THE ROYAL STATISTICAL SOCIETY

2008 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

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. $\log_{10}$.

Higher Certificate, Module 4, 2008. Question 1

(i) (a) The null hypothesis is that there are no differences between the population mean numbers of miles per gallon (mpg) for the fuel additives. The alternative hypothesis is that at least two of these means differ.

The grand total is 185 + 195 + 340 = 720. The sum of squares of all 20 observations is 26078 (this is given in the question).

"Correction factor" is $\frac{720^2}{20} = 25920$.

Therefore total SS = 26078 – 25920 = 158.

SS for additives = $\frac{185^2}{5} + \frac{195^2}{5} + \frac{340^2}{10} - 25920 = 90$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows.

SOURCE	DF	SS	MS	*F* value
Additives	2	90	45	11.25 Compare $F_{2,17}$
Residual	17	68	4	$= \hat{\sigma}^2$
TOTAL	19	158		

A level of significance for a formal test is not specified in the question. However, the upper 0.1% point of $F_{2,17}$ is 10.66, and the *F* value from the analysis of variance exceeds even this. So the additives effect is very highly significant. There is very strong evidence to reject the null hypothesis that all the additives lead to the same mean mpg.

The means of mpg for the additives are 185/5 = 37, 195/5 = 39 and 340/10 = 34. This suggests that additive C (the current standard additive) is distinctly worse in this regard than the other two, and perhaps B is better than A.

(b) We have $\bar{y}_A - \bar{y}_C = 3.0$, and the standard error of this estimate is $\sqrt{\frac{\hat{\sigma}^2}{n_A} + \frac{\hat{\sigma}^2}{n_C}}$ $= \sqrt{4\left(\frac{1}{5} + \frac{1}{10}\right)} = 1.095$. The two-sided 5% critical value for t_{17} is 2.110, so a 95% confidence interval for the true population mean difference is given by

$$3.0 \pm (2.110 \times 1.095) \quad \text{or} \quad 3.0 \pm 2.31, \quad \text{i.e. } (0.69, 5.31).$$

The interpretation is in terms of repeated sampling: 95% of all intervals calculated in this way from sets of experimental data would contain the true value of $\mu_A - \mu_C$.

Solution continued on next page

(ii) (a) The assumptions are that the residuals are independent, identically distributed $N(0, \sigma^2)$ random variables.

These assumptions can be checked by calculating the observed residuals and checking for absence of patterns (e.g. serial correlation or some form of time sequence if it is known which observation was made on which day) and for apparent underlying Normality. Equality of the within-treatments variances can also be checked for. Details of procedures or of formal tests are not expected in this module.

(b) We would expect the degree of congestion to affect mpg, as more time will be spent idling or driving very slowly in heavily congested traffic. As congestion is likely to show well-established variation through the day, test drives should be made at a fixed time of day.

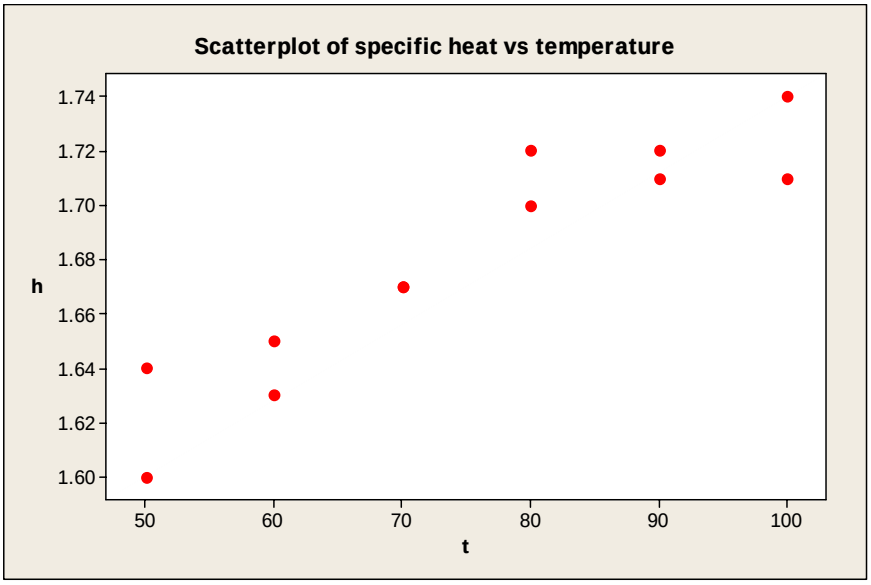
Rush-hour periods, which present the risk of exceptional conditions, should be avoided.

There may also be variations in traffic flow associated with days of the week. With 5, 5 and 10 trials, it would be reasonable to use each of the days Monday to Friday once for each of A and B and twice for C.

Atypical weeks (e.g. involving public holidays) should be avoided.

Higher Certificate, Module 4, 2008. Question 2

(i) The dependent (y) variable is h, specific heat in calories per gram. The explanatory (x) variable (also often referred to as the predictor variable or the regressor variable) is t, the temperature in degrees Celsius. A scatter plot is as follows (note that the "false origin" has *not* been corrected in this display).



The data follow a broadly linear increasing trend with roughly constant scatter. It is reasonable to assume that the temperature is preset by the experimenters without error. These three features are consistent with the assumptions for simple linear regression analysis (see (ii)(a)).

(ii) (a) [Candidates were expected merely to quote (not derive) the formulae relating to the estimates of the slope and the intercept. There are many equivalent forms for these formulae.]

The slope estimate is $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ in a familiar notation, identifying x with t and y with h as stated above.

$$S_{xy} = \sum th - \frac{\left(\sum t\right)\left(\sum h\right)}{n} = 1519.9 - \frac{900 \times 20.16}{12} = 7.9 .$$

$$S_{xx} = \sum t^2 - \frac{\left(\sum t\right)^2}{n} = 71000 - \frac{900^2}{12} = 3500 .$$

$$\therefore \hat{\beta}_1 = 7.9/3500 = 0.002257 .$$

Solution continued on next page

The intercept estimate is (again x is identified with t and y with h)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.68 - (0.002257 \times 75) = 1.511.$$

The assumptions are that, for some constants a and b, the underlying model is $y_i = a + bx_i + e_i$, for $i = 1, 2, \ldots, n$. Here the x values are preset without error (or, for inference in the context of repeated sampling, we condition on the observed values of the xs as fixed), and the es are uncorrelated and identically distributed random variables with zero mean and constant variance. If formal inference and tests are required (as in part (ii)(b)), the es are also required to be Normally distributed.

$\hat{H}(85) = 1.511 + (0.002257 \times 85) = 1.703$, or 1.70 to 3 significant figures.

(b) The estimate of σ^2 is the residual mean square in the usual analysis of variance of regression. So we need to calculate the elements of this analysis.

Total SS = $\sum h^2 - (\sum h)^2/n = 33.8894 - 20.16^2/12 = 33.8894 - 33.8688 = 0.0206$.

Regression SS = $\hat{\beta}_1^2 S_{xx}$ [see above] = $0.002257^2 \times 3500 = 0.01783$.

$\therefore$ Residual SS = $0.0206 - 0.01783 = 0.00277$, with $n - 2 = 10$ df.

$\therefore$ Residual MS, s^2 say, is $0.00277/10 = 0.000277$, with 10 df.

The variance of the estimator of the slope is $\sigma^2/S_{xx} = \sigma^2/3500$. Our estimate of this is 0.000277/3500.

We note that the double-tailed 1% point of t_{10} is 3.169. Thus we have that the critical region for testing for zero slope at the 1% level against a two-sided alternative is given by

$$|t| = \left| \frac{\hat{\beta}_1}{\sqrt{0.000277/3500}} \right| > 3.169.$$

We have $|t| = \dfrac{0.002257}{0.0002813} = 8.023$, which is (very much) greater than 3.169, so the null hypothesis of zero slope is decisively rejected and we conclude that there is very strong evidence of an increasing trend.

Solution continued on next page

(c) The differences d (first result − second result) are 0.04, −0.02, 0, 0.02, −0.01, −0.03 respectively.

The sample variance of these could be regarded as either $\Sigma d^2/6$ (on the basis that their true population mean is known to be zero) or as $\Sigma(d-\bar{d})^2/5$ (on the usual definition of "s^2"). These give 0.0005667 and 0.00068 respectively.

This is an estimate of $2\sigma^2$ (see the question), so σ^2 is estimated as either 0.000283 or 0.00034.

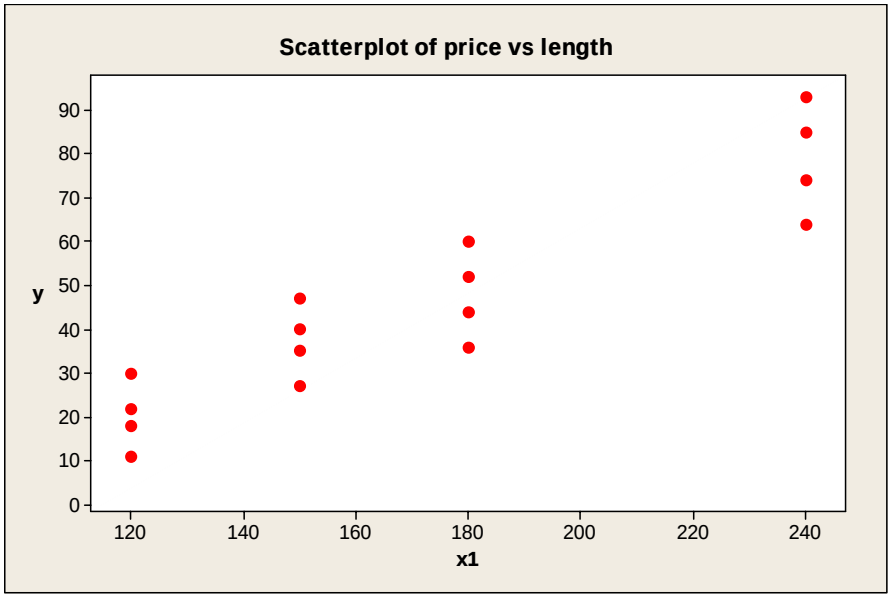
Either of these is a "pure error" estimate of σ^2, not dependent on the model used. The residual variance from the regression model includes both pure error and potential lack of fit, so if nonlinearity of trend were important we would (subject to sampling variation) intuitively expect s^2 (the estimate from the regression model) to be larger. s^2 was 0.000277, which is very close to (in fact slightly less than) the "pure error" results. This suggests that there is no obvious lack of fit: the regression model may be a good one.

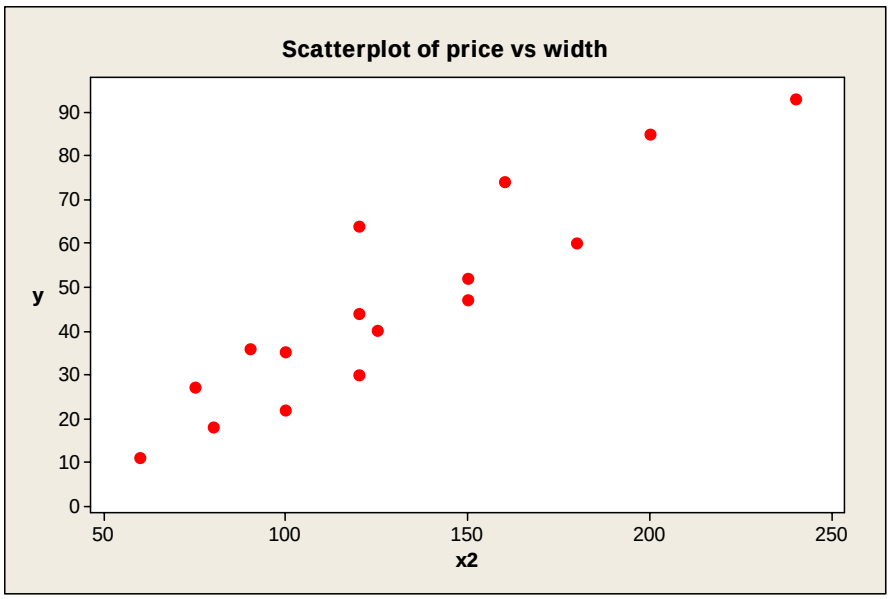
Higher Certificate, Module 4, 2008. Question 3

(i) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, for $i = 1, 2, \ldots, n$.

The $\{\varepsilon_i\}$ are independent Normally distributed residuals (or "errors") with mean 0 and constant variance σ^2 (if formal inference and tests are not required, it is sufficient to take these as uncorrelated rather than independent Normally distributed). β_0 is the overall mean response when x_1 and x_2 are both zero. β_1 and β_2 represent the expected increase in Y for unit increases in x_1 and x_2 respectively when the other x variable is kept constant.

(ii) (a) Scatter plots are as follows (note that the "false origin" has *not* been corrected in these displays).





Solution continued on next page

The graphs show a tendency for Y to increase roughly linearly as either x_1 or x_2 increases. There may be more scatter in the plot against x_1 and a slight tendency for this to increase with x_1, but these features may be due to the dependence of Y on x_2.

(ii) (b) The constant term is β_0 (in the model as stated in part (i)), and β_1 and β_2 are the respective coefficients of x_1 (length) and x_2 (width), so the fitted model is

$$\hat{Y} = -52.671 + 0.32356\, x_1 + 0.44383\, x_2 .$$

The estimated standard deviations of the estimated coefficients are as given in the output (5.34500 etc), as are the values of the test statistics and the p-values for t tests of the (separate) hypotheses $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 0$. The number of degrees of freedom for the residual is 13 (= number of data points − number of estimated coefficients), so the tests are based on t_{13}.

The t values are all large, and the corresponding p-values are very small (all zero to at least 3 decimal places). Each p-value measures the probability of obtaining an estimated coefficient at least as large in absolute value as is observed if in fact the true value of the coefficient is zero. These results strongly suggest that all three terms need to be in the model.

The residual mean square is given in the analysis of variance as 28.4; this is the estimate of the error variance σ^2. The square root of this number is 5.32611 [note: "28.4" is clearly a rounding to 1 decimal place], as given in the output as the value of s.

R^2 (given as 97.9%) is the percentage of the variation in the Y values that is explained by the fitted model. (Mathematically, R^2 is given by the quotient (SS Regression)/(SS Total).) As R^2 is near to 100%, the model appears to explain the variability in Y very well.

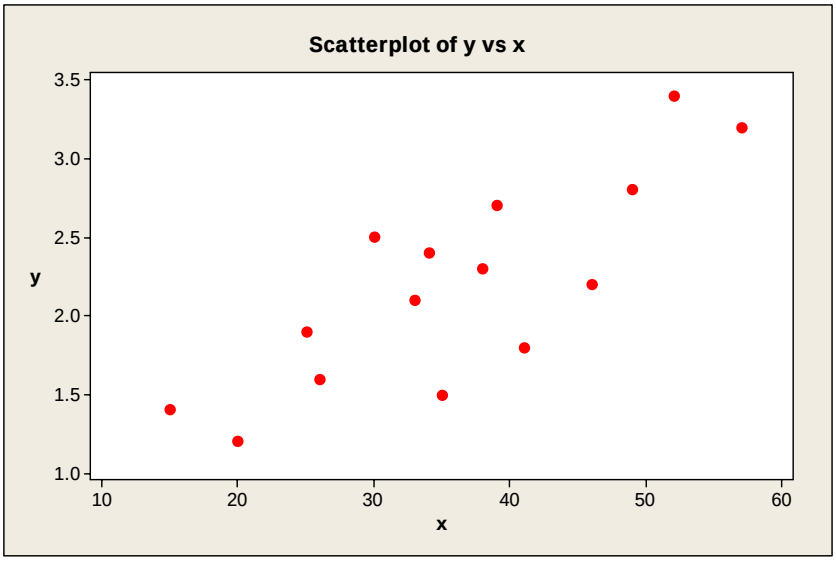
The F value of 300.43 in the analysis of variance is extremely high. It can be referred formally to $F_{2,13}$, and the corresponding p-value (zero to at least 3 decimal places) measures the probability of obtaining such a high level (or higher) of explanation by chance if in fact the true values of both β_1 and β_2 were simultaneously zero.

Putting $x_1 = 200$ and $x_2 = 150$, we find $\hat{Y} = 78.62$ (i.e. £78.62).

Negative predicted prices will be given for sufficiently small carpets, since the constant term is −(£)52.671, and such results are obviously unreasonable. This illustrates the danger of extrapolation, or assuming that the model necessarily holds for values of the predictor variables well away from the observed data used to fit it.

Higher Certificate, Module 4, 2008. Question 4

(i) (a) The scatter diagram is as follows (note that the "false origin" has *not* been corrected in this display). x is temperature and y is thrust.



The data show an approximately linear increasing trend, with some scatter. The context makes it plausible that both x and y measurements are subject to error. This and the linearity make r an appropriate measure of the association between x and y.

(i) (b) There are many equivalent forms of the expression for r. Here we use

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}} = \frac{1276.6 - \frac{540 \times 33}{15}}{\sqrt{\left\{21412 - \frac{540^2}{15}\right\}\left\{78.54 - \frac{33^2}{15}\right\}}} = 0.8186.$$

For the test, we must assume that the underlying population is bivariate Normally distributed.

The Society's *Statistical tables for use in examinations* (Table 8) give that the upper 5% critical point for the required one-sided test with $n = 15$ is 0.4409. As our observed value of r is 0.8186, the null hypothesis is rejected and we have evidence of positive correlation, i.e. an increasing linear relationship (the evidence is in fact very strong; even at the 0.5% level, the critical value of 0.6411 is comfortably exceeded).

Solution continued on next page

(ii) Spearman's rank correlation coefficient should be used. We separately rank the x and y data items, find the difference between the ranks, d, for each pair and then calculate the coefficient using the formula

$$1-\frac{6\Sigma d^2}{n(n^2-1)}.$$

(Note. This may alternatively be calculated as the product-moment correlation coefficient of the paired ranks.)

x	15	19	25	26	30	33	34	35	38	39	40	45	49	52	57
y	1.4	1.2	1.9	1.6	2.5	2.1	2.4	1.5	2.3	2.7	1.8	2.2	2.8	3.4	3.2
Rank of x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Rank of y	2	1	6	4	11	7	10	3	9	12	5	8	13	15	14
d	−1	1	−3	0	−6	−1	−3	5	0	−2	6	4	0	−1	1

Thus $\sum d^2 = 1 + 1 + 9 + \ldots + 1 = 140$, so the value of Spearman's coefficient is $1 - (6\times140)/3360 = 1 - 0.25 = 0.75$.

Again using Table 8, the required critical value for a 5% one-sided test is 0.4464, so the null hypothesis here is rejected. We may conclude that there is evidence for an increasing, not necessarily linear, trend. (Again, the evidence is in fact strong, as the 1% critical point is given in the table as 0.6036.)

The results of the tests based on the product-moment and rank coefficients are similar. This may be expected, since the scatter plot is reasonably linear and also suggests that underlying bivariate Normality is a reasonable assumption.