EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2015

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. (i) A simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

is to be fitted to some data. What assumptions are usually made about the term representing experimental error (ε_i)?

(2)

(ii) By minimising a suitable function, show that the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 are given by

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where $S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \Sigma(x_i - \bar{x})^2$.

(8)

A clinician recorded the age in years (x) and total cholesterol level of blood (y) for 20 patients suffering from a certain disease. Summary statistics for the data are Σx_i = 809, Σy_i = 68.3, S_{xx} = 3630.95, S_{xy} = 201.665, S_{yy} =12.9455.

(iii) Find the equation of the fitted simple linear regression model.

(3)

(iv) Obtain the analysis of variance and hence test the hypothesis that $\beta_1 = 0$.

(7)

2. It is required to compare the durability of four experimental carpet products. Four samples of each product are produced, giving 16 samples altogether. One of these samples is assigned to each of 16 homes, using a completely randomised design. A measure of durability, *y*, is made for each sample after 60 days, with the results below.

Carpet product			
1	2	3	4
18.95	10.06	10.92	10.46
12.62	7.19	13.28	21.40
11.94	7.03	14.52	18.10
14.42	14.66	12.51	22.50

(i) The usual model for the one-way analysis of variance is given by

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} .$$

State any associated assumptions about the term representing experimental error (ε_{ij}) and state a suitable constraint on the parameters that will enable them to be estimated.

(3)

(ii) Carry out an analysis of variance of these data, using the fact that $\Sigma\Sigma y_{ij}^2 = 3350.27,$ and test at the 5% significance level the null hypothesis that the mean durability is the same for the different carpet products.

(11)

(iii) You are now told that carpet product 2 is made from a synthetic material and carpet product 4 is made from wool alone. Perform an appropriate *t* test at the 0.5% significance level to investigate whether there is a difference in mean durability between these two carpet products.

(6)

3. (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* (pmcc).

(3)

(ii) A new variable z is calculated as $z_i = ax_i + b$, i = 1, 2, …, n, where a, b are constants. Show that when $a > 0$ the pmcc between z and y is the same as that between x and y. What is the pmcc between z and y when $a < 0$?

(6)

The number of chirps (pulses of sound) per second made by a striped ground cricket was measured at various temperatures. Because crickets are cold-blooded there is reason to believe that temperature affects aspects of their behaviour such as chirp frequency. The data are given in the table below and may be regarded as a random sample from a population of crickets.

Chirps/Second y	Temperature (°F) x
14.4	76.3
14.7	69.7
15.0	79.6
15.4	69.4
15.5	75.2
15.7	71.5
16.0	71.6
16.1	80.5
16.3	83.3
17.0	83.5
17.1	80.6
17.2	82.6
18.4	84.3
19.8	93.3
20.0	88.6

You are given that

$$\Sigma x_i = 1190.0,\ \Sigma y_i = 248.6,\ \Sigma x_i^2 = 95\,098,\ \Sigma y_i^2 = 4161.1,\ \Sigma x_i y_i = 19\,862.6.$$

(iii) Calculate the pmcc for these data, and test at the 1% significance level the null hypothesis that, in the underlying population, the number of chirps and temperature are uncorrelated, against the alternative of a positive correlation.

(6)

(iv) Calculate Spearman's rank correlation coefficient for these data (note that the data for y are already listed in rank order). Use it to test, at the 1% significance level, the null hypothesis that there is no association between the number of chirps and temperature in the underlying population, against the alternative of positive association.

(5)

4. Write down the multiple linear regression model for a response variable Y and explanatory variables x_1, x_2, …, x_p, stating any necessary assumptions.

(4)

In an ecological study 35 bears, regarded as a random sample, were temporarily anaesthetised, and their bodies were measured and weighed. One goal of the study was to estimate the weight of a bear based on other measurements. This would be used because in the forest it is easier to measure the length of a bear, for example, than it is to weigh it.

The following variables were recorded: Age (age, in months); Head (length of the head, in inches); Neck (neck girth, distance around the neck, in inches); Chest (chest girth, in inches); Weight (weight, in pounds).

The researchers wished to build a multiple linear regression model with Weight as the dependent variable and the other four variables as explanatory variables. Summary output from building this model is given below.

(i) State what hypothesis is tested using the analysis of variance table. What is your conclusion from this test using a 5% significance level? Explain the meaning of the statement 'R–Sq = 96.3%'.

(6)

(ii) Carry out a test at the 5% level of the significance of the variable Head in the regression, stating the hypothesis that you are testing. Calculate also a 90% confidence interval for the coefficient of Chest.

(8)

(iii) Use the regression equation to calculate the predicted weight of a bear with measurements Age = 71, Head = 7, Neck = 27 and Chest = 44. You may assume that these values are well within the range of observations from which the model has been estimated.

(2)

```
Regression Analysis: Weight versus Age, Head, Neck, Chest

The regression equation is
Weight = - 203 + 0.655 Age - 9.01 Head + 11.8 Neck + 6.26 Chest

Predictor      Coef    SE Coef
Constant    -202.52      35.05
Age          0.6555     0.1905
Head         -9.013      4.693
Neck         11.815      3.559
Chest         6.255      1.677

S = 25.7088   R-Sq = 96.3%   R-Sq(adj) = 95.8%

Analysis of Variance

Source            DF       SS      MS      F
Regression         4   513741  128435  194.3
Residual Error    30    19828     661
Total             34   533570
```

BLANK PAGE

BLANK PAGE

BLANK PAGE