THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 4

SPECIMEN PAPER A

AND SOLUTIONS

The time for the examination is 1½ hours. The paper contains four questions, of which candidates are to attempt **three**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

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

1. (i) Write down the model for, and standard assumptions of, simple linear regression analysis.

(3)

(ii) (a) Suppose now that the intercept parameter in the regression model is known to be zero, so that the model becomes

$$y_i = \beta x_i + e_i ,$$

where the usual assumptions apply to e_i. Show that the least squares estimator of β is $\dfrac{\sum x_i y_i}{\sum x_i^2}$.

(7)

(b) Over a period of one month, a survey was made on each of ten main roads in a large city. Each road was observed for a one-hour period randomly chosen during the working day. For each road, the mean traffic flow, x_i (in vehicles per minute), and the number of speed limit violations, y_i, $i = 1, 2, \ldots, 10$, were recorded. Plot the data shown below on a suitable graph and comment on the suitability of the above model. Fit the model to the data and hence estimate the expected number of violations on a road with an average traffic flow of 20 vehicles per minute. *Without* any further calculation, comment on the suggestion that an intercept should be included in the model.

Flow, x	5	5	5	10	10	15	25	25	30	50
Violations, y	2	1	1	4	2	5	8	2	5	10

(10)

2. For a study into the density of population around a large city, a random sample of 10 residential areas was selected, and for each area the distance from the city centre and the population density in hundreds per square kilometre were recorded. The following table shows the data and also the log of each measurement.

distance, x (km)	*population density, y*	*log x*	*log y*
0.4	149	–0.916	5.004
1.0	141	0.000	4.949
3.1	102	1.131	4.625
4.5	46	1.504	3.829
4.7	72	1.548	4.277
6.5	40	1.872	3.689
7.3	23	1.988	3.135
8.2	15	2.104	2.708
9.7	7	2.272	1.946
11.7	5	2.460	1.609

(i) By plotting three separate graphs, decide which of the following regressions is best represented by a straight line.

(a) y on x (b) y on $\log x$ (c) $\log y$ on x

(7)

(ii) On the basis of the regression results **on the next page**, which regression do you consider to be best? Justify your answer by reference to the diagnostic criteria given in the output and relating these to your plots in (i). Would you consider regressing $\log y$ on $\log x$? If not, why not?

(5)

(iii) For the model you consider to be best in (ii), obtain an expression for y in terms of x.

(3)

(iv) Using your chosen model, estimate the density of the population at a distance of 5 km from the city centre.

(2)

(v) State any reservations you have about using the model to predict population density.

(3)

The regression results for this question are on the next page

Regression results for question 2

Regression Analysis: y versus x

The regression equation is y = 140 - 14.0x

Predictor	Coef	SE Coef	T	P
Constant	139.70	11.12	12.56	0.000
x	-13.958	1.663	-8.39	0.000

S = 18.2834 R-Sq = 89.8% R-Sq(adj) = 88.5%

Observation 10 has an unusually large positive residual

Regression Analysis: y versus logx

The regression equation is y = 127 - 48.0logx

Predictor	Coef	SE Coef	T	P
Constant	126.990	9.147	13.88	0.000
logx	-47.980	5.293	-9.07	0.000

S = 17.0492 R-Sq = 91.1% R-Sq(adj) = 90.0%

Observation 1 has an unusually large negative residual

Regression Analysis: logy versus x

The regression equation is logy = 5.41 - 0.322x

Predictor	Coef	SE Coef	T	P
Constant	5.4133	0.1621	33.40	0.000
x	-0.32157	0.02425	-13.26	0.000

S = 0.266544 R-Sq = 95.6% R-Sq(adj) = 95.1%

3. Three types of watch dial were tested on 21 subjects under simulated conditions. One dial was assigned at random to each subject and the number of errors the subject made in reading this dial during a standardised series of tests was recorded. The results are shown in Table 1 below.

Table 1

Dial type		
1	*2*	*3*
42	62	56
30	53	36
21	61	43
47	47	58
34	45	46
22	59	24
42		31
38		

Incomplete results of a one-way analysis of variance of the data are shown in Table 2.

Table 2

One-way ANOVA: type 1, type 2, type 3

```
Analysis of Variance
Source     DF        SS
Dial type   2      1377
Error      18      1858
Total      20      3234
```

(i) Complete the analysis and interpret your results, stating any assumptions you have made in reaching a conclusion.

(6)

(ii) Estimate the difference between the mean numbers of errors that would be made by subjects reading dials of type 1 and of type 2, and find a 95% confidence interval for this difference. Explain what is meant by describing your interval as a "95% confidence interval". You may take it that any assumptions needed for your analysis are satisfied.

(10)

(iii) How could you investigate the assumptions needed for the one-way analysis of variance of the data in Table 1? If you were unwilling to accept these assumptions, explain briefly how you might proceed. (Do not actually do so.)

(4)

4. (i) Two explanatory variables x_1, x_2 are used to predict a dependent variable Y. Write down the linear model which can be used as a basis for the analysis, and explain carefully the meaning and properties of the terms in it.

(5)

(ii) The data in the following table show the price Y (£) for double-glazed windows of width x_1 (cm) and height x_2 (cm).

x_1	66	183	124	239	66	109	196	251	165	81	249	142	170	254
x_2	122	122	122	122	30	58	61	76	86	81	117	61	30	30
Y	66	78	75	90	45	57	73	83	71	59	95	61	53	64

(a) Draw a graph of Y plotted against x_1, and a graph of Y plotted against x_2. Discuss briefly what information these graphs give.

(5)

(b) An extract from computer output of the analysis of Y on x_1 and x_2 is as follows.

```
Predictor      Coef     StDev        t          p
Constant     24.823     2.909     8.53      0.000
X1          0.13800   0.01288    10.72      0.000
X2          0.27226   0.02392    11.38      0.000

s = 3.132     Rsq = 95.9%

Analysis of variance

Source       df       SS       MS        F        p
Regression    2   2515.0   1257.5   128.2   0.000
Residual     11    107.9      9.8
Total        13   2622.9
```

Explain this output fully, in suitable terms for a non-expert, and write down the regression equation of Y on x_1 and x_2.

(7)

(c) A simple linear regression of Y on x_1 only gives a sum of squares for regression of 1244.0. Explain what further information can be obtained from this.

(3)

SOLUTIONS

Question 1

(i) $Y_i = a + bx_i + e_i, \quad i = 1, 2, \ldots, n.$

The $\{e_i\}$ are uncorrelated random variables with mean 0 and constant variance σ^2.

[If the analysis is to proceed to *inference* for the regression coefficients, Normality of the $\{e_i\}$ is required. This is taken to be the case in question 4 in this paper.]

(ii)(a) For $Y_i = \beta x_i + e_i$, we minimise $S = \sum e_i^2 = \sum (y_i - \beta x_i)^2$.

We have $\dfrac{dS}{d\beta} = -2\sum x_i (y_i - \beta x_i)$ which on setting equal to zero gives $\sum x_i y_i = \beta \sum x_i^2$, so the least squares estimate is $\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$.

[Consideration of $\dfrac{d^2S}{d\beta^2}$ confirms that this is a minimum: $\dfrac{d^2S}{d\beta^2} = 2\sum x_i^2 > 0$.]

(b) See scatter plot at foot of page. It shows an increasing trend, roughly linear; but there seems to be some increase in variability as x increases. There are not enough data points to be sure.

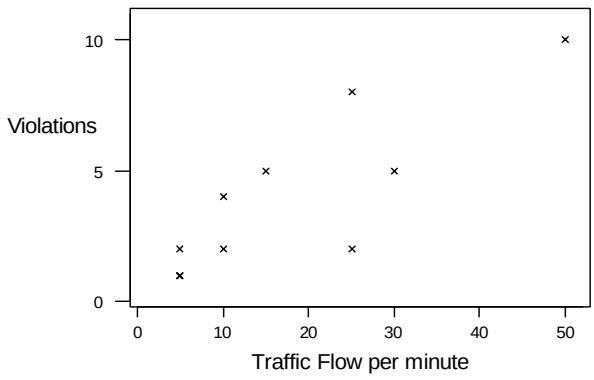
The usual summary statistics (not all required for the zero intercept model) are

$$n = 10, \ \Sigma x_i = 180, \ \Sigma y_i = 40, \ \Sigma x_i^2 = 5150, \ \Sigma y_i^2 = 244, \ \Sigma x_i y_i = 1055.$$

$\therefore \hat{\beta} = 1055/5150 = 0.205$. So the fitted line is $y = 0.205x$.

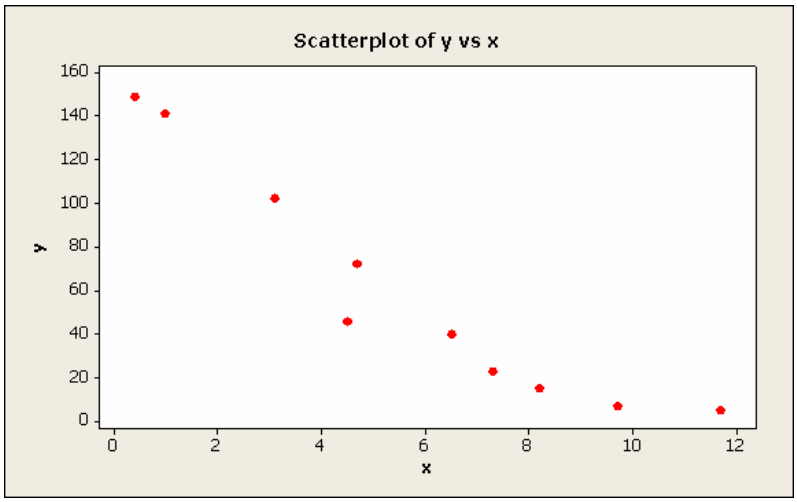
Hence the estimated expected number of violations for $x = 20$ is $0.205 \times 20 = 4.1$.

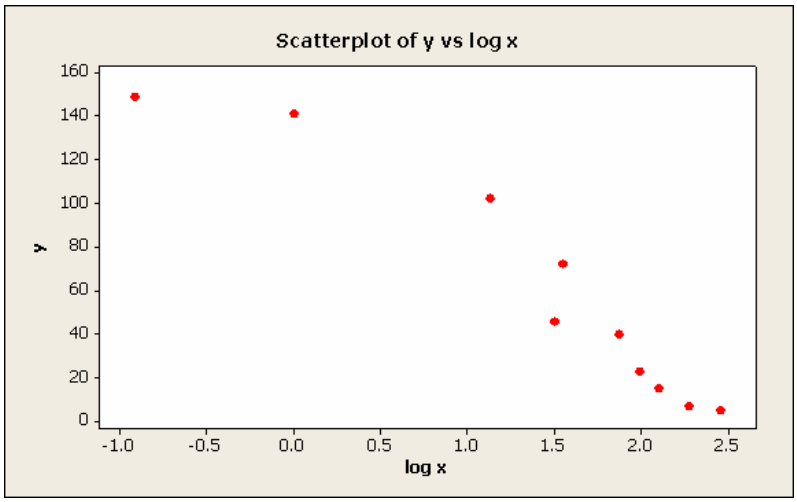
Logically, zero traffic flow should imply zero speed violations, so that y should be 0 when x is 0, i.e. the zero intercept model seems reasonable. The scatter plot does not contradict this.

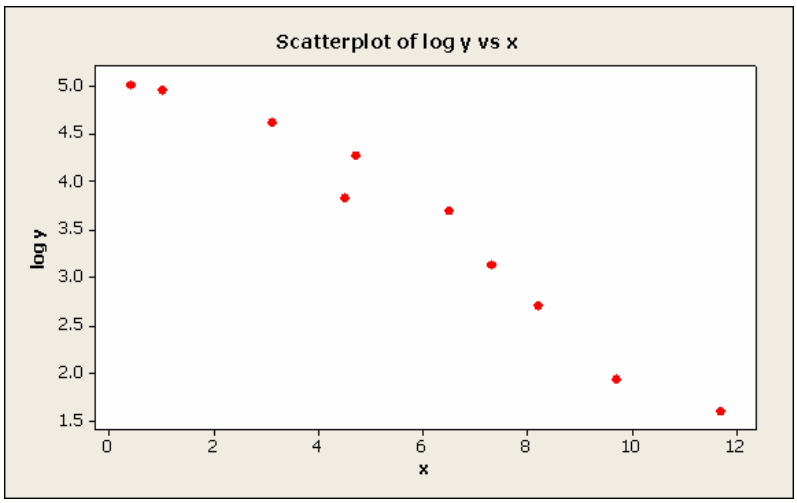


Question 2

(i)







Solution continued on next page

There is some curvature in all these plots; log y on x is slightly "straighter" than y on x. Using log x looks worst. So use log y on x.

(ii) (c) has the highest R^2 – though all are good.

(c) also has the highest t values for the coefficients – though again all are good.

(c) also has the lowest residual variance relative to the mean response.

(c) is the only one without a "large" residual.

(c) appeared (marginally) the best plot.

It does not seem sensible to regress log y on log x; it looks as if this would increase the curvature.

(iii) Using (c), we have $\log y = 5.41 - 0.322x$, so

$$y = \exp\left(5.41 - 0.322x\right) = e^{5.41}e^{-0.322x} = 223.63e^{-0.322x}.$$

(iv) From the expression in part (iii), inserting $x = 5$, $y = 223.63e^{-1.61} = 44.7$. This is in hundreds per square kilometre. So the estimate is 4470 per square kilometre.

(v) Prediction within the range of the data may be adequate, except perhaps near the upper end because of the tendency for curvature there. Extrapolation to values of x outside the data will, for similar reasons, be unreliable, and the linear model is likely to underestimate density. Where is the next city or town centre? Interaction with that is very likely unless it is a long distance away. There may also be directional effects, i.e. densities changing more or less slowly according to the direction from the centre.

Question 3

Summary:	Dial type	1	2	3	
	Total	276	327	294	
	Number of tests n_i	8	6	7	Total 21
	Mean number of errors $\bar{x}_i$	34.5	54.5	42.0	

(i) Analysis of variance

Source of variation	df	SS	Mean Square	*F* ratio
Dial type	2	1377	688.50	6.67
Residual (Error)	18	1858	103.22	
Total	20	3234		

The F ratio of 6.67 is referred to $F_{2,18}$ and is very highly significant (the upper 1% point is 6.01). We reject the null hypothesis that the mean numbers of errors with the three dial types are all the same. We deduce that at least one mean is different from the other two.

We have assumed that all sets of data come from Normally distributed populations with the same variance σ^2.

(ii) We have $\bar{x}_2 - \bar{x}_1 = 20.0$, and the standard error of this estimate is $\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}$

$= \sqrt{103.22\left(\frac{1}{8} + \frac{1}{6}\right)} = 5.487$. [Thus the difference between the means for dial types 1 and 2 is very highly significant: test statistic is 20.0/5.487 = 3.645, refer to t_{18}.] The two-sided 5% critical value for t_{18} is 2.101, so a 95% confidence interval for the true population mean difference $\mu_2 - \mu_1$ is given by

$$20.0 \pm (2.101 \times 5.487) \quad \text{or} \quad 20.0 \pm 11.53, \quad \text{i.e. } (8.47, 31.53).$$

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 $\mu_2 - \mu_1$.

(iii) Residuals could be calculated for the 21 observations, and their pattern studied either as a Normal probability plot or by plotting residuals against fitted values.

The variances within the three dial types could be checked for equality, but no good, sensitive, test exists for small amounts of data such as we have here.

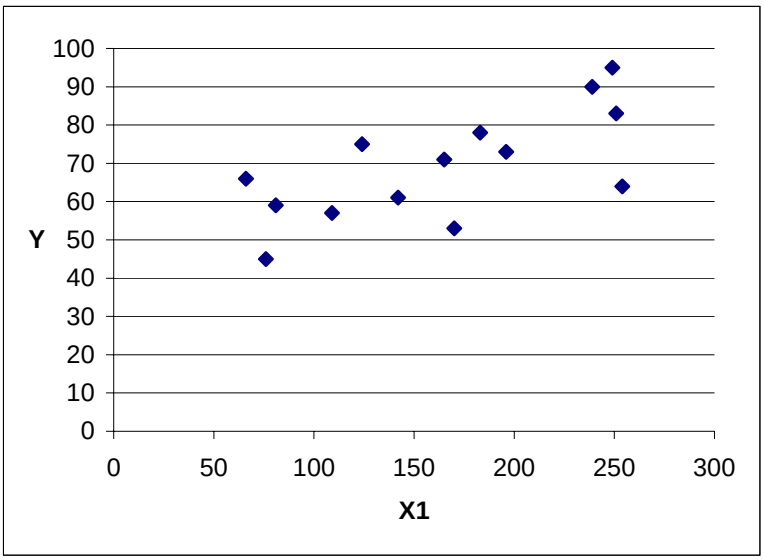
Outliers, if any, could be checked for possible recording error or change in background conditions. Any outliers could be removed from the data before re-doing an analysis.

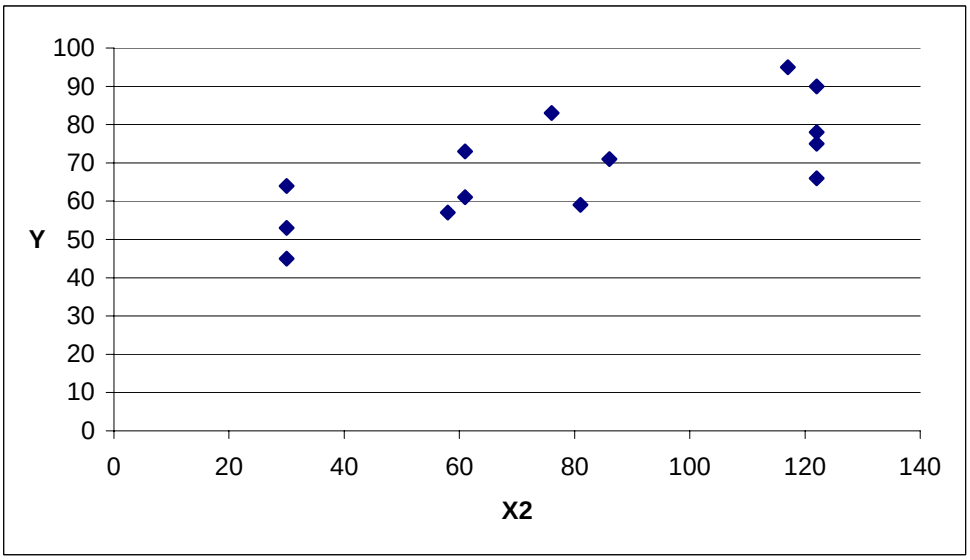
Sometimes a transformation (such as log) will make data behave more like data from Normal homoscedastic distributions.

Question 4

(i) $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ where $\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$. $\beta_0, \beta_1, \beta_2$ are parameters (constants) to be estimated from the data. The set $\{\varepsilon_i\}$ are independent Normally distributed residuals ("errors") with mean 0 and common variance σ^2. β_0 is the overall mean response, and β_1, β_2 show the increase in Y for unit increases in x_1, x_2 respectively with the other x variable kept constant.

(ii)(a)





From the graphs, there appears to be a trend for Y to increase, roughly linearly, as either x_1 or x_2 increases, but there is considerable variability. A regression on one of x_1 and x_2 is not likely to be very good, but it is worth trying multiple regression on x_1 and x_2 together.

Solution continued on next page

(b) The coefficients of the "predictors" are $\beta_0, \beta_1, \beta_2$ so the fitted equation is

$$Y = 24.82 + 0.1380\,x_1 + 0.2723\,x_2.$$

The standard deviations of each estimated coefficient are given in the computer output, as are the results of t tests to examine the hypotheses "$\beta_0 = 0$" etc. The t values here are all very high. They are formally tested by reference to t_{11} because the residual df is 11. The p-values (all zero to at least three decimal places) are given. These mean that, on the null hypothesis "$\beta_j = 0$" (for $j = 0, 1, 2$), each of these tests gives a value that has very small probability indeed of being obtained (or exceeded). Hence all of $\beta_0, \beta_1, \beta_2$ need to remain in the model.

The residual mean square given in the analysis of variance is 9.8. This is our estimate of σ^2, the experimental error. The square root of 9.8 is 3.132; this is also given in the output, as the value of s.

R^2 (given as 95.9%) is the proportion of total variation explained by the model, (SS regression)/(SS total). (R is often called the "coefficient of determination".) This is a very high proportion (percentage), indicating that the model with x_1 and x_2 included is a very good explanation of the data.

The F value in the analysis of variance table, 128.2, can be referred formally to $F_{2,11}$. It is a very large value, very highly significant (the p-value is zero to at least three decimal places). This is very unlikely to have occurred if x_1 and x_2 do <u>not</u> together "predict" Y, thus giving another indication that the model with both of them is a good one.

The value for the constant, 24.823, is a "baseline" cost, perhaps for general overheads.

(c) When only x_1 is used, only $1224.0 / 2622.9 = 47.43\%$ of the total variation in Y is explained. Adding x_2 is essential for a reliable prediction.

[Notes. (1) If only x_2 is used, the results are quite similar to these: x_1 needs to be added too.

(2) The product x_1x_2 could be used as a single predictor, being area of glass (this proves to be not so good)].