STAT2008/STAT2014/STAT6014

Tutorial 2

Question 1. The data file **Lubricant.csv** (available on Wattle) contains 53 measurements of the viscosity of a particular lubricating agent at various temperatures and pressures. The names of the three variables in the data are **viscos, pressure** and **tempC**. At the end of this question, remember to save the related R code as we will be using it again in next tutorial.

(a) Use **lm()** to perform a simple linear regression with viscosity as the response and pressure as the predictor variable. What are the least-squares estimates of the slope and intercept?

(b) Plot viscosity against pressure and use **abline()** to superimpose the estimated regression line. Use the estimated coefficients of the regression line to predict what the viscosity of the lubricant would be at a pressure of 1,000? Also predict what the viscosity of the lubricant would be at a pressure of 10,000? Locate these predictions on your plot and comment on whether or not they appear to be sensible predictions.

(c) Use R to find the means of both pressure and viscosity and check that together the two means form a point (called the centroid of the data) which is located on the estimated regression line.

Solution:

I have created an R commands file associated with this question in "Tutorial2.R" (available on Wattle). This includes all the R codes you will need to answer the questions along with extensive comments, which include the answers to the questions. To follow these solutions, you will need to download a copy of this file from Wattle and run the code (preferably line by line), so that you can see the R output and then read the associated comments.

Question 2. Show the following equations:

(a) $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 0$.

(b) $E(b_1) = \beta_1$ and $Var(b_1) = \frac{\sigma^2}{S_{xx}}$.

(c) $E(b_0) = \beta_0$ and $Var(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]$.

Solution:

(a) This result is a key part to show the partition of variation ($SSTO = SSR+SSE$).

$$
\begin{aligned}
& \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \\
= & \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})e_i \\
= & \sum_{i=1}^{n}\hat{Y}_i e_i - \bar{Y}\sum_{i=1}^{n} e_i \quad \text{by the 1st and 5th properties of fitted regression line} \\
= & \ 0.
\end{aligned}
$$

(b) In Week 1's lecture, it has been shown that

$$b_1 = \sum_{i=1}^{n} k_i Y_i$$

where $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{X_i - \bar{X}}{S_{xx}}$ and these constants k_i have the following properties

$$\sum_{i=1}^{n} k_i = 0, \quad \sum_{i=1}^{n} k_i X_i = 1, \quad \sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{1}{S_{xx}}.$$

Then,

$$
\begin{aligned}
E(b_1) &= E\left(\sum_{i=1}^{n} k_i Y_i\right) = \sum_{i=1}^{n} k_i E(Y_i) \\
&= \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i X_i \\
&= 0 + \beta_1 \times 1 \\
&= \beta_1.
\end{aligned}
$$

As responses Y_i are uncorrelated,

$$\begin{aligned} Var(b_1) &= Var\left(\sum_{i=1}^n k_i Y_i\right) = \sum_{i=1}^n k_i^2 Var(Y_i) \\ &= \sum_{i=1}^n k_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n k_i^2 \\ &= \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

(c) As we know $b_0 = \bar{Y} - b_1 \bar{X}$, we have

$$\begin{aligned} E(b_0) &= \frac{1}{n}\sum_{i=1}^n E(Y_i) - \bar{X}E(b_1) \\ &= \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 X_i + \varepsilon_i) - \beta_1 \bar{X} \\ &= \beta_0 + \beta_1 \bar{X} + \frac{1}{n}\sum_{i=1}^n E(\varepsilon_i) - \beta_1 \bar{X} \\ &= \beta_0 + (\beta_1 \bar{X} - \beta_1 \bar{X}) + 0 \\ &= \beta_0. \end{aligned}$$

We can rewrite b_0 as

$$b_0 = \frac{1}{n}\sum_{i=1}^n Y_i - \sum_{i=1}^n k_i \bar{X} Y_i = \sum_{i=1}^n c_i Y_i$$

where $c_i = \frac{1}{n} - k_i \bar{X}$. Then,

$$\begin{aligned} \sum_{i=1}^n c_i^2 &= \sum_{i=1}^n \left(\frac{1}{n^2} - 2\bar{X}k_i + \bar{X}^2 k_i^2\right) \\ &= \frac{1}{n} - 2\bar{X}\sum_{i=1}^n k_i + \bar{X}^2 \sum_{i=1}^n k_i^2 \\ &= \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}. \end{aligned}$$

As responses Y_i are uncorrelated,

$$\begin{aligned} Var(b_0) &= Var\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i^2 Var(Y_i) \\ &= \sum_{i=1}^n c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n c_i^2 \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]. \end{aligned}$$