MATH382301

This question paper consists of 5 printed pages, each of which is identified by the reference **MATH3823**.

Only approved basic scientific calculators may be used.

Examination for the Module MATH3823
(May / June 2006)

GENERALIZED LINEAR MODELS

Time allowed: **2 hours**

Attempt not more than THREE questions.
All questions carry equal marks.

1. (a) Let $Y_1, \ldots, Y_n$ be a sample from a one-parameter exponential family distribution with density $f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}$. Show that the maximum likelihood estimate $\hat{\theta}$ of θ is given by $b'(\hat{\theta}) = \sum_i y_i/n$.

(b) Now assume that $Y \sim Exp(\lambda)$, $\lambda > 0$ so that

$$f_Y(y;\lambda) = \begin{cases} \lambda e^{-\lambda y} & y \geqslant 0 \\ 0 & y < 0. \end{cases}$$

Find the exponential family form of Y and use this form to find the expectation and variance of Y. Use the result from part (a) to show that the maximum likelihood estimate of λ given a sample $Y_1, \ldots, Y_n$ is $\hat{\lambda} = n/\sum_i Y_i$.

(c) In a test to determine the average lifetime of lightbulbs, 25 lightbulbs were left on until they failed. The lifetimes in thousands of hours is recorded below. Assuming that an exponential distribution is appropriate to model these data, use the result from part (b) to find the value of $\hat{\lambda}$ for these data.

5.50	3.90	3.75	5.98	4.02
0.11	2.43	2.00	13.95	4.35
4.51	4.60	1.27	6.61	6.88
1.09	4.59	16.79	0.05	9.97
5.52	3.49	2.23	3.25	2.49

A sample of 25 lightbulbs of an improved design were also tested. Explain how a generalized linear model with a log link function could be constructed to model the lifetimes of the two samples of lightbulbs and give the design matrix X for this model. Write down an equation for λ in terms of regression parameters β_1 and β_2.

The average lifetime of the bulbs in the second sample was 5349.2 hours. Find the regression parameter estimates for your generalized linear model.

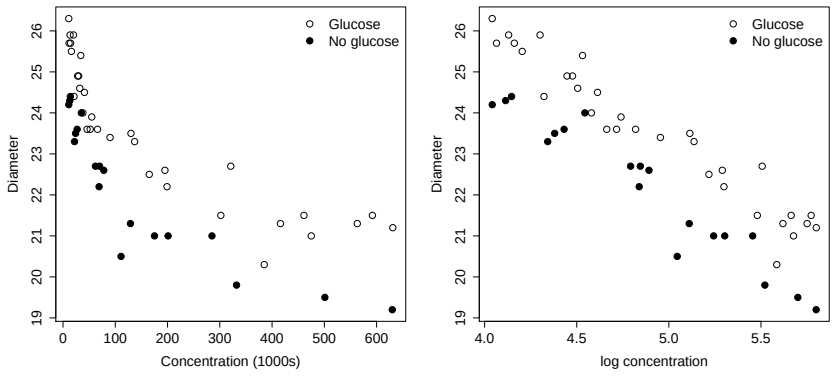
2. The growth of *Tetrahymena* cells was thought to depend on two factors; the initial concentration of cells in a sample and the presence or absence of glucose. Thirty two samples were prepared with glucose and 19 without. For each sample, the initial concentration of cells was measured and the sample left to grow in controlled conditions. Once the cells were mature, the average cell diameter was measured.

(a) Let y be the average cell diameter, x_1 be a factor taking value 1 if the sample contained glucose and 2 if the sample contained no glucose, and x_2 be the initial cell concentration. Assume that samples $i = 1, \ldots, 32$ are those with glucose present.

Construct the design matrices corresponding to the models $y \sim x_1 + x_2$ and $y \sim x_1 * x_2$.

(b) The plots below show y against x_2 (left plot) and y against $\log_{10} x_2$ (right plot). Explain why in this case it is appropriate to use $x_3 = \log_{10} x_2$ rather than x_2 as an explanatory variable. Comment on any features of the data that are apparent from these plots.

Interpret the models $y \sim 1$, $y \sim x_1$, $y \sim x_3$, $y \sim x_1 + x_3$, and $y \sim x_1 * x_3$ in terms of lines on the plot of y against x_3.



(c) Fitting the various models from part (b) in R gave the following deviances.

Model number	Model	Deviance
1	$y \sim 1$	165.72000
2	$y \sim x_1$	143.70612
3	$y \sim x_3$	36.18442
4	$y \sim x_1 + x_3$	10.15089
5	$y \sim x_1 * x_3$	10.07332

For each model, write down the number of free parameters r.

Use appropriate hypothesis tests to determine which model is most appropriate. The following R code gives upper 5% points of various F distributions which may be of use.

```
> qf(0.05, df1 = 1, df2 = c(1, 5, 46, 51), lower.tail=F)
[1] 161.448   6.608   4.047   4.030
```

3. (a) Let $Y^* = Y/m$ where $Y \sim Bin(m, p)$. Show that the mass function of Y^* is

$$f_{Y^*}(y^*) = \binom{m}{y} p^y (1-p)^{m-y}.$$

Hence express Y^* as an exponential family random variable.

(b) Find the deviance of a binomial generalized linear model for data $y_i \sim Bin(m_i, p_i), i = 1, \ldots, n$ with fitted parameters $\hat{p}_i$. You may use the facts that for any generalized linear model the deviance is defined to be $D = 2\phi\{l(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) - l(\hat{\boldsymbol{\theta}}|\boldsymbol{y})\}$ where $\tilde{\boldsymbol{\theta}}$ are the parameters of the saturated model and the log-likelihood of parameter vector $\boldsymbol{\theta}$ given data $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}.$$

For two candidate models M_1 and M_2 explain when and how the deviances can be used to test which model provides a better description of the data.

(c) A study was conducted to investigate how the incidence of coronary heard disease (chd) in $n = 462$ patients was associated with age, obesity, and previous family history of heart disease (famhist). Age and obesity were recorded as quantitative variables and famhist as a two-level factor.

The following R code gives the deviances of a sequence of models fitted to these data. Use the deviances to determine which is the most appropriate model and comment on your result. You may find the percentage points given at the end of the R output helpful in conducting the necessary hypothesis tests.

```
> deviance(glm(chd ~ age*famhist*obesity, family=binomial))
[1] 505.5772
> deviance(glm(chd ~ age+famhist+obesity, family=binomial))
[1] 506.6312
> deviance(glm(chd ~ age+famhist, family=binomial))
[1] 506.6582
> deviance(glm(chd ~ age+obesity, family=binomial))
[1] 525.5529
> deviance(glm(chd ~ famhist+obesity, family=binomial))
[1] 559.5388
> deviance(glm(chd ~ age, family=binomial))
[1] 525.5623
> deviance(glm(chd ~ famhist, family=binomial))
[1] 561.8944
> deviance(glm(chd ~ obesity, family=binomial))
[1] 591.5284

> qchisq(0.05, df = c(1, 2, 3, 4, 5, 6), lower = F)
[1]  3.841  5.991  7.815  9.488 11.070 12.592
```

4. (a) A generalized linear model consists of a random component Y, a linear predictor η and a link function g. Explain how the link function connects Y to η.

Given a specific distribution for Y, explain how the canonical link function is defined.

Let $Y \sim Po(\lambda)$. Write Y in exponential family form and find the canonical link function for this choice of error distribution. What is the main advantage of the canonical link function over an identity link function in this case?

(b) Consider a two-way contingency table with count Y_{ij} in row i and column j for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Assume that $Y_{ij} \sim Po(\lambda_{ij})$ with all counts being mutually independent. Let $Y_{i\bullet} = \sum_j Y_{ij}$, $y_{i\bullet} = \sum_j y_{ij}$, and $\lambda_{i\bullet} = \sum_j \lambda_{ij}$. Show that

$$P\{Y_{ij} = y_{ij} \text{ for } j = 1, \ldots, J | Y_{i\bullet} = y_{i\bullet}\} = K(y_{i1}, \ldots, y_{iJ}) \prod_{j=1}^{J} p_{ij}^{y_{ij}},$$

giving explicit expressions for $K(y_{i1}, \ldots, y_{iJ})$ and p_{ij} and an intuitive explanation of what p_{ij} represents.

Assuming that the Poisson parameters are modelled by $\log \lambda_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, show that

$$p_{ij} = \frac{\exp\{\beta_j + (\alpha\beta)_{ij}\}}{\sum_{j'} \exp\{\beta_{j'} + (\alpha\beta)_{ij'}\}}.$$

(c) In 2000, the Australian Government Statistical Service surveyed 1170 people as to whether they would accept cuts in the standard of living to help the environment on a scale of 1: very willing to 5: very unwilling. The data are tabulated below.

Sex	Willingness					Total
	1	2	3	4	5	
Female	34	149	160	142	168	653
Male	30	131	152	98	106	517

The data were analysed in R with the following edited results.

```
> glm(count ~ sex + will, family=poisson)

Call:  glm(formula = count ~ sex + will, family=poisson)

Coefficients:
(Intercept)         sex2        will2
     3.5757      -0.2335       1.4759
      will3        will4        will5
     1.5841       1.3218       1.4542
```

Explain why it is appropriate to condition on the row totals in this case.

Assuming that this model is appropriate, explain why we can conclude that women and men have the same attitudes regarding their willingness to accept lower standard of life to help the environment.

Calculate the probabilities of people to be very willing (1) or very unwilling (5) to accept cuts in the standard of living to help the environment.

END