Categorical Data Analysis: HW 4

1. In class, we showed that random variables following normal, Poisson and Bernoulli distributions are members of the exponential family. What about the binomial? Since the binomial mean is $n_i\pi_i$ and we don't want the n_i appearing in the mean, we need to look at the success *proportions*. So, let n_iY_i have a binomial(n_i, π_i) distribution, where Y_i is now the success proportion out of n_i Bernoulli trials.

   (a) Show that the distribution of Y_i belongs to the exponential family and find the expressions for θ_i, $b(\theta_i)$, $a(\phi_i)$ and $c(y_i; \phi_i)$. Hint: $P(Y_i = y_i) = P(niY_i = n_iy_i) = \binom{n_i}{n_iy_i}\pi_i^{n_iy_i}(1-\pi_i)^{n_i-n_iy_i}$ (since n_iY_i is binomial and all you have to do is to write this in exponential family form.)

   (b) Find the expression for the deviance under this setup.

2. In class, we showed that the likelihood equation for independent Bernoulli(π_i) random components in a GLM with link $g(\pi_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$ have the form

$$\sum_{i=1}^{n}(y_i - \pi_i)x_{ij} = 0, j = 1, 2, \ldots, p.$$

   Find the likelihood equation when fitting a GLM to Poisson random components with mean μ_i and link function $g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$.

3. Refer to the logit model $\text{logit}(\pi_i) = \alpha + \beta x_i$ for the success probability of independent Binomial responses $Y_i \sim Bin(n_i, \pi_i)$.

   (a) Write down the log-likelihood under this model and find the sufficient statistics for α and β. Find the expected values of the sufficient statistics under the model.

   (b) Set up the likelihood equations and show that the likelihood equations for the logit model equate the sufficient statistics for α and β to their expected values (as in any GLM with canonical link).

   (c) Show that the information matrix for (α, β) does not depend on y_i. (Hence, the observed information matrix (the Hessian) equals the expected (Fisher) information matrix, as in any GLM with canonical link.)

4. Refer to the Horseshoe crab data set analyzed in class.

   (a) Using just a crabs carapace *width* as a predictor for having satellites, fit a logistic regression model and interpret the parameter associated with width.

   (b) Give the results of an appropriate test that tests if width is a significant predictor.

   (c) Give and interpret a 95% confidence interval for the effect of width.

(d) Plot the original 0-1 responses versus width and overlay the fitted logistic response curve.

(e) Predict the probability of having a satellite for a crab of mean width (i.e., a crab whose width equals the observed mean width).

(f) Obtain a 95% confidence interval for the probability of having a satellite for a crab of mean width.

(g) In the plot from part (d), include (pointwise) confidence limits for the probability of having a satellite.

(h) Check goodness of fit by forming 8 width intervals. I actually found the R command `cut` that can do this quit efficiently and generate the table of the number of successes and failures in each width category:

```
> width1=cut(width, breaks = c(min(width)-1, 23.25, 24.25, 25.25,
26.25, 27.25, 28.25, 29.25, max(width)+1))
> table(width1,sat)
              sat
width1         FALSE TRUE
  (20,23.2]        9    5
  (23.2,24.2]     10    4
  (24.2,25.2]     11   17
  (25.2,26.2]     18   21
  (26.2,27.2]      7   15
  (27.2,28.2]      4   20
  (28.2,29.2]      3   15
  (29.2,34.5]      0   14
```

For each width category, find the estimated proportion of having a satellite at the midpoint of the interval using the fitted model, and then compare observed and fitted values using the X^2 or G^2 statistic.

(i) Add spine condition and color to the model (treat both as categorical). Test if spine condition is really needed in the model.

(j) Refer to the model that includes width and color. Plot the estimated probabilities of having a satellite versus width, for each color.

(k) Find the confidence interval for the odds of having a satellite for dark (color=4) versus medium dark(color=3) crabs of the same weight.