STAT 8620 — Categorical Data Analysis and Generalized Linear Models
Homework 4 – Due Thursday, Oct. 18, 2012
SHOW ALL WORK

- Homework is due by 4:30 on the due date specified above. You may turn it in to me during class, slip it under my door or send it to me via e-mail. I will post homework solutions shortly after all homeworks have been collected. **No late homeworks will be accepted without permission granted prior to the due date.**

- Use only standard (8.5×11 inch) paper.

- Homework should show enough detail so that the reader can clearly understand the procedures of the solutions. This is **absolutely essential** for you to receive full credit for your answer.

- Problems should appear in the order that they were assigned.

Assignment:

1. The file visc.dat contains data from an experiment performed in a two-way layout without replication; that is, one observation was obtained at each combination of the levels of two factors: naphthenic oil level (P) and filler level (F), which have 4 and 6 levels, respectively. The response was the viscosity (V) of a compound manufactured at the levels of these two factors.

    a. Consider a model of the form

    $$z_{ij} = \mu + \alpha_i + \gamma_j + e_{ij}, \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2) \qquad (*)$$

    where z_{ij} is a Box-Cox transformation of y_{ij}, the viscosity at the ith level of factor P, jth level of factor F. Plot the profile likelihood $p\ell(\lambda)$ for the Box-Cox transformation parameter λ, obtain an approximate 90% confidence interval for λ and suggest an appropriate transformation for y. Apply this transformation, fit the model, and report an appropriate summary of the fitted model.

    b. Using the transformation found in part (a), now fit the model

    $$z_{ij} = \beta_0 + \beta_1 P_{ij} + \beta_2 F_{ij} + e_{ij}, \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2) \qquad (**)$$

    where P_{ij} and F_{ij} are the numeric values of P and F for the i, jth experimental unit. That is, now treat P and F as continuous covariates in a multiple regression model for z rather than factors in an ANOVA model for z. If possible, test whether model (**) fits no better than model (*) using an F test for nested linear models and using an asymptotic likelihood ratio

test based on Wilks' theorem. Compare the results of the two tests and explain how the testing approaches differ.

c. Now consider the model

$$w_{ij} = \beta_0 + \beta_1 P_{ij} + \beta_2 F_{ij} + e_{ij}, \quad \{e_{ij}\} \overset{iid}{\sim} N(0, \sigma^2) \qquad (\dagger)$$

where w_{ij} is a Box-Cox transformation of y_{ij} chosen based on model (†) rather than model (*). Estimate the transformation parameter λ for this model, fit the model and, if possible, test this model against model (*) (i.e., test if it fits no better than model (*)). Comment on the transformation chosen here as compared with in part (b). Is it surprising that the two transformations differ?

2. Consider a dose-response problem where y_i = the number of dead insects out of m_i insects exposed to a dose x_i, $i = 1, \ldots, N$. Assume $y_1, \ldots, y_N$ are independent with $y_i \sim \text{Bin}(m_i, \pi_i)$ where

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, N. \qquad (*)$$

We are interested in estimating the LD-p_0 where $p_0 = .5$ for the LD50, $p_0 = .90$ for the LD90, etc. For the given p_0 let x_0 be the corresponding dose value (that is, let x_0 denote the value of LD-p_0).

a. On pp.160–161 of the course lecture notes we used Fieller's method to derive an approximate $100(1-\alpha)\%$ confidence interval for the LD50. Now generalize this method to derive an approximate $100(1-\alpha)\%$ confidence interval for LD-p_0 for any given value of p_0.

b. Argue that the profile loglikelihood function for x_0 can be constructed by fitting a series of logistic regressions of the form

$$\text{logit}(\pi_i) = \beta_1(x_i - x_0), \quad i = 1, \ldots, N,$$

to the data, one model for each value of x_0 along a grid spanning a range of plausible values for x_0, each with offset equal to $\text{logit}(p_0)$. A plot of the maximized loglikelihood values reported for each of these regressions versus the series of fixed values taken for x_0 then forms the profile loglikelihood function for x_0. Explain **clearly and thoroughly** why this approach is correct.

c. Using the budworm data discussed in class, compute an approximate 95% confidence interval for the LD90 for dose (not ldose) using (i) a Wald interval, (ii) an interval based on Fieller's method, and (iii) a profile likelihood interval. Base your intervals on model m1: $y_{ij} \overset{ind}{\sim} Bin(m, \pi_{ij})$ where

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{ldose}_{ij} + \beta_2 \text{male}_{ij} + \beta_3 \text{male}_{ij}\text{ldose}_{ij} \qquad (\text{m1})$$

d. Will the point estimate of the LD50 in the linear logit model (*) be the same as in the linear probit model? How about the model with complementary log-log link? Explain your answer.

3. Do problem 5.2 in Agresti. In part (b) also compute a 95% confidence interval for the requested probability. In part (d), also conduct Hosmer & Lemeshow's goodness of fit test and summarize both how it was conducted and its results.

4. Exercise 5.4 in Agresti.

5. Exercise 5.16 in Agresti.

6. Exercise 5.19 in Agresti.

7. Exercise 5.37 in Agresti.

8. Exercise 6.1 in Agresti.