Qualifying Exam: CAS MA 576.

Boston University, Spring 2010

Question 3

Let Y_i, $i = 1, \ldots, n$ be a r.v. giving us the number of successes in a set of m_i trials with probability of success π_i, i.e. $Y_i \sim Binomial(m_i, \pi_i)$. Moreover, let the success probability π_i depend on the the p covariates of the ith covariate class $\boldsymbol{x_i}$. We seek to model the the above relation in a GLM framework.

(a) Assuming the m_i's to be known show that the logit link is the canonical link for the above model.

(b) Name and write the expressions for two other link functions by which we can provide a realistic model for π_i as a function of the covariates $\boldsymbol{x_i}$.

(c) Write the likelihood of this model (for arbitrary link functions) and show that the residual deviance can be written as

$$D(y, \hat{\pi}) = 2 \sum_{i=1}^{n} \left(y_i \log(y_i / m_i \hat{\pi}_i) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i(1 - \hat{\pi}_i)} \right) \right),$$

where $\hat{\pi}_i$'s are the MLE's of π_i's.

(d) Show that the the Pearson Chi-square statistic for Y_i is

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i \hat{p}_i)^2}{(m_i \hat{p}_i (1 - \hat{p}_i))}$$

(e) Use the first two terms of the Taylor Series Expansion of $s \ \log(s/t)$ at $s = t$ given by

$$s \ \log(s/t) = (s - t) + \frac{1}{2t}(s - t)^2 + \ldots$$

to show that the Pearson Chi-square statistic and deviance statistics are asymptotically equivalent.

Question 4

A groups of 20 snails were held for periods of 1, 2, 3 or 4 weeks in carefully controlled conditions of temperature and relative humidity. There were two species of snail, A and B, and the experiment was designed as a 4 by 3 by 4 by 2 completely randomized design. At the end of the exposure time the snails were tested to see if they had survived; the process itself is fatal for the animals. The object of the exercise was to model the probability of survival in terms of the stimulus variables, and in particular to test for differences between species. The data frame contains the following variables:

```
Species     Snail species A (1) or B (2)
Exposure    Exposure in weeks
Rel.Hum     Relative humidity (4 levels)
Temp        Temperature, in degrees Celsius (3 levels)
Deaths      Number of deaths
N           Number of snails exposed
```

Here is a subset of the data

```
   Species Exposure Rel.Hum Temp Deaths  N
1        A        1    60.0   10      0 20
2        A        1    60.0   15      0 20
3        A        1    60.0   20      0 20
4        A        1    65.8   10      0 20
5        A        1    65.8   15      1 20
6        A        1    65.8   20      2 20
.        .        .      .    .      . .
.        .        .      .    .      . .
93       B        4    70.5   20      9 20
94       B        4    75.8   10      4 20
95       B        4    75.8   15      5 20
96       B        4    75.8   20      7 20
```

First we fit an additive model (model 1) to check Species difference. Note that we consider the explanatory variable `Exposure, Rel.Hum, Temp` to be continuous.

```
Call:
glm(formula = cbind(Deaths, N - Deaths) ~ Species + Exposure +
    Rel.Hum + Temp, family = "binomial", data = snails)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40495    0.97070  -1.447    0.148
SpeciesB     1.30864    0.16350   8.004 1.20e-15 ***
Exposure     1.50339    0.10235  14.689  < 2e-16 ***
Rel.Hum     -0.10684    0.01388  -7.699 1.37e-14 ***
Temp         0.09404    0.01927   4.881 1.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 539.72  on 95  degrees of freedom
Residual deviance:  55.07  on 91  degrees of freedom
AIC: 223.93

Number of Fisher Scoring iterations: 5
```

(a) Is there enough evidence of difference in Survival rates of the two species? Justify your answer

(b) Using the above model find the odds ratio for Survival for Species A vs Species B.

(c) Provide a 95% confidence interval for the above odds ratio.

(d) Does the CI include the value 1? Can you draw any inference on species difference based on the the confidence interval.

(e) What is the increase in odds of Survival (multiplicatively), for a unit change in `Exposure` rate.

(f) Is their any indication of overdispersion? Justify your answer.

The above model assumes that the effect of covariates `Exposure, Rel.Hum, Temp` is the same for the two species. A natural way to enrich the above model is to fit a partial interaction model (model 2) assuming different covariate effects for the two species using the following model

```
Call:  glm(formula = cbind(Deaths, N - Deaths) ~
            Species * (Exposure +      Rel.Hum + Temp),
                    family = "binomial", data = snails)

Coefficients:
      (Intercept)           SpeciesB           Exposure            Rel.Hum
        -2.504682           3.055430           1.463156          -0.088566
             Temp  SpeciesB:Exposure   SpeciesB:Rel.Hum      SpeciesB:Temp
         0.096270           0.068108          -0.029156          -0.002968

Degrees of Freedom: 95 Total (i.e. Null);  88 Residual
Null Deviance:      539.7
Residual Deviance: 53.99        AIC: 228.9
```

(g) Which model 1 or 2, would you prefer based solely on the AIC values?

Now we conduct an anova analysis for the two models.

```
Analysis of Deviance Table

Model 1: cbind(Deaths, N - Deaths) ~ Species + Exposure + Rel.Hum + Temp
Model 2: cbind(Deaths, N - Deaths) ~ Species * (Exposure + Rel.Hum + Temp)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        91     55.070
2        88     53.989  3    1.081     0.782
```

(h) Justify why we can use the Chi Square test for comparing the above set of models

(i) Which model would you prefer based on the above test and why?

Now suppose in a follow up experiment the number of snails studied under a different set of conditions change from 20 to 200, whereas the number of deaths remain same as before i.e. the dataset looks like

	Species	Exposure	Rel.Hum	Temp	Deaths	N
1	A	1	60.0	10	0	200
2	A	1	60.0	15	0	200
3	A	1	60.0	20	0	200
4	A	1	65.8	10	0	200
5	A	1	65.8	15	1	200
6	A	1	65.8	20	2	200
.	.	.	.	.	.	.
.	.	.	.	.	.	.
93	B	4	70.5	20	9	200
94	B	4	75.8	10	4	200
95	B	4	75.8	15	5	200
96	B	4	75.8	20	7	200

(j) Would you have used the same binomial glm to model the relationship of death with the covariates. If yes: defend your answer. If not: provide alternative models for modeling the variable death with the covariates. Write the R-code to implement the calculation for your proposed model.