ST3131 Regression Analysis - Tutorial 3

1. 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 and normally distributed.

   (i) [Decomposition of variance]
   Show that $SS_T = SS_R + SS_{Res}$ where
   $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$
   $SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
   $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$.

   (ii) Show that $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$

   (iii) Show that $E(MS_{Res}) = \sigma^2$

   (iv) According to linear model theory, $(n-2)MS_{Res}/\sigma^2$ follows a chi-square distribution with degrees of freedom $n-2$. Show that a $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\frac{(n-2)MS_{Res}}{\chi^2_{\alpha/2,n-2}} \le \sigma^2 \le \frac{(n-2)MS_{Res}}{\chi^2_{1-\alpha/2,n-2}}.$$

2. Consider the simple linear regression model through the origin

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

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 and normally distributed.

   (i) Show that the least-squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(ii) Find $E(\hat{\beta}_1)$.

(iii) Find $Var(\hat{\beta}_1)$.

(iv) Derive the probability density function of $\hat{\beta}_1$.

(v) [Decomposition of variance] Show that

$$\sum_{i=1}^{n} y_i^2 = SS_{Res} + SS_R$$

where $SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $SS_R = \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \hat{y}_i^2$.

(vi) Find $E(SS_R)$ and $E(SS_{Res})$.

(vii) Based on the results in part (vi), suggest a test statistic for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

3. Suppose there is a response variable y and two regressor variables x_1 and x_2. Further suppose the true model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$, and ϵ's are independent. If the simple linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ is fitted instead, show that $\hat{\beta}_1$ is a biased estimator of β_1.

4. The data set data-table-B3.csv contains data on the gasoline mileage performance of different automobiles. The second variable *displacement* in the data set is the capacity of an engine (in cubic in.) and the third variable *horsepower* is the horsepower in hp. We are interested to study the relationship between *horsepower* and *displacement*. [do not do any conversion of units here]

(i) Make a plot of *horsepower* against *displacement*. Comment on any relationship found.

(ii) Fit a simple linear regression model

$$horsepower = \beta_0 + \beta_1 \, displacement + \epsilon$$

and plot the least-squares line on the plot in part (i). Comment on the fit of this line. Construct 95% confidence intervals for β_0 and β_1 and comment.

(iii) It is not unreasonable to assume that $horsepower$ is zero if the $displacement$ is zero. Fit a simple linear regression model passing through the origin

$$horsepower = \beta_1 \, displacement + \epsilon$$

and plot the least-squares line on the plot in part (i). Comment on the fit of this line. Construct 95% confidence intervals for β_1 and comment.

(iv) Which model do you prefer? Explain.

5. The data set data-table-B3.csv contains data on the gasoline mileage performance of different automobiles. The second variable *displacement* in the data set is the capacity of an engine (in cubic in.) and the third variable *horsepower* is the horsepower in hp. [do not do any conversion of units here]

   (i) Construct a scatter matrix plot for $mileage, displacement, horsepower$, and calculate the correlation coefficient of $displacement$ and $horsepower$. Comment.

   (ii) Fit a simple linear regression model

   $$y = \beta_0 + \beta_1 \; displacement + \epsilon$$

   Is β_1 significantly different from zero?

   (iii) Fit a simple linear regression model

   $$y = \beta_0 + \beta_2 \; horsepower + \epsilon$$

   Is β_2 significantly different from zero?

   (iv) Fit a multiple linear regression model

   $$y = \beta_0 + \beta_1 \; displacement + \beta_2 \; horsepower + \epsilon.$$

   Test $H_0 : \beta_1 = \beta_2 = 0$ and state your conclusion.

   (v) Fit a multiple linear regression model

   $$y = \beta_0 + \beta_1 \; displacement + \beta_2 \; horsepower + \epsilon.$$

   Use t tests to assess the contribution of each regressor variable to the model. Discuss your findings.