EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2013

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 2.

There are 4 questions altogether in the paper.

1. (a) For what experimental design is the following model appropriate?

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad j = 1, ..., r; \quad i = 1, ..., k.$$

Explain the meaning of each of the symbols $y_{ij}, \mu, \alpha_i, e_{ij}, r, k$.

(8)

(b) A new drug for reducing blood pressure (BP) in hypertensive patients is administered over a two-week period to three groups, each of four patients, and the reductions in BP (in millimetres of mercury) are noted. The three groups are labelled L, M and H, indicating low, medium and high dose levels of the drug respectively. The results of the trial are given in the following table.

Dose Level	*Reduction in BP*	*Row Total*
L	8, 6, 3, 3	20
M	11, 7, 10, 8	36
H	15, 13, 11, 13	52

The sum of squares of the observations is 1136.

(i) Test at the 5% significance level whether the mean reduction in BP is the same for the three dose levels. State any assumptions you make, and report your conclusions in terms that a non-statistician would understand.

(10)

(ii) You are now given that the dose levels of the drug used in groups L, M and H are 5, 10 and 15 mg respectively. Without making any further calculations, identify a new model which takes account of this information and might be useful for further analysis.

(2)

2. (i) State the standard model and assumptions for simple linear regression analysis, defining any notation that you use.

(5)

(ii) (a) A flame photometer is designed to register the sodium concentration in a sample of material as a scale reading. 10 samples of material were made up with preassigned sodium contents (in milligrams per litre, or mg/l) and these values (x) and the resulting readings (y) are shown in the table below.

x	50	50	100	100	150	150	200	200	250	250
y	21	19	43	36	53	56	84	81	102	95

Plot the data on a scatter diagram. Use this diagram to comment on the suitability of linear regression analysis for predicting y from x.

(6)

(b) You are given that

$\Sigma x = 1500$, $\Sigma y = 590$, $\Sigma x^2 = 275\,000$, $\Sigma y^2 = 42\,938$, $\Sigma xy = 108\,500$.

Calculate the least squares regression line of y on x.

(4)

(c) Estimate the error variance σ^2, and hence test the regression for significance at the 5% level.

(3)

(d) Calculate a point estimate for the photometer reading corresponding to a sodium content of 200 mg/l. A physicist asks you to estimate the photometer reading corresponding to a sodium content of 300 mg/l. How would you answer him?

(2)

3. A manufacturer of luxury cosmetics has recently put a new product on the market. This product is initially being offered at a wide range of prices, and the company has made a survey of its sales y (in 100s) and prices x (in £) across a random sample of stores in which it is sold. It wishes to examine whether, on the whole, increased price is associated with reduced sales. The results are shown in the following table.

Store	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Price x (£)	27	30	37	47	55	62	70	80	95	99
Sales y (100s)	110	79	69	48	51	44	29	32	26	30

(i) Plot the data on a scatter diagram and comment on the relationship, if any, between x and y.

(4)

(ii) A research assistant suggests calculating the product-moment correlation coefficient, r, between sales and prices. Carry out this calculation and test at the 1% significance level the null hypothesis of zero correlation against an appropriate one-sided alternative. You are given that

$$\Sigma x = 602, \quad \Sigma x^2 = 42\,202, \quad \Sigma y = 518, \quad \Sigma y^2 = 33\,384, \quad \Sigma xy = 25\,712.$$

(6)

(iii) A statistician in the market research department suggests calculating instead Spearman's rank correlation coefficient, r_s. Calculate r_s for these data, and test at the 1% level the null hypothesis of no association between prices and sales against an appropriate two-sided alternative.

(6)

(iv) Comment on the tests used in parts (ii) and (iii), stating with a reason which you prefer.

(2)

(v) Given that each item costs £15 to make, suggest without further calculation how the data might be analysed with a view to maximising profit.

(2)

4. An empirical relationship of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

is suggested for predicting the reading of an aneroid barometer. In this formula,

y = (aneroid barometer reading of atmospheric pressure in mm of mercury) – 740;

x_1 = (mercury barometer reading of atmospheric pressure in mm of mercury) – 740;

x_2 = temperature in degrees Celsius;

x_3 = relative humidity (%);

e is an error term.

10 observations of y, x_1, x_2 and x_3 were made, and edited computer output of an analysis of this relationship, and possible simplifications of it, is shown on the following **two** pages.

(i) Explain what is meant by the statement `'R-Sq = 99.8%'` in the output for Model 1 and comment on the scatter diagram provided for Model 1.

(3)

(ii) Carry out a partial t test of the significance at the 5% level of x_2 in the regression for Model 1.

(3)

(iii) Carry out an F test at the 5% level of the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ in the regression for Model 1.

(2)

(iv) Carry out a partial t test of the significance at the 5% level of x_3 in the regression for Model 2.

(2)

(v) Test at the 5% level the hypothesis $H_0 : \beta_1 = 1$ in Model 3.

(3)

(vi) The F statistic for the regression for Model 3 exceeds that for Model 2 although Model 3 only uses one of the regressor variables in Model 2. What can be said about the relative merits of Models 2 and 3?

(3)

(vii) State with reasons which of Models 1, 2 and 3 you regard as providing the most satisfactory fit to the data.

(4)

The computer output begins on the next page

Model 1: Regression of y on x1, x2 and x3

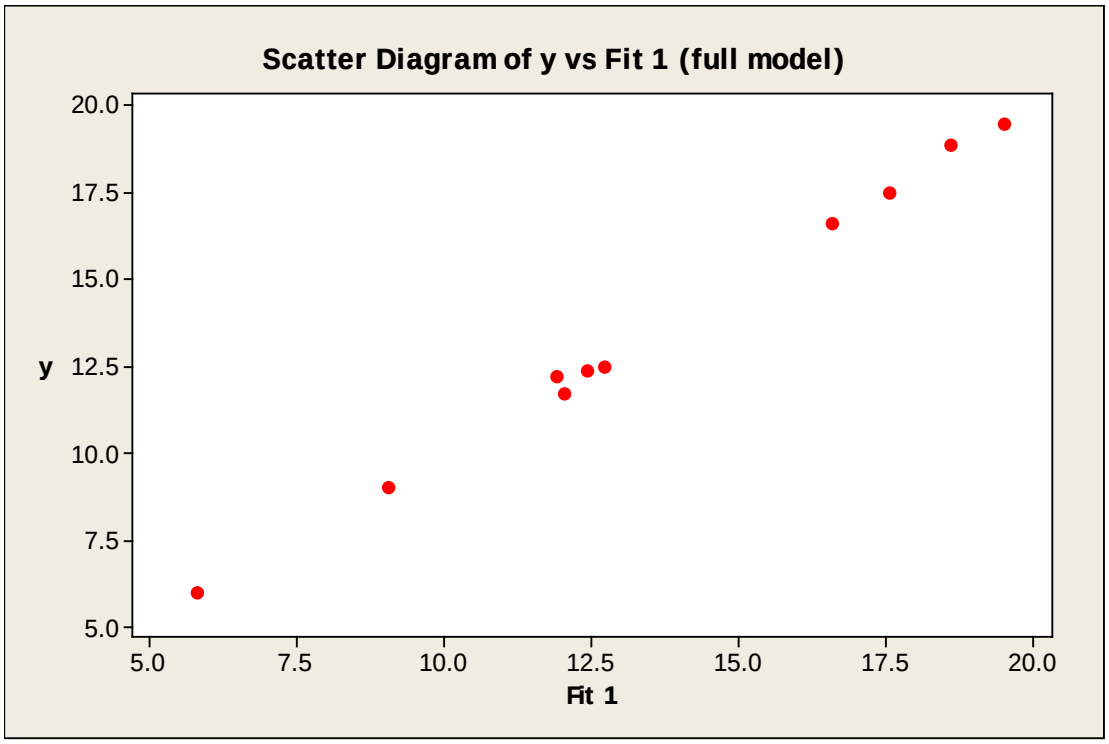
The regression equation is
ŷ = 3.88 + 0.946x1 + 0.0190x2 + 0.0117x3?(y = Fit 1, full model)

Predictor	Coef	SE Coef
Constant	3.8842	0.4392
x1	0.94594	0.01859
x2	0.01896	0.01205
x3	0.011740	0.005664

R-Sq = 99.8%

Analysis of Variance

Source	DF	SS	MS
Regression	3	173.855	57.952
Residual Error	6	0.386	0.064
Total	9	174.241	



Model 2: Regression of y on x1 and x3

The regression equation is
$\hat{y} = 4.30 + 0.943x1 + 0.00828x3$

Predictor	Coef	SE Coef
Constant	4.3017	0.3851
x1	0.94327	0.02037
x3	0.008276	0.005742

R-Sq = 99.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	173.696	86.848	1114.99	0.000
Residual Error	7	0.545	0.078		
Total	9	174.241			

Model 3: Regression of y on x1

The regression equation is
$\hat{y} = 4.77 + 0.937x1$

Predictor	Coef	SE Coef	T	P
Constant	4.7693	0.2210	21.58	0.000
x1	0.93664	0.02114	44.31	0.000

R-Sq = 99.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	173.53	173.53	1963.53	0.000
Residual Error	8	0.71	0.09		
Total	9	174.24			

BLANK PAGE