MATH 523 - ASSIGNMENT 3

To be handed in not later than 5pm, Tuesday 27th March 2012.
Please hand in during lectures, to Burnside 1225, or to the Mathematics Office Burnside 1005

1 The data set contained in the file Food.csv represents the results of a study where a cohort of 190 male and 199 female subjects were asked to recall how many times within the last month they had consumed a particular food product. The data are formatted into three columns

- **Column 1 - FoodCount:** This lists the numbers of times that the food product had been eaten, as reported by the cohort members.
- **Column 2 - Female:** This records the numbers of times a particular value was reported by female subjects.
- **Column 3 - Male:** This records the numbers of times a particular value was reported by male subjects.

For example, eleven females and two males reported having eaten the food product ten times.

(a) Using an appropriate statistical model, test for a difference in the expected (per month) consumption counts of this food product in males and females, and construct an approximate 95 % confidence interval for the difference in expected counts between males and females.

8 MARKS

(b) Consider a model for count data fitted using a Poisson GLM, using first a **log** link, and secondly an **identity** link function; consider in particular the parameter estimates and standard errors under the two assumptions. Are the estimates and standard errors identical under the two link functions ? Justify your answers.

4 MARKS

2 Consider the *logistic regression model* for individual level data and a single continuous covariate X; we have that $Y_i \sim Bernoulli(\pi_i)$, where

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

By means of a simulation study, demonstrate that, in large samples,

(a) the maximum likelihood (ML) estimator of β_1 is approximately normally distributed.

4 MARKS

(b) the distribution of the deviance statistic depends, in general, on the true values of the parameters.

4 MARKS

For a fixed large n, and a range of parameter values, perform analyses on replicate data sets and extract the ML estimates and deviance statistics; then study the empirical distributions of these quantities. Note that the `R` functions `coef` and `deviance` extract the relevant quantities.

Pseudocode for carrying out such a simulation study is given on page 2.

```
#Define a range of beta1 values to be studied
be<-seq(lower.beta,upper.beta,by=incr.beta)

Define a beta0 value, be0

Define a sample size, n

Define a number of replications, nreps

Define a matrix to store the results

#Loop across the beta1 values
for(ibe in 1:length(be)){

        #Extract the beta1 to be studied on this loop
        be1<-be[ibe]

    #Loop across the replications
    for(irep in 1:nreps){

                Generate some covariate values, X

                Compute the linear predictor, eta

                Define the Bernoulli probability, pv

                #Generate the response data, Y
                Y<-rbinom(n,1,pv)

                #Extract the estimated coefficient, beta1.hat
                beta1.hat<-coef(glm(Y~X,family=binomial))[2]

                #Extract the deviance
                dev.stat<-deviance(glm(Y~X,family=binomial))


                Store the computed values

        }
}

Summarize the statistics using a histogram, boxplot etc.
```