Homework 3 Statistics 8320 (Wikle) February 12, 2016

**DUE: February 25, 2016**

1. Consider the negative binomial distribution for random variable $Y$ in the form

$$f(y) = \left( \begin{array}{c} y+k-1 \\ k-1 \end{array} \right) p^k (1-p)^y, \ \ y = 0, 1, 2, \ldots$$

(a) Given that if the parameter $k$ is assumed fixed and known, this distribution is a member of the natural exponential family, show $b(\theta) = -k\log(1-e^{\theta})$ (note, all logarithms in this problem are natural logs), $E(Y) = k(1-p)/p$, $var(Y) = \mu + \mu^2/k$, and report the values of $\theta$, $a(\phi)$, $c(y, \phi)$.

(b) For this distribution, the canonical link is $\log(\mu/(k+\mu))$. Justify this.

(c) Show that the deviance under the canonical link can be written:

$$D = 2\sum_{i=1}^{n} y_i \log\left[\frac{y_i k + y_i \mu_i}{\mu_i k + y_i \mu_i}\right] + k\log\left[\frac{k+\mu_i}{k+y_i}\right]$$

(d) Assuming a (natural) log link ($\eta_i = \log(\mu_i)$) and observations $i = 1, \ldots, n$ and $j = 1, \ldots, J$ parameters in $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, show that

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{k(y_i - \mu_i)}{(k+\mu_i)} x_{ij}$$

[Note: it is uncommon to use the canonical link for negative binomial models - most people use a log link.]

(e) As in part (d), find the $(j,k)$th element of the Fisher information matrix (as defined in the class notes) and show that it is equal to

$$\sum_{i=1}^{n} \frac{\mu_i k}{k+\mu_i} x_{ij} x_{ik}.$$

(f) Now, assume we have the following 7 observations and associated covariates:

| y | $x_1$ | $x_2$ |
|---|---|---|
| 3 | 0 | 35 |
| 16 | 1 | 60 |
| 12 | 1 | 25 |
| 1 | 0 | 20 |
| 18 | 1 | 50 |
| 8 | 0 | 55 |
| 9 | 1 | 30 |

In addition, we are given the initial parameter estimates $\hat{\boldsymbol{\beta}}^{(0)} = (2.5, -0.5, -.01)'$ and $k = 15$. Implement one step of the Fisher scoring algorithm with the log link (note that there is an intercept in the model):
$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{F}'\boldsymbol{V}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$.

Report $\hat{\boldsymbol{\beta}}^{(1)}$ and $\mathbf{F}$, $\mathbf{V}$, $\boldsymbol{\mu}$ (evaluated at $\hat{\boldsymbol{\beta}}^{(0)}$). Using the Wald test, evaluate the two linear hypotheses that $\beta_1 = 0, \beta_2 = 0$. Note: for this test, use the $\hat{\boldsymbol{\beta}}^{(0)}$ values as your estimates (and the associated $\mathbf{F}$ and $\mathbf{V}$ matrices); note: I'm only asking this so that everyone has a common answer - in reality, you would use the converged estimates form the Fisher scoring algorithm! Report $\mathbf{L}$.

(g) School administrators were interested in predicting the attendance behavior of high school juniors at two schools. Predictors of the number of days of absence include gender of the student and standardized test scores in math and language arts. We have attendance data on 316 high school juniors from two urban high schools in the Excel data file `nbreg.csv` on the class Blackboard site. The response variable of interest is days absent. The variables **math** and **langarts** give the standardized test scores for math and language arts respectively. The variable **male** is a binary indicator of student gender (1 if male, 0 if female). Fit a negative binomial model to these data. Note, in SAS output, the dispersion parameter estimate is the estimate of $1/k$ from the distribution presented above. Some things you should consider: how do you interpret your output relative to the goal of the experiment, justify that there is a model better than the "null model" (intercept only model), and state whether there is evidence of over dispersion in these data.

2. As part of a larger study on the effects of various chemicals on the germination of seeds under various temperature regimes, four different concentrations of a chemical were used for treating seeds stored in four temperature regimes. For each of the 16 combinations, four replicate dishes of 50 seeds were stored and the germination of all 64 sets of seeds tested under standard conditions. The numbers of seeds (out of 50) germinating in each dish were observed.

   (a) If seeds are assumed to germinate independently within each set of 50, and with no differences in germination rates among dishes treated identically, then a binomial distribution plus a logit link might be expected to be an appropriate model. Write out this model in the context of this problem and make sure to define all terms in the model. Assume that there could be interactions in the independent variables. (Note, there is no blocking of the four replicates.)

   (b) A sequence of models is fitted to assess the relative importance of the concentration and temperature regime effects and their interactions. The results are as follows:

```
Model  Model Fit                          Deviance         df
-----  --------------------------------   -------------    --
 1     Overall mean                       1193.8           63
 2     Mean + Temp                         430.1           60
 3     Mean + Conc                         980.1            ?
 4     Mean + Conc + Temp                  148.1            ?
 5     Mean + Conc + Temp + Conc x Temp     55.6            ?
-------------------------------------------------------------
```

   From these results decide whether the interaction term is needed in the model. Assume that the dispersion parameter is equal to 1.0 when you evaluate these models.

   (c) Based on the model you select above, is there evidence of lack of fit?

   (d) Consider the multiple logistic regression model with a linear predictor:

$$\mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

   Derive an expression for the odds ratio for $X_1$. Does $\exp(\beta_1)$ have a different meaning here compared to a model containing no interaction term? Explain.

   (e) Is there a difference in how you would interpret a coefficient for the "Conc" variable in model 4 and 5? Explain.

3. Consider a dataset that has information about the gas mileage (in miles per gallon, or mpg) of various automobiles. In particular, the dataset consists of the following attributes (columns):

   1. mpg: continuous 2. cylinders: multi-valued discrete 3. displacement: continuous 4. horsepower: continuous 5. weight: continuous 6. acceleration: continuous 7. model year: multi-valued discrete 8. origin: multi-valued discrete

   This is a famous dataset that is often used to evaluate new statistical learning methodologies. Here, we will use logistic regression and try to predict whether a car gets high or low gas mileage. There are two datasets on the Blackboard website, a training data set `auto_mpg_data2_train.dat` (with 342 observations) and a test data set `auto_mpg_data2_test.dat` (with 50 observations).

   (a) Create a binary variable (say, `mpg01`) from the mpg column that contains a value above its median, and a 0 if it is below the median. Explore the data graphically in order to investigate the relationship between `mpg01` and the other variables – particularly, which ones are likely to help in predicting this variable. Scatterplots and box plots may be useful tools to answer this question. Describe your findings.

   (b) Use the training set to build the best logistic regression model that you can to predict `mpg01`. Be sure to justify your model and give the confusion matrix for the training data and evaluate model assumptions. Note, you may not need all of the variables, or you may need to consider interactions, etc. This is part of the model building process.

   (c) Now, use the model that you fit on the training data to predict the test dataset. Note, you are NOT allowed to refit the model to these data set - they are to be used to evaluate your model only. You can check out the SAS help to figure out how to save the model details from the training data (hint: I believe the `outmodel` statement will do this in `proc logistic`) and then evaluate the new data by reading in the training model (hint: use the `inmodel` statement) and scoring the test data set (hint: use the `score` command). Report your results via a confusion matrix and discuss the overall quality of your prediction. Include the relevant SAS code commands for these steps in your answer. [Note: when you make the new `mpg01` variable for the test dataset, you MUST use the medians from the training dataset!]

4. Biologists are interested in studying endangered fish species on the Missouri river. In particular, the Pallid Sturgeon is listed as an endangered species in the Missouri river watershed. The Pallid Sturgeon is a "benthic" fish, meaning that it lives on the bottom of the river. In order to better understand this species, and to potentially catch them for measurement, a study was conducted on various benthic fish, based on their habitat conditions (i.e., "macro habitat") and based on what gear was used to catch them. The data in the file `benthicfish.dat` includes 174 samples of a benthic fish species. The first column of the dataset gives the counts of the fish in the sample, the second column gives a code for macrohabitat (1-4), and the third column is a code for the gear type (1-5).

   (a) Consider a Poisson regression model with fish count as the response and macrohabitat type as the covariates (note: make these indicator variables). Which, if any, macrohabitats are significant? Is there evidence of over or under dispersion? (Ignore the gear type in this analysis).

   (b) Note that if you examine a histogram of the counts, there are many more zeros than one would expect from a Poisson distribution. This is not surprising since one could observe a zero for two different reasons: (1) this fish species was not present at the location at the time of the sample, or (2) the fish was present but was not caught by the gear used in that sample. It thus makes sense to try to account for both types of zeros. One way to do this is by using what is called a "zero-inflated Poisson" model. Such a model has a density function that is a mixture of a point mass at zero (with a specific probability) and a Poisson distribution (with probability 1 minus the probability of being a zero). Proc GENMOD can accommodate such models (see the user manual). We are interested in modeling the zero inflation probability as a function of gear type. The easiest way to do this is to assume that the probability of a zero is transformed by a logistic function. As an example, see a similar analysis at `http://www.ats.ucla.edu/stat/sas/dae/zipreg.htm`. Thus, rerun the fish count data as a zero-inflated Poisson (ZIP) model with a log link in the Poisson portion of the model (with covariates the macrohabitat indicator variables) and a logistic link in the zero model portion, with gear type indicator variables as the covariates in the zero model. Interpret the output from this model and compare to the pure Poisson model from part (a).