there a correlation?
- If so, positive or negative?
- Weak or strong?
- Can you guess the correlation coefficient (between -1 and 1)?



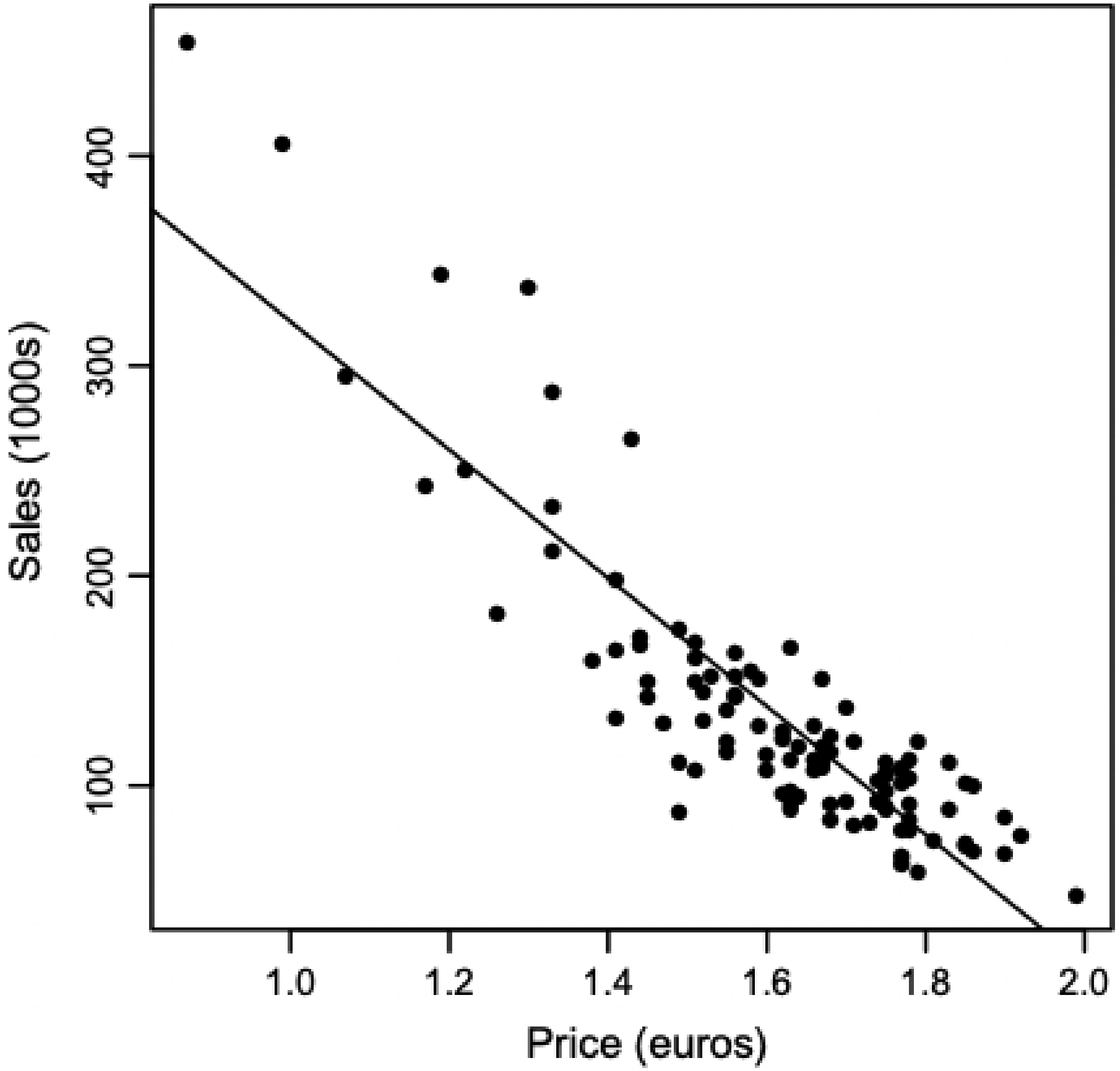
Predicting sales from price

Now we can create a predictive model!

Type in R script the formula you obtained:

Y =626-(305,6*x)

Now, try to add a price in x



Predicting sales from price

Now we can create a predictive model!

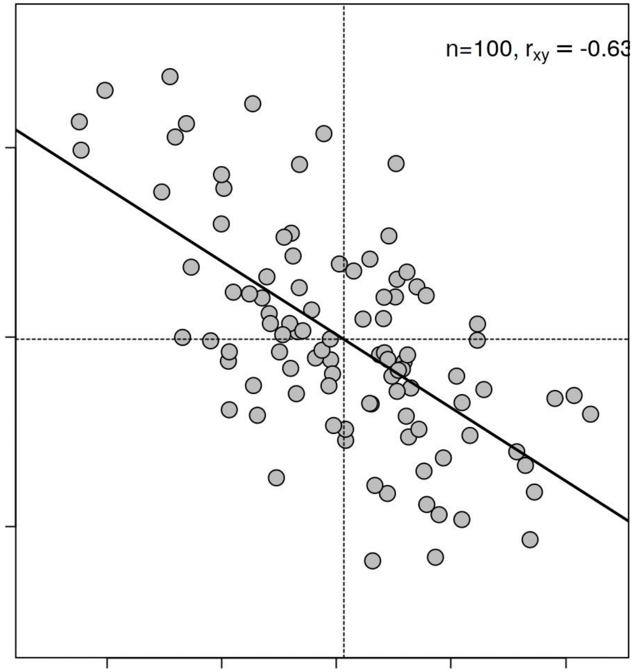
Type in a blank cell the formula you obtained (either through functions or through the graph):

=626-(305,6*F2)

Now, try to add a price in x

- What happens to our sales if we decrease the price to 0,1 euro? Do they go up or down?
- **And if we increase it to 0,2 euros from the original price? Do they go up or down?**

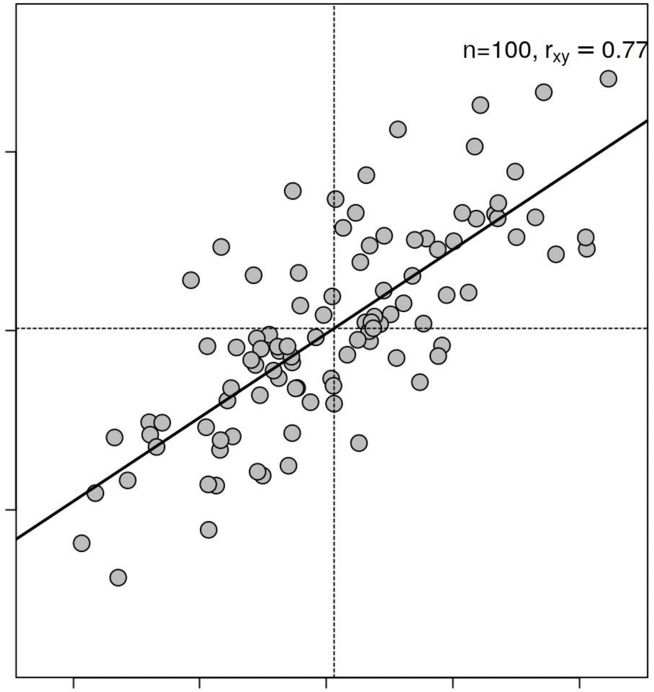
Fundamentals of Econometrics Models



Vicenç Soler

v.soler@tbs-education.org

~~vincent.soler@tbs-education.org~~


n=100, r_{xy} = 0.77