Department of Statistics

Paper 475.330 **1998 Exam Solutions** **Date: Oct 21**

1. (a) E(Y) is a linear function of the regressors, errors are N(0, σ^2) and independent.

   (b) (1) Linearity of response surface
   (2) Normality of errors
   (3) Constant variance

   (c) Non constant variance. Weights should be proportional to $1/\sigma_i^2$.

   (d) A partial regression indicates whether a regressor (X_j say) is needed in the model and if it should be transformed.
   (1)Regress X_j on all other regressors.
   (2)Regress Y on all regressors except X_j.
   Plot residuals from (2) versus residuals from (1).

   (e) p is the power used in the Box-Cox transformation
   $$Y^{(p)} = \frac{y^p - 1}{p}$$
   RSS$^{(p)}$ is the Residual sum of squares for $Y^{(p)}$.
   The plot is used to determine a suitable transformation to make observations approximately normal.

2. (a) (i)This model has logit(π) being a linear function of CS_2. The same line applies to both replications.

   (ii)This model is like (i) except we have seperate lines for rep1 and rep2 but these lines must be parallel.

   (iii)This is like model (ii) but now the lines do not have to be parallel.

(b) The P-value for CS_2 tests that CS_2 is needed in the model (ignoring reps and CS_2:reps).

The P-value for reps tests that reps is needed in the model given that CS_2 is in the model but CS_2:reps is not in the model.

The P-value for CS_2:reps tests that CS_2:reps is needed in the model given that both reps and CS_2 are in the model.

(c) We work our way up from the bottom of the table. Doing this we conclude CS_2:reps can be dropped, from the model that contains all 3 terms. Then we conclude that reps can be dropped from the model that contains just CS_2 and reps. Now we just have CS_2 in the model and the first line indicates that we should not drop CS_2.

(d)

$$\hat{\pi} = \frac{\exp(-14.808 + 0.24917CS_2)}{1 + \exp(-14.808 + 0.24917CS_2)}$$

for cs2=70

$$\hat{\pi} = \frac{\exp(-14.808 + 17.442)}{1 + \exp(-14.808 + 17.442)}$$
$$= 0.933$$

(e) A constructed variable plot is used to determine the type of power transformation that may be useful for a numerical regressor. Ths slope of the line indicates the power (p).

$$p = slope + 1$$

For our plot we have a slope of 2.8 with a s.e. of ≈ 1. Therefore we should consider transformations between 2 and 6. (Usually try to use a transformation at the lower end of this scale.

3. (a) The model that contains the 3 factors and the dept:gender and dept:result interactions.
Association Graph

(b) Gender and result are conditionally independent (conditional on dept). This means that for each department gender and result are not related. However overall gender may be related to result because the probability of being admitted is different for different departments and the proportion of male/female applicants is also different for different departments.

(c) A 2-way table involving gender and result would be misleading in this situation. It woud indicate that there is a relationship between gender and department but obscure that this is due to the differences in number of applicants to the different departments for different genders.

(d) A 2-way table involving dept and result would be reasonable since for each department gender and result are independent.

4. (a)

$$\text{Odds ratio} \;=\; \frac{7 * 31530}{86 * 951} = 2.70$$

The odds of an infant congenital malformation are 2.7 times as large if the mother drinks $\geq$1 drinks per day during the first 3 months of pregnancy than if she drinks $<$1 drink per day.

(b) A 95% CI for log(odds ratio) is :
$\log(2.70) \pm 1.96\sqrt{1/86 + 1/31530 + 1/7 + 1/951}$
$(0.220, 1.766)$
Therefore a 95% CI for the odds ratio is :
$(\exp(0.220), \exp(1.776))$
$(1.25, 5.85)$

(c) It is possible to estimate the relative risk (not a case-control study).

$$\text{relative risk} \;=\; \frac{7/(951+7)}{86/(31530+86)} = 2.69$$

5. (a) We have grouped data so a goodness of fit test is valid. We test
H_0:The model is valid

The test statistic is
χ^2_0 = residual deviance = 4.430

To get the p-value find
P-value = $\Pr(\chi^2 \geq 4.430)$
where $\chi^2 \sim$ Chi-square (df=8)

(b) $\text{logit}(\hat{\pi}) = 2.522 - 2.984 + 0.3757 + 0 + 1.327 = 1.2407$

$$\hat{\pi} = \frac{\exp(1.2407)}{1 + \exp(1.2407)} = 0.776$$

(c) No points have unusually large residuals or are high leverage points. However both points 3 and 12 are influential points since they standout on the Cooks Distance plot and the Deviance changes plot.

(d) The DFFITS value for observation 12 indicates that if this point is deleted the fitted value for the regression surface will change by 4.286 s.e.'s at that point (low moisture, s.temp=62, g.temp=21).
The DFBETAS value for g.temp indicates that if point 12 is deleted the coefficient for g.temp will decrease by 1.865 std. errors.

6. (a) The R^2 value indicates that $\approx$ 81% of the variability in log(oxy) can be explained by this model. The F-test has a very small p-value indicating that the model clearly does have some predictive value for the response. However, most of the P-values for the individual coefficients are quite large (only 1 is less than 0.10). This indicates that not all the regressors are needed and suggests we may have multicollinearity among the regressors.

(b) There is some indication of curvature in the plot of residuals vs fitted values. The normal plot is acceptable.

Observation 1 has a somewhat large residual and observation 17 has a large leverage (h_{ii}). We note that observation 17 has a very unusual value for TVS which should be checked.

(c) VIF's are used to detect multicollinearity amng the regressors. There are 3 moderately large VIF's in this set corresponding to BOD, TS and COD.This indicates there is one or more near linear relationships involving these 3 regressors.

(d) My preferred model would contain TS and COD as regressors. It has the lowest C_p value and close to the maximum value of adjusted R^2 (also close to the minimum value of s^2). One other model is worth considering and that contains TKN, TS and COD as regressors. This model has the maximum value of adjusted R^2 (and the minimum value of s^2) and only a slightly higher value of C_p than the previous model.

(e) Data that is collected over time may be serially correlated. We can check by plotting r_i vs r_{i-1} (residuals vs lagged residuals). If serial correlation is not present then this plot should just be random scatter. If it is present then a trend (usually positive) should be indicated.