EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

HIGHER CERTIFICATE IN STATISTICS, 2009

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 8 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. The Devon Motor Racing Grand Prix takes place every five years. Winning average lap speeds (in miles per hour) in the last nine events are shown in the table below.

Year x	1965	1970	1975	1980	1985	1990	1995	2000	2005
Speed y	109	114	116	117	114	127	131	138	141

You are given that

$$\bar{x} = 1985, \quad \sum(x-\bar{x})^2 = 1500, \quad \sum y = 1107, \quad \sum y^2 = 137233, \quad \sum(x-\bar{x})y = 1200.$$

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

(4)

(b) Fit a simple linear regression model and state its equation. Also compute the total sum of squares and regression sum of squares for this regression, and deduce the error mean square.

(6)

(ii) It is later noted that driving conditions in 1985 were affected by a freak thunderstorm which caused partial flooding of the track. The 1985 values were therefore omitted and the regression was recalculated. Results are shown in the computer output below. Compare this analysis with your own results and say with reasons which you regard as the more satisfactory.

(3)

```
The regression equation is    y = - 1464 + 0.800 x

Predictor       Coef   SE Coef        T       P
Constant    -1463.87     95.60   -15.31   0.000
x            0.80000   0.04816    16.61   0.000

S = 1.86525   R-Sq = 97.9%   R-Sq(adj) = 97.5%

Analysis of Variance
Source           DF       SS       MS        F       P
Regression        1   960.00   960.00   275.93   0.000
Residual Error    6    20.87     3.48
Total             7   980.87
```

(iii) Use the analysis of part (ii) to obtain point estimates of

(a) the expected winning speed in 1985,

(2)

(b) the expected winning speed in 2010,

(1)

(c) the time by which a winning speed of 160 mph might be expected.

(2)

Mention any reservations you might have about your answers.

(2)

2. The table below shows breaking strains, y kg, of 5 samples of steel wire from each of 4 suppliers, A, B, C and D.

Supplier	*Breaking strains*	$\sum y$	$\sum y^2$
A	137, 133, 136, 134, 130	670	89810
B	136, 134, 140, 133, 137	680	92510
C	134, 142, 137, 143, 139	695	96659
D	144, 140, 143, 147, 141	715	102275

(i) Carry out an analysis of variance to test for differences between the mean breaking strains of the steel wire manufactured by these four suppliers. State your conclusions clearly.

(12)

(ii) State the assumptions made in your analysis and briefly indicate, without further calculation, how you would check them.

(4)

(iii) It is known that supplier A is the cheapest and supplier B the next cheapest. Calculate a two-sided 90% confidence interval for the difference between the mean breaking strains for suppliers A and B, and interpret this interval. Using your confidence interval, or otherwise, test at the 5% significance level the null hypothesis that the mean breaking strains for A and B are equal, against the alternative that the mean breaking strain for B exceeds that for A.

(4)

Turn over

3. (a) Draw scatter diagrams to illustrate the following features of bivariate data.

(i) Strong positive association, appropriately reflected by the product-moment correlation coefficient.

(2)

(ii) Weak negative association, appropriately reflected by the product-moment correlation coefficient.

(2)

(iii) Strong negative association, appropriately reflected by Spearman's rank correlation coefficient but less satisfactorily by the product-moment correlation coefficient.

(2)

(iv) Strong association with a non-monotonic trend.

(2)

(b) State the statistical assumptions underlying

(i) a hypothesis test on the product-moment correlation coefficient,

(2)

(ii) a hypothesis test on the slope parameter of a simple linear regression that includes a constant term.

(2)

(c) The following table shows diastolic (DBP) and systolic (SBP) blood pressure measurements (in mm Hg) for 10 cardiac patients.

DBP	55	60	70	75	80	85	90	95	105	110
SBP	125	115	120	135	105	145	130	200	190	150

(i) Plot the data and suggest a suitable measure of association for these data, justifying your choice.

(3)

(ii) Given that the sample product-moment correlation coefficient of these data is 0.673, test at the 1% level the null hypothesis that $\rho = 0$ against the alternative hypothesis that $\rho > 0$, where ρ is the population value of the product-moment correlation coefficient. Also calculate the value of Spearman's rank correlation for these data, and carry out the corresponding test. State your conclusions clearly.

(5)

Turn over

4. The accompanying edited computer output **(on the next three pages)** shows analyses of three different regression models for the progress in sales in hundreds (y) of a newly developed electronic component over time in months (x) since the launch of this product. These regression models may be written as shown below.

Model 1: $E(Y \mid x) = \alpha + \beta x$

Model 2: $E(Y \mid x) = \alpha + \beta x + \gamma x^2$

Model 3: $E\left(\log_{10} Y \mid x\right) = \alpha + \beta x$

Use the output to answer the following questions.

(i) In the output for Model 1, the p-values for the partial t tests for the slope and intercept parameters are missing. Making any necessary assumptions, use the available information to test these parameters for statistical significance at the 5% level.

(3)

(ii) In the light of Plot 1, comment on the adequacy of Model 1 for the data.

(2)

(iii) Test the significance of the coefficient of x in Model 2 and compare the outcome with the result of your test for the coefficient of x in Model 1. Interpret the statement "R-Sq = 98.1%" in the output for Model 2.

(3)

(iv) With reference to Plot 2, what standard assumption about the distribution of the error term may be called into question in Model 2?

(2)

(v) Critically compare Models 1 and 3 with regard to their success in fitting the data. Why does the total sum of squares for Model 3 differ from the total sum of squares for Models 1 and 2?

(5)

(vi) Use each of these models to give a point estimate of sales 10 months after launching the product. State with reasons which of the three estimates you think is the best.

(5)

The edited computer output is on the next three pages

Edited Computer Output for question 4

The output is given on this page and the next two pages

Model 1

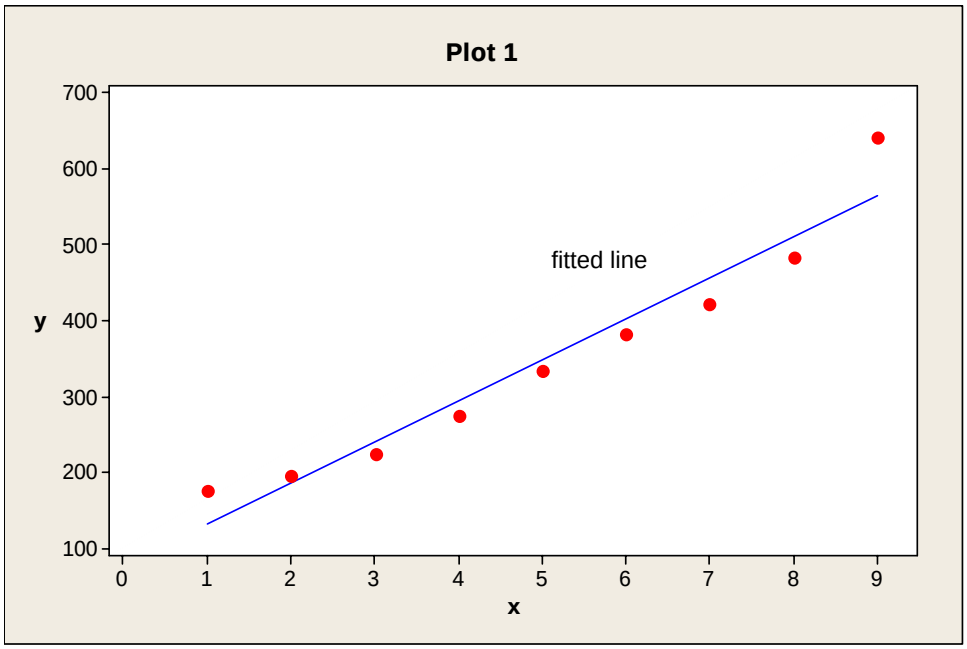
```
The regression equation is    y = 78.3 + 54.0 x

Predictor      Coef   SE Coef
Constant      78.33     29.01
x            54.000     5.155


S = 39.9285   R-Sq = 94.0%   R-Sq(adj) = 93.1%


Analysis of Variance

Source           DF       SS       MS        F       P
Regression        1   174960   174960   109.74   0.000
Residual Error    7    11160     1594
Total             8   186120
```



Output continued on next page

Model 2

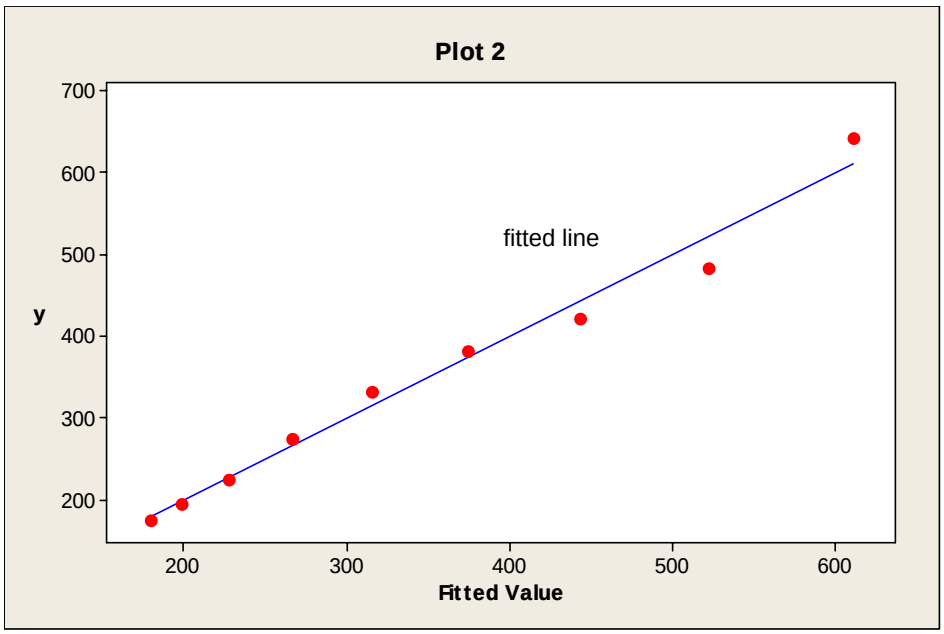
The regression equation is y = 170 + 4.0 x + 5.00 x-sq

Predictor	Coef	SE Coef	T	P
Constant	170.00	30.56	5.56	0.001
x	4.00	14.03	0.29	0.785
x-sq	5.000	1.368	3.65	0.011

S = 24.0139 R-Sq = 98.1% R-Sq(adj) = 97.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	182660	91330	158.38	0.000
Residual Error	6	3460	577		
Total	8	186120			



Output continued on next page

Model 3

The regression equation is log10(y) = 2.16 + 0.0689 x

Predictor	Coef	SE Coef	T	P
Constant	2.16086	0.01422	151.93	0.000
x	0.068910	0.002527	27.26	0.000

S = 0.0195777 R-Sq = 99.1% R-Sq(adj) = 98.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.28492	0.28492	743.35	0.000
Residual Error	7	0.00268	0.00038		
Total	8	0.28760