ST3131 Regression Analysis - Tutorial 11

(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.

$y = mileage$
$x_1 = displacement$
$x_2 = horsepower$
$x_3 = torque$
$x_4 = compression\_ratio$
$x_5 = real\_axel\_ratio$
$x_6 = length$
$x_7 = width$
$x_8 = weight$

(i) Scale the response and all the regressor variables to unit length. Fit the multiple linear regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon$$

Is the model significant?

(ii) Construct a matrix scatter plot of the regressor variables and find their correlation coefficients. Is there any evidence of multicollinearity?

(iii) Find the condition number and condition indices of the $\boldsymbol{X'X}$ matrix. Is there any evidence of multicollinearity?

(iv) Find all the variance inflation factors. Is there any evidence of multicollinearity?

(v) For the model in part (i), construct a ridge trace for $k = 0.01, 0.02, ..., 0.50$. Hence, determine an appropriate value of k for which the regression coefficients have stabilized. Find the ridge estimates of the regression coefficients based on this value of k and their standard errors. Compare them with those obtained using the ordinary multiple linear regression model.

(vi) For the model in part (i), display all the principal components. Perform the principal-component regression by deleting the two principal components with the smallest eigenvalues. State the regression coefficients obtained.

(2) A multiple linear regression model for the acetylene data (Acetylene.csv) is given as

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2 + \epsilon$

where $y = Conversion$, $x_1 = Temperature$, $x_2 = Hratio$ and $x_3 = Contact$.

The response is the percentage of conversion of n-heptane to acetylene and it is non-negative.

(i) Unit length scale the response and all the 9 regressor variables.

(ii) Fit an ordinary least-squares (OLD) regression model and use it to make prediction of the regressor values shown in Table 1. The last four rows of predicted values are based on extrapolation.

(iii) Use ridge regression with k = 0.032 to find estimates of the regression coefficients. Use the model to make prediction of the regressor values shown in Table 1. The last four rows of predicted values are based on extrapolation.

(iv) Use principal component regression to obtain estimates of the regression coefficients by deleting three principal components associated with the three smallest eigenvalues. Use the model to make prediction of the regressor values shown in Table 1. The last four rows of predicted values are based on extrapolation.

(v) Do the three models produce consistent predictions?

(vi) Which is your preferred model and why?

Table 1: Acetylene Data

Conversion	Temperature	HRatio	Contact	Principal Component	Ridge	OLS
49	1300	7.5	0.012			
50.2	1300	9	0.012			
50.5	1300	11	0.0115			
48.5	1300	13.5	0.013			
47.5	1300	17	0.0135			
44.5	1300	23	0.012			
28	1200	5.3	0.040			
31.5	1200	7.5	0.038			
34.5	1200	11	0.032			
35	1200	13.5	0.026			
38	1200	17	0.034			
38.5	1200	23	0.041			
15	1100	5.3	0.084			
17	1100	7.5	0.098			
20.5	1100	11	0.092			
29.5	1100	17	0.086			
	1200	15	0.095			
	1300	6	0.095			
	1100	12	0.012			
	1200	5	0.012			

(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). Use the following procedures to determine the sets of regressor variables. Use $\alpha_{IN} = 0.1$ and $\alpha_{OUT} = 0.1$.

(i) forward selection,

(ii) backward elimination,

(iii) stepwise regression.

(4) The data set data-table-B3.csv contains 30 samples of car. The variables measured are: mileage (y), displacement (x_1), horsepower (x_2), torque (x_3), compression ratio (x_4), real axel ratio (x_5), length (x_6), width (x_7), and weight (x_8). Use the following procedures to determine the sets of regressor variables. Use $\alpha_{IN} = 0.2$ and $\alpha_{OUT} = 0.2$.

(i) forward selection,

(ii) backward elimination,

(iii) stepwise regression.