**ST3131 Regression Analysis - Tutorial 6**

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)

   Using all the knowledge you have about multiple linear regression, find the "best" model. Explain why you think your model is the best?

2. The data set wine.csv contains 38 samples of wine collected from three different regions. The variables measured are: Quality ($y$), Flavor ($x_1$), Aroma ($x_2$), Body ($x_3$), Oakiness ($x_4$), Clarity ($x_5$), and Region ($x_6$). Do not use Region because this is a categorical variable and we will learn how to fit a model with categorical variables as regressors in a later chapter.

   Using all the knowledge you have about multiple linear regression, find the "best" model. Explain why you think your model is the best?

3. Let $\boldsymbol{a}$ be a $k \times 1$ vector of constants, and $\boldsymbol{y}$ be a $k \times 1$ random vector with non-singular variance-covariance matrix $Var(\boldsymbol{y}) = V$. Show that $Var(\boldsymbol{a'y}) = \boldsymbol{a'Va}$.

4. Suppose we have the multiple linear regression model

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

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 $\boldsymbol{\epsilon \sim N(0,}$ $\sigma^2 \boldsymbol{I})$.

The residual $e_i = y_i - \hat{y}_i, i = 1, 2, ..., n$. Let $\boldsymbol{e} = (e_1 \quad e_2 \quad ... \quad e_n)'$. Show that $Var(\boldsymbol{e}) = \sigma^2(\boldsymbol{I - H})$. Are the residuals $e_1, e_2, ..., e_n$ independent?

5. Give an example to show that two vectors are linearly independent but not orthogonal.

6. Suppose the multiple linear regression model is

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i, \ i = 1, 2, ..., n.$

Let $y, x_1, x_2, ..., x_k$ be scaled to unit length, that is

$$y_i^{\#} = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad w_{i1} = \frac{x_{i1} - \bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}}, ..., w_{ik} = \frac{x_{ik} - \bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}.$$

The model becomes

$y_i^{\#} = b_1 w_{i1} + b_2 w_{i2} + ... + b_k w_{ik} + \epsilon_i, \ i = 1, 2, ..., n.$

or

$$\boldsymbol{y^{\#} = Wb + \epsilon}$$

(i) Prove that the least-squares estimate of $b_0$ is zero.

(ii) Show that the off-diagonal elements of $\boldsymbol{W'W}$ are the correlation coefficients of the regressor variables:

$$\boldsymbol{W'W} = \begin{bmatrix} 1 & r_{12} & r_{13} & ... & r_{1k} \\ r_{21} & 1 & r_{23} & ... & r_{2k} \\ r_{31} & r_{32} & 1 & ... & r_{3k} \\ . & . & & & . \\ . & . & & & . \\ . & . & & & . \\ r_{k1} & r_{k2} & r_{k3} & ... & 1 \end{bmatrix}$$