Department of Statistics

COURSE STATS 330

Final Exam, 2002. Model Answers

1. Short answer questions.

(a) Answer

(i) A normal plot of the residuals is designed to detect non-normal errors. When there is no problem the points on the plot cluster about a straight line. When the data are not normal, the points cluster about a curve.
(ii) A plot of residuals versus lagged residuals is designed to detect serial correlation in the errors. Of there is no serial correlation, the plot has no pattern. When there is serial correlation, the plot clusters about a line.

(b) Answer:

(i) Multicollinearity occurs when there are near linear relationships between the explanatory variables.
(ii) It is detected by examining the variance inflation factors: if these are large (say more than 10-20) then there is a problem.

(c) Answer:

(i) $\pi = \exp(-40.971 + 23.040\ \text{Dose}) / (1 + \exp(-40.971 + 23.040\ \text{Dose}))$, $(\log \pi / (1-\pi) = -40.971 + 23.040\ \text{Dose}$.
(ii) Multiplies by $\exp(23.040 \times 0.1) = 10.01416$, so increases the odds approximately 10-fold.
(iii) A probability of 0.8 corresponds to a logit of $\log(.8/(1-0.8)) = 1.386294$, so $-40.971 + 23.040\ \text{Dose} = 1.386294$, or $\text{Dose} = 1.838$.

(d) Answer:

The plots indicate that there is a problem with the fitted model. The deviance residual plot, which is designed to detect points with big residuals, shows that the model does not fit well at points 6,7 and 8. The leverage plot (designed ti detect points that are outliers in Dose) suggests that none of the points have great influence, although the points at either end of the Dose range naturally have the greatest influence. The Cooks distance plot and the deviance change plot (both designed to show the changes in the regression when one point is deleted) show that it is point 8 that is causing the trouble, it has a big effect on the fitted regression. The remedy to the problem would be to modify the fitted regression model, possibly by adding a quadratic term in Dose, as there is a suggestion that the effect of the gas on mortality increases and then decreases slightly.

2.

(a) *Answer*:

(i) Zero mean, same variance, normally distributed, independent

(ii) Least squares: coefficients are chosen to minimize the sum of squared residuals.

(b) *Answer*:

(i) Estimated coefficient indicates that the mean response (SO2) increases by 0.06497 for each extra plant.

(ii) Interval is estimate ± standard error × $t_{df}(\alpha/2)$. In this case, the estimate is 0.06497, the standard error is 0.01823, the df are 33, and α is 0.05. We could calculate the t percentage point in R using qt(0.975,33).

(c) *Answer*:

(i) The hypothesis being tested is that the true regression coefficient of manu is zero.

(ii) The studentised residuals have the same variance.

(iii) The plot seems to indicate a funnel effect, implying that the variances of the observations increase with the mean.

(iv) When transforming the response, the distribution, the variances and the relationship between the mean and the covariates are all affected. In this case we hope the variances will be more similar.

(d) Answer:

The two possible models are model A, deleting pop and days, and model B, deleing days. Model B has the largest adjusted R^2 and smallest Cp, but model A has one fewer parameter and the Cp closest to p+1. On balance, I favour model A.

3.

(a) Answer:

(i) This is the estimated error standard deviation i.e. the estimate of σ. It measures the scatter about the fitted plane, realatively small in comparison with the means

(ii) This is the value of R^2: the ratio of the regression sum of squares to the total sum of squares. It measures the goodness of fit (how palanar theb data are). Here the value is 97.8%, indicating a good fit.

(iii) This is the test statistic for testing the hypothesis that all the regression coefficients (except the constant) are zero. The zero p-value indicates this hypothesis is decisively rejected.

(b) Answer:

The interaction term is significant, and positive, indicating that the effect of ncreasing distance is not linear.

(c) Answer:

The first interval is a confidence interval for the mean delivery time for *all* trips involving 25 cases being transported over 50 meters. The second interval is the prediction interval for predicting how long one individual trip will be.

(d) Answer:

(i) This indicates that if the regression is refitted without point 9, the new regression will be very different from the old.

(ii) The influence statistics indicate that point 9 has a huge influence on all aspects of the regression. The average hat matrix diagonal is 4/25, so the "3 times average" cutoff is 12/25=0.48. The HMD for point 9 is 6 times the average. This is because point 9 has very large values for both cases and distance. The covariance ratio is very large (should be between about 0.5 and 1.5) indicating that point 9 has a big effect on the standard errors. The DFBETAS should be less than $2/\sqrt{n}$ or about 0.4, so the effect on the interaction in particular is very large. The answer given in part (b) is almost completely determined by point 9 and cannot be trusted.

4.

(a) Answer:

(i) In Test 1, the hypothesis being tested is that that dose and time are unrelated to the response, i.e. that the probability of death is the same for all observations. The alternative hypothesis is that the logistic model dose + time is the correct model. In Test 2, the hypothesis being tested is that the logistic model dose + time is the correct model. The alternative is that the probability of death is arbitrary.

(ii) Test 2 shows that the fitted model is inadequate: the deviance is too large. Test 1 shows that there is a relationship between the probability of death and dose and/ or time.

(b) Answer:

Using log time gives a much smaller residual deviance: the p-value is considerably more than 0.05, so the model is acceptable.

(c) Answer:

Note that the output refers to a confidence interval for the log-odds. The interval for the log-odds is $0.7288 \pm 0.0839*1.96$ or (0.5644, 0.8933). The corresponding interval for the probability is
$(\exp(0.5644)/(1+\exp(0.5644)) , \exp(0.8933)/(1+\exp(0.8933)))$
or (0.6375, 0.7096).

(d) Answer:

The odds ratio is the ratio of the odds for high and med doses, at time t say. The odds for high dose are $\exp(-3.03858 + 3.16095 + 0.5499 \log(t)) = 1.1302 \exp(0.5499 \log(t))$. The odds for medium dose are $\exp(-3.03858 + 2.119920 + 0.5499 \log(t)) = 0.3990 \exp(0.5499 \log(t))$. The ratio is $1.1302/0.3990=2.8326$. This doesn't depend on t as the terms in t cancel out.

5.

(a) Answer:

(i) A:C + A:M + M:C + M:G

(ii) A*C*M + M:G

(b) *Answer*:

(i) Gender is related to marijuana use (there is a M:G interaction)

(ii) Given marijuana use, gender is conditionally independent of alcohol use. Only paths from G to A pass through M.

(c) Answer:

This is a 20x topic, we didn't cover it.

(d) Answer:

(i) No, because marijuana use is not independent of A and C.

(ii) Could use logistic regression to examine the distribution of marijuana use, conditional on gender, alcohol and cigarette use. (we didn't discuss this in class).