Final exam 2011: Model answers to part B

26 (a) High leverage point: Has an extreme value of one or more covariates [2 marks]
Influential point: makes a big difference to the regression if the point is removed [2 marks]

(b) DEBETAS: measures change in an individual coefficient
DFFITS: measures change in the predicted value of a response
Cook's D: measures overall change in the coefficients
COVRATIO: measures change in the standard errors
Hat matrix diagonal: measures leverage of a point
[1 mark for each of these. An extra mark was given for any attempt to define these mathematically.]

(c) Thresholds:
Cook's D: The red dotted lines on the leverage-residual plot, representing the values 0.5 and 1
DFBETAS: |DFBETAS|>1
DFFITS: |DFFITS|>3sqrt(p/(n-p))
COVRATIO: {COVRATIO-1|>3p/n
HMD: HMD>3p/n
[1 mark for each, an extra mark of you mentioned the connection between Cooks D and the F-distribution.]

(d) For this example, n=60, p=5 so 3p/n = 0.25 and the DFITTS threshold is 0.905. Point 5 had a big DFITTS value and all the points exceeded the COVRATIO threshold. All the other measures were OK.
[2 marks for the thresholds, 1 each for the conclusions]

27 (a) Define the deviance: let L_{MAX} be the maximum of the log-likelihood under the maximal model (ie substitute the sample proportions for the p's) and let L_{MOD} be the maximum of the log-likelihood under the logistic model (ie substitute the fitted logistic probabilities for the p's). Then the deviance is $2(L_{MAX}-L_{MOD})$.
[2 marks for the definitions of L_{MAX}, L_{MOD} and two marks for the deviance formula.]

Can use the deviance as a goodness-of-fit measure when the number of covariate patterns is small and the number of observations in each covariate pattern is large. [2 marks]

(b) The residual deviance p-value is 1-pchisq(13;362,12) = 0.3432871 [2 marks]. Since this is large, there is no evidence the model is not OK. [2 marks]

(c) Can do the test both for grouped and ungrouped data [1 mark]. The test statistic is the difference of the residual deviances for the two models. [2 marks] When the smaller model is adequate, the difference in the deviances has a chisquare distribution, with df equal to the difference in the dfs for the two residual deviances. [1 mark]

(d) The model fits well, as the p-value of the residual deviance is large. [2 marks]. The data are grouped so we can make this interpretation, and also interpret the residuals [1 mark]. The residual plot shows no suspicious residuals [1 mark]. The plot of the logits versus age is reasonably straight so we don't have to transform age. [2 marks].

28 (a) A and B are independent if P(A=i and B=j) = P(A=i)P(B=j), or equivalently if the population odds ratio is 1. [2 marks]. Can test independence by fitting a saturated Poisson regression to the counts and seeing if the interaction is significant.

(b) In this case the interaction p-value is significant (p=0.00259) [3 marks] so the hypothesis of independence is rejected. [2 marks]

(c) The odds ratio is exp(-0.37858) = 0.68. [2 marks] Thus the odds of being a smoker in the surviving group is less than for the non-surviving group. [2 marks] This is counter-intuitive since smoking supposedly kills. [1 mark] (an equivalent interpretation is that the odds of surviving are lower in the smoking group).

(d) When we condition on age, a different picture emerges. The OR is now exp(0.42741)=1.533, significantly different from 1. [2 marks] Now the odds of being a smoker in the surviving group is more than for the non-surviving group, as we would expect[1 mark]. The association reversal is caused by the strong dependence of smoking on age (older people smoke less) and survival on age (older people have a lower chance of survival). [2 mark].