ST3131 Regression Analysis - Tutorial 10

(1) A data set (mercedes_car_prices.csv) was collected to study the car prices of used Mercedes-Benz S Class. Twenty eight cars were sampled. Let y be the price of a car in thousands of dollars and x be the age of the car in years. Cars that are less than five years old are of the latest models. Cars that are more than five years old are of the old models.

(i) Plot y against x_1 with the points differentiated by two colours representing new and old models of cars. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$ and display it on the plot of y against x_1. Comment on the fitted model.

(ii) Fit a multiple linear regression model using y, x_1 and the categorical variable x_2 $(x_2 = 0 \text{ if new model}, x_2 = 1 \text{ if old model})$, without interaction between x_1 and x_2. Display the two lines on the plot in part (i). Comment on the model fitted.

(iii) Test to see if the categorical variable in part (ii) is significant.

(iv) Fit a multiple linear regression model using y, x_1 and the categorical variable x_2 $(x_2 = 0 \text{ if new model}, x_2 = 1 \text{ if old model})$, with interaction between x_1 and x_2. Display the two lines on the plot in part (i). Comment on the model fitted.

(v) Test to see if the interaction in part (iv) is significant.

(vi) Based on the analysis in parts (i) to (v), suggest the best model based on y, x_1 and the categorical variable x_2.

(2) The data set data-table-B3.csv contains data on the gasoline mileage performance of different automobiles. All the variables are shown below. Do not do any conversion of units. Use the statement `newdat <- na.omit(data_set_name)` to remove observations containing missing values before you analyse the data.

Mileage is the number of miles per gallon. Displacement is engine capacity. Torque is the force responsible for making a car accelerate quickly. Horsepower is a measurement used to calculate how quickly the force is produced from a car's engine. Compression ratio is the ratio between the maximum and minimum cylinder volume when the piston is at the bottom and top of its stroke respectively. Real axle ratio refers to the number of times a driveshaft turns in relation to the number of times the driven wheel turns.

$y = mileage$
$x_1 = displacement$
$x_2 = horsepower$
$x_3 = torque$
$x_4 = compression\_ratio$
$x_5 = real\_axel\_ratio$
$x_6 = length$
$x_7 = width$
$x_8 = weight$
$x_9 = transmission$ (1 for Automatic and 0 for manual)
$x_{10} = carburetor$ (number of carburetors)

$x_{11} = n\_speeds$ (number of speeds)

The $real\_axel\_ratio$ can be classified as

$$raratio = \begin{cases} \text{low}, & real\_axel\_ratio < 2.6, \\ \text{low-medium}, & 2.6 \leq real\_axel\_ratio < 3, \\ \text{medium}, & 3.0 \leq real\_axel\_ratio < 3.5, \\ \text{high}, & real\_axel\_ratio \geq 3.5. \end{cases}$$

(i) Plot y against x_1 with the points differentiated by four colours representing the four levels of $raratio$.

(ii) Fit the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon,$$

and display the fitted model on the plot in part (i). Comment on the model fitted.

(iii) Fit a multiple linear regression model using y, x_1 and the categorical variable $raratio$, without interaction between x_1 and $raratio$. Display the four lines on the plot in part (i). Comment on the model fitted.

(iv) Test to see if the categorical variable in part (iii) is significant.

(v) Fit a multiple linear regression model using y, x_1 and the categorical variable $raratio$, with interaction between x_1 and $raratio$. Display the four lines on the plot in part (i). Explain why the four lines appear to be linear even though we have the quadratic term x_1^2 in the model. Comment on the model fitted.

(vi) Test to see if the interaction in part (v) is significant.

(vii) Based on the analysis in parts (i) to (vi), suggest the best model based on y, x_1 and the categorical variable $raratio$.

(3) A experiment was conducted to see if there is any difference among four types of oil used in preparing popcorns. Five samples of popcorns were prepared using each type of oil. The same number of kernels and same amount of oil were used each time and the number of unpopped kernels was counted. The data set popcorn.csv contains the data.

(i) Contruct boxplots to compare the four types of oil. Which type of oil yield the least unpopped kernels?

(ii) Analyse the data using the design of experiment approach. State your findings.

(iii) Analyse the data using the regression analysis approach. State your findings.