Department of Statistics

STATS 330/762

Model answers for part B, 2012 Examination

26. (a) The initial model would include interactions between X and A and W and A. [1 mark].

The code to fit this is lm(y~A*X + A*W) [1 mark].

The model fits three non-parallel planes [1 mark].

Could simplify by dropping interactions [1 mark].

Test using anova function and examining p-values for the interaction terms. Could also compare the model with interactions to the model without interactions using the anova function. [1 mark]

[5 marks altogether]

(b) There seems to be a significant improvement in the model when dgdg is added (see the p-value in the anova table), so there seems to be a relationship between dadg and adg for at least one of the breeds. [2 marks] In the anova table there is no evidence of interaction although there is weak evidence in the summary table that the slope of the line for Breed 2 is different to that of Breed 1. [2 marks] There is strong evidence that the intercepts of the lines are different in both the summary table (the coefficient for Breed 2 has a p-value of 0.0148) and the anova table (p=6.577e-13). [2 marks]

From the summary table it seems that the lines for Breed 2 and Breed 1 are different, so that the effect of the ADG of the dam does depend on the breed. [2 marks]

[8 marks altogether]

(c) Cook's D is a measure of the overall effect on the estimated regression coefficients of deleting a data point. It is defined mathematically by

$$D = \frac{\left(\hat{\beta}-\hat{\beta}(-i)\right)X^T X\left(\hat{\beta}-\hat{\beta}(-i)\right)}{ps^2}$$

where $\hat{\beta}$ is the estimate , $\hat{\beta}(-i)$ is the estimate with the ith data point removed, and p is the number of regression coefficients. [1 mark]. Points are influential if Cooks D exceeds the 10% or the 50% percentage point of an F distribution with p and n-p degrees of freedom. [1 mark]

[2 marks altogether]

(d) The flagged points, together with the reason for flagging them and the effects are shown in the table below. Each line is worth 1 mark.

Point	Reason	Effect
1	DFBETAS all large	On all regression coefficients
2	DFBETAS for dadg large, COV ratio large	Coef of dadg, std errors
3	COV ratio large	std errors
13	COV ratio large	std errors
16	COV ratio large	std errors

[5 marks altogether]

Total for Q 26: 20 marks

27 (a) The two definitions are contained in the lecture slides [2 marks each]. A saturated model (aka maximal model) is one that places no restrictions on the parameters. It has deviance zero since the model being compared to the maximal model is in fact the same maximal model.

[6 marks altogether]

(b) We need to test if the probabilities of a successful free throw are the same for each of the 23 games. [2 marks]. The model corresponding to equal probabilities is the null model, so we can test if the probabilities remain the same using the null model deviance, which is 33.376 on 22 df.[2 marks]. The p-value is 0.056, so there is weak evidence that the probabilities are not the same. Thus the criticism is only very weakly justified. [2 marks]. The model used is the one-way binary anova model (logistic regression). [2 marks]

[8 marks altogether]

(c) Under and over-dispersion occur when the variance of the number of successes is less than (more than) that given by the binomial distribution. Can be caused by the binomial trials not being independent. [2 marks]. Could be happening here (learning effect, loss of confidence). [2 marks] Likely effect is to inflate the true standard errors so the estimated standard errors will be too low (exaggerated significance, so the real evidence for inconsistency may be weaker than that suggested above). [2 marks]

[6 marks altogether]

Total for question 27: 20 marks

28. (a) Split te joint table into separate tables, one for each level of C. The conditional odds ratios are the OR's in the conditional tables corresponding to the different levels of C. [4 marks] [If A and B are conditionally independent given C the conditional OR's will be equal to 1. [2 marks]

[6 marks altogether]

(b) See Lecture 30.

[4 marks altogether]

(c) The indicated model is the homogeneous association model (since the 3-factor interaction is not significantly different from zero). [3 marks]. In this model, the odds ratios in the conditional table corresponding to Eject = Yes are the same as the odds ratios in the conditional table corresponding to Eject = No. (same conclusion if we condition on Seatbelt of Injury) [2 marks]

[5 marks altogether]

(d) The odds ratio conditional on Eject = Yes is the same as the odds ratio conditional on Eject = No.

We can interpret it as (odds of not wearing a seatbelt for the fatal injury group)/ (odds of not wearing a seatbelt for the non-fatal injury group).

The estimated log-odds is 1.71732 with a standard error of 0.05402 (we use the Seatbeltyes:InjuryNonfatal line in the table).

The CI is exp(1.71732 +/- 1.96*0.05402) or (1.611, 1.823) to 3 decimal places.

[5 marks altogether]

Total for question 28: 20 marks.