Categorical Data Analysis

(STAT343)

Assignment 3

Due April 16, 2019

1. The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. Let $\mu_i = E(Y_i)$ for a response variable Y_i. Consider the following functions to define GLMs:

   $g(\mu_i) = \mu_i$

   $g(\mu_i) = |\mu_i|$

   $g(\mu_i) = \mu_i^2$

   $g(\mu_i) = \log(\mu_i)$

   $g(\mu_i) = \log(\mu_i/(1-\mu_i))$

   $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ^{-1} is the inverse of CDF for the standard normal distribution.

   $g(\mu_i) = \log(-\log(1-\mu_i))$

   $g(\mu_i) = \log(-\log(\mu_i))$

   (a) Choose all the link functions for normal data. Justify your choices.

   (b) Choose all the link functions for binary data. Justify your choices.

   (c) Choose all the link functions for count data. Justify your choices.

2. (GLM for count data, including overdispersion parameter) Consider the horseshoe crab data with width (W) as a predictor and the number of satellites a response. You can use SAS and/or R to answer the following questions.

   (a) Fit a Poisson regression model to the data and test the significance of the regression coefficients without and with the dispersion parameter.

(b) Fit a negative binomial (NB) regression model to the data and test the significance of the regression coefficients.

3. (GLM for count data, including an offset variable and overdispersion parameter) Consider the collision data involving trains in Great Britain. The rates, the number of train collisions divided by log(train-km), can be investigated by the number of years since 1975. You can use SAS and/or R to answer the following questions.

   (a) Fit a Poisson regression model to the data with an offset variable, log(train-km). Test the significance of the regression coefficients without and with the dispersion parameter and interpret.

   (b) Fit a negative binomial (NB) regression model to the data with an offset variable, log(train-km). Test the significance of the regression coefficients and interpret.