UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday 1 June 2001 9 to 12

PAPER 28

APPLIED STATISTICS

*Attempt any **FOUR** questions. The questions carry equal weight.*

Candidates will be handed appropriate pages of S-Plus5 output at the start of the examination.

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 S. Chinn and R.J. Rona published data on the prevalence of overweight children in 1974, 1984 and 1994 in England and Scotland, in their article in the British Medical Journal of 6 January 2001. A subset of their data is reproduced in the S-Plus output attached, together with the results of S-Plus analysis. Here

$p =$ proportion of children overweight, and
$n =$ number in the sample.

Explain carefully the resulting S-Plus output, and illustrate the interaction by an appropriate sketch graph.

How might you use a Poisson regression for these data?

2 The Cambridge University Reporter Special of 8 December 2000 gives, as Table 9, the data shown in the attached S-Plus output on Applications & Acceptances in 2000 and 1999 classified by College: the figure shown is the percentage from *maintained* schools and F.E. institutions. The last line of the Table corresponds to 'Mature Student Colleges', i.e., Hughes Hall, Lucy Cavendish, St Edmund's and Wolfson.

(i) Discuss carefully the regression model applied, and its output. What diagnostic checks would you wish to try?

(ii) The published data is of course a summary of the original 25 2×2 contingency tables. Suppose that for the year 2000, we have the original frequencies denoted, for the jth college, as

		Accepted: yes	Accepted: no
Applications	maintained	n_{11j}	n_{12j}
	other	n_{21j}	n_{22j}
		n_j	

If you had this set of $2 \times 2 \times 25$ frequencies, with the factors "app", "acc" and "college" set up appropriately, with $2, 2, 25$ levels respectively, what would you learn from the result of

$$\text{glm}(n \sim (\text{app} + \text{acc}) * \text{college}, \text{poisson})$$

applied to the vector of frequencies (n_{abj})?

3 The dataframe UScrime gives aggregate data on 47 states of the USA for 1960. The response variable y, given in the final column, is the rate of crime in a particular category per head of population. There are 15 explanatory variables; most of these, and also y, have been rescaled to convenient numbers. The 15 explanatory variables are

M	percentage of males aged 14-24
So	indicator variable for a southern state
Ed	mean years of schooling
Po1	police expenditure in 1960
Po2	police expenditure in 1959
LF	labour force participation rate
M.F	number of males per 1000 females
Pop	state population
NW	number of nonwhites per 1000 people
U1	unemployment rate of urban males 14-24
U2	unemployment rate of urban males 35-39
GDP	gross domestic product per head
Ineq	income inequality
Prob	probability of imprisonment
Time	average time served in state prisons

(a) Discuss the interpretation of the model

`first.lm`

in the S-Plus output attached.

(b) The Venables and Ripley library(MASS) function stepAIC has been applied to `first.lm` to reduce this model. Noting that

$$\text{AIC } = n \log \frac{(\text{residual sum of squares})}{n} + 2p,$$

where p = number of parameters in the linear model, interpret the resulting (edited) output, and the final model `last.lm`.

4 See the attached edited S-Plus output, which will be handed to you at the start of the examination.

(a) A client comes to you with data from a randomised-controlled trial to investigate the effectiveness of a new chemotherapy treatment for colorectal cancer. There are 35 patients in the new chemotherapy group and 21 patients in the standard treatment group. He has calculated the two-year and five-year survival probabilities for the group receiving the new chemotherapy treatment and for the group receiving the standard treatment. He has found that the two-year and five-year survival probabilities for the new chemotherapy group are 0.879 and 0.524 respectively. The corresponding survival probabilities for the standard treatment group are 0.680 and 0.367 respectively. From these findings he concludes that the new chemotherapy treatment works. You have plotted the Kaplan–Meier curves for the two treatments with their corresponding survival tables.

How do you respond to your client's conclusion? Do you agree or disagree? (Give reasons.) What do you advise? Why is it preferable to compare two Kaplan–Meier curves overall rather than at specific time points? What are the median survival times for the two groups?

(b) Also given is the output from an analysis of data, unrelated to (a) above, on the survival times (in months) for rectal cancer patients treated with either a low dose ("GROUP=0") or a high dose ("GROUP=1 ") radiotherapy regimen. (Data-set taken from page 17 of Harris and Albert (1991) - Survivorship Analysis for Clinical Studies. Males are coded 0 and females are coded 1. The age of the patient is recorded at the time of entry into study.)

Comment on the S-Plus commands, the analysis done and the results (**Detailed derivations of the underlying techniques, e.g. coxph( ), are not required.**)

5 Venables and Ripley (1999) show a 4-way classification of 1681 householders in Copenhagen, who were surveyed on

(i) the type of rental accommodation they occupied

(ii) the degree of contact they had with other residents

(iii) their feeling of influence on apartment management, and

(iv) their level of satisfaction with their housing conditions - this variable will be treated as the response variable.

You may assume that, for each (i, j, k), the distribution of the frequencies (n_{ijkl}) is such that $(n_{ijkl}) \mid n_{ijk+}$ is a multinomial, parameters n_{ijk+} and (p_{ijkl}),
where $\sum_l p_{ijkl} = 1$ for each i, j, k
and i, j, k, l correspond, respectively to
Influence, Type, Contact and Satisfaction.

(These are independent multinomial variables, as i, j, k vary.) Discuss carefully the attached SPlus5 output, and derive the sufficient statistics for the two models discussed.

How might binomial logistic regression be applied to these data?