ST3131 Regression Analysis - Tutorial 4

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)

   In this problem, we will consider the multiple linear regression model with five regressor variables

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i,\ i = 1, 2, ..., n, \\ &= \beta_0 + \sum_{j=1}^{5} \beta_j x_{ij} + \epsilon_i,\ i = 1, 2, ..., n. \end{aligned}$$

(i) What is n after observations containing missing values are removed?

(ii) The $\boldsymbol{X'X}$ matrix is an important matrix in multiple linear regression. From the $\boldsymbol{X'X}$ matrix, we can obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{\boldsymbol{-1}}\boldsymbol{X'y}$$

and

$$Var(\hat{\boldsymbol{\beta}}) = Var\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{bmatrix} = \sigma^2(\boldsymbol{X'X})^{-1} = \sigma^2\boldsymbol{C} = \sigma^2\begin{bmatrix} C_{00} & C_{01} & C_{02} & C_{03} & C_{04} & C_{05} \\ C_{10} & C_{11} & C_{12} & C_{13} & C_{14} & C_{15} \\ C_{20} & C_{21} & C_{22} & C_{23} & C_{24} & C_{25} \\ C_{30} & C_{31} & C_{32} & C_{33} & C_{34} & C_{35} \\ C_{40} & C_{41} & C_{42} & C_{43} & C_{44} & C_{45} \\ C_{50} & C_{51} & C_{52} & C_{53} & C_{54} & C_{55} \end{bmatrix}.$$

Use R to find $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{X'y}, \boldsymbol{X'X}, (\boldsymbol{X'X})^{\boldsymbol{-1}}, \hat{\boldsymbol{\beta}}$ and $SS_{Res} = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X'y}$. Hence find

(1) $\hat{\sigma}^2$ or MS_{Res},

(2) $Var(\hat{\beta}_2)$ and $\widehat{Var(\hat{\beta}_2)}$,

(3) $Cov(\hat{\beta}_2, \hat{\beta}_3)$ and $\widehat{Cov(\hat{\beta}_2, \hat{\beta}_3)}$.

(iii) Fit the model using the $lm()$ function and check that the values in parts (1) and (2) of (ii) are found correctly.

(iv) What is the R function for finding $\widehat{Var(\hat{\boldsymbol{\beta}})}$?

(v) Test the overall fit of the model by testing

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$

and state you conclusion.

(vi) Which regressor variable is significant after including the other four variables in the model?

(vii) Find the adjusted R^2 of this model and explain what it means. Is this a good model?

(viii) Find the correlation matrix of the variables: y, x_1, x_2, x_3, x_4 and x_5 and make a scatter matrix plot of these variables. Which regressor variables are highly correlated? Explain why they are highly correlated.

(ix) When two or more regressor variables are highly correlated, this is called multicollinearity. To understand the effect of multicollinearity, create two regressor variables that are nearly perfectly correlated:

```
x2 <- x1

x2[1] <- 351
```

and fit the same model again. Comment on the estimated regresssion coefficients and their variances, and explain why such shocking and disastrous results are obtained.

2. Consider the multiple linear regression model with k regressor variables:

   $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$

   (i) Show that $((\boldsymbol{X'X})^{-1})' = \boldsymbol{X'X}$.

   (ii) Show that $Var(\hat{\boldsymbol{y}}) = \sigma^2\boldsymbol{H}$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$

   (iii) Show that $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent.

   (iv) The rank of matrix $\boldsymbol{H}$ is the number of linearly independent rows or columns. Find the rank of the matrix $\boldsymbol{H}$.

   (v) Show that $0 \leq h_{ii} \leq 1$ for all i where h_{ii} is the ith diagonal element of $\boldsymbol{H}$.

   (vi) Suppose the $\boldsymbol{X}$ matrix is partitioned as $\boldsymbol{X} = [\boldsymbol{X_1} \quad \boldsymbol{X_2}]$. Show that (a) $\boldsymbol{HX_1} = \boldsymbol{X_1}$, (b) $\boldsymbol{HX_2} = \boldsymbol{X_2}$, (c) $\boldsymbol{X_1'H} = \boldsymbol{X_1'}$ and (d) $\boldsymbol{X_2'H} = \boldsymbol{X_2'}$.

   (vii) Show that $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ and $\boldsymbol{1'e} = 0$.

   (viii) Show that $\hat{\boldsymbol{y}}'\boldsymbol{e} = 0$.