EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2005

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

(2)

(ii) Define an MA(1) process, and derive the mean, variance and autocorrelation function (ACF) of an MA(1) process.

(7)

(iii) Explain what a partial autocorrelation coefficient is. State, without giving detailed working, the form of the partial autocorrelation function (PACF) for an MA(1) process.

(3)

(iv) Describe how you would use the sample ACF and PACF to recognise an MA(1) process.

(2)

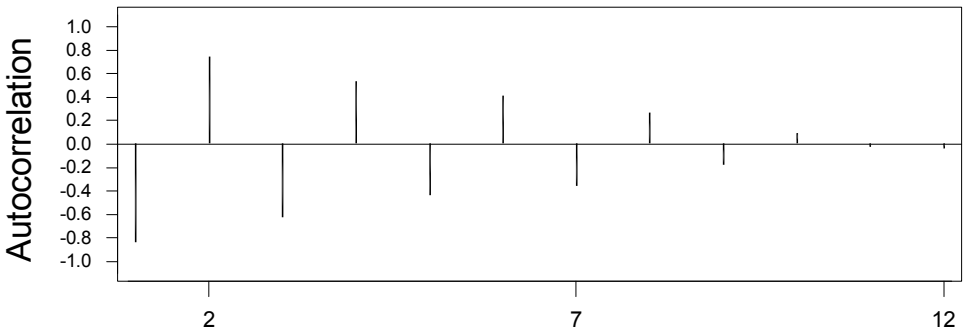
(v) The charts **shown on the next three pages** are derived from 50 observations from each of three time series. Which of the three do you think is an MA(1) process? Justify your answer.

Suggest models for the other two time series, giving reasons for your answers.

(6)

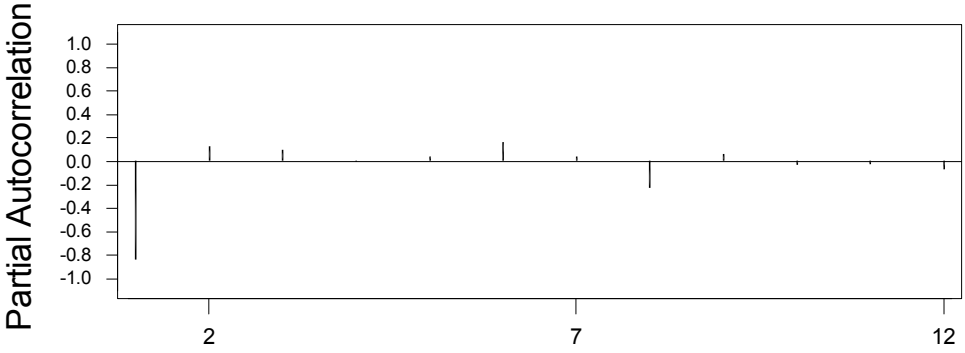
Charts for question 1 follow on the next three pages

Autocorrelation Function for seriesA



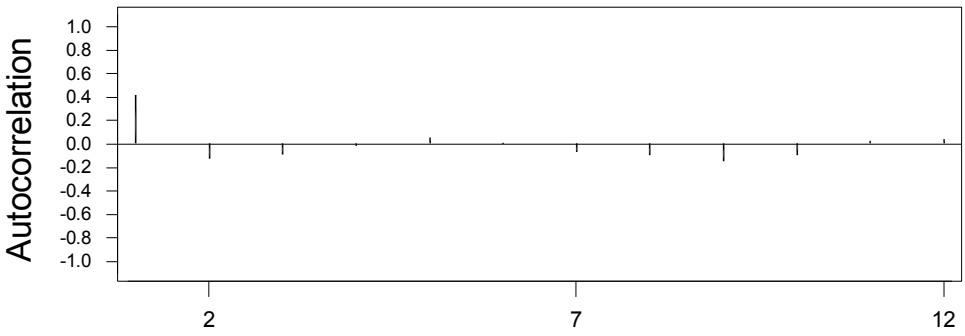
Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	-0.84	-5.95	37.54	8	0.26	0.77	138.56
2	0.74	3.39	67.60	9	-0.18	-0.51	140.56
3	-0.63	-2.36	89.31	10	0.09	0.27	141.13
4	0.53	1.82	105.42	11	-0.03	-0.08	141.18
5	-0.44	-1.41	116.58	12	-0.04	-0.11	141.27
6	0.41	1.27	126.58				
7	-0.36	-1.07	134.30				

Partial Autocorrelation Function for seriesA



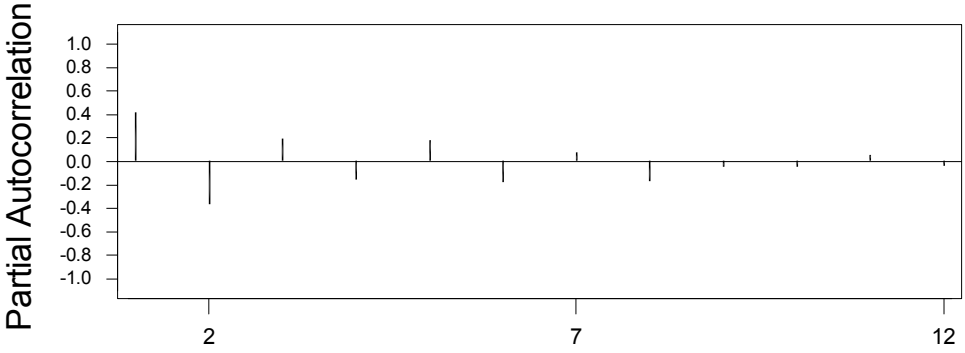
Lag	PAC	T	Lag	PAC	T
1	-0.84	-5.95	8	-0.23	-1.60
2	0.13	0.91	9	0.06	0.45
3	0.10	0.68	10	-0.03	-0.25
4	-0.00	-0.01	11	-0.02	-0.17
5	0.04	0.28	12	-0.07	-0.47
6	0.16	1.15			
7	0.04	0.29			

Autocorrelation Function for seriesB



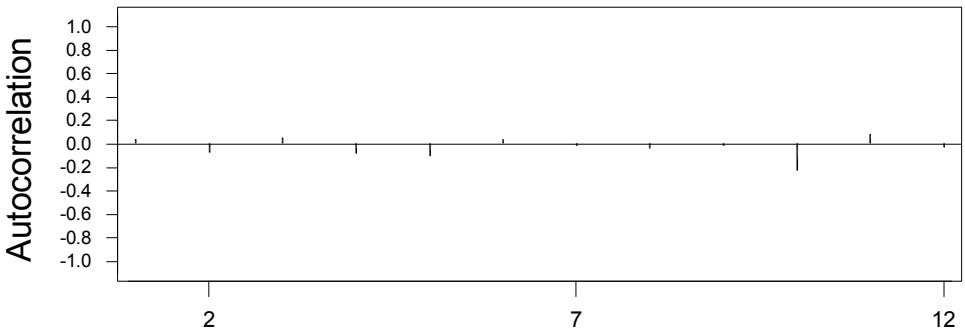
Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.42	2.98	9.39	8	-0.09	-0.57	11.84
2	-0.13	-0.80	10.32	9	-0.15	-0.87	13.21
3	-0.09	-0.54	10.77	10	-0.10	-0.58	13.84
4	-0.02	-0.12	10.79	11	0.02	0.14	13.88
5	0.05	0.32	10.96	12	0.04	0.25	14.01
6	0.00	0.03	10.96				
7	-0.07	-0.43	11.27				

Partial Autocorrelation Function for seriesB



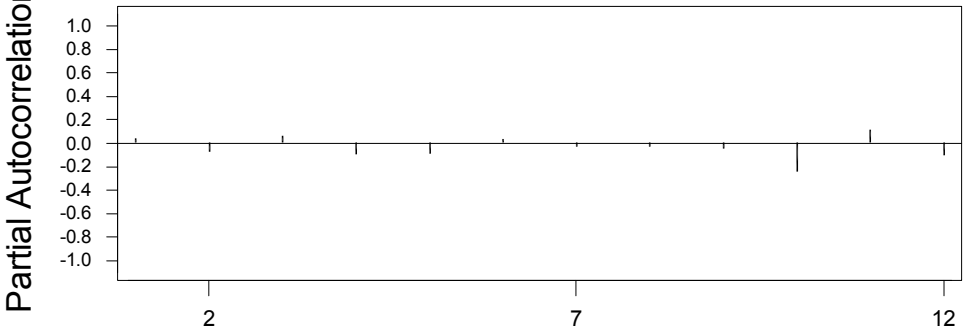
Lag	PAC	T	Lag	PAC	T
1	0.42	2.98	8	-0.17	-1.21
2	-0.37	-2.62	9	-0.05	-0.33
3	0.19	1.38	10	-0.04	-0.32
4	-0.15	-1.09	11	0.05	0.38
5	0.18	1.26	12	-0.04	-0.27
6	-0.18	-1.30			
7	0.08	0.57			

Autocorrelation Function for seriesC



Lag	Corr	T	LBQ	Lag	Corr	T	LBQ
1	0.04	0.27	0.08	8	-0.04	-0.29	1.83
2	-0.08	-0.55	0.40	9	-0.02	-0.15	1.86
3	0.05	0.37	0.55	10	-0.23	-1.58	5.34
4	-0.08	-0.59	0.95	11	0.08	0.53	5.77
5	-0.10	-0.73	1.59	12	-0.03	-0.23	5.86
6	0.04	0.29	1.70				
7	-0.02	-0.14	1.72				

Partial Autocorrelation Function for seriesC



Lag	PAC	T	Lag	PAC	T
1	0.04	0.27	8	-0.03	-0.23
2	-0.08	-0.56	9	-0.04	-0.31
3	0.06	0.42	10	-0.24	-1.72
4	-0.10	-0.68	11	0.11	0.79
5	-0.09	-0.63	12	-0.10	-0.74
6	0.03	0.25			
7	-0.03	-0.22			

2. (a) A property investor is trying to model current property prices in terms of the age and type of property. He has data on 55 properties for sale in his area. These data consist of price in thousands of UK pounds, age of property in years, and type of property. The type is represented by the following coding:

detached house = 1
semi-detached house = 2
terraced house = 3.

(i) Contrast the terms *factor* and *continuous variable* as used in linear modelling.

(2)

(ii) The property investor has done some linear modelling, using multiple regression, with both age and type as predictor variables. Part of the output is shown below. Say, with reasons, whether "type" has been coded as a factor or a continuous variable. Comment on whether you think this is a sensible choice.

(4)

```
Analysis of Variance

SOURCE       DF          SS          MS          F         p
Regression    2      5717.5      2858.7       6.56     0.003
Error        52     22662.3       435.8
Total        54     28379.8
```

(iii) After receiving statistical advice, he runs a further multiple regression analysis using two different packages, and appears to get different output.

Data for the first three observations are **shown on the next page**, together with the coefficients and the first three rows of the design matrix for package A. Part of the output from package B is also shown.

Write down the first three rows of the design matrix used in package B. Show that the parameter estimates obtained by the two packages are consistent with each other. State any other ways in which the two sets of output would differ.

(9)

Question 2 is continued on the next page

Sample Data (first three observations)

Observation	Price	Age	Type
1	54.0	58	3
2	54.3	19	2
3	55.2	10	1

Output from package A

```
Parameter estimates:
78.8603   11.2249   -0.5764   -0.4180

First 3 rows of Design Matrix
   1   -1   -1   58
   1    0    1   19
   1    1    0   10
```

Output from package B

```
Intercept     68.212
Age           -0.418
[type=1]      21.873
[type=2]      10.072
[type=3]       0
```

(b) A health psychologist is studying fear of falling in old people. He is examining the relationships between three psychometric scales using a random sample of 64 people. Scale A measures fear of falling, where a high score corresponds to high fear. Scale B measures confidence doing tasks where there is a risk of falling; here a high score corresponds to high confidence. Scale C is a measure of general anxiety, where a high score corresponds to high anxiety.

Pearson correlations from his sample are as follows.

	Scale A	Scale B
Scale B	-0.771	
Scale C	0.637	-0.328

Coefficients from multiple regression, modelling Scale B as the dependent variable, are as follows.

Dependent variable: Scale B

	Coefficient	Standard error
Constant:	126.28	2.976
Scale A:	-1.486	0.165
Scale C:	1.439	0.566

Given the negative correlation between Scales B and C, the psychologist expected the coefficient of Scale C to be negative. Provide a possible explanation for the positive coefficient.

(5)

3. (i) A multiple linear regression model can be written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the error term. The parameter $\boldsymbol{\beta}$ is commonly estimated by $\hat{\boldsymbol{\beta}}$, given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

(a) Show that, subject to assumptions which you should state clearly, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Show also that the variance of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, where σ^2 should be defined. (7)

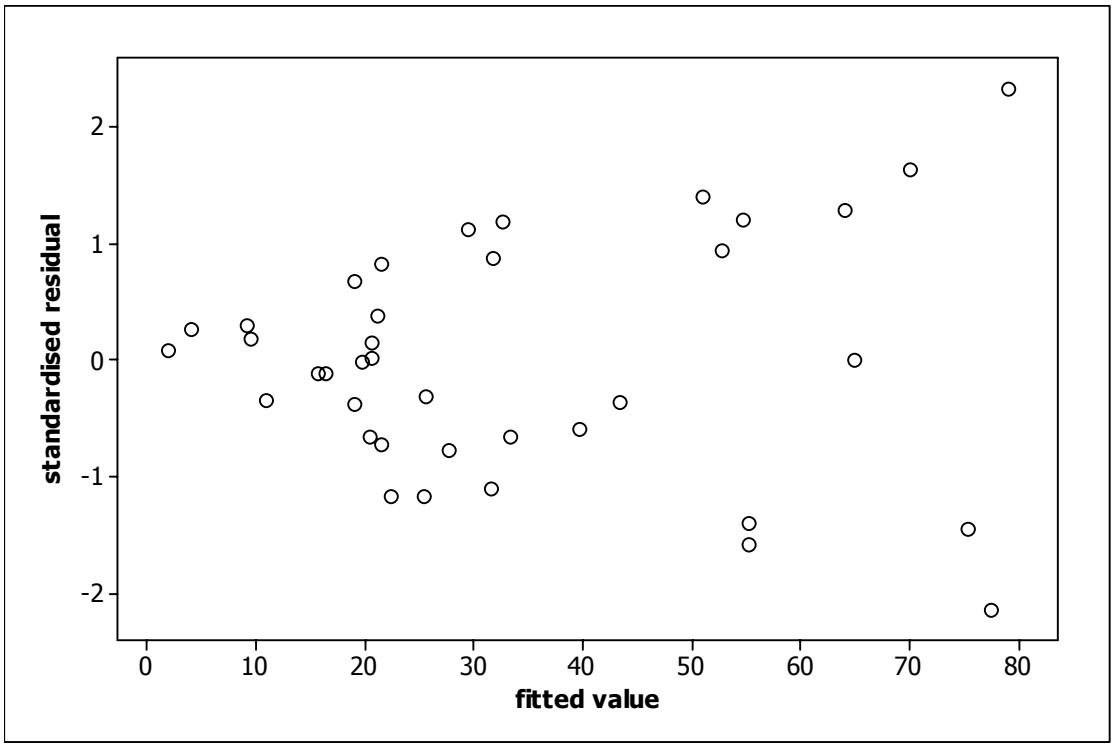
(b) State the Gauss-Markov theorem. (2)

(c) Suppose now that the error term $\boldsymbol{\varepsilon}$ is Normally distributed. Comment briefly on the implications of this for inference about the parameter $\boldsymbol{\beta}$. (2)

(ii) (a) State two main purposes of transforming the dependent variable in linear regression. (2)

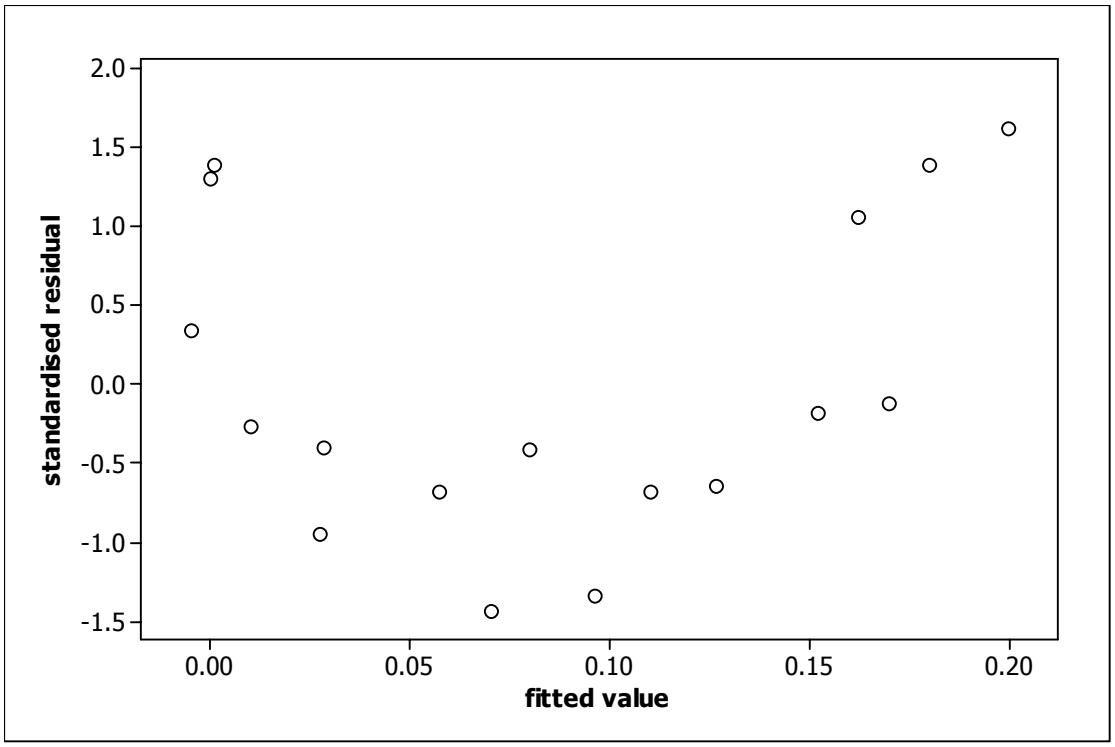
(b) The following graphs (**on this page and the next**) show plots of standardised residuals against fitted values, for three regression analyses. In each case, say whether you consider that any of the assumptions stated in part (i) may not be valid. Also suggest any transformation of the dependent variable, or other appropriate action, that might improve the validity of an analysis. Justify your answers. (7)

Graph A

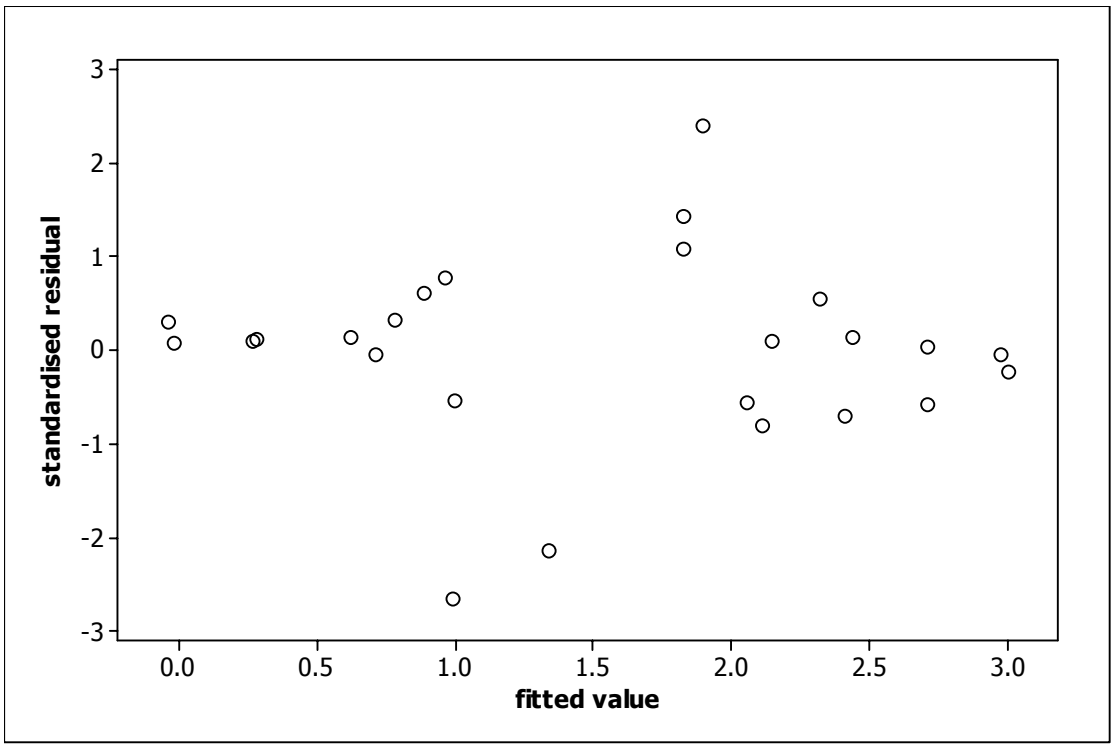


The remaining graphs for question 3 are on the next page

Graph B



Graph C



4. When petrol is pumped into the tank of a car, hydrocarbon vapours are forced from the tank into the atmosphere, causing pollution. An experiment was carried out to determine the relationship between the amount of hydrocarbon vapour emitted from a tank and other variables that can be measured more easily.

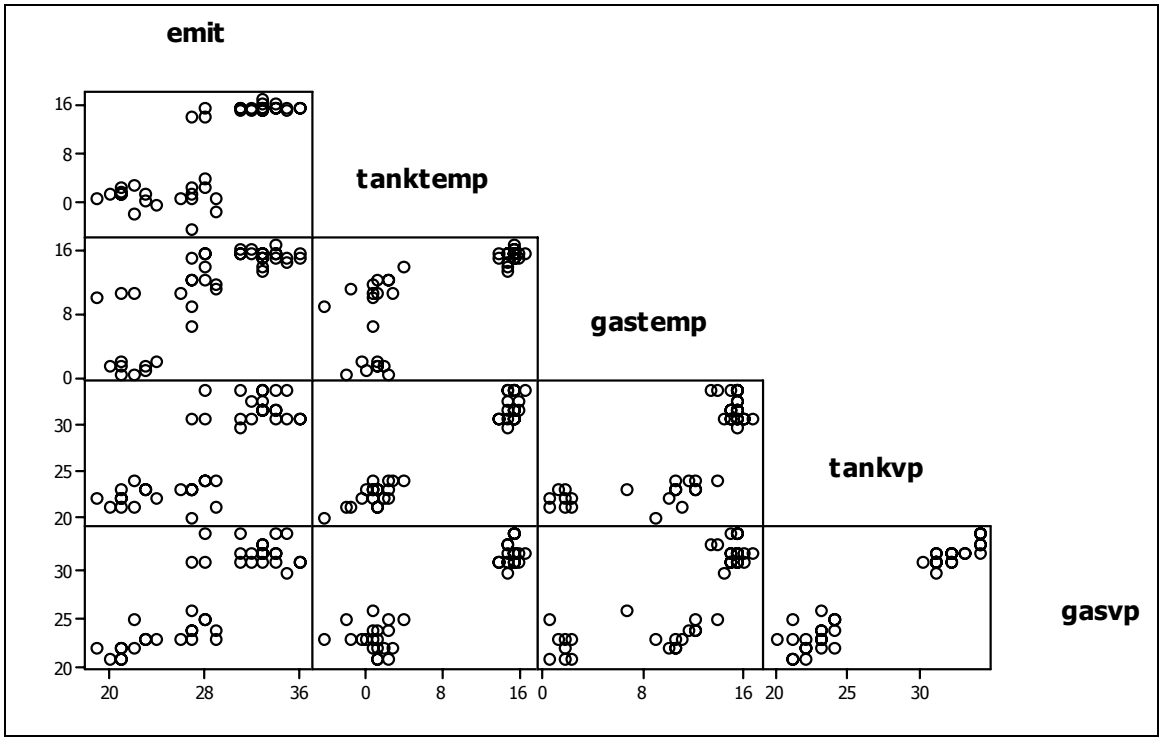
The measured variables were:

tanktemp:	the initial tank temperature (degrees Celsius)
gastemp:	the temperature of the dispensed petrol (degrees Celsius)
tankvp:	the initial vapour pressure in the tank (kPa)
gasvp:	vapour pressure of the dispensed petrol (kPa)
emit:	emitted hydrocarbons (gm).

There were 44 observations.

(i) The output below shows a matrix plot of these 5 variables. What do these diagrams tell you about the use of these variables in a multiple regression model?

(4)



(ii) The table **shown on the next page** summarises a number of multiple regression models, from a "best subsets" analysis. On the basis of this information, which model would you choose? Justify your answer.

(4)

Question 4 is continued on the next page

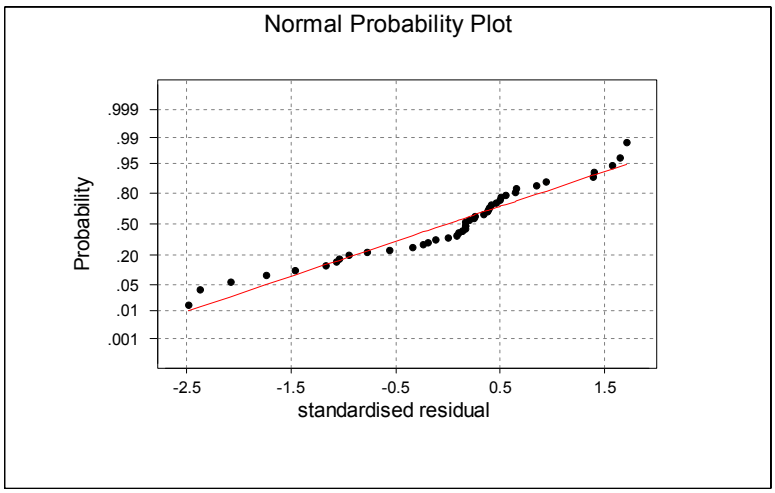
Variables in model (in addition to constant)	*R-squared*	*Adjusted R-squared*	*Mallows' C_p*	*s*
gasvp	74.0	73.4	8.5	2.5972
tankvp	69.1	68.4	17.7	2.8320
gastemp, gasvp	78.6	77.6	2.0	2.3854
gastemp, tankvp, gasvp	78.9	77.4	3.4	2.3968
tanktemp, gastemp, gasvp	78.6	77.0	3.9	2.4143
tanktemp, gastemp, tankvp, gasvp	79.1	77.0	5.0	2.4165

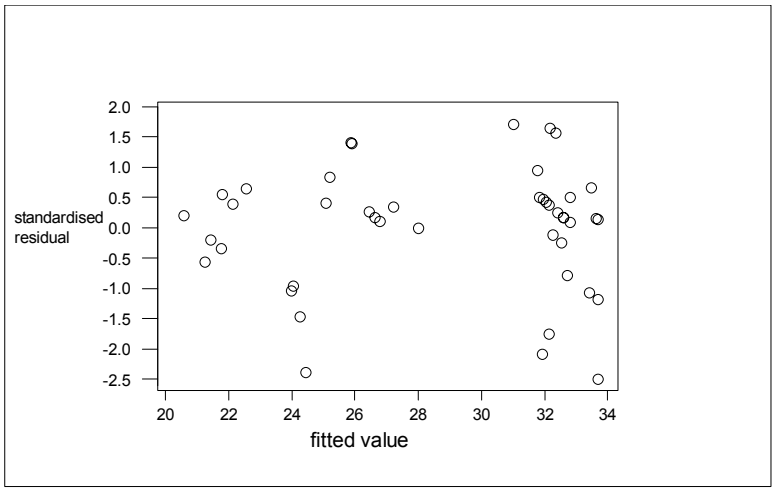
(iii) Someone has suggested using a stepwise procedure to select a suitable model. Explain what is meant by *stepwise* and state, with reasons, whether you think this is a good approach for these data.

(6)

(iv) The following plots were obtained from the model that includes all four predictor variables. Describe the main features of these plots and the implications for the model.

What steps would you take next so as to model these data satisfactorily?





(6)

5. The random variables Y_i, $i = 1, 2, \ldots, k$, have independent binomial (n_i, π_i) distributions, where each n_i is known and the π_i are unknown.

(i) Define the term *odds* in the context of this statistical model. Explain how the odds may be estimated following the fitting of a logistic model. State the form of the model fitted in terms of the odds. (4)

(ii) Eight groups of insects were exposed to various doses of carbon disulphide and the numbers of insects R_i out of N_i, $i = 1, 2, \ldots, 8$, that were dead after 5 hours were recorded.

The data are shown below, together with a summary of output from fitting generalised linear models with logit link.

Log_{10} of dose (x_i)	1.6905	1.7245	1.7558	1.7843	1.8116	1.8364	1.8500	1.8995
N_i	58	21	62	57	64	57	60	58
R_i	5	12	19	26	50	51	59	58

Terms in model	*Parameter estimate*	*Scaled deviance*
Constant only	0.5785	237.20
Constant x	–55.0777 31.1093	28.32
Constant x x^2	444.1290 –532.4059 158.9245	19.32

			Model			
			Constant, x		Constant, x, x^2	
x_i	N_i	R_i	*Linear predictor*	*Standard error of linear predictor*	*Linear predictor*	*Standard error of linear predictor*
1.6905	58	5	–2.4874	0.3056	–1.7303	0.3452
1.7245	21	12	–1.4297	0.2196	–1.3793	0.1987
1.7558	62	19	–0.4559	0.1560	–0.7314	0.1814
1.7843	57	26	0.4307	0.1317	0.1293	0.1676
1.8116	64	50	1.2800	0.1544	1.1960	0.1718
1.8364	57	51	2.0515	0.2023	2.3702	0.2761
1.8500	60	59	2.4746	0.2341	3.0972	0.3789
1.8995	58	58	4.0145	0.3639	6.2396	0.9614

(a) Choose the best model of those presented, justifying your choice. Comment on the goodness of fit of this model. (4)

(b) Illustrating for the lowest dose, where $x_i = 1.6905$, show how the linear predictor is generated by your chosen model, and how it is related to the odds.

Calculate also an approximate 95% confidence interval for the odds of death at this lowest dose. (6)

(c) For your chosen model, estimate the dose that kills half the insects. (5)

(d) Using the first table above, draw attention to any characteristic of the raw data which may reduce the precision of any fitted model. (1)

6. (i) Explain the purpose of cluster analysis.

(2)

(ii) A nurse is trying to do some cluster analysis using a statistical package. The package asks her to choose from options for a *distance measure* and a *linkage method*. It also gives the user the option of *standardising* the variables.

(a) Explain what each of these options means and why they are important.

(5)

(b) The nurse has analysed trainee nurses' written answers to an essay question asking them to discuss how they would deal with a particular clinical problem. She has identified several ideas and coded them as "present" or "absent" in each trainee's answer.

She wishes to classify the trainees into clusters on the basis of her analysis of their essays.

She suggests the following distance measure:

> The distance between answers A and B is a count of the number of ideas that are present in A but not in B plus those that are present in B but not in A.

Is this a valid distance measure on theoretical grounds? Justify your answer.

(5)

(c) A colleague of the nurse has looked at her data and commented that she believes there to be some overlap between the ideas that have been identified. For example, she suggests that the two ideas "reassure family" and "encourage the relatives" might really be the same basic idea. If this is the case, what effect will this have on the distance matrix used in the cluster analysis? What advice would you give the nurse to overcome this problem?

(4)

(d) The nurse asks you how to identify the number of clusters in the data. What advice would you give?

(4)

7. A scientist has carried out a randomised block experiment using three treatments in three blocks. He has measured three response variables, X1, X2 and X3, on each experimental unit, and intends to analyse the data using multivariate analysis of variance (MANOVA).

(i) Assuming that MANOVA is appropriate, write down a suitable model for the data, stating the distributional assumptions.

(4)

(ii) Describe how you would attempt to check these assumptions, and any limitations of such checks.

(4)

(iii) Describe the advantages of using MANOVA rather than three univariate analyses on the responses.

(2)

(iv) The table below shows part of the output from a MANOVA.

(a) Explain how the entries in the SSCP matrix for treatment are obtained. (You may illustrate this by showing how the entries 2.149 and 1.582 in the matrix would have been found.)

(4)

(b) Explain why there are three test statistics to test for a treatment effect, and why the corresponding *p*-values are different.

(4)

MANOVA for treatment s = 2 m = 0.0 n = 0.0

CRITERION	TEST STATISTIC	F	DF	P
Wilks'	0.00645	7.633	(6, 4)	0.035
Lawley-Hotelling	119.57301	19.929	(6, 2)	0.049
Pillai's	1.21558	1.550	(6, 6)	0.304

SSCP Matrix for treatment

	x1	x2	x3
x1	2.149	1.582	-1.443
x2	1.582	1.182	-1.127
x3	-1.443	-1.127	1.207

(v) Briefly describe any further analyses you might carry out on these data to investigate the nature of the treatment effects.

(2)

8. A manager in a factory knows little about statistics. On the basis of statistical advice, he carried out an experiment to decide which of three machines was best to buy. Six employees were chosen at random. Each employee operated each machine twice. Each piece of work was rated on its quality, where a high score is better than a low score. The manager wanted to choose a machine on which employees would produce high quality work.

The 36 runs in the experiment were carried out in random order.

The data are given below, together with some related statistics.

Machine	*Employee*	*Rating*	*Machine*	*Employee*	*Rating*	*Machine*	*Employee*	*Rating*
1	1	52.4	2	1	64.5	3	1	67.5
1	1	50.4	2	1	63.8	3	1	67.2
1	2	51.9	2	2	59.4	3	2	61.4
1	2	52.6	2	2	59.2	3	2	61.9
1	3	60.2	2	3	55.0	3	3	70.5
1	3	61.5	2	3	65.7	3	3	70.6
1	4	51.3	2	4	62.4	3	4	64.9
1	4	52.4	2	4	62.3	3	4	66.5
1	5	64.5	2	5	64.9	3	5	72.4
1	5	63.5	2	5	65.1	3	5	71.3
1	6	46.6	2	6	43.8	3	6	62.5
1	6	49.5	2	6	44.6	3	6	60.4

MEANS

Machine	N	Rating
1	12	54.733
2	12	59.225
3	12	66.425

Employee	N	Rating
1	6	60.967
2	6	57.733
3	6	63.917
4	6	59.567
5	6	66.950
6	6	51.233

TOTALS

	Employee						
Machine	1	2	3	4	5	6	All
1	102.8	104.5	121.7	103.7	128.0	96.1	656.8
2	128.3	118.6	120.7	124.7	130.0	88.4	710.7
3	134.7	123.3	141.1	131.4	143.7	122.9	797.1
All	365.8	346.4	383.5	359.8	401.7	307.4	2164.6

(Note: $102.8^2 + 104.5^2 + \ldots + 122.9^2 = 264308.12$.
You may also use the fact that the corrected total sum of squares for all 36 observations is 2071.99.)

(i) Write down an appropriate model for this design, explaining all terms and stating the necessary assumptions. Indicate which (if any) of the terms are "fixed effects" and which (if any) are "random effects". (6)

(ii) State the expected mean squares of the effects in your model. (3)

(iii) Complete an analysis of variance for these data, and write a report on the results. (You may use the tables of means and totals above as required.) (11)