Stats 539 Homework 5: Due Thursday Mar. 9 by 10:50am

1. Agresti Exercise 6.1 (p. 223-4). The multivariate generalization of the exponential dispersion family is

$$f(\mathbf{y}_i; \boldsymbol{\theta}_i, \phi) = \exp\{[\mathbf{y}_i^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)]/a(\phi) + c(\mathbf{y}_i, \phi)\},$$

where $\boldsymbol{\theta}_i$ is the natural parameter (a $(c-1) \times 1$ vector).

2. The data in `Alligators.csv` (on course webpage) is from a study of factors influencing the primary food choice of alligators. The study captured 219 alligators in four Florida lakes. The nominal response variable is the primary food type, in volume, found in an alligator's stomach: F = fish, I = invertebrate, R = reptile, B = bird, O = other. (The other category consisted of amphibian, mammal, plant material, stones or other debris, or no food or dominant type.) The study also classified the alligators according to the lake captured (1 = Hancock, 2 = Oklawaha, 3 = Trafford, 4 = George), gender (1 = male, 2 = female), and size (1 = small ($\leq$2.3 meters long), 2 = large (>2.3 meters long)).

   (a) Produce two plots that illuminate the relationship between one or more of the explanatory variables and primary food type. (Since these are all categorical variables, you may need to be creative!) Write a few sentences describing what each plot shows you about these data.

   (b) Fit the baseline-category logit model for alligator food choice based on an indicator variable for size ($s = 1$ if small, $s = 0$ if large) and indicator variables for each lake except Lake George (L_H, L_O, L_T, L_G). Use fish as the baseline category. Write the equation of the fitted model. Choose two of the estimated coefficients and write a sentence interpreting each of the chosen coefficients.

   (c) Calculate and interpret a 95% confidence interval for the effect of size on the conditional odds π_I/π_R adjusting for lake, where π_I is the probability an alligator's primary food type is invertebrate, and π_R is the probability an alligator's primary food type is reptile.

   (d) What is the estimated probability that a small alligator in Lake Oklawaha has invertebrates as the primary food choice?

   (e) An alternative fitting approach for the baseline-category logit model fits binary logistic models separately for the $c-1$ pairings of responses. The estimates have larger standard errors than the maximum likelihood estimates for simultaneous fitting of the $c-1$ logits, but Begg and Gray (1984) showed that the efficiency loss is minor when the response category having highest prevalence is the baseline. Illustrate, by showing that the fit using categories fish and invertebrate alone is

$$\log\left(\frac{\hat{\pi}_I}{\hat{\pi}_F}\right) = -1.69 + 1.66\, s - 1.78 L_H + 1.05 L_O + 1.22 L_T$$

   with standard error values $(0.43, 0.62, 0.49, 0.52)$ for the effects. Compare with the model in part (b).

3. For the alligator food choice data in the previous problem, at one of the lakes the alligators' actual length (rather than a size indicator variable) was measured in meters. Download the data from this lake here:

   http://www.stat.ufl.edu/∼aa/glm/data/Alligators3.dat.

   Read the data into R and fit the baseline-category logit model for alligator food choice based on length. (Note that only fish, invertebrates, and other food choice categories were recorded at this lake).

   (a) Choose one of the "length" coefficients and write a sentence interpreting this coefficient.

   (b) Produce a single well-labeled plot of the estimated multinomial probabilities (y-axis) versus length (x-axis) by food choice category (line type and/or color with legend). Turn in the R code you used to fit the model and create the plot. Write a few sentences describing the features of the plot.

4. Agresti Exercise 6.20 (p. 226-7)

5. Agresti Exercise 6.21 (p. 227). Note that the data are shown in the first part of the R output on p. 227.