ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2015 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 2015 Solutions

Each question has marks out of 20.

Note: Instructions on what to do about any odd half-marks. Our usual rule is to round up any odd half in the overall mark for a question - note that this is not done in the mark for a part of the question, and conversely it is not the case that any odd halves are carried forward and only rounded up once at the end of the whole paper.

Question 1 (i)

$$\varepsilon_i \sim N(0, \sigma^2) \text{ independent}$$

Normal **[0.5]**
Independent or uncorrelated **[0.5]**
Mean 0 and variance σ^2 **[0.5+0.5]**

(ii) To find the least squares estimators we must minimise S, where

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

To do this we find the solutions to the equations

$$\frac{\partial S}{\partial \beta_0} = 0 \text{ and } \frac{\partial S}{\partial \beta_1} = 0.$$

[1]

Now

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i), \quad [1] \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i). \quad [1] \end{aligned}$$

So the least squares estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are solutions to the normal equations

$$\begin{aligned} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \quad \text{(A)} \quad [1] \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \quad \text{(B)}. \quad [1] \end{aligned}$$

If we multiply (B) by n and (A) by $\sum x_i$ and subtract we obtain the estimate for β_1 as

$$\begin{aligned} \widehat{\beta}_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}. \quad [1] \\ &= \frac{S_{xy}}{S_{xx}} \quad [1] \end{aligned}$$

To find the estimate for β_0 we divide (A) by n and rearrange to obtain

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}. \quad [1]$$

(iii)

$$\widehat{\beta}_1 = \frac{201.665}{3630.95} = 0.0554 \quad [1]$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} = \frac{68.3}{20} - 0.0554 \times \frac{809}{20} = 1.168 \quad [1]$$

Fitted model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 1.168 + 0.056 x_i$$

(iv) Total $SS = S_{yy} = 12.9455$.

Reg $SS = \widehat{\beta}_1^2 S_{xx} = \frac{S_{xy}{}^2}{S_{xx}} = \frac{201.665^2}{3630.95} = 11.201$

Residual SS = Total SS − RegSS = $12.9455 - 11.201 = 1.7449$

This information can all be summarised in an Analysis of Variance (ANOVA):

Source	SS	d.f.	MS	F
Regression	11.201	1	11.201	115.55
Residual	1.7449	18	$s^2 = 0.0969$	
Total	12.9455	19		

where SS stands for sums of squares and MS for mean square.

SS column **[2]**
(Give full 2 marks if correct and [1] for 1 error.)
df column **[1]**
MS column **[1]**
F value **[1]**

Test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$.

$$F^* = 115.55 \sim F^1_{18} \text{ if } H_0 \text{ is true, } [0.5]$$

$$F^1_{18}(0.001) = 15.38\,[0.5]$$

give half a mark for the hypothesis and half for the use of the F distribution.

Hence $P < 0.001$, OR strong evidence against H_0, OR need X in the model. **[1]**

Marks given for correct conclusion.

Total marks [7]

Question 2 (i)

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Marks Assumptions (ε_i):
independent or uncorrelated **[0.5]**
Normal **[0.5]**
Mean 0, Variance σ^2 **[0.5+0.5]**
Any valid constraint eg $\mu = 0$, $\sum \alpha_i = 0$, $\alpha_1 = 0$, etc **[1]**

(ii)

$$SS(\text{carpet products}) = \frac{\sum T_i^2}{t} - \frac{G^2}{n}$$

where $G = 220.56$.

$$\begin{aligned} SS(\text{carpet products}) &= \frac{57.93^2 + 38.94^2 + 51.23^2 + 72.46^2}{4} - \frac{220.56^2}{16} \\ &= \frac{12747.374}{4} - 3040.4196 \\ &= 146.374 \quad [1] \end{aligned}$$

Total $SS = S_{yy} = \sum y_{ij}^2 - \frac{G^2}{n} = 3350.27 - \frac{220.56^2}{16} = 309.8504$ **[1]**

Residual SS = Total $SS - SS$(carpet products) = 309.8504−146.374 = 163.476 **[1]**

ANOVA table

Source	SS	d.f.	MS	F
Carpet products	146.374	3 [0.5]	48.791 [1]	3.582 [1]
Residual	163.476	12 [0.5]	$s^2 = 13.623$ [1]	
Total	309.850	15		

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ all carpet products are equally durable vs H_1 : at least one non-zero **[1]**

$$F^* = 3.582 \sim F_{12}^3(0.05) = 3.49 \Rightarrow P < 0.05$$

[1]

Rejection region $\{F : F > 3.49\}$ **[1]**

Can reject H_0 at the 5% significance level. **[1]**

(iii) Carpet 2 Synthetics Carpet 4 Wool alone

Test $H_0 : \alpha_2 = \alpha_4$ vs $H_1 : \alpha_2 \neq \alpha_4$ **[1]**

$$
\begin{aligned}
t^* &= \frac{\bar{y}_2 - \bar{y}_4 - 0}{s\left(\frac{1}{n_2} + \frac{1}{n_4}\right)^{\frac{1}{2}}} \quad [1] \\
&= \frac{\frac{38.94}{4} + \frac{72.46}{4}}{13.623^{\frac{1}{2}}\left(\frac{1}{4} + \frac{1}{4}\right)^{\frac{1}{2}}} \\
&= \frac{-8.38}{2.610} = 3.211 \quad [1]
\end{aligned}
$$

$$t^* \sim t_{12} \text{ if } H_0 \text{ is true} \quad [1]$$

$$t_{12}(0.005) = 3.055 \Rightarrow P < 0.005 \quad [1]$$

EITHER Strong evidence against H_0 OR There is a significant difference between Carpet 2 and Carpet 4. **[1]**

Question 3 (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) $z_i = ax_i + b$, $a > 0$

$$r_{zy} = \frac{S_{zy}}{\sqrt{S_{zz}S_{yy}}}$$

[1]

where

$$
\begin{aligned}
S_{zy} &= \sum_i (ax + b - (a\bar{x} + b))(y_i - \bar{y}) \quad [1] \\
&= a\sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
&= aS_{xy}, \quad [1]
\end{aligned}
$$

$$
\begin{aligned}
S_{zz} &= \sum_i (ax_i + b - (a\bar{x} + b))^2 \\
&= a^2 \sum_i (x_i - \bar{x})^2 \quad [1]
\end{aligned}
$$

S_{yy} stays the same.

Hence

$$r_{zy} = \frac{S_{zy}}{\sqrt{S_{zz}S_{yy}}} = \frac{aS_{xy}}{\sqrt{a^2 S_{xx}S_{yy}}} = r_{xy} \text{ for } a > 0 \quad [1]$$

and for $a < 0$, $r_{zy} = -r_{xy}$. **[1]**

(iii)

$$S_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = 19862.6 - \frac{1190.0 \times 248.5}{15} = 140.333 \quad [1]$$

$$S_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = 95098 - \frac{1190.0^2}{15} = 691.337 \quad [1]$$

$$S_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = 4161.1 - \frac{248.5^2}{15} = 40.9693 \quad [1]$$

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{140.333}{\sqrt{691.337 \times 40.9693}} = 0.834. \quad [1]$$

$H_0 \ : \ \rho_{XY} = 0$ vs $H_1 \ : \ \rho_{XY} > 0$

1% point 0.5923 **[1]**

At 1% significance level there is strong evidence against H_0, which shows that there is a strong positive correlation between number of chirps and temperature. **[1]**

(iv) Spearman rank correlation coefficient

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

[1]

y	Rank y	x	Rank x	d
14.4	1	76.3	6	-5
14.7	2	69.7	2	0
15.0	3	79.6	7	-4
15.4	4	69.4	1	3
15.5	5	75.2	5	0
15.7	6	71.5	3	3
16.0	7	71.6	4	3
16.1	8	80.5	8	0
16.3	9	83.3	11	-2
17.0	10	83.5	12	-2
17.1	11	80.6	9	2
17.2	12	82.6	10	2
18.4	13	84.3	13	0
19.8	14	93.3	15	-1
20.0	15	88.6	14	1

Correct ranks **[1]**
(The correct follow through from incorrect ranks get full marks.)
Correct d_i's and $\sum d_i^2 = 86$. **[1]**

$$r_S = 1 - \frac{6 \times 86}{15(15^2 - 1)} = 1 - 0.1536 = 0.8464 \quad [1]$$

1% point $0.6036 \Rightarrow$ Strong evidence against H_0, which shows that there is a strong association between the number of chirps and temperature. **[1]**

Question 4 (i) A multiple regression model for p explanatory variables $x_1, x_2, \ldots, x_p$ by

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{p,i} + \varepsilon_i.$$

Marks to be given with or without the subscript i.
Correct coefficients $\beta_0, \beta_1, \ldots, \beta_p$ of the explanatory variables **[1]**
Correct use of response and error term in model above. **[1]**
where Y_i is the response variable.
This model has $p + 1$ unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$. ε_i are independent and normally distributed or uncorrelated **[0.5+0.5]**
errors have mean zero and constant variance σ^2. **[0.5+0.5]**

(i) Test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (i.e. all coefficients except β_0 are zero) versus H_1 : at least one of the coefficients is non-zero. **[1]**

Consider the statistic $F^\star$ defined by

$$F^\star = \frac{\text{regression MS}}{s^2} = \frac{128435}{661} = 194.30$$

[1]

$$F^\star \sim F^4_{30} \text{ under } H_0$$

[1]

$$F^4_{30}(0.05) = 2.69 \quad [1]$$

$F^\star$ is very large $\Rightarrow$ very strong evidence against H_0. **[1]**

So 96.3% of the variation is explained by this model, so it suggests that the model fits well. **[1]**

(ii) $H_0 \;:\; \beta_{Head} = 0$ vs $H_1 \;:\; \beta_{Head} \neq 0$ **[1]**

$$t = \frac{-9.013}{4.693} = -1.92 \sim t_{30} \text{ under } H_0 \quad [1]$$

$t_{30}(0.025) = 2.042$ **[1]** At the 5% significance level no evidence against H_0 **[1]** $\Rightarrow$ exclude the variable Head from the model. **[1]**

90% confidence interval for β_{Chest}

$$\hat{\beta}_{Chest} \pm t_{30}(0.05) se(\hat{\beta}_{Chest}) \quad [1]$$

$$6.255 \pm 1.697 \times 1.677$$

[1]

$$(3.409, 9.101)$$

[1]

(iii) The regression equation is

$$\hat{y} = -203 + 0.655Age - 9.01Head + 11.8Neck + 6.26Chest$$

$$\hat{y} = -203 + 0.655 \times 71 - 9.01 \times 7 + 11.8 \times 27 + 6.26 \times 44 \quad [1]$$

$$\hat{y} = 374.475 \quad [1]$$

If you use the coefficients given to the number of dp in the question, this is 375.1545. Please accept anything that rounds to 374 or 375.