EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2014

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 table below shows some data for the nine top-ranked countries in the medal table for the London 2012 Olympic Games according to

http://data.london.gov.uk/datastore/package/alternative-olympics-2012-medal-table (accessed on 15 June 2013).

Country	*Rank*	*Gold*	*Silver*	*Bronze*
USA	1	46	29	29
China	2	38	27	22
G Britain	3	29	17	19
Russia	4	24	25	33
S Korea	5	13	8	7
Germany	6	11	19	14
France	7	11	11	12
Italy	8	8	9	11
Hungary	9	8	4	5

(i) Plot on a scatter diagram the paired data (x_i, y_i), $i = 1, \ldots, 9$, where x_i and y_i respectively denote the silver and bronze medal counts for the country ranked ith in the table above, $i = 1, 2, ..., 9$. Comment briefly on the relationship between the silver and bronze medal counts, and on the suitability of the product moment correlation coefficient (pmcc) for summarising this relationship.

[7]

(ii) Calculate the pmcc, r_{xy}, for these data, test at the 2% level the null hypothesis of zero correlation against the two-sided alternative, and give your conclusion clearly.

[You are given that $\Sigma x = 149$, $\Sigma y = 152$, $\Sigma x^2 = 3127$, $\Sigma y^2 = 3310$, $\Sigma xy = 3156$.]

[4]

(iii) Calculate Spearman's rank correlation coefficient, r_{xy}^s, for these data, test at the 2% significance level the null hypothesis of no association against the two-sided alternative, and give your conclusion clearly. State with reasons which of r_{xy} and r_{xy}^s you consider to be the better measure of association between the silver and bronze medal counts.

[9]

2. Gross mean weekly earnings (y, in £ per week) for a sample of male clerical workers of varying ages (x, in completed years) in a large company are as follows.

Earnings y	215	259	348	387	534	660	726	$\Sigma y = 3129$	$\Sigma y^2 = 1\,632\,011$
Age x	18	20	23	28	35	45	55	$\Sigma x = 224$	$\Sigma x^2 = 8312$

You are also given that $\Sigma xy = 116\,210$.

(i) Plot a scatter diagram of these data and comment on their suitability for simple linear regression analysis.

[4]

(ii) Write down the models for

(a) simple linear regression of y on x,

(b) simple linear regression of x on y.

Define your notation and explain clearly which model is better suited to fit the variables and data as defined in the table above.

[5]

(iii) (a) Fit the simple linear regression model of y on x to the data above, find the equation of the fitted regression line, draw this line on your scatter diagram, and use the equation to estimate the mean weekly earnings at age 50.

[7]

(b) You are given that the estimated standard error of the slope of the regression line is 1.128. Use this result to test for the significance of the regression slope at the 5% level, and state your conclusion clearly.

[3]

(c) Your line manager asks you to use your model to estimate the mean weekly earnings at age 70. How would you answer him?

[1]

3. Counts of lesions (y, in hundreds) of Aucuba mosaic virus after timed exposures (x, in minutes) to X-rays are shown in the table below. The table also shows values of $\log_{10} y$.

y	271	226	209	108	59	29	12
x	0	3	7.5	15	30	45	60
$\log_{10} y$	2.43297	2.35411	2.32015	2.03342	1.77085	1.46240	1.07918

(i) Plot the (x, y) data on a scatter diagram and comment briefly.

[4]

(ii) It is thought that these data conform to an underlying relationship of the form

$$y = a \times 10^{-bx},$$

where a and b are positive constants to be estimated. Show how a logarithmic transformation may be applied to give a linear relationship of the form

$$\log_{10} y = A + Bx,$$

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

[3]

(iii) Plot the $(x, \log_{10} y)$ data on a scatter diagram and comment briefly.

[3]

(iv) You are given that

$$\Sigma x = 160.5, \quad \Sigma x^2 = 6815.25, \quad \Sigma(\log_{10} y) = 13.4531, \quad \Sigma(\log_{10} y)^2 = 27.4182.$$

Fit a linear regression to the $(x, \log_{10} y)$ data, and hence write down a formula for the estimated count of lesions (in hundreds) after exposure for x minutes. Calculate this estimate in the case $x = 10$. Use your regression to estimate the X-ray exposure time which would halve the count of lesions observed.

[10]

4. State the model and assumptions for the one-way analysis of variance, defining your notation.

[8]

Specimens of milk from dairies in three different districts are assayed for their concentrations of the radioactive isotope Strontium-90. The results, in picocuries per litre, are as shown in the table below.

District	*Observations*	*Sum*	*Sum of Squares*
1	6.7 6.1 6.8 8.0	27.6	192.34
2	7.5 10.3 11.6 10.9 6.9 9.2	56.4	547.96
3	10.3 9.8 12.9	33.0	368.54

Carry out an analysis of variance of these data, conducting your significance test at the 5% level. Report your conclusion clearly, in terms that a non-statistician would understand.

[9]

Test the null hypothesis H_0: there is no difference between the mean concentrations in Districts 1 and 2, against the two-sided alternative using a 5% significance level, and give your conclusion.

[3]

BLANK PAGE

BLANK PAGE

BLANK PAGE