EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2002

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 silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

Note that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$ *and that* ln *stands for* $\log_e$.

This examination paper consists of 18 printed pages. This front cover is page 1. The reverse of the front cover, which is intentionally left blank, is page 2. Question 1 starts on page 3.

There are 8 questions altogether in the paper.

1. (i) Write down the general forms of equations for *AR(p)* and *MA(q)* stationary time series processes, taking care to define all terms and symbols used and to state any necessary assumptions.

(5)

(ii) For each of the following time series, in which the symbols have their usual meanings, calculate the mean, variance and autocorrelation function:

(a) $Y_t = 95 + \varepsilon_t - 0.7\varepsilon_{t-1}$

(b) $Y_t = 68 - 0.5Y_{t-2} + \varepsilon_t$

[Note. The second term on the right has subscript $t - 2$, **not** $t - 1$.]

(12)

(iii) Express the model

$$Y_t = 90 - 0.8Y_{t-1} + \varepsilon_t$$

as an infinite moving average process.

(3)

[Note. Standard results for ARMA processes may be assumed without proof.]

2. An educationalist is studying the effect of feedback on children's performance. He chooses two classes of children from a school. Each class does a test on two occasions, giving a pre-score and a post-score. One class (the treatment group) gets positive feedback after the first test; the other (the control group) gets no feedback.

(i) Briefly describe the major problem with this design.

(1)

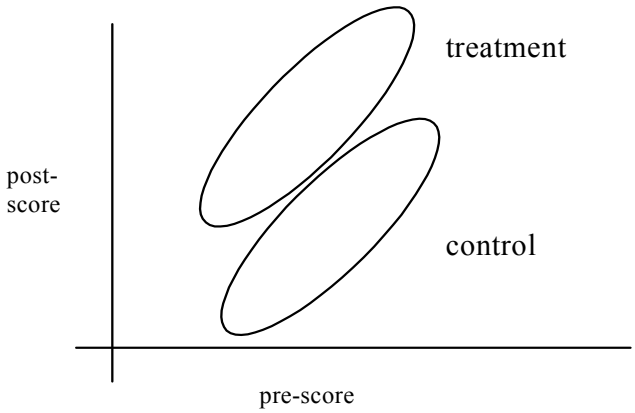
(ii) The figure below shows a diagrammatic representation of the case where the two classes have a similar distribution of pre-score marks, but where the average post-score for the treatment group is higher.

(a) Explain why a simple one-way analysis of variance can be used to estimate the differential treatment effect.

(b) Write down the form of this analysis of variance model, and explain the meaning of each term in it.

(c) Copy the figure into your answer book and annotate it to illustrate your answers to (a).

(6)



(iii) (a) Write down the form of the model for an analysis of covariance where the response is the post-score and the covariate is the pre-score.

(b) Draw a figure to illustrate a case where such an analysis of covariance would give a different estimate of treatment effect from that obtained using an analysis of variance.

(c) When using analysis of covariance, would there be any advantage in using the difference (post-score – pre-score) as the response variable? Justify your answer.

(8)

Question 2 is continued on the next page

(iv) (a) What does analysis of covariance assume about the slope of the regression lines for the two groups?

(b) Write down an extension to your model in (iii)(a) which does not make this assumption.

(c) Draw a figure to illustrate a case in which the model in (iv)(b) might be useful, and describe why.

(d) How would you estimate the differential treatment effect in this case? Carefully justify your answer.

(5)

3. On the next three pages are shown sets of residual plots from three case studies in regression, which are outlined below. In each study, the regression is to be used for prediction purposes.

Case Study 1 (plots for this are on page 7)

The response variable is the normal average minimum temperature (degrees Celsius (Centigrade)) for 56 cities in the USA. The predictor variable currently in the model is latitude (degrees north). Another possible predictor variable is longitude (degrees west), and it has also been suggested that "distance from the sea" might be a useful predictor variable.

Case Study 2 (plots for this are on page 8)

The response variable is the strength of a timber beam and the predictor variable is the specific gravity, for 10 beams.

Case Study 3 (plots for this are on page 9)

The response variable is consumer expenditure and the predictor variable is gross national product (GNP), for 18 years in the USA. The data are in $ billion, all at 1958 prices.

For each study, and making reference to the plots, answer the following.

(i) Describe the residual plots, clearly identifying any apparent problems with the assumptions that are required for validity of the analysis.

(5)

(ii) For each problem identified, describe other tests, diagnostic plots or methods that would help to identify the cause of the problem.

(7)

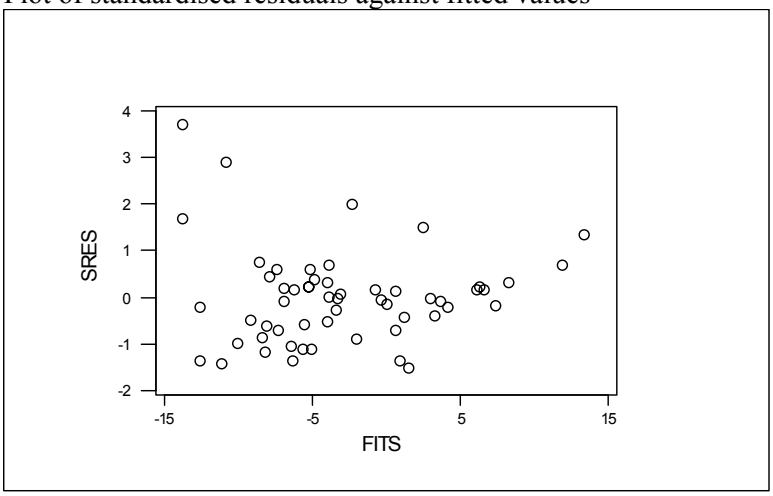
(iii) State what you would do to try to overcome the problems.

(8)

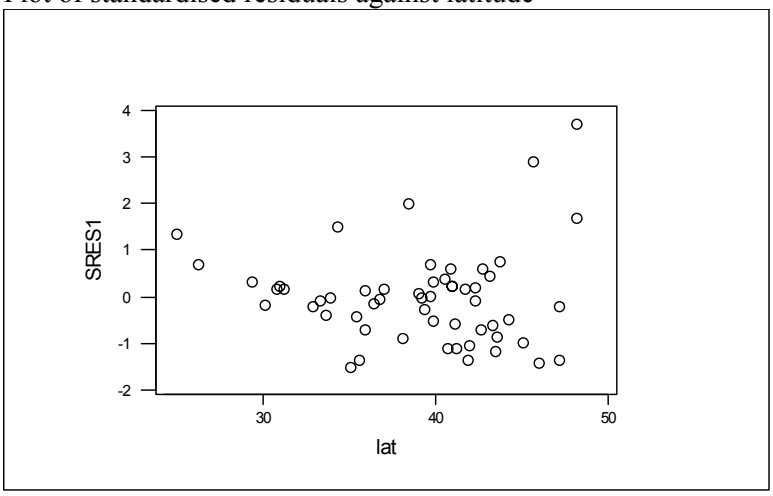
Plots for the case studies in question 3 are shown on the following pages

Case study 1

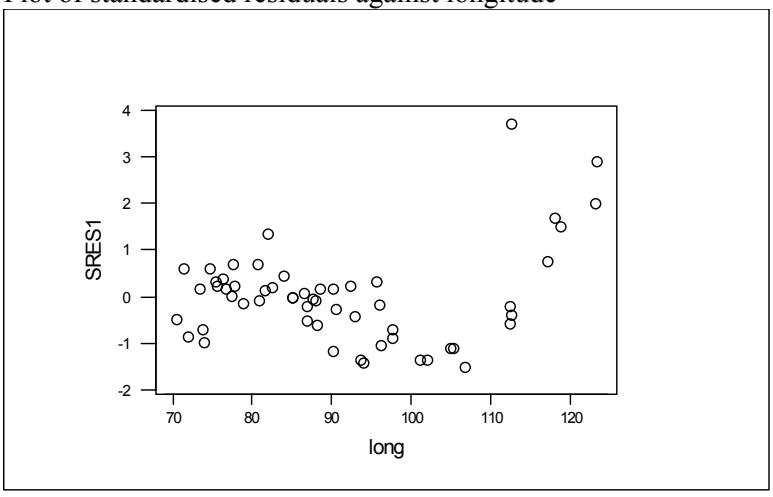
Plot of standardised residuals against fitted values



Plot of standardised residuals against latitude



Plot of standardised residuals against longitude



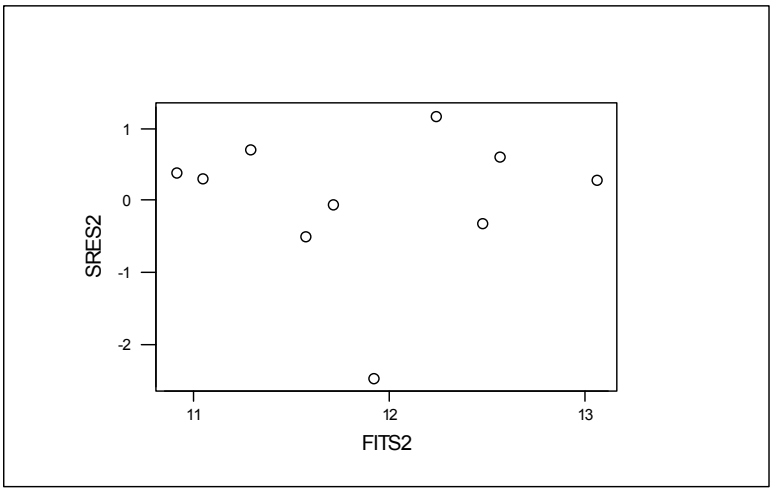
Plots for the other case studies in question 3 are shown on the following pages

Turn over

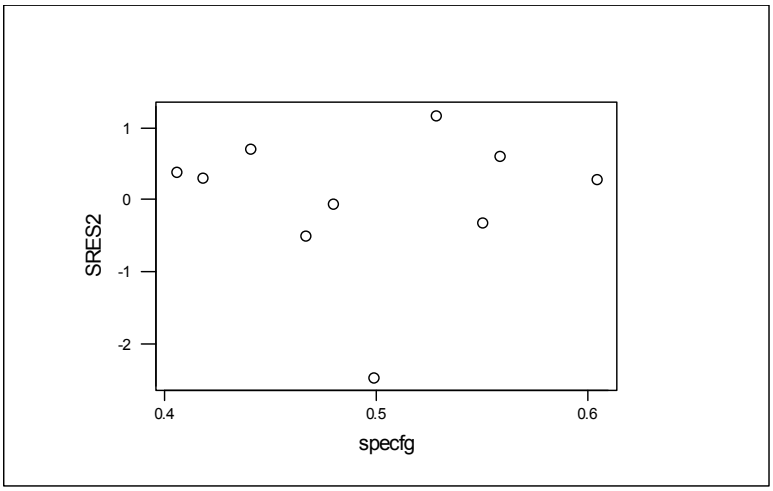
Case study 2

[Reminder. The response variable is the strength of a timber beam and the predictor variable is the specific gravity, for 10 beams.]

Plot of standardised residuals against fitted values



Plot of standardised residuals against specific gravity

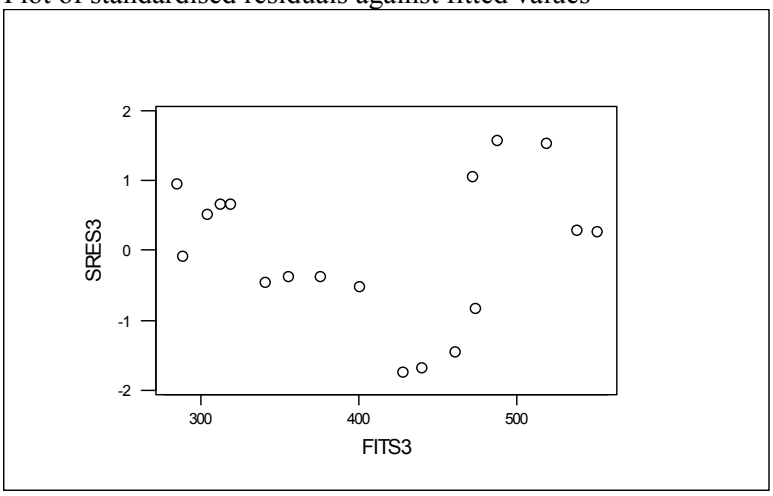


Plots for the remaining case study in question 3 are shown on the following page

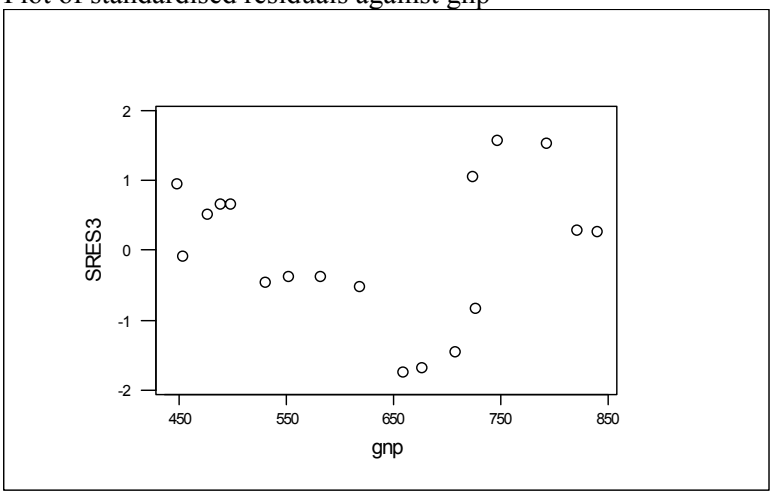
Case study 3

[Reminder. The response variable is consumer expenditure and the predictor variable is gross national product (GNP), for 18 years in the USA. The data are in $ billion, all at 1958 prices.]

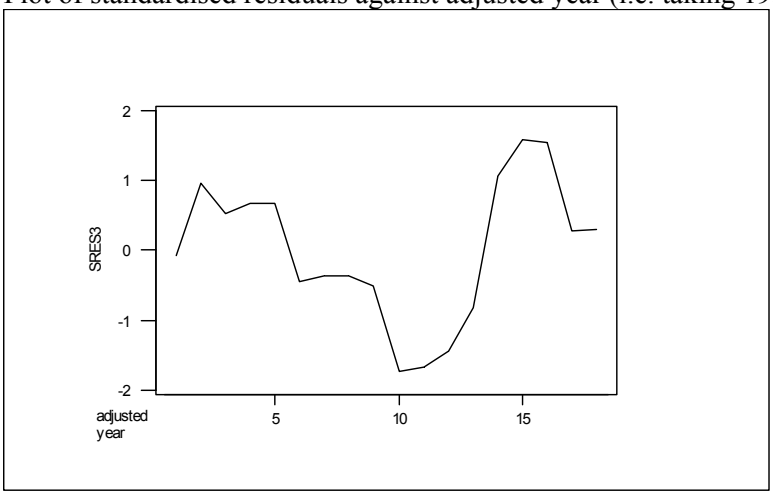
Plot of standardised residuals against fitted values



Plot of standardised residuals against gnp



Plot of standardised residuals against adjusted year (i.e. taking 1958 as 0)



Turn over

4. (i) (a) Define the term *odds ratio*.

(b) Suppose you have fitted a binomial response to some data using a generalised linear modelling package, and that the model includes a particular factor with two levels but no interactions involving that factor. Explain how you would use the fitted model and its associated statistics to calculate an approximate 95% confidence interval for the odds ratio for level 2 of that factor relative to level 1 of the factor.

(5)

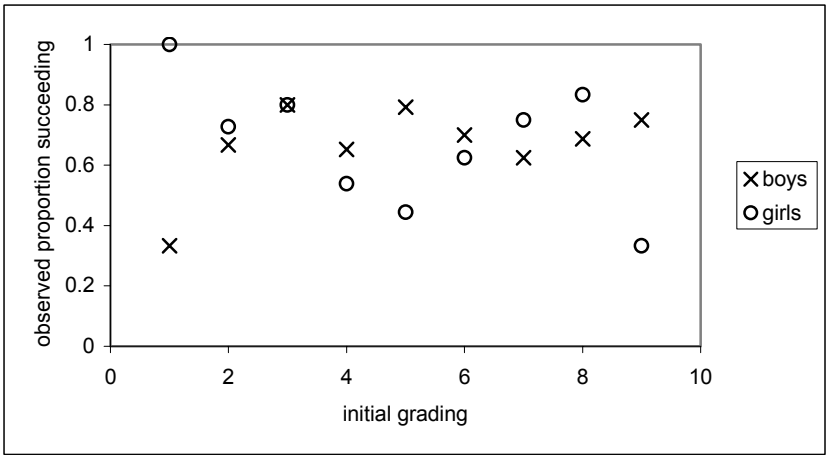
(ii) The manager of a fitness club is investigating the relationship between success rates of boys and girls in a test of fitness and their initial fitness grading by their school teacher before joining the club. The table below gives the number of children with each initial grading, sub-divided by sex, and their success rates.

	Boys		**Girls**	
Initial grading	*Total number*	*Number who succeed*	*Total number*	*Number who succeed*
1 (lowest)	3	1	1	1
2	9	6	11	8
3	10	8	15	12
4	23	15	13	7
5	24	19	9	4
6	20	14	16	10
7	16	10	12	9
8	16	11	6	5
9 (highest)	16	12	3	1

(a) The following graph shows the observed proportions succeeding plotted against the initial grading. Interpret the graph.

(2)

Plot of observed proportions succeeding against initial grading



Question 4 is continued on the next page

(b) Copy the table below into your answer book, and complete it by including a column showing the numbers of degrees of freedom. Use forward selection to choose a parsimonious well-fitting model for the success rate. You should note that in this modelling "Initial grading" has been coded as a covariate, not a factor. Explain your model selection process and justify your final model being "well-fitting".

(7)

Table summarising generalised linear models fitted to the data

Predictor Variables in Model	*Scaled Deviance*
–	12.348
Sex	11.997
Initial grading	12.346
Initial grading + Sex	11.993

(c) As a statistician reporting to the manager, write a short statement on your findings. You should view the manager as having only a basic understanding of statistics.

(3)

(d) When one of your colleagues sees the analysis, he tells you that the formal analysis was unnecessary because the conclusion is obvious from the graph. Discuss the validity of this criticism.

(3)

5. A set of data has been collected in which y is the response variable and $x1$, $x2$, …, $x16$ are possible predictor variables. The intention is to produce a multiple regression model, and there are 12 observations.

(i) Briefly explain the major difficulties with the problem, as presented.

(2)

(ii) The table below contains details of the correlations between the variables. Describe the patterns of correlations.

(4)

Correlations (Pearson) of all variables

	x1	x2	x3	x4	x5	x6	x7	x8
x2	0.986							
x3	0.995	0.987						
x4	0.504	0.504	0.538					
x5	0.203	0.160	0.175	0.212				
x6	0.062	-0.001	0.027	0.180	0.898			
x7	0.559	0.555	0.591	0.996	0.229	0.173		
x8	0.107	0.184	0.116	-0.115	-0.595	-0.649	-0.112	
x9	0.086	0.157	0.090	-0.182	-0.547	-0.607	-0.176	0.989
x10	0.133	0.209	0.143	-0.091	-0.583	-0.647	-0.087	0.999
x11	0.391	0.454	0.406	0.165	-0.399	-0.565	0.185	0.889
x12	0.597	0.630	0.614	0.421	-0.097	-0.342	0.454	0.561
x13	0.654	0.662	0.671	0.540	0.131	-0.134	0.577	0.246
x14	0.610	0.593	0.603	0.346	0.805	0.553	0.391	-0.389
x15	0.508	0.469	0.487	0.437	0.743	0.576	0.461	-0.392
x16	0.230	0.182	0.202	0.381	0.425	0.403	0.374	-0.261
y	0.293	0.342	0.311	0.058	0.143	-0.306	0.110	0.173

	x9	x10	x11	x12	x13	x14	x15	x16
x10	0.988							
x11	0.875	0.907						
x12	0.546	0.594	0.877					
x13	0.233	0.285	0.662	0.941				
x14	-0.385	-0.364	-0.075	0.273	0.479			
x15	-0.391	-0.360	0.005	0.420	0.652	0.834		
x16	-0.264	-0.233	0.084	0.426	0.605	0.380	0.827	
y	0.182	0.196	0.416	0.572	0.599	0.494	0.324	0.040

(iii) Explain why a subset of principal components of the predictor variables might be suggested as a way of helping in such a situation.

(2)

(iv) The table following (at the top of the next page) contains details of the first six principal components (based on the correlation matrix) of the predictor variables. Give a brief interpretation of the first four principal components, describing any other information you might find useful and why.

(4)

Question 5 is continued on the next page

Principal Component Analysis of 16 predictor variables

Eigenanalysis of the Correlation Matrix

Eigenvalue	6.4774	5.7594	1.4154	1.2606	0.6665	0.3288
Proportion	0.405	0.360	0.088	0.079	0.042	0.021
Cumulative	0.405	0.765	0.853	0.932	0.974	0.994
Eigenvalue	0.0689	0.0140	0.0073	0.0010	0.0008	0.0000
Proportion	0.004	0.001	0.000	0.000	0.000	0.000
Cumulative	0.999	0.999	1.000	1.000	1.000	1.000
Eigenvalue	0.0000	-0.0000	-0.0000	-0.0000		
Proportion	0.000	-0.000	-0.000	-0.000		
Cumulative	1.000	1.000	1.000	1.000		

Variable	PC1	PC2	PC3	PC4	PC5	PC6
x1	-0.342	-0.040	0.301	-0.227	-0.129	-0.286
x2	-0.340	-0.070	0.305	-0.230	-0.090	-0.195
x3	-0.343	-0.048	0.324	-0.192	-0.130	-0.194
x4	-0.276	0.053	0.216	0.534	0.307	0.053
x5	-0.150	0.318	-0.188	-0.265	0.405	0.205
x6	-0.071	0.344	-0.128	-0.159	0.510	-0.398
x7	-0.291	0.050	0.231	0.492	0.275	0.093
x8	-0.021	-0.399	-0.091	-0.088	0.273	-0.074
x9	-0.011	-0.392	-0.131	-0.148	0.291	-0.084
x10	-0.035	-0.399	-0.100	-0.083	0.260	-0.062
x11	-0.179	-0.358	-0.181	-0.016	0.099	0.065
x12	-0.302	-0.229	-0.232	0.064	-0.108	0.195
x13	-0.345	-0.104	-0.234	0.111	-0.238	0.258
x14	-0.279	0.205	-0.015	-0.348	0.052	0.503
x15	-0.301	0.205	-0.323	-0.071	-0.099	0.006
x16	-0.220	0.134	-0.527	0.236	-0.218	-0.503

(v) The table below contains information about two regression analyses of the data. In each case comment critically on the choice of predictor variables and on the fit of the model.

(4)

Model selection	*Variables in model*	*R-squared*	*Adjusted R-squared*
Forward selection	*x*13, *x*16	52.3%	41.7%
Direct entry	first 5 principal components	35.4%	*

* The standard formula for Adjusted *R*-squared gives a value of –18.4%.

(vi) A summary of a regression analysis using just five predictor variables x_4, x_5, x_6, x_7 and x_{16} is as follows.

R-squared	*Adjusted R-squared*
100%	100%

(a) Describe the apparent anomaly between this analysis and those summarised in (v).

(b) Explain why the methods used in (v) did not produce this solution to the problem.

(c) Critically discuss any methods you might use to find the "best-fitting" regression model for these data, relating your answer to the problems described in (i).

(4)

6. A company sells products in a large number of sales regions, each of which is assigned to a single sales representative. Data have been collected, comprising the following variables:

SALES	total sales in units credited to the sales representative
TIME	length of time employed by the company (in months)
POTEN	market potential: total sales for industry in units for region
ADV	advertising expenditure in region
SHARE	market share (weighted average for the last 4 years)
CHANGE	change in market share over the last 4 years
ACCTS	number of accounts assigned to sales representative
WORK	work load: a weighted index based on annual purchases etc
RATING	sales representative overall rating on 8 performance indicators
AREA	social/industrial coding for region (1 = poor, …, 4 = rich)

The (Pearson) correlations for the first 9 of these, omitting AREA, are:

	SALES	TIME	POTEN	ADV	SHARE	CHANGE	ACCTS	WORK
TIME	0.623							
POTEN	0.598	0.454						
ADV	0.596	0.249	0.174					
SHARE	0.484	0.106	–0.211	0.264				
CHANGE	0.489	0.251	0.268	0.377	0.085			
ACCTS	0.754	0.758	0.479	0.200	0.403	0.327		
WORK	–0.117	–0.179	–0.259	–0.272	0.349	–0.288	–0.199	
RATING	0.402	0.101	0.359	0.411	–0.024	0.549	0.229	–0.277

The objective is to produce a linear model to predict SALES from suitable predictor variables.

(i) Comment on the nature and quality of the variables, any apparent relationships between them, and the implications for the modelling process. Why do you think AREA has been omitted from the analysis?

(5)

(ii) Explain carefully how you would go about selecting a suitable set of predictor variables.

(5)

Question 6 is continued on the next page

(iii) Statistical packages contain various influence diagnostics.

(a) Explain what the *hat matrix* **H** is in the analysis of the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \left(\mathrm{Var}\left(\boldsymbol{\varepsilon}\right) = \sigma^2\mathbf{I}\right).$$

Obtain an expression for $\hat{\boldsymbol{\varepsilon}}$ in terms of **H**.

[You may assume that the OLS estimator of $\boldsymbol{\beta}$ is

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

State the properties of the residuals, and explain how the diagonal elements of the hat matrix are used to decide whether or not a data point ($\mathbf{x}_i$, y_i) is *influential*. Explain how this helps in constructing a model.

(6)

(b) Explain in detail how <u>any **two**</u> of the following can be useful in constructing a model:

(1) Studentised residuals and deleted residuals

(2) Cook's distance

(3) DFFITS

(4) DFBETAS

(5) COVRATIO.

(4)

7. A group of archaeologists are studying remains found at a number of sites. They have measurements on a set of 32 skulls, 17 of which were found at one archaeological site and the other 15 at another site. They believe that each of these two sites was inhabited by a different tribe of people. They are now working on other sites in the same region and wish to decide which tribe the skulls they are finding belong to.

The measurements comprised 5 different dimensions of the skulls, all in mm.

(i) Explain how linear *discriminant analysis* might be useful in studying these data, and describe the way in which the results may be applied to data from the new sites.

(3)

(ii) State the assumptions that would need to be made about the data in order for linear discriminant analysis to be valid. Describe the checks that you could do to investigate these assumptions, stating any limitations on the methods.

(4)

(iii) Summary statistics for each of the two groups of skulls are shown below. Describe the main features in relation to your answers to (i) and (ii).

(4)

Means for the 5 variables

Variable	*Site 1 (n = 17)*	*Site 2 (n = 15)*
x_1	174.82	185.73
x_2	139.35	138.73
x_3	132.00	134.77
x_4	69.82	76.47
x_5	130.55	137.50

Variance-covariance matrices:

Site 1

$$\begin{pmatrix} 45.53 & 25.22 & 12.39 & 22.15 & 27.97 \\ 25.22 & 57.81 & 11.88 & 7.52 & 48.06 \\ 12.39 & 11.88 & 36.09 & -0.31 & 1.41 \\ 22.15 & 7.52 & -0.31 & 20.94 & 16.77 \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{pmatrix}$$

Question 7 is continued on the next page

Site 2

$$\begin{pmatrix} 74.42 & -9.52 & 22.74 & 17.79 & 11.13 \\ -9.52 & 37.75 & -11.26 & 0.70 & 9.46 \\ 22.74 & -11.26 & 36.32 & 10.72 & 7.20 \\ 17.79 & 0.70 & 10.72 & 15.30 & 8.66 \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{pmatrix}$$

(iv) Below is shown a summary of two discriminant analyses on these data. Describe the differences between the two sets of results. Which model do you prefer and why?

(9)

OUTPUT FROM LINEAR DISCRIMINANT ANALYSIS

METHOD 1 – including all 5 variables

Linear discriminant function is: $-0.09x_1 + 0.16x_2 + 0.01x_3 - 0.18x_4 - 0.18x_5$

Classification Results

			Predicted group membership		Total
		SITE	1	2	
Original	Count	1	14	3	17
		2	3	12	15
	%	1	82.4	17.6	100.0
		2	20.0	80.0	100.0
Cross-validated	Count	1	12	5	17
		2	6	9	15
	%	1	70.6	29.4	100.0
		2	40.0	60.0	100.0

METHOD 2 – stepwise, using forward selection

Discriminant function: $-0.37x_4$

Classification Results

			Predicted group membership		Total
		SITE	1	2	
Original	Count	1	12	5	17
		2	3	12	15
	%	1	70.6	29.4	100.0
		2	20.0	80.0	100.0
Cross-validated	Count	1	12	5	17
		2	3	12	15
	%	1	70.6	29.4	100.0
		2	20.0	80.0	100.0

8. The table below summarises an analysis of data from a balanced three-factor experiment with replication.

Source	*SS*	*df*
A	60.75	1
B	6.00	2
C	18.75	1
A × B	0.00	2
A × C	0.75	1
B × C	24.00	2
A × B × C	6.00	2
Within	171.00	36
Total	287.25	47

(i) For each of the two cases below, complete the ANOVA table, showing the expected values of the mean squares, based on a suitable linear model which you should specify fully. State your conclusions regarding statistical significance and the practical conclusions that can be made.

(a) A, B, C are all fixed factors.

(7)

(b) A and B are fixed factors; C is a random factor.

(8)

(ii) In each case state what further analysis you would do, if any, and why.

(2)

(iii) How and when would you decide which of the two analyses was the more appropriate?

(3)