EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2008

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 19 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) Define the terms *stationarity* and *weak stationarity* in the context of a time series.

Explain what is meant in practice when a time series is described as being stationary.

(4)

(ii) Derive the autocorrelation function (ACF) for each of the following stationary time series models.

(a) Moving average of order 2, i.e. MA(2).

(b) Autoregressive of order 1, i.e. AR(1).

(7)

(iii) The five plots **on the next three pages** show the ACFs for five different time series, labelled ts1, ts2, …, ts5. Two of these are non-stationary series, and three are stationary. Of the three stationary series, one is MA(2), one is AR(1) and the third is white noise.

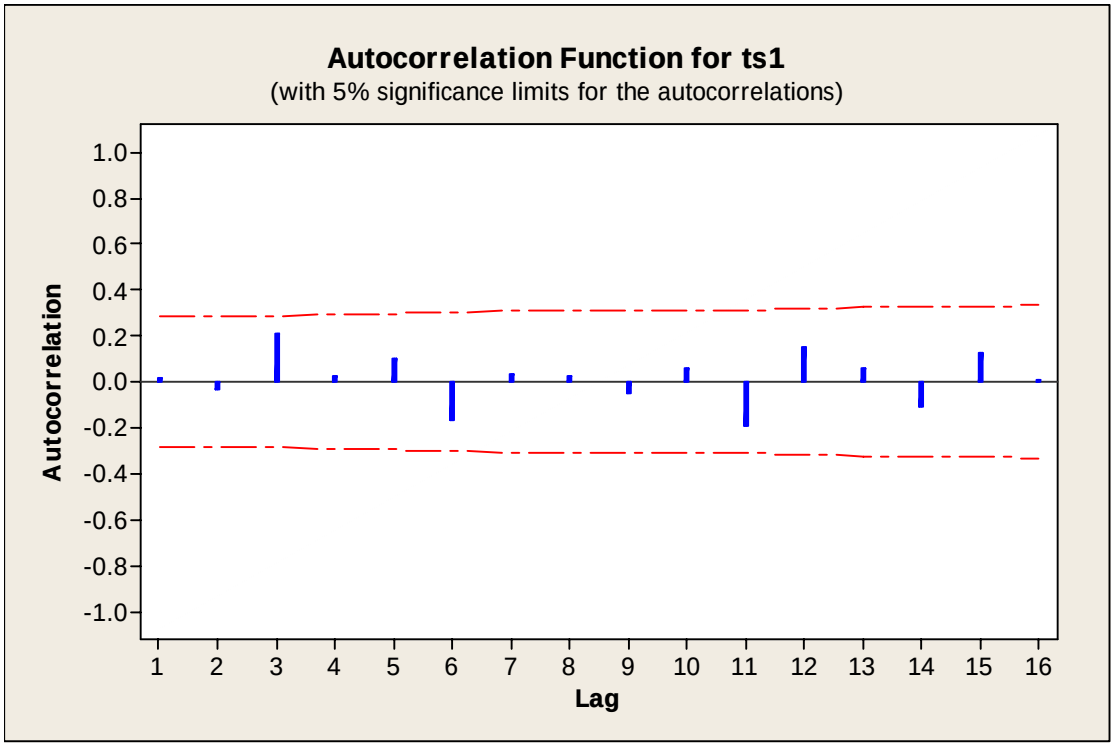
(a) Identify which of the plots correspond to the two non-stationary series, justifying your answer.

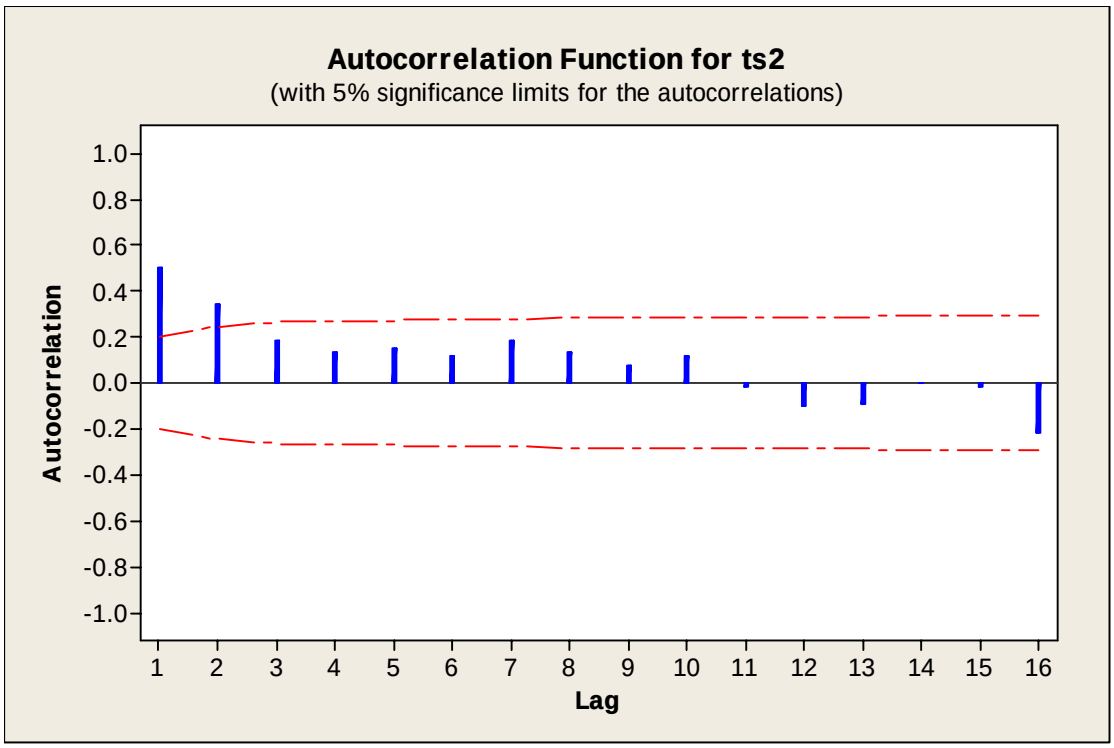
(2)

(b) Identify which plot corresponds to each model for the three stationary series, and for each model sketch the form of the partial autocorrelation function.

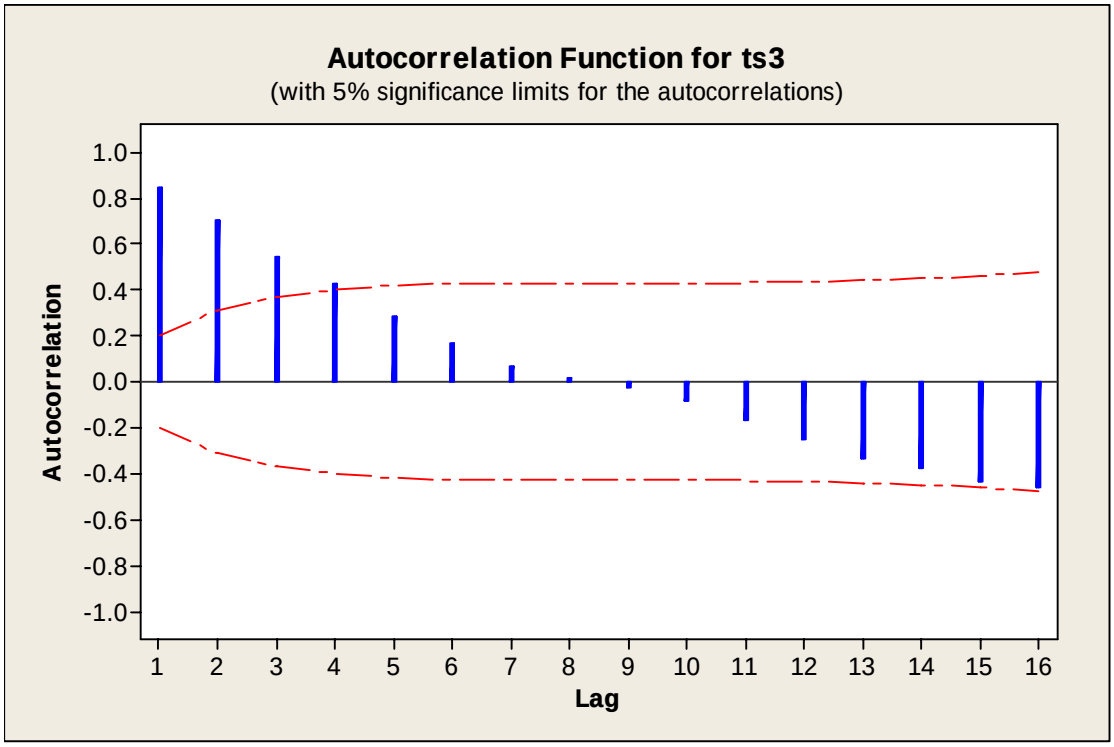
(7)

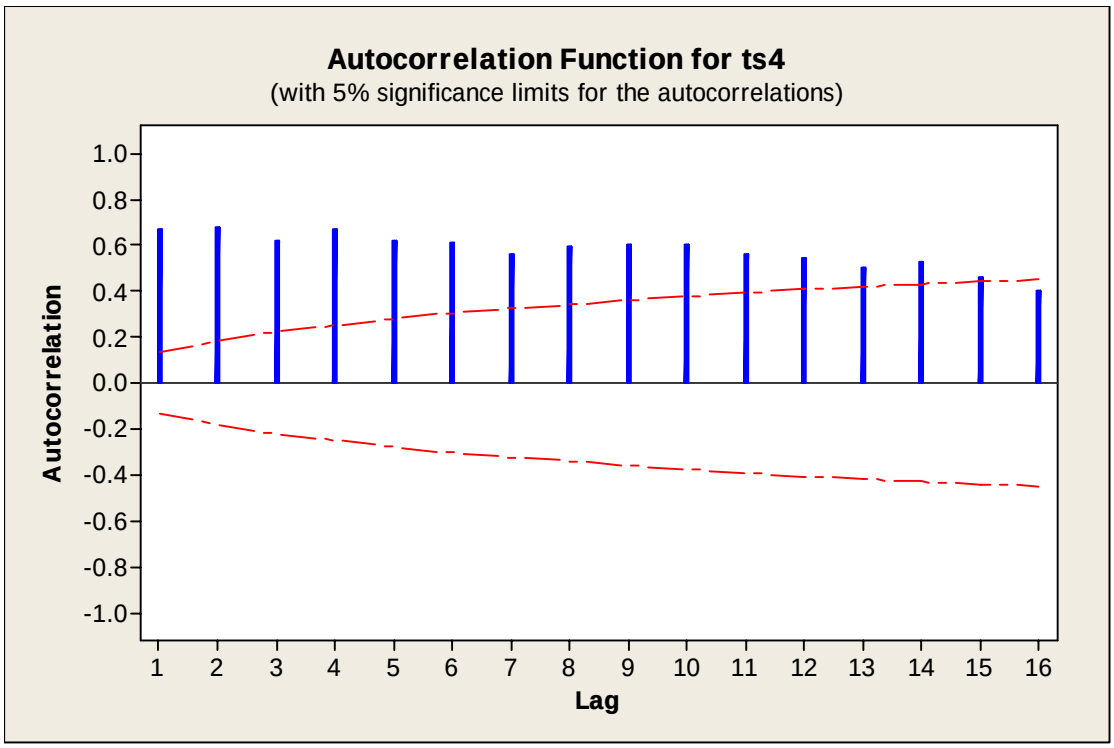
The five plots of ACFs are on the next three pages


Autocorrelation Function for ts1
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag

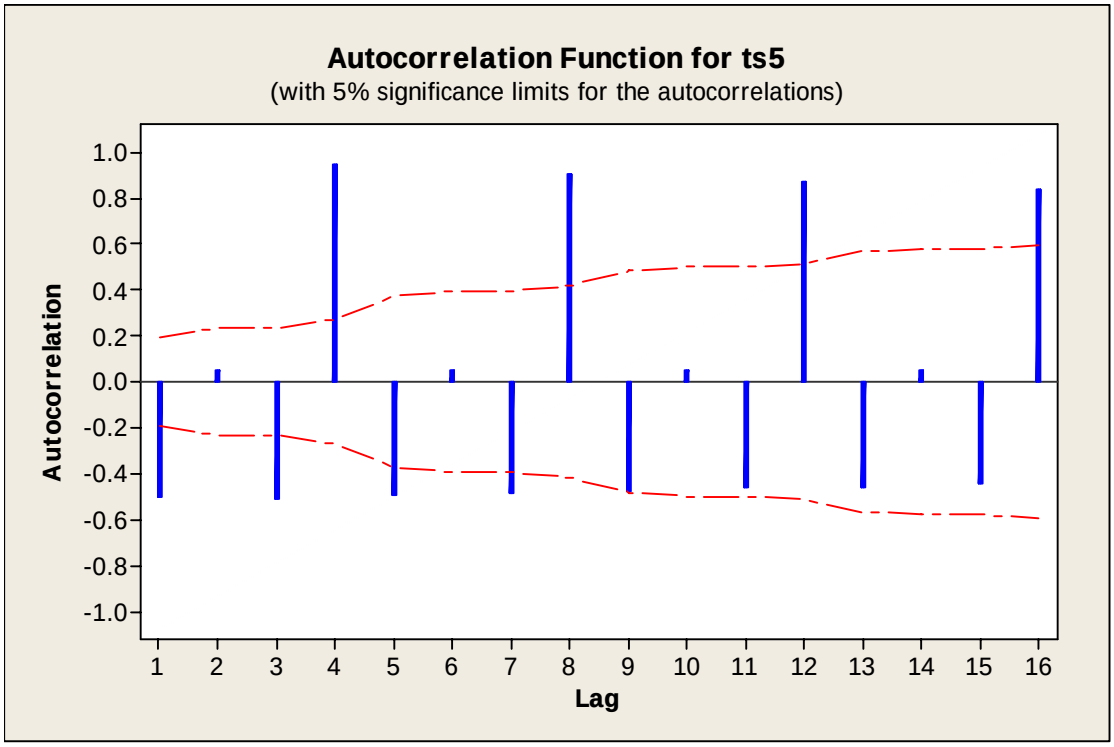

Autocorrelation Function for ts2
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag

Turn over


Autocorrelation Function for ts3
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag


Autocorrelation Function for ts4
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag

Turn over


Autocorrelation Function for ts5
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag

Turn over

2. Data were collected from students at school in the USA. Students were asked about their use of 13 types of psychoactive substances. The substances were cigarettes, beer, wine, spirits, cocaine, tranquillizers, drug store medications, heroin, marijuana, hashish, glue, LSD, stimulants. At the time of the study the legal substances were cigarettes, beer, wine, spirits.

The students were asked to rate how often they used each substance using a 5-point scale as follows.

1 (never tried)
2 (tried only once)
3 (tried a few times)
4 (tried many times)
5 (tried regularly)

The (Pearson) correlation matrix for the data is as follows.

	Cigarettes	Beer	Wine	Spirits	Cocaine	Tr'qu'zers
Cigarettes	1.000					
Beer	0.447	1.000				
Wine	0.422	0.619	1.000			
Spirits	0.435	0.604	0.583	1.000		
Cocaine	0.114	0.068	0.053	0.115	1.000	
Tranquillizers	0.203	0.146	0.139	0.258	0.349	1.000
Medications	0.091	0.103	0.110	0.122	0.209	0.221
Heroin	0.082	0.063	0.066	0.097	0.321	0.355
Marijuana	0.513	0.445	0.365	0.482	0.186	0.315
Hashish	0.304	0.318	0.240	0.368	0.303	0.377
Glue	0.245	0.203	0.183	0.255	0.272	0.323
LSD	0.101	0.088	0.074	0.139	0.279	0.367
Stimulants	0.245	0.199	0.184	0.293	0.278	0.545

	Medic'ns	Heroin	Marijuana	Hashish	Glue	LSD	Stim'nts
Cigarettes							
Beer							
Wine							
Spirits							
Cocaine							
Tranquillizers							
Medications	1.000						
Heroin	0.201	1.000					
Marijuana	0.150	0.154	1.000				
Hashish	0.163	0.219	0.534	1.000			
Glue	0.310	0.288	0.301	0.302	1.000		
LSD	0.232	0.320	0.204	0.368	0.340	1.000	
Stimulants	0.232	0.314	0.394	0.467	0.392	0.511	1.000

Question 2 is continued on the next page

(i) Comment on the appropriateness of Pearson's correlations to summarise this type of data, and describe the main correlation structure.

(3)

(ii) The researchers carried out a principal component analysis on this correlation matrix. The eigenvalues and eigenvectors are given below.

Eigenvalues

4.380, 2.046, 0.953, 0.817, 0.766, 0.687, 0.644, 0.615, 0.561, 0.399, 0.395, 0.374, 0.363

Eigenvectors

1	2	3	4	5	6	7
-0.278	-0.280	-0.060	-0.016	-0.316	-0.463	0.123
-0.286	-0.397	0.130	-0.101	0.177	0.150	-0.110
-0.265	-0.392	0.220	-0.142	0.308	0.160	-0.063
-0.318	-0.325	0.053	-0.063	0.178	0.164	0.002
-0.208	0.288	0.052	-0.591	-0.438	0.356	-0.295
-0.293	0.259	-0.172	-0.086	0.125	0.054	0.551
-0.176	0.189	0.727	0.330	-0.248	0.261	0.338
-0.202	0.315	0.147	-0.532	0.324	-0.392	0.162
-0.339	-0.164	-0.236	0.109	-0.358	-0.146	0.131
-0.329	0.051	-0.355	0.118	-0.247	0.247	-0.124
-0.276	0.169	0.314	0.195	-0.084	-0.506	-0.472
-0.248	0.329	-0.109	0.288	0.358	0.146	-0.402
-0.328	0.232	-0.232	0.264	0.210	0.019	0.144

8	9	10	11	12	13
0.028	-0.620	-0.135	0.142	-0.218	0.201
-0.054	0.047	-0.104	0.072	0.637	0.492
0.072	-0.092	-0.422	-0.215	-0.160	-0.565
0.120	0.200	0.623	0.159	-0.489	0.149
0.226	-0.180	0.086	-0.158	0.027	0.007
0.482	0.078	-0.119	0.431	0.174	-0.140
-0.209	-0.070	0.006	0.005	-0.035	0.041
-0.502	0.106	0.022	-0.040	-0.057	0.048
-0.253	0.141	0.375	-0.193	0.379	-0.467
-0.375	0.348	-0.447	0.221	-0.291	0.128
0.354	0.368	-0.079	0.063	0.019	-0.036
-0.189	-0.480	0.185	0.289	0.107	-0.171
0.178	-0.058	-0.035	-0.720	-0.085	0.296

(a) Write down an equation showing how the correlation matrix, eigenvalues and eigenvectors are related.

(b) Interpret the first two principal components.

(c) How many principal components would you choose to describe the data? Justify your answer.

(8)

Question 2 is continued on the next page

(iii) (a) Carry out a hierarchical cluster analysis on the 13 variables. Use single linkage on the correlation matrix.

(b) Draw a dendrogram to show the clusters.

(c) Interpret the results of the cluster analysis.

(9)

3. (i) Briefly explain how **any three** of the following regression diagnostic statistics are used in linear regression analysis.

Leverages

Deleted *t* residuals (i.e. Studentized deleted residuals)

Standardised residuals

Cook's distance

(4)

(ii) An experiment was carried out to investigate how effective radiation is in killing bacteria. Fourteen plates containing bacteria were exposed to different doses of radiation, and the researchers measured the proportions of bacteria surviving.

The researchers hypothesised that the logarithm of the proportion surviving should be linearly related to dose. They therefore created a variable *logprop*, the natural logarithm of the proportion, and regressed this variable on dose.

The results are given in the table below (in coded units).

Dose	Proportion of bacteria surviving, *prop*	*logprop*
1.175	0.4400	–0.8210
1.175	0.5500	–0.5978
2.350	0.1600	–1.8326
2.350	0.1300	–2.0402
4.700	0.0400	–3.2189
4.700	0.0196	–3.9322
4.700	0.0612	–2.7936
7.050	0.0050	–5.2983
7.050	0.0032	–5.7446
9.400	0.0011	–6.8124
9.400	0.0002	–8.5172
9.400	0.0002	–8.5172
14.100	0.0070	–4.9618
14.100	0.0001	–9.2103

Some output from statistical software is shown below.

```
The regression equation is
logprop = -0.851 - 0.572 dose

Predictor      Coef   SE Coef       T       P
Constant   -0.8513    0.7851   -1.08   0.300
dose       -0.5715    0.1012   -5.65   0.000

S = 1.57530     R-Sq = 72.7%     R-Sq(adj) = 70.4%
```

Question 3 is continued on the next page

(a) Draw a suitable graph to explore the possible relationship between the two variables. Describe its main features.

(4)

(b) Write down the statistical model that would correspond to the regression output if *prop* instead of *logprop* were the response.

(2)

(c) Further output from the analysis is given below. Using the computer output above, together with this extra information, discuss the adequacy of the model.

Observation	Standardised residual	Deleted t residual	Leverage	Cook's distance
1	0.495	0.480	0.191	0.029
2	0.653	0.636	0.191	0.050
3	0.248	0.238	0.144	0.005
4	0.106	0.101	0.144	0.001
5	0.211	0.203	0.086	0.002
6	-0.262	-0.252	0.086	0.003
7	0.494	0.478	0.086	0.011
8	-0.275	-0.264	0.072	0.003
9	-0.570	-0.553	0.072	0.013
10	-0.395	-0.381	0.105	0.009
11	-1.539	-1.645	0.105	0.139
12	-1.539	-1.645	0.105	0.139
13	3.011	5.827	0.307	2.008
14	-0.229	-0.220	0.307	0.012

(6)

(d) What further analysis would you do on these data? Justify your answer.

(4)

4. (i) Consider the linear model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$\mathbf{Y}$ is an $(n \times 1)$ vector of observations,
$\mathbf{X}$ is an $(n \times p)$ design matrix of known form,
$\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters,
$\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of errors.

(a) State the *Gauss-Markov theorem* for least squares estimators. (3)

(b) Write down the expression for the (residual) error sum of squares in matrix form and derive the normal equations. Hence show that the least squares estimator for this model is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y},$$

and state the conditions under which this result is valid. (6)

(c) If the errors are Normally distributed, what further useful properties can you state about $\hat{\boldsymbol{\beta}}$? (2)

(ii) A researcher has done an experiment to compare the effects of fasting times prior to anaesthetic for a short operation. She has randomly divided a set of patients into two groups. One group has been instructed to fast for a long period of time, but the other group has been told that they can drink up to four hours before the operation. She has measured the volume of residual gastric juice for each patient just prior to the operation. She has done a two-sample *t* test to compare the two groups.

(a) Write down a linear regression model that could also be used for the analysis of these data, stating the design matrix. Explain how the two-sample *t* test done by the researcher is equivalent to a *t* test that would arise from the linear regression model.

(4)

(b) Although the researcher randomised the patients to the groups, she is concerned that the groups appear to be imbalanced with regard to two important variables: sex and whether or not the patient is a smoker. She believes that the volume of gastric juice is related to sex and smoking, and she is especially concerned that the results for female smokers may be different from the others. She wants to use a test for the effect of fasting time that takes account of sex and smoking status.

Describe a suitable model for analysing the data. Explain and justify your choice of parameters, and state how you would expect the results from this model to be different from those of the initial two-sample *t* test.

(5)

5. (i) Briefly compare and contrast the statistical methods of linear discriminant analysis and logistic regression.

(4)

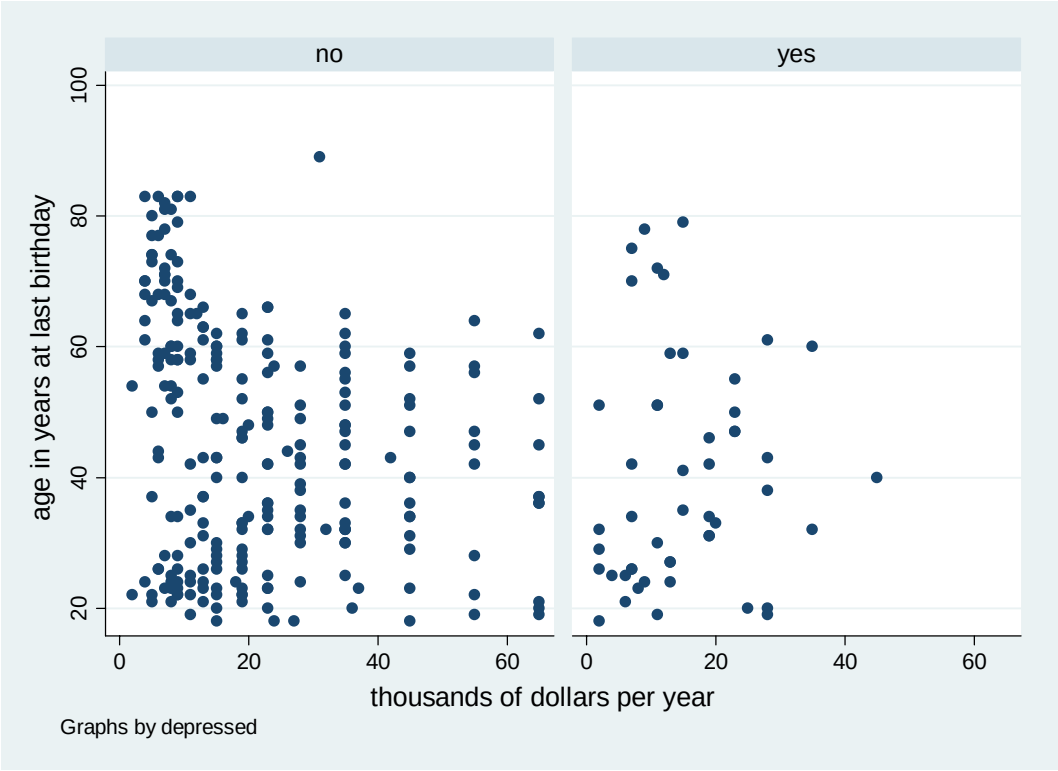
(ii) Data were collected on 294 patients presenting at a clinic. Of the 294, fifty were classified as being depressed, and the other 244 were normal (i.e. not depressed). Other available data were the age in years at the last birthday, and income in thousands of dollars per year for each patient.

(a) Summary statistics and scatter plots are given below. Describe the data, and discuss the suitability of discriminant analysis and logistic regression to analyse these data.

(4)

Depressed	*Age Mean(SD)*	*Income Mean(SD)*
Yes	40.38 (17.40)	15.20 (9.84)
No	45.24 (18.15)	21.68 (15.98)

Plots of age against income for each of the two groups (depressed and not depressed)



Question 5 is continued on the next page

(b) A researcher has run both methods on the data, and the results are shown **on the next page**.

Interpret each set of output.

(6)

(c) How would you explain to the researcher why neither method is particularly suitable for this data set?

(2)

(d) Describe what further analysis you would do to assess how the two methods would perform on a new, similar, set of data.

(4)

The results for the researcher's run of both methods are on the next page

Discriminant Analysis

Prior probabilities = 0.500

Summary of classification

	True Group	
Put into Group	depressed	normal
depressed	30	90
normal	20	154
Total N	50	244

Linear Discriminant Function

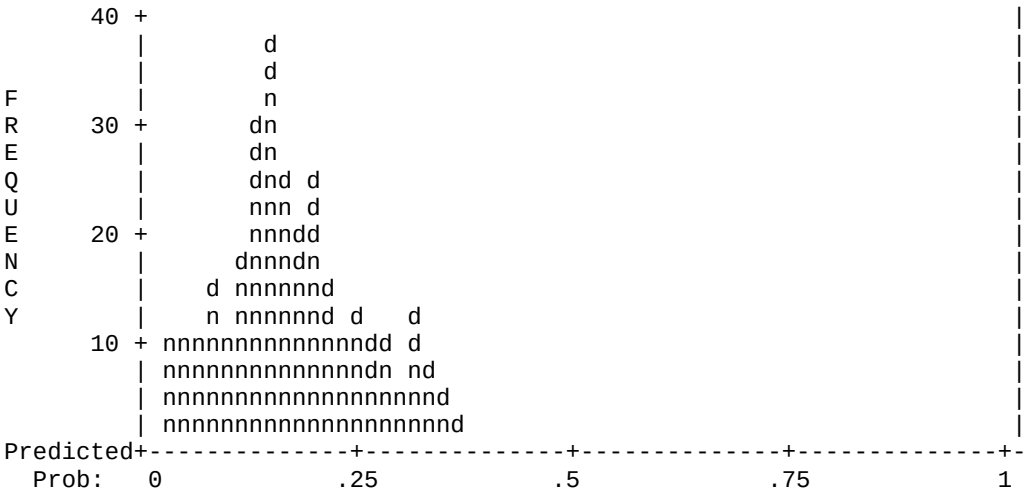
Constant	1.5157
age	-0.0209
income	-0.0336

Logistic regression

Scaled deviance = 254.03

cases	Coef.	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
income	-0.0413479	0.0140587	-2.94	0.003	-0.0689025	-0.0137933
age	-0.0201671	0.0088966	-2.27	0.023	-0.0376041	-0.0027301
constant	0.0279773	0.4872007	0.06	0.954	-0.9269184	0.9828731

Observed Groups and Predicted Probabilities



Group: nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnddddddddddddddddddddddddddddddd

Predicted Probability is of Membership for depressed

The Cut Value is 0.50

Symbols: n - normal
d - depressed
Each Symbol Represents 2.5 Cases.

Turn over

6. (i) Derive the natural link function for a Poisson distribution, and explain how this is used in generalised linear modelling.

(4)

(ii) Quality inspectors have expressed a concern that students on a particular year-long course do not get enough help and attention from their tutors. Data were collected over three successive years from students on the course. Each student was asked whether the level of help and attention offered was low, medium or high. The researchers were particularly interested in trends over time. The data are given below.

	Reported level of help/attention			
Year	*low*	*medium*	*high*	*Total*
2004	10	72	43	125
2005	16	37	41	94
2006	12	38	35	85
				304

(a) Convert the data in the nine cells of the table into suitable proportions, and discuss whether there are any evident trends over time. Explain why a log-linear model may be appropriate to analyse these data.

(3)

(b) A log-linear model was fitted, using the following parametrisation for the nine cells of the table.

$$\begin{array}{ccc} \mu & \mu+\alpha_2 & \mu+\alpha_3 \\ \mu+\beta_2 & \mu+\alpha_2+\beta_2+(\alpha\beta)_{22} & \mu+\alpha_3+\beta_2+(\alpha\beta)_{32} \\ \mu+\beta_3 & \mu+\alpha_2+\beta_3+(\alpha\beta)_{23} & \mu+\alpha_3+\beta_3+(\alpha\beta)_{33} \end{array}$$

State the scaled deviance and degrees of freedom for this model. Justify your answer.

(2)

(c) Modify the parametrisation in part (b) so that the distribution of level of help/attention is the same for each year.

(2)

(d) A further model is fitted with the following parametrisation.

$$\begin{array}{ccc} \mu & \mu+\alpha_2 & \mu+\alpha_3 \\ \mu+\beta_2 & \mu+\alpha_2+\beta_2+C & \mu+\alpha_3+\beta_2 \\ \mu+\beta_3 & \mu+\alpha_2+\beta_3+C & \mu+\alpha_3+\beta_3 \end{array}$$

State a hypothesis that is consistent with this parametrisation.

(2)

Question 6 is continued on the next page

(e) The scaled deviance for the model in part (c) is 9.18. That for the model in part (d) is 1.89. Of the three models described in parts (b), (c) and (d), which do you prefer? Do you think there is evidence of a trend over time? Justify your answers.

(4)

(f) What other analyses would you do to check your chosen model, or to examine a different model?

(3)

7. Data were collected on 256 children under five years old in an African country. The aim was to predict N_score (defined below) from other variables. Several variables were recorded for each child, and stepwise regression selected the nine listed below.

N_score	a score showing how well-nourished the child was (higher scores describe a child that is better nourished)
Sex	male = 1 female = 0
Age	age of child in months
Mother's education	no formal education = 0 some formal education = 1
Marital status of mother	widowed = 1 not widowed (i.e. married, never married, divorced or separated) = 0
Malaria	present = 1 absent = 0
Fever	present = 1 absent = 0
Skip	how often the child had skipped meals in the previous two months often = 1 rarely or never = 0
SES	socio-economic status of family (the higher the value, the better the status)
Dependency	ratio of number of wage earners in household earning to total number in household

The researchers report that "Stepwise linear regression was used to predict N_score from the other variables, using a significance level of 5%". Some output is shown below.

Variable	*Estimate of beta from model*
Sex	0.184
Age	0.240
Mother's education	0.197
Marital status	–0.290
Malaria	0.651
Fever	0.461
Skip	0.729
SES	0.192
Dependency	–0.520

Where appropriate, use this output to answer the following questions about the analysis.

Question 7 is continued on the next page

(i) Explain what is meant by *stepwise regression*, and comment critically on its usefulness in multiple regression.

(2)

(ii) The researchers say that they "checked all continuous variables for Normality" and that they "produced graphs to explore pairs of variables and to identify univariate and bivariate outliers". Discuss the usefulness of these checks for a multiple regression analysis.

(4)

(iii) Comment on the way in which marital status has been coded.

(2)

(iv) The researchers say that "the prevalence of under-nourishment rose with age but peaked at 18.3 months and then dropped". Comment on the implications of this pattern of under-nourishment for the regression model.

(2)

(v) The researchers make the following statements.

> "Children whose mothers had some formal education were less likely to be badly nourished. A higher percentage of boys were malnourished than girls. The prevalence of malnutrition was significantly reduced with an increase in SES. Although the adjusted R-squared is low, this could be a result of a small sample size. However, it was possible to identify important trends. In particular, the impact of malaria, child's age, dependency and skipping meals appeared to be greater than sex, fever, mother being widowed, mother's education level and SES."

Comment critically on the validity of these statements.

(6)

(vi) What else would you do to check the fit of the model?

(4)

8. A new procedure has been implemented in a hospital, and the managers want to study employees' satisfaction with it. In particular they are interested in differences between the satisfaction of two professional groups: doctors and nurses. They also wish to check whether there is any statistical interaction between the professional groups and the wards where they work. The managers use a questionnaire which grades satisfaction from 0 to 20, where 20 is the best possible satisfaction and 0 is the worst possible. They select 3 wards in the hospital and then randomly select 5 doctors and 5 nurses from each ward to answer the questionnaire. In this way they obtain the satisfaction scores for 15 doctors and 15 nurses.

They plan to analyse these scores using analysis of variance.

(i) Comment on the suitability of analysis of variance for this response variable.

(2)

(ii) Is professional group (i.e. doctor or nurse) a fixed factor or a random factor? Justify your answer.

(2)

(iii) Describe two situations, the first where "ward" would be a fixed factor, and the second where "ward" would be a random factor.

(2)

(iv) For each of the two situations in part (iii), write down the model for the analysis of variance, explaining each of the terms and the assumptions of the model.

(5)

(v) The table below shows the data and some summary statistics for the data. Assuming that the model is correct, carry out an analysis of variance for the case where "ward" is a random factor and interpret the results.

(7)

Ward	*Doctors*	*Nurses*
1	12, 14, 10, 11, 9 (sum=56)	14, 12, 14, 16, 17 (sum=73)
2	10, 9, 8, 12, 16 (sum=55)	15, 16, 17, 15, 9 (sum=72)
3	12, 13, 9, 8, 12 (sum=54)	7, 8 ,6, 8, 14 (sum=43)

$$\sum y_{ijk} = 353$$

$$\sum y_{ijk}^{\ 2} - \left(\sum y_{ijk}\right)^2 / 30 = 301.3667$$

(v) How would your conclusions be different if "ward" was a fixed factor?

(2)