Department of Statistics

COURSE STATS 330

Model answers for Final Exam, 2004

Multiple Choice:

Question	Answer		Question	Answer
1	3		11	1
2	1		12	1
3	5		13	5
4	5		14	3
5	3		15	2
6	4		16	5
7	3		17	4
8	1		18	2
9	2		19	4
10	4		20	2

Section B

Q1

a) The indicated model fits six parallel lines, one to each panel of the coplot. This seems more or less consistent with the patterns in the coplot, particularly for the panels where the data extend across the width of the panel.

b) The output suggests that the there is no interaction between HeadWt and the other factors. This confirms that the parallel line model in (a) is reasonable. In fact an even simpler model, ignoring Date, also seems reasonable. We would want do a test to confirm this.

c) The model here is fitting 6 arbitrary (i.e. not necessarily parallel) lines, one to each panel. The large dfbetas all correspond to terms determining the slope and intercept of the line corresponding to Line 52, Date 20, the middle plot in the top row. The influential points are the ones whose removal would have a big change on the line. These will probably be the points at the extreme ends of the weight range in that panel. Take your pick!

d) In this new model the lines are now constrained to have equal slopes. All 60 points now contribute to the estimation of this slope. Removing one point out of 60 will have a smaller effect than removing 1 point out of 10.

e) There is a strong suggestion that date has no effect on the response, as all terms involving data have insignificant p-values. Line 52 cabbages have significantly higher contents. This

effect does depend on HeadWt, as this is highly significant in the summary. The heavier the head, the lower the ascorbic acid content, for a fixed line.

Q2

a) Initially fit a saturated model of the form

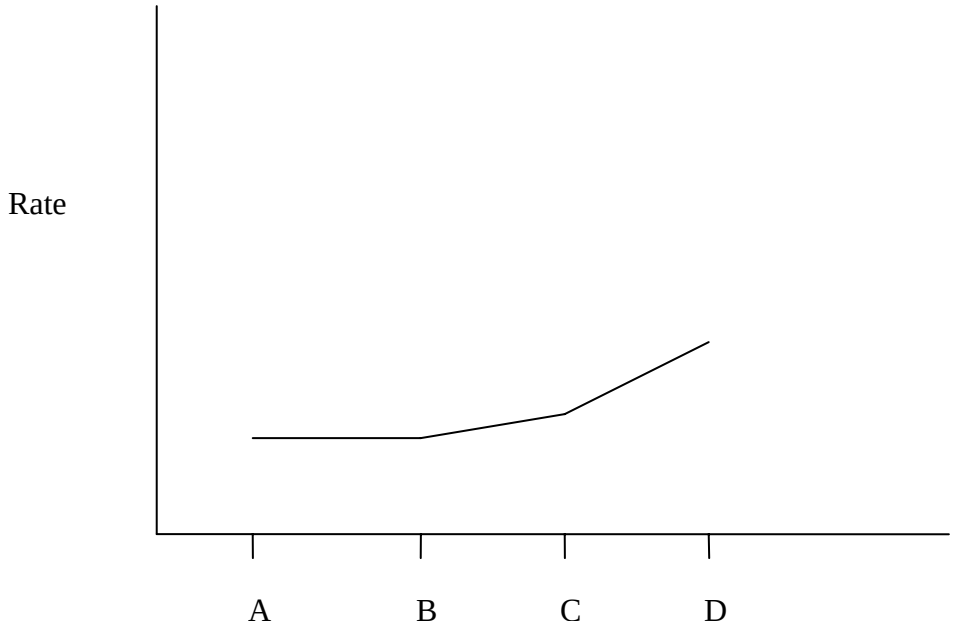
```
cbind(num,tot-num) ~ sex*serum*age
```

Use anova, stepwise to select a suitable submodel. Can use the residual deviance to check fit of a submodel, and diagnostic plots to see which factor-level combinations are poorly fit by the chosen model.

b) Inspection of the diagnostic plots reveals that points 5 and 13 are poorly fitted by the model. Inspection of the data reveals that the rates for males are at least twice the rates for females, except for 5 and 13 (age group 50-62, Serum A) This suggests that the male and female data may have been transposed for these two groups.

c) To discuss the rates and the effect of the factors, a graph is helpful:

This could be drawn by hand, eg one of the form



with one line for each age-sex combination, using the predicted data supplied. We can then see that

- age is the most important factor (has the biggest coefficient in the summary)
- Sex is the second most important factor, followed by serum

- As age increases, probability of CHD goes up.
- Males have higher risk than females
- Serum A and serum B about equal, but C and D have higher probabilities, with D higher than C

Q3

a) and b) taken from lecture on Simpson's paradox.

c) In the marginal table, collapsing over Departments, there is a strong association between Gender and Admission. The p-value is tiny corresponding to 93.45 on 1 df.

d) When Department is taken into account, the gender:Admit and 3-factor interactions are insignificant, suggesting that gender and admission are conditionally independent, given department. Thus the conclusion from the marginal table is different from that based on the conditional tables. This is because the condition for collapsing over Depts is not met (the Dept-gender and dept-admit interactions are highly significant).