EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2006

Applied Statistics I

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal 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 13 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

©RSS 2006

1. (i) Write down the general forms of AR(p) and MA(q) time series models, stating fully the standard stochastic assumptions.

(3)

(ii) For each of the following time series calculate the mean, variance, and autocorrelation function. State, without working, the form of the partial autocorrelation function.

(a) $Y_t = 5 + \varepsilon_t - 0.3\varepsilon_{t-1}$

(b) $Y_t = 34 - 0.1Y_{t-2} + \varepsilon_t$

[The symbols all have their usual meanings, and the series in (b) is stationary.]

(14)

(iii) Express the model

$$Y_t = 54 - 0.2Y_{t-1} + \varepsilon_t$$

as an infinite moving average process.

[The symbols all have their usual meanings.]

(3)

Turn over

2. Some years ago, data describing road usage were collected in 46 areas of the USA. The variables were as follows.

Drivers	number of drivers $\times 10^{-4}$
Popn	number of people per square mile
Lengthrd	miles of road $\times 10^{-3}$
Fuel	total fuel consumption (gallons $\times 10^{-6}$)
Deaths	number of road deaths in one year

(i) A principal component analysis of the correlation matrix of the first four variables (i.e. excluding Deaths) yields the following results.

	Principal component			
	1	2	3	4
Drivers	–0.61	–0.14	0.59	–0.50
Popn	–0.10	–0.82	–0.50	–0.24
Lengthrd	–0.48	0.53	–0.63	–0.31
Fuel	–0.62	–0.14	–0.02	0.77
Eigenvalue	2.23	1.33	0.25	0.19

(a) Explain whether or not it would be appropriate to carry out the analysis on the covariance matrix, and why.

(2)

(b) Many statistical packages also give values for the "cumulative proportion". Explain what this is, and calculate the corresponding values for this analysis.

(2)

(c) Interpret the results, giving interpretations of the components where appropriate.

(6)

Question 2 is continued on the next page

(ii) In a subsequent analysis a stepwise regression is performed using Deaths as the dependent variable and the four principal component scores (PC1, PC2, PC3 and PC4) as the predictor variables. The package stops when the introduction of further variables is not significant at a nominal 5% level. The computer output is shown below.

STEP	1	2	3
constant	968.5	968.5	968.5
PC1	-545	-545	-545
t-ratio	-13.29	-17.08	-24.64
PC3		519	519
t-ratio		5.45	7.68
PC4			-517
t-ratio			-6.89
S	411	319	221
R-squared	80.1	88.2	94.5

(a) Someone suggests that the output might be more easily interpreted if the original variables were in units of kilometres and litres, rather than miles and gallons, where appropriate. Comment on this, justifying your answer.

(3)

(b) Using the PC analysis and the stepwise regression, discuss which of the variables affected the number of road deaths, and in what ways.

(3)

(iii) People sometimes recommend using principal component scores of a "suitable" subset of principal components instead of original variables when the number of predictor variables is large. Discuss the advantages and disadvantages of this approach.

(4)

3. (i) Distinguish between discriminant analysis and cluster analysis. Give an example of a situation in which discriminant analysis would be an appropriate method of analysis, but where cluster analysis would not be appropriate.

(5)

(ii) A market researcher wishes to investigate whether eight products A – H fall into distinct categories (i.e. a few closely associated groups) with respect to customer perception. She has conducted a survey and a statistician has helped her to derive a meaningful measure of distance between the products. The distances are given in the matrix below.

	Product							
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A	0.0	3.9	5.5	8.7	9.3	3.7	8.9	1.0
B	3.9	0.0	3.0	4.9	5.6	6.7	7.3	3.7
C	5.5	3.0	0.0	4.6	5.7	7.1	4.6	5.2
D	8.7	4.9	4.6	0.0	1.7	11.0	7.3	8.4
E	9.3	5.6	5.7	1.7	0.0	11.7	8.1	9.1
F	3.7	6.7	7.1	11.0	11.7	0.0	9.1	3.9
G	8.9	7.3	4.6	7.3	8.1	9.1	0.0	8.7
H	1.0	3.7	5.2	8.4	9.1	3.9	8.7	0.0

(a) State the conditions necessary for a measure to be a suitable distance measure, and show that for products A, B and C these conditions are indeed met.

(4)

(b) Carry out a hierarchical clustering method using single linkage cluster analysis, and draw a dendrogram of your results.

(8)

(c) Do you think there is any evidence of the "distinct categories" proposed by the researcher? Justify your answer.

(3)

4. Suppose that independent random variables $Y_1, Y_2, \ldots, Y_n$ follow binomial distributions, that is

$$Y_i \sim \mathrm{B}(m_i, \pi_i),\ i = 1, 2, \ldots, n, \quad \text{where } P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}.$$

(i) (a) Show that the natural canonical link function for this distribution in the context of generalised linear models is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

where $\eta_i = \mathbf{x_i}^T \boldsymbol{\beta}$ is the linear predictor.

(2)

(b) Define the terms *odds* and *log-odds* in the context of such a model. Explain how these values may be estimated after such a model has been fitted. Explain also how approximate 95% confidence intervals may be calculated for these quantities, given that the necessary standard errors for the linear predictor values are available.

(3)

(ii) The following data show the perinatal mortality rates for children born to a group of women classified by their age and the length of the gestation period.

Gestation period (days)	*Mother's age (years)*	*Mortality/Total Births*
197 – 260	<30	59/414
	≥30	45/203
>260	<30	30/4501
	≥30	15/1633

The data were analysed by fitting a binomial model with logit link. The variables used were:

GEST coded as 0 for 197 – 260, and 1 for >260

AGE coded as 0 for <30, and 1 for ≥30.

The results are summarised below.

Question 4 is continued on the next page

Variables in model	*Scaled deviance*	*Parameter estimate*	*(standard error)*
constant	346.25	–3.7912	(0.0828)
constant AGE	334.08	–3.9931 0.6054	(0.1070) (0.1693)
constant GEST	6.88	–1.5959 –3.3117	(0.1075) (0.1843)
constant AGE GEST	0.3140	–1.7659 0.4677 –3.2886	(0.1296) (0.1798) (0.1846)

(a) Briefly describe two situations, one where the method of sampling would be appropriate to using the above model for the analysis of these data, and one where it would not. Justify your answers.

(3)

For the rest of this question, you should assume that it is valid to use the above model.

(b) Using forward selection, discuss with reasons which combination of variables in the model best describes the data.

(4)

(c) Using the output for your chosen combination, estimate the odds of mortality for the group "mother's age less than 30 years and gestation period 197 – 260 days"; estimate also the probability of mortality for this group. Calculate approximate 95% confidence intervals for both these quantities.

(5)

(d) Calculate the odds ratio for the mortality in the group in (c) compared to the "less than 30 years and at least 260 days" group, and give an approximate 95% confidence interval for this quantity.

(3)

5. A scientist has collected a set of data (x_i, y_i) in a situation in which he believes that the underlying model is of the form

$$y = ae^{bx}.$$

He has read about two ways of fitting such a model and does not know how to proceed.

(i) One method is to use log(y) rather than y as the response variable.

(a) Write down the appropriate statistical form of the model, stating the assumptions with regard to the distribution of errors for the *untransformed* data.

(2)

(b) Derive the normal equations for this model and hence derive expressions for the parameter estimates.

(5)

(ii) The second method uses the statistical model

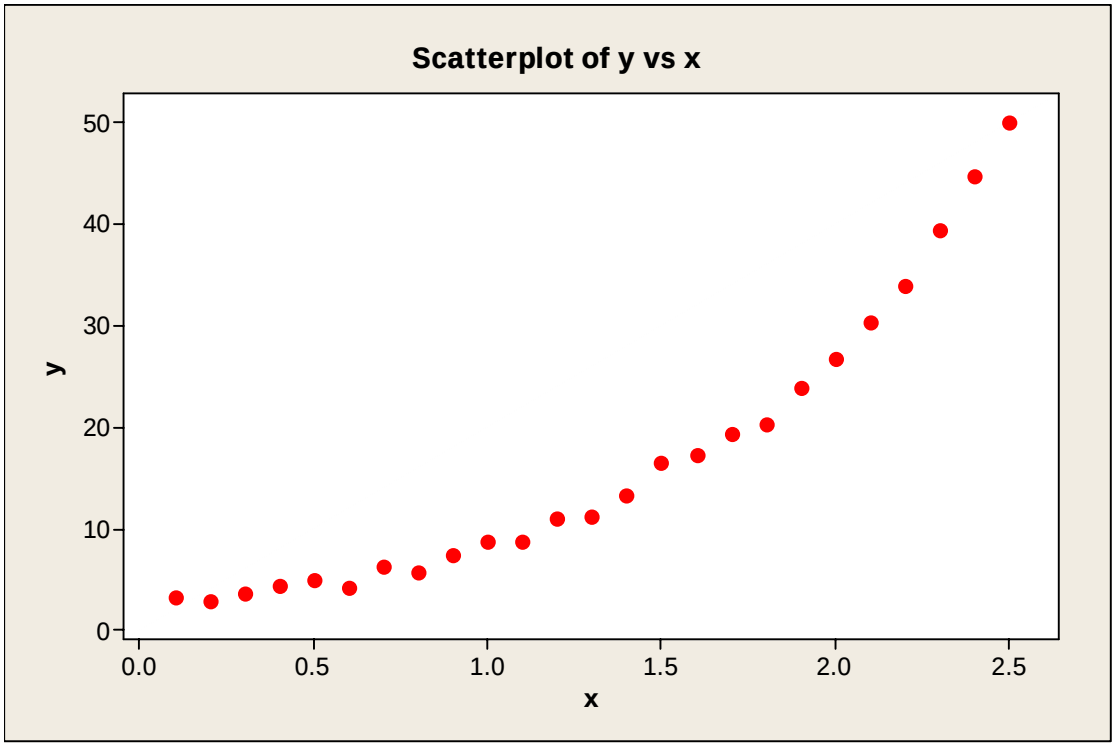
$$y_i = ae^{bx_i} + \varepsilon_i .$$

For a non-linear model $y_i = f(x_i) + \varepsilon_i$, normal equations for any parameters can be obtained by differentiating $S = \Sigma\{y_i - f(x_i)\}^2$ suitably and equating the derivatives to zero. Derive the normal equations for the model $y_i = ae^{bx_i} + \varepsilon_i$, and hence derive a nonlinear equation for the estimate of the parameter b, and an expression for the estimate of a in terms of that of b.

(6)

(iii) The scientist proceeds to fit each model to his data. He has 25 data points, shown in the scatterplot below.

Question 5 is continued on the next page



The resulting equations are

$\log(y) = 0.918 + 1.19x$ from the first model

and

$y = 2.45e^{1.20x}$ from the second model.

(a) Compare the estimates of a and b given by the two models.

(1)

(b) Computer output has provided some residual plots, which are given **on the next page**. Based on all of the available information, which model would you prefer, and why?

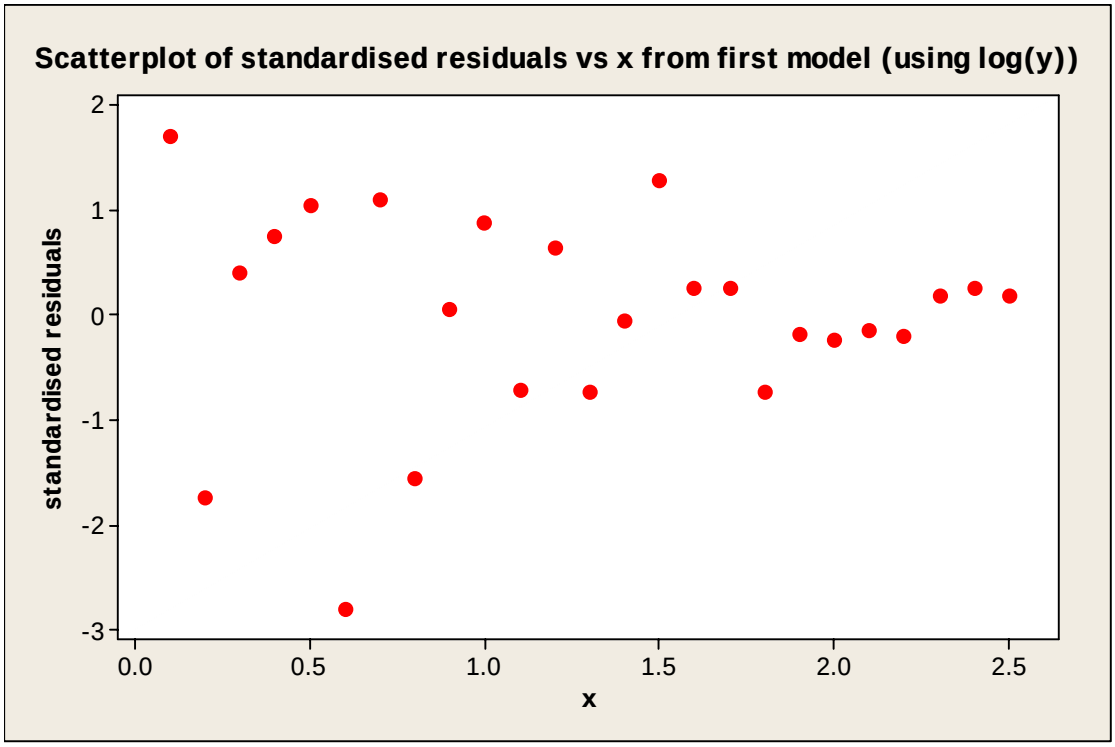
(3)

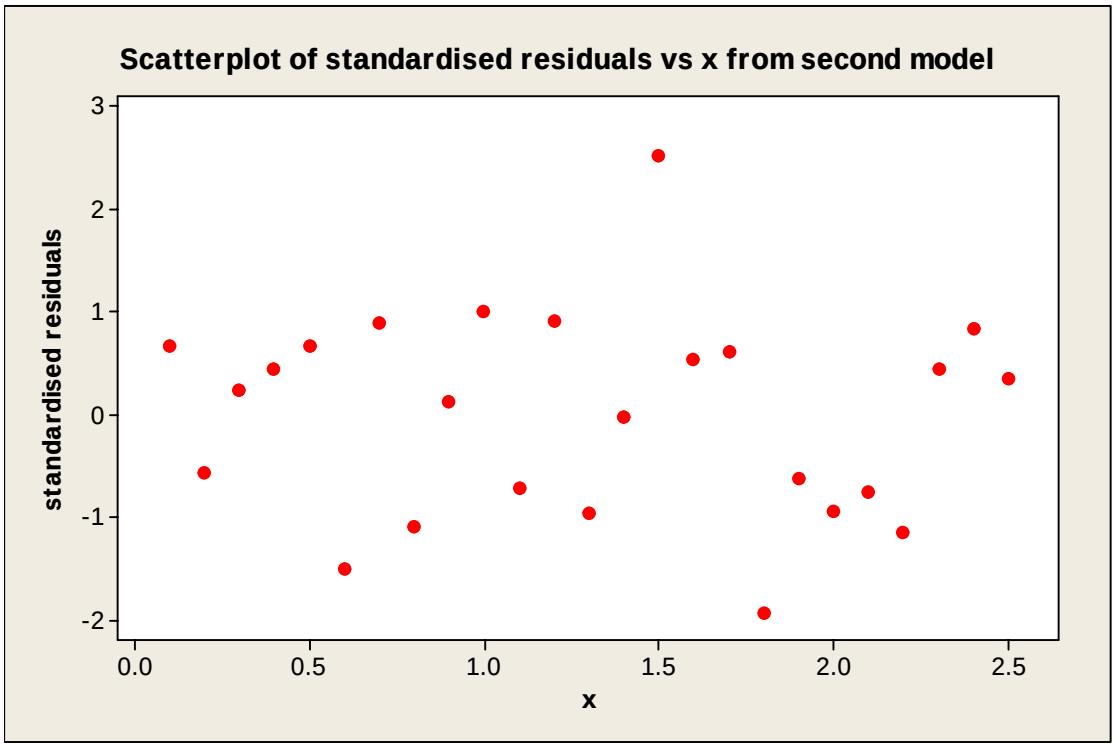
(c) What further information would you wish to have in order to decide which is the better model?

(3)

The residual plots are on the next page

Turn over


Scatterplot of standardised residuals vs x from first model (using log(y))
standardised residuals
2
1
0
-1
-2
-3
0.0
0.5
1.0
1.5
2.0
2.5
x


Scatterplot of standardised residuals vs x from second model
standardised residuals
3
2
1
0
-1
-2
0.0
0.5
1.0
1.5
2.0
2.5
x

Turn over

6. A researcher is trying to find a "good" regression model for a set of 13 data values. He has one response variable y and 4 possible predictor variables, x_1, x_2, x_3 and x_4. You may assume that a model including standard assumptions of Normality is acceptable.

(i) Using relevant statistics from the table below, apply forward selection, testing at the 5% level, to find a suitable model.

(10)

Predictors in model apart from intercept	Residual sum of squares	R^2	Adjusted R^2	Mallows' C_p
–	2715.764			
x_1	1265.687	0.534	0.492	202.55
x_2	906.336	0.666	0.636	142.49
x_3	1939.401	0.286	0.221	315.16
x_4	883.867	0.675	0.645	138.73
x_1, x_2	57.905	0.979	0.974	2.68
x_1, x_3	1227.072	0.548	0.458	198.10
x_1, x_4	74.762	0.972	0.967	5.50
x_2, x_3	415.443	0.847	0.816	62.44
x_2, x_4	868.880	0.680	0.616	138.23
x_3, x_4	175.738	0.935	0.922	22.37
x_1, x_2, x_3	48.111	0.982	0.976	3.04
x_1, x_2, x_4	47.973	0.982	0.976	3.02
x_1, x_3, x_4	50.836	0.981	0.975	3.50
x_2, x_3, x_4	73.815	0.973	0.964	7.34
x_1, x_2, x_3, x_4	47.864	0.982	0.974	5.00

(ii) Would you expect to get the same model if you used backward elimination? (You should explain your answer briefly, but there is no need to do any working.)

(2)

(iii) Explain what Mallows' C_p is, and how it is used in model selection.

(2)

(iv) What other models would you suggest based on the statistics presented in the table?

(3)

(v) Explain why in practice it would be important to know about the practical situation and what the response and predictor variables are in order to select a model of practical usefulness.

(3)

7. A researcher is fitting a regression model predicting an observed variable y from another variable x, the values of which the researcher can control. The available data are given in the table, presented here in ascending order of values of x.

y	2.2	1.9	2.7	1.6	2.3	1.9	1.8	3.6	1.6	2.8	2.8
x	1.2	1.2	2.1	2.1	2.8	3.4	3.4	3.8	3.8	4.1	4.1

y	2.1	3.3	1.8	1.9	2.9	2.2	3.4	3.5	3.3	3.1	3.0
x	4.1	4.6	4.6	5.1	5.2	5.2	5.4	5.8	6.1	6.1	6.2

(i) Draw a scatter plot of the data. Using this, discuss the nature of any association between the variables and the nature of the variability in the data.

(5)

(ii) Explain how the information about the table helps to justify the researcher's decision to use a linear regression model to analyse the data, splitting the residual error term into two parts, for "lack of fit" and "pure error".

State the linear model underlying this analysis, and the properties of the terms in it.

(5)

(iii) (a) Show that the pure error SS from repeats at $x = 1.2$ is 0.045 and state the associated degrees of freedom.

(b) Calculate the pure error SS from repeats at $x = 4.1$ and state its associated degrees of freedom.

(c) Without further working state how the total pure error SS is computed.

(4)

(iv) Using the results of a linear regression, presented below, together with the fact that the pure error SS is 4.3717, carry out a lack of fit test, and state your conclusions.

(4)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	2.6723	2.6723	8.03	0.010
Residual Error	20	6.6554	0.3328		
Total	21	9.3277			

(v) Describe how lack of fit can be detected when there are no repeated observations of y at any x value.

(2)

8. The body of the following table gives data from a replicated two-factor experiment. The last row in the table gives totals for the six factor combinations. (The data have been coded, so Normality can be assumed as required.)

A1		A2		A3	
B1	B2	B1	B2	B1	B2
6	3	10	4	10	3
3	8	6	8	12	7
7	4	10	4	11	4
11	6	11	6	14	3
5	3	13	6	16	3
32	24	50	28	63	20

$$\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{5} y_{ijk} = 217\,; \quad \sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{5} y_{ijk}^2 = 1977$$

(i) Factor A is *fixed*. For each of the cases below, complete the ANOVA table, which should include showing the expected values of the mean squares, based on a suitable linear model including main effects and interaction. You should fully specify your models. State your conclusions in terms of statistical significance and practical meaning.

(a) B is a *fixed* factor.

(8)

(b) B is a *random* factor.

(8)

(ii) For each of the cases in (i), briefly describe a practical situation for which it would be appropriate.

(4)