Biometry 726
Fall 2010
Homework 3
Due Tuesday 2nd November 2010

1. Let Y_i be distributed as a Poisson(λ_i) random variable for $i = 1, \ldots, n$ where $\lambda_i > 0$. Therefore

$$\Pr(Y_i = j) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!},$$

for $j = 0, 1, 2, \ldots$.

   (a) Show Y_i is a member of a one-parameter exponential family.
   (b) What is the canonical parameter, θ_i?
   (c) What is the scale parameter, ψ?
   (d) What is $c(Y_i, \psi)$?
   (e) What is $b(\theta_i)$?
   (f) Derive the mean of Y_i based on Part 1e.
   (g) Derive the variance of Y_i based on Part 1e.
   (h) What is the canonical link function?

2. Repeat question 1 assuming Y_1 has an exponential distribution; that is,

$$f(Y_i) = \lambda_i \exp(-\lambda_i Y_i)$$

for $Y_i > 0$.

3. The iteratively re-weighted least squares algorithm can be described by the following sequence of steps.

   (a) Given current estimates of $\boldsymbol{\beta}$, estimate the linear predictor, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
   (b) Given the current estimate of $\boldsymbol{\eta}$, estimate the mean using the fact that $\mathbf{g}^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$ (i.e. $g(\mu_i)$ is the link function).
   (c) Given the current estimate of $\boldsymbol{\mu}$, estimate $\mathbf{Z} = \mathbf{g}(\boldsymbol{\mu}) + (\mathbf{Y} - \boldsymbol{\mu})\mathbf{g}'(\boldsymbol{\mu})$.
   (d) Given the current estimate of $\boldsymbol{\mu}$, estimate $\mathbf{W} = \text{diag}((g'(\mu_i))^2 V(\mu_i))$.
   (e) Given the current estimates of $\mathbf{Z}$ and $\mathbf{W}$, estimate $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Z}$.
   (f) Repeat steps 3a through 3e to convergence.

   Specify the components of each of these steps for logistic regression assuming use of the canonical link. You can assume $Y_1, Y_2, \ldots, Y_n$ are mutually independent and distributed Binomial$(1, p_i)$ = Bernoulli(p_i).

4. A toy data set is available on the class website. The data contains 40 binary observations, 20 from group 1 (coded 1) and 20 from group 2 (coded 0). Assume interest lies in estimating the probability of an event (i.e. an outcome of 1) given group assignment. Using the steps you've detailed in question 3, write your own logistic regression code to work through one iteration of IRWLS for the appropriate model applied to this toy data to estimate an intercept and a group effect. For the intercept, use a starting value of the

overall proportion of events. For the group effect, use a starting value of 0. What are the estimates for the intercept and group effect after one iteration? Check your results against the estimated intercept and group effect obtained using the `glm` function in `R`. They should be extremely close, even after one iteration. Use both your estimates after one iteration and the estimates obtained from `glm` to construct estimates of the probability of an event for subjects in group 1 and subjects in group 2.

Bonus: Write your own logistic regression function that fully iterates to convergence.