Department of Statistics
COURSE STATS 330
Model answer for Final Exam, 2003

Part A

In the exam on the Web page, the **first alternative** is always the correct answer.

Part B

Note that full marks can be gained by brief answers!

Question 1

(a) All of the following contain some information about transforming the explanatory variables:

- Plot of residuals versus fitted values
- Plot of residuals versus explanatory variables
- Added variable plots
- Partial residual plots
- GAM plots

The last 2 are usually the most informative, particularly the last one.

(b) The Box-Cox plot: a plot of the negative of the profile likelihood versus the transformation power. The power corresponding to the minimum of the graph gives the correct power p for the transformation y^p.

(c) Yes, as the R2 has increased from 0.46 to 0.62.

(d) There are a few outliers/influential points. The worst are points 6 and 16, with bad DFBETAS/COV RATIO etc. Delete and refit.

Question 2

(a) Fit a saturated (Maximum) model with all interactions. Then use the R functions anova and step to find a simpler model with fewer interactions. To check that the chosen model fits well, we can use graphical diagnostics (residuals, Cooks D, leave-one out deviance change etc) and also the residual deviance (OK since the data are grouped.)

(b)

- The probability of death
- Is higher for Species B than Species A
- Goes up as exposure increases
- Goes down as relative humidity increases
- Goes up as temperature increases.

(c) Under Model 1, the effect of exposure (with all other factors at baseline) is as follows:

Exposure 2 weeks: log odds is -4.2147
Exposure 3 weeks: log odds is -4.2147 + 2.2390 = -1.9757
Exposure 4 weeks: log odds is -4.2147 + 3.1841 = -1.0306

Under Model 2, the effect of exposure (with all other factors at baseline) is as follows:

Exposure 2 weeks: log odds is 2.1138 -12.4662/2 = -4.1193
Exposure 3 weeks: log odds is 2.1138 -12.4662/3 = -2.0416
Exposure 4 weeks: log odds is 2.1138 -12.4662/4 = -1.00275

Thus the results from the two models are very similar. Note that since there are no interactions, the effect of changing exposure is the same for all levels of the other factors.

There seems to be little to choose between the models. Model 2 is a reasonable choice (the deviance difference has a p-value of 0.4909).

Question 3

(a) A is independent of B and C if

$$\text{Prob}(A=i, B=j, C=k) = \text{Prob}(A=i)\ \text{Prob}(B=j, C=k)$$

for all possible factor levels i, j and k. This is equivalent to all interactions between A on the one hand and B and C on the other being zero, or, the model $A + B*C$.

(b) A table can be collapsed over a factor if the factor is independent of the rest.

(c) From the anova table, the model Age + Paralysis*Vaccine fits well. This is the model corresponding to Age being independent of Paralysis and Vaccine, so the relationship between Paralysis and Vaccine does not depend on age.

(d) The interactions are significant (p=0.00016) so paralysis is not independent of vaccine. OK to ignore age since Age is independent of the other two factors, so we can collapse the table over Age.