

Bio 753: Homework 1

January 28, 2014

See the class site for due date, guidance on how to answer, and notice of any corrections or clarifications. For problems which require computation, please append neat and commented R code as an appendix to your homework.

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 $Y_i|\theta \sim_{i.i.d.} p(\cdot|\theta, \phi), i = 1, \dots, n$ from the following distributions:

- Poisson: $p(y|\theta, \phi) = \frac{e^{-\lambda} \lambda^y}{y!}$ for $y = 0, 1, 2, \dots$
- 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 $\text{Var}(Y|\theta, \phi)$.
- (f) Determine the canonical link function for each distribution
- (g) The Altham et al. 1991 data available from <http://biostat.jhsph.edu/~jleek/teaching/2011/754/data/altham.txt> are T_4 cells/mm³ 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 $i = 1, \dots, n = 20$ and $x_i = [11]$ for $i = n + 1, \dots, 2n = 40$ 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 $\hat{\beta} \sim N(\beta, I(\hat{\beta})^{-1})$ 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

$$\text{logit}(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 X_i + \gamma_2 Z_i \quad (1)$$

$$\text{logit}(\mathbb{E}[Y_i]) = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \gamma_3 Z_i X_i \quad (2)$$

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. **[Earthquakes!]** Download earthquake data for the last 30 days from Data.gov (<http://earthquake.usgs.gov/earthquakes/feed/v1.0/csv.php>). The question of interest is to quantify the association (if any) between earthquake magnitude and earthquake depth. Fit a GLM with appropriate choice for distribution and link to these data. All model fitting must be performed “by hand”. You must write your own code to maximize the likelihood with neat (read: readable, reproducible) code and derivations included as supplementary information. Use the `glm` and `sandwich` functions to confirm your code works.