University of Technology Sydney
School of Mathematical and Physical Sciences

Mathematical Statistics (37262) – Tutorial 7 SOLUTIONS

1.

i) $\bar{x} = \frac{-5-4-...+4+5}{11} = 0$ and $\bar{y} = \frac{-1.6+12.3+...-0.9}{11} = \frac{288.1}{11}$

ii) $\sum_{i=1}^{n} \varepsilon_i = 0$ hence $\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$.

This gives $\sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} x_i = 0$ so

$\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\alpha} - \hat{\beta} \frac{\sum_{i=1}^{n} x_i}{n} = 0$ which gives $\hat{\alpha} = \bar{y} - \beta\bar{x}$.

iii) We minimise $\sum_{i=1}^{n} \varepsilon_i{}^2$ by setting the derivative with respect to β and setting the resulting derivative to zero.

$$\frac{\partial}{\partial\beta} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \beta x_i)^2 = \frac{\partial}{\partial\beta} \sum_{i=1}^{n} (y_i - (\bar{y} - \beta\bar{x}) - \beta x_i) = 0$$

$$= \frac{\partial}{\partial\beta} \sum_{i=1}^{n} \left((y_i - \bar{y})^2 - 2\beta(x_i - \bar{x})(y_i - \bar{y}) + \beta^2 (x_i - \bar{x})^2 \right).$$

$$\frac{\partial}{\partial\beta} \sum_{i=1}^{n} \left((y_i - \bar{y})^2 - 2\beta(x_i - \bar{x})(y_i - \bar{y}) + \beta^2 (x_i - \bar{x})^2 \right)$$

$$= \sum_{i=1}^{n} \left(-2(x_i - \bar{x})(y_i - \bar{y}) + 2\beta(x_i - \bar{x})^2 \right) = 0$$

hence $\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$.

iv) Here, we obtain

$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (-5)^2 + (-4)^2 + ... + (4)^2 + (5)^2 = 110$ and

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = (-5)\left(-1.6 - \frac{288.1}{11}\right) + ... + (5)\left(-0.9 - \frac{288.1}{11}\right) = -3.3$$

Together, these give $\hat{\beta} = \frac{-3.3}{110} = -0.03$ and $\hat{\alpha} = \bar{y} - \beta\bar{x} = \frac{288.1}{11}$ hence $y_i = \frac{288.1}{11} - 0.03x_i + \varepsilon_i$.

v) $SSE = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y})^2 \approx 3137.49$ and

$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \approx 3137.589$.

This then gives $R^2 = \frac{SSR}{SST} = \frac{3137.589 - 3137.49}{3137.49} \approx 0.00003$.

This model fits very poorly, explaining only around 0.003% of the variation in the response variable.

2.

i) The trend is clearly non-linear. We can see this, for example, in the fact that the difference between the responses for $x = 10$ and $x = 8$ is much smaller than the difference between the responses for $x = 2$ and $x = 1$. The differences are greater for smaller values of *x*.

ii) Set $\frac{1}{x} = z$. We also calculate $\bar{y} = 5.0875$ and $\bar{z} = 0.3625$.

y	4.3	4.375	4.6	4.75	5.5	7
x	10	8	5	4	2	1
z	0.1	0.125	0.2	0.25	0.5	1
$(z_i - \bar{z})$	-0.2625	-0.2375	-0.1625	-0.1125	0.1375	0.6375
$(y_i - \bar{y})$	-0.7875	-0.7125	-0.4875	-0.3375	0.4125	1.9125

These give $\hat{\beta} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} = 3$ and hence

$\hat{\alpha} = \bar{y} - \beta\bar{x} = 5.0875 - 3(0.3625) = 4$. The model is then $y_i = 4 + 3\left(\frac{1}{x_i}\right) + \varepsilon_i$

iii) In this case, we can see that all points lie exactly on this line with no residual error. This would give an R-squared value of 1.

3.

i) There are multiple ways to do this. If we take Class A as the reference category, we rewrite $y_i = \alpha + \beta_A x_{Ai} + \beta_B x_{Bi} + \beta_C x_{Ci} + \varepsilon_i$ as
$y_i = (\alpha + \beta_A) + (\beta_B - \beta_A)x_{Bi} + (\beta_C - \beta_A)x_{Ci} + \varepsilon_i$.

ii) Now, the estimate of the intercept term $(\alpha + \beta_A)$ must just be the mean response for the reference category – here Class A.

This is $\frac{98 + 76 + ... + 55}{8} = 78.125$.

The other classes have mean responses 75.75 (Class B) and 64.375 (Class C).

This gives slope estimates for $(\beta_B - \beta_A)$ of $75.75 - 78.125 = -2.375$ and for $(\beta_C - \beta_A)$ of $64.375 - 78.125 = -13.75$.

The model is then $y_i = 78.125 - 2.375x_{Bi} - 13.75x_{Ci} + \varepsilon_i$.

iii) We calculate *SST* by calculating the sum of squared differences between each observation and the overall mean of 72.75. (This is also the mean of the three group means.)

We calculate *SSE* by calculating the sum of squared differences between each observation and its class mean (of 78.125 for Class A, 75.75 for Class B or 64.375 for Class C).

This gives $SST = 4372.5$ and $SSE = 3508.25$ so
$SSR = 4372.5 - 3508.25 = 864.25$

Our *F*-statistic for testing the null hypothesis that
$(\beta_B - \beta_A) = (\beta_C - \beta_A) = 0$ (i.e. no differences between classes) is then
$F = \frac{\frac{SSR}{2}}{\frac{SSE}{21}} \approx 2.59$.(Note, we are testing $k = 2$ slope parameters using $n = 24$ total observations hence the $n - k - 1 = 21$ degrees of freedom in the denominator.)

$P(F_{2,21} > 2.59) \approx 0.10$ hence we do not reject the null hypothesis. There is no reason to believe that the mean scores differ between classes.