STATS 500 - Practice Exam 2 Solutions

April 20, 2011

1. There are some highly correlated predictors (`Ex0` and `Ex1`, `X` and `W`), which suggests collinearity will be a problem. Also, the number of predictors ($p = 12$) is somewhat large relative to the sample size ($n = 47$).

2. Pick the model with `Age, Ed, Ex0, U2, W` and `X` – it has the lowest C_p statistic.

3. They have more bias than OLS estimates but less variance, typically combining so that the mean squared error is lower.

4. 7 components.

5. $0.1 \times 135.82 = 13.582$ days.

6. $2.65 - (-19.95) = 22.60$ days longer for the fly with one pregnant female.

7. The easiest way to do this in practice would be to change the reference level to one of these two levels, say to 10. Then the p-value in the output for `treat81` tests the hypothesis of no difference between groups 10 and 81.

8. Even though `thorax` is uncorrelated with the predictor of interest, it is known to be correlated with response, so omitting it would worsen the fit and inflate the residual standard error, making comparisons less precise and potentially losing signficance.

9. The main feature is heteroscedasticity, which suggests a transformation of the response.

10. Must be missing in a non-informative way.

11. Square root – this is close to the maximum (at about 0.45) yet still interpretable.

12. No. Because the relationship would no longer be linear, we would need to know the treatment and the thorax length to answer this question. You could find the difference in `longevity` on the transformed scale but not in the original scale.

13. 1358.2.

14. $H_0 : \beta_m = 0$, i.e., there is no difference in mean concentration between clones a and m.

15. c – that's why the SE for c is larger than all the other SEs. The SE in one-way ANOVA is determined by group size.

16. a, c and l.

17. The 95% confidence interval on the difference between f and ℓ is $(-712.80 - (-290.0)) \pm 2.01 * 151.0$ or $(-726.31, -119.29)$. Since 0 is not in the interval, there is evidence that the there is a difference between these two clones.

18. **Your are not responsible for this problem – material not covered in class**

    In this case we consider the contrast of $\frac{\alpha_j+\alpha_k}{2} - \frac{\alpha_h+\alpha_i+\alpha_m}{3}$. The Scheffe's confidence interval is given by

$$\frac{\alpha_j + \alpha_k}{2} - \frac{\alpha_h + \alpha_i + \alpha_m}{3} \pm \sqrt{12 \cdot 1.95} \cdot 238.8 \cdot \sqrt{\frac{1}{5}\left(2 \cdot \left(\frac{1}{2}\right)^2 + 3\left(\frac{1}{3}\right)^2\right)}$$

or the result is $(-194.1, 749.0)$ and so we do not have evidence (at a significance level of .05) that there is a difference between the two types of potatoes.

19. The deviance cannot be used to do a goodness-of-fit – assumptions for the deviance to be valid in testing goodness-of-fit for GLM binomial model is that the values of n_i are not too small. In this case, they are all 1, so that clearly violates the assumption.

20. The difference in odds is given by

$$\begin{aligned}\text{odds(third quartile)} &= \exp\left((36.6 - 27.30) * (.0877)\right) \cdot \text{odds(first quartile)} \\ &= 2.26 \cdot \text{odds(first quartile)}\end{aligned}$$

i.e., the odds of testing positive for diabetes increased by factor of 2.26 in moving from the first to the third quartile in bmi (body mass index), assuming all other predictors were held constant.

21. The diastolic pressure does appear to be significant in the regression model and it has a negative coefficient suggesting that higher diastolic pressure corresponds to lowering the odds for testing positive for diabetes, given everything else is held constant. This can be consistent with the plot since the other variables are potential confounding variables. As an example, high blood pressure and high bmi may be correlated and we see that bmi clearly has the effect of increasing odds of testing positive for diabetes.

22. The value of η (i.e., logit of p) is given by

$$\eta = -8.240 + .125 * 1 + .0335 * 100 - .0135 * 70 + .0877 * 25 + .896 * .6 = -2.98$$

and utilizing the inverse logit function, we have

$$p = \frac{\exp(\eta)}{\exp(\eta) + 1} = .048.$$

23. **This problem was deleted – based on material not covered**

24. Let W be diagonal matrix with $W_{ii} = \frac{1}{\kappa i}$. Then the best linear unbiased estimator is

$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

where X is $n \times 2$ matrix with i^{th} row being $[1 \;\; x_i]$ and y is vector of y_i's. The covariance matrix of $\hat{\beta}$ is $(X^T W X)^{-1}$.