EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2010

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* ${}^nC_r$.

This examination paper consists of 8 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. (a) (i) Briefly explain the principles of *randomisation* and *replication*, in the context of a completely randomised experimental design. Your answer should make reference to an example which is different from the application given in part (b).

(4)

(ii) Write down the model equation for a completely randomised design having equal numbers of replicates in all treatment groups, defining all the symbols that you use.

(4)

(b) (i) Observations are taken to assess whether the tensile strength (TS) of a synthetic fibre depends on the percentage of cotton in the fibre. Five samples of fibre are tested at each of four levels of this percentage, with the following results.

% cotton	*TS (grams per mm^2)*					*Row Totals*	*Row Sums of Squares*
15	7	7	15	11	10	50	544
20	10	17	12	18	18	75	1181
25	14	20	13	16	22	85	1505
30	19	25	16	19	21	100	2044

Carry out an analysis of variance of these data and report your conclusion in terms that a non-scientist would understand.

(7)

(ii) An assistant suggests that a regression analysis of the data might be useful. Give a reason for this suggestion. Outline without further calculation how the analysis would proceed, and indicate what information might be gained from it.

(5)

2. A certain metal discolours when exposed to air. To protect the metal against discoloration, it is coated with a chemical. In an experiment, coatings of varying thickness, x mm, of the chemical were applied to standard samples of the metal, and the times, t hours, for the metal to discolour were noted. The results are as shown.

x	1.8	3.0	4.0	5.7	7.2	8.4	10.3
t	3.4	5.9	7.0	8.7	9.5	10.4	11.1

(i) You are given that the least squares regression line for these data is

$$t = 3.027 + 0.8617x.$$

Draw a scatter diagram of the data. Plot this regression line on your diagram and comment on the appropriateness of a simple linear regression model for the dependence of t on x.

(5)

(ii) A researcher suggests that the theoretical relationship between t and x should be of the form

$$\exp(t) = Ax^B,$$

where A and B are constants. Show that this relationship may be expressed in the form

$$t = a + b\log x,$$

where a and b are functions of A and B respectively, which you should identify.

(2)

(iii) You are given that

$$\Sigma \log x = 11.2476, \quad \Sigma t = 56, \quad \Sigma (\log x)^2 = 20.3687, \quad \Sigma t \log x = 100.101.$$

Use these results to calculate the least squares regression line of t on $\log x$, and plot this line and the data on a scatter diagram with values of $\log x$ on the horizontal axis.

(6)

(iv) State with reasons which model you prefer. For each of the two models, calculate the predicted value of t when $x = 6$, and comment briefly.

(7)

3. (a) You are given a set of bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ which can be assumed to be a random sample from a bivariate population. Define the sample product-moment correlation coefficient. Write down a formula for Spearman's (sample) rank correlation coefficient that has a similar form to that for the sample product-moment correlation coefficient. Explain the circumstances in which each of these quantities is useful.

(5)

(b) The amounts of excise duty (in pence per litre) levied on unleaded petrol and diesel in 10 European countries are given in the following table (dated May 2008).

Country	*Unleaded (x)*	*Diesel (y)*
Austria	36	28
Denmark	42	29
Estonia	23	19
Germany	52	37
Greece	26	22
Hungary	30	25
Italy	45	34
Poland	33	23
Spain	31	24
United Kingdom	57	57

You are given that

$$\Sigma x = 375, \quad \Sigma y = 298, \quad \Sigma x^2 = 15\,193, \quad \Sigma y^2 = 9974, \quad \Sigma xy = 12\,191.$$

(i) Construct a scatter diagram of y against x and comment on the relationship, if any, between y and x.

(4)

(ii) Calculate the product-moment correlation coefficient for these data, and test at the 1% significance level the null hypothesis that x and y are uncorrelated, against the alternative of a positive correlation.

(4)

(iii) Calculate Spearman's rank correlation coefficient for the above data. Use it to test, at the 1% significance level, the null hypothesis that there is no association between x and y in the underlying population, against the alternative of positive association.

(4)

(iv) Comment on the results of parts (ii) and (iii) and say with a reason which analysis you think is better here. Mention any reservations you may have about this analysis.

(3)

4. (i) In the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad i = 1, \ldots, n,$$

what assumptions are usually made about the residual (error) terms $e_1, \ldots, e_n$?

(3)

(ii) The following **three** pages of edited computer output show three regression analyses (Models A, B and C) of Computer Aptitude Score (y) on one or both of Verbal Ability (x_1) and Arithmetic Ability (x_2) using the data below.

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Mean*
y	22	24	34	48	58	76	82	96	55
x_1	53	52	58	60	57	61	66	65	59
x_2	50	45	51	55	62	70	73	74	60

Use the output to answer the following questions.

(a) Briefly comment on the scatterplots of observed against fitted y for the three models. Relate your comments in each case to S, the square root of the mean square error.

(3)

(b) In Models B and C, test for the significance of the coefficients of x_1 and x_2 respectively.

(3)

(c) In Model A, test at the 5% level for the significance of x_1 in the presence of x_2, and for the significance of x_2 in the presence of x_1. Also test for the global significance of the regression on x_1 and x_2. Interpret the statement "R-Sq = 96.7%" and explain how this quantity is calculated.

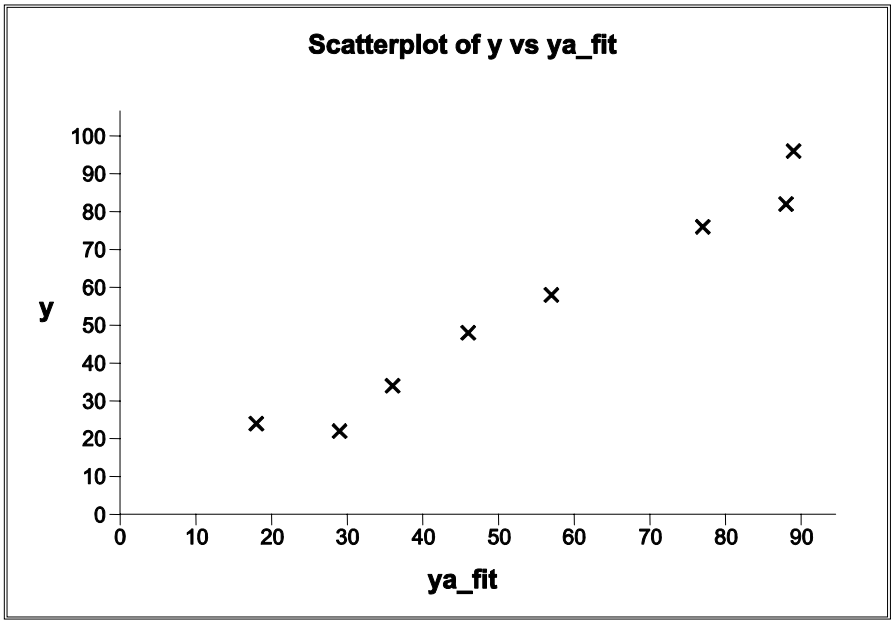
(8)

(d) Which of the three models do you consider best describes the data? Justify your answer.

(3)

The computer output begins on the next page

Model A



Regression Analysis: y versus x1, x2

The regression equation is y = -124 + 1.00 x1 + 2.00 x2

Predictor	Coef	SE Coef
Constant	-124.00	37.06
x1	1.0000	1.0000
x2	2.0000	0.4472

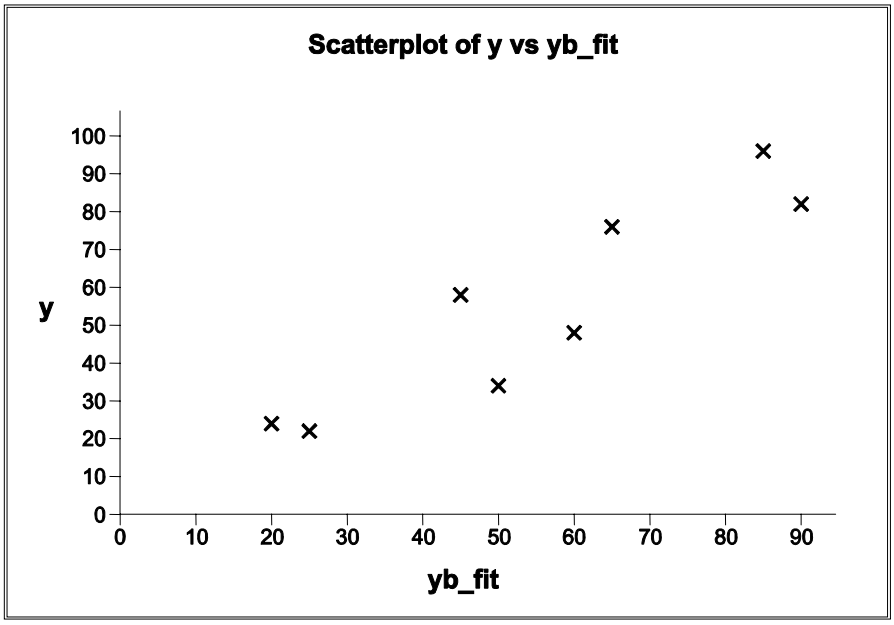
√(mean square error) is S = 6.0000; R-Sq = 96.7%

Analysis of Variance

Source	DF	SS	MS
Regression	2	5220.0	2610.0
Residual Error	5	180.0	36.0
Total	7	5400.0	

Turn over

Model B



Regression Analysis: y versus x1

The regression equation is y = -240 + 5.00 x1

Predictor	Coef	SE Coef
Constant	-240.00	54.03
x1	5.0000	0.9129

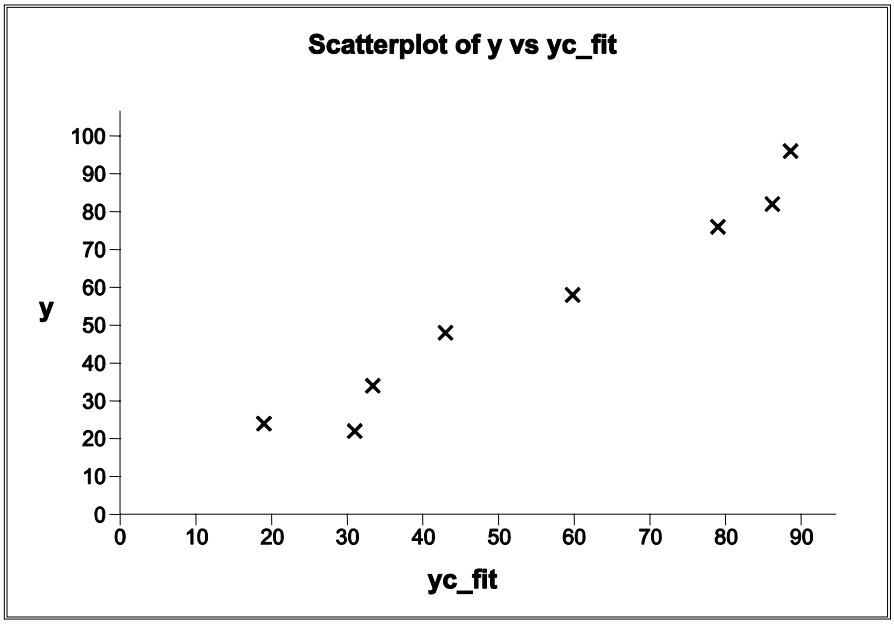
√(mean square error) is S = 12.2474; R-Sq = 83.3%

Analysis of Variance

Source	DF	SS	MS
Regression	1	4500.0	4500.0
Residual Error	6	900.0	150.0
Total	7	5400.0	

Turn over

Model C



Regression Analysis: y versus x2

The regression equation is y = -89.0 + 2.40 x2

Predictor	Coef	SE Coef
Constant	-89.00	12.19
x2	2.4000	0.2000

√(mean square error) is S = 6.0000; R-Sq = 96.0%

Analysis of Variance

Source	DF	SS	MS
Regression	1	5184.0	5184.0
Residual Error	6	216.0	36.0
Total	7	5400.0	