ST3131 Regression Analysis - Tutorial 8

(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) Construct a plot of y against x_1 and the fitted regression line of the the model $y = \beta_0 + \beta_1 x_1 + \epsilon$. Also, construct the studentized residual plot agianst $\hat{y}$ and x_1, and comment on the residual plots obtained. State the MS_{Res} of this model.

(ii) Consider the model $y^{(\lambda)} = \beta_0 + \beta_1 x_1 + \epsilon$. Use R to find the MS_{Res} for $\lambda = -2.00, -1.99$, ..., 1.99, 2.00 and make a plot of MS_{Res} against λ. Hence determine an approximate value of λ that minimizes the MSRes.

Note: For the MS_{Res} to be comparable for different $y^{(\lambda)}$, the following transformation must be used:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda\, \tilde{y}^{\lambda-1}}, & \lambda \neq 0, \\ \tilde{y} \log(y), & \lambda = 0. \end{cases}$$

where $\tilde{y}$ is the geometric mean defined as $\tilde{y} = e^{[\frac{1}{n}\sum_{i=1}^{n} \log(y_i)]}$.

(iii) Once the λ is obtained in part (ii), then either one of the transformations can be used:

$$y^{(\lambda)} = \begin{cases} y^{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0. \end{cases}$$

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda\, \tilde{y}^{\lambda - 1}}, & \lambda \neq 0, \\ \tilde{y} \log(y), & \lambda = 0. \end{cases}$$

In practice, the first one is more commonly used because of its simpler expression.

Using the λ obtained in part (ii), construct a plot of y^{λ} against x_1 and the fitted regression line of the model $y^{\lambda} = \beta_0 + \beta_1 x_1 + \epsilon$. Also, construct the studentized residual plot agianst $\hat{y}$ and x_1, and comment on the residual plots obtained. State the MS_{Res} of this model.

(2) The Box-Cox transformation is given as

$$y^{(\lambda)} = \begin{cases} y^{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0. \end{cases}$$

It is also often written as

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \log(y), & \lambda = 0. \end{cases}$$

Show that $\lim\limits_{\lambda \to 0} \dfrac{y^{\lambda} - 1}{\lambda} = \log(y)$.

(3) 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.
Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$ and construct 95% joint Bonferroni confidence region for β_1, β_2 and β_4.

(4) Consider the model

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

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$. Assume that $\boldsymbol{V}$ is known but not σ^2. Show that

$$(\boldsymbol{y'V^{-1}y - y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y})/(n-p)$$

is an unbiased estimate of σ^2.

(5) 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 $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \epsilon$. Find the Cook's distance of every observation and identify all influential observations using the Cook's distance.