EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2011

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 7 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 4 questions altogether in the paper.

1. (i) State the model and standard assumptions for one-way analysis of variance, defining clearly all your notation.

(4)

(ii) A trainee statistician at an educational research centre is asked to analyse data from a trial of three different methods of teaching a school mathematics course. 15 children of approximately equal ability and mathematical knowledge are randomly allocated to the methods, five to be taught by each method. At the end of the course, the children are tested on what they have learnt, their marks obtained in the test being as follows.

Method	*Marks obtained*					*Row Totals*	*Row Sums of Squares*
A	0	17	12	15	16	60	914
B	23	24	15	20	18	100	2054
C	17	16	25	21	21	100	2052

Briefly comment on the data in the light of the assumptions of part (i).

Making these assumptions, test at the 5% significance level the null hypothesis that the teaching methods do not differ in mean effectiveness, and state your conclusions clearly.

(9)

(iii) The trainee is later told that the score of zero is wrong, as the child in question was ill on the day of the test and did not take it. Reanalyse the data without the zero score, and give your final conclusions.

(7)

Turn over

2. State the model and standard assumptions for simple linear regression analysis, defining clearly all your notation.

(4)

The organisation and methods unit of a company is reviewing its in-house consulting service to other sections of the business. It is particularly interested in the relationship between invoiced cost and job duration. A random sample of consulting service projects is taken from the company's records, yielding data for job duration x in hours and invoiced cost y in £ as follows.

x	1	1.5	3	3.5	3.5	4.5	5	6	8
y	40	74	80	140	180	220	175	209	331

You are given that $\Sigma x = 36$, $\Sigma y = 1449$, $\Sigma x^2 = 182$, $\Sigma y^2 = 297\,743$, $\Sigma xy = 7278$.

(i) Plot a scatter diagram of the data and comment briefly.

(3)

(ii) Calculate the least squares regression equation of y on x.

(5)

(iii) Calculate the residual mean square error for this regression equation and hence test the regression slope for significance at the 5% level.

(6)

(iv) Without carrying out any further calculations, comment on the adequacy of this model.

(2)

3. In the processing of flax into yarn, it is necessary to expose it to moisture for a suitable period to soften and separate the fibrous core of the plant from the outer layer. This process is known as "retting". An experiment was conducted to investigate the effects of the three explanatory variables mean daily rainfall (x_1), retting period in days (x_2) and mean daily temperature during retting period in degrees Fahrenheit (x_3) on the percentage ret loss of flax (y). Annual observations were taken from past growing seasons, and edited computer output of some regression analyses of the results is given **on the next two pages**.

(i) Comment on the success of the fit of Model 1 and test the whole regression of Model 1 for significance at the 5% level.

(6)

(ii) Calculate the partial t values for the three explanatory variables and the constant term in Model 1, and hence test each of these for significance.

(5)

(iii) Compare and contrast the results of fitting Model 2 with the results for Model 1, and state with a reason which of these models you prefer.

(6)

(iv) Use the output for each model to calculate a point estimate of the true percentage ret loss of flax when $x_1 = 4.5$, $x_2 = 70$ and $x_3 = 68$, and comment briefly.

(3)

The computer output begins on the next page

Model 1

Regression Analysis of y on x1, x2 and x3

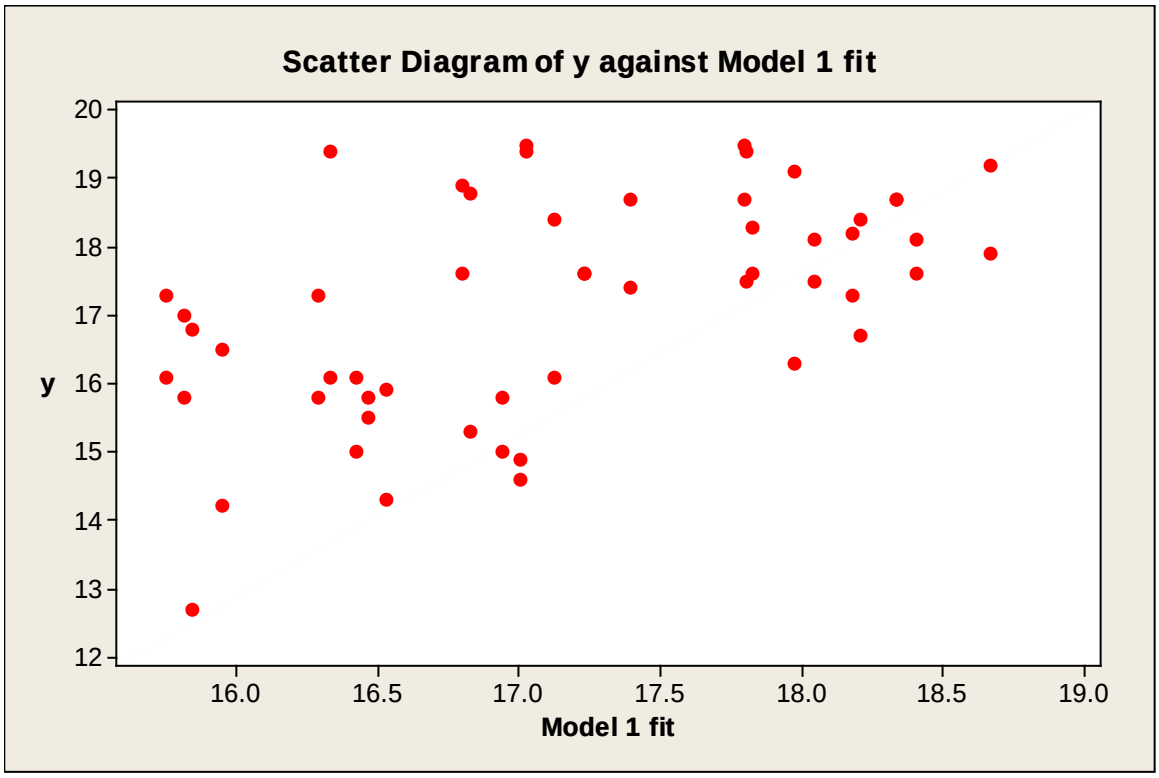
The regression equation is y = 5.41 + 0.905 x1 + 0.0544 x2 + 0.0463 x3

Predictor	Coef	SE Coef
Constant	5.409	3.247
x1	0.9051	0.2439
x2	0.05444	0.02621
x3	0.04632	0.04159

S = 1.39020 R-Sq = 29.6%

Analysis of Variance

Source	DF	SS	MS
Regression	3	40.642	13.547
Residual Error	50	96.632	1.933
Total	53	137.275	



Turn over

Model 2

Regression Analysis of y on x1 and x2

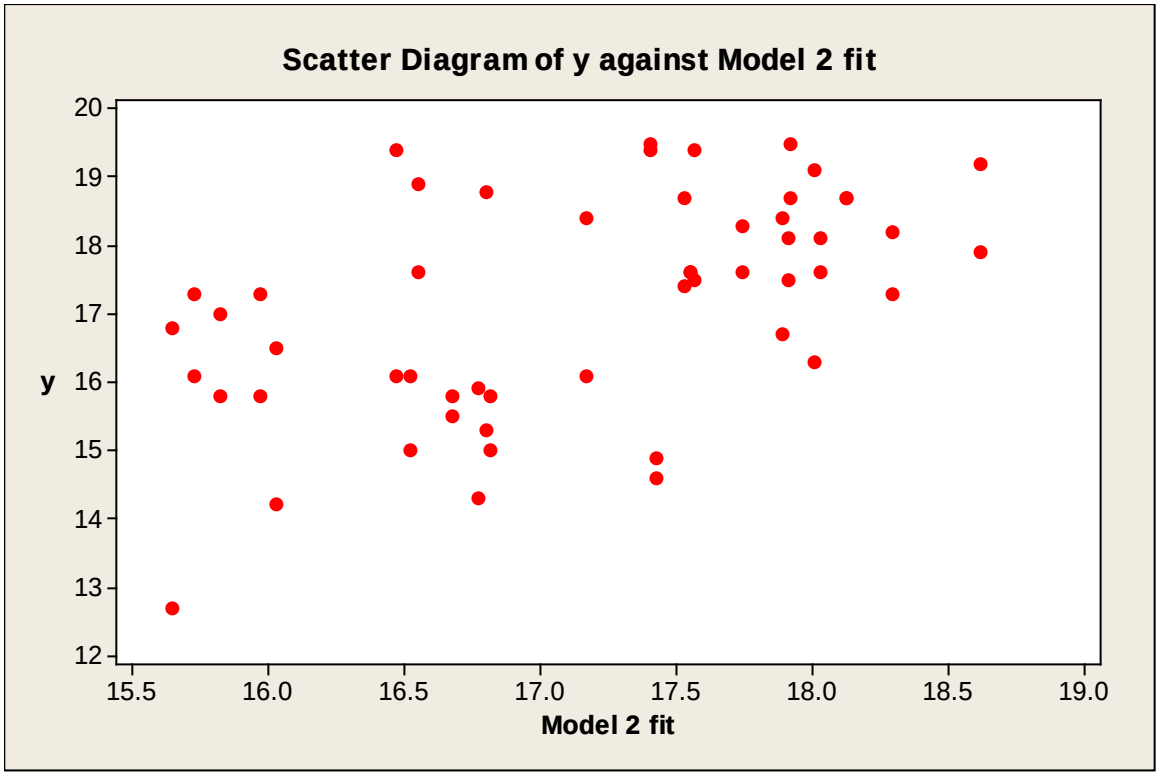
The regression equation is y = 8.21 + 0.793 x1 + 0.0703 x2

Predictor	Coef	SE Coef	P
Constant	8.208	2.062	0.000
x1	0.7926	0.2225	0.001
x2	0.07029	0.02206	0.002

S = 1.39347 R-Sq = 27.9% R-Sq(adj) = 25.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	38.245	19.123	9.85	0.000
Residual Error	51	99.030	1.942		
Total	53	137.275			



Turn over

4. In a traffic survey of a busy road, observations are made of the speeds of vehicles v (in kilometres per hour) and their respective distances s (in metres) from the next vehicle in front. You may assume that successive pairs of observations are sufficiently separated in space and time that they may be taken as independent. The results are shown in the following table.

v	8	9	11	14	17	18	22	24	27	30
s	10	9	17	16	35	32	51	110	104	200

(i) Plot a scatter diagram of the data, and comment on the nature of the relationship between v and s.

(5)

(ii) You are given that

$$\sum v = 180, \quad \sum s = 584, \quad \sum v^2 = 3764, \quad \sum vs = 14\,313, \quad \sum s^2 = 68\,492.$$

Use these results to calculate the sample product-moment correlation coefficient between v and s. Test at the 2.5% level the null hypothesis of zero population correlation between speed and distance, H_0: $\rho_{vs} = 0$, against an appropriate one-sided alternative hypothesis H_1 to be stated. Comment on your results, mentioning any reservations you may have about this analysis.

(7)

(iii) Rank the data in the table, calculate Spearman's rank correlation coefficient between v and s, and similarly test the null hypothesis of zero association between v and s against an appropriate one-sided alternative. Compare your result with that of part (ii), and say with a reason which analysis you prefer.

(8)