UNIVERSITY OF SURREY©

B. Sc. Undergraduate Programmes in Mathematical Studies
M. Math. Undergraduate Programmes in Mathematical Studies

Level HE2 Examination

Module MS236 GENERAL LINEAR MODELS

Time allowed – 2 hours Autumn Semester 2006

Attempt THREE questions. If any candidate attempts more than THREE questions only the best THREE solutions will be taken into account.
A formula sheet will be provided.
Cambridge Statistical Tables will be provided.

SEE NEXT PAGE

Question 1

(a) Consider the simple linear regression model

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n.$$

Stating the usual assumptions about the errors (ε_i), write down the distribution of y_i. Obtain a representation of $\hat{\beta}$ in the form $\Sigma a_i y_i$, where $a_i = \frac{x_i - \bar{x}}{\Sigma(x_i - \bar{x})^2}$. Hence

(i) show that $\hat{\beta}$ is an unbiased estimator of β;

(ii) show that the variance of $\hat{\beta}$ is $\frac{\sigma^2}{\Sigma(x_i - \bar{x})^2}$, where σ^2 is suitably defined;

(iii) write down the distribution of $\hat{\beta}$.

[10]

(b) A botanical researcher wants to estimate the number of larvae of a particular species of beetle within brackets of the birch bracket fungus. When the brackets are stored in the laboratory, the larvae within them mature and the beetles emerge to be counted. A sample of 25 brackets was collected. Their weights in grams (x) and the numbers of beetles (y) that they were shown to contain are given in the folowing table.

x	y	x	y	x	y	x	y	x	y
62	16	226	50	175	26	255	73	226	95
99	40	150	40	25	3	200	49	77	0
178	56	283	91	192	48	125	25	177	61
307	64	42	0	99	10	122	1	88	15
201	98	183	60	63	0	296	98	162	15

Summary statistics for these data are

$S_{xx} = 152086.24$, $S_{xy} = 53839.32$, $S_{yy} = 26151.76$, $\Sigma x_i = 4013$, $\Sigma y_i = 1034$, $\Sigma x_i^2 = 796253$.

(i) Fit a simple linear regression model to these data. Write down $\hat{\alpha}$, $\hat{\beta}$ and the equation of the fitted model. [4]

(ii) Find a 95% confidence interval for α. [4]

(iii) Calculate $(X^T X)^{-1}$, where X is the design matrix for the experiment. Hence, show that the leverage of the ith observation corresponding to a bracket of weight x_i is:

$$\frac{1}{3802156}(25x_i^2 - 8026x_i + 796253)$$

Obtain the leverage of the observation with $x_i = 307$ and determine whether or not this observation can be said to be influential. [7]

SEE NEXT PAGE

Question 2

(a) (i) Write down models for a Latin square design and for a Græco Latin square design. In each case, state clearly any assumptions that are made. [4]

(ii) A set of data arising from a Græco Latin square design has a coresponding ANOVA table. Outline the differences in the ANOVA table that would be obtained if the data were treated as if it came from a Latin square design rather than from a Græco Latin square design. [4]

(b) An engineer was investigating the effect of four tyre treatments (A, B, C, D) on the mileage that a tyre covers before it is considered to be worn. Four vehicles were selected for the study. The engineer believes that tyre position is significant. Here is the Latin square design that was used and the data (in units of 10,000km) that were obtained.

	Tyre Position				
Vehicle	1	2	3	4	Total
1	$C = 11$	$B = 10$	$D = 14$	$A = 18$	53
2	$B = 8$	$C = 12$	$A = 10$	$D = 12$	42
3	$A = 9$	$D = 11$	$B = 7$	$C = 15$	42
4	$D = 9$	$A = 8$	$C = 18$	$B = 6$	41
Total	37	41	49	51	178

The totals for each tyre treatment are: $\Sigma y_A = 45$, $\Sigma y_B = 31$, $\Sigma y_C = 56$ and $\Sigma y_D = 46$. The sum of squares of the observations is 2174.

Produce an ANOVA table corresponding to the data and state your conclusions clearly. [9]

(c) Four drivers completed the driving in the above experiment. The engineer suggested that the driver of a vehicle may represent an additional source of variation. A fourth factor 'driver' (α, β, γ, δ) was introduced, and inclusion of this 'driver' factor gave the following Græco Latin square.

	Tyre Position				
Vehicle	1	2	3	4	Total
1	$C\beta = 11$	$B\gamma = 10$	$D\delta = 14$	$A\alpha = 18$	53
2	$B\alpha = 8$	$C\delta = 12$	$A\gamma = 10$	$D\beta = 12$	42
3	$A\delta = 9$	$D\alpha = 11$	$B\beta = 7$	$C\gamma = 15$	42
4	$D\gamma = 9$	$A\beta = 8$	$C\alpha = 18$	$B\delta = 6$	41
Total	37	41	49	51	178

The totals for each driver are: $\Sigma y_\alpha = 55$, $\Sigma y_\beta = 38$, $\Sigma y_\gamma = 44$ and $\Sigma y_\delta = 41$.

Amend the ANOVA table obtained in part (b) to take account of the fourth factor. Is 'driver' a significant factor? Has including the extra factor affected any conclusion that was made in part (b)? [4]

(d) The engineer would like to extend this experiment by including yet one more factor at four levels, namely 'road surface'. Suggest how the design of part (c) could be changed to take account of this extra factor. [4]

SEE NEXT PAGE

Question 3

An industrial experiment is carried out to determine factors that affect the viscosity of a polymer. 30 observations are collected. Four possible explanatory variables under investigation are: $X1$ reaction temperature, $X2$ catalyst feed rate, $X3$ pressure and $X4$ stirring rate. The dependent variable, Y, is a measure of viscosity. The following table gives the model deviances on fitting all possible combinations of the explanatory variables.

Model	Deviance	Model	Deviance
Null	1139.1	$X2 + X3$	288.94
$X1$	415.84	$X2 + X4$	289.60
$X2$	316.89	$X3 + X4$	292.73
$X3$	935.89	$X1 + X2 + X3$	284.32
$X4$	310.52	$X1 + X2 + X4$	287.01
$X1 + X2$	308.67	$X1 + X3 + X4$	282.56
$X1 + X3$	370.30	$X2 + X3 + X4$	270.23
$X1 + X4$	302.09	$X1 + X2 + X3 + X4$	268.41

(a) (i) Obtain an estimate of the variance based on the full model. [2]

(ii) Find the best model using backwards elimination. [7]

(iii) Find the best model using forwards fitting. [4]

(b) The data presented is part of a larger study involving 11 possible explanatory variables. Explain why it would not be sensible to produce a table showing the deviances for all possible regression models in this case. [2]

(c) Define Mallow's C_p statistic and R^2 and comment on the use of these statistics in general. Calculate both statistics for the two models selected in part (a). Comment on the values that you obtain and hence on the chosen models. [6]

(d) Give two possible disadvantages of the backwards elimination method. [4]

SEE NEXT PAGE

Question 4

(a) Write down the linear model for a one-way analysis of variance. Explain what each term in the model represents and state any assumptions necessary for the analysis to be valid. [4]

A study on the effect of different types of traffic signals on traffic delay was carried out at road junctions. Three types of traffic signals were used in the study, namely; pretimed, partially traffic-activated and fully traffic-activated. Five junctions were used for each type of traffic signal. The measure of traffic delay in the study, y_{ij}, was the average stopped time per vehicle at a junction. The following data were obtained.

Pretimed	Partially Traffic-activated	Fully Traffic-activated
36.6	17.5	15.0
39.2	20.6	10.4
30.4	18.7	18.9
37.1	25.7	10.5
34.1	22.0	15.2

For this set of data $\Sigma_i \Sigma_j y_{ij}^2 = 9596.03$.

(b) Construct the analysis of variance table and test for differences between the three types of traffic signals. [6]

(c) Write down two orthogonal contrasts that can be used to provide estimates of:

(i) a difference between the Partially and Fully Traffic-activated signals

(ii) a difference between the Pretimed signals and any signal involving Traffic-activation.

Show that the contrasts are orthogonal. [4]

(d) Compute the sum of squares for each contrast, and show that their sum is equal to the treatment sum of squares in the analysis of variance. Test separately that the diffences detailed in part (c) are zero. State your conclusions in context. [7]

(e) Data are available on another five junctions at which there is no traffic control, adding a fourth column to the three in the table above. Explain how you could break the new treatment sum of squares up into 3 components to correspond to the differences detailed in part (c) and to a comparison of three traffic control methods versus no traffic control method. [4]

SEE NEXT PAGE

Question 5

(a) The tables below give cell means for four different 2×4 factorial experiments.

(i)

	b_1	b_2	b_3	b_4
a_1	10	12	14	16
a_2	8	10	12	14

(ii)

	b_1	b_2	b_3	b_4
a_1	10	14	12	16
a_2	12	10	8	14

(iii)

	b_1	b_2	b_3	b_4
a_1	10	12	14	16
a_2	14	12	10	8

(iv)

	b_1	b_2	b_3	b_4
a_1	12	12	8	8
a_2	8	8	12	12

In each case sketch an appropriate plot and hence comment with reasons on whether or not an A effect, a B effect and an $A \times B$ interaction are present. [8]

(b) A clinical trial involving 24 patients is carried out to investigate if the progress of patients recovering from hip replacement operations is affected by the physiotherapy regime used and by the therapist carrying out the treatment. There are two regimes of physiotherapy that are under consideration. Four therapists each take a group of three randomly selected patients on one regime and another group of three patients on a different regime. The patients are assessed by recording a measure of increase in mobility gained by the end of the course of treatment. The results for the kth patient, treated under the ith regime, by the jth therapist, y_{ijk}, are as follows. The numbers in brackets indicate the total for each set of three patients.

Regime	Therapist 1			Therapist 2			Therapist 3			Therapist 4		
1	9	11	12	17	18	18	22	23	25	11	13	13
		(32)			(53)			(70)			(37)	
2	12	13	15	19	21	22	22	25	26	15	15	16
		(40)			(62)			(73)			(46)	

For this set of data $\Sigma_i \Sigma_j \Sigma_k y_{ijk}^2 = 7695$.

(i) Write down the appropriate model, explaining what each term represents. [2]

(ii) Produce the Analysis of Variance table. [7]

(iii) Do the two factors, Regime and Therapist, interact? [2]

(iv) Is there any indication that either Regime or Therapist influences recovery, as measured by increase in mobility? [4]

(v) What conclusions do you draw from parts (iii) and (iv) about the Regimes and the Therapists? [2]

INTERNAL EXAMINER: J.D. Godolphin
EXTERNAL EXAMINER: W. Krzanowski