Computational Statistics and Statistical Modelling: Mathematical Tripos, Part IIA, questions from 1997 onwards

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

July 5, 2005

I/12M

i) Assume that the n-dimensional observation vector Y may be written

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

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.

If now $X\beta$ is written as $X\beta = X_1\beta_1 + X_2\beta_2$, where $X = (X_1 : X_2)$, $\beta^T = (\beta_1^T : \beta_2^T)$, and β_2 is of dimension p_2, state without proof the form of the $F-$test for testing $H_0 : \beta_2 = 0$.

ii) The data in the GLIM analysis shown were obtained as part of a health survey in the US investigating cholesterol levels in women from two states, Iowa and Nebraska (denoted by 1,2 respectively). Discuss carefully the GLIM analysis below, using sketch graphs to illustrate the two models being fitted. Which is your preferred model? What F-test would be appropriate?

```
$un 30 $data state age chol $read
1 46 181 1 52 228 1 39 182 1 65 249 1 54 259
1 33 201 1 49 121 1 76 339 1 71 224 1 41 112 1 58 189
2 18 137 2 44 173 2 33 177 2 78 241 2 51 225 2 43 223
2 44 190 2 58 257 2 63 337 2 19 189 2 42 214 2 30 140
2 47 196 2 58 262 2 70 261 2 67 356 2 31 159 2 21 191
2 56 197
$fact state 2 $yvar chol $fit state*age$d e $
$fit - state.age$d e $
```

(nb these data came from the Glim Manual, p368. I have corrected a typo in the 2nd line of the data; ie 232 has been replaced above by 121. This error would not have affected the candidates, but it would certainly affect later students if they were seeking to check all the estimates.)

```
$factor state 2 $yvar chol $fit state*age $d e $
deviance= 48395
df      = 26

     estimate   se   parameter
   1   35.81   55.12   1
   2   65.49   61.98   STAT(2)
   3   3.238   1.009   AGE
   4  -0.7177  1.163   STAT(2).AGE
 scale parameter taken as 1861
```

```
$fit - state.age$d e$
deviance = 49104 (change = +709.1)
  df     = 27    (change = +1)

     estimate   se   parameter
   1   64.49   29.30   1
   2   28.65   16.54   STAT(2)
   3   2.698   0.4960  AGE
 scale parameter taken as 1819
```

Outline solution

i)

With $Q(\beta) = (Y - X\beta)^T(Y - X\beta)$, we see that $Q(\beta)$ is minimised with respect to β by $\hat{\beta}$, the solution to

$$\frac{\partial Q(\beta)}{\partial \beta} = 0$$

ie

$$X^T(Y - X\beta) = 0.$$

Note that $\text{rank}(X) = \text{rank}(X^TX)$, hence the $p \times p$ matrix X^TX is of full rank, hence $(X^TX)^{-1}$ exists, hence $\hat{\beta}$ is the unique solution to

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

and by simple manipulation, we see that

$$Q(\hat{\beta}) = Y^TY - Y^THY.$$

H is of course the usual 'hat' matrix $X(X^TX)^{-1}X^T$.

With $X\beta = X_1\beta_1 + X_2\beta_2$, let us write R_Ω, R_ω as the residual sums of squares fitting the models

$$\Omega : E(Y) = X\beta, \quad \omega : E(Y) = X_1\beta_1$$

respectively.

Note that $\omega \subseteq \Omega$, and $dim(\Omega) - dim(\omega) = p_2$.

Facts: to be quoted without proof,

on $\Omega, R_\Omega/\sigma^2 \sim \chi^2_{n-p}$

and on $\omega, (R_\omega - R_\Omega)/\sigma^2 \sim \chi^2_{p_2}$

and these are independent random variables.

So to test ω against Ω, we refer

$$\frac{(R_\omega - R_\Omega)/p_2}{R_\Omega/(n-p)} \text{ to } F_{p_2,n-p}$$

.

All of the above is standard book work.

ii) The model we are fitting is say

$$\Omega : y_{ij} = \mu + \theta_i + \beta x_{ij} + \gamma_i x_{ij} + \epsilon_{ij}, \text{ for } 1 \le j \le n_i, \ 1 \le i \le 2.$$

Here y_{ij} is the cholesterol reading for the jth woman, ith state,

and $i = 1, 2$ for Iowa, Nebraska respectively,

x_{ij} is the corresponding age

and $\theta_1 = 0, \gamma_1 = 0$ (the usual GLIM constraints)

and $\epsilon_{ij} \sim NID(0, \sigma^2)$.

So we see that we are fitting 2 lines, which have possibly different slopes and different intercepts.

We find that the fitted lines are

for Iowa, $chol = 35.81 + 3.238 \times age$

and for Nebraska, $chol = (35.81 + 65.49) + (3.238 - 0.7177) \times age$

with our estimate of σ^2 as deviance/df = 48395/26.

But we see, from .7177/1.163 (referred to t_{26}, but there is no need for formal use of tables) that we can DROP the state.age interaction, and so we fit two parallel straight lines, resulting in the fitted lines

$chol = 64.49 + 2.698 \times age$ for Iowa,

$chol = (64.49 + 28.65) + 2.698 \times age$ for Nebraska,

with σ^2 estimated by 49104/27.

The appropriate F-test , which tests whether the 2 slopes are indeed the same, is to refer ((709.1/1)/(48395/26)) to $F_{1,26}$ and it would presumably be non- significant.

In fact (look at (28.65/16.54)) we could probably use just one line for the whole data-set.

paper2/11M

i) Suppose that $Y_1, ..., Y_n$ are independent binomial observations, with

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

where $t_1, ..., t_n$ and $x_1, ..., x_n$ are given. Discuss carefully the estimation of β.

ii) A new drug is thought to check the development of symptoms of a particular disease. A study on 338 patients who were already infected with this disease yielded the data below.

		Symptoms	
Race	Drug use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

You see below the corresponding GLIM analysis. Discuss its interpretation carefully.

```
$units 4 $data y n $read
14 93 32 81 11 52 12 43 $
$yvar y $cal tot=y+n$err b tot $
$factor race 2 drug 2 $fit race + drug$d e $
scaled deviance = 1.3835
             df = 1

   estimate    se     parameter
 1  -1.1738   0.2404  1
 2  -0.05548  0.2886  race(2)
 3   0.7195   0.2790  drug(2)
```

Solution

i) With $f(y_i|\pi_i) \propto \pi_i^{y_i}(1-\pi_i)^{t_i - y_i}$ we see that

$$\ell(\beta) = \sum (y_i \log(\pi_i/(1-\pi_i)) + t_i \log(1-\pi_i)).$$

Substitute for π_i in terms of β to give

$$\ell(\beta) = \beta^T \sum x_i y_i - \sum t_i \log(1 + \exp(\beta^T x_i)) + \text{constant}.$$

Hence

$$\frac{\partial \ell}{\partial \beta} = \sum x_i y_i - \sum t_i x_i \pi_i$$

and so

$$-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = \sum t_i x_i {x_i}^T \pi_i(1-\pi_i) = (V(\beta))^{-1}$$

say. The rest of the solution consists of describing the iterative solution of

$$\frac{\partial \ell}{\partial \beta} = 0$$

and the large-sample distribution of $\hat{\beta}$ which is of course

$$N(\beta, V(\beta)).$$

ii) The fitted model is

$$y_{ij} \sim B(tot_{ij}, \pi_{ij}),\ 1 \leq i,j \leq 2$$

with $i = 1, 2$ corresponding to Race (White, Black) and $j = 1, 2$ corresponding to Drug Use (Yes, No).

We fit

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

with $\alpha_1 = \beta_1 = 0$, the usual GLIM constraints.

Thus, using Wilks' theorem, we may test the adequacy of ω by referring 1.385 to ${\chi^2}_1$, so that our model ω clearly fits well.

Furthermore, $\hat{\alpha}_2/se(\hat{\alpha}_2)$ is clearly non-significant when referred to $N(0,1)$, so that Race is not significant in its effect on Symptoms(Yes/No).

However, (7195/.2790) is clearly in the tail of $N(0,1)$, showing that [Drug Use = No] increases the probability of [Symptoms = Yes]; the drug use is effective in reducing the probability of Symptoms.

Note that we could have tried to fit

```
$fit race*drug$
```

allowing for a possible interaction between Race and Drug use; this model would have given us a perfect fit (zero deviance), but it is in any case obvious from the fact the the model ω fits so well that the race.drug term must be non-significant.

4/13M 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) $\text{var}(Y_i) = \phi b''(\theta_i) = V_i$ 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 GLIM finds the maximum likelihood estimator $\hat{\beta}$, and discuss its application for Y_i independent Poisson random variables, with mean μ_i, and

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

Solution
This is 'the calculus at the heart of GLIM': see your lecture notes for the full story.
The example has $\phi = 1$
and

$$\ell(\beta) = -\sum \exp(\beta^T x_i) + \beta^T \sum x_i y_i + \text{constant}$$

so that GLIM will solve, by iteration, the simultaneous equations

$$\frac{\partial \ell}{\partial \beta} = 0.$$

Computational Statistics
Mathematical Tripos 1998, Part IIA, questions and solutions

The numerical parts of the questions have been edited somewhat, as you will see below. (They have been recast in R, but are essentially asking the same as in the original version of the question. The R output is given in slightly reduced form.)
PAPER A1. 13D
(i) Suppose $Y_1, ..., Y_n$ are independent observations, with

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

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 $V_i = var(Y_i),\ 1 \leq i \leq n$.

(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?

SOLUTION.
(i) Dropping the suffix i, we see that
$logf(y|\theta,\phi) = (y\theta - b(\theta))/\phi +$ term free of θ.
Thus

$$\frac{\partial logf(y|\theta)}{\partial\theta} = (y - b'(\theta))/\phi.$$

Now apply the well-known results (suppressing the known constant ϕ) that since $\int f(y|\theta)dy = 1$ for all θ,

$$E(\frac{\partial logf(y|\theta)}{\partial\theta})) = 0,$$

and

$$E(\frac{-\partial^2 logf(y|\theta)}{\partial\theta^2}) = var(\frac{\partial logf(y|\theta)}{\partial\theta})$$

and apply the chain-rule, to give the desired expression for

$$\partial\ell/\partial\beta.$$

(ii) The model that we are fitting is $r_i \sim$ independent $Bin(n_i, \pi_i)$, for $1 \le i \le 20$, where (since the logit link is the default for the binomial)

$$log(\pi_i/(1-\pi_i)) = \mu + sex_{j(i)} + age.group_{k(i)} + psych_{l(i)}$$

and, for example, $j(i) = 1, 1, 1, ...2, 2, 2,$ (ie as in the first column of the data). We know that R will assume the usual parameter identifiability conditions:

$$sex_1 = 0,\ age.group_1 = 0, psych_1 = 0,$$

so that in the output, each factor level is effectively being compared with the $first$ corresponding factor level.

By Wilks' theorem, we know that the deviance of 14.803 can be compared to χ^2_{13}, and this comparison shows that the model fits well, since 14.803 is only slightly bigger than the expected value of χ^2_{13}.

We also know that, approximately, each (mle/its standard error) can be compared with $N(0,1)$ to test for significance of that parameter.

So we see that a female is significantly more likely than a comparable male to be on drugs,

and the probability of being on drugs increases as the age.group increases (more or less, since the last 2 age.groups have almost the same parameter estimate)

and those who are psychiatric cases are more likely than those who are *not* psychiatric cases to be on drugs.

If the term 'sex' is dropped from the model, the deviance increases by what is obviously a hugely significant amount, so it was clearly wrong to try to reduce the model in this way (as we should expect, from the original est/se for sex).

PAPER A2. 11D

(i) Suppose that $Y_1, \ldots, Y_n$ are independent Poisson random variables, with $E(Y_i) = \mu_i$, $1 \leq i \leq n$. Let H be the hypothesis $H : \mu_1, \ldots, \mu_n \geq 0$.

Show that D, the deviance for testing

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

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?

(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
```

SOLUTION

(i) We have $f(y_i|\mu_i) \propto e^{-\mu_i}\ {\mu_i}^{y_i}$
from which we see that the loglikelihood is

$$\sum log f(y_i|\mu_i) = -\sum \mu_i + \sum y_i log \mu_i + \text{constant}.$$

Clearly this is maximised under H by

$$\hat{\mu}_i = y_i,\ 1 \leq i \leq n.$$

Under H_0, we see that the loglikelihood is now $\ell(\mu, \beta)$, where

$$\ell(\mu, \beta) = -\sum e^{\mu + \beta^T x_i} + \mu \sum y_i + \beta^T \sum x_i y_i.$$

Hence, taking partial derivatives with respect to μ, β respectively, we obtain the equations

$$\sum e^{\mu + \beta^T x_i} = \sum y_i$$

$$\sum x_i e^{\mu + \beta^T x_i} = \sum x_i y_i,$$

which is a set of equations for $(\hat{\mu}, \hat{\beta})$, which we could solve iteratively by glm().

The given expression for D is twice the difference between the loglikelihood maximised under H, H_0, respectively. (Observe that the $\sum \hat{\mu}_i$ term will cancel.)

Use of D: Wilks' theorem tells us that for large n, on H_0, D is approximately distributed as χ^2_f, where f is the difference in dimension between H and H_0: let us call this $n - 1 - p$.

We see that H_0 will be a good fit to the data if we find that $D \leq n - 1 - p$, (recalling that the expected value of a χ^2 variable is its d.f.)

(ii) Throughout we assume the model $n_i \sim$ independent $Po(\mu_i)$ for $i = 1, \ldots, 9$.

The log link is the default for the Poisson. The first model we try is say

$$H_L : log(\mu_i) = \mu + \beta\ i,\ i = 1, \ldots, 9.$$

This has a deviance which is nearly twice its d.f, showing that H_L is not a good fit. Note that under H_L, the estimate of the slope β is clearly positive: compare $(0.3106/0.0382)$ to $N(0, 1)$.

The next model we try is say

$$H_Q : log(\mu_i) = \mu + \beta\ + \gamma\ i^2,\ i = 1, \ldots, 9.$$

Although the deviance is reduced (by $13.218 - 11.098$), this model still has a deviance nearly twice its d.f. Inspection of $\hat{\gamma}$, -0.02030, and its se, shows that there *may* be a significant quadratic effect. But the next model we try, which extends H_L by one more parameter, but in a different way from H_Q, produces a much better fit. It corresponds to

$H_S : log(\mu_i) = \mu + \beta\ i, i = 1, \ldots, 5,$
and $log(\mu_i) = \mu + special + \beta\ i, i = 6, \ldots, 9.$

This time the deviance is only a little bigger than its d.f. Furthermore, comparing the estimate of 'special' with its se $(0.662/0.2984)$, we see that 'special' (ie the 'jump' in the line) is clearly significant.

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 :

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

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

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

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

SOLUTION
This model may be rewritten in matrix form

$$y = X\beta + \epsilon$$

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

Much of the solution is essentially contained in your lecture notes: here are points that you should cover in your essay, possibly with appropriate sketch diagrams:

a) The estimation of β, σ^2, the joint distribution of these estimates, and how to construct confidence intervals for elements of β. (Remember, you don't need to prove any of the distributional results.)

b) What to do if $X\beta = X_1\beta_1 + X_2\beta_2$, and we want to test, say, $\beta_2 = 0$.

c) The relevance of projections.

d) The relevance of (and of course the definition of) parameter orthogonality.

e) How to check the assumption $\epsilon \sim N(0, \sigma^2 I)$. (ie what to do with *residuals*.)

The two hypotheses ω_1, ω_2 can be used to illustrate some of the above points. Note that if $n_{ij} = u$ say, for all i, j then we have a *balanced* two-way design, for which the standard 'two-way anova' is appropriate.

Mathematical Tripos 1999, Part IIA, questions

Paper 1. 13D (i) Suppose Y_i are independent Binomial variables, and

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

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

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

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

Paper 2. 12D
(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 $-\infty < \theta < \infty$, $\phi > 0$

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

(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?

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,

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

Mathematical Tripos, Part IIA, 2000
I/13
(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 $\hat{\epsilon}$.
(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 $= 37.89(41df)$. 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?

2/12
(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 $\nu = 1, ..., p$.

(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 $\hat{\beta}$.

4/14
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

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

Mathematical Tripos, Part IIA, 2001

1/13.
(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.
(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

$$\frac{1}{\sigma^2}\text{var}(\hat{Y}_i) = \frac{1}{n} + ax_i^2 + 2bx_i z_i + cz_i^2, \ 1 \leq i \leq n.$$

2/12
(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

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

and hence

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

(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
```

4/14

(i) Assume that the independent observations $Y_1, \ldots, Y_n$ are such that

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

where $x_1, \ldots, x_n$ are given covariates. Discuss carefully how to estimate β, and how to test that the model fits.

(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.

P.M.E.Altham's Part IIA Tripos questions, June 2002.
Computational Statistics and Statistical Modelling.
Question 1.
(i) Suppose $Y_1, \ldots, Y_n$ are independent Poisson variables, and

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

where α, β are unknown parameters, and $x_1, \ldots, x_n$ are given covariates, each of dimension p. Obtain the maximum likelihood equations for α, β, and explain briefly how you would check the validity of this model.
(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.
(i) Suppose that the random variable Y has density function of the form

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

where $\phi > 0$. Show that Y has expectation $b'(\theta)$ and variance $\phi b''(\theta)$.
(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 $1 \leq i \leq n$, 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.
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
```

Notes for solution on the R output.
Here, we are fitting

$$log(Y_i) = \mu + \beta_1 X1_i + \ldots + \beta_6 X6_i + \epsilon_i$$

for $i = 1, \ldots, 41$ with the usual assumption that $\epsilon_i \sim NID(0, \sigma^2)$.
We see that $R^2 = 0.6541$ (not a bad fit, but still a lot of scatter). Note that R^2 = (ss due to regression)/("total" ss),
and the F-statistic of 10.72 is closely related to this: specifically
F-statistic = [(ss due to regression)/6]/[(residual ss)/34],
and if the null hypothesis $H : \beta_1 = \ldots = \beta_6 = 0$ is true, this quantity has the distribution F, with $6, 34$ degrees of freedom. Evidently 10.72 is well out in the right-hand tail of this $F-$ distribution, the corresponding p-value is tiny ($1.126e-06$ in fact, and we don't need this ridiculous accuracy, but that's computers for you.)
We reject the hypothesis H.
More interestingly, we can assess the significance of each of the coefficients $\beta_1, \ldots, \beta_6$ in turn, from the corresponding $t-$values. For example, for β_1,

the t-value is $-0.0599017/0.0190138 = -3.150$).
We see that $\beta_3, \beta_5, \beta_6$ can probably be dropped from the model.
Note that because the parameters are almost certainly *non-orthogonal*, when we fit

```
lm(log(Y) ~ X1 + X2 + X4)
```

which would be the natural next step in the fitting process, our estimates for $\beta_1, \beta_2, \beta_4$ may change quite markedly (and so too will their se's, generally reducing a bit).
It appears that (back in 1969-71) the amount of pollution (sulphur dioxide)
decreased as the average annual temperature increased,
increased as the amount of industry increased,
and decreased as the wind speed increased.
When these 3 variables are taken into account, the other 3 variables (namely population size, total rainful p.a., and total number of rainy days p.a.) have no significant effect.

Part IIA, June 2003.
Paper 1, number 13.
(i) Suppose $Y_i,\ 1 \leq i \leq n$, are independent binomial observations, with $Y_i \sim Bi(t_i, \pi_i),\ 1 \leq i \leq n$, 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 $\hat{\beta}$.
(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 $Y_{ij} \sim Bi(t_{ij}, \pi_{ij},\ 1 \leq i \leq 3, 1 \leq j \leq 2$. 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.
(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.
(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 anay 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
```

Notes for solution on the R output.
Here we are fitting the model $Y_i \sim Po(\mu_i)$, independent, with Y_i as the total for $year_i$ and we assume that

$$\omega : log(\mu_i) = \mu + \beta \times year_i$$

for $i = 1, \ldots, 7$. The log-link is of course the default link for the Poisson, since it is the canonical link.
The null deviance, which is the deviance when we fit ω with $\beta = 0$, is 20.77: refer this to χ^2 with 6 df (and recalling that this distribution has expectation = its df), we see that this is a terrible fit. But, the deviance when fitting ω is only 2.7931, which is well below the expected value (2) of χ^2 with 2 df, so we infer that ω fits well.
$\hat{\beta} = .2055(se = .04973)$ and so we compute the corresponding z-value as
$.2055/.04973 = 4.133$ and refer this to the $N(0, 1)$ distribution: the p-value given in the table shows us what we know already, that 4.133 is WAY out in the tail of $N(0, 1)$. We reject the hypothesis $\beta = 0$ in favour of $\beta > 0$, and conclude that the incidence of this illness has increased significantly from 1994 to 2000. (We can also see that $\mu < 0$, but that is less interesting.)
It would be very easy to predict the number of cases in say, 2001, as $exp(\mu + \beta \times 2001)$, replacing the parameters by their estimates.
The reason why the output says
'(Dispersion parameter for poisson family taken to be 1)'
is because, in the general glm formulation, we take $\phi = 1$ for the Poisson. This has the effect that

$$E(Y_i) = \mu_i, var(Y_i) = \phi\mu_i$$

with $\phi = 1$: we might prefer a less stringent model, in which we use the data to estimate ϕ, but in fact that is not necessary for this particular dataset, since ω fits so well.
(I note that the separate figures for men and women, not given here, show a significant **gender** effect (men have a higher incidence than women) but, in the interests of going for a quiet life, I thought it politically inadvisable to include this in the examination question.)

Paper 4, number 14
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.]

Notes for solution on the R output.
We are fitting the model

$$y_{ij} = \mu + \beta x_{ij} + \theta_i + \epsilon_{ij}$$

where we take $i = 1, 2$ corresponding to Gothic, Romanesque, respectively,
$j = 1, \ldots, n_i$,
$\theta_1 = 0$ for parameter identifiability (the default constraint in R), and
$\epsilon_{ij} \sim N(0, \sigma^2)$ independently, with σ^2 unknown.
The R output shows us that
i) $R^2 = 0.5384$: this is the (ss due to regression)/("total" ss), and it would be 1 for a perfect fit. So this model does not fit very well (and to investigate this further in practice we would look at the residuals: note that
residual = observed value - fitted value.
It may be that the lack of fit is caused by one or two 'anomalous' cathedrals: this could be interesting if so, but the question does not provide us with the data on this point, so we can't tell.
ii) The parameter estimates are, say, $\hat{\mu}, \hat{\beta}, \hat{\theta}_2$ and these are
$44.298(81.648), 4.712(1.058), 80.393(32.306)$
(the corresponding standard error being given in brackets). To test, for example, whether
$\theta_2 = 0$ (ie no difference in the intercepts for the 2 Styles), we refer
$80.393/32.306 = 2.488$ to the t-distribution with 22 df. Note that 22= total number of observations - 3, and we are fitting 3 parameters in the model. Conveniently, R tells us the corresponding 2-tail probability: it is .0209, so well below .05, thus at level .05, we reject the hypothesis $\theta_2 = 0$.
The appropriate sketch graph is of y against x: two parallel lines, of slope 4.712, with the line for Romanesque above that of Gothic. (Don't ask me to explain this in terms of cathedral architecture, but someone must know why this is.)
iii) Our estimate of σ^2 is (residual ss)/df, and here this gives us
$\hat{\sigma} = 77.53$.

Part IIA, June 2004.
Paper 2, q12
(i) Suppose we have independent observations $Y_1, \ldots, Y_n$, and we assume that for $i = 1, \ldots, n$, Y_i is Poisson with mean μ_i, and $log(\mu_i) = \beta^T x_i$, where $x_1, \ldots, x_i$ are given covariate vectors each of dimension p, where β is an unknown vector of dimension p, and $p < n$. Assuming that $\{x_1, \ldots, x_n\}$ span R^p, find the equation for $\hat{\beta}$, the maximum likelihood estimator of β, and write down the large-sample distribution of $\hat{\beta}$.
(ii) A long-term agricultural experiment had 90 grassland plots, each $25m \times 25m$, differing in biomass, soil pH, and species richness (the count of species in the whole plot). While it was well-known that species richness declines with increasing biomass, it was not known how this relationship depends on soil pH, which for the given study has possible values 'low', 'medium' or 'high', each taken 30 times. Explain the commands input, and interpret the resulting output in the (slightly edited) R output below, in which 'species' represents the species count.
(The first and last 2 line of the data are reproduced here as an aid. You may assume that the factor pH has been correctly set up.)

```
> species
     pH    Biomass Species
1  high 0.46929722      30
2  high 1.73087043      39
  ......................
  ......................
89  low 4.36454121       7
90  low 4.87050789       3

>  summary(glm(Species ~Biomass, family = poisson))
Call:
glm(formula = Species ~ Biomass, family = poisson)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.184094   0.039159   81.31  < 2e-16
Biomass     -0.064441   0.009838   -6.55 5.74e-11

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 452.35  on 89  degrees of freedom
Residual deviance: 407.67  on 88  degrees of freedom

Number of Fisher Scoring iterations: 4


> summary(glm(Species ~pH*Biomass, family = poisson))
Call:
glm(formula = Species ~ pH * Biomass, family = poisson)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.76812    0.06153  61.240  < 2e-16
pHlow         -0.81557    0.10284  -7.931 2.18e-15
pHmid         -0.33146    0.09217  -3.596 0.000323
Biomass       -0.10713    0.01249  -8.577  < 2e-16
pHlow:Biomass -0.15503    0.04003  -3.873 0.000108
pHmid:Biomass -0.03189    0.02308  -1.382 0.166954

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 452.346  on 89  degrees of freedom
Residual deviance:  83.201  on 84  degrees of freedom

Number of Fisher Scoring iterations: 4
```