AMS 274 – Generalized Linear Models (Fall 2016)

Homework 2 (due Thursday October 27)

1. Let y_i be realizations of independent random variables Y_i with Poisson(μ_i) distributions, where $\mathrm{E}(Y_i) = \mu_i$, for $i = 1, ..., n$.
   (a) Obtain the expression for the deviance for comparison of the full model, which assumes a different μ_i for each y_i, with a reduced model defined by a Poisson GLM with link function $g(\cdot)$. That is, under the reduced model, $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ (with $p < n$) is the vector of regression coefficients corresponding to covariates $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$.
   (b) Show that the expression for the deviance simplifies to $2 \sum_{i=1}^n y_i \log(y_i/\hat{\mu}_i)$, for the special case of the reduced model in part (a) with $g(\mu_i) = \log(\mu_i)$, and linear predictor that includes an intercept, that is, $\eta_i = \beta_1 + \sum_{j=2}^p x_{ij}\beta_j$, for $i = 1, ..., n$.

2. Let y_i, $i = 1, ..., n$, be realizations of independent random variables Y_i following gamma(μ_i, ν) distributions, with densities given by

$$f(y_i \mid \mu_i, \nu) = \frac{(\nu/\mu_i)^\nu y_i^{\nu-1} \exp(-\nu y_i/\mu_i)}{\Gamma(\nu)}, \qquad y_i > 0;\ \ \nu > 0,\ \mu_i > 0,$$

   where $\Gamma(\nu) = \int_0^\infty t^{\nu-1} \exp(-t) \mathrm{d}t$ is the Gamma function.
   (a) Express the gamma distribution as a member of the exponential dispersion family.
   (b) Obtain the scaled deviance and the deviance for the comparison of the full model, which includes a different μ_i for each y_i, with a gamma GLM based on link function $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ $(p < n)$ is the vector of regression coefficients corresponding to a set of p covariates.

3. Consider the data set from:
   `http://www.stat.columbia.edu/~gelman/book/data/fabric.asc`
   on the incidence of faults in the manufacturing of rolls of fabric. The first column contains the length of each roll (the covariate with values x_i), and the second contains the number of faults (the response with means μ_i).
   (a) Use R to fit a Poisson GLM, with logarithmic link,

$$\log(\mu_i) = \beta_1 + \beta_2 x_i \tag{1}$$

   to explain the number of faults in terms of length of roll.
   (b) Fit the regression model for the response means in (1) using the quasi-likelihood estimation method, which allows for a dispersion parameter in the response variance function. (Use the quasipoisson "family" in R.) Discuss the results.
   (c) Derive point estimates and asymptotic interval estimates for the linear predictor, $\eta_0 = \beta_1 + \beta_2 x_0$, at a new value x_0 for length of roll, under the standard (likelihood) estimation method from part (a), and the quasi-likelihood estimation method from part (b). Evaluate the point and interval estimates at $x_0 = 500$ and $x_0 = 995$. (Under both cases, use the asymptotic bivariate normality of $(\hat{\beta}_1, \hat{\beta}_2)$ to obtain the asymptotic distribution of $\hat{\eta}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0$.)

4. This problem deals with data collected as the number of Ceriodaphnia organisms counted in a controlled environment in which reproduction is occurring among the organisms. The experimenter places into the containers a varying concentration of a particular component of jet fuel that impairs reproduction. It is anticipated that as the concentration of jet fuel grows, the number of organisms should decrease. The problem also includes a categorical covariate introduced through use of two different strains of the organism.

   The data set is available from the course website
   `https://ams274-fall16-01.courses.soe.ucsc.edu/node/4`
   where the first column includes the number of organisms, the second the concentration of jet fuel (in grams per liter), and the third the strain of the organism (with covariate values 0 and 1).

   Build a Poisson GLM to study the effect of the covariates (jet fuel concentration and organism strain) on the number of Ceriodaphnia organisms. Use graphical exploratory data analysis to motivate possible choices for the link function and the linear predictor. Use classical measures of goodness-of-fit and model comparison (deviance, AIC and BIC), as well as Pearson and deviance residuals, to assess model fit and to compare different model formulations. Provide a plot of the estimated regression functions under your proposed model.