QUEEN MARY UNIVERSITY OF LONDON

MTH5120 Statistical Modelling I

Solution to Exercise Sheet 5

1. A chemist studied the concentration of a solution (Y) over time (x). Fifteen identical solutions were prepared. The solutions were randomly divided into five sets of three, and the five sets were measured, respectively after 1, 3, 5, 7, and 9 hours.

   Without making any plots the chemist entered the data into R, fitted a simple linear regression model and then carried out a goodness of fit test. The following is the Analysis of Variance table she produced but with some figures missing.

```
Analysis of Variance Table

Response: y
              Df  Sum Sq Mean Sq F value
x              1 12.5971
Residuals     13
  Lack of fit     2.770
  Pure error
Total         14 15.5218
```

   (a) In order to complete the Analysis of Variance Table, we need to compute different elements. First of all, we need to compute the SS_E, which is the difference between SS_T and SS_R. Thus

$$SS_E = SS_T - SS_R = 15.5218 - 12.5971 = 2.9247$$

   Then we can compute the MS_R and MS_E, which are

$$MS_R = \frac{SS_R}{1} = 12.5971 \qquad MS_E = \frac{SS_E}{n-2} = \frac{2.9274}{13} = 0.2251846$$

   Then F-value is equal to the ratio between MS_R and MS_E previously computed

$$F = \frac{MS_R}{MS_E} = \frac{12.5971}{0.2251846} = 55.94121$$

   Moving to the lack of fit and pure error part, we firstly compute the SS_{PE} since we know all the other elements:

$$SS_{PE} = SS_E - SS_{LoF} = 2.9247 - 2.770 = 0.1547$$

   In our case, $m = 5$ is the number of measured and then $d.f.$ of the lack of fit is $m - 2 = 5 - 2 = 3$, while the $d.f.$ of the pure error is $n - m = 15 - 5 = 10$. Moving to the Mean square, we have:

$$MS_{LoF} = \frac{SS_{LoF}}{m-2} = \frac{2.770}{3} = 0.9233333 \qquad MS_{PE} = \frac{SS_{PE}}{n-m} = \frac{0.1547}{10} = 0.01547$$

Finally, the F of residuals is

$$F = \frac{MS_{LoF}}{MS_{PE}} = \frac{0.923333}{0.01547} = 59.68539$$

Thus we have completed the table.

(b) Firstly we have $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ and F follows a F^1_{13} if H_0 is true. The observed F is equal to 55.94, while the p-value is given by

$$1 - pf(55.94, 1, 13) = 4.635011e - 06$$

So overwhelming evidence against H_0.

The second possible F test is H_0 model fits well versus it does not fit well. In this case, F is distributed as F^3_{10} if H_0 is true. The observed value of F is equal to 59.68. We compute the p-value given by

$$1 - pf(59.68, 3, 10) = 1.096306e - 06$$

So overwhelming evidence against H_0.

2. Write the following models in the form of a general linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Hence find the least squares estimators of the parameters.

(a) The model with just a constant (p=1)

$$y_i = \beta_0 + \varepsilon_i \quad i = 1, 2, \ldots, n.$$

can be written as a GLM with

$$\boldsymbol{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \beta_0$$

Thus we have

$$\boldsymbol{X}^t \boldsymbol{y} = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} y_i$$

On the other hand, $\boldsymbol{X}^t \boldsymbol{X} = n$. Hence, the inverse of $(\boldsymbol{X}^t \boldsymbol{X})^{-1} = n^{-1}$. Then we have that the least square estimate of β_0 is

$$\widehat{\beta}_0 = \frac{\sum y_i}{n} = \bar{y}$$

The variance of $\widehat{\beta}_0$ is $\sigma^2 (\boldsymbol{X}^t \boldsymbol{X})^{-1}$ so σ^2 / n.

(b) The linear regression model through the origin (p=1):

$$y_i = \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, n.$$

can be written as a GLM model with

$$\boldsymbol{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \beta_1$$

Thus we have

$$\boldsymbol{X}^t \boldsymbol{y} = \begin{pmatrix} x_1 & x_2 & \ldots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^{n} x_i y_i$$

On the other hand, $\boldsymbol{X}^t \boldsymbol{X} = \sum_{i=1}^n x_i^2$. Hence, the inverse of $(\boldsymbol{X}^t \boldsymbol{X})^{-1} = (\sum_i x_i^2)^{-1}$. Then we have that the least square estimate of β_1 is

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum_i x_i^2}$$

The variance of $\widehat{\beta}_1$ is $\sigma^2 (\boldsymbol{X}^t \boldsymbol{X})^{-1}$ so $\sigma^2 / (\sum_i x_i^2)$.

3. We have

$$\operatorname{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \qquad \operatorname{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{S_{xx}} \qquad \operatorname{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Hence

$$\begin{aligned} \operatorname{Var}((\widehat{\beta}_0 + \widehat{\beta}_1 x_0) &= \operatorname{Var}((\widehat{\beta}_0) + x_0^2 \operatorname{Var}((\widehat{\beta}_1) + 2x_0 \operatorname{cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x_0^2}{S_{xx}} - 2x_0 \frac{\bar{x}}{S_{xx}} \right) = \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$