ST3131 Regression Analysis - Tutorial 1

1. The data set data-table-B3.csv contains data on the gasoline mileage performance of different automobiles. Write R codes to read in the data set and answer the following questions. [Hand calculations are not encouraged. Use R.]

   (i) How many variables are there in the data set?

   (ii) The first variable *mileage* is the number of miles per gallon and the second variable *displacement* is the capacity of an engine in cubic in. Which should be the response variable and which should be the regressor variable? Make a plot of *mileage* against *displacement*. Comment on any relationship found.

   (iii) Convert the unit of mileage to km per litre using 1 mile per gallon = 0.425 km per litre, and unit of displacement to cubic centimetre (cc) using 1 cubic in = 16.387 cc. Make a plot of mileage against displacement. Is there any difference between this plot and the plot obtained in part (ii)? Why?

   (iv) Use the R function $lm()$ to fit a simple linear regression model

   $$mileage = \beta_0 + \beta_1 \; displacement + \epsilon$$

   and plot the least-squares line on the plot obtained in part (iii). Comment on the fit of this line.

   (v) Let $y = mileage$ and $x = displacement$. Write R codes to calculate $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy} and use these to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$. Compare your answers obtained using the R function $summary()$.

   (vi) Write R codes to calculate the residual $e_i = y_i - \hat{y}_i, i = 1, ..., n$. Calculate the sums $\sum_{i=1}^{n} e_i$ and $\sum_{i=1}^{n} e_i^2$. Is it possible to find another straight line that yields a smaller value than $\sum_{i=1}^{n} e_i^2$?

   (vii) Use R to construct a plot of e_i against x_i. Draw a horizontal line at $residual = 0$ on the plot. What can you learn from this plot?

   (viii) A Toyota Camry has a displacement of 1998 cc. Predict the mileage using the regression line. From Toyota's website, it is stated that the fuel consumption is 7.3 L/100 km. Comment. [Hint: You may use the $predict()$ function]

   (ix) A Mercedes Benz E250 has a displacement of 1991 cc. Predict the mileage using the regression line. From Mercedes Benz's website, it is stated that the fuel consumption is 5.3-6.9 L/100 km. Comment.

2. (i) Show that $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

(ii) Show that $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i(y_i - \bar{y})$

3. The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2,\ i = 1, 2, ..., n$ and ϵ_i's are independent. Show that the sum of residuals $\sum_{i=1}^{n} e_i$ from the least-squares regression line is zero, where $e_i = y_i - \hat{y}_i$.

4. The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2,\ i = 1, 2, ..., n$ and ϵ_i's are independent.

(i) Show that $Cov(\bar{y}, \hat{\beta}_1) = 0$.

(ii) Show that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2 / S_{xx}$.