THE ROYAL STATISTICAL SOCIETY

2010 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

MODULE 4

LINEAR MODELS

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

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.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Part (a)

(i) The idea of randomisation is that the treatments (e.g. fertilisers) which the experiment is intended to compare should be allocated *at random* to the experimental material (e.g. wheat of a given variety). Using this example as the context, any differences in crop yields between treatments will then tend to be due to the different treatments (fertilisers) used, as the random variations between individual wheat plants of the same variety will tend to average out. Allocation at random should also help eliminate any (possibly unsuspected) sources of bias, such as a consistent "fertility gradient" in the natural fertility of the soil in the field where the experiment is carried out. Analysis of the data focuses on comparing the variation of yields between treatments with the random variation within treatments: the greater the between-treatments variation relative to the within-treatments variation, the more likely it is to be due to real differences between the effects of the treatments than to have arisen by chance.

To improve the accuracy of the analysis, it is necessary to apply each treatment to several *replications* (repetitions which are expected to be identical apart from random variation: for example, standard plot areas or individual plants), to assist the averaging out of random variations within the experimental material assigned to each treatment.

The examples given by candidates in the examinations were expected to clearly identify *experimental material* (wheat in the example above) and *treatments* (fertilisers)**.** Answers were expected to mention the key aim of comparing systematic variation between treatments with random variation reflecting the random allocation of treatments to experimental material.

(ii) $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}\,, \quad i = 1, \ldots, k; \quad j = 1, \ldots, r$

$i = 1, \ldots, k$ indexes the treatments

$j = 1, \ldots, r$ indexes the replications

μ = the true (i.e. population) overall mean yield

α_i = the true (population) mean effect of the ith treatment relative to the overall mean

ε_{ij} = independent $N(0, \sigma^2)$ error associated with the jth replication of the ith treatment

Solution continued on next page

Part (b)

(i) The grand total is 50 + 75 + 85 + 100 = 310 (equivalently, the grand mean = 310/20 = 15.5). The sum of squares of all 20 observations is 544 + 1181 + 1505 + 2044 = 5274.

"Correction factor" is $\frac{310^2}{20} = 4805$.

Therefore total SS = 5274 – 4805 = 469, with 19 df.

SS for treatments (i.e. % cotton in fibre)

$$= \frac{50^2}{5} + \frac{75^2}{5} + \frac{85^2}{5} + \frac{100^2}{5} - 4805 \; = \; 265 \text{, with 3 df.}$$

The residual SS and df are obtained by subtraction: SS is 469 – 265 = 204, with 19 – 3 = 16 df.

Hence the analysis of variance table is as follows.

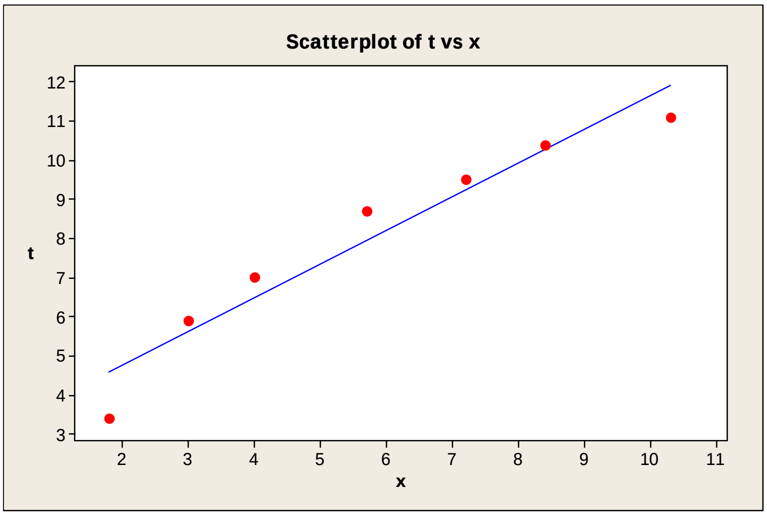
SOURCE	DF	SS	MS	*F* value
% cotton	3	265	88.33	6.93 Compare $F_{3,16}$
Residual	16	204	12.75	$= \hat{\sigma}^2$
TOTAL	19	469		

A level of significance for a formal test is not specified in the question. However, the upper 0.5% point of $F_{3,16}$ is 6.30, and the *F* value from the analysis of variance exceeds this. So the suppliers effect is very highly significant**.** There is very strong evidence that varying the percentage of cotton in the fibre does affect the tensile strength.

(ii) As % cotton is a factor based on an interval scale, we might expect that there will be a trend in tensile strength (TS) as % cotton increases. The table of data clearly shows that average TS increases with % cotton, and it would be natural to test whether the increases are linear. The procedure would be to regress the 20 observations of TS (the dependent variable) on % cotton (the independent variable), noting that there will be 5 observations of TS for each distinct value of % cotton. The sum of squares (SS) for this regression will be that attributable to a linear trend in the effects of % cotton. The amount by which this SS falls short of the SS for % cotton in the overall analysis of variance in part (b)(i), i.e. 265, represents the SS due to nonlinear (quadratic, cubic, etc) dependence on % cotton. The adequacy of the simple linear regression model (or indeed of a proportional model) can be tested; details of this were not expected from the candidates in the examination.

Higher Certificate, Module 4, 2010. Question 2

(i)



[Note "false origin".]

Values of t show a strong increasing trend with x, but the trend appears to be a curve with decreasing gradient, rather than a straight line. It appears that the straight line model will underestimate the value of t in the middle of the range of x but overestimate the value of t for low or high values of x.

(ii) Taking logarithms (base e) of the relationship $\exp(t) = Ax^B$, we find

$$t = \log A + B\log x.$$

Matching this with $t = a + b\log x$, we find $a = \log A$, $b = B$.

(iii) Putting the data of the question into the standard simple linear regression formulae, we have

$$\hat{b} = \frac{n\Sigma t\log x - \Sigma t\Sigma\log x}{n\Sigma(\log x)^2 - \left(\Sigma\log x\right)^2} = \frac{(7\times 100.101) - (56\times 11.2476)}{7\times 20.3687 - 11.2476^2} = \frac{70.8414}{16.0724} = 4.4076$$

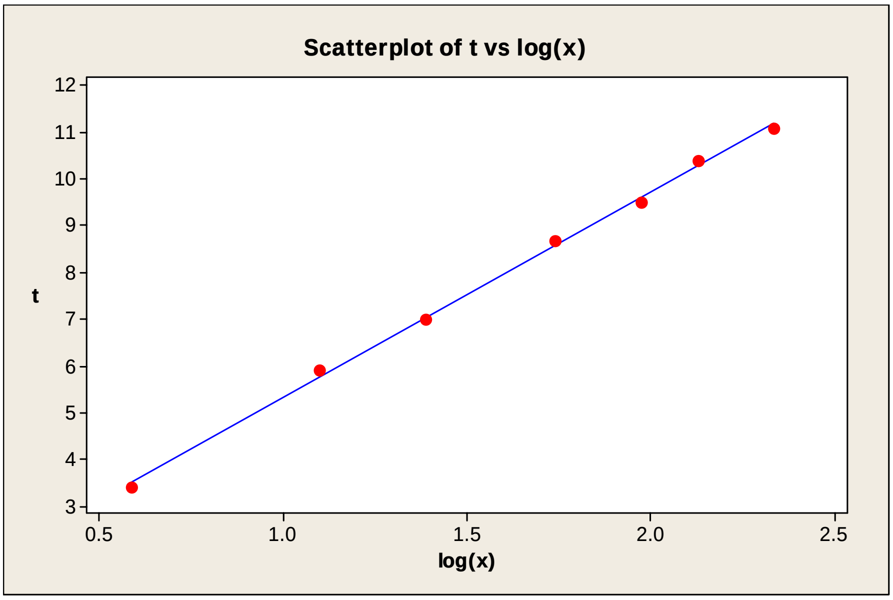
$$\hat{a} = \bar{t} - \hat{b}\,\overline{\log x} = 8 - \left(4.4076\times\frac{11.2476}{7}\right) = 0.917(87).$$

[0.9178 is used in the sequel – this is the value of $\hat{a}$ to 4 d.p. as calculated without rounding $\hat{b}$.]

So the regression line using $\log x$ as the independent variable is

$$t = 0.9178 + 4.4076\log x.$$

Solution continued on next page



[Note "false origin".]

(iv) As noted above, the original scatterplot follows an increasing nonlinear trend with decreasing gradient: the fit exceeds the data at the extremes of the plot but is below the data in the middle. However, the data points show a very good linear trend when $\log x$ is used as the independent variable. All 7 points lie very close to the fitted line and are haphazardly above or below it, so the scatter is small and seems random, and also seems reasonably constant across different values of $\log x$. The second model is preferable.

When $x = 6$, the first model predicts $t = 3.027 + (0.8617 \times 6) = 8.20$ (to 3 significant figures).

The second model predicts $t = 0.9178 + 4.4076 \log 6 = 8.815$.

As expected for this middle-range value of x, the first model under-predicts compared with the second, by about 0.6. The discrepancy far exceeds the observable scatter for the second model, reinforcing the superiority of the second model.

Higher Certificate, Module 4, 2010. Question 3

(a) The sample product moment correlation coefficient is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

where $\bar{x}$ and $\bar{y}$ denote the respective sample mean values of $x_1, x_2, ..., x_n$ and of $y_1, y_2, ..., y_n$, and all summations are over the data.

Note the equivalent formulae, usually used for (hand) calculation:

$$r = \frac{\Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\left(\Sigma x_i^2 - \dfrac{(\Sigma x_i)^2}{n}\right)\left(\Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n}\right)}} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{\sqrt{\left(n\Sigma x_i^2 - (\Sigma x_i)^2\right)\left(n\Sigma y_i^2 - (\Sigma y_i)^2\right)}}.$$

For i = 1, 2, …, n, let $R(x_i)$ be the rank of x_i when x_1, x_2, …, x_n are placed in ascending order, and similarly let $R(y_i)$ be the rank of y_i when they are placed in ascending order. Then Spearman's (sample) rank correlation coefficient r_s is given by replacing each x_i by $R(x_i)$ and y_i by $R(y_i)$ in the above formula for r, noting that the mean values $\bar{x}, \bar{y}$ both become $(n + 1)/2$, the mean of the numbers (ranks) 1, 2, …, n.

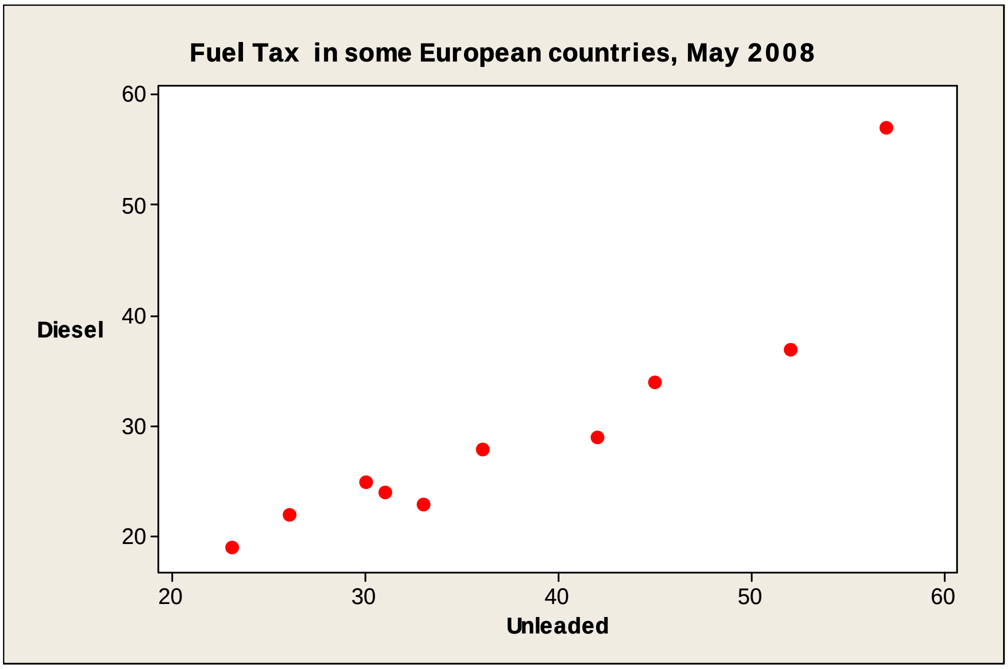
r measures, for a sample, the mutual linear association between two quantities or variables, x and y say. It is useful if the two variables are thought to have an underlying linear relationship.

r_s measures, for a sample, the strength of association between two quantities or variables, x and y say, which are thought to have an underlying monotonic relationship (which may or may not be linear). However, in the case of a relationship which is monotonic but nonlinear, r will tend to understate the true strength of the association between the variables.

Solution continued on next page

Part (b)

(i)



[Note "false origin".]

For the countries listed in the table, there is a quite clear monotonic increasing relationship. However, the rightmost point (the UK) seems to be somewhat off the (roughly linear) trend of the rest; the data suggest either a linear trend with a possible outlier, or perhaps a shallow curvilinear relationship.

(ii) Using, for convenience, the third of the above formulae for r, we have

$$r = \frac{(10\times 12191)-(375\times 298)}{\sqrt{(10\times 15193-375^2)(10\times 9974-298^2)}}$$

$$= \frac{10160}{\sqrt{11305\times 10936}} = \frac{10160}{11119} = 0.914.$$

From the Society's *Statistical tables for use in examinations*, the critical value for r for a sample of size 10 at the 1% level is 0.7155 for a one-sided test. Since 0.914 > 0.7155, we reject the hypothesis of zero correlation in the underlying population and conclude that there is evidence of a positive correlation between the fuel taxes levied on diesel and on unleaded petrol.

Solution continued on next page

(iii)

Unleaded (x)	*R(x)*	*Diesel (y)*	*R(y)*	*d = R(x) − R(y)*
36	6	28	6	0
42	7	29	7	0
23	1	19	1	0
52	9	37	9	0
26	2	22	2	0
30	3	25	5	−2
45	8	34	8	0
33	5	23	3	2
31	4	24	4	0
57	10	57	10	0

Calculating r_s from the formula $r_s = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)}$, we obtain $r_s = 1 - (48/990)$ = 0.9515. From the tables, the critical value for a sample of size 10 at the 1% level is 0.7455 for a one-sided test. Since 0.9515 > 0.7455, we reject the null hypothesis of no association between *x* and *y* in the underlying population and conclude that there is evidence of a positive association (monotonic relationship) between the fuel taxes levied on diesel and on unleaded petrol.

(iv) The scatter diagram of part (b)(i) casts some doubt on the linearity of the relationship between *x* and *y*, although it is clear that the relationship is monotonic. We note that $r_s > r$. The results of the two tests agree, but the rank test is more appropriate for these data. However, the strength of association is strong enough for the less appropriate test to give a significant result also.

The principal reservation is that the analysis implicitly assumes that the countries whose data are tabulated are a random sample from a defined population. The natural "population" here would be the countries of Europe, but selection would have to be without replacement. However, assumptions such as these are often made in the statistical analysis of economic data.

Higher Certificate, Module 4, 2010. Question 4

Part (i)

Usual assumptions are that the residual (error) terms should be independent, identically distributed, have zero mean and constant variance, and, if the usual inferences and tests are to be made, be Normally distributed.

Part (ii)

(a) All three scatter plots appear to be roughly linear; that for model B shows more scatter than those for A and C (which are similar in this regard), and these comments are in line with the respective values of S (B: 12.25; A: 6; C: 6).

(b) In Model B, the value of the test statistic for testing the null hypothesis that the coefficient of x_1 is zero is (5.0000 – 0)/0.9129 = 5.48, with null distribution t_6.

From tables, the critical value for a two-sided test at the 5% level is 2.447; the value of the test statistic exceeds this, so the null hypothesis is rejected, in favour of the alternative that the coefficient is non-zero, at the 5% level. It would also be rejected at the 1% level (critical value is 3.707), and the value of the test statistic is in fact not far short of the 0.1% critical value of 5.959. So here the null hypothesis is decisively rejected and there is strong evidence that this coefficient is non-zero.

Similarly in Model C, the value of the test statistic for testing the null hypothesis that the coefficient of x_2 is zero is (2.4000 – 0)/0.2000 = 12.00, again with null distribution t_6. So here the null hypothesis is very decisively rejected and there is extremely strong evidence that this coefficient is non-zero.

Solution continued on next page

(c) In Model A, the value of the test statistic for the partial t test for the significance of x_1 in the presence of x_2 is (1.0000 – 0)/1.0000 = 1.00, with null distribution t_5. The critical value for a two-sided test at the 5% level is 2.571 and the value of the test statistic is well below this critical value. It is therefore reasonable to conclude that, in the presence of x_2, x_1 can be omitted.

The test for the significance of x_2 in the presence of x_1 follows similarly: the value of the test statistic is (2.0000 – 0)/0.4472 = 4.47, again with null distribution t_5. This easily exceeds 2.571, indeed it exceeds the critical point (4.032) at the 1% level, so we may quite strongly conclude that, in the presence of x_1, x_2 is still needed.

The test for the global significance of the regression on x_1 and x_2, i.e. a test of the null hypothesis that both the coefficients are zero against the alternative that at least one of them is not, is given by considering F = (regression mean square)/(residual mean square). The value of F here is 2610.0/36.0 = 72.5. The null distribution of this is $F_{2,5}$. 72.5 is greatly in excess of all the usual critical points (e.g. the upper 5% point is 5.79), so the null hypothesis is very decisively rejected.

"R-Sq = 96.7%" means that 96.7% of the total variation of y is attributable to (can be explained by) a multiple linear regression of y on x_1 and x_2 [an alternative interpretation is that the square of the correlation between y and the best linear predictor in terms of x_1 and x_2 is 0.967].

R-Sq is calculated as $\dfrac{\text{regression sum of squares}}{\text{total sum of squares}} = \dfrac{5220}{5400}$.

(d) In each of Models B and C separately, we have found strong evidence that the single variable (x_1 and x_2 respectively) is important. Model A contains both x_1 and x_2 but we have shown that it is reasonable, in the presence of x_2, to omit x_1. However, we have also shown in Model A that, in the presence of x_1, there is strong evidence that x_2 is also needed, so this indicates that Model B (x_1 on its own) is inadequate. In Model C, the coefficient of x_2 is significantly different from zero (and so also is the constant), and in comparison with the more complicated Model A it achieves almost as good an "explanation" (R-sq) and the same mean square error. So choose Model C.