ST3131 Regression Analysis - Tutorial 2

1. The data set data-table-B3.csv contains data on the gasoline mileage performance of different automobiles. Write R codes to read in the data set and answer the following questions.

   (i) Verify that $\sum_{i=1}^{n} x_i e_i = 0$ and $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ for this data set.

   (ii) Construct a plot of y against x and show the residuals on the plot. Is there any evidence that σ^2 is not constant against x?

   (iii) Calculate SS_T, SS_{Res}, SS_R and $\hat{\sigma}^2$ or MS_{Res}.

   (iv) Construct a 99% confidence interval for β_1. Give an interpretation of the confidence interval.

   (v) Construct a 99% confidence interval for β_0. Give an interpretation of the confidence interval.

   (vi) Construct a 99% confidence interval for σ^2. Give an interpretation of the confidence interval.

   (vii) Construct a 99% confidence interval for the mean mileage at $x = 2000$ cc. Give an interpretation of the confidence interval.

   (viii) Calculate the R^2 and give an interpretation.

   (ix) Calculate the correlation coefficient r between *displacement* and *mileage* and give an interpretation. Explain why the correlation coefficient is appropriate for this problem. What is the relationship between r and R^2?

   (x) Construct a 99% confidence interval for the population correlation coefficient ρ. Give an interpretation of the confidence interval.

2. For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n,$$

   the least-squares estimate $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

   (i) Show that $\hat{\beta}_1 = \sum_{i=1}^{n} c_i\, y_i$, where $c_i = \dfrac{x_i - \bar{x}}{S_{xx}}, i = 1, 2, ..., n$.

   (ii) Show that $\sum_{i=1}^{n} c_i = 0$.

   (iii) Show that $\sum_{i=1}^{n} c_i\, x_i = 1$.

3. For the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n.$$

The residual $e_i = y_i - \hat{y}_i, i = 1, 2, ..., n.$

(i) Show that $\sum_{i=1}^{n} x_i\, e_i = 0.$

(ii) Show that $\sum_{i=1}^{n} \hat{y}_i\, e_i = 0.$

4. **Distribution theory**

The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2,\ i = 1, 2, ..., n$ and ϵ_i's are independent.

Assume further that ϵ_i's are normally distributed.

The following results are true:

(i) $\hat{\beta}_1$ follows the normal distribution with mean $E(\hat{\beta}_1)$ and variance $Var(\hat{\beta}_1)$.

(ii) $\hat{\beta}_0$ follows the normal distribution with mean $E(\hat{\beta}_0)$ and variance $Var(\hat{\beta}_0)$.

(iii) $\dfrac{(n-2)MS_{Res}}{\sigma^2}$ follows the χ^2 distribution with $n-2$ degrees of freedom where $E(MS_{Res}) = \sigma^2$ [$\because$ linear model theory]

(iv) $\hat{\beta}_1$ and MS_{Res} are independent. [$\because$ linear model theory]

(v) $\hat{\beta}_0$ and MS_{Res} are independent. [$\because$ linear model theory]

(vi) $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}}$ follows the t distribution with degrees of freedom $n - 2$.

(vii) $\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}}$ follows the t distribution with degrees of freedom $n - 2$.

Prove (i), (ii), (vi) and (vii). Use result (vi) to derive a 95% confidence interval for β_1.

5. (a) Let $x_1, x_2, ..., x_n$ be a random sample from the normal distribution with mean μ and variance σ^2. State the distribution of $\dfrac{(n-1)S^2}{\sigma^2}$ where S^2 is the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

(b) The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, 2, ..., n,$$

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2,\ i = 1, 2, ..., n$ and ϵ_i's are independent.

Assume further that ϵ_i's are normally distributed.
$\dfrac{(n-2)MS_{Res}}{\sigma^2}$ follows the χ^2 distribution with $n-2$ degrees of freedom according to linear model theory. Explain why the degrees of freedom is $n - 2$?