

EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY



GRADUATE DIPLOMA, 2007

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  ${}^nC_r$ .*

This examination paper consists of 15 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.

1. (i) Write down the equations defining AR(1) and MA(1) models as used in time series analysis. Define all the terms used. (2)
- (ii) Show how an AR(1) model can be written as an infinite moving average model. (3)
- (iii) For each of the AR(1) and MA(1) models, derive the autocorrelation function and state, without working, the form of the partial autocorrelation function. (6)
- (iv) A stationary time series is given by

$$Y_t = 20 + 0.8Y_{t-1} + \varepsilon_t - 0.2\varepsilon_{t-1} ,$$

where the symbols have their usual meanings and properties.

- (a) Some time series are AR, some are MA, and some are a mixture of these two types. How may this time series be classified?
- (b) Derive its mean and variance.
- (c) Sketch the form of the autocorrelation and partial autocorrelation functions for this time series.
- (d) Explain how, in practice, given the sample autocorrelation and partial autocorrelation functions for a time series, you would set about identifying a suitable model. (9)

2. A dataset consists of five socio-economic variables for 12 areas in Los Angeles. The variables are as follows.

POP	total population in area
SCHOOL	median of the number of years spent in education at school
EMPLOY	total number of adults in employment
SERVICES	number of professional services available
HOUSE	median house value

- (i) Briefly explain the main purpose of principal component analysis for such data. (2)

- (ii) The correlation matrix for these data is given below.

	POP	SCHOOL	EMPLOY	SERVICES	HOUSE
POP	1.00	0.0097	0.9724	0.4389	0.0224
SCHOOL		1.00	0.1543	0.6914	0.8631
EMPLOY			1.00	0.5147	0.1219
SERVICES				1.00	0.7777
HOUSE					1.00

- (a) Describe the correlation structures in the variables. Hence suggest possible groupings of variables such that each group contains variables that are associated with each other.
- (b) Explain why for these data it is appropriate to carry out the principal component analysis on the correlation matrix rather than on the covariance matrix. (6)

- (iii) A statistical package gives the following output.

```

2 components were retained by the criterion 'eigenvalue > 1'
Eigenvalue    2.8733    1.7966
Coefficients:  POP          0.5810    0.8064
                SCHOOL       0.7670   -0.5448
                EMPLOY       0.6724    0.7261
                SERVICES     0.9324   -0.1043
                HOUSE       0.7912   -0.5582

```

- (a) Explain, without doing any numerical working, how these values are derived from the correlation matrix.
- (b) Interpret the two principal components presented in the output.
- (c) How much of the standardised variance is explained by these two components?
- (d) Explain why only two components have been presented, and comment on how sensible you believe this to be.
- (e) Is the criterion "eigenvalue > 1" always the best method of determining the number of components to retain? Justify your answer.

(12)

3. An archaeologist has collected data on a number of skulls. The following measurements (in mm) were taken on each skull.

X1      greatest length  
 X2      upper face height  
 X3      face breadth

The archaeologist believes that 17 of the skulls have come from one geographical area (Group 1) and that the remaining 15 are from a different area (Group 2). The archaeologist wishes to find a formula that will enable her to predict group membership for any skulls found in the future.

- (i) Comment critically on the use of multiple linear regression, with Group (coded as in part (iv) below) as response variable and X1, X2, X3 as regressor variables, as an approach to such prediction. (2)
- (ii) Describe the techniques of linear discriminant analysis and logistic regression, and explain under what circumstances they would be appropriate for these data. (5)
- (iii) Some output from a linear discriminant analysis is shown **on the next page**.
- (a) State how the pooled covariance matrix is derived from the covariance matrices of the two groups. Illustrate this by finding the estimate of the covariance between X1 and X2.
- (b) Use the information supplied to determine the linear discriminant function. (3)
- (iv) The output from a logistic regression is given below, where Group 1 is coded as 1 and Group 2 is coded as 0. What model has been fitted? Say, with reasons, whether you would want to fit alternative models. (3)

var1	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
x1	-.1017463	.0828389	-1.228	0.219	-.2641076 .0606150
x2	-.2198467	.1523823	-1.443	0.149	-.5185104 .0788171
x3	-.0974476	.1094728	-0.890	0.373	-.3120104 .1171151
_cons	47.69997	19.7242	2.418	0.016	9.041248 86.3587

- (v) For each of the two techniques (linear discriminant analysis and logistic regression), predict the group membership for a skull with X1 = 175, X2 = 71 and X3 = 133. (5)
- (vi) Which of the two techniques would you prefer for these data? Justify your answer. (2)

**Output from linear discriminant analysis is on the next page**

### Output from linear discriminant analysis for question 3

Variable	Pooled Mean	Means for Group	
		1	2
x1	179.94	174.82	185.73
x2	72.938	69.824	76.467
x3	133.70	130.35	137.50

Variable	Pooled Stdev	Stdev for Group	
		1	2
x1	7.682	6.748	8.627
x2	4.279	4.576	3.912
x3	6.610	8.137	4.238

Pooled Covariance Matrix

	x1	x2	x3
x1	59.01		
x2	20.12	18.31	
x3	20.11	12.99	43.70

Covariance Matrix for Group 1

	x1	x2	x3
x1	45.529		
x2	22.154	20.936	
x3	27.972	16.769	66.211

Covariance Matrix for Group 2

	x1	x2	x3
x1	74.424		
x2	17.794	15.302	
x3	11.125	8.661	17.964

Inverse of Pooled Covariance Matrix

0.0278	-0.0273	-0.0047
-0.0273	0.0959	-0.0160
-0.0047	-0.0160	0.0300

4. Data have been collected by a motor insurance company during a three-month period. They consist of the number of policy-holders who were exposed to a particular risk and the total number of claims during the three-month period which were associated with that risk. The data are classified by three 4-level factors, as described below.

DIST	the district in which the policy holder lived
CAR	the insurance group of the policy owner's car
AGE	the age of the policy holder

A generalised linear model is to be fitted, using a Poisson distribution.

- (i) The data are arranged in a table as defined by the three factors. Assume that accidents happen randomly and independently to the different policy-holders. Consider a cell of the table corresponding to a combination of levels of DIST, CAR and AGE. Let this be cell  $i$ , and let  $\lambda_i$  represent the mean number of claims made by a policy-holder in cell  $i$ . If there are  $N_i$  policy-holders in cell  $i$  and  $C_{ij}$  is the number of claims made by the  $j$ th policy holder in cell  $i$ , write down the mean for the number of claims in cell  $i$ , justifying your answer. (2)
- (ii) Someone suggests using a binomial model instead of a Poisson model. Comment on the appropriateness of this, giving your reasons. (2)
- (iii) A Poisson model is fitted, using a log link and an offset. Explain why the log link is the natural link for this model, and why an offset is required. (2)
- (iv) A series of models is fitted to the number of claims, where the explanatory variables are chosen from the three 4-level factors defined above. Some interactions between these factors are also considered. The order of entry of variables and interactions is determined by the prior knowledge of someone in the company. The results are summarised below.

<i>Model</i>	<i>Variables in model</i>	<i>Scaled deviance</i>
1	Constant only	238.16
2	Constant, DIST	225.57
3	Constant, DIST, CAR	137.39
4	Constant, DIST, CAR, AGE	51.01
5	Constant, DIST, CAR, AGE, CAR*AGE	40.06
6	Constant, DIST, CAR, AGE, CAR*AGE, DIST*CAR	32.64
7	Constant, DIST, CAR, AGE, CAR*AGE, DIST*CAR, DIST*AGE	26.52

On the basis of the information given, select the most parsimonious model, justifying your answer.

(10)

**Question 4 is continued on the next page**

- (v) A selection of the output from model 4 is given below. Here, Idista\_2 is a dummy variable for DIST = 2, Idista\_3 is a dummy variable for DIST = 3, and so on.

```

Poisson distribution, log link, offset log(n)
-----
      c |      Coef.   Std. Err.      z    P>|z|
-----+-----
Idista_2 |   .0232983   .0430152    0.542   0.588
Idista_3 |   .0362438   .0505071    0.718   0.473
Idista_4 |   .2318156   .0616708    3.759   0.000
  Icar_2 |   .1615003   .0505323    3.196   0.001
  Icar_3 |   .3970087   .0549985    7.219   0.000
  Icar_4 |   .5635516   .0723149    7.793   0.000
  Iage_2 |  -.1820974   .0828569   -2.198   0.028
  Iage_3 |  -.3448876   .0813732   -4.238   0.000
  Iage_4 |  -.5364148   .0699548   -7.668   0.000
   _cons |  -1.821786   .0767843  -23.726   0.000
      n | (exposure)
-----

```

Based on this output, suggest suitable re-codings of the factors DIST and AGE, justifying your answer.

(4)

5. (i) Briefly describe the advantages and disadvantages of the backward elimination method of model selection in multiple linear regression. (4)

- (ii) A dataset of 13 observations contains four predictor variables (X1, X2, X3, X4) and one response variable (Y), and the following statistics are available.

Variables in linear model	Model Sum of Squares
X1, X2, X3, X4	2667.90
X2, X3, X4	2641.95
X1, X3, X4	2664.93
X1, X2, X4	2667.79
X1, X2, X3	2667.65
X3, X4	2540.00
X2, X4	1846.88
X2, X3	2300.30
X1, X4	2641.00
X1, X3	1488.70
X1, X2	2657.90
X4	1831.90
X3	776.40
X2	1809.40
X1	1450.10
Total Sum of Squares	2715.76

Apply a backward elimination method to select the set of predictor variables that you consider "best" model the data. (8)

- (iii) Explain how your method of model selection might be different if you knew what the predictor variables were. (4)

- (iv) A journal gives the following advice to authors.

"Automated stepwise techniques often produce wildly unreliable results. This includes not only forward and backward automated selection but also 'best subset' approaches. Manuscripts that employ these techniques will not be considered unless the model is supported by a validation procedure."

- (a) Do you agree with the first sentence? Justify your answer.
- (b) Suggest possible validation procedures that could be used for the model chosen in part (ii). (4)



6. A scientist is studying air pollution and wishes to investigate how ozone concentration is related to wind speed, air temperature and the intensity of solar radiation.

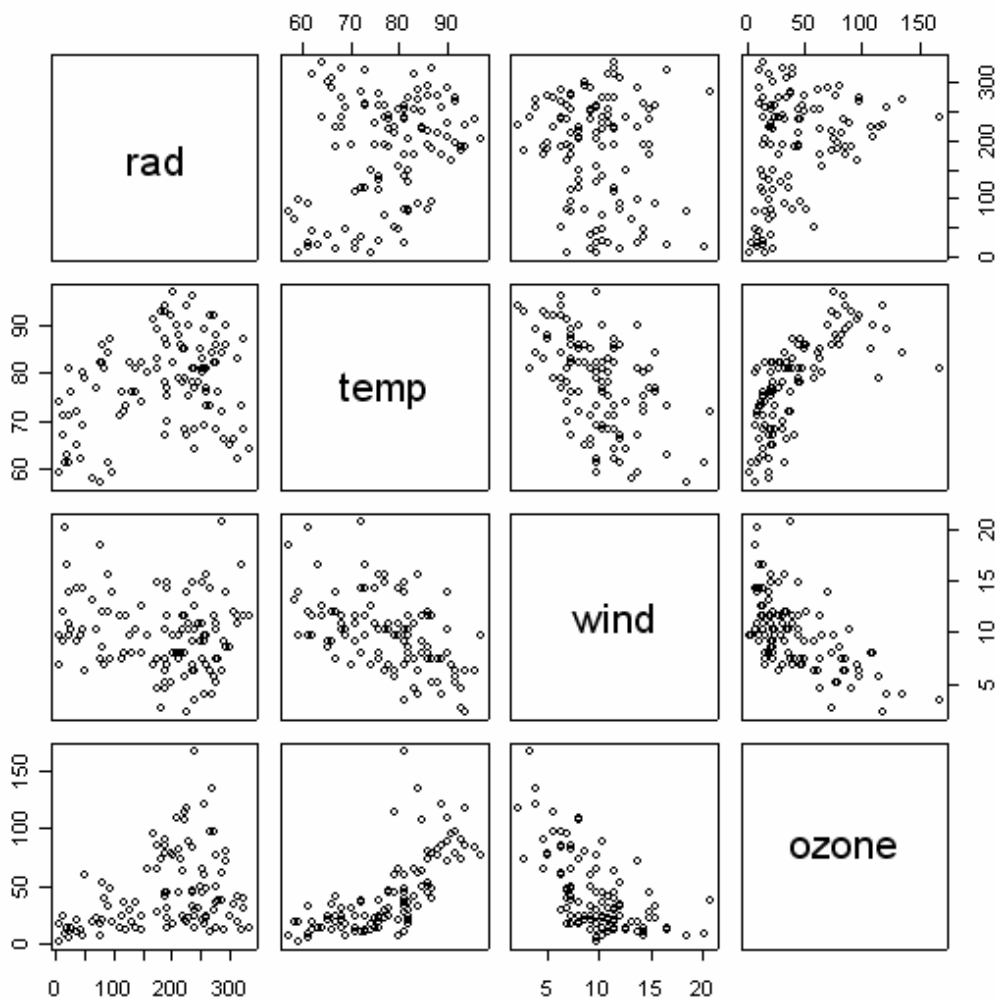
The variables are as follows.

rad	intensity of solar radiation (watts per square metre)
temp	air temperature (degrees Fahrenheit)
wind	wind speed (miles per hour)
ozone	ozone concentration (micrograms per cubic metre)

The dataset contains 111 observations.

- (i) Describe the information contained in the matrix plot below, and hence suggest possible relationships relevant to the scientist's question.

(3)



Question 6 is continued on the next page

- (ii) The scientist suspects that there might be some interactions between the predictor variables when trying to model air pollution.
- (a) Explain what is meant by an interaction in this context.
- (b) Write down the multiple regression model which includes all interactions, explaining each term in the model. (5)
- (iii) A colleague then suggests that there may also be quadratic effects of temperature, radiation and wind.

The output below shows the results of a multiple regression from such a model.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.683e+02	2.073e+02	2.741	0.00725
temp	-1.076e+01	4.303e+00	-2.501	0.01401
wind	-3.237e+01	1.173e+01	-2.760	0.00687
rad	-3.117e-01	5.585e-01	-0.558	0.57799
rad-squared	-3.619e-04	2.573e-04	-1.407	0.16265
temp-squared	5.833e-02	2.396e-02	2.435	0.01668
wind-squared	6.106e-01	1.469e-01	4.157	6.81e-05
temp:wind	2.377e-01	1.367e-01	1.739	0.08519
temp:rad	8.402e-03	7.512e-03	1.119	0.26602
wind:rad	2.054e-02	4.892e-02	0.420	0.67552
temp:wind:rad	-4.324e-04	6.595e-04	-0.656	0.51358

Residual standard error: 17.82 on 100 degrees of freedom  
 Multiple R-Squared: 0.7394, Adjusted R-squared: 0.7133  
 F-statistic: 28.37 on 10 and 100 DF, p-value: < 2.2e-16

In the table, the notation e indicates a power of 10. So, for example, 5.683e+02 means 568.3; and 6.106e-01 means 0.6106.

Explain how you would set about simplifying this model. (2)

- (iv) The scientist decides that the most parsimonious model contains main effects of radiation, temperature and wind speed, quadratic effects of temperature and wind speed only, and no interactions.

Based on the output **on the next two pages**, discuss the fit of this model. (6)

- (v) Describe how you would set about trying to improve the model, and the diagnostics you would look at to check your model. (4)

**Output for part (iv) is on the next two pages**

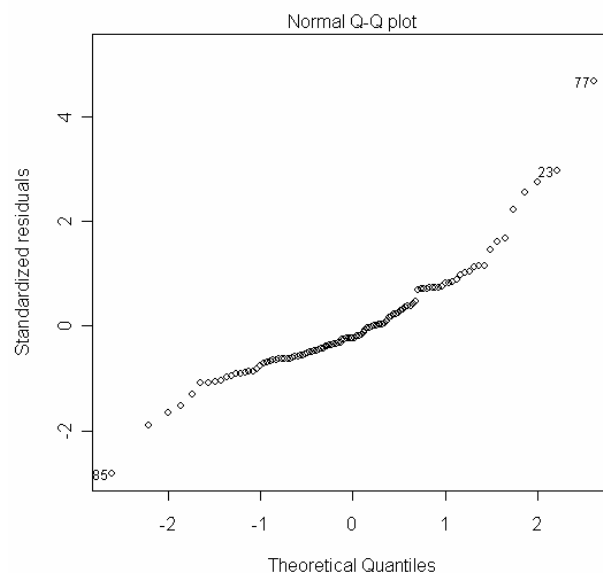
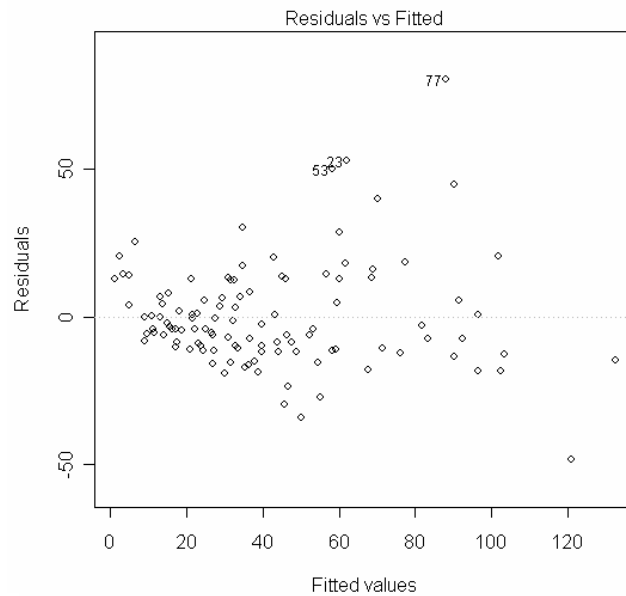
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	291.16758	100.87723	2.886	0.00473
rad	0.06586	0.02005	3.285	0.00139
temp	-6.33955	2.71627	-2.334	0.02150
wind	-13.39674	2.29623	-5.834	6.05e-08
temp-squared	0.05102	0.01774	2.876	0.00488
wind-squared	0.46464	0.10060	4.619	1.10e-05

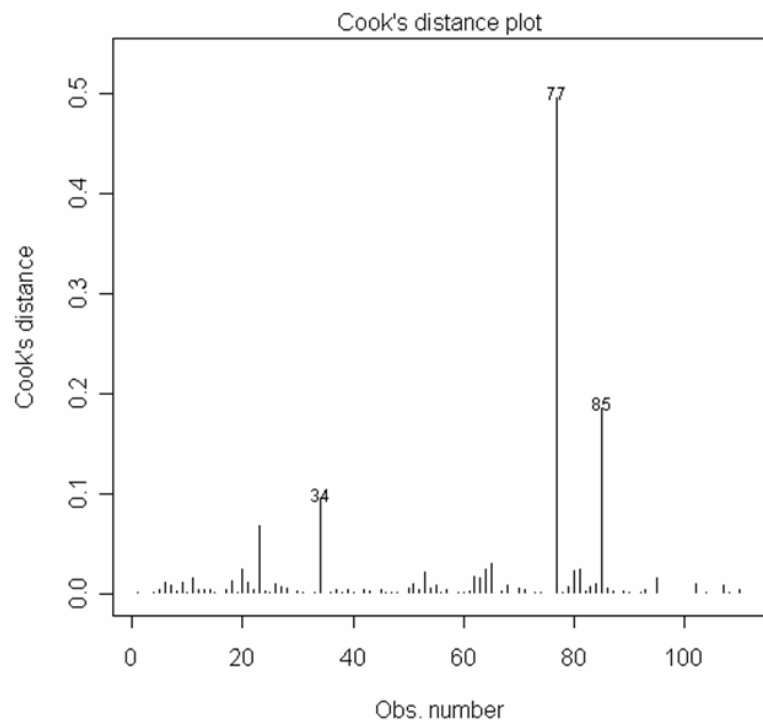
Residual standard error: 18.25 on 105 degrees of freedom

Multiple R-Squared: 0.713, Adjusted R-squared: 0.6994

F-statistic: 52.18 on 5 and 105 DF, p-value: < 2.2e-16



**Further output for part (iv) is on the next page**



7. An economist wishes to model the following data by fitting two straight lines with different slopes. The predictor variable is  $x$  and the response variable is  $Y$ . The first five observations ( $x = 1, 2, 3, 4, 5$ ) are to be fitted by one line, and the last five observations ( $x = 5, 6, 7, 8, 9$ ) by the other line; the point at  $x = 5$  is to be included in both lines.

$x$	$Y$
1	2.3
2	3.8
3	6.5
4	7.4
5	10.2
6	10.5
7	12.1
8	13.2
9	13.6

- (i) Plot the data, and describe the data briefly. (3)
- (ii) Explain why the following design matrix could be used to fit a suitable model, and interpret the coefficients in the model. (4)

$$\begin{pmatrix} 1 & -4 & 0 \\ 1 & -3 & 0 \\ 1 & -2 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 2 \\ 1 & 0 & 3 \\ 1 & 0 & 4 \end{pmatrix}$$

**Question 7 is continued on the next page**

- (iii) Someone else suggests the following design matrix. Is this equivalent to that in part (ii)? Justify your answer. (2)

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 0 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 5 & 3 \\ 1 & 5 & 4 \end{pmatrix}$$

- (iv) Derive the normal equations for the model in part (ii), and hence estimate the slopes of the two lines. (4)

- (v) In the model of part (ii), the two lines intersect at  $x = 5$ . Now suppose instead that you have **no** information about the point of intersection of the two lines.

- (a) Explain, without any working, how you might try to fit a model consisting of two intersecting straight lines. (3)

- (b) How would you decide which of models (A) and (B) was better?

(A) The model in part (v)(a).

(B) A curve (such as a quadratic).

(4)

8. A medical research worker wished to compare two different ways of treating a common type of skin rash. Eight medical clinics were chosen from those in a large area to take part in the study. The choice was made at random from clinics where there were at least three doctors available to take part. Four of the chosen clinics used Treatment A and the other four used Treatment B. Three doctors in each clinic (chosen at random if there were more than three available) took part, and each doctor chose four patients at random. These patients were examined again a week after treatment, and their recovery was assessed on a numerical scale.
- (i) The research worker in charge of the study decided to present the results in terms of the components of variance due to each stage of the study. Write down the model for this analysis, and state carefully the properties of each term in it. (4)
- (ii) Corrected sums of squares between treatments, clinics and doctors were respectively 4240.04, 6839.53 and 14269.11. The total (corrected) sum of squares was 39505.99. Write down the full analysis of variance table for this analysis, and use it to extract the components of variance the research worker required. (10)
- (iii) The research worker says he knows the treatments differ because he has carried out an  $F_{1,94}$  test and obtained a value near the borderline of statistical significance at 0.1%. Explain how he could have found this result, and comment on how valid it is. (2)
- (iv) Discuss critically the design of the study and the choice of analysis made by the research worker. (4)