ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2017

MODULE 4 : Linear models

Time allowed: One and a half hours

Candidates should answer **THREE** *questions.*

Each question carries 20 marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$.

This examination paper consists of 8 printed pages.
This front cover is page 1.
Question 1 starts on page 3.

There are 4 questions altogether in the paper.

BLANK PAGE

1. A gas generation plant distills liquid air to produce oxygen. The percentage purity of the oxygen is thought to be linearly related to the amount of impurities in the air, as measured by the "pollution count" in parts per million by volume (ppm). The following data were collected on 15 successive days.

Purity (%) y	93.3	92.0	92.4	91.7	94.0	94.6	93.6	93.1
Pollution count (ppm) x	1.10	1.45	1.36	1.59	1.08	0.75	1.20	0.99

Purity (%) y	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution count (ppm) x	0.83	1.22	1.47	1.81	2.03	1.75	1.68

(i) Fit a linear regression to the data using $\sum x_i = 20.31$, $\sum y_i = 1387.9$, $S_{xx} = 1.959\,56$, $S_{xy} = -5.6846$ and $S_{yy} = 18.8693$, where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

(4)

(ii) Construct the ANOVA table for this model and perform the appropriate hypothesis test using the 0.1% significance level. Hence write down the value of s^2, the estimate of the error variance.

(9)

(iii) Find a 95% confidence interval for the slope of the regression equation.

(3)

(iv) Find a 90% confidence interval for the mean purity on a day when the pollution count is 1.00.

(4)

[You may use the fact that the estimated variance for the predicted mean response $\hat{\alpha} + \hat{\beta}x_0$ is $s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$.]

2. (a) Sketch scatter diagrams to illustrate the following features of bivariate data. Comment briefly on each of your plots.

(i) Strong positive association, appropriately reflected by the product moment correlation coefficient.

(2)

(ii) Weak negative association, appropriately reflected by the product moment correlation coefficient.

(2)

(iii) Strong negative association, appropriately reflected by Spearman's rank correlation coefficient, but less satisfactorily by the product moment correlation coefficient.

(2)

(iv) Strong association with a non-monotonic trend.

(2)

(b) The following table shows diastolic (DBP) and systolic (SBP) blood pressure measurements (in mm Hg) for 10 randomly chosen cardiac patients.

DBP	55	60	70	75	80	85	90	95	105	110
SBP	125	115	120	135	105	145	130	200	190	150

$S_{xx} = 2962.5, \quad S_{xy} = 3437.5$ and $S_{yy} = 8802.5,$ where $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$,

$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ and $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

(i) Calculate the sample product-moment correlation coefficient of these data. Test at the 1% level the null hypothesis that $\rho = 0$ against the alternative hypothesis that $\rho > 0$, where ρ is the population value of the product moment correlation coefficient. State any assumptions made in performing the test.

(6)

(ii) Calculate the value of Spearman's rank correlation coefficient for these data, and carry out the corresponding test. State your conclusions clearly.

(6)

3. An experiment was conducted to examine the effect of different lighting conditions on the number of eggs laid by a certain breed of chickens. The treatments were

O : control (natural daylight),

E : extended day (natural daylight extended by artificial light to a total of 14 hours),

F : flashlight (natural daylight plus flashes of light every 20 seconds through the night).

Twelve pens each containing 6 chickens were randomly allocated to the three treatments. The total number of eggs laid in a given period was recorded as follows.

O	330	288	295	313
E	372	340	343	341
F	359	337	373	302

You are given that $\sum_{i=1}^{3}\sum_{j=1}^{4} y_{ij}^2 = 1\,337\,535$, where y_{ij} is the number of eggs laid in the *j*th pen under the *i*th treatment.

(i) Write down an appropriate model for these data, explaining fully all the terms in the model. State any assumptions that are made for this model.

(4)

(ii) Draw up an Analysis of Variance table for this model and test for differences in the treatments at the 10% significance level.

(11)

(iii) Test whether there are treatment differences between the extended day (*E*) and flashlight (*F*) at the 5% significance level.

(5)

4. Data for the first year box office receipts (Y) have been collected for a number of movies. A project to model these receipts collects data on the total production costs (X_1), promotional costs (X_2) and any associated book sales (X_3), all data being measured in millions of US dollars. Consider the edited computer output from three regression models, labelled A, B and C, as given **below and on the next page**.

(i) Briefly comment on the scatter plots, in Figures 1, 2 and 3, of the observed against fitted y for the three models. Relate your comments in each case to s, the square root of the mean square error.

(4)

(ii) In Model A, test for the global significance, at the 5% level, of the regression model, stating clearly the null and alternative hypotheses that you are testing. State the null distribution of the test statistic. Interpret the statement "Multiple R-Squared: 0.9668" and explain how this quantity is calculated.

(5)

(iii) By considering the output from all three models, say which of the explanatory variables should be included in the model and justify your answer using appropriate t tests.

(6)

(iv) Which of the three models do you consider best describes the data? Justify your answer.

(5)

Model A:

```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)    7.6760        6.7602      1.135      0.2995
x1             3.6616        1.1178      3.276      0.0169 *
x2             7.6211        1.6573      4.598      0.0037 **
x3             0.8285        0.5394      1.536      0.1754
---
Signif. codes:   0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 7.541 on 6 degrees of freedom
Multiple R-squared: 0.9668,    Adjusted R-squared: 0.9502
F-statistic: 58.22 on 3 and 6 DF,    p-value: 7.913e-05
```

Model B:

```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)    11.848         6.765      1.751     0.12334
x1              4.228         1.153      3.667     0.00800 **
x2              7.436         1.806      4.117     0.00448 **
---
Signif. codes: 0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 8.241 on 7 degrees of freedom
Multiple R-squared: 0.9537,    Adjusted R-squared: 0.9405
F-statistic: 72.14 on 2 and 7 DF,    p-value: 2.131e-05
```

Model C:

```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)    23.163         9.625      2.407      0.0427 *
x2             12.669         1.771      7.155    9.66e-05 ***
---
Signif. codes: 0  ***  0.001  **  0.01  *  0.05  .  0.1    1

Residual standard error: 13.17 on 8 degrees of freedom
Multiple R-squared: 0.8648,    Adjusted R-squared: 0.8479
F-statistic: 51.19 on 1 and 8 DF,    p-value: 9.665e-05
```

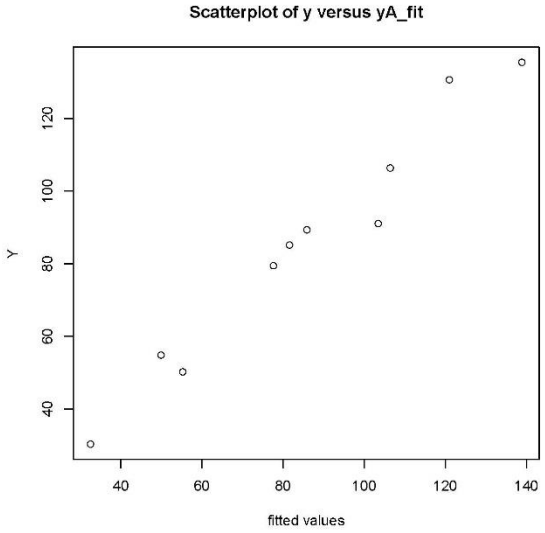

Figure 1
Model A – Scatterplot of *Y* vs fitted values

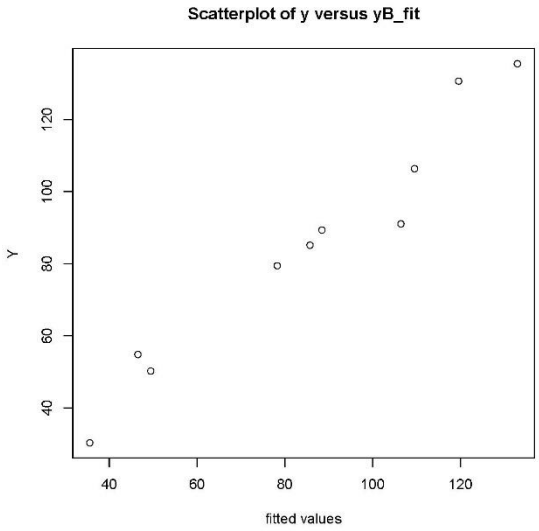

Figure 2
Model B – Scatterplot of *Y* vs fitted values

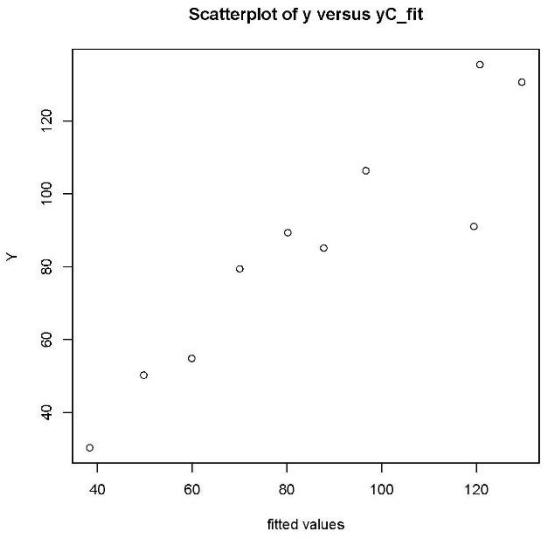

Figure 3
Model C – Scatterplot of *Y* vs fitted values

BLANK PAGE