GR 2054

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER II (SUMMER) EXAMINATIONS, 2003/2004

VISITING STUDENT EXAMINATION

1EM1 – ERASMUS

Module Code:	**MA557**
Module:	**STATISTICAL MODELLING**
External Examiner	Dr. T. C. Bailey
Internal Examiner	Prof. J. P. Hinde

Instructions:

Duration: **Two** Hours.

Answer any *Three* of the four questions.

Relevant tables are supplied.

Question 1 is on the next page

1. The standard normal linear regression model for the response variable y and p explanatory variables is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

where the ϵ_i are independent $N(0, \sigma^2)$ random errors.

(a) i) Writing X as the $n \times (p+1)$ design matrix (*i.e.* $X^T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$), show that the maximum likelihood estimate of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \tag{1}$$

where $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$.

Show that $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ and has covariance matrix $\sigma^2 (X^T X)^{-1}$
Note: Assume throughout that X has full rank.

ii) Suppose now that the ϵ_i are independent normal but do not have constant variance. Assuming that $\mathrm{Var}\,(\epsilon_i) = v_i \sigma^2$, where the v_i are known, show that the usual least-squares estimator given in (1) is unbiased for $\boldsymbol{\beta}$ and find the covariance matrix $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right)$.

Writing V as the matrix $\mathrm{diag}\{v_1, v_2, \ldots, v_n\}$, use the fact that $V^{-1/2}\boldsymbol{\epsilon}$ has variance $\sigma^2 I_n$ to show that the true maximum likelihood estimate is given by

$$(X^T V^{-1} X)^{-1} X^T V^{-1} \mathbf{y}.$$

(b) i) Show that the hat-matrix $H = X(X^T X)^{-1} X^T$ is idempotent and symmetric, *i.e.* that

$$H\,H = H \quad \text{and} \quad H^T = H.$$

Hence, prove that the vector of fitted values, $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}$, is orthogonal to the vector of residuals

$$\boldsymbol{e} = \mathbf{y} - \widehat{\mathbf{y}}.$$

ii) Suppose that the true model is $\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, but that the model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is fitted, giving $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ with hat matrix $H = X(X^T X)^{-1} X^T$ and residuals $\boldsymbol{e} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}$. Show that

$$\boldsymbol{e} = (I - H)\,\mathbf{y} = (I - H)\,Z\boldsymbol{\gamma} + (I - H)\,\boldsymbol{\epsilon},$$

and hence that $\mathbb{E}[\boldsymbol{e}] = (I - H)\,Z\boldsymbol{\gamma}$. What happens if Z lies in the space spanned by the columns of X?

2. (a) Define the three basic components of a generalized linear model. How do these extend the class of models from the standard normal linear regression model?

(b) A random variable, Y, in the single parameter exponential family, has probability density function

$$f(y) = \exp\{y\theta - b(\theta) + c(y)\}$$

where $b()$ and $c()$ are specified functions. Show that the moment generating function for Y is given by

$$M_Y(t) = \exp\{b(\theta + t) - b(\theta)\}.$$

Hence, obtain expressions for the mean and variance of Y.

Show that for known r the negative binomial distribution with probability density function

$$f(y;p) = \binom{y+r-1}{r-1} p^r (1-p)^y, \qquad y = 0, 1, \ldots; \quad 0 \leq p \leq 1$$

is in the above exponential family and obtain expressions for the mean and variance. What is the *natural* link function for the negative binomial distribution? What is the possible problem about using this link function in modelling? Suggest an alternative link function that may be be preferable here.

(c) The following extracts of S-PLUS output are from an analysis for a study of how exposure to air pollution, and duration of exposure, might affect the response to an infection. Mice were exposed to three different dose levels (1.5, 3.5, and 7.0*ppm*) of NO_2 (`dose`) for variable amounts of time before being challenged by exposure to a bacterial infection. For each dose and time combination the number of mice tested (`n`) and the number who died (`died`) were recorded. In the analysis `time` represents the exposure time.

i) Interpret each of the models that have been fitted. Construct analysis of deviance tables and use appropriate tests to select and justify an adequate model.

ii) For the model with only main effects, test whether there is a significant difference between dose levels with 3.5 and 7.0*ppm* of NO_2. Note that you will need to find the standard error for the difference of the relevant parameter estimates.

/output...

```
Call: glm(formula = cbind(died, n - died) ~ 1, family = binomial)

Residual Deviance: 200.0901 on 16 degrees of freedom

Call: glm(formula = cbind(died, n - died) ~ log(time),
                                                family = binomial)

Residual Deviance: 165.6371 on 15 degrees of freedom

Call: glm(formula = cbind(died, n - died) ~ dose, family = binomial)

Residual Deviance: 195.1399 on 14 degrees of freedom

Call: glm(formula = cbind(died, n - died) ~ log(time) + dose,
                                                  family = binomial)
Residual Deviance: 16.01728 on 13 degrees of freedom

Coefficients:
                Value Std. Error   t value
(Intercept) -3.0385795 0.25612713 -11.86356
  log(time)  0.5499537 0.04365483  12.59777
    dose3.5  2.1199201 0.20237581  10.47517
      dose7  3.1609548 0.26889491  11.75535

Correlation of Coefficients:
          (Intercept)  log(time)    dose3.5
log(time) -0.9023909
  dose3.5 -0.9413922  0.7827892
    dose7 -0.9388475  0.8443945  0.8848344


Call: glm(formula = cbind(died, n - died) ~ log(time) * dose,
                                                  family = binomial)

Residual Deviance: 15.41356 on 11 degrees of freedom

Coefficients:
                     Value Std. Error     t value
      (Intercept) -2.94211258  0.9215924 -3.19242291
        log(time)  0.53173272  0.1728267  3.07668164
          dose3.5  2.03643008  0.9268162  2.19723177
            dose7  3.05196129  0.9264316  3.29431908
log(time)dose3.5  0.01042954  0.1789646  0.05827713
  log(time)dose7  0.15287096  0.2493551  0.61306534
```

3. Consider the following Poisson regression model for independent random variables Y_i:

$$Y_i \sim \text{Pois}(\mu_i), \quad i = 1, \ldots, n$$

$$\mu_i = \beta_0 + \beta_1 x_i = \boldsymbol{\beta}^T \mathbf{x}_i$$

Using the Fisher-scoring method, show that $\widehat{\boldsymbol{\beta}}$, the maximum likelihood estimate of $\boldsymbol{\beta}$, can be found by using the following iteratively reweighted least-squares algorithm:

$$\boldsymbol{\beta}^{(r+1)} = \left(X^T W_r X\right)^{-1} X^T W_r \mathbf{z}_r$$

where $r = 0, 1, \ldots$, refers to the cycle of iteration, $X^T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, and $W_r = \text{diag}\left(\mu_1^{(r)}, \ldots, \mu_n^{(r)}\right)$. Give a clear expression for the form of the adjusted dependent variable $\mathbf{z}_r$.

Write down an expression for an estimate of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

Show that for a model with fitted values $\widehat{\mu}_i$ the scaled deviance can be written as

$$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right\}.$$

By considering this and the Pearson χ^2 statistic, suggest two possible forms of residual for a Poisson regression model.

Often in analysing count data we may have recorded the counts over time periods of different lengths. Explain how the basic Poisson regression model is extended to allow for this, giving brief details of how such models are fitted in S-PLUS.

Note: The Poisson distribution with mean μ has probability density function

$$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!} \qquad y = 0, 1, 2, \ldots$$

4. A study of student admissions to two departments at a certain university resulted in the following table:

	Whether admitted (A)	
Gender (G)	Yes	No
Male	473	412
Female	219	399

Compute the *odds-ratio* and explain the implication of your result for evidence of sexual discrimination.

The following extracts of S-PLUS output present the results of fitting two log-linear models to this 2×2 table. Explain each of the models as specified in the `glm` calls and describe the resulting output.

Extract A

```
Call: glm(formula = count ~ gender * admitted, family = poisson)
Coefficients:

                    Value Std. Error     t value
     (Intercept)  6.1590954 0.04598005  133.951475
          gender -0.7700237 0.08173356   -9.421144
        admitted -0.1380720 0.06738953   -2.048865
gender:admitted  0.7379617 0.10776746    6.847723
(Dispersion Parameter for Poisson family taken to be 1 )
    Null Deviance: 105.0865 on 3 degrees of freedom

Residual Deviance: 0 on 0 degrees of freedom
```

Extract B

```
Call: glm(formula = count ~ gender + admitted, family = poisson)
Coefficients:

                 Value Std. Error     t value
(Intercept)  6.0099552 0.04369667  137.538064
     gender -0.3590991 0.05241682   -6.850838
   admitted  0.1586821 0.05174635    3.066537
(Dispersion Parameter for Poisson family taken to be 1 )
    Null Deviance: 105.0865 on 3 degrees of freedom

Residual Deviance: 47.97097 on 1 degrees of freedom
```

Use these results to test the significance of the association in this table and calculate an approximate 95% confidence interval for the odds-ratio.

/question continued...

Taking account of the two departments (D), the results of fitting a sequence of Poisson log-linear models to the full $2 \times 2 \times 2$ table are presented below.

Model	Deviance	df
constant	883.83	7
A	874.40	6
A + G	826.71	5
A + G + D	752.32	4
A + G + D + A.G	704.35	3
A + G + D + D.G	116.18	3
A + G + D + A.D	637.15	3
A + G + D + A.D + A.G	589.17	2
A + G + D + A.D + D.G	1.01	2
A + G + D + A.G + D.G	68.21	2
A + G + D + A.D + A.G + G.D	0.58	1

Interpret each of the models that have been fitted and recommend an appropriate model for the data, giving formal justification.

Use these model fitting results to explain why the conclusions from the 2-way table of A against G are misleading. Also, how could these results be used to obtain the test of the association between A and G in the 2-way table? Show that you obtain the same test statistic.

Which of the fitted models is not a graphical model, and why not?