STATS 330

Model answers for Final Exam 2006, Part B

B1 (a) We detect departures by (i) plotting residuals versus fitted values and looking for a funnel effect, or (ii) plotting squared residuals against fitted values and smoothing.

Two ways to deal with the problem: (i) transform the response to stabilise the variance, or (ii) use weighted least squares.

Consequence: standard errors are wrong.

B1(b) Leave one out diagnostics are

DFBETAS: standardised difference in coefficients – measures effect on a particular coefficient
Cook's D: measures overall effect on the coefficients
DFFITS: standardised difference in fitted value – measures effect on the fitted value
COV RATIO: measures effect on the standard errors

Note HMD's are not leave one out diagnostics, they measure leverage.

B1(c) Point 19: Minor effect on the constant term, coeff of lime, standard errors. A big residual, but not high leverage.
Point 21: Minor effect on the cow.den, fitted value. A small residual, but high leverage.
Point 33: Effect on constant, prop, fitted value. A big residual, but low leverage.
Point 42: Slight effect on prop. A big residual, but low leverage.
Point 66: Effect on cow.den. Low leverage.

Points 19, 21, 33 seem to be having the biggest effects, although they are not catastrophic. However, there is no really strong reason to remove them, unless they can be shown to be the result of typographical errors. Removing them makes the regression appear to be better than it really is.

Another possibility is that the outliers may cease to be a problem after a transformation. There is a hint of curvature in the residuals/fitted value plot. A transformation may mean that the points are no longer outlying after the transformation.

B2(a) The model survived~age.gp*pclass*sex seems to fit very well, as an even smaller model is chosen by the stepwise algorithm. Note that data are in ungrouped form, so the p-value is not reliable. It is clear from the anova output that the 3-factor action is

not required, and that the interaction between age and class is also not significant, although it is included in the model selected by step. Thus, either of the models survived~av.age.gp + pclass + sex+ av.age :pclass + sex: av.age + sex:pclass or survived~av.age + pclass + sex+ sex: av.age + sex:pclass
seem OK.

B2(b) For a first-class passenger, the log-odds of survival is 2.48940 + 0.01283*av.age , as the other main effects and interaction terms vanish as sex and class are at their baselines. For age group 30-34, the log-odds of survival is is 2.48940 + 0.01283*33.84137 = 2.923585, and the probability is exp(2.923585)/(1+ exp(2.923585)) = 0.949.

B2(c) If we treat age as a factor, then the effect of age on the log-odds is no longer constrained to be linear. This means that the probabilities are far less constrained. If age is treated as a numeric variable, the model consists of 6 straight lines, one for each class/sex combination. This is not as general as treating age as a factor. In the case of the model survived~age.gp *pclass *sex, the log-odds and probabilities are completely unconstrained.

B2(d) The model treating age as a numeric variable would constrain the traces in the Figure to be straight lines. The model treating age a factor allows the traces to assume any shape suggested by the data. We can see that straight lines are not appropriate for some of the traces, e.g. female 1^{st} class, male 2^{nd} class.

B3(a) Fit a Poisson regression model counts~V*S using the counts as responses and V and S (sex) as the explanatory variables. To test if the male and female populations differ in their voting behaviour, test for a zero interaction in the model.

B3(b) Fit a Poisson regression model using the counts as responses and V, I and S (sex) as the explanatory variables. To test if the joint distribution of V and I is the same in the two populations, test if the model counts ~ S + V*I fits well (i.e. if V and I are independent of S)

B3(c) Party preference does not differ between the sexes, as the affil:sex interaction is not significant.

B3(d) We need to test if support is independent of sex and status i.e. if the model count ~ support + sex*status fits well. The output indicates that it doesn't, so support does depend on sex and status.