Bio 754: Homework 2

April 16, 2013

Due April 30th at 10am by email.

1. **[GLM Basics]**

   (a) Define the following components of a GLM for a member of the exponential family: link function, canonical link, natural parameter, linear predictor, iteratively re-weighted least squares.

   (b) What are the advantages of using the canonical link function? Describe a case where you would not use the canonical link.

   (c) In general terms, describe the type of data for which each of the following families would be sensible: Gaussian, Poisson, Gamma, Inverse-Gaussian, Binomial.

   (d) Suppose we have a sample $Y_i|\theta \sim_{i.i.d.} p(\cdot|\theta,\phi), i = 1,\ldots,n$ from the following distributions:

   - Poisson: $p(y|\theta,\phi) = \frac{e^{-\lambda}\lambda^y}{y!}$ for $y = 0,1,2,\ldots$
   - Gamma: $p(y|\theta,\phi) = \frac{b^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-by}$ for $y > 0$
   - Inverse Gaussian: $p(y|\theta,\phi) = \left(\frac{\delta}{2\pi y^3}\right)^{1/2} \exp\left[\frac{-\delta(y-\lambda)^2}{2\lambda^2 y}\right]$ for $y > 0$

   Show that each of these distributions is a member of the exponential family and identify θ, ϕ, $b(\theta)$, $a(\phi)$, and $c(y,\phi)$.

   (e) Identify $\mathbb{E}[Y|\theta,\phi]$ and $\mathrm{Var}(Y|\theta,\phi)$.

   (f) Determine the canonical link function for each distribution

   (g) The Altham et al. 1991 (http://biostat.jhsph.edu/ jleek/teaching/2011/754/data/altham.txt) data are T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies. The question of interest is: is there any difference in the distribution of cell counts between the two diseases? Using the R function `glm()` fit the above models to the cell count data assuming the canonical link with $g(\mu_i) = x_i\beta$ where $x_i = [10]$ for $i = 1,\ldots,n = 20$ and $x_i = [11]$ for $i = n+1,\ldots,2n = 40$ and $\beta' = (\beta_0,\beta_1)$.

(h) The question of interest here is whether the means of the two groups are equal? Express this question in terms of β_0 and β_1. For what transformation of β is this question answered on the scale of the original data?

(i) Using the asymptotic distribution of the MLE, that is $\hat{\beta} \dot{\sim} N(\beta, I(\hat{\beta})^{-1})$ give 90% confidence intervals for each parameter. Under each of the distributional assumptions, would you conclude that the means of the two groups are equal?

2. [**Interpreting coefficients**] Suppose you have observed an outcome Y_i and covariates X_i, Z_i (all univariate) and fit the models

$$\begin{aligned} \text{logit}(\mathbb{E}[Y_i]) &= \beta_0 + \beta_1 X_i + \gamma_2 Z_i & (1) \\ \text{logit}(\mathbb{E}[Y_i]) &= \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 Z_i X_i & (2) \end{aligned}$$

What are the interpretations of the coefficients for each of these models, in language suitable for a non-statistical audience, under both the parametric and non-parametric assumptions? What transformation of the coefficients may be more interpretable to a non-statistical audience? How do you interpret the parameters after this transformation?
Hint: It may be helpful to make up plausible hypothetical X and Y - see problem 1(c).

3. [**Kaggle contest**]

Participate in one of the following Kaggle contests:

(a) Yelp recruiting - http://www.kaggle.com/c/yelp-recruiting

(b) Facial expression recognition - http://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge

(c) Handwriting stroke recovery (due date soon!) - http://www.kaggle.com/c/icdar2013-stroke-recovery-from-offline-data

You must turn in a report (please keep to ¡ 5 pages and ¡ 3 figures) plus supplementary code explaining what you did. You are also required to submit at least one solution to the actual competition.