ST3131 Regression Analysis - Tutorial 9

(1) The data data-table-B13.csv contains data on the thrust of a jet engine.

$y =$ thrust
$x_1 =$ primary speed of rotation
$x_2 =$ secondary speed of rotation
$x_3 =$ fuel flow rate
$x_4 =$ pressure
$x_5 =$ exhaust temperature
$x_6 =$ ambient temperature at time of test

Fit a regression model to the data using all the regressors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Identify all influential observations with respect to

(i) $\boldsymbol{\beta}$

(ii) $\beta_i, i = 0, 1, 2, ..., 6$

(iii) $\hat{y}_i$

(iv) $\hat{\sigma}^2$

(2) A data set (mercedes_car_prices.csv) was collected to study the car prices of used Mercedes-Benz S Class. Twenty eight cars were sampled. Let y be the price of a car in thousands of dollars and x be the age of the car in years. Cars that are less than five years old are of the latest models. Cars that are more than five years old are of the old models.

(i) Construct a plot of y against x. Comment on the relationship between y and x.

(ii) Fit the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$ and display it on a plot of y against x. Comment on the fitted model.

(iii) (a) Fit a quadratic spline model with one knot at $x = 5$, and continuous at the knot.

(b) Fit a quadratic spline model with one knot at $x = 5$, and with no continuity restriction.

(c) Display both models on a plot of y against x. Which model is preferred?

(iv) (a) Fit a linear spline model with one knot at $x = 5$, and continuous at the knot.

(b) Fit a linear spline model with one knot at $x = 5$, and with no continuity restriction.

(c) Display both models on a plot of y against x. Which model is preferred?

(v) Fit a nonparameteric regression model and display it on a plot of y against x.

(vi) Fit a loess model and display it on a plot of y against x.

(vii) Compare the models in parts (ii), (iii), (iv), (v) and (vi). Which model do you prefer?

(3) The data set carbonation.csv contains 12 samples of carbonation level (y), temperature (x_1), and pressure (x_2) of a soft drink. The objective of this experiment is to find a setting of temperature and pressure that yields the highest carbonation level.

(i) (a) Construct a 3D plot of carbonation level against the temperature and pressure.

(b) Construct a matrix scatter plot of carbonation level, temperature and pressure.

(c) Comment on any relationship found.

(ii) Fit a second order polynomial model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$

and test for significance of the model.

(iii) Test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ and state your conclusion.

(iv) Is there any interaction between pressure and temperature in affecting the carbonation level?

(v) Explain what LOF is with respect to the second order model. Test for LOF and state your conclusion.

(vi) Construct a contour plot of the second order polynomial model.

(vii) Construct a response surface of the second order polynomial model.

(viii) From the contour plot, what setting of temperature and pressure yielded the highest level of carbonation?

(4) Consider the nonparametric kernel regression model with bandwidth b and the triangle kernel function

$$K(t) = \begin{cases} 1 - \dfrac{|t|}{c}, & |t| \le c, \\ 0, & |t| > c. \end{cases}$$

Determine the relationship between b and c.

(5) Consider the nonparametric kernel regression model with bandwidth b and kernel functions

Box kernel 1:

$$K_1(t) = \begin{cases} 1, & |t| \le 0.5, \\ 0, & |t| > 0.5. \end{cases}$$

Box kernel 2:

$$K_2(t) = \begin{cases} 2, & |t| \le 0.5, \\ 0, & |t| > 0.5. \end{cases}$$

A nonparametric regression model is fitted with Box kernel 1 and another model is fitted with Box kernel 2. Will the same model be obtained? Is it a must that $\int_{-\infty}^{\infty} K(t)\, dt = 1$?