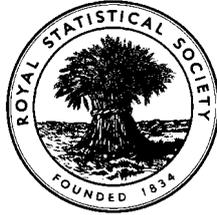**EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY**

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1997

Applied Statistics I

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.

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* $\log_{e}$.

1. An experiment was performed to compare two methods of instruction. Six instructors, selected at random from a large population of instructors, were allocated to the methods, three to each. Three trainees were then taught by each instructor and at the end of the training period were assessed for their proficiency. The proficiency scores achieved (x) were as follows:

	Method 1						Method 2					
	Instructors						Instructors					
	1		**2**		**3**		**4**		**5**		**6**	
	Trainee	**Score**	**Trainee**	**Score**	**Trainee**	**Score**	**Trainee**	**Score**	**Trainee**	**Score**	**Trainee**	**Score**
	1	47	4	73	7	68	10	43	13	42	16	36
	2	56	5	64	8	57	11	49	14	40	17	25
	3	49	6	69	9	52	12	60	15	39	18	30
$\sum x$		152		206		177		152		121		91
$\sum x^2$		7746		14186		10577		7850		4885		2821

(i) Write down an appropriate model for these data, explain the meaning of each term and state any assumptions made.

(ii) Complete the analysis including estimates of all the parameters in the model. Obtain an approximate 95% confidence interval for the difference in mean proficiencies between the two methods.

2. (a) Describe the concepts of leverage and influence in regression analysis.

(b) By considering the linear regression of the dependent variable y on a single regressor variable x, construct scatter plots demonstrating points with (i) high leverage and high influence; (ii) high leverage and low influence; and (iii) low leverage and low influence.

(c) To investigate factors affecting the cost of a service, a regression analysis with three regressor variables has been performed. The regression diagnostics provided by the computer package are given in the table below. The cases are numbered in the order in which they were observed.

Case Number	Observed Cost	Fitted Cost	Residual	Standardised Residual	Leverage	Cook's distance
1	13320	8817	4503	2.39	0.48	0.88
2	2850	1246	1604	1.06	0.69	0.38
3	5580	2900	2680	1.62	0.60	0.66
4	5055	5944	-889	-0.47	0.47	0.03
5	6060	7956	-1896	-1.01	0.48	0.16
6	4800	5553	-753	-0.34	0.27	0.01
7	4290	4410	-120	-0.07	0.61	0.00
8	4875	5720	-845	-0.42	0.41	0.02
9	3018	4686	-1668	-0.73	0.23	0.03
10	3180	4473	-1293	-0.87	0.68	0.26
11	4140	4933	-793	-0.43	0.49	0.03
12	3339	3868	-529	-0.32	0.60	0.03

Interpret these statistics.

3. (a) Describe the main features of the generalised linear model and its role in unifying the theory underlying a variety of statistical models.

(b) A random sample of voters has been classified by their sex, whether they are employed or not and if they intend to vote for party A. The results were:

	Intend to vote for party A			
	YES		NO	
	Male	*Female*	*Male*	*Female*
Employed	21	16	49	4
Not employed	24	18	6	42

A number of log linear models have been fitted to these data which gave the following deviances.

Terms in model	*Deviance*
μ (constant)	80.28
μ , Y, S, E	75.36
μ , Y, S, E, Y.S	75.25
μ , Y, S, E, Y.E	74.79
μ , Y, S, E, S.E	37.97
μ , Y, S, E, Y.S, Y.E	74.68
μ , Y, S, E, Y.S, S.E	37.86
μ , Y, S, E, Y.E, S.E	37.41
μ , Y, S, E, Y.S, Y.E, S.E	36.84

where S is the sex factor, E the employment factor and Y the voting factor.

(i) Report on whether voting intention varies with the other two factors.

(ii) How could these data be analysed for the same purpose using logistic regression? You are not expected to perform any calculations.

Turn over

3

4. A general linear model relating n independent observations of a response variable y to a predictor variable x is given by

$$E(y) = X\boldsymbol{\beta}$$

where X is an $n \times 2$ design matrix of full rank and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is a vector of parameters, the first parameter being an intercept.

(i) Three observations (y_i, x_i) $i = 1, 2, 3$ have been taken on this model with the results (4, 1), (6, 2) and (9, 3). Assuming $Var(y) = I\sigma^{2.}$, where I is the identity matrix, obtain the least squares estimate $\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ and also its variance-covariance matrix.

Calculate a 95% confidence interval for β_1, the regression slope. What assumption have you made?

(ii) It has just been discovered that $y_3 = 9$ is the mean of two observations taken at $x = 3$. Unfortunately the two individual values for y are not available. Calculate the weighted least squares estimator of $\boldsymbol{\beta}$ allowing for this extra information. You may assume that all observations are independent. State any further assumptions that you make. What is the advantage of knowing the individual values for y at $x = 3$ rather than their mean?

5. A time series X_t is generated by

$$X_t = \phi X_{t-1} + a_t - \theta a_{t-1},$$

where $|\phi| < 1$, $|\theta| < 1$ and a_t is white noise with zero mean and variance σ^2.

(i) By obtaining $E(a_t X_t)$ and $E(a_{t-1} X_t)$ show that

$$\rho_k = \frac{[1 - \phi\theta][\phi - \theta]}{1 - 2\phi\theta + \theta^2} \cdot \phi^{k-1} \qquad k \geq 1$$

where ρ_k is the autocorrelation function of lag k.

(ii) In each of the following cases, obtain the autocorrelation function and hence describe the series generated by the given values of ϕ and θ (i.e. specify the autoregressive moving average model).

(a) $\phi = -0.5, \theta = 0$

(b) $\phi = 0, \theta = 0.5$

(c) $\phi = \theta = 0.5$

6. A supermarket chain is interested in the effect on sales of the display position of a commodity. Nine supermarkets each note the sales of this commodity when placed in one of four different heights above a base level. Each height is used for one week at every supermarket, the order of heights being selected at random. The results below are the sales achieved at each height.

Supermarket	Heights above base level (cm)			
	0	30	60	120
1	92	96	94	86
2	106	110	116	108
3	86	89	85	85
4	78	95	85	78
5	124	128	120	118
6	98	100	96	100
7	68	72	73	67
8	75	79	76	74
9	106	100	104	104
Mean	92.6	96.6	94.3	91.1

Several models have been fitted to these data. A brief description of each model together with a partial analysis of variance is given below.

Model A: linear regression of sales on height.

Source	Sum of squares	Degrees of freedom
Regression	34.7	1
Residual	9377.6	34
Total	9412.3	35

Model B: Linear regression of sales on height and supermarket using supermarket number as a quantitative variable.

Source	Sum of squares	Degrees of freedom
Regression	399.7	2
Residual	9012.6	33
Total	9412.3	35

Model C: Linear regression of sales on height and supermarket with supermarkets coded using eight indicator variables

Source	Sum of squares	Degrees of freedom
Regression	9001.2	9
Residual	411.1	26
Total	9412.3	35

Model D: as Model C but including a quadratic term for height

Source	Sum of squares	Degrees of freedom
Regression	9083.2	10
Residual	329.1	25
Total	9412.3	35

Model E: cross-classified two-factor analysis of variance

Source	Sum of squares	Degrees of freedom
Supermarkets	8966.6	8
Heights	149.0	3
Residual	296.8	24
Total	9412.3	35

(i) Comment on the suitability of each model.

(ii) Using whichever models you feel appropriate, report on the effect, if any, of height on sales.

(iii) Explain, in non-statistical terms, the implication of there being no interaction term in model E.

7. Under certain circumstances, the horizontal distance (range) travelled by a projectile launched at an angle of 45° is given by $r = u^2/g$ where r is the range, u is the initial velocity and g is a constant.

(i) Assuming u has a distribution with mean μ and variance σ^2 show that, to a first approximation, $E(r) = \mu^2/g$ and find the variance of r.

(ii) In a series of launchings it is known that $\mu = 500$ metres/sec, $\sigma = 5$ metres/sec and $g = 9.81$ metres/sec^2. Using your approximation obtain the probability that the ranges of two successive launches differ by less than 1000 metres. You may assume that the ranges have a normal distribution. What other assumption have you made?

(iii) Calculate the required values for μ and σ which would result in the approximate probability in (ii) being 0.99 if the mean range is to be 10,000 metres.

8. A small study was conducted to investigate whether three methods of speech therapy were equally effective in teaching children with a speech disorder. Twelve children with the disorder were randomly allocated to the three methods, four children to each method. After the therapy each child was scored on technical ability (x_1) and self appraisal (x_2) with the following results:

	Method 1			Method 2			Method 3	
Child	x_1	x_2	*Child*	x_1	x_2	*Child*	x_1	x_2
1	87	52	5	86	62	9	67	78
2	91	67	6	82	69	10	64	90
3	99	40	7	88	58	11	75	64
4	90	59	8	73	65	12	73	78

The corrected sums of squares and crossproducts matrix for between methods, S_{method}, and the partially completed matrix for within methods, S_{within}, are as follows

$$S_{method} = \begin{bmatrix} 974 & -1022 \\ -1022 & 1074.7 \end{bmatrix} \qquad S_{within} = \begin{bmatrix} 290.3 & \\ & 797 \end{bmatrix}$$

(i) Calculate the missing elements in S_{within}.

(ii) For each of the following problems give a suitable model, explaining carefully the meaning of each of the terms:

   (a)   It is required to determine whether there is a difference in mean levels of x_1 between the three methods.

   (b)   Using a single test, it is required to determine whether there is a difference in mean levels of x_1 or x_2 between the three methods.

(iii) Give an example of a statistic which would be appropriate for performing the test in (b) and calculate its value. What assumptions have you made concerning the data?