BS1a: Applied Statistics  Problem Sheet 5  Michaelmas Term 2010

(1) Consider a GLM with link function equal the canonical link. Show that the score equations \( \partial \ell / \partial \beta = 0 \) (notation as lectures) reduce to
\[
X^T y = X^T \mu
\]
with \( \mu_i = \mathbb{E}(Y_i) \) a function of \( \beta \) and \( \mu = (\mu_1, ..., \mu_n) \). Calculate the iteration equations for iteratively reweighted least squares in the canonical case and show that they have a fixed point at \( X^T y = X^T \mu \). Explain why the method of scoring, and Newton-Raphson, are identical in the canonical case. Show that the iteration (and \( \hat{\beta} \)) does not depend on the dispersion \( \phi \).

(2) Consider a Binomial GLM with response \( Y_i \sim \text{Binomial}(m_i, \mu_i/m_i) \) \( i = 1, 2, ..., n \), linear predictor \( \eta_i = x_i \beta \), \( \beta = (\beta_1, ..., \beta_p) \) and logistic link \( \log(\mu_i/(m_i - \mu_i)) = \eta_i \). Calculate the variance function \( V(\mu) \), the expected information, and the IRLS equations.

(3) The data below are the babyfood data described in JJ Faraway (06) Chapter 2. For each gender and feed type a certain total number \( (m_i \text{disease}[i] + \text{healthy}[i], i = 1, 2, ..., 6) \) of children were followed for their first year of life, and the number developing some form of respiratory disease recorded.
\[
\begin{align*}
&\text{bf<-data.frame(disease=c(77,19,47,48,16,31),} \\
&\text{healthy=c(381,128,447,336,111,433),} \\
&\text{gender=c('M','M','M','F','F','F'),} \\
&\text{feed=c('Bottle','Suppl','Breast','Bottle','Suppl','Breast'))} \\
&> \text{sum}\text{mary(glm(cbind(disease,healthy)~gender+feed,} \\
&\text{family=binomial,data=bf))} \\
&\text{...} \\
&\text{Coefficients:} \\
&\text{(Intercept) -1.9253 0.1286 -14.971 < 2e-16 ***} \\
&\text{genderM 0.3126 0.1410 2.216 0.0267 *} \\
&\text{feedBreast -0.6693 0.1530 -4.374 1.22e-05 ***} \\
&\text{feedSuppl -0.1725 0.2056 -0.839 0.4013} \\
&\text{...} \\
&\text{Null deviance: 26.37529 on 5 degrees of freedom} \\
&\text{Residual deviance: 0.72192 on 2 degrees of freedom} \\
&\text{(a) Calculate the odds for disease in male and female infants for each feed type. Give a 95\% CI for the log-odds ratio for disease in breast against bottle fed infants and test for a difference in odds at 5\%.} \\
&\text{(b) Write down the linear predictor for the null model and carry out a test for any explanatory relation between the response (disease, healthy) and the variables gender and feed.} 
\end{align*}
\]
(c) The residual deviance for the model with linear predictor $1 + \text{gender}$ is 20.899. Test for an effect due to \textit{feed}.

(d) Show that the model with an interaction between feed and gender (linear predictor $\text{feed*gender}=1+\text{feed+gender+feed:gender}$) is just the saturated model, and give the residual deviance for this model (!). Test for an interaction between feed and gender at 5%.

(e) Using your result in Question (2), adapt the IRLS R-code given in lectures to compute the \textit{Estimate} and \textit{Std. Error} columns of the R-output above. [Some hints: use $X<-\text{model.matrix(cbind(disease,healthy)~gender+feed}$ to generate the design matrix, and $\hat{\mu}<-\text{diag(c(mu*(m-mu)/m))}$ with $m<-\text{disease+healthy}$ to set up the weight matrix]