Regression Modelling

(STAT2008/STAT4038/STAT6038)

Solutions to Tutorial 1 – Simple Linear Regression

Questions One and Two

For the questions involving **R** in this tutorial, I have created an **R** commands file called **Tutorial1.R** (available on Wattle). This includes all the **R** code 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 Three (optional extra – there will be no question like this one on the final exam)

Recall that any matrix, A, is called a projection if it satisfies the identities: $A^{\mathrm{T}} = A$ and $A^2 = A$ (see page 9 of the lecture notes). Also, recall that the hat matrix is defined as $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

(a) Show that the hat matrix, H, and the matrix $I - H$ are projections.

SOLUTION: Since $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, we have

$H^T = \{X(X^TX)^{-1}X^T\}^T = (X^T)^T\{(X^TX)^{-1}\}^TX^T = X\{(X^TX)^T\}^{-1}X^T = X(X^TX)^{-1}X^T = H$

$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$

$(I - H)^T = I^T - H^T = I - H$

$(I - H)^2 = (I - H)(I - H) = II - HI - IH + HH = I - H - H + H = I - H$

(b) The data frame **protpreg** (from the Protein in Pregnancy example in lectures) has two columns, the first relating to the amount of a certain protein and the second to the number of weeks of gestation for a group of pregnant women. Create a vector named **gest** containing the gestation data. Now, create the design matrix, X by attaching an initial column of ones to gest using the command:
X <− cbind(rep(1, length(gest)), gest)

(c) Now, use **R**'s matrix multiplication capabilities to construct the hat matrix H. Multiply H by itself and construct its transpose to check that it is indeed a projection. Also, use matrix calculations to find the least-squares regression estimate $b = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$, where Y is the vector of response values in the first column of **protpreg**. Check the results against the values given for this problem in the example discussed in lectures.

SOLUTION: Parts b and c require more **R** coding, so I have included some appropriate code in the file **Tutorial1.R**

Question Four (optional extra – there will be no question like this one on the final exam)

Recall (from page 10 of the lecture notes) the breakdown of the total sum of squares SS_{Total}, into the sum of $SS_{Regression}$ and SS_{Error}:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

(a) In demonstrating this breakdown, we used the fact that: $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) = 0$. Show that this fact is indeed true.
[Hint: Recall that the residuals, $e_i = Y_i - \hat{Y}_i$, from a least-squares regression satisfy: $\sum_{i=1}^{n} e_i = 0$; $\sum_{i=1}^{n} x_i e_i = 0$.]

SOLUTION: To show this fact, we note that:

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \bar{Y}\right) &= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} e_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n} e_i \\ &= \sum_{i=1}^{n} e_i\hat{Y}_i = \sum_{i=1}^{n} e_i(b_0 + b_1 x_i) = b_0\sum_{i=1}^{n} e_i + b_1\sum_{i=1}^{n} x_i e_i = 0\end{aligned}$$

Since $\sum_{i=1}^{n} e_i$ and $\sum_{i=1}^{n} x_i e_i$ are both zero as noted in the hint.

(b) Recall that the coefficient of determination, R^2, and the sample correlation coefficient, r, are defined as:

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = \frac{1}{SS_{Total}}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot SS_{Total}}}$$

Show that $R^2 = r^2$.

SOLUTION: From the definition of R^2, we have:

$$\begin{aligned}R^2 &= \frac{1}{SS_{Total}}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \frac{1}{SS_{Total}}\sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{Y})^2 \\ &= \frac{1}{SS_{Total}}\sum_{i=1}^{n}(\bar{Y} - b_1\bar{x} + b_1 x_i - \bar{Y})^2 = \frac{1}{SS_{Total}}\sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2 \\ &= \frac{1}{SS_{Total}} b_1^2 S_{xx} = \frac{S_{xx}}{SS_{Total}}\left(\frac{S_{xy}}{S_{xx}}\right)^2 = \frac{S_{xy}^2}{S_{xx} \cdot SS_{Total}} = \left(\frac{S_{xy}}{\sqrt{S_{xx} \cdot SS_{Total}}}\right)^2 = r^2\end{aligned}$$

Question Five (optional extra – there will be no question like this one on the final exam)

Refer to page 12 of the lecture notes. Show that the expectation of the mean square for the regression is:

$$E\left(MS_{Regression}\right) = E\left\{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2\right\} = \sigma^2 + \beta_1^2 S_{xx}$$

[Hint: Recall that $\hat{Y} = b_0 + b_1 x_i$; $b_0 = \bar{Y} - b_1\bar{x}$; and for any random variable Z, $E(Z^2) = Var(Z) + \{E(Z)\}^2$.]

SOLUTION: Using the hint, we have:

$$\begin{aligned} E(MSR) &= E\left\{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2\right\} = E\left\{\sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{Y})^2\right\} = E\left\{\sum_{i=1}^{n}(\bar{Y} - b_1\bar{x} + b_1 x_i - \bar{Y})^2\right\} \\ &= E\left\{\sum_{i=1}^{n}(\bar{Y} - b_1\bar{x} + b_1 x_i - \bar{Y})^2\right\} = E\left\{b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} = S_{xx}E(b_1^2) \\ &= S_{xx}[Var(b_1) + \{E(b_1)\}^2] = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx} \end{aligned}$$