Saint Joseph University - Year 2023-2024

Data Science License - Statistical analysis of data

TD2 Sheet – Simple Linear Regression

EXERCISE 1

1. Show that the sum of the cross products

$$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

can also be written

$$S_{XY} = \sum_{i=1}^{n} X_i Y_i - n\overline{XY}.$$

2. Show that the sum of the cross products

$$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

can also be written

$$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})Y_i.$$

3. Find the least squares estimators of the slope and the intercept by canceling the partial derivatives with respect to β_0 and β_1 of the sum of the squares of the errors:

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

EXERCISE 2

In the Europe.xlsx file, you will find the population and area of 27 European countries.

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. Calculate the residuals and verify the property that the mean of the residuals is zero.
4. Calculate the error variance estimator.
5. Calculate the variance estimators of β_0 and β_1.
6. Construct a 95% confidence interval for the parameter β_0.
7. Construct a 95% confidence interval for the parameter β_1.
8. Test the hypothesis $H_0: \beta_0 = 0$ (against $H_1: \beta_0 \neq 0$) with a significance threshold $\alpha = 5\%$.
9. Test the hypothesis $H_0: \beta_1 = 0$ (against $H_1: \beta_1 \neq 0$) with a significance threshold $\alpha = 5\%$.
10. Establish the ANOVA table associated with this regression. What can we conclude about parameter β_1?

EXERCISE 3

Eight simultaneous realizations of variables X and Y were observed. We obtain the results:

$$\sum_{i=1}^{8} x_i = 37, \quad \sum_{i=1}^{8} y_i = 85, \quad \sum_{i=1}^{8} x_i y_i = 430, \quad \sum_{i=1}^{8} x_i^2 = 193, \quad \sum_{i=1}^{8} y_i^2 = 968.$$

1. Estimate the parameters β_0 and β_1 of the model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

using the least squares method.

With these estimators, the sum of the squares of the residuals is: $SC_{reg} = 2.7143$.

2. Give a confidence interval for parameter β_1 for a confidence level of 95%.
3. Test the hypothesis $H_0: \beta_0 = 2$ (against $H_1: \beta_0 \neq 2$) with a significance threshold $\alpha = 5\%$.
4. Calculate the coefficient of determination R^2 of this regression.

EXERCISE 4

Given the data presented in the table below, it is the number of calories consumed per day (y_i) and the percentage of agricultural population (x_i) in 11 countries.

x_i	4	5.7	4.9	3	14.8	69.6	63.8	26.2	38.3	24.7	67.5
y_i	3432	3273	3049	3642	3394	2628	2204	2643	2192	2687	2159

1. Graph Y as a function of X.
2. Estimate the parameters β_0 and β_1 of the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

3. Represent the regression line on the graph.
4. Construct a 95% confidence interval for parameter β_1.
5. Establish the ANOVA table associated with this regression. What can we conclude about parameter β_1?