EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

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. The manager of an oil refinery measures the percentage yield of petroleum spirit y and the specific gravity of crude oil x on seven separate occasions. The data, arranged in order of increasing x-values, are as follows.

x	30.2	32.8	32.9	35.1	42.3	45.5	46.0
y	6.8	10.1	14.3	19.3	10.2	20.0	23.7

(i) Draw a scatter diagram of the data and comment briefly on the suitability of carrying out a simple linear regression analysis on these data.

(5)

(ii) Fit a simple linear regression model $E(Y) = \alpha + \beta x$ to these data, showing details of your calculations.

(6)

(iii) Stating any assumptions you must make and showing details of your calculations, find a 95% confidence interval for β.

(7)

(iv) Give a point prediction for the percentage yield of petroleum spirit when $x = 40$.

(2)

2. (a) (i) You are given a set of bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Define the sample product-moment correlation coefficient.

(3)

(ii) A chocolate manufacturer is interested in the effect of advertising on sales. He selects a random sample of eight of the products he makes. The value *Y* of sales, in tens of thousands of pounds, of each product in a certain period and the amount of money *X*, in thousands of pounds, spent on advertising each product were recorded. Using the following information, calculate the sample product-moment correlation coefficient between the variables 'sales' and 'advertising costs'.

$$\sum_{i=1}^{8} x_i = 386 \qquad \sum_{i=1}^{8} y_i = 460 \qquad \sum_{i=1}^{8} x_i^2 = 25\,426$$

$$\sum_{i=1}^{8} y_i^2 = 28\,867 \qquad \sum_{i=1}^{8} x_i y_i = 26\,161$$

(4)

(iii) Examine, at the 1% significance level, whether there is evidence of positive correlation between 'sales' and 'advertising costs', stating the critical value and your conclusion.

(4)

(b) At a cat show, two judges put 8 Siamese cats, labelled A to H, in the following orders, from best to worst.

Position	1st	2nd	3rd	4th	5th	6th	7th	8th
Judge 1	B	C	E	A	D	F	G	H
Judge 2	C	B	E	D	F	A	G	H

(i) Construct a table showing the ranks of the 8 Siamese cats for each judge.

(2)

(ii) Calculate Spearman's rank correlation coefficient between the sets of ranks. Examine, at the 1% level, whether there is evidence of a positive association between the two judges' ranks of Siamese cats.

(4)

(iii) Does the test using Spearman's rank correlation coefficient rely on any distributional assumptions? State one advantage and one disadvantage of this coefficient over the product-moment correlation coefficient.

(3)

3. (i) Briefly explain the reasons for using 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 below.

(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)

Four different marine paints were compared for their ability to protect ships in a sea-going environment. Sixteen ships were used, each treated with one of the four paints. Each of the ships was deployed for 6 months and on the ship's return a score was assigned according to the amount of chipping, peeling and average remaining paint thickness. A higher score indicated a better state of repair. The scores are given in the following table.

Paint 1	80	73	72	90
Paint 2	81	82	88	84
Paint 3	93	80	80	97
Paint 4	89	86	96	99

(iii) The following is part of the analysis of variance table for a one-way model with a factor 'Paint'.

```
Source    d.f.   sum of squares    mean square    F
Paint
Error                577.5
Total                983.8
```

Complete the table and test whether there are differences between the four paints.

(8)

(iv) State two problems with the design of this experiment and two improvements that could be made to the experiment.

(4)

4. (i) Two explanatory variables, X_1 and X_2, are used to predict a dependent variable Y. Write down a multiple linear regression model which can be used as a basis for the analysis of data containing Y, X_1 and X_2 as described above, and explain the meanings and properties of the terms in the model.

(5)

An experimental investigation was made into the heat evolved during the hardening of cement, considered as a function of the chemical composition of the cement. The data recorded were the heat evolved (Y) after 180 days of hardening measured in calories per gram of cement, and the percentages of tricalcium aluminate (X_1) and tricalcium silicate (X_2).

The data were read into a statistical package for analysis. The relevant output follows **at the end of this question**. Use the output to answer the following questions.

(ii) Test the overall regression for significance at the 1% level, and explain the results in terms that a non-statistician would understand.

(4)

(iii) Write down the fitted regression equation of Y on X_1 and X_2 as defined above. Use it to predict the heat evolved during hardening of similar cement with $X_1 = 8$ and $X_2 = 35$.

(5)

(iv) In the output for the fitted model, the p-values for the partial t tests for the regression parameters are missing. Test these parameters for statistical significance at the 0.1% level, quoting the critical value. What do the results imply about the effects of tricalcium aluminate and tricalcium silicate on the heat evolved in hardening?

(4)

(v) Use the value of R^2 to comment on the overall fit of the model.

(2)

```
Regression Analysis: y versus xl and x2

Analysis of Variance Table

Response: y
             Df    Sum Sq    Mean Sq    F value     Pr(>F)
Regression    2    2657.9    1328.95     229.53    4.404e-09
Residuals    10      57.9       5.79
Total        12    2715.8

Coefficients:
                Estimate     Std. Error   t value      Pr(>|t|)
(Intercept)     52.57735      2.28617      23.00       5.46e-10
x1               1.46831      0.12130      12.11
x2               0.66225      0.04585      14.44

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared: 0.9787
```

BLANK PAGE

BLANK PAGE

BLANK PAGE