UNIVERSITY OF ABERDEEN **SESSION 2003–2004**

DEGREE EXAMINATION

MX4532 Modelling 2

Friday 28 May 2004 (12noon to 2pm)

Only calculators approved by the Department of Mathematical Sciences may be used in this examination. Calculator memories must be clear at the start of the examination.

Marks may be deducted for answers that do not show clearly how the solution is reached.

Answer THREE questions. All questions carry equal weight.

1. **(a)** Suppose that in a particular regression problem the statistician assumes a model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$$

when in fact the correct model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2).$$

Find expressions for the bias and variance-covariance matrix of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ in terms of $\boldsymbol{\beta}_2$, $(\mathbf{X}^T\mathbf{X})^{-1}$ and the matrix $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}_2$.

(b) Write brief notes on **four** different diagnostic plots that one could use test the suitability of a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, and sketch plots to illustrate deviations from the linear model assumptions.

2. **(a)** Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2).$$

(i) Show that the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is equal to the estimate obtained by minimising the residual sum of squares.

(ii) Assuming that the rank of $\mathbf{X}$ is p (i.e. full rank case) derive an expression for $\hat{\boldsymbol{\beta}}$.

(iii) Calculate the mean and the variance-covariance matrix of the fitted values $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

(b) Observations on Y are made at $x = -5, -4, \ldots, 4, 5$ and it is required to fit a *continuous* relation

$$Y = \begin{cases} \alpha_1 + \beta_1 x + \epsilon & \text{for } x \leq 0 \\ \alpha_2 + \beta_2 x + \epsilon & \text{for } x \geq 0 \end{cases},$$

where the ϵ are independent and identically distributed $\mathrm{N}(0, \sigma^2)$.

(i) Show how to set up the problem in terms of a linear model with suitable dummy variables.

(ii) Describe briefly how you would test whether $\beta_1 = \beta_2$ in the model above.

3. **(a)** Write notes to describe how one would use a generalised linear model from a Poisson family with a log link to analyse contingency tables. You should include (i) a description of the difference between response and explanatory variables, (ii) the difference in interpretation of results with response and explanatory variables and (iii) possible reasons for departure from null models in your notes.

(b) Surveys were carried out in the USA in two different years, 1956 and 1971. The main question asked was concerned with how the country should be run, and the possible answers to this question were:

(1) Changes should rarely be made.

(2) We should be cautious about making changes.

(3) We should feel free to make changes.

(4) We must constantly make changes.

The respondents were also asked their political affiliations: Republicans (R), Democrats (D), or Independents (I). The results were as follows:

	1956			1971		
Response	R	D	I	R	D	I
1	4	9	1	2	11	1
2	94	211	46	85	216	87
3	74	164	28	41	166	80
4	28	47	15	31	116	74

A generalised linear model was used to model these count data with a Poisson model and a log link. Affiliation, year and response were taken as factors. The deviance and residual degrees of freedom (d.f.) for a selection of possible models are given below.

	model	deviance	d.f.
1	affiliation*year*response	0	0
2	affiliation*year + response*affiliation + year*response	8.69	6
3	affiliation*year + year*response	22.71	12
4	affiliation*response + year*response	63.32	8
5	affiliation*year + affiliation*response	39.49	9
6	affiliation + year*response	84.18	14
7	year + affiliation*response	100.96	11
8	response + year*affiliation	60.35	15
9	year*affiliation	957.77	18
10	response*affiliation	122.91	12
11	response*year	488.10	16
12	response + year*affiliation	60.35	15
13	year + affiliation + response	121.82	17
14	year + affiliation	1019.24	20
15	response + year	525.74	19
16	response + affiliation	143.77	18
17	response	547.69	20
18	year	1423.16	22
19	affiliation	1041.19	21

(i) What model would you choose to explain these data? Give reasons for your answer.

(ii) Write a brief paragraph to describe the interpretation of your chosen model to a non-statistician.

4. **(a)** Show that the binomial distribution is a member of the exponential family of distributions with probability density function of the form:

$$\exp\left[\frac{(y\theta - b(\theta))}{\phi} + c(y,\phi)\right],$$

and hence show that the canonical link for the binomial distribution is the logit function, $\log(\frac{p}{1-p})$.

(b) (i) Define the deviance for a generalised linear model, and show that the deviance for the binomial distribution can be written as

$$2\sum o_i \log(o_i/e_i),$$

where o_i is the observed data and e_i is its predicted value under the model, for $i = 1, \ldots, n$.

(ii) Show the deviance of part (i) is approximately equal to the χ^2 goodness-of-fit statistic, $\sum_i (o_i - e_i)^2/e_i$, if all the values of $(o_i - e_i)/e_i$ are small.

(c) The following R session attempts to analyse some data on whether or not patients experienced pain relief after treatment for back pain. The data are y, which has value 1 if the patient experienced relief and 0 if not, and the following covariates: tr, which is 1 if they were treated and 0 if not; age, which is the age of the patient in years; g, which has values 0 for male and 1 for female; and dur, which is the number of months for which symptoms were present before treatment started.

```
> pain <- read.table("pain.csv",sep=",",header=T)
> pain
   y tr age g dur
1  1  1  76 0  36
2  1  1  52 0  22
3  0  0  80 1  33
4  0  1  77 0  33
5  0  1  73 1  17
6  0  0  82 1  84
7  0  1  71 0  24
8  0  0  78 1  96
9  1  1  83 1  61
10 1  1  75 1  60
11 0  0  62 0   8
12 0  0  74 1  35
13 1  1  78 1   3
14 1  1  70 1  27
15 0  0  72 0  60
16 1  1  71 1   8
17 0  0  74 1   5
18 0  0  81 1  26
```

(continued on next page)

```
> model1 <- glm(y~tr+age+g+dur,family=binomial,data=pain)
> coefficients(model1)
(Intercept)          tr         age           g         dur
-17.3345396  23.7926081  -0.1149064   2.5706350   0.0534949
> model2 <- step(model1);
Start:  AIC= 19.47
 y ~ tr + age + g + dur

coefficients(model2)
       Df Deviance     AIC
- dur   1  10.3902 18.3902
- age   1  10.5782 18.5782
<none>      9.4698 19.4698
- g     1  11.7260 19.7260
- tr    1  22.5326 30.5326

Step:  AIC= 18.39
 y ~ tr + age + g

       Df Deviance    AIC
- age   1   10.952 16.952
- g     1   12.133 18.133
<none>      10.390 18.390
- tr    1   22.558 28.558

Step:  AIC= 16.95
 y ~ tr + g

       Df Deviance    AIC
- g     1   12.217 16.217
<none>      10.952 16.952
- tr    1   23.939 27.939

Step:  AIC= 16.22
 y ~ tr

       Df Deviance    AIC
<none>      12.217 16.217
- tr    1   24.057 26.057


> coefficients(model2)

(Intercept)          tr
  -19.56607    20.41337
```

(i) Briefly describe the commands used in this session.

(ii) Using model 2, estimate the probability that a 70 year old male patient, who has suffered symptoms for 12 months, will experience relief of pain if treated.