

# Predictive Data Analysis

Lecturer: Marina Iantorno

E-mail: [miantorno@cct.ie](mailto:miantorno@cct.ie)





- ❑ Correlation Coefficient
- ❑ Linear Regression
- ❑ Practice in Python

# Linear Regression

Looking for relationships and making predictions is one of the staples of Data Analysis. We want to answer questions like:

- Could I predict how many units I'll sell if I spend  $x$  amount of money in advertising?
- Does drinking more diet soft drink really relate to more weight gain?
- Can we predict the number of women hired in a company based in their level of education?

# Linear Regression

There is a statistic measure that help us to understand if there is a relationship between two given variables, and if so, how that relationship is. This is called “Correlation Coefficient” and its main goal is determining the degree of the correlation between two variables. The creator was Karl Pearson, and this is why this measure is also known as a Pearson Coefficient or Pearson Product-Moment Correlation.

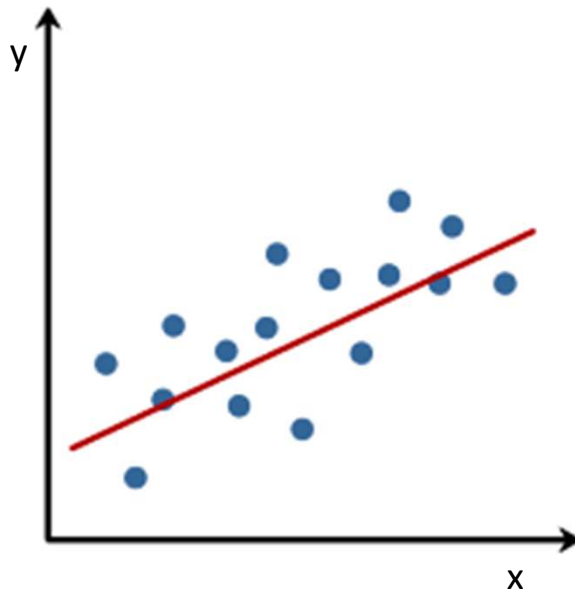
How does it work? The coefficient is represented by the letter “ $r$ ” and could take any value between -1 and 1.

# Linear Regression

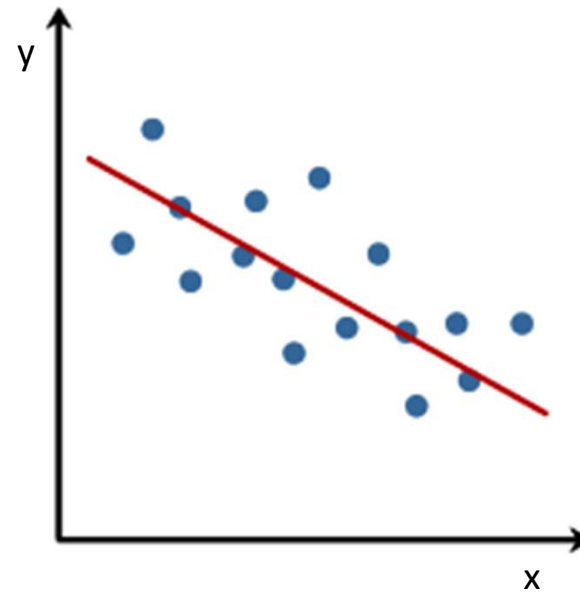
Suppose that we have two variables, name them  $X$  and  $Y$ , and we want to know if there is any correlation between them, we can calculate the Correlation Coefficient ( $r$ ) and analyse the result.

- The sign of  $r$  will tell us whether the relationship is positive or negative. If  $r$  is positive, it means that when  $X$  increases so does  $Y$ , and when  $X$  decreases so does  $Y$ . If  $r$  is negative, it means that when  $X$  increases,  $Y$  decreases, and the other way around.

# Linear Regression



Positive relationship



Negative relationship

# Linear Regression

Now let's focus on the number because it will tell us the strength of the relationship between the variables.

➤ As closer  $r$  is to 1 or to -1, as stronger is the relationship between the variables.

Usually we will take as a precise model the one whose  $r$  is greater or equal to 0.70.

# Linear Regression

We can use this table as an orientation to interpret our results.

Correlation Coefficient	Interpretation of the result
$ 0.80 $ to $ 1 $	Very strong correlation
$ 0.70 $ to $ 0.80 $	Strong correlation
$ 0.40 $ to $ 0.70 $	Moderate correlation
$ 0.20 $ to $ 0.40 $	Weak correlation
0 to $ 0.20 $	Very weak or non correlation



# Linear Regression

Some important concepts about  $r$ :

- A correlation value of zero means that you can find no linear relationship between  $x$  and  $y$  (maybe there is a relationship but it's not linear).
- A correlation value of  $+1$  or  $-1$  indicates that the points fall in a perfect, straight line (Negative values indicate a downhill relationship; positive values indicate an uphill relationship).
- A correlation value close to  $1$  or  $-1$  means a strong relationship.

# Linear Regression

As we saw in the previous graphs, the relationship between the variables is visible through a line, and that is why it is called “Linear regression”. This model tries to represent the relationship between two or more variables. The equation of the model is as follow:

$$y = mx + b$$

Let's have a closer look!

# Linear Regression

Analysing the model:

$$y = mx + b$$



*Y: this is known as the dependent variable. If there are changes in X, the result will be reflected in Y. In other words, Y depends on X.*

# Linear Regression

Analysing the model:

$$y = mx + b$$



*m: this is the slope of the line and indicates the inclination of the line. When m is positive, the line increases, and when m is negative, the line decreases. m and r have the same sign.*

# Linear Regression

Analysing the model:


$$y = mx + b$$



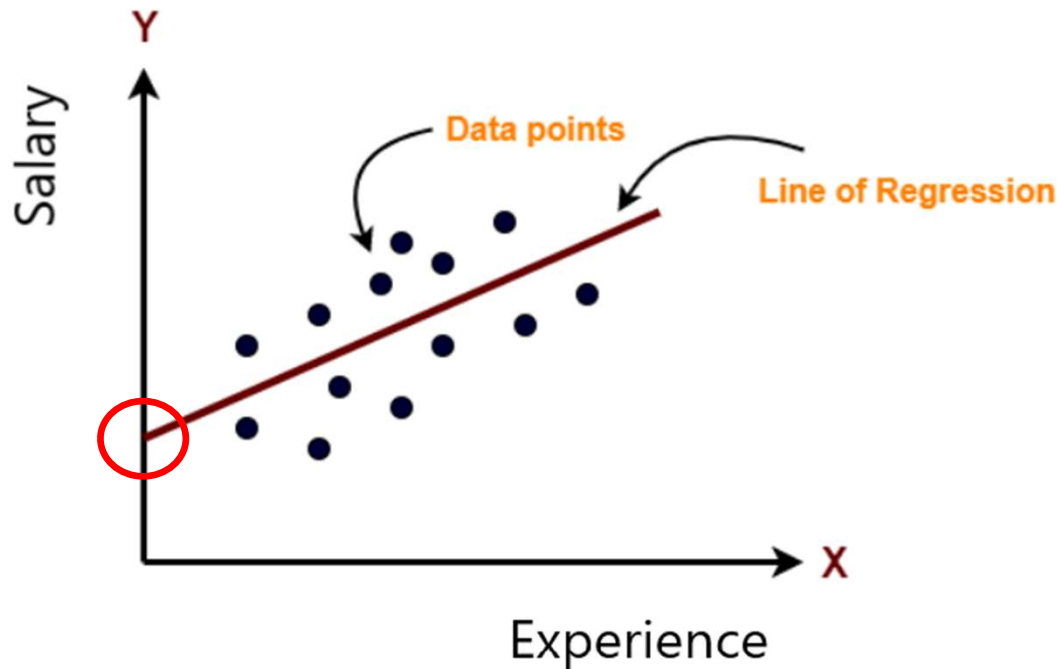
*X: this is known as the independent variable. X could take any value within a specific domain, and will modify the result of Y.*

# Linear Regression

Suppose that we want to analyse the relation between the salary that people receive and their professional experience. We could think that the experience determine the salary, therefore experience is the independent variable (x) and the salary is the dependent variable (y). Something as follow:

$$y = mx + b$$

$$\text{Salary} = m * \text{Experience} + b$$

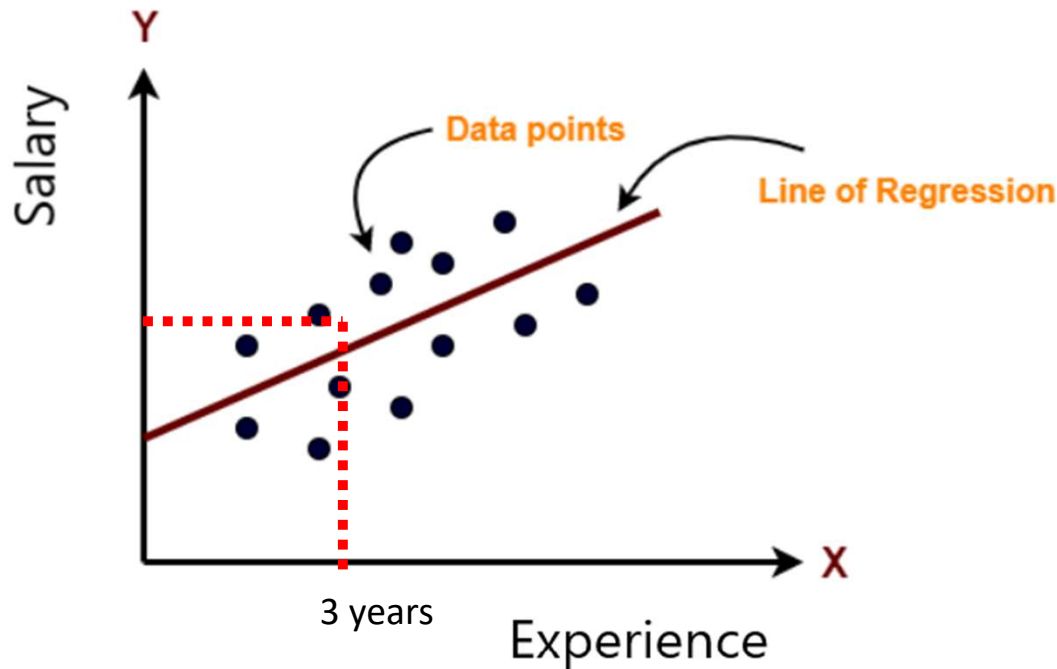
# Linear Regression



$$\text{Salary} = m * \text{Experience} + b$$

*b: this is known as the intersection of the line because this is the point in which the line will cross the Y axis.*

# Linear Regression



$$\text{Salary} = m * \text{Experience} + b$$

*m*: this is the slope of the line, and it is always multiplying the independent variable. Suppose that a person has 3 years of experience, we can project what the salary would be like.



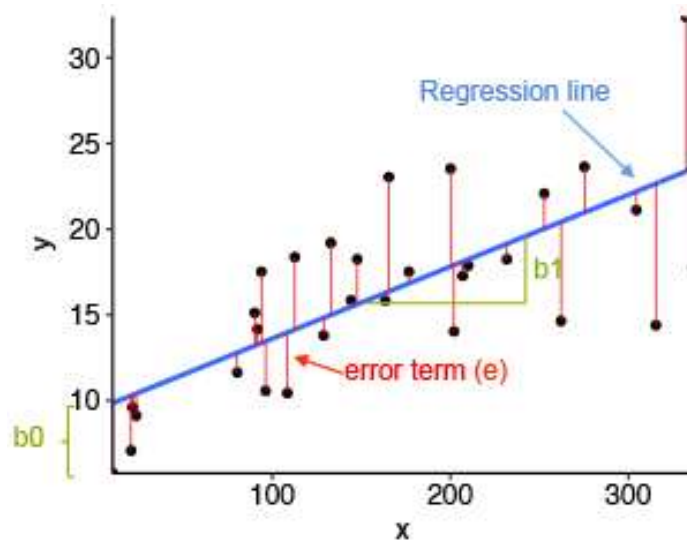
# Linear Regression

Before looking ahead to predicting a value of  $y$  from  $x$  using a line you need to:

- Establish that you have a legitimate reason to do so by using a straight line.
- Feel confident that using a line to make that prediction will actually work well.

Keep in mind that unlike in Mathematics, in Statistics our line will not be perfect, there is some variability and error involved, but the real result should be close to the line we predict.

# Linear Regression



When we build the model, we find the line that best fit for the datapoints. We will see some distance between the points and the line, and this is known as the error, which means, something that is not really explained by this model.

# Linear Regression

## Determination Coefficient

If we want to know the precision of the model, we calculate the determination coefficient.

The only thing we have to do is to square the correlation coefficient   $r^2$

Suppose that we have this result:

$$r^2 = 0.90$$

How do we read this result?

# Linear Regression

## Determination Coefficient

$$r^2 = 0.90$$

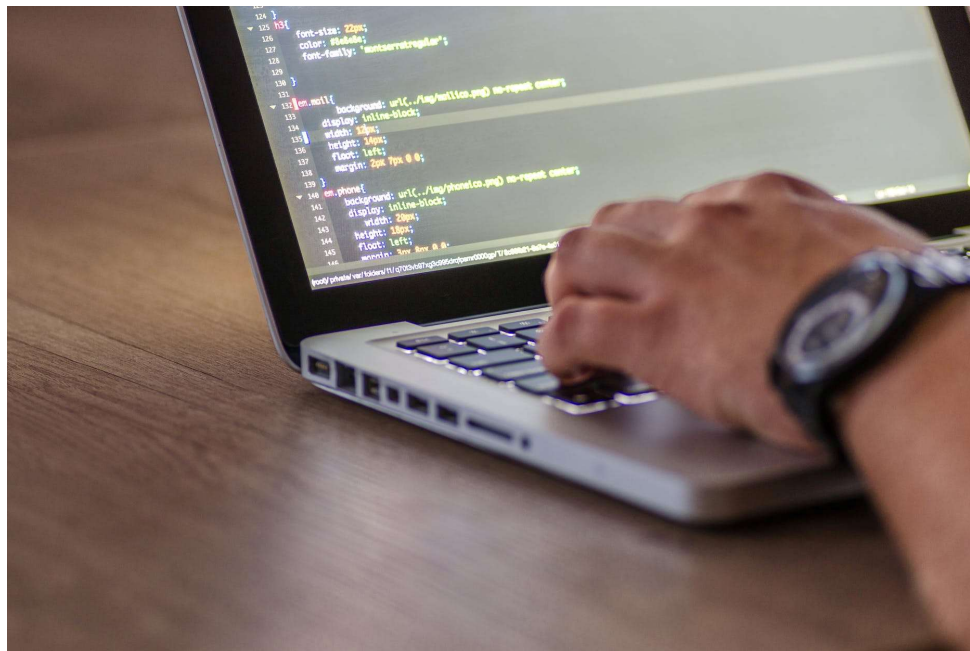
90% of the changes in Y are explained by the changes in X. In other words, the precision of the model is 90%.

The 10% left is not explained by this model, there may be other variables that are affecting Y, but they are not in this analysis.

# Linear Regression

In Data Analysis, we should never make any conclusions about a relationship between two variables based solely on either the correlation or the scatterplot alone. The two elements are to be examined together. Sometimes the scale on the axes could make the graph look better or worse, therefore we should calculate the correlation coefficient to determine whether or not a linear relationship exists.

# Linear Regression



Let's try to code it in Python!

**THAT'S ALL FOR TODAY**

**THANK YOU**

