# EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

## HIGHER CERTIFICATE IN STATISTICS, 2008

**(Modular format)**

## 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 5 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 trial is undertaken to investigate the effect on fuel economy of 3 fuel additives A, B and C, where A and B are new and C is the current standard additive. The same driver drives the same car on a fixed test route during 20 working days. The additive used on each day is randomly assigned so that A and B are each used for 5 days and C is used for 10 days. The response variable measured each day is $Y$, the number of miles per gallon (mpg) achieved.

The results are shown in the following table.

| *Additive* | *y* | *Total* |
|---|---|---|
| A | 39, 35, 37, 36, 38 | $\sum y_A = 185$ |
| B | 36, 41, 39, 40, 39 | $\sum y_B = 195$ |
| C | 37, 33, 30, 34, 36, 34, 31, 36, 34, 35 | $\sum y_C = 340$ |

You are given that the sum of squares of the observations is 26078.

(i) (a) Carry out an analysis of variance to test for differences between the effects on $Y$ of the additives. State clearly your null and alternative hypotheses and present your conclusions.

(11)

(b) Give a 95% confidence interval for the mean difference in mpg between A and C, and interpret this interval.

(3)

(ii) (a) State the assumptions needed for the analysis you have done in part (i), and suggest how you might check them (but do not actually do so).

(4)

(b) What advice might you offer about scheduling the test drives in any future trial so as to minimise possible influence of varying traffic conditions?

(2)

2. An experiment was carried out to study the variation of the specific heat $H$ in calories per gram of a certain compound with $T$, its temperature in degrees Celsius. The specific heat was measured twice at each of a series of chosen temperatures, and the results are shown in the following table.

| $t$ | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|
| $h$ | 1.64 | 1.63 | 1.67 | 1.72 | 1.71 | 1.71 |
| | 1.60 | 1.65 | 1.67 | 1.70 | 1.72 | 1.74 |

You are given that $\sum t = 900$, $\sum h = 20.16$, $\sum t^2 = 71000$, $\sum h^2 = 33.8894$, $\sum th = 1519.9$.

(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) (a) Fit a simple linear regression model to the data, stating any assumptions made for the purpose of the analysis. Also give a point prediction for the specific heat when $T = 85$.

(5)

(b) Let $\sigma^2$ denote the residual variance. Estimate $\sigma^2$ and test at the 1% significance level the null hypothesis that the slope parameter in the regression model is zero against a two-sided alternative. State your conclusion clearly.

(6)

(c) A statistical research adviser notes that, under the usual assumptions of regression analysis, the difference (first result − second result) found for any given temperature is distributed with zero mean and variance $2\sigma^2$, the six differences being independent across temperatures. Use this result to obtain a second estimate of $\sigma^2$, compare it with your regression-based estimate and comment briefly.

(4)

3. (i) Two explanatory variables 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, and explain the meanings and properties of the terms in the model.

(4)

(ii) The data in the following table show the values of price $Y$ (£) for individually patterned Persian carpets of length $x_1$ (cm) and width $x_2$ (cm).

| $y$ | 14 | 20 | 37 | 36 | 31 | 42 | 54 | 64 | 38 | 66 | 64 | 77 | 79 | 93 | 119 | 135 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | |
| $x_1$ | 120 | 120 | 120 | 120 | 150 | 150 | 150 | 150 | 180 | 180 | 180 | 180 | 240 | 240 | 240 | 240 |
| $x_2$ | 60 | 80 | 100 | 120 | 75 | 100 | 125 | 150 | 90 | 120 | 150 | 180 | 120 | 160 | 200 | 240 |

(a) Plot scatter diagrams of price against each of length and width. What do these graphs show?

(5)

(b) A multiple regression model of price on length and width was fitted to the data given in the table. Edited computer output of the results is as follows.

```
Predictor       Coef    SE Coef         T         P
Constant    -52.671    5.34500    -9.85     0.000
Length      0.32356    0.04250     7.61     0.000
Width       0.44383    0.04012    11.06     0.000

S = 5.32611      R-Sq = 97.9%

Analysis of Variance
Source          DF       SS       MS         F       P
Regression       2  17045.2   8522.6    300.43   0.000
Residual Error 13    368.8     28.4
Total          15  17413.9
```

Interpret these results fully, in terms that a non-statistician would understand. Write down the fitted regression equation of $Y$ on $x_1$ and $x_2$, and use it to predict the price of a similar carpet of length 200 cm and width 150 cm. To what extent would you rely on the model to predict the prices of carpets of dimensions outside the sizes observed in the above table (for example, much smaller carpets)?

(11)

4. The following coded pairs of measurements were taken of the temperature ($X$) and thrust ($Y$) of a jet engine while it was being tested under uniform operating conditions.

| $x$ | 15 | 20 | 25 | 26 | 30 | 33 | 34 | 35 | 38 | 39 | 41 | 46 | 49 | 52 | 57 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 1.4 | 1.2 | 1.9 | 1.6 | 2.5 | 2.1 | 2.4 | 1.5 | 2.3 | 2.7 | 1.8 | 2.2 | 2.8 | 3.4 | 3.2 |

You are given that $\sum x = 540$, $\sum x^2 = 21412$, $\sum y = 33.0$, $\sum y^2 = 78.54$, $\sum xy = 1276.6$.

(i) (a) Plot the data on a scatter diagram and comment on the suitability of the Pearson product-moment correlation coefficient, $r$, as a measure of the association between thrust and temperature.

(6)

(b) Calculate $r$ for these data, and test at the 5% significance level the hypothesis of zero correlation against the alternative that thrust and temperature are positively correlated. State clearly (but do not prove) any formulae that you have used, and list the assumptions you have made in the test.

(6)

(ii) A colleague wishes to test at the 5% significance level the hypothesis of no trend against the alternative of an increasing trend, without assuming that the trend is necessarily linear. State what measure of association he should use, calculate it for the above data and carry out the desired test. Compare the result of this test with your findings in part (i).

(8)