Example sheet 4. Statistical Modelling: Mathematical Tripos, Part IIC

P.M.E.Altham, Statistical Laboratory, University of Cambridge.

January 18, 2005

These questions are all adapted from recent Part IIA Tripos questions. There may be some overlap between questions, eg on basic bookwork matters.
You are not intended to do **all** these questions for 1 supervision. Be selective. Keep some questions for later revision purposes.

1998/4/13M (long question) Suppose that $Y_1, ..., Y_n$ are independent random variables, and that Y_i has probability density function

$$f(y_i|\theta_i, \phi) = exp[(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)].$$

Assume that $E(Y_i) = \mu_i$, and that there is a known link function g such that

$$g(\mu_i) = \beta^T x_i, \text{where } x_i \text{ is known and } \beta \text{ is unknown.}$$

Show that
(a) $E(Y_i) = b'(\theta_i)$,
(b) $\mathrm{var}(Y_i) = \phi b''(\theta_i) = V_i$ say, and hence
(c) if $\ell(\beta, \phi)$ is the log-likelihood function from the observations $(y_1, ..., y_n)$ then

$$\frac{\partial \ell(\beta, \phi)}{\partial \beta} = \sum_1^n \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i}.$$

Describe briefly how glm finds the maximum likelihood estimator $\hat{\beta}$, and discuss its application for Y_i independent Poisson random variables, with mean μ_i, and

$$\log \mu_i = \beta^T x_i, \ 1 \le i \le n.$$

1998/PAPER A1. 13D
Short question.
(i) Suppose $Y_1, ..., Y_n$ are independent observations, with

$$E(Y_i) = \mu_i, \ g(\mu_i) = \beta^T x_i, \ 1 \le i \le n,$$

where $g(\cdot)$ is a known function. Suppose also that Y_i has a probability density function

$$f(y_i|\theta_i, \phi) = exp[(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)]$$

where ϕ is known. Show that if $\ell(\beta)$ is defined as the corresponding log likelihood, then

$$\frac{\partial \ell}{\partial \beta} = \sum \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i}$$

where $V_i = var(Y_i), \ 1 \le i \le n$.
Long question
(ii) Murray *et al.* (1981) in a paper "Factors affecting the consumption of psychotropic drugs" presented the data on a sample of individuals from West London in the table below:

sex	age.group	psych	r	n
1	1	1	9	531
1	2	1	16	500
1	3	1	38	644
1	4	1	26	275
1	5	1	9	90
1	1	2	12	171
1	2	2	16	125
1	3	2	31	121
1	4	2	16	56
1	5	2	10	26
2	1	1	12	588
2	2	1	42	596
2	3	1	96	765
2	4	1	52	327
2	5	1	30	179
2	1	2	33	210
2	2	2	47	189
2	3	2	71	242
2	4	2	45	98
2	5	2	21	60

Here r is the number on drugs, out of a total number n. The variable 'sex' takes values 1, 2 for males, females respectively, and the variable 'psych' takes values 1, 2, according to whether the individuals are not, or are, psychiatric cases.
Discuss carefully the interpretation of the R-analysis below. (You need not prove any of the relevant theorems needed for your discussion, but should quote them carefully.)

```
data = read.table("data", header=T)
attach(data)
sex = factor(sex); psych = factor(psych)
age.group = factor(age.group)
summary(glm(r/n ~ sex + age.group + psych, binomial, weights=n))
 deviance = 14.803
      d.f.= 13

Coefficients:
             Value    Std.Error
 (Intercept) -4.016   0.1506
  sex         0.6257  0.09554
  age.group2  0.7791  0.1610
  age.group3  1.323   0.1476
  age.group4  1.748   0.1621
  age.group5  1.712   0.1899
  psych       1.417   0.09054
```

The term 'sex' is dropped from the model above, and the deviance then increases by 45.15 (corresponding to a 1 d.f. increase) to 59.955 (14 d.f.). What do you conclude?
1998/PAPER A2. 11D
Short question
(i) Suppose that $Y_1, \ldots, Y_n$ are independent Poisson random variables, with $E(Y_i) = \mu_i$, $1 \le i \le n$.
Let H be the hypothesis $H : \mu_1, \ldots, \mu_n \ge 0$.
Show that D, the deviance for testing

$$H_0 : log\mu_i = \mu + \beta^T x_i,\ 1 \le i \le n,$$

where $x_1, \ldots, x_n$ are given covariates, and μ, β are unknown parameters, may be written

$$D = 2[\sum y_i log y_i - \hat{\mu} \sum y_i - \hat{\beta}^T \sum x_i y_i],$$

where you should give equations from which $(\hat{\mu}, \hat{\beta})$ can be determined.
How would you make use of D in practice?

Long question
(ii) A.Sykes (1986) published the sequence of reported new cases per month of AIDS in the UK for each of 36 consecutive months up to November 1985. These data are used in the analysis below, but have been grouped into 9 (non-overlapping) blocks each of 4 months, to give 9 consecutive readings.
It is hypothesised that for the logs of the means, *either*, there is a quadratic dependence on i, the block number *or*, the increase is linear, but with a 'special effect' (of unknown cause) coming into force after the first 5 blocks.
Discuss carefully the analysis that follows below, commenting on the fit of the above hypotheses.

```
n = scan()
3 5 16 12 11 34 37 51 56

i = scan()
1 2 3 4 5 6 7 8 9

summary(glm(n~i,poisson))
deviance = 13.218
    d.f. = 7
Coefficients:
             Value  Std.Error
(intercept)  1.363  0.2210
i            0.3106 0.0382

ii = i*i ; summary(glm(n~ i + ii, poisson))
deviance = 11.098
     d.f.=  6

Coefficients:
             Value   Std.Error
(Intercept)  0.7755  0.4845
i            0.5845  0.1712
ii         -0.02030  0.0141

 special = scan()
1 1 1 1 1 2 2 2 2

special = factor(special)
summary(glm(n~ i + special, poisson))
deviance = 8.2427
     d.f.= 6
Coefficients:
            Value  Std.Error
(intercept) 1.595   0.2431
i           0.2017  0.0573
special     0.6622  0.2984
```

Long question
1998/PAPER 4, 14D
Write an essay on fitting the model

$$\omega : y_i = \beta^T x_i + \epsilon_i, \ 1 \leq i \leq n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent normal, mean 0, variance σ^2, and where β, σ^2 are unknown, and $x_1, \ldots, x_n$ are known covariates. Include in your essay discussion of the following special cases of ω :

$$\omega_1 : y_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \ 1 \le i \le n,$$

$$\omega_2 : y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, 1 \le k \le n_{ij}, 1 \le i \le r, 1 \le j \le c,$$

where $\sum \sum n_{ij} = n$.

[Any distribution results that you need may be quoted without proofs.]

1999/Paper 1. 13D

Short question

(i) Suppose Y_i are independent Binomial variables, and

$$Y_i \sim Bi(n_i, p_i), \quad 1 \le i \le k.$$

Discuss carefully the maximum likelihood estimation of the parameters (α, β) in the model

$$\omega : log(p_i/(1 - p_i)) = \alpha + \beta x_i, \quad 1 \le i \le k,$$

where $x_1, \cdots, x_k$ are given covariates. How would you assess the fit of the model ω in practice?

Long question

(ii) You see below a table of data analysed in R via glm(.).

```
A  B    n    r
1  1  796  498
1  2 1625  878
2  1  142   54
2  2  660  197
```

With A and B each defined as factors,

```
glm(r/n ~ A+ B, family=binomial, weights=n)
```

found that the deviance was .00019, with 1 df, and the estimates for $A(2), B(2)$ were respectively $-1.015(se = 0.0872)$, $-0.3524(se = 0.0804)$, and "intercept" $0.5139(se = 0.687)$. What is the model that is being fitted here? Does it fit well? How do you interpret the parameter estimates? How would you compute the fitted values of r/n for $A = 1, B = 1$?

[In the original data set, A and B correspond to race and sex respectively, and r/n was the observed proportion of a certain type of success.]

1999/Paper 2. 12D

Short question

(i) Suppose that the random variable Y has probability density function

$$f(y|\theta, \phi) = exp[(y\theta - b(\theta))/\phi + c(y, \phi)]$$

for $-\infty < y < \infty$. Show that for $-\infty < \theta < \infty, \ \phi > 0$

$$E(Y) = b'(\theta), \ var(Y) = \phi b''(\theta).$$

Long question

(ii) Suppose that we have independent observations $Y_1, \cdots, Y_n$ and that we assume the model
$\omega : Y_i$ is Poisson, parameter μ_i, and $log(\mu_i) = \beta_0 + \ \beta_1 x_i$,
where $x_1, \cdots, x_n$ are given scalar covariates.

Find the equations for the maximum likelihood estimators $\hat\beta_0, \hat\beta_1$, and state without proof the asymptotic distribution of $\hat\beta_1$.

If, for a particular Poisson model you found that the deviance obtained on fitting ω was 29.3, where $n = 35$, what would you conclude?

Long question

1998/Paper 4. 14D

Consider the linear regression

$$Y = X\beta + \epsilon,$$

where Y is an n-dimensional observation vector, X is an $n \times p$ matrix of rank p, and ϵ is an n-dimensional vector with components $\epsilon_1, \cdots, \epsilon_n$, where $\epsilon_1, \cdots, \epsilon_n$ are normally and independently distributed, each with mean 0 and variance σ^2. We write this as $\epsilon \sim N_n(0, \sigma^2 I_n)$.
(a) Let $\hat{\beta}$ be the least-squares estimator of β. Show that

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

(b) Define $\hat{Y} = X\hat{\beta}$ and $\hat{\epsilon} = Y - \hat{Y}$. Show that $\hat{Y}$ may be written

$$\hat{Y} = HY,$$

where H is a matrix to be defined.
(c) Show that $\hat{Y}$ is distributed as $N_n(X\beta, H\sigma^2)$, and $\hat{\epsilon}$ is distributed as $N_n(0, (I_n - H)\sigma^2)$.
(d) Show that if h_i is defined as the ith diagonal element of H, then $0 \leq h_i \leq 1$, for $i = 1, \cdots, n$.
(e) Why is h_i referred to as the "leverage" of the ith point? Sketch a graph as part of your answer.
Hint: You may assume that if the n-dimensional vector Z has the multivariate normal distribution, mean μ, and covariance matrix V, so that we may write

$$Z \sim N_n(\mu, V),$$

then for any constant $q \times n$ matrix A,

$$AZ \sim N_q(A\mu, AVA^T).$$

2000/I/13
Short question
(i) Consider the linear regression

$$Y = X\beta + \epsilon,$$

where Y is an n-dimensional observation vector, X is an $n \times p$ matrix of rank p, and ϵ is an n-dimensional vector with components $\epsilon_1, \cdots, \epsilon_n$. Here $\epsilon_1, \cdots, \epsilon_n$ are normally and independently distributed, each with mean 0 and variance σ^2; we write this as $\epsilon \sim N_n(0, \sigma^2 I_n)$.
(a) Define $R(\beta) = (Y - X\beta)^T (Y - X\beta)$. Find an expression for $\hat{\beta}$, the least squares estimator of β, and state without proof the joint distribution of $\hat{\beta}$ and $R(\hat{\beta})$.
(b) Define $\hat{\epsilon} = Y - X\hat{\beta}$. Find the distribution of $\hat{\epsilon}$.
Long question
(ii) We wish to investigate the relationship between n, the number of arrests at football matches in a given year, and a, the corresponding attendance (in thousands) at those matches, for the First and Second Divisions clubs in England and Wales. Thus, we have data

$$(n_{ij}, a_{ij}) \;\; j = 1, \cdots, N_i, \;\; i = 1, 2,$$

where $N_1 = 21$ and $N_2 = 23$. We fit the model

$$H_0 : log(n_{ij}) = \mu + \beta log(a_{ij}) + \theta_i + \epsilon_{ij} \;\; j = 1, \cdots, N_i, \;\; i = 1, 2,$$

with $\theta_1 = 0$, and we assume that the ϵ_{ij} are distributed as independent $N(0, \sigma^2)$ random variables.
We find the following estimates, with standard errors given in brackets:
$\hat{\mu} = -0.9946(2.1490)$
$\hat{\beta} = 0.8863(0.3647)$
$\hat{\theta}_2 = 0.5261(0.3401)$
with residual sum of squares $= 37.89(41 df)$. The residual sum of squares if we fit H_0 with β and θ_2 each set to 0 is 43.45.
Give an interpretation of these results, using an appropriate sketch graph.
How could you check the assumptions about the distribution of (ϵ_{ij})?
What linear model would you try next?
2000/2/12
Short question.
(i) Suppose that $Y_1, \cdots, Y_n$ are independent observations, with $E(Y_i) = \mu_i$, $g(\mu_i) = \beta^T x_i$, where

$g(\cdot)$ is the known "link" function, β is an unknown vector of dimension p, and $x_1, ..., x_n$ are given covariate vectors. Suppose further that the log-likelihood for these data is $\ell(\beta)$, where we may write

$$\ell(\beta) = \frac{(\Sigma_1^p \beta_\nu t_\nu(y) - \psi(\beta))}{\phi} + \text{constant},$$

for some function $\psi(\beta)$. Here $t_1(y), ..., t_p(y)$ are given functions of the data $y = (y_1, \cdots, y_n)$, and ϕ is a known positive parameter.
(a) What are the sufficient statistics for β?
(b) Show that $E(t_\nu(Y)) = \frac{\partial \psi}{\partial \beta_\nu}$, for $\nu = 1, ..., p$.
Long question.
(ii) With the same notation as in Part (i), find an expression for the covariance matrix of $(t_1(Y), ..., t_p(Y))$, and hence show that $\ell(\beta)$ is a concave function. Why is this result useful in the evaluation of $\hat{\beta}$, the maximum likelihood estimator of β?
Illustrate your solution by the example

$$Y_i \sim Bi(1, \mu_i) \text{ where } 0 < \mu_i < 1,$$

$$log \frac{\mu_i}{(1 - \mu_i)} = \beta x_i, \ 1 \leq i \leq n,$$

with $x_1, ..., x_n$ known covariate values, each of dimension 1. Your solution should include a statement of the large-sample distribution of $\hat{\beta}$.
2000/4/14
Long question
In an actuarial study, we have independent observations on numbers of deaths $y_1, ..., y_n$ and we assume that Y_i has a Poisson distribution, with mean $\mu_i t_i$, for $i = 1, ..., n$. Here $(t_1, .., t_n)$ are given quantities, for example "person-years at risk".
(a) Find the maximum likelihood estimators $\hat{\mu}_1, ..., \hat{\mu}_n$.
(b) Now consider the model

$$\omega : log\mu_i = \beta^T x_i, \ 1 \leq i \leq n,$$

where $x_1, ..., x_n$ are given vectors, each of dimension p. Derive the equations for $\hat{\beta}$, the maximum likelihood estimator of β, and briefly discuss the method of solution used by the function glm( ) in R to solve this equation.
(c) How is the deviance for ω computed ? If you found that this deviance took the value 27.3, and you knew that $n = 37, p = 4$, what would you conclude about ω?
(d) Discuss briefly how your answers to the above are affected if the model ω is replaced by the model

$$\omega_I : \mu_i = \beta^T x_i, \ 1 \leq i \leq n.$$

2001/ 1/13.
Short question.
(i) Assume that the n-dimensional observation vector Y may be written as

$$Y = X\beta + \epsilon \ ,$$

where X is a given $n \times p$ matrix of rank p, β is an unknown vector, and

$$\epsilon \sim N_n(0, \sigma^2 I).$$

Let $Q(\beta) = (Y - X\beta)^T(Y - X\beta)$. Find $\hat{\beta}$, the least-squares estimator of β, and show that

$$Q(\hat{\beta}) = Y^T(I - H)Y,$$

where H is a matrix that you should define.
Long question
(ii) Show that $\Sigma_i H_{ii} = p$. Show further for the special case of

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i, \ 1 \leq i \leq n,$$

where $\Sigma x_i = 0, \Sigma z_i = 0$, that

$$H = \frac{1}{n}\mathbf{1}\mathbf{1}^T + axx^T + b(xz^T + zx^T) + czz^T \ ;$$

here, $\mathbf{1}$ is a vector of which every element is one, and a, b, c, are constants that you should derive. Hence show that, if $\hat{Y} = X\hat{\beta}$ is the vector of fitted values, then

$$\frac{1}{\sigma^2}\mathrm{var}(\hat{Y}_i) = \frac{1}{n} + ax_i^2 + 2bx_iz_i + cz_i^2, \ 1 \le i \le n.$$

2001/2/12
Short question.
(i) Suppose that $Y_1, \ldots, Y_n$ are independent random variables, and that Y_i has probability density function

$$f(y_i|\theta_i, \phi) = \exp[(y_i\theta_i - b(\theta_i))/\phi + c(y_i, \phi)].$$

Assume that $E(Y_i) = \mu_i$, and that $g(\mu_i) = \beta^T x_i$, where $g(.)$ is a known 'link' function, $x_1, \ldots, x_n$ are known covariates, and β is an unknown vector. Show that

$$E(Y_i) = b'(\theta_i), \ \mathrm{var}(Y_i) = \phi b''(\theta_i) = V_i, \ \text{say},$$

and hence

$$\frac{\partial l}{\partial \beta} = \sum_i \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i}, \ \text{where } l = l(\beta, \phi) \text{ is the log} - \text{likelihood}.$$

Long question
(ii) The table below shows the number of train miles (in millions) and the number of collisions involving British Rail passenger trains between 1970 and 1984. Give a detailed interpretation of the *R* output that is shown under this table:

```
   year collisions miles
1  1970          3   281
2  1971          6   276
3  1972          4   268
4  1973          7   269
5  1974          6   281
6  1975          2   271
7  1976          2   265
8  1977          4   264
9  1978          1   267
10 1979          7   265
11 1980          3   267
12 1981          5   260
13 1982          6   231
14 1983          1   249

Call:
glm(formula = collisions ~ year + log(miles), family = poisson)

Coefficients:

              Estimate   Std. Error z value Pr(>|z|)
(Intercept) 127.14453  121.37796   1.048    0.295
year         -0.05398    0.05175  -1.043    0.297
log(miles)   -3.41654    4.18616  -0.816    0.414

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 15.937  on 13  degrees of freedom
```

```
Residual deviance: 14.843  on 11  degrees of freedom

Number of Fisher Scoring iterations: 4
```

2001/4/14
Short question
(i) Assume that the independent observations $Y_1, \ldots, Y_n$ are such that

$$Y_i \sim \text{Binomial}(t_i, \pi_i), \text{ and } \log \frac{\pi_i}{1-\pi_i} = \beta^T x_i \text{ for } 1 \le i \le n \ ,$$

where $x_1, \ldots, x_n$ are given covariates. Discuss carefully how to estimate β, and how to test that the model fits.
Long question
(ii) Carmichael *et al.* (1989) collected data on the numbers of 5-year old children with "dmft", i.e. with 5 or more decayed, missing or filled teeth, classified by social class, and by whether or not their tap water was fluoridated or non-fluoridated. The numbers of such children with dmft and the total numbers, are given in the table below:

```
                       dmft
Social Class  |Fluoridated Non-fluoridated
--------------|---------------------------
 I            | 12/117      12/56
 II           | 27/170      48/146
 III          | 11/52       29/64
 Unclassified | 24/118      49/104
```

A (slightly edited) version of the *R* output is given below. Explain carefully what model is being fitted, whether it does actually fit, and what the parameter estimates and Std. Errors are telling you. (You may assume that the factors SClass (social class) and Fl (with/without) have been correctly set up.)

```
Call:
glm(formula = Yes/Total ~ SClass + Fl, family = binomial, weights = Total)

Coefficients:
            Estimate Std. Error z value
(Intercept)  -2.2716     0.2396  -9.480
SClassII      0.5099     0.2628   1.940
SClassIII     0.9857     0.3021   3.262
SClassUnc     1.0020     0.2684   3.734
Flwithout     1.0813     0.1694   6.383

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.53785  on 7  degrees of freedom
Residual deviance:  0.64225  on 3  degrees of freedom

Number of Fisher Scoring iterations: 3
```

Here 'Yes' is the vector of numbers with dmft, taking values $12, 12, \ldots, 49$, 'Total' is the vector of Total in each category, taking values $117, 56, \ldots, 118, 104$ and SClass, Fl are the factors corresponding to Social class and Fluoride status, defined in the obvious way.
2002/Question 1.
Short question.
(i) Suppose $Y_1, \ldots, Y_n$ are independent Poisson variables, and

$$E(Y_i) = \mu_i \ , \log \mu_i = \alpha + \beta^T x_i \ , \ 1 \le i \le n$$

where α, β are unknown parameters, and $x_1, \ldots, x_n$ are given covariates, each of dimension p. Obtain the maximimum likelihood equations for α, β, and explain briefly how you would check the

validity of this model.
Long question.
(ii) The data below show $y_1, \ldots, y_{33}$, which are the monthly accident counts on a major US highway for each of 12 months of 1970, then for each of 12 months of 1971, and finally for the first 9 months of 1972. The data-set is followed by the (slightly edited) R output. You may assume that the factors 'Year' and 'month' have been set up in the appropriate fashion. Give a careful interpretation of this R output, and explain
a) how you would derive the corresponding standardised residuals, and
b) how you would predict the number of accidents in October 1972.

```
52 37 49 29 31 32 28 34 32 39 50 63
35 22 27 27 34 23 42 30 36 56 48 40
33 26 31 25 23 20 25 20 36
```

```
>first.glm = glm(y~ Year + month, poisson) ; summary(first.glm)

Call:
glm(formula = y ~ Year + month, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.81969    0.09896  38.600  < 2e-16 ***
Year1971    -0.12516    0.06694  -1.870 0.061521 .
Year1972    -0.28794    0.08267  -3.483 0.000496 ***
month2      -0.34484    0.14176  -2.433 0.014994 *
month3      -0.11466    0.13296  -0.862 0.388459
month4      -0.39304    0.14380  -2.733 0.006271 **
month5      -0.31015    0.14034  -2.210 0.027108 *
month6      -0.47000    0.14719  -3.193 0.001408 **
month7      -0.23361    0.13732  -1.701 0.088889 .
month8      -0.35667    0.14226  -2.507 0.012168 *
month9      -0.14310    0.13397  -1.068 0.285444
month10      0.10167    0.13903   0.731 0.464628
month11      0.13276    0.13788   0.963 0.335639
month12      0.18252    0.13607   1.341 0.179812
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 101.143  on 32  degrees of freedom
Residual deviance:  27.273  on 19  degrees of freedom

Number of Fisher Scoring iterations: 3
```

Question 2.
Short question
(i) Suppose that the random variable Y has density function of the form

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

where $\phi > 0$. Show that Y has expectation $b'(\theta)$ and variance $\phi b''(\theta)$.
Long question.
(ii) Suppose now that $Y_1, \ldots, Y_n$ are independent negative exponential variables, with Y_i having density function

$$f(y_i|\mu_i) = (1/\mu_i)e^{-y_i/\mu_i}$$

for $y_i > 0$. Suppose further that $g(\mu_i) = \beta^T x_i$ for $1 \le i \le n$, where $g(\ )$ is a known 'link' function, and $x_1, \ldots, x_n$ are given covariate vectors, each of dimension p. Discuss carefully the problem of finding $\hat{\beta}$, the maximum likelihood estimator of β, firstly for the case $g(\mu_i) = 1/\mu_i$, and secondly for the case $g(\mu_i) = \log \mu_i$.
(Any standard theorems used need not be proved.)
Paper 4 question.
Long question.
Assume that the $n-$dimensional observation vector Y may be written as

$$Y = X\beta + \epsilon$$

where X is a given $n \times p$ matrix of rank p, β is an unknown vector, with $\beta^T = (\beta_1, \ldots, \beta_p)$, and

$$\epsilon \sim N_n(0, \sigma^2 I) \qquad *$$

where σ^2 is unknown. Find $\hat{\beta}$, the least-squares estimator of β, and describe (without proof) how you would test

$$H_0 : \beta_\nu = 0$$

for a given ν.
Indicate briefly two plots that you could use as a check of the assumption *.
Sulphur dioxide is one of the major air pollutants. A data-set presented by Sokal and Rohlf (1981) was collected on 41 US cities in 1969-71, corresponding to the following variables:
Y = Sulphur dioxide content in micrograms per cubic metre
$X1$ = average annual temperature in degrees Fahrenheit
$X2$ = number of manufacturing enterprises employing 20 or more workers
$X3$ = population size (1970 census) in thousands
$X4$ = Average annual wind speed in miles per hour
$X5$ = Average annual precipitation in inches
$X6$ = Average annual number of days with precipitation per year.
Interpret the R output that follows below, quoting any standard theorems that you need to use.

```
>next.lm <- lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6)
>summary(next.lm)

Call:
lm(formula = log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6)

Residuals:
     Min       1Q   Median       3Q      Max
-0.79548 -0.25538 -0.01968  0.28328  0.98029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
X1          -0.0599017  0.0190138  -3.150  0.00339 **
X2           0.0012639  0.0004820   2.622  0.01298 *
X3          -0.0007077  0.0004632  -1.528  0.13580
X4          -0.1697171  0.0555563  -3.055  0.00436 **
X5           0.0173723  0.0111036   1.565  0.12695
X6           0.0004347  0.0049591   0.088  0.93066
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'

Residual standard error: 0.448 on 34 degrees of freedom
Multiple R-Squared: 0.6541
F-statistic: 10.72 on 6 and 34 degrees of freedom,p-value: 1.126e-06
```

2003/Paper 1, number 13.
Short question.
(i) Suppose $Y_i,\ 1 \le i \le n$, are independent binomial observations, with $Y_i \sim Bi(t_i, \pi_i),\ 1 \le i \le n$, where $t_1, \ldots, t_n$ are known, and we wish to fit the model

$$\omega : log(\pi_i/(1-\pi_i)) = \mu + \beta^T x_i, \text{ for each } i,$$

where $x_1, \ldots, x_n$ are given covariates, each of dimension p. Let $\hat{\mu}, \hat{\beta}$ be the maximum likelihood estimators of μ, β. Derive equations for $\hat{\mu}, \hat{\beta}$ and state without proof the approximate distribution of $\hat{\beta}$.
Long question.
(ii) In 1975, data were collected on the 3-year survival status of patients suffering from a type of cancer, yielding the following table

```
                         survive?
age in years malignant yes no
under 50        no      77  10
under 50        yes     51  13
50-69           no      51  11
50-69           yes     38  20
70+             no       7   3
70+             yes      6   3
```

Here the second column represents whether the initial tumour was no malignant or was malignant. Let Y_{ij} be the number surviving, for age group i and malignancy status j, for $i = 1, 2, 3$ and $j = 1, 2$, and let t_{ij} be the corresponding total number. Thus $Y_{11} = 77, t_{11} = 87$. Assume $Y_{ij} \sim Bi(t_{ij}, \pi_{ij},\ 1 \le i \le 3, 1 \le j \le 2$. The results from fitting the model

$$log(\pi_{ij}/(1-\pi_{ij})) = \mu + \alpha_i + \beta_j$$

with $\alpha_1 = 0, \beta_1 = 0$ give $\hat{\beta}_2 = -0.7328(se = 0.2985)$, and deviance $= 0.4941$. What do you conclude?
Why do we take $\alpha_1 = 0, \beta_1 = 0$ in the model?
What "residuals" should you compute, and to which distribution would you refer them?
Paper 2, number 12.
Short question
(i) Suppose $Y_1, \ldots, Y_n$ are independent Poisson variables, and

$$E(Y_i) = \mu_i, \quad log(\mu_i) = \alpha + \beta t_i, \quad \text{for } i = 1, \ldots, n,$$

where α, β are two unknown parameters, and $t_1, \ldots, t_n$ are given covariates, each of dimension 1. Find equations for $\hat{\alpha}, \hat{\beta}$, the maximum likelihood estimators of α, β, and show how an estimator of $var(\hat{\beta})$ may be derived, quoting any standard theorems you may need.
Long question
(ii) By 31 December 2001, the number of new vCJD patients, classified by reported calendar year of onset, were

$$8, 10, 11, 14, 17, 29, 23$$

for the years

$$1994, \ldots, 2000 \text{ respectively.}$$

Discuss carefully the (slightly edited) R output for these data given below, quoting any standard theorems you may need.

```
> year
[1] 1994 1995 1996 1997 1998 1999 2000
> tot
[1]  8 10 11 14 17 29 23
> first.glm <- glm(tot ~ year, family=poisson)
```

```
> summary(first.glm)

Call:
glm(formula = tot ~ year, family = poisson)

Coefficients:
             Estimate     Std.Error z-value Pr(>|z|)
(Intercept) -407.81284   99.35709  -4.105 4.05e-05
year           0.20556    0.04973   4.133 3.58e-05

     (Dispersion parameter for poisson family taken to be 1)

        Null deviance: 20.7753  on 6  degrees of freedom
    Residual deviance:  2.7931  on 5  degrees of freedom
```

Paper 4, number 14
Long question
The nave height, x and the nave length, y, for 16 Gothic-style cathedrals and 9 Romanesque-style cathedrals, all in England, have been recorded, and the corresponding R output (slightly edited) is given below.

```
> first.lm <- lm(y~x + Style); summary(first.lm)

  Call:
 lm(formula = y ~ x + Style)

  Residuals:
     Min      1Q  Median      3Q     Max
 -172.67  -30.44   20.38   55.02   96.50

      Coefficients:
             Estimate  Std.Error t-value Pr(>|t|)
 (Intercept)   44.298     81.648   0.543   0.5929
 x              4.712      1.058   4.452   0.0002
 Style2        80.393     32.306   2.488   0.0209

  Residual standard error: 77.53 on 22 degrees of freedom
  Multiple R-Squared: 0.5384
```

You may assume that x, y are in suitable units, and that ‘Style’ has been set up as a factor with levels 1, 2 corresponding to Gothic, Romanesque respectively.
(a) Explain carefully, with suitable graph(s) if necessary, the results of this analysis.
(b) Using the general model $Y = X\beta + \epsilon$ (in the conventional notation) explain carefully the theory needed for (a).
[Standard theorems need not be proved.]