EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2012

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) Give the principal features of a *balanced completely randomised design*, and explain the role of replication in such a design. State the statistical model for this design, define the terms in the model and state the standard assumptions made about the error term.

(7)

(ii) In a completely randomised trial of three fertilisers based respectively on nitrogen, phosphorus and potash (N, P and K), each fertiliser was applied to four plots of winter wheat. At the time of harvest, the yields of wheat in kg from the 12 plots are noted, the results being as follows.

Fertiliser	*Yields of wheat*	*Row total*
Nitrogen (N)	15, 13, 11, 13	52
Phosphorus (P)	9, 10, 5, 8	32
Potash (K)	12, 7, 9, 8	36

You are also given that the sum and sum of squares of all twelve observations are 120 and 1292 respectively.

(a) Carry out an analysis of variance of these data, and test at the 5% level of significance the null hypothesis that there are no mean differences between the effects of N, P and K on the yield.

(10)

(b) Obtain a 95% confidence interval for the true mean difference between the effects of N and P on the yield, explaining your method clearly.

(3)

2. (i) State the standard model and assumptions for a multiple linear regression with two regressor variables. Explain the meaning of all the terms in the model.

(4)

(ii) In a hydrological study, observations are made for 13 small rivers of one year's rainfall (x_1, in mm), catchment area (x_2, in km^2) and average runoff (y, in thousands of litres per second). A third explanatory variable is constructed, $x_3 = x_1x_2$: this corresponds approximately to the total volume of rainfall. It is required to model runoff using some or all of these regressor variables. The **next two pages** show some edited computer analysis of three possible regression models (1, 2 and 3). Use the output to answer the following questions.

(a) In Model 1, test the multiple regression for global significance, and explain the meaning of the statement '`R-Sq = 48.9%`'. Also carry out a partial t test of significance for each of x_1 and x_2.

(9)

(b) For each of the three models, calculate the predicted runoff for an observation with $x_1 = 1111$ and $x_2 = 5.38$. You may assume that these values are well within the range of observations from which the regressions have been calculated.

(3)

(c) Study the output for all three models carefully, and suggest any further data, information or analyses that might help you choose between them. On the basis of the output, state with reasons which model you prefer.

(4)

The computer output begins on the next page

Model 1

Regression Analysis: y versus x1, x2

```
The regression equation is
y = - 141 + 0.496 x1 + 6.27 x2

Predictor     Coef    SE Coef
Constant    -140.8      255.5
x1          0.4957     0.2164
x2           6.274      2.432

S = 44.1447     R-Sq = 48.9%

Analysis of Variance
Source          DF      SS     MS
Regression       2   18625   9313
Residual Error  10   19488   1949
Total           12   38113

Source   DF   Seq SS
x1        1     5662
x2        1    12963

Unusual Observations
Obs     x1       y     Fit   SE Fit   Residual
 13   1176   383.0   469.7     15.3      -86.7
```

Model 2

Regression Analysis: y versus x1, x3

```
The regression equation is
y = - 73 + 0.439 x1 + 0.00523 x3

Predictor      Coef    SE Coef       T       P
Constant      -73.3      251.5   -0.29   0.777
x1           0.4392     0.2146    2.05   0.068
x3         0.005232   0.002089    2.51   0.031

S = 44.6508     R-Sq = 47.7%

Analysis of Variance
Source          DF      SS     MS      F       P
Regression       2   18176   9088   4.56   0.039
Residual Error  10   19937   1994
Total           12   38113

Source   DF   Seq SS
x1        1     5662
x3        1    12514

Unusual Observations
Obs     x1       y     Fit   SE Fit   Residual
 13   1176   383.0   470.2     15.4      -87.2
```


Turn over

Model 3

Regression Analysis: y versus x2, x3

The regression equation is
y = 443 - 30.6 x2 + 0.0309 x3

Predictor	Coef	SE Coef	T	P
Constant	442.96	25.40	17.44	0.000
x2	-30.57	24.08	-1.27	0.233
x3	0.03089	0.02086	1.48	0.169

S = 49.3588 R-Sq = 36.1%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	13750	6875	2.82	0.107
Residual Error	10	24363	2436		
Total	12	38113			

Source	DF	Seq SS
x2	1	8404
x3	1	5345

Unusual Observations

Obs	x2	y	Fit	SE Fit	Residual
8	19.1	551.0	577.4	42.7	-26.4

Turn over

3. An insurance company is studying the relationship between the size of house insurance claims y, in £, and the current insured value x, in £. Data from a sample of recent claims are given below.

x	50 000	100 000	150 000	200 000	250 000	300 000	350 000	400 000
y	3100	5750	11 700	9050	18 700	12 800	18 150	35 300

(i) Plot the values on a scatter diagram and comment on the suitability of a simple linear regression of y on x for modelling these data.

(5)

(ii) An actuary suggests that the data might reasonably be fitted by a model of the form

$$y = A\exp(Bx + e),$$

where A and B are constants and e is a zero-mean, constant-variance error term. Show how a linearising transformation may be applied to the above model to express it in the form

$$f(y) = a + bx + e,$$

where f, a and b are functions of y, A and B respectively, which you should identify. Plot the values of $f(y)$ and x on a different scatter diagram from that of part (i) and comment on the actuary's suggestion.

(4)

(iii) You are given that

$$\Sigma x = 1.8\times10^{6},\ \Sigma(\log y) = 74.7455,\ \Sigma x^{2} = 5.1\times10^{11},\ \Sigma x(\log y) = 17\,411\,993.$$

Calculate the least squares simple linear regression of $\log y$ on x and plot this line on the scatter diagram in part (ii). Use this regression to calculate a point estimate of the average claim size for a house of current insured value £250 000.

(7)

(iv) You are given that the simple linear regression of y on x is

$$y = -1771 + 0.071\,51x.$$

Plot this regression on the scatter diagram in part (i) and use it to find a point estimate of the average claim size for a house of current insured value £250 000, compare your result with that of part (iii) and say with reasons which estimate you consider to be the more reliable.

(4)

4. The table below shows the height H in metres and the weight W in kilograms of a random sample of 10 diabetes patients at a teaching hospital.

Patient	1	2	3	4	5	6	7	8	9	10
H	1.66	1.91	1.78	2.01	1.55	1.82	1.97	1.74	1.92	1.69
W	75.2	97.5	88.8	128.0	79.5	92.2	116.1	83.4	105.3	86.9

(i) Plot the data on a scatter diagram.

(3)

(ii) Some medical students are considering four ways in which the data could be used to investigate the relationship between H and W for these patients:

(a) simple linear regression of weight on height;

(b) simple linear regression of height on weight;

(c) Pearson (product-moment) correlation coefficient;

(d) Spearman's rank correlation coefficient.

For each of these possibilities, briefly state the assumptions that are made about the population that is sampled. In the light of your scatter diagram in part (i), indicate with reasons which you consider to be the most appropriate.

(5)

(iii) You are given that

$$\Sigma H = 18.05,\ \Sigma W = 952.9,\ \Sigma H^2 = 32.7781,\ \Sigma W^2 = 93\,326.3,\ \Sigma HW = 1740.3.$$

Calculate the Pearson correlation between H and W and test at the 1% level of significance the hypothesis of zero correlation against the alternative of positive correlation.

(6)

(iv) Calculate Spearman's rank correlation between H and W and test the hypothesis of zero rank correlation against the hypothesis of positive rank correlation at the 1% level of significance.

(5)

(v) Comment briefly on your findings in parts (iii) and (iv).

(1)