ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2016 EXAMINATIONS − SOLUTIONS

HIGHER CERTIFICATE – MODULE 4

The Society is providing these solutions to assist candidates preparing for the examinations in 2017.

The solutions are intended as learning aids and should not be seen as "model answers".

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

RSS HC Paper 4 2016 Solutions

Each question has marks out of 20.

Slightly different values for various quantities might be seen in the scripts and would be given credit provided they are clearly within the limits of rounding.

Question 1 (i) The simple linear regression plot of the oil data: **[3]**
1 mark for correct axes (including false origin) and labels, 2 for correct points (lose 1 mark per wrong point).

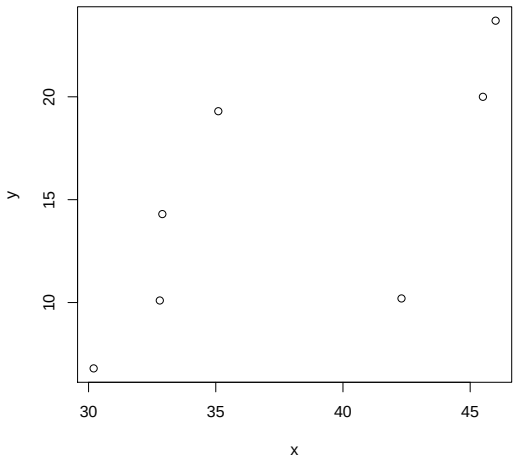

Figure 1: Model 1 - Simple Linear Regression model of Y on x

The plot shows a weak correlation, **[1]**

so a linear regression model may not be a particularly good fit. **[1]**

(Reward other sensible comments.)

(ii) Calculate:

$$\begin{aligned} S_{xx} &= \sum_i^n x_i^2 - \frac{(\sum_i^n x_i)^2}{n} \\ &= 10277.84 - \frac{264.8^2}{7} \end{aligned}$$

$$= 260.8343 \quad [1]$$

$$\begin{aligned} S_{xy} &= \sum_i^n x_i y_i - \frac{(\sum_i^n x_i)(\sum_i^n y_i)}{n} \\ &= 4116.2 - \frac{264.8 \times 104.4}{7} \\ &= 166.8971 \quad [1] \end{aligned}$$

$$\begin{aligned} \widehat{\beta} &= \frac{S_{xy}}{S_{xx}} \quad [1] \\ &= \frac{166.8971}{260.8343} \\ &= 0.6398589 \quad [1] \end{aligned}$$

$$\begin{aligned} \widehat{\alpha} &= \bar{y} - \widehat{\beta}\bar{x} \quad [1] \\ &= \frac{104.4}{7} - 0.6398589\frac{264.8}{7} \\ &= -9.290662 \quad [1] \end{aligned}$$

(iii) Assume that the errors are normally distributed **[0.5]**
with mean zero **[0.5]**
and variance σ^2 **[0.5]**
and that they are independent. **[0.5]**

$$\text{Var}[\widehat{\beta}] = \frac{\sigma^2}{S_{xx}}$$

$$\begin{aligned} \text{Total SS} &= S_{yy} \\ S_{yy} &= \sum_i^n y_i^2 - \frac{(\sum_i^n y_i)^2}{n} \\ &= 1790.6 - \frac{104.4^2}{7} \\ &= 233.9086 \, [0.5] \end{aligned}$$

$$\begin{aligned} \text{Regression SS} &= \frac{S_{xy}^2}{S_{xx}} \\ &= \frac{166.8971^2}{260.8343} \\ &= 106.7906 \quad [1] \end{aligned}$$

$$\begin{aligned}
\text{Residual SS} &= \text{Total SS} - \text{Regression SS} \\
&= 233.9086 - 106.7906 \\
&= 127.1180 \,[0.5]
\end{aligned}$$

$$\begin{aligned}
s^2 &= \frac{\text{Residual SS}}{7-2} \\
&= 25.424 \\
\text{Var}\,\widehat{\beta} &= \frac{s^2}{S_{xx}} \\
&= \frac{25.424}{260.8} \\
&= 0.0975 \quad [1]
\end{aligned}$$

A 95% Confidence interval for β is:

$$\widehat{\beta} \pm t_5(0.025)\sqrt{\frac{s}{S_{xx}}} \quad [1]$$

$$0.6399 \pm 2.571 \times 0.3122$$

$$0.6399 \pm 0.8027$$

$$(-0.1628, 1.4426) \quad [1]$$

(iv) A point prediction is

$$\begin{aligned}
\widehat{\alpha} + \widehat{\beta}x &= -9.290662 + 0.6399 \times 40 \quad [1] \\
&= 16.305 \quad [1]
\end{aligned}$$

Question 2 (a) (i) Sample product moment correlation coefficient

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad [1]$$

(or equivalent expression) where $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$ **[1]**
$S_{xx} = \sum_i (x_i - \bar{x})^2$, $S_{yy} = \sum_i (y_i - \bar{y})^2$. **[1]**

(ii)

$$\sum_{i=1}^{8} x_i = 386 \quad \sum_{i=1}^{8} y_i = 460 \quad \sum_{i=1}^{8} x_i^2 = 25426$$

$$\sum_{i=1}^{8} y_i^2 = 28867 \quad \sum_{i=1}^{8} x_i y_i = 26161$$

$$\begin{aligned} S_{xx} &= \sum_i^n x_i^2 - \frac{(\sum_i^n x_i)^2}{n} \\ &= 25426 - \frac{386^2}{8} \\ &= 6801.5 \quad [1] \end{aligned}$$

$$\begin{aligned} S_{yy} &= \sum_i^n y_i^2 - \frac{(\sum_i^n y_i)^2}{n} \\ &= 28867 - \frac{460^2}{8} \\ &= 2417 \quad [1] \end{aligned}$$

$$\begin{aligned} S_{xy} &= \sum_i^n x_i y_i - \frac{(\sum_i^n x_i)(\sum_i^n y_i)}{n} \\ &= 26161 - \frac{386 \times 460}{8} \\ &= 3966 \quad [1] \end{aligned}$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{3966}{\sqrt{6801.5 \times 2417}} = \frac{3966}{4054.5315} = 0.978. \quad [1]$$

(iii) Let H_0 : $\rho = 0$ and H_1 : $\rho > 0$ **[1]**
The 1% critical value from Table 14 is 0.7887. **[1]**
The observed value is 0.9782, **[1]**
which is larger than the critical value,
so strong evidence against the null hypothesis at the 1% significance level/there is a significant positive correlation between 'sales' and 'advertising'. Either comment **[1]**

(b) (i) Let y be Judge 1's ranking and x be Judge 2's ranking.

Cat	Rank y	Rank x
A	4	6
B	1	2
C	2	1
D	5	4
E	3	3
F	6	5
G	7	7
H	8	8

Correct ranks each judge **[1,1]**

(ii) (The correct follow through from incorrect ranks get full marks.)
Spearman rank correlation coefficient

$$r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

[1]

Cat	Rank y	Rank x	d	d_i^2
A	4	6	-2	4
B	1	2	-1	1
C	2	1	1	1
D	5	4	1	1
E	3	3	0	0
F	6	5	1	1
G	7	7	0	0
H	8	8	0	0

Correct d_i's and $\sum d_i^2 = 8$. **[1]**

$$r_S = 1 - \frac{6 \times 8}{8(8^2-1)} = 1 - \frac{6}{63} = 0.9048 \quad [1]$$

1% point 0.8333 **[1]**
$\Rightarrow$ Strong evidence against H_0, which shows that there appears to be a strong association between the judges rankings. **[1]**

(iii) The Spearman's rank correlation coefficient does not rely on any distributional assumptions. **[1]**
Possible advantages are that it does not rely on the variables being normally distributed and it is very easy to calculate. **[1]**
Possible disadvantage is that it does not use the actual values and only the rank of the values, which loses information. **[1]**
Give 1 mark for each of any sensible advantages/disadvantages.

Question 3 (i) Randomisation reduces bias. **[1]**
Need replication to measure variation **[1]**
and estimate quantities of interest. **[1]**
Give an example. **[1]**

(ii) For the completely randomised design we use the so-called one-way model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad [1]$$

where $j = 1, 2, \ldots, r_i$, $i = 1, 2, \ldots, t$ and $\sum_{i=1}^{t} r_i = n$. In the case where all the $r_i = r$, then $n = rt$.
y_{ij} is the response of the jth unit receiving the ith treatment **[0.5]**

μ is an overall mean effect, **[0.5]**
α_i is the effect due to the ith treatment **[1]**
and ε_{ij} is random error. We have the usual least-squares assumptions and normality assumption that the

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

and are uncorrelated (or independent). **[1]**
Lose 0.5 marks for each missing assumption to a maximum of 1 lost mark.

(iii) Incomplete table with marks to award for each omitted value:

Source	d.f.	sum of squares	mean square	F
Paint	(1)	(1)	(1)	(1)
Error	(1)	577.5	48.125	
Total	(1)	983.8		

Give marks for each omitted value (as in brackets) in the table **[6]**

Source	d.f.	sum of squares	mean square	F
Paint	3	406.3	135.43	2.814
Error	12	577.5	48.13	
Total	15	983.8		

$0.05 < P < 0.1$. **[1]**
So there is weak evidence for a difference in the 4 paints. **[1]**

(iv) Each paint was used entirely on 4 ships. This may induce a bias, for example if the 4 ships with one type of paint travel in particular rough seas and 4 ships in calm seas, this may not be a useful comparison. If the size of ships is not the same could also induce another difficulty as paint exposure might be different.
For potential problems with the randomisation **[2]**
It might have been better that each type of paint was used on each ship randomising over areas of the ship that was used.
For ideas of what might be an improved design. **[2]**

Question 4 (i) Multiple regression with two explanatory variables X_1 and X_2. **[1]**

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \quad \text{for } i = 1, \ldots, n \quad [1]$$

This model has $p = 3$ unknown parameters β_0, β_1 and β_2. **[1]**
We make assumptions that the errors ε_i are independent and normally distributed with mean zero and constant variance σ^2 **[2]**
Give 0.5 marks for each assumption.

(ii) To test $H_0 \;:\; \beta_1 = \beta_2 = 0$ vs $H_1 \;:\;$ at least one non zero, **[1]**
we calculate the F statistic

$$F^* = 229.525 \sim F^2_{10} \quad [1]$$

As the quoted P-value is less than 0.01 or using

$$F^2_{10}(0.01) = 7.56$$

we can reject H_0 at the 1% level. **[1]**

We need to include at least one of the variables in the model. **[1]**

(iii) The fitted value is

$$\hat{y} = 52.577 + 1.468X_1 + 0.662X_2 \quad [3]$$

1 mark for each correct coefficient
The predicted value at $X_1 = 8$ and $X_2 = 35$ is

$$\hat{y} = 52.577 + 1.468 \times 8 + 0.662 \times 35 = 87.491 \quad [2]$$

1 marks for each of the last two equalities.

(iv) Testing the coefficients $H_0 \;\; \beta_1 = 0$ and $H_0 \;\; \beta_2 = 0$, versus alternative hypotheses that they are non zero: **[1]**

$$t_{10}(0.0005) = 4.587 \quad [1]$$

12.11>4.587 so we reject $H_0 \;:\; \beta_1 = 0$ at the 0.10% significance level. **[0.5]**
14.44>4.587 so we reject $H_0 \;:\; \beta_2 = 0$ at the 0.10% significance level. **[0.5]**
This means that variables percentages of tricalcium aluminate (X_1) and tricalcium silicate (X_2) are both significant and affect the heat evolved during the hardening of cement. **[1]**
Marks may be assigned for sensible conclusions.

(v) The R^2 value is 0.9787, which means that 97.87% of the variation is explained by the variables X_1 and X_2, **[1]**
so the model fits well. **[1]**