ST3131 Regression Analysis - Tutorial 7

1. 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)

   Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \epsilon$.

   (i) Plot the raw residual against $\hat{y}$, x_1, and x_4.

   (ii) Plot the standardized residual against $\hat{y}$, x_1, and x_4.

   (iii) Plot the studentized residual against $\hat{y}$, x_1, and x_4.

   (iv) Plot the PRESS residual against $\hat{y}$, x_1, and x_4.

   (v) Plot the R-student residual against $\hat{y}$, x_1, and x_4.

   (vi) Comment on the residual plots and identify all possible outliers.

   (vii) Construct a normal Q-Q probability plot of the raw residuals. Comment on the plot.

   (viii) Test the normality of the residuals using the Shapiro-Wilk test.

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)

   Fit the model $y = \beta_0 + \beta_5 x_5 + \epsilon$.

   (i) Plot y against x_5. Show the least-squares regression line on the same plot. Comment on the fit.

   (ii) Plot the raw residual against $\hat{y}$ and x_5. Comment on the plots.

   (iii) Construct an ANOVA table that includes the lack of fit (LOF) and pure error sum of squares. Test the LOF and comment.

3. The data set vapor.csv contains the vapor pressure of water for various temperatures.

   (i) Construct a plot of vapor pressure against the temperature. Fit a straight line model and plot it on the same plot. Comment on the fit. Also check the model using residual plots.

   (ii) According to the Clausius-Clapeyron equation,

   $$ln(\text{vapor pressure}) \propto -\frac{1}{\text{temperature}}$$

   Using this information, suggest an appropriate transformation and fit the new model.

4. The data set chemical_yield.csv contains the yield from a chemical process and the time.

   (i) Construct a plot of yield against the time. Fit a straight line model and plot it on the same plot. Comment on the fit. Also check the model using residual plots.

   (ii) Is there a need to transform the variables? If so, find an appropriate transformation. Support your choice with a model fitted and residual plots.

5. Show that
$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i-\hat{y}_i)^2.$$

6. If X is a continuous random variable with cumulative distribution function $F(x)$, show that $F(X)$ has the $uniform(0,1)$ distribution. Hence, generate 100 random variates from the standard normal distribution using the R functions `qnorm()` and `runif()` and construct a normal Q-Q plot.

7. Suppose $U_1, U_2, ..., U_n$ is a random sample from the $uniform(0,1)$ distribution. Let $U_{(1)}, U_{(2)}, ..., U_{(n)}$ be ordered from the smallest to the largest.

   (i) Show that $U_{(i)} \sim beta(i, n+1-i)$.

   (ii) Show that $E(U_{(i)}) = \dfrac{i}{n+1}$.