Saint Joseph University - Year 2023-2024

Data Science License - Statistical analysis of data

TD1 Sheet - Linear Regression Analysis

EXERCISE 1

We give the following pairs of observations:

x_i	18	7	14	31	21	5	11	16	26	29
y_i	55	17	36	85	62	18	33	41	63	87

1. The first step is to obtain the data. To do this, you can download them and then save them to your PCs.
2. Draw the scatter diagram of the couples (x_i, y_i). Looking at this diagram, can we suspect a linear relationship between these two variables?
3. Determine the least squares line for these observations, i.e. give the coefficients of the least squared line.
4. Give the ordinates of the y_i calculated by the least squares line corresponding to the different values of the x_i.
5. Then draw the line on the same graph.
6. What is a plausible estimate of Y at $x_i = 21$?
7. What is the difference between the observed value of Y at $x_i = 21$ and the value estimated with the least squares line? What do we call this gap?
8. Does the least squares line obtained in question 3 pass through the point (x, y).? Can we generalize this conclusion to any regression line?

EXERCISE 2

We study the influence of an antibiotic on a bacterial culture. Equal volumes of culture added with a quantity X of antibiotic are distributed into 10 tubes, and the optical density D is measured after incubation. The results are as follows.

X	0.2	0.2	0.4	0.4	0.6	0.6	0.8	0.8	1.0	1.0
D	19	21	35	38	64	66	115	130	200	210

1. Does a linear adjustment seem justified? What coefficient should you calculate with R?
2. Determine a regression equation by specifying what are the explanatory variable and the explained variable?
3. Using R, give a prediction of D for a quantity of antibiotic X = 0.5.

EXERCISE 3

The Transport Bertrand company wants to establish a maintenance policy for the trucks in its fleet. All are of the same model and used for similar transport. Company management believes that a statistical link between the direct cost of travel (cents per km) and the amount of time since that truck was last inspected would be useful. We therefore collected a certain amount of data on these two variables. We want to use linear regression as statistical modeling.

coût direct	10	18	24	22	27	13	10	24	25	8	16
Nombre de mois	3	7	10	9	11	6	5	8	7	4	6
coût direct	20	28	22	19	18	26	14	20	26	30	12
Nombre de mois	9	12	8	10	9	11	6	8	10	12	5

1. Which variable should we identify as dependent variable Y and which should we identify as explanatory variable X.
2. Plot the scatter diagram of these observations. Does the scatterplot suggest a particular form of connection?
3. Calculate the equation of the least squares line.
4. Determine the total change in direct travel cost.
5. Calculate the explained variation.
6. Calculate the residual variation.
7. Calculate the coefficient R^2 and interpret the result.