UNIVARIATE LINEAR REGRESSION

We will use the data in file data_lecture4.xlsx[sheet_name = lin_regression]

STRATEGY TO UNDERSTAND THE MODEL

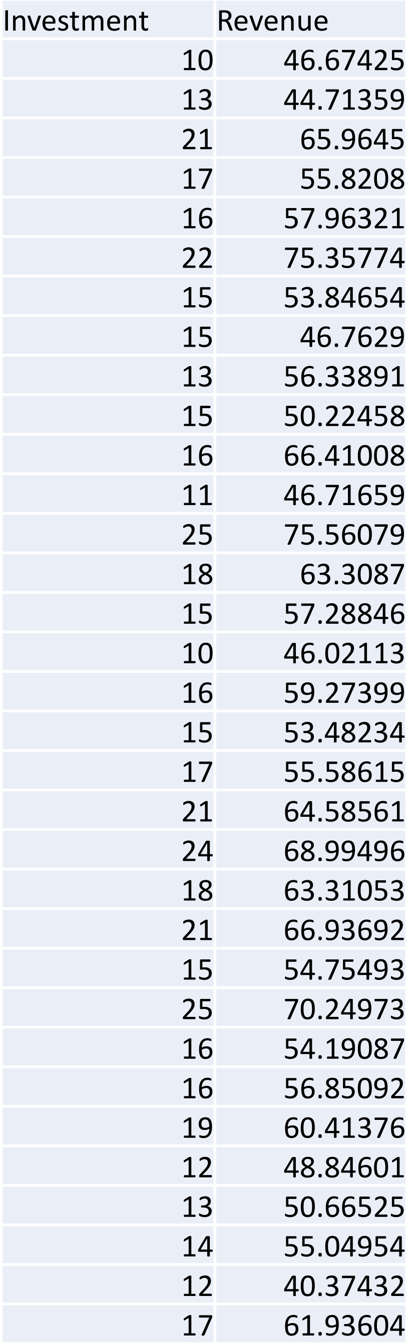
Chose independent and the dependent variables; for instance,

X = Investment, Y = Revenue

Write a straight-line equation Y = aX +b, for some parameter. **a**, and **b**.

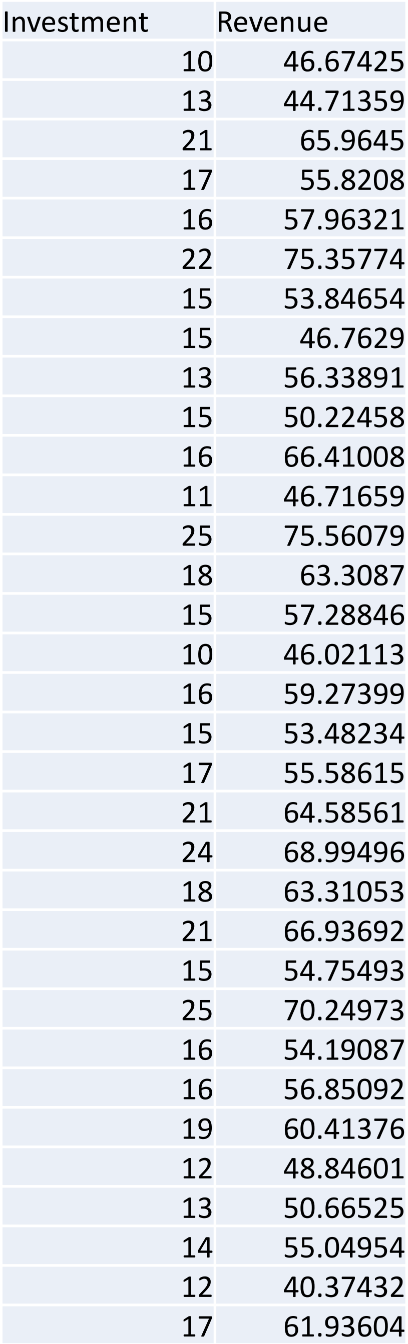
- **a** is also called coefficient or slope.
- **b** is sometimes called intercept.

For every X value, apply the equation to find the predicted value and plot the predicted points on the same graph



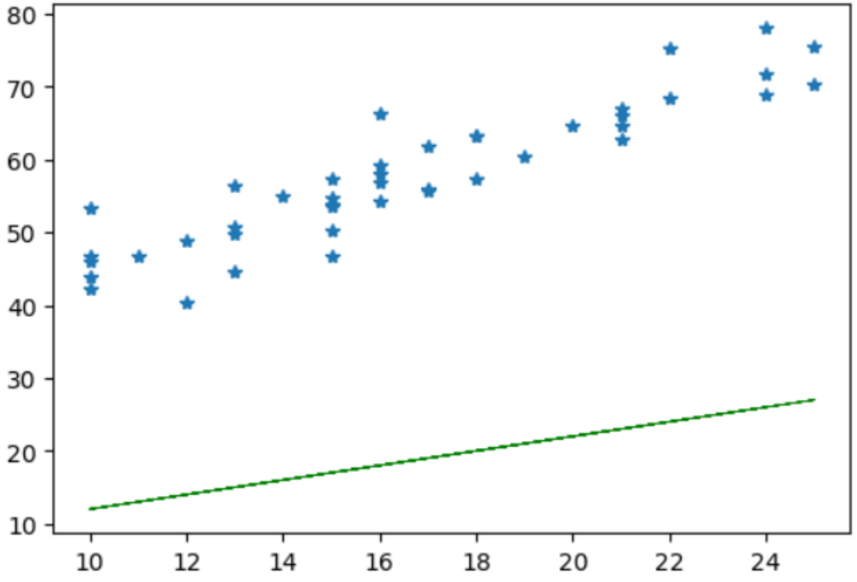
Investment	Revenue
10	46.67425
13	44.71359
21	65.9645
17	55.8208
16	57.96321
22	75.35774
15	53.84654
15	46.7629
13	56.33891
15	50.22458
16	66.41008
11	46.71659
25	75.56079
18	63.3087
15	57.28846
10	46.02113
16	59.27399
15	53.48234
17	55.58615
21	64.58561
24	68.99496
18	63.31053
21	66.93692
15	54.75493
25	70.24973
16	54.19087
16	56.85092
19	60.41376
12	48.84601
13	50.66525
14	55.04954
12	40.37432
17	61.93604

STRATEGY TO UNDERSTAND THE MODEL

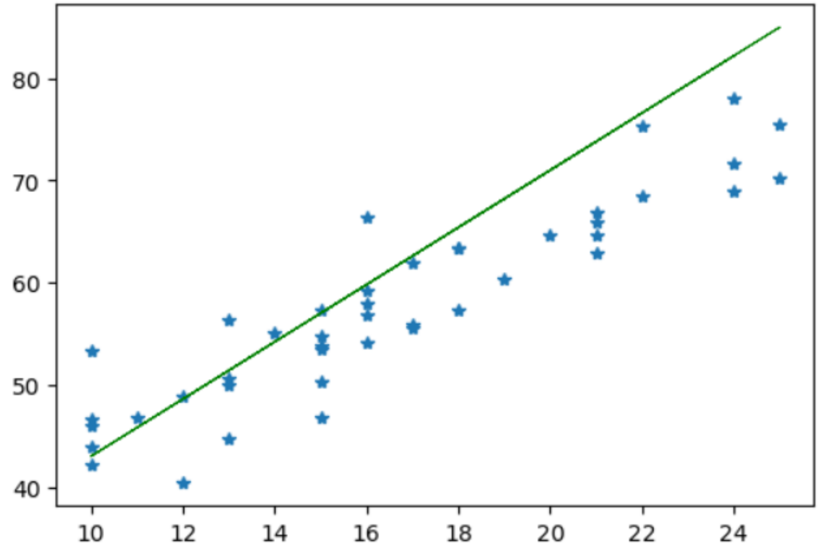


Investment	Revenue
10	46.67425
13	44.71359
21	65.9645
17	55.8208
16	57.96321
22	75.35774
15	53.84654
15	46.7629
13	56.33891
15	50.22458
16	66.41008
11	46.71659
25	75.56079
18	63.3087
15	57.28846
10	46.02113
16	59.27399
15	53.48234
17	55.58615
21	64.58561
24	68.99496
18	63.31053
21	66.93692
15	54.75493
25	70.24973
16	54.19087
16	56.85092
19	60.41376
12	48.84601
13	50.66525
14	55.04954
12	40.37432
17	61.93604

a=1, b=2
Terrible fit



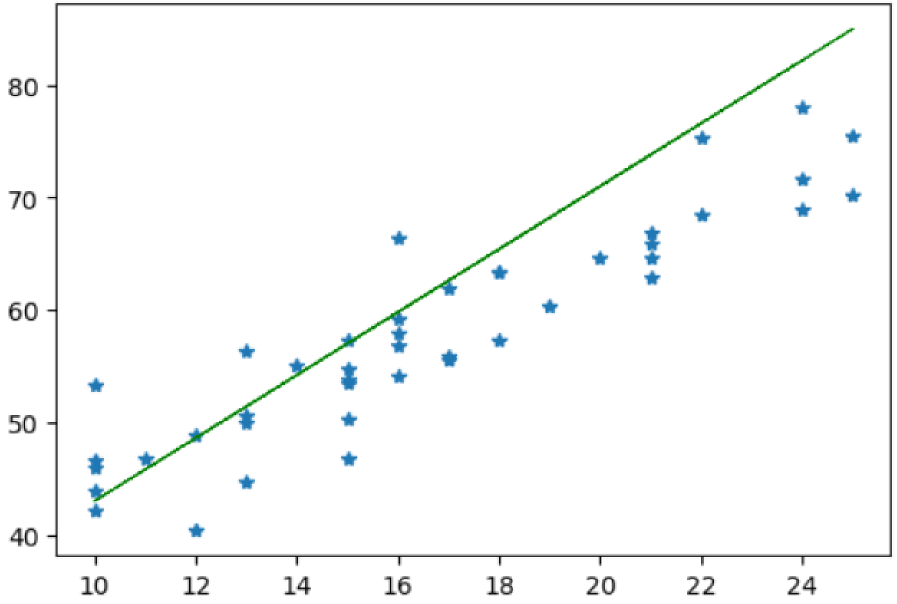
a=2,8, b=15
Better fit, but can be improved



STRATEGY TO UNDERSTAND THE MODEL

We ***declare that*** a model is a better fit if the sum of the squares of the differences from the measured point to the line is smaller. The ***best fit*** is, by this definition, the model that minimises that sum of the squares of the differences.

Why? It can be interpreted as: ‘*the model minimises on average the risk of giving a very wrong prediction.*



Warning:

Are there alternative definitions of best fit?
Yes!!

This topic is beyond the scope of this course and represents an active field of research when the models are Neural Networks (the Nonlinear multivariate statistical model).

REGRESSION IN PYTHON

Let’s go now to Python to see how to solve this linear model in practice.

There are two ways of doing it.
1- Using a complicated formula (that you can find in any textbook)
2- Using one of the available statistical libraries to get the solution.

We will use the pragmatical approach, and we will explore two libraries:

- scikit-learn
- statsmodel

We will use the data in file data_lecture4.xlsx[sheet_name = lin_regression]

For your info, these are the equations; however, you DON’T need to memorise them:

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$