EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2003

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.*

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 12 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) The table below gives details of the autocorrelation function (ACF) and partial autocorrelation function (PACF) for three time series each of length 50. Suggest, giving reasons, suitable forms for each of the time series.

	Series 1		**Series 2**		**Series 3**	
lag	*ACF*	*PACF*	*ACF*	*PACF*	*ACF*	*PACF*
1	–0.03	–0.03	0.63	0.63	0.88	0.88
2	–0.08	–0.08	0.50	0.18	0.83	0.21
3	0.17	0.17	0.31	–0.09	0.82	0.25
4	0.03	0.04	0.31	0.15	0.75	–0.15
5	–0.05	–0.02	0.12	–0.21	0.68	–0.10
6	–0.02	–0.05	0.08	0.02	0.62	–0.12
7	0.02	0.00	–0.01	–0.04	0.57	0.06
8	–0.23	–0.23	–0.10	–0.19	0.49	–0.18
9	0.03	0.03	–0.20	–0.05	0.44	0.09
10	0.20	0.18	–0.30	–0.19	0.39	–0.06
11	–0.26	–0.19	–0.37	–0.13	0.31	–0.10
12	–0.09	–0.09	–0.37	0.02	0.26	0.02
13	–0.02	–0.13	–0.39	–0.13	0.22	0.00
14	–0.01	0.02	–0.42	–0.13	0.18	0.08
15	0.00	0.06	–0.40	–0.03	0.14	–0.01

(12)

(ii) Consider the second-order moving average process $\{X_t\}$ where

$$X_t = Z_t + 0.8Z_{t-1} - 0.4Z_{t-2}$$

and $\{Z_t\}$ is a white noise process with $E(Z_t) = 0$ and $\text{Var}(Z_t) = \sigma_z^2$.

Obtain the mean, variance and autocorrelation function of $\{X_t\}$, carefully justifying your working.

(8)

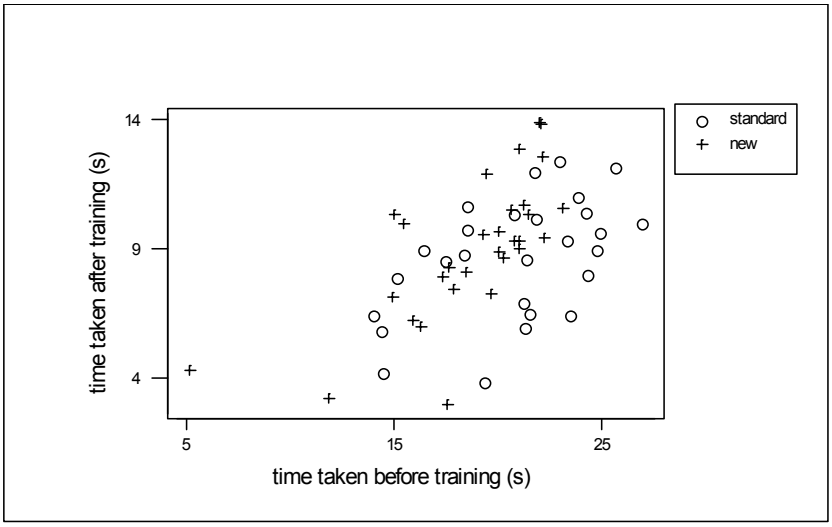
2. One of the roles of the applied statistician is to help clients by interpreting computer output for them. For each of the following scenarios, provide explanations that would be suitable for a client with a basic knowledge of statistics.

(a) A psychologist has run a study comparing two methods of training children to solve puzzles. He had 57 children whom he randomly divided into two groups. Each child was timed doing a puzzle. One group was then given standard training and the other group had a new training method. The groups were kept separate throughout. After training, each child was given a second, similar puzzle. The response of interest was the time taken to solve the puzzle, recorded in seconds.

The psychologist has done two-sample *t* tests on the final times and on the changes in times, and he has also done an analysis of covariance on the final times using initial times (mean-centred) as a covariate. He thinks the results are contradictory because of the *p*-values obtained, and wants to know the correct interpretation of his experiment. Note that negative changes in times correspond to a reduction in the time taken to solve the puzzle, and so an improvement in performance.

A graph showing the data, and a table summarising the output, are given below.

(10)



Method	*Estimate of group difference (new – standard)* [seconds]	*Standard error* [seconds]	*p-value*
t test on **final times**	0.37	0.67	0.58
t test on **changes in times**	2.46	0.79	0.0028
Analysis of covariance	1.24	0.56	0.032

Question 2(b) is on the next page

(b) It is sometimes necessary to pass a catheter into the vein or artery of a small child. The required length of catheter depends on the child's physiology. In a study of 12 children, their heights and weights were recorded and the corresponding catheter lengths were measured. The objective was to see how accurately weight and height could predict catheter length.

Three general linear models have been fitted with catheter length as the response variable. Summary output is shown below, where *S* is the square root of the error mean square. The clinician cannot interpret the output from the third model, and does not know which model to choose. Explain the apparent differences between the models and outline the advice you would give him.

(10)

Terms in regression model	*Parameter estimate*	*Standard error*	*S*	*R-squared*
Constant	12.1	4.3	4.0	0.78
Height	0.60	0.10		
Constant	25.6	2.0	3.8	0.80
Weight	0.28	0.04		
Constant	21.0	8.8	3.9	0.81
Height	0.20	0.36		
Weight	0.19	0.17		

3. The optimum conditions for extruding plastic film were examined using a technique called evolutionary operation. In the course of this study, three responses (X_1 = tear resistance, X_2 = gloss, X_3 = opacity) were measured at two levels (high and low) of each of the factors **rate of extrusion (extr)** and **amount of an additive (addit)**. Measurements were replicated 5 times at each combination of the factor levels.

(i) Explain why it is sensible to consider multivariate analysis of variance (MANOVA) for these data. State the distributional assumptions that you would need to make.

(5)

(ii) Partial output from a MANOVA is shown **on the next page**. Explain why several sets of test statistics are given in the output.

(4)

(iii) Wilks' Lambda (Λ) can be calculated from a ratio of determinants derived from relevant SSCP matrices. Write down the expression for Λ for the factor **rate of extrusion**. (There is no need to compute the values of the determinants.)

(4)

(iv) Use the output to carry out a MANOVA. State your hypotheses and conclusions clearly.

(4)

(v) Describe further analysis you could do to investigate the nature of any significant effects found in part (iv).

(3)

NOTE. When the number of degrees of freedom associated with the factor being tested is 1, the following result holds exactly under the null hypothesis:-

$$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{r-p+1}{p}\right) \sim F_{p,\,r-p+1}\,,$$

where Λ is the value of Lambda, p is the number of response variables and r is the number of degrees of freedom associated with the residual matrix.

Output for question 3 is on the next page

Output for question 3

MANOVA for extr

CRITERION	TEST STATISTIC	F
Wilks'	0.38186	7.554
Lawley-Hotelling	1.61877	7.554
Pillai's	0.61814	7.554
Roy's	1.61877	

MANOVA for addit

CRITERION	TEST STATISTIC	F
Wilks'	0.52303	4.256
Lawley-Hotelling	0.91192	4.256
Pillai's	0.47697	4.256
Roy's	0.91192	

MANOVA for extr*addit

CRITERION	TEST STATISTIC	F
Wilks'	0.77711	1.339
Lawley-Hotelling	0.28683	1.339
Pillai's	0.22289	1.339
Roy's	0.28683	

SSCP Matrix for extr

	x1	x2	x3
x1	1.740	-1.504	0.8555
x2	-1.504	1.300	-0.7395
x3	0.855	-0.739	0.4205

SSCP Matrix for Error

	x1	x2	x3
x1	1.764	0.0200	-3.070
x2	0.020	2.6280	-0.552
x3	-3.070	-0.5520	64.924

Turn over

4. An economist wants to develop a scale to measure doctors' attitudes to the cost of health care. She designs a questionnaire comprising 15 questions about attitude. Each of the responses is coded on a 5-point scale, where 1 means "strongly agree" and 5 means "strongly disagree". Completed questionnaires have been returned by 149 doctors.

(i) The economist wishes to devise subscales consisting of combinations of answers from subsets of the questions, and intends to use principal component analysis on the correlation matrix. Explain why this technique might be useful, and comment on the suitability of the method for these data.

(5)

(ii) Describe another multivariate method that could be used to explore relationships between questions on the questionnaire.

(2)

(iii) Unfortunately there are many missing values in the data. In fact nine questions have a response rate of less than 50%. The economist decides to restrict her analysis to the remaining six questions, which have been answered by all 149 doctors. Comment on whether you think this is a sensible strategy.

(3)

(iv) The six questions are as follows.

Q1 Best health care is expensive
Q2 Cost is a major consideration
Q3 I would determine the cost of tests before referral
Q4 I would monitor likely complications only
Q5 I would use all means irrespective of cost
Q6 I prefer unnecessary tests to omitting tests.

[You are reminded that low values indicate agreement with the statements given in the questions.]

A summary of the first three eigenvalues and associated principal components using the correlation matrix is given in the table below.

Eigenvalue	3.24	1.38	0.52
Q1	0.27	0.50	–0.29
Q2	–0.37	0.45	–0.51
Q3	–0.41	0.51	0.13
Q4	–0.35	0.26	0.68
Q5	0.49	0.34	0.42
Q6	0.52	0.32	–0.04

Interpret those principal components that you consider to be meaningful, justifying your answer. Would you want to see any further principal components from this analysis?

(6)

(v) What advice would you give to the economist about the scores she should use? Justify your answer.

(4)

Turn over

5. (i) (a) State the *Gauss-Markov* theorem.

(2)

(b) State the conditions under which you might consider using weighted least squares rather than ordinary least squares in simple linear regression.

(4)

(ii) Data were collected about the performance of children in 14 schools. The children took an aptitude test before going to the schools, and then they sat a national test in mathematics five years later. The researchers expected there to be a relationship between a child's scores in these two tests. The mean scores were presented for each of the schools, and are given in the table below.

School	*Number of pupils*	*Mean score on aptitude test*	*Mean score on mathematics test*
1	12	70.6	44.6
2	35	61.5	51.0
3	24	67.3	59.2
4	10	79.3	85.7
5	32	65.2	53.6
6	25	64.3	56.1
7	21	67.5	42.8
8	20	60.2	52.2
9	15	66.0	39.1
10	7	80.5	74.7
11	27	52.4	41.6
12	37	70.4	40.4
13	19	64.2	53.7
14	27	68.4	39.6

(a) Draw a scatter plot, describe the data, and comment on any apparent relationship between the schools' mean scores from the two tests.

(6)

(b) The two sets of output **on the next page** give the results of ordinary linear regression and weighted linear regression, where the weights are the numbers of pupils from each school. Compare and contrast the two analyses and state which one you think is more appropriate, justifying your answer.

(8)

Output for question 5 is on the next page

Turn over

Output for question 5

Regression Analysis

The regression equation is
maths = -26.1 + 1.17 aptitude

Predictor	Coef	Stdev	t-ratio	p
Constant	-26.14	29.14	-0.90	0.387
aptitude	1.1732	0.4327	2.71	0.019

R-sq = 38.0% R-sq(adj) = 32.8%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	921.4	921.4	7.35	0.019
Error	12	1504.2	125.3		
Total	13	2425.6			

Weighted analysis using number of children as weight

The regression equation is
maths = 8.1 + 0.636 aptitude

Predictor	Coef	Stdev	t-ratio	p
Constant	8.09	30.32	0.27	0.794
aptitude	0.6363	0.4611	1.38	0.193

R-sq = 13.7% R-sq(adj) = 6.5%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	4413	4413	1.90	0.193
Error	12	27809	2317		
Total	13	32222			

6. (i) Give a brief description of the *forward selection procedure* for selecting a multiple linear regression model. Your description should include the following.

Reasons for using model selection procedures.
The relative merits of the method.
The steps required to carry out the procedure.

(6)

(ii) A metal alloy was being developed to have a prescribed percentage elongation under stress. Varying amounts of additives *X1*, *X2* and *X3* were used to produce 24 experimental pieces of alloy, and their elongations were measured. Multiple linear regression models were fitted as shown in the table below.

Terms in model (in addition to intercept)	*Error sum of squares*
–	170.85
X1	165.97
X2	117.17
X3	122.43
X1, *X2*	116.30
X1, *X3*	121.75
X2, *X3*	90.007
X1, *X2*, *X3*	88.453

(a) Using the forward selection procedure, find the model which best explains the data. At each step, state clearly the hypothesis being tested, the value of the test statistic and the outcome of the test, specifying an appropriate significance level.

(7)

(b) Mallows' C_p statistic for a model containing s parameters is defined by

$$C_p(s) = \frac{SS_E}{\hat{\sigma}^2} - (n - 2s)$$

where SS_E is the error sum of squares for the model and $\hat{\sigma}^2$ is the variance estimated from the full model (which is the final model in the table above). Calculate $C_p(s)$ for each of the models in the table and comment, referring to desirable values of $C_p(s)$. Does the full model appear to provide the best fit to the data? Justify your answer.

(7)

7. (i) Explain what is meant by a *transformation to stabilise variance*, and give an example of where this might be useful in linear regression.

(3)

(ii) A random variable Y has mean μ and standard deviation σ, and X is defined as a function $X = h(Y)$ of Y.

Use a suitable expansion to derive approximate expressions for $E(X)$ and $\mathrm{Var}(X)$ in terms of the expected value and variance of Y. Hence, or otherwise, show that if σ is a function $f(\mu)$ of the mean μ, an appropriate variance-stabilising transformation for Y might be $h(Y)$ defined through the relationship

$$\frac{dh(Y)}{dY} = \frac{1}{f(Y)} .$$

(3)

(iii) Use the result of part (ii) to find a suitable transformation for each of the following cases.

The standard deviation is proportional to the mean.
The standard deviation is proportional to the square of the mean.
The variance is proportional to the mean.

(3)

(iv) An experiment was performed to examine the relationship between calcium and strength of fingernails. A sample of 32 students were given calcium supplements in amounts of either 10mg, 20mg, 30mg or 40mg. Eight students were allocated at random to each amount. After a set period, fingernail strength was measured. The results are given in the table.

10mg	*20mg*	*30mg*	*40mg*
13	114	48	79
42	75	72	314
28	27	104	75
12	136	133	108
67	100	57	90
62	83	112	175
28	56	87	108
29	37	64	74

Calculate suitable summary statistics and hence sketch a graph to help to examine possible transformations to stabilise variance for these data. Describe the practical problems of applying the theory from part (iii) when analysing the results of this experiment.

(5)

(v) Discuss how you would proceed to fit a suitable model. How would you decide between models that had the predictor variable (calcium supplement amount) as a covariate or a factor?

(6)

Turn over

8. An experimenter was investigating the effect of the order of presentation of a set of stimulus cards on the mean response time, X, of subjects. Three different orders were used, and the two other factors in the experiment were the sex and age of subjects. Each subject was tested only once. The data were as follows.

Response time, x (seconds)

Order	**Under 20**		**21 – 30**		**Over 30**		**Totals**
	Male	*Female*	*Male*	*Female*	*Male*	*Female*	
1	7 7 4	8 12 10	8 8 6	10 8 6	12 6 10	12 15 13	162
2	8 9 10	6 12 9	6 8 9	11 10 7	10 9 6	12 10 10	162
3	8 6 13	7 8 9	9 14 13	8 6 6	14 12 11	12 10 14	180
Totals	72	81	81	72	90	108	504

$\sum_{i=1}^{54} x_i^2 = 5068$.

(i) Copy and complete the Analysis of Variance table below.

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARE	*F* VALUE
Order (*O*)				
Sex (*S*)				
Age (*A*)				
$O \times S$		61.778		
$O \times A$		21.667		
$S \times A$		21.000		
$O \times S \times A$		16.556		
Residual				
TOTAL				

(7)

(ii) Examine the significance of main effects and interactions, and draw suitable interaction diagrams to help explain the results.

(7)

(iii) The experimenter has limited understanding of statistics. Write a report that will help him interpret the results of the experiment.

(6)