Saint Joseph University - Year 2023-2024

Data Science License - Statistical analysis of data

TD5 Sheet – Model diagnostics

EXERCISE 1

R's LifeCycleSavings database contains information on 50 different countries. These data are averages over 1960-1970 (to eliminate business cycle or other short-term fluctuations). dpi is per capita disposable income in US dollars, ddpi is the percentage rate of change in per capita disposable income, sr is aggregate personal savings divided by disposable income. The percentage of population under 15 (pop15) and over 75 (pop75) are also recorded. Data are from Belsley, Kuh and Welsch (1980).

We seek to explain sr as a function of pop15, pop75, dpi and ddpi.

1. Graph sr as a function of pop15, pop75, dpi and ddpi.
2. Construct the multiple linear regression of sr on pop15, pop75, dpi and ddpi.
3. Give the estimated values of the unknown coefficients β_0, β_1, β_2, β_3 and β_4.
4. Construct 95% confidence intervals for the parameters β_1, β_2, β_3 and β_4. Conclude.
5. Determine the coefficient of determination R^2 and interpret the result.
6. Calculate the correlation coefficients between these variables taken two by two. Conclude.
7. Test the following null hypotheses, at the significance level α = 5%, using an appropriate F ratio:
   a) $H_0: \beta_1 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \varepsilon$
   b) $H_0: \beta_2 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \varepsilon$
   c) $H_0: \beta_3 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \varepsilon$
   d) $H_0: \beta_4 = 0$ in the model $sr = \beta_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 dpi + \beta_3 ddpi + \varepsilon$

8. Test the hypothesis $H_0: \rho(\mathrm{sr}, \mathrm{pop15}) = 0$ (against $H_1: \rho(\mathrm{sr}, \mathrm{pop15}) \neq 0$) with a significance threshold α = 5%.
9. Test the hypothesis $H_0: \rho(\mathrm{sr}, \mathrm{pop75}) = 0$ (against $H_1: \rho(\mathrm{sr}, \mathrm{pop75}) \neq 0$) with a significance threshold α = 5%.
10. Calculate the residuals and verify the property that the residuals are normally distributed.
11. Graph the model residuals.
12. Which countries correspond to the largest and smallest residual.
13. The multiple regression model can be written in matrix form as follows:

$$Y = \beta X + \varepsilon$$

Give the matrix X.

14. Calculate the levers of the observations:

$$h_{ii} = x_i(X'X)^{-1}x_i' \text{ avec } i = 1, \cdots, 50.$$

15. Graph the levers.
16. Prove that the sum of the levers is equal to p + 1.
17. Give the number of levers which are greater than $2\frac{p+1}{n}$. Name the countries that correspond to these levers.
18. Calculate the internal studentized residuals. How many points are suspected?
19. Calculate the external studentized residuals. How many points are suspected?
20. Calculate Cook's distance.

EXERCISE 2

The R mtcars database contains information on 32 automobiles (1973-74 models). The data was taken from the 1974 American magazine Motor Trend and includes fuel consumption and 10 aspects of automobile design and performance: “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.

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Type "?mtcars" to find out more information.

We seek to explain mpg as a function of hp, wt, and the iteration between wp and hp (hp:wt).

1. Construct the multiple linear regression of mpg on hp, wt and hp:wt.
2. Give the estimated values of the unknown coefficients β_0, β_1, β_2, and β_3.
3. Construct 95% confidence intervals for the parameters β_1, β_2, and β_3. Conclude.
4. Determine the coefficient of determination R^2 and interpret the result.
5. Calculate the residuals and verify the property that the residuals are normally distributed.
6. Graph the model residuals.
7. Which automobile corresponds to the largest and smallest residue?
8. The multiple regression model can be written in matrix form as follows:

$$Y = \beta X + \varepsilon$$

   Give the matrix X.
9. Calculate the levers of the observations:

$$h_{ii} = x_i(X'X)^{-1}x_i' \text{ avec } i = 1, \cdots, 50.$$

10. Graph the levers.
11. Prove that the sum of the levers is equal to p + 1.
12. Give the number of levers which are greater than $2\frac{p+1}{n}$. Name the countries that correspond to these levers.
13. Calculate the internal studentized residuals. How many points are suspected?
14. Calculate the external studentized residuals. How many points are suspected?
15. Test the hypothesis H_0: Homoscedasticity of errors (against H_1: Heteroscedasticity of errors) with a significance threshold α = 5%.
16. Propose a remedy for the heteroskedasticity problem.

EXERCISE 3

The productivity Y of a variety of trees is evaluated according to its population density X (in plants per m2). The following table provides a summary of the values (x_i, y_i).

X	1.11	1.22	1.49	2.01	2.46	3	3.22	3.67	4.02
Y	1.73	1.49	1.1	0.7	0.52	0.39	0.31	0.2	0.17

1. Draw the corresponding cloud of points and explain why we do not consider a linear adjustment.
2. Construct the simple linear regression of Y as a function of X.
3. Graph the residuals against the estimated values. What do we notice?

We would like to estimate productivity as a function of density using a relationship of the form $Y = \alpha X^{\beta}$. To do this, we set Z = ln(X) and T = ln(Y).

a) Draw the cloud of points associated with the series (Z,T). What do we notice?
b) Determine the linear correlation coefficient between Z and T. Interpret.
c) Estimate the coefficients of the fitting line from T to Z. Give the regression equation.
d) Plot the QQ-plot of the residuals.
e) Derive a relationship allowing productivity to be a function of density.
f) Can we predict the productivity value for a density of 5m2.