ST3131 Regression Analysis - Tutorial 5

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)

   (i) Fit the simple linear regression model and test $H_0 : \beta_1 = 0$. State your findings.

   $$y = \beta_0 + \beta_1 x_1 + \epsilon.$$

   (ii) Fit the multiple linear regression model and test $H_0 : \beta_2 = \beta_3 = 0$. State your findings.

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

   (iii) Construct a plot of mileage against displacement and comment on the relationship found.

   (iv) Fit the multiple linear regression model and test $H_0 : \beta_2 = \beta_3 = 0$. State your findings.

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon.$$

(v) Construct a plot of mileage against displacement and superimpose on it the straight line, quadratic and cubic curves. Which line seems to fit the data better?

(vi) Find the adjusted R^2 of the following models and comment.

$$y = \beta_0 + \beta_1 x_1 + \epsilon.$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon.$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon.$$

(vii) Fit the multiple linear regression model and test H_0: $\beta_6 = \beta_7$. State your findings.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_6 x_6 + \beta_7 x_7 + \epsilon.$$

(viii) Fit the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \epsilon.$$

(a) Construct 95% confidence intervals for the four regression coefficients.

(b) Construct a 95% confidence interval for the mean mileage of a car with a displacement of 300 cubic in.

(c) Construct a 95% prediction interval for the mileage of a car with a displacement of 300 cubic in.

2. Suppose we have the multiple linear regression model

$$\boldsymbol{y_{n\times 1}} = \boldsymbol{X_{n\times p}\beta_{p\times 1}} + \boldsymbol{\epsilon_{n\times 1}}$$

where $\boldsymbol{\beta'} = (\beta_0, \beta_1, ..., \beta_k)$, $p = k+1$, $\boldsymbol{E(\epsilon) = 0}$ and $\boldsymbol{Var(\epsilon)} = \sigma^2 \boldsymbol{I_n}$. Assume $\boldsymbol{\epsilon \sim N(0}$, $\sigma^2 \boldsymbol{I)}$. We consider the problem of testing

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0.$

Let $SS_T = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$, $SS_{Res} = \sum\limits_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $SS_R = \sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

(i) Show that $\boldsymbol{y'y} = SS_{Res} + \boldsymbol{\hat{\beta}'X'y}$

(ii) Show that $SS_T = SS_{Res} + SS_R$.

(iii) Show that $SS_{Res} = \boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{y}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

(iv) Show that $\boldsymbol{I} - \boldsymbol{H}$ is idempotent and hence find the rank of $\boldsymbol{I} - \boldsymbol{H}$.

(v) Under H_0, show that SS_{Res}/σ^2 follows the χ^2 distribution with $n - p$ degrees of freedom.

(vi) Show that $SS_R = \boldsymbol{y}'[\boldsymbol{H} - \boldsymbol{1}(\boldsymbol{1}'\boldsymbol{1})^{-1}\boldsymbol{1}']\boldsymbol{y}$ where $\boldsymbol{1}' = [1, 1, ..., 1]$.

(vii) Show that $\boldsymbol{H} - \boldsymbol{1}(\boldsymbol{1}'\boldsymbol{1})^{-1}\boldsymbol{1}'$ is idempotent and hence find the rank of $\boldsymbol{H} - \boldsymbol{1}(\boldsymbol{1}'\boldsymbol{1})^{-1}\boldsymbol{1}'$.

(viii) Under H_0, show that SS_R/σ^2 follows the χ^2 distribution with k degrees of freedom.

(ix) Show that SS_{Res} and SS_R are independent.

(x) Let $MS_{Res} = SS_{Res}/(n - p)$ and $MS_R = SS_R/k$. Under H_0, show that MS_R/MS_{Res} follows the F distribution with k and $n - p$ degrees of freedom.