Saint Joseph University - Year 2023-2024

Data Science License - Statistical analysis of data

TD4 Sheet – Correlation

EXERCISE 1

We want to model the average price of houses in Canada as a function of time and we obtain the following data (unfortunately incomplete...):

Année (X)	1980	1981	⋯	2010
Prix moyens des maisons ($) ($Y$)	74721	76236	⋯	249017

An analysis of this data shows that a linear model could potentially be applied to explain house prices from year to year. We calculate the sample covariance and correlation, and we obtain the following quantities:

$$Cov(X,Y) = 374225$$
$$r_{xy} = 0.77$$

1. We decide to express the price of houses in thousands of dollars rather than in dollars. What happens to covariance and correlation?
2. Let's keep the initial prices (in dollars). We now want to express time in number of years since 1980. What happens to the covariance and the correlation?

EXERCISE 2

R's Orange database contains information on 35 orange trees. We are interested in the following 2 variables:

- "age" represents the age of the trees, it is a continuous quantitative variable. It is measured in days, the youngest tree is 118 days old and the oldest 1582 days old.
- "circumference" represents the circumference of the shaft, it is a continuous quantitative variable, measured in mm. the smallest circumference is 30 mm and the largest is 214 mm.

1. Does a linear adjustment seem justified? What coefficient should you calculate with R?
2. Calculate the residuals and verify the property that the residuals are normally distributed.

3. Calculate the coefficient of determination R^2 and interpret the result.
4. Calculate the correlation coefficient r and interpret the result.
5. Specify the link between R^2 and r.
6. Calculate the estimator of β_1 using the correlation coefficient r.
7. Test the hypothesis H_0: ρ: (age,circumference) = 0 (against H_1: ρ (age,circumference)≠ 0) with a significance threshold $\alpha = 5\%$
8. Test the hypothesis H_0: $\beta_1 = 0$ (against H_1: $\beta_1 \neq 0$) with a significance threshold $\alpha = 5\%$.
9. Establish the ANOVA table associated with this regression. What can we conclude about parameter β_1?
10. Construct a 95% confidence interval for parameter 1. What can we conclude about parameter β_1

EXERCISE 3

We recorded for different countries the GDP per capita in 2004 X (in dollars) and the gross enrollment rate of those under 24 in the same year Y (in percentage). The results are as follows:

$$\sum_{i=1}^{8} x_i = 39457, \quad \sum_{i=1}^{8} y_i = 509, \quad \sum_{i=1}^{8} x_i y_i = 2763685, \quad \sum_{i=1}^{8} x_i^2 = 245474957, \quad \sum_{i=1}^{8} y_i^2 = 33685.$$

We seek to explain the schooling rate as a function of GDP.

1. Identify the variable to be explained and the explanatory variable.
2. Specify the conditions for applying simple linear regression and give the equation of the theoretical model.
3. Calculate the covariance between X and Y then the correlation coefficient r. Interpret the result.
4. Give the estimated values of the unknown coefficients β_0 and β_1.
5. Determine the coefficient of determination R^2 and interpret the result.
6. Test the hypothesis H_0: ρ(X,Y) = 0 (against H_1: $\rho(X, Y) \neq 0$) with a significance threshold $\alpha = 5\%$