Introduction to Generalized Linear Modelling, Example Sheets 1, 2 and 3, with solutions

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

Contents

Preface

These three Examples Sheets, and their solutions, have been built up from 1992, when I first gave this lecture course, to 2005, when I gave it for the last time. I am grateful to many users, particularly undergraduates, for their comments and questions. There is some overlap between these examples and the examples in my lecture notes. An expanded version of these notes may be seen at `http://www.statslab.cam.ac.uk/~pat/All.pdf`

Chapter 1

Example Sheet 1

1.1 Example Sheet 1: questions

(This is meant to be a very easy sheet, to get you started.)

1. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson random variables with parameter μ.
Show that $\hat{\mu} = \Sigma X_i/n$, and $\mathrm{var}\,(\hat{\mu}) = \mu/n$.
What is

$$\mathbb{E}(-\frac{\partial^2 L}{\partial \mu^2})?$$

What is the exact distribution of $(n\hat{\mu})$? What is the asymptotic distribution of $\hat{\mu}$?
2. Suppose we have n independent trials, and the outcome of each trial is
Red with probability θ_1,
or White with probability θ_2,
or Blue with probability θ_3,
where $\theta_1 + \theta_2 + \theta_3 = 1$.
Let (X, Y, Z) be the total number of (Red, White, Blue) trials in the sequence of n; write $X = \sum_1^n X_i$, $Y = \sum_1^n Y_i$ for suitably defined (X_i, Y_i).
Find $\mathbb{E}(X)$, $\mathrm{var}\,(X)$, and show that

$$\mathrm{cov}\,(X, Y) = -n\theta_1\theta_2.$$

Find $\binom{\hat{\theta}_1}{\hat{\theta}_2}$, and find the mean vector, and covariance matrix, of its asymptotic distribution (which is of course bivariate normal).
3. Suppose Y_i independent Poisson, mean μ_i, and our model is

$$H : \log(\mu_i) = \alpha + \beta x_i$$

where (x_i) are given.
Write down the log likelihood $\log f(y|\alpha, \beta)$ and hence find

(i) the sufficient statistics for (α, β);
(ii) equations for $(\hat{\alpha}, \hat{\beta})$, the maximum likelihood estimator (mle), and
(iii) an expression for

$$\max_{\beta=0} f(y|\alpha, \beta).$$

Show how you would use this, together with Wilks' theorem, to test $H_0 : \beta = 0$.

4. Suppose

$$f(y_i|\beta) = (1/2) \exp -|y_i - \beta|, \ -\infty < y_i < \infty.$$

Let $L(\beta) = \sum_1^n \log f(y_i|\beta)$. Sketch this as a function of β, and hence find an expression for $\hat{\beta}$, the mle, given that the ordered observations are $y_{(1)} < \ldots < y_{(n)}$. What is $\frac{\partial L}{\partial \beta}$?

5. Suppose $f(y_i|\theta) = \theta e^{-\theta y_i}$, $y_i > 0$, $1 \le i \le n$.
Show that this is an exponential family form distribution, with natural parameter $\pi = -\theta$. Find the sufficient statistic and its distribution, and find the mle for each of π, θ.

6. Suppose $f(y_i|\alpha, \beta)$ is the pdf of $N(\alpha + \beta x_i, \sigma^2)$, where σ^2 is known, and $x_1, \ldots, x_n$ are known. Show that the loglikelihood function $L_n(\alpha, \beta)$ is a concave function of (α, β).

7. Suppose

$$f(y_i|\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp -y_i^2/2\theta,$$

for $i = 1, \ldots, n$. Show that the mle of θ is $\hat{\theta} = 1/n \sum_1^n y_i^2$, and that $\frac{n\hat{\theta}}{\theta} \sim \chi_n^2$. Find the exact mean and variance of $\hat{\theta}$, and compare these with the asymptotic mean and variance obtained from general likelihood theory.

8. Suppose $Y_i \sim Bi(1, \pi_i)$ independent, $(i = 1, \ldots, n)$ (i.e. $P(Y_i = 1) = \pi_i$, $P(Y_i = 0) = 1 - \pi_i$). Suppose also $\log(\pi_i/(1 - \pi_i)) = \alpha + \beta x_i$, where $x_1, \ldots, x_n$ are given.
Write down $L_n(\alpha, \beta)$, the loglikelihood function, and find equations for $(\hat{\alpha}, \hat{\beta})$ the mle of (α, β).

9. Suppose $Y_i \sim NID(\beta_1 + \beta_2 x_i + \beta_3 P_2(x_i), \sigma^2)$, $1 \le i \le n$, where $\Sigma x_i = 0$ and $P_2(x_i)$ is a given quadratic function of x_i such that

$$\Sigma P_2(x_i) = \Sigma x_i P_2(x_i) = 0.$$

Find

$$\frac{\partial \ell}{\partial \beta}, \ \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} \text{ and } \mathbb{E}\frac{\partial^2 \ell}{\partial \beta \partial \beta^{\mathrm{T}}}\ .$$

Find $\hat{\beta}_3$ and var $(\hat{\beta}_3)$, and show how you would test $H_0 : \beta_3 = 0$, when σ^2 is known.
If $P_2(x_i) = x_i^2 + a x_i + b$, find expressions for a, b in terms of (x_i). Can you see any advantages in fitting the model with

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i + \beta_3 P_2(x_i)$$

as above rather than the model written as

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i + \gamma x_i^2$$

say?
(The purpose of this question is to introduce you to ORTHOGONAL polynomials.)
10. Suppose $Y_i \sim Bi(n_i, p_i)$, $1 \le i \le k$, independent, and $g(p_i) = \alpha + \beta x_i$ where $g(\cdot)$ is a known link function. Write down

$$\frac{\partial \ell}{\partial \alpha}, \ \frac{\partial \ell}{\partial \beta} \ .$$

Write down the sufficient statistics for (α, β)
(i) if $g(p) = \log(p/(1-p))$ (link=logit)
(ii) if $g(p) = \log(-\log(1-p))$ (link= cloglog).

1.2 Solutions to Example Sheet 1

1. The likelihood is

$$f(x|\mu) = \Pi e^{-\mu} \ \mu^{x_i}/x_i!$$

giving the loglikelihood as

$$L = log(f(x|\mu)) = -n\mu + \Sigma x_i \ log(\mu) + constant.$$

Hence

$$\frac{\partial L}{\partial \mu} = -n + \Sigma x_i/\mu$$

and

$$\frac{\partial^2 L}{\partial \mu^2} = -\Sigma \ x_i/\mu^2 \ (\ < 0)$$

so that L is maximised at

$$\hat{\mu} = \Sigma x_i/n.$$

Clearly $\mathbb{E}(X_i) = \mu = var(X_i)$. Thus $\mathbb{E}(\hat{\mu}) = \mu$

and $var(\hat{\mu}) = \mu/n$, and

$$-\mathbb{E}\frac{\partial^2 L}{\partial \mu^2} = n/\mu.$$

The exact distribution of $\hat{\mu}n = \Sigma X_i$ is $Po(n\mu)$.
The asymptotic distribution of $\hat{\mu}$, by the Central Limit Theorem, is $N(\mu, \mu/n)$.

2. (This is the 3-cell multinomial distribution). With

$$f(x,y,z|\theta) = n!\frac{\theta_1^x\theta_2^y\theta_3^z}{x!y!z!}$$

for $x, y, z = 0, 1, 2...$, and $x+y+z=n$, we have $X = \Sigma X_i$ say, where X_i=1 if ith trial results in a Red, and X_i =0, otherwise, $Y = \Sigma Y_i$, and $Y_i = 1$ if ith trial results in a White, Y_i=0 otherwise.
Clearly, $P(X_i = 1) = \theta_1$, and X is Bi(n,θ_1)
so that var(X)= n$\theta_1(1-\theta_1)$, $\mathbb{E}(X) = n\theta_1$.
Further

$$cov(X,Y) = \Sigma cov(X_iY_i) = n(\mathbb{E}(X_1Y_1) - \mathbb{E}(X_1)\mathbb{E}(Y_1)).$$

Clearly, $\mathbb{E}(X_1Y_1) = 0$, so $cov(X,Y) = -n\theta_1\theta_2$.
Now

$$L = logf(x,y|\theta) = xlog(\theta_1) + ylog(\theta_2) + zlog(\theta_3) + \text{constant}$$

which is maximised subject to $\theta_1 + \theta_2 + \theta_3$ =1 (use a Lagrange multiplier)
by $\hat{\theta}_1 = x/n, \hat{\theta}_2 = y/n, \hat{\theta}_3 = z/n.$
Hence $\mathbb{E}(\hat{\theta}_i) = \theta_i$ for $i = 1, 2, 3$. Now

$$\frac{\partial L(\theta)}{\partial\theta_1} = (x/\theta_1) - (z/\theta_3)$$

$$\frac{\partial L(\theta)}{\partial\theta_2} = (y/\theta_2) - (z/\theta_3).$$

Hence minus the matrix of 2nd derivatives of L is

$$\begin{pmatrix} x/\theta_1^2 + z/\theta_3^2 & z/\theta_3^2 \\ z/\theta_3^2 & y/\theta_2^2 + z/\theta_3^2 \end{pmatrix}$$

Substituting for $\mathbb{E}(x), \mathbb{E}(y), \mathbb{E}(z)$, we see that the expectation of the above matrix is

$$\begin{pmatrix} n(1-\theta_2)/\theta_1\theta_3 & n/\theta_3 \\ n/\theta_3 & n(1-\theta_1)/\theta_2\theta_3 \end{pmatrix}.$$

It now remains for you to check that the inverse of this 2×2 matrix is

$$\begin{pmatrix} \theta_1(1-\theta_1)/n & -\theta_1\theta_2/n \\ -\theta_1\theta_2/n & \theta_2(1-\theta_2)/n \end{pmatrix}.$$

This is what the general formula for the **asymptotic** covariance matrix gives us. In this case, it agrees exactly with the **exact** covariance matrix.

3.(i)
With $f(y_i|\mu_i)$ proportional to $e^{-\mu_i}{\mu_i}^{y_i}$
and $\mu_i = exp(\alpha + \beta x_i)$ we see that the likelihood for (α, β) is proportional to

$$[exp - \Sigma e^{\alpha+\beta x_i}]exp\ [\alpha\ t_1 + \beta\ t_2]$$

where t_1 is defined as Σy_i , and t_2 as $\Sigma x_i y_i$.
Hence, by the factorisation theorem, (t_1, t_2) are sufficient for (α, β).
The log likelihood is

$$L(\alpha, \beta) = -\Sigma e^{\alpha+\beta x_i} + \alpha\ t_1 + \beta\ t_2 + constant.$$

ii)Thus

$$\frac{\partial L}{\partial \alpha} = 0 \text{ for } t_1 = \Sigma e^{\alpha+\beta x_i}$$

$$\frac{\partial L}{\partial \beta} = 0 \text{ for } t_2 = \Sigma x_i e^{\alpha+\beta x_i}.$$

These are the equations for $(\hat{\alpha}, \hat{\beta})$. To verify that this is indeed **the maximum**, we should check that **(minus the matrix of 2nd derivatives)** is positive- definite at $(\hat{\alpha}, \hat{\beta})$.
The equations for $(\hat{\alpha}, \hat{\beta})$ do not have an explicit solution, but we could solve them iteratively to find $(\hat{\alpha}, \hat{\beta})$, and hence we could evaluate the maximum of L.
iii) Now, if $\beta = 0, L(\alpha, \beta) = -\Sigma e^{\alpha} + \alpha\ t_1$. It is easily seen that this is maximised with respect to α by α^* say, where $\alpha^* = log(t_1/n)$.
We know, by Wilks' theorem,that to test $H_0 : \beta = 0$ against $H_1 : \beta$ arbitrary, we should refer

$$2[L(\hat{\alpha}, \hat{\beta}) - L(\alpha^*, 0)] \text{ to } {\chi^2}_1.$$

4.(An example of how things can be tricky when we are not in the glm family).

$$logf(y|\beta) = -\sum |y_i - \beta| + \text{constant} = -g(\beta) + \text{constant say.}$$

Defining $y_{(1)}, ..., y_{(n)}$ as the ordered sample values, as instructed in the question, we see that

$$g(\beta) = \sum (y_{(i)} - \beta) \text{ for } \beta < y_{(1)}$$

(this is a straight line of slope $-\ n$)

$$g(\beta) = -(y_{(1)} - \beta) + \sum_2^n (y_{(i)} - \beta) \text{ for } y_{(1)} < \beta < y_{(2)}$$

(this is a straight line of slope $-\ (n+2)$)
and so on....
Finally,

$$g(\beta) = \sum_1^n (\beta - y_{(i)}), \text{ for } \beta > y_{(n)}$$

This is a straight line of slope n. Thus, sketching $-g(\beta)$ we see that the log-likelihood function is concave in shape, consisting of straight line segments, and

if n is odd, say $n = 2m + 1$, then $\hat{\beta} = y_{(m+1)}$,
if n is even, say $n = 2m$, then $\hat{\beta}$ is anywhere between $y_{(m)}$ and $y_{(m+1)}$.
$\partial L/\partial \beta$ is of course not defined at $y_{(1)}, ..., y_{(n)}$.

Otherwise it is the slope of the appropriate linear segment.

In this example we could find the asymptotic distribution of $\hat{\beta}$ by going back to first principles (it would make for quite a tough exercise). But we CANNOT find it by quoting the general theorem for the asymptotic distribution of mle's: the appropriate regularity conditions do not hold.

5. $f(y|\theta) = \theta^n exp{-}\theta\ \Sigma y_i$
which is of exponential family form, with
$t(y) = \Sigma y_i$ as our sufficient statistic, having gamma(n, θ) distribution,
and loglikelihood L $= n\ log\theta - \theta\Sigma y_i$. Thus
$\partial L/\partial\theta = n/\theta - t(y)$.
Hence $\hat{\theta} = n/t(y)$, and $\hat{\pi} = -n/t(y)$.

6. $logf(y|\alpha, \beta) =\ -\Sigma(y_i - \alpha - \beta\ x_i)^2/2\sigma^2 + constant = L_n(\alpha, \beta)$ say.
Now find [minus the matrix of 2nd derivatives] : show that it is positive-definite.
Hence $L_n(\alpha, \beta)$ is a concave function.

7. $L_n(\theta) = -(n/2)log(\theta) - \Sigma y_i^2/(2\theta) + constant.$
Thus

$$\partial L/\partial\theta = -(n/2\theta) + \Sigma y_i^2/(2\theta^2)$$

Hence

$$\hat{\theta} = \Sigma y_i^2/n$$

and

$$n\hat{\theta}/\theta = \Sigma y_i^2/\theta,$$

where $y_i/\sqrt{\theta}$ are $NID(0, 1)$.
Hence $n\hat{\theta}/\theta$ is dist'd as χ^2_n hence has mean n, variance $2n$.
Thus $\mathbb{E}(\hat{\theta}) = \theta$, as we would hope.
Now,

$$\frac{\partial^2 L}{\partial\theta^2} = n/(2\theta^2) - \Sigma(y_i^2/\theta^3)$$

giving

$$\mathbb{E}(-\frac{\partial^2 L}{\partial\theta^2}) = n/(2\theta^2).$$

So the asymptotic variance of $\hat{\theta}$ is $(2\theta^2)/n$, which is the same as the exact variance.

8. With Y_i independent $Bi(1, \pi_i)$ as given, we see that

$$f(y|\alpha, \beta) = \Pi{\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}$$

Hence,

$$L_n(\alpha, \beta) = \Sigma y_i \ log(\pi_i/(1-\pi_i)) - \Sigma log(1-\pi_i)$$

Substituting for (π_i) gives

$$L_n(\alpha, \beta) = \Sigma(\alpha + \beta \ x_i)y_i - \Sigma log(1 + e^{\alpha+\beta \ x_i})$$

thus,

$$L_n(\alpha, \beta) = \alpha \ y_+ + \beta \ \Sigma x_i y_i - \Sigma log(1 + e^{\alpha+\beta \ x_i})$$

Hence we can write down the equations

$$\partial L_n/\partial\alpha = 0, \ \partial L_n/\partial\beta = 0.$$

Observe that there is no closed form solution to these equations. We can only find the mle's by an iterative solution.

9. With Y_i distributed as $NID(\beta_1 + \beta_2 x_i + \beta_3 P_2(x_i), \sigma^2)$
where $\Sigma x_i = 0, \Sigma P_2(x_i) = 0, \Sigma x_i P_2(x_i) = 0$
we see that the loglikelihood is $l(\beta)$ + constant, where

$$l(\beta) = -\Sigma(y_i - \beta_1 - \beta_2 \ x_i - \beta_3 \ P_2(x_i))^2/2\sigma^2.$$

This gives

$$\partial l/\partial\beta_1 = \Sigma(y_i - \beta_1 - \beta_2 \ x_i - \beta_3 \ P_2(x_i))/\sigma^2 = \Sigma(y_i - \beta_1)/\sigma^2.$$

Similarly,

$$\partial l/\partial\beta_2 = \Sigma x_i(y_i - \beta_2 x_i)/\sigma^2$$

$$\partial l/\partial\beta_3 = \Sigma P_2(x_i) \ (y_i - \beta_3 P_2(x_i))/\sigma^2.$$

Hence

$$\frac{\partial^2 l}{\partial\beta\partial\beta^T} = (-1/\sigma^2) \ diag(n, \Sigma x_i^2, \Sigma(P_2(x_i))^2)$$

$$= \mathbb{E}(\partial^2 l/\partial\beta\partial\beta^T).$$

Solving $\partial l/\partial\beta = 0$, gives $\hat{\beta}$, in particular

$$\hat{\beta}_3 = \Sigma y_i \ P_2(x_i)/\Sigma(P_2(x_i))^2.$$

Now, since $\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i + \beta_3 P_2(x_i)$, we can see that
$\mathbb{E}(\hat{\beta}_3) = \beta_3$.
Further, Y_i are NID, each with variance σ^2, hence
$\hat{\beta}_3$ has variance $\sigma^2/\Sigma(P_2(x_i))^2$, and is normally distributed.
We can test $H_0 : \beta_3 = 0$, by referring
$\hat{\beta}_3/\sqrt{(its \ variance)}$ to $N(0, 1)$.

Given $P_2(x_i) = {x_i}^2 + a\ x_i + b$, we find a, b by solving the pair of equations

$$\Sigma({x_i}^2 + a\ x_i + b) = 0, \Sigma({x_i}^2 + a\ x_i + b)x_i = 0$$

giving

$$\Sigma x_i^2 + n\ b = 0, \Sigma x_i^3 + a\ \Sigma x_i^2 = 0.$$

The advantage of parametrising the model in ths way is that the parameters $\beta_1, \beta_2, \beta_3$ are **orthogonal**. Thus for example, if we want to fit

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2\ x_i$$

we find that the least squares estimators for β_1, β_2 are the same as in the full model, ie the same as they were when we included the term β_3 .

10. We may write

$$l(\alpha, \beta) = \Sigma y_i log(p_i/(1 - p_i)) + \Sigma n_i log(1 - p_i)$$

from which we may find the expressions for the partial derivatives; for general link function $g(\ )$ these do not simplify.
i) The sufficient statistics for (α, β) for the logit link are

$$(\Sigma y_i, \Sigma y_i x_i).$$

The logit link is of course the **canonical** link for the binomial distribution.
ii) For the complementary log log link, there is no reduction in dimensionality from n for the sufficient statistics: we will still need the original data (y_i, x_i) to construct the likelihood function.

Chapter 2

Example Sheet 2

2.1 Example Sheet 2: questions

1. If Y_i are independent Poisson, means $\exp \beta^T x_i$, $1 \leq i \leq n$, how would you evaluate $\hat{\beta}$ and its asymptotic covariance matrix? What does 'scale parameter taken as 1.000' mean in the corresponding glm output?
2 * If the loglikelihood can be written

$$\ell(\beta) = (\beta^T t - \psi(\beta))/\phi \quad \text{where } \phi > 0,$$

and $t = t(\mathbf{y})$ is a p-dimensional vector, show that the covariance matrix of $t(\mathbf{y})$ is

$$\phi \left(\frac{+\partial^2 \psi}{\partial \beta \partial \beta^T} \right)$$

and hence that $\ell(\beta)$ is a strictly concave function of β. What is the practical application of this result in estimation of β? Illustrate your answer for either the binomial or the Poisson distribution.
3. We can say that an asymptotic 90 % confidence region for $\hat{\beta}$ is derived from

$$(\hat{\beta} - \beta)^T (V(\hat{\beta}))^{-1} (\hat{\beta} - \beta) \sim \chi^2_p \quad \text{(approximately)}.$$

Show that if $p = 2$, the resulting region is an ellipse, and finds its equation in the case where

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_2} \equiv 0\,.$$

Why is the remark $\int_c^\infty e^{-x} dx = e^{-c}$ relevant in this context?
4. In the least-squares fit of the model

$$y_{ij} = \mu + \theta_i + \epsilon_{ij},\ 1 \leq j \leq n_i, 1 \leq i \leq t$$

with $\theta_1 = 0$ (i.e. the glm constraint), and the usual assumption that ϵ_{ij} are $NID(0, \sigma^2)$, show that

$$\hat{\mu} = y_{1+}/n_1, \hat{\theta}_i = -y_{1+}/n_1 + y_{i+}/n_i \ (i \neq 1)$$

and

$$\text{var}\,(\hat{\theta}_i) = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_i} \right) \ (i \neq 1).$$

Find $\text{cov}\,(\hat{\theta}_i, \hat{\theta}_\ell)$ for $i \neq \ell$, $(i, \ell \neq 1)$. (Here y_{i+} is defined as $\sum_{j=1}^{n_i} y_{ij}$).
Show that the 'fitted value' of y_{ij} is y_{i+}/n_i, for $i = 1, \ldots, t$.
Hint: it's easier to do the algebra with the 'sum to zero' constraint, and then transform back to the 'corner-point' constraint'.
5. In the least-squares fit of the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},\ 1 \leq k \leq r, 1 \leq i \leq I, 1 \leq j \leq J$$

with $\alpha_1 = 0$, $\beta_1 = 0$, and the usual assumption that ϵ_{ijk} are $NID(0, \sigma^2)$, show that, for $i \neq 1$,

$$\hat{\alpha}_i = \frac{1}{Jr} \left(\sum_{j,k} y_{ijk} - \sum_{j,k} y_{1jk} \right),$$

with the corresponding expression for $\hat{\beta}_j$. How is your answer affected if the condition $1 \leq k \leq r$ is replaced by the condition $1 \leq k \leq r_{ij}$?
6. Why would you expect, in an experiment with IJ observations, that the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij},$$

where

$$\epsilon_{ij} \sim NID(0, \sigma^2),\ 1 \leq i \leq I, 1 \leq j \leq J$$

gives a perfect fit to the data (i.e. zero deviance)?
7. In the usual model

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

with $\epsilon_i \sim NID(0, \sigma^2)$ explain why displaying the estimated covariance matrix of $\hat{\beta}$ is a method of finding out about $(X^T X)^{-1}$.
8. The data-set below is taken from the Minitab student Handbook(1976) by Ryan,T., Joiner, B. and Ryan, B. and is also discussed in the book by Aitkin et al. (See Chapter 3 of the book.) For the 31 cherry trees, the table below shows d, h, v. These are defined by d is the diameter (in inches), at a height of 4.5 feet from the ground, h is the height (in feet) of the trees, v is the volume of useable wood, in cubic feet.

Reminder: 1 foot= 12 inches.

This is one of the datasets already in R: try

```
data(trees); attach(trees); trees[1,]
```

(but you need to take 'Girth' as d, which is confusing, I know).
The order is (d, h, v).

```
 8.3  70 10.3,    8.6 65 10.3,    8.8 63 10.2,    10.5 72 16.4,
 10.7 81 18.8,    10.8 83 19.7,   11.0 66 15.6,   11.0 75 18.2,
 11.1 80 22.6,    11.2 75 19.9,   11.3 79 24.2,   11.4 76 21.0,
 11.4 76 21.4,    11.7 69 21.3,    12.0 75 19.1,    12.9 74 22.2,
 12.9 85 33.8,    13.3 86 27.4,    13.7 71 25.7,    13.8 64 24.9,
 14.0 78 34.5,    14.2 80 31.7,    14.5 74 36.3,    16.0 72 38.3,
 16.3 77 42.6,    17.3 81 55.4,    17.5 82 55.7,    17.9 80 58.3,
 18.0 80 51.5,    18.0 80 51.0,    20.6 87 77.0.
```

Take $lv = log(v)$, with ld, lh defined similarly.
Verify that the following models give the estimates (with se's) and deviances below, and discuss the fit of these models.

$$M_0 : \mathbb{E}(lv) = \mu$$

for which

$$\hat{\mu} = 3.273(0.0945)$$

and deviance $= 8.3087(df = 30)$;

$$M_1 : \mathbb{E}(lv) = \mu + \beta\ ld$$

for which

$$\hat{\mu} = -2.353(0.2307), \hat{\beta} = 2.200(0.0898)$$

and deviance$= 0.38324(df = 29)$;

$$M_2 : \mathbb{E}(lv) = \mu + \beta\ ld + \gamma\ lh$$

for which

$$\hat{\mu} = -6.632(0.7998), \hat{\beta} = 1.983(0.0750), \hat{\gamma} = 1.117(0.2044)$$

and deviance$= 0.18546(df = 28)$;

$$M_3 : \mathbb{E}(lv) = \mu + \gamma\ lh$$

for which

$$\hat{\mu} = -13.96(3.755), \hat{\gamma} = 3.982(0.8677)$$

and deviance$= 4.8130(df = 29)$.
What are the numerical consequences of the non-orthogonality of the parameters β, γ?
The volume of a cylinder of length ℓ, diameter d, is $(\pi d^2 \ell)/4$, and the volume of a cone of height ℓ and base diameter d is $(\pi d^2 \ell)/12$. Are these cherry trees more like cylinders than cones?
9. 'The Independent' (21/12/88) gave the 'League Table of football-related arrests', printed in the table below. This list details a total of 6147 football-related arrests in the 1987-8 season, and is compiled by the Association of Chief Constables. It does not differentiate between Home-fans and Away-fans.
There are 4 Divisions (with Division 1 containing the best clubs) and these have 21, 23, 24 and 24 clubs respectively, each Division corresponding to a pair of columns in the Table below.

The columns below give (a, n) where
$a=$ attendance, in thousands, and
$n=$ number of arrests,
for each of the 4 soccer divisions, with the order (reading across the rows) being (a, n) for Division 1, (a, n) for Division 2, (a, n) for Division 3, (a, n) for Division 4.
YOU ARE NOT INTENDED TO TYPE IN THIS DATA: ASK ME TO EMAIL THE SET TO YOU.

a1	n1	a2	n2	a3	n3	a4	n4
325	282	404	308	116	99	71	145
409	271	286	197	401	80	227	132
291	208	443	184	105	72	145	90
350	194	169	149	77	66	56	83
598	153	222	132	63	62	77	53
420	149	150	126	145	50	74	46
396	149	321	110	84	47	102	43
385	130	189	101	128	47	39	38
219	105	258	99	71	39	40	35
266	91	223	81	97	36	45	32
396	90	211	79	205	34	53	29
343	86	215	78	106	32	51	28
518	74	108	68	43	28	51	27
160	49	210	67	59	22	115	21
291	43	224	60	88	22	52	21
783	38	211	57	226	21	67	21
792	33	168	55	61	21	52	17
314	32	185	44	91	21	52	17
556	24	158	38	140	20	72	15
174	14	429	35	85	18	49	12
162	1	226	29	127	11	101	10
NA	NA	150	20	59	5	90	8
NA	NA	148	19	87	4	50	5
NA	NA	NA	NA	79	3	41	0

Let (n_{ij}, a_{ij}) be the observations for the ith club of the jth Division, for $j = 1, ..., 4$.
Making the standard assumption that the errors are $NID(0, \sigma^2)$ consider how to fit the following models in R or Splus, and sketch graphs to show what these models represent:
a) $\mathbb{E}(n_{ij}) = \mu$,
b) $\mathbb{E}(n_{ij}) = \mu + \alpha\ a_{ij}$,
c) $\mathbb{E}(n_{ij}) = \mu + \beta_j + \alpha\ a_{ij}$,
d) $\mathbb{E}(n_{ij}) = \mu + \beta_j + \alpha_j\ a_{ij}$.
You will find that the deviances for these 4 models are, respectively
$371056(91 df),\ 296672(90 df),\ 272973(87 df),\ 241225(84 df)$.
Now plot n against a. What do you conclude about your assumption of constant

error variances?
Now repeat the model-fitting exercise with n, a replaced by $log(n), log(a)$.
Can you now think of a way of identifying certain clubs as 'rogues', or indeed as 'saints' within their Division?
10. The data-set below which is also discussed by Agresti (1995, p101) is based on a study of British doctors by R.Doll and A.B.Hill(1966) and gives the number of coronary deaths for smokers and non-smokers, for each of 5 different age-groups, with the corresponding 'person-years', ie the total time at risk. Thus for example, in the youngest age-group, in the non-smoking category, the total time at risk was 18793 years, and during this time there were 2 coronary deaths in this particular class. The age-groups are 35-44, 45-54, 55-64, 65-74, 75-84 years.
Define d_{ij} as the number of deaths in age group i, smoking group j, where $i = 1, \ldots, 5, j = 1, 2$ with $j = 1$ for non-smokers, $j = 2$ for smokers. Assume that (d_{ij}) are distributed as independent Poisson variables, with $\mathbb{E}(d_{ij}) = \mu_{ij}$ and $\mu_{ij} = \theta_{ij} p_{ij}$ where p_{ij} = total person-years at risk, for age i, smoking group j. Hence $log(\mu_{ij}) = log(\theta_{ij}) + log(p_{ij})$ for all i, j. This is why we take $log(p_{ij})$ as the 'offset' in the glm analysis below. We model $log(\theta_{ij})$, the parameter of interest.
Verify and interpret the results from the models fitted below.

```
 pyears <-  scan()
18793  52407
10673  43248
 5710  28612
 2585  12663
 1462   5317
          # BLANK LINE
deaths  <-  scan()
2 32
12 104
28 206
28 186
31 102
          # BLANK LINE
sm  <-rep(c(1,2),times=5)
age <- c(1,1,2,2,3,3,4,4,5,5)
#these are crude but easy-to-understand ways of setting up the factor levels
sm  <- factor(sm)  ;  age  <-  factor(age)
prop <- deaths/pyears ; tapply(prop,list(age,sm),mean)   <- log(pyears)
summary(glm(deaths ~  age + offset(l),poisson),cor=F)
summary(glm(deaths  ~ age + sm + offset(l),poisson),cor=F)
```

It should now be obvious to you that
(i) smoking is bad for you and
(ii) so is getting old.
But you will also see that this final model does not fit well (its deviance is

12.134, with 4 df). Can you suggest a way of improving the fit?
11. With $y_1, \ldots, y_n$ independent observations, and y_i having pdf from the usual glm family, with $g(\mu_i) = \beta^T x_i$ as usual, find the expectation of the second derivative of the log-likelihod function with respect to β, ϕ, and hence show that $\hat{\beta}, \hat{\phi}$ are asymptotically independent.
12. New for 2005: introduction to the inverse Gaussian distribution.
Consider the density

$$f(y|\theta, \phi) = exp[(y\theta - b(\theta))/\phi + c(y, \phi)], \quad y > 0,$$

where

$$\phi = \sigma^2, b(\theta) = -(-2\theta)^{1/2} \quad ,$$

and

$$-2c(y, \phi) = log(2\pi\phi y^3) + (1/\phi y).$$

Show that $\mathbb{E}(Y) = (-2\theta)^{-1/2} = \mu$, say. Check that $1/\mu^2$ is the canonical link function for this glm, and that $var(Y) = \mu^3\sigma^2$. **nb, no fancy integration required, at all.**

2.2 Solutions to Example Sheet 2

1. Here Y_i are distributed as independent $Po(\mu_i)$, with $\log(\mu_i) = \beta^T x_i$. Thus

$$f(y_i|\beta) \propto (\exp -\mu_i)\ (\mu_i)^{y_i}$$

so that

$$\log f(y|\beta) = \sum_1^n [-\exp(\beta^T x_i) + y_i \beta^T x_i] + constant$$

ie $L_n = -\sum_1^n [\exp(\beta^T x_i)] + \beta^T \sum_1^n y_i x_i + constant$.
Thus

$$\frac{\partial L_n}{\partial \beta} = \sum_i^n [-x_i \exp(\beta^T x_i) + y_i x_i]$$

and minus the matrix of second derivatives of L_n is say J, where

$$J = +\sum_1^n x_i x_i^T \exp(\beta^T x_i).$$

This is $+\sum_1^n x_i x_i^T \mu_i$, where $\mu_i > 0$.
Hence J is a positive definite matrix, and so $\hat{\beta}$ (the mle) is the solution of

$$\frac{\partial L_n}{\partial \beta} = 0,$$

which we may write as

$$\sum x_i \mu_i = \sum x_i y_i$$

ie the observed and the expected values of the sufficient statistics $\sum x_i y_i$ agree exactly at $\beta = \hat{\beta}$.
The equation $\frac{\partial L_n}{\partial \beta} = 0$ has to be solved iteratively. The asymptotic covariance matrix of $\hat{\beta}$ is the inverse of $\mathbb{E}(J)$, which here is

$$\sum_1^n \mu_i x_i x_i^T.$$

(There is no simple formula for the inverse.)
"Scale parameter taken as 1.000 " means that in writing the pdf in glm formulation, ie as

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

then we take ϕ as 1.

2. The loglikelihood is

$$l(\beta) = (\beta^T t - \psi(\beta))/\phi \text{ where } \phi > 0.$$

Thus

$$\frac{\partial}{\partial \beta} l(\beta) = (t - \frac{\partial}{\partial \beta}\psi(\beta))/\phi$$

and as always

$$\mathbb{E}(\frac{\partial l}{\partial \beta}) = 0.$$

so that $\mathbb{E}(t) = \frac{\partial}{\partial \beta}\psi(\beta)$. Further the matrix of 2nd derivatives of the loglikelihood is - ϕ^{-1} times the matrix of 2nd derivatives of ψ and, using

$$\int f(y|\beta)dy = 1 \ for \ all \ \beta$$

again, we have

$$\mathbb{E}(-\text{matrix of 2nd derivatives of } l) = \mathbb{E}(UU^T), \ where \ U = \frac{\partial}{\partial \beta} l$$

Thus covariance matrix of t(y) is

$$\phi \times \ matrix \ of \ 2nd \ derivatives \ of \ \psi$$

which must therefore be a positive definite matrix, so that $l(\beta)$ is a STRICTLY CONCAVE function. Thus any solution $\hat{\beta}$ say, of $\frac{\partial}{\partial \beta} l = 0$ must be THE MAXIMUM of $l(\beta)$. Thus, eg if we solve $U = 0$ by the Newton-Raphson algorithm, we will unerringly home in on the right solution, rather than something nasty and irrelevant like a local minimum.

3. We know (using $\sim$ to mean “approximately distributed as”)

$$\hat{\beta} \sim N(\beta, V(\beta))$$

and so

$$(\hat{\beta}-\beta)^T(V(\beta))^{-1}(\hat{\beta}-\beta) \sim \chi_p^2$$

and hence

$$(\hat{\beta}-\beta)^T(V(\hat{\beta}))^{-1}(\hat{\beta}-\beta) \sim \chi_p^2.$$

Hence from χ^2 tables we can choose c, given α, such that

$$Pr[(\hat{\beta}-\beta)^T(V(\hat{\beta}))^{-1}(\hat{\beta}-\beta) \leq c] \simeq 1-\alpha.$$

Write $V = V(\hat{\beta}))$ for short; our $(1-\alpha)-$ confidence region is

$$(\hat{\beta}-\beta)V^{-1}(\hat{\beta}-\beta) \leq c$$

which (because V^{-1} is positive-definite) is an ELLIPSE, centred on $\hat{\beta}$. For

$$\frac{\partial^2 \ell}{\partial\beta_1\partial\beta_2} = 0$$

we have V a diagonal matrix, with diagonal entries v_1, v_2 say and the ellipse is

$$(\beta_1-\hat{\beta}_1)^2/v_1 + (\beta_2-\hat{\beta}_2)^2/v_2 = c.$$

The relevance of the final remark is that χ_p^2 has, for $p=2$ the probability density function

$$(1/2)exp(-x/2),\ x > 0.$$

4. Given

$$y_{ij} = \mu + \theta_i + \epsilon_{ij}\ ,$$

for $i = 1 \ldots t,\ j = 1 \ldots n_i$ with $\theta_1 = 0$
equivalently

$$y_{ij} = \mu_i + \epsilon_{ij}$$

with $\mu_1 = \mu$, and $\mu_i = \mu + \theta_i$ for $i > 1$. We see that

$$\Sigma\Sigma(y_{ij} - \mu - \theta_i)^2$$

is minimised, equivalently $\Sigma\Sigma(y_{ij} - \mu_i)^2$ is minimised with respect to (μ_i) by

$$\hat{\mu}_i = \frac{y_{i+}}{n_i}.$$

giving

$$\hat{\mu} = \frac{y_{1+}}{n_1} \text{ and}$$

$$\hat{\theta}_i = -\frac{y_{1+}}{n_1} + \frac{y_{i+}}{n_i} \text{ for } i > 1,$$

as required. Clearly, since ϵ_{ij} are NID(0,σ^2),

$$var(\frac{y_{i+}}{n_i}) = \frac{\sigma^2}{n_i}$$

and

$$cov(y_{1+}, y_{i+}) = 0 \ for \ i > 1$$

hence $var(\hat{\theta}_i)$ is as given, and

$$cov(\hat{\theta}_i, \hat{\theta}_l) = var(\frac{y_{1+}}{n_1}) \ for \ i \neq l.$$

The fitted value of y_{ij} is $\hat{\mu}_i$, ie y_{i+}/n_i.

5.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, k = 1 \ldots r, \ i = 1 \ldots I, j = 1, \ldots J.$$

We can reparametrise this as

$$y_{ijk} = m + a_i + b_j + \epsilon_{ijk} \ ,$$

with $\Sigma a_i = 0, \Sigma b_j = 0$. Then $\mu = m + a_1 + b_1$, $\mu + \alpha_i = m + a_i + b_1$, $\mu + \beta_j = m + a_1 + b_j$, so that
$\alpha_i = a_i - a_1 \ , and \ \beta_j = b_j - b_1$.
Straightforward minimisation of

$$\Sigma\Sigma\Sigma(y_{ijk} - m - a_i - b_j)^2$$

subject to the constraints $\Sigma a_i = 0 \ , \Sigma b_j = 0$ gives

$$\hat{m} = \bar{y}, \hat{a}_i = \frac{y_{i++}}{n_i} - \bar{y} \ , etc.$$

Hence, returning to the original model, we see that

$$\hat{\alpha}_i = \frac{y_{i++}}{n_i} - \frac{y_{1++}}{n_1} \ .$$

If now $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, with

$$k = 1 \ldots r_{ij}, \ i = 1, \ldots, I, \ j = 1, \ldots, J$$

then we find that $\hat{\alpha}_i$ is no longer given by the above simple formula .
This is best seen by the following simple example. Suppose the observations (y_{ijk}) are
23.9 , 7.2 for $(i, j) = (1, 1)$,

3.6 for $(i, j) = (1, 2)$,
10.4 for $(i, j) = (2, 1)$,
29.7 for $(i, j) = (2, 2)$.

Note that you will different estimates for, say α_i, depending on whether or not β_j is in the model, because the design is unbalanced.

6. Here the model is

$$y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij}$$

for $i = 1 \ldots I, j = 1 \ldots J$, which is equivalent to

$$y_{ij} = \mu_{ij} + \epsilon_{ij},$$

and the lse(=mle) of (μ_{ij}) are obtained by minimising

$$\Sigma\Sigma(y_{ij} - \mu_{ij})^2$$

with respect to μ_{ij}. This gives $\mu_{ij} = y_{ij}$ so that the minimised residual sum of squares is 0, trivially.

Further, we have no degrees of freedom left to estimate σ^2. This is called a "saturated" model: it is saturated with parameters.

7.

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

with ϵ distributed as $N(0, \sigma^2 I)$, gives

$$(Y - X\beta)^T(Y - X\beta)$$

is minimised by $\hat{\beta}$ such that $X^T X\hat{\beta} = X^T Y$

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

Hence $\hat{\beta}$ is distributed as $N(0, \sigma^2(X^T X)^{-1})$.

If we display the the estimated covariance matrix of $\hat{\beta}$ it is of course

$$s^2(X^T X)^{-1}$$

where $s^2 = deviance/df$. Since s^2 is therefore known, we can evaluate $(X^T X)^{-1}$.

8. Using an obvious notation, you need to try

```
lm(lv ~ 1) ; lm(lv ~ lh) ; lm(lv ~ ld) ; lm(lv~lh + ld)
```

Observe that we never expected M_0 to fit: we always start by fitting it as our "baseline". Observe also that M_1 fits much better than M_0, as we would expect, and that $\hat{\beta}$ is clearly significant. (The ratio 2.2/.0898 is much greater than 2).

Observe that the estimate of β changes between M_1 and M_2 because β and γ are not mutually orthogonal. [There is a non-zero correlation between the vectors (ld),(lh)].

The coefficients $\hat{\beta}$ and$\hat{\gamma}$ in M_2 are clearly both significant.

Observe: the reduction in deviance in moving from M_0 to M_1 the "ss due to γ" is different from the reduction in deviance in moving from M_1 to M_2 , the "ss due to γ, allowing for β ".
This is another consequence of the non-orthogonality of these 2 parameters.
Our final model is thus

$$lv_i = -6.632(.7998) + 1.117(.2044)\ lh_i + \ 1.983(.07501)\ ld_i$$

Observe that both a cylinder & a cone would give

$$lv_i = C + 1\ log(h_i) + 2\ log(d_i/12)$$

where $C = log(\pi/4)$ for a cylinder, $log(\pi/12)$ for a cone. [warning : d_i is actually measured in inches. Do today's metric students know about feet and inches?]
The rest of the solution is up to you. Not surprisingly, trees do turn out to be more like cones than cylinders.
9. Note that the linear regression of n, the number of arrests on a, the attendance, although having a significant positive slope (as we would expect) is really rather a poor fit: it has R^2, the proportion of the total deviance "explained by" the regression as only 0.2005. (By definition, R^2 lies between 0 and 1, with R^2=1 for a perfect fit.)
If we plot the $residuals = (observed - fitted\ values)$ against the corresponding $fitted\ values$, we get a very pronounced 'fanning-out' effect, suggesting that $var(n_i)$ increases as $\mathbb{E}(n_i)$ increases, which is what we would expect, since n_i might well be Poisson-like in its behaviour. Thus the analysis presented, which fails to START by doing some simple plots, is not very sensible. Its big drawback is that it relies on the assumption that $var(n_i)$ is constant over i.
This is what the analysis does. First fit

$$n_{ij} = \mu + \epsilon_{ij},\ for\ j = 1 \ldots 4, i = 1 \ldots n_j\ .$$

with the usual assumption that ϵ_{ij} is $NID(0, \sigma^2)$.
Then fit

$$n_{ij} = \mu + \alpha\ a_{ij} + \epsilon_{ij}$$

ie the same line for all 4 divisions. Then fit

$$n_{ij} = \mu + \beta_j + \alpha\ a_{ij} + \epsilon_{ij}$$

ie parallel lines for all 4 divisions. Lastly try

$$n_{ij} = \mu + \beta_j + \alpha_j\ a_{ij} + \epsilon_{ij}.$$

There isn't any appreciable improvement in fit after the second model.
And, as we have already said, the assumption of homogeneity of error variance looks very implausible.
However, if we try

$$log(n_{ij}) = \mu + \alpha\ log(a_{ij}) + \epsilon_{ij}$$

it does now seem reasonable to assume $var(\epsilon_{ij})$ is constant. It seems that the same line will fit all 4 divisions, namely

$$log(n_{ij}) = -0.01365(.6475) + \ 0.7506(.1285) log(a_{ij})$$

which has $R^2 = .2748$ (just a little better than before). Indeed the constant can be dropped, to show that a straight line through the origin fits quite well.
The question about 'rogues and saints' clearly relates to picking out those clubs with high (positive or negative) standardised residuals.
10. Let p_{ij} ='pyears' = person-years at risk, and let d_{ij}= the number of deaths, for $i = 1, \dots, 5$, where i corresponds to age, $j = 1$(for non-smoker), $j = 2$ (for smoker).
We assume d_{ij} is distributed as indep. Po(μ_{ij}), with the default link for μ, which is log( ). Further, we take $\mu_{ij} = \theta_{ij}\ p_{ij}$, so that

$$log(\mu_{ij}) = log(\theta_{ij}) + log(p_{ij}).$$

This is the interpretation of $log(p_{ij})$ as the *offset*. We try various models for $log(\theta_{ij})$.
The table suggests that θ_{ij} increases with i, for each fixed j.
The first model fitted is in effect

$$log(\theta_{ij}) = \mu + \ \alpha_i, \text{for } i = 1 \dots 5, j = 1, 2$$

with α_1 =0. This model includes an age effect but not a smoking effect. This model has deviance= 23.99($df = 5$), but the fit is not good enough.
But clearly this model shows that $\hat{\alpha}_i$ increases with i: this makes sense.
Next we try the model

$$log(\theta_{ij}) = \mu + \alpha_i + \beta_j.$$

In this model, both smoking and age effects are present, but they operate additively (thus the difference between smokers and non-smokers is constant over all 5 ages). The fit is now much improved, and shows a clear effect of smoking($\hat{\beta}_2/(its\ se) > 0$), but the fit is still not acceptable (refer 12.13 to ${\chi^2}_4$).
The model corresponding to

```
glm(deaths ~ age:sm + offset(l), poisson)
```

would give a perfect fit (since it is the saturated model) so therefore to improve the fit, we include an interaction term in a 'weaker' way as our final model. We have
$log(\theta_{ij}) = \mu + \alpha_i$ for sm =1, ie non-smoker,
$log(\theta_{ij}) = \mu + \alpha_i + \beta_j + \gamma\ i$ for sm=2, ie smoker.
This model fits very well (compare the deviance of 1.54 with ${\chi^2}_3$). It shows that although smoking is bad for you (compute 1.445/.3729), there is an interaction between smoking and age: the discrepancy between the mortality rates for

smokers and non-smokers DECLINES with age.
11. Write down $\ell(\beta, \phi)$, and find

$$\frac{\partial^2 \ell}{\partial \beta \partial \phi}.$$

It is easy to show that this has expectation zero.
12. Use the glm facts that $\mathbb{E}(Y) = b'(\theta) = \mu$ and $var(Y) = \phi b''(\theta)$. (Note added November 2007: Wikipedia is also aware of this example.)

Chapter 3

Example Sheet 3

3.1 Example Sheet 3: questions

(Warning: somehow most people find questions 2, 3 to be hard. There is a practical point behind these 2 questions – and the others too!)

1. The observed waiting times $t_1, \ldots, t_n$ are independent, with T_i having pdf

$$f(t_i \mid \alpha, \beta) = (t_i/\mu_i)^{\nu-1}\, e^{-t_i/\mu_i} \frac{1}{\mu_i} \frac{1}{\Gamma(\nu)},$$

for $t_i > 0$, where ν is a known parameter, and μ_i depends linearly on a known covariate x_i through the following link function:

$$\frac{1}{\mu_i} = \alpha + \beta x_i, \text{ for } \mu_{\mathrm{i}} > 0.$$

Let $a_1 = \Sigma\theta_i, a_2 = \Sigma x_i\theta_i$. Show that (a_1, a_2) are sufficient for (α, β).
The observations $t_1, \ldots, t_m$ are times between consecutive earthquakes in Mexico City, and the observations $t_{m+1}, \ldots, t_n$ are the times between consecutive earthquakes in Turkey. Take $x_1 = \ldots = x_m = 0$ and $x_{m+1} = \ldots = x_n = 1$, and discuss the estimation of β when m, n are large, quoting any general asymptotic likelihood results needed for your solution.
This introduces you to another 'error distribution' available in glm: the gamma.
2. Your client gives you data

$$(y_{ij}, x_{ij}, 1 \leq j \leq n_i, 1 \leq i \leq t)$$

and asks you to fit the model

$$y_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij},$$

with $\epsilon_{ij} \sim NID(0, \sigma^2)$, σ^2 unknown. He has arranged the experiment so that

$$x_{ij} = x_i,\ 1 \leq j \leq n_i, 1 \leq i \leq t.$$

Find expressions for $(\hat{\alpha}, \hat{\beta})$, the least squares estimators.
Your client now observes that a consequence of the model above is that

$$\mathbb{E}(\overline{y}_i) = \alpha + \beta x_i,\ 1 \leq i \leq t,$$

where $\overline{y}_i = \sum_j y_{ij}/n_i$).
He suggests that some of your (highly paid) time could be saved by reading in the data as the t pairs $(\overline{y}_i, x_i)$, $1 \leq i \leq t$ instead of the original $(n_1 + \cdots + n_t)$ pairs of points. How do you advise him? Give reasons for your answer.

[Hint: write down the likelihood given by
a) the full set of data (y_{ij}, x_{ij}), and
b) the reduced set of data $(\overline{y}_i, x_i)$.
Show that the maximum likelihood estimates of (α, β) are the same for a) and for b).]

3. Another client gives you data for binary regression, consisting of observations (y, x) with the following structure:

$$(0, x_1), (1, x_1), (0, x_1), (1, x_2), (1, x_2),$$

$$(1, x_2), (0, x_2), (1, x_3), (1, x_3), (0, x_3),$$

$$(1, x_4), (1, x_4), (1, x_4), (1, x_4), (0, x_4).$$

Thus there are 15 independent observations, with the first digit of each pair being 1 or 0 with probabilities $p(x), q(x)$ respectively. She asks you to fit the model:

$$\log(p(x)/q(x)) = \alpha + \beta x$$

and to use the appropriate difference in deviances to test the hypothesis $\beta = 0$.
You observe that the data can in fact be compressed and read in as four independent values

$$(1, 3, x_1), (3, 4, x_2), (2, 3, x_3), (4, 5, x_4).$$

(eg $(1, 3, x_1)$ means that of the 3 readings at $x = x_1$, exactly 1 has value 1.)
Would this approach result in misleading your client? Give reasons for your answer.

[Hint: think LIKELIHOODS, as in question 2 above.]

4. Suppose

$$Y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I)$$

where Y is $n \times 1$, X is $n \times p$ and of rank p, and β is the unknown vector of parameters.
(a) Show that $\hat{\beta} = (X^T X)^{-1} X^T Y$ is the lse of β.

(b) Show that $\hat{Y} = X\hat{\beta} = HY$ say, where $H = H^T$ and $HH = H$ ($\hat{Y}$ is the vector of 'fitted' values).
(c) Suppose $e = Y - \hat{Y}$ (the vector of *residuals*). Show that

$$e \sim N(0, \sigma^2(I - H))$$

and hence that

$$e_i \sim N(0, \sigma^2(1 - h_i)),\ 1 \leq i \leq n$$

where h_i is the i^{th} diagonal element of H.
(d) Show that if λ is an eigenvalue of H, then λ is either 1 or 0. Hence show $\sum_1^n h_i = p$. (The quantity h_i is called the 'influence' or 'leverage' of the i^{th} data point x_i where $y_i = \beta^T x_i + \epsilon_i,\ \ 1 \leq i \leq n$.)
(e) Show that $0 \leq h_i \leq 1$, and find h_i for simple linear regression, i.e. for

$$y_i = \beta_1 + \beta_2(x_i - \overline{x}) + \epsilon_i.$$

(f) How do you interpret the statement "Values of x_i corresponding to large leverage exert a pronounced effect on the fit of the linear model at (x_i, y_i)"?
(g) Use R to construct leverages (also called 'influence values') and qqplots for simple (y, x) data-sets.
5. (a) Suppose $(y_i) \sim Mn(n, (p_i))$, $\sum_1^k p_i = 1$. Consider testing

$$H_0 : \log p_i = \mu + \beta x_i,\ 1 \leq i \leq k$$

where (x_i) is given, and μ is such that $\Sigma p_i = 1$, thus

$$e^\mu = 1/\Sigma e^{\beta x_i}.$$

Show that $\hat{\beta}$ is the solution of

$$(*) \qquad \Sigma y_i x_i = n\Sigma x_j e^{\beta x_j}/\Sigma e^{\beta x_j}.$$

Let $e_i = np_i(\hat{\beta})$ ('expected values under the null hypothesis').
(b) Now suppose instead that

$$y_i \sim \text{independent Poisson}(\mu_\text{i}), \text{with } \mu_\text{i} \geq 0.$$

Consider testing

$$HP_0 : \log \mu_i = \mu' + \beta x_i,\ 1 \leq i \leq k.$$

Find $\ell(\beta, \mu')$ the loglikelihood function, and hence show that the mle for β is given by $\hat{\beta}$, as in equation $*$ above. Show also that

$$\sum_1^k \mu_i(\hat{\beta}, \hat{\mu}') = y_+$$

(i.e. the observed and expected values of y_+ agree exactly at the mle). Comment on the glm application of the Poisson distribution for problem (a).

S=1	D=1	89	2	4	1
S=1	D=2	8	4	3	1
S=2	D=1	70	6	2	0
S=2	D=2	1	0	1	1
		A=1	A=2	A=1	A=2
		B=1	B=1	B=2	B=2

Table 3.1: A 4-way contingency table

6. A random sample of 193 individuals, classified according to four 2-level factors, S, D, A, B respectively, gave the following $2 \times 2 \times 2 \times 2$ contingency table as Table 3.1.
Let $p_{ijk\ell}$ denote the corresponding underlying cell probabilities where i, j, k, ℓ correspond to the factors S, D, B and A respectively. With the Poisson distribution for n, the cell frequencies, and log link, glm (in S-Plus) finds, in the following order,
(a)$n \sim S * D * A * B$ gives deviance = 0, df = 0,
(b)$n \sim (S + D + A + B) \wedge 3$ gives deviance 2.72, df = 1,
(c) $n \sim (S + D + A + B) \wedge 3 - S : A : B - D : A : B$ gives deviance 3.42, df = 3,
(d) $n \sim (S + A + B) * D$ gives deviance 8.48, df = 8.
Write down the model for $p_{ijk\ell}$ for each of (a), (b), (c) and (d), give an interpretation of the deviance at each stage, and give an interpretation in terms of conditional independence for the model in (d).
How would you check the fit of the model S, D, A, B mutually independent?
7. *Which?* (August 1980, p. 436) gives the data in Table 3.2 on lager (available to you on catam stats). The columns are price per half pint, o.g. (original gravity), percent alcohol, calories per half pint, and 'experts' rating'. The original gravity is described by *Which?* as "another guide to strength; it's a measure of what has gone into the beer besides water, and is used to calculate the duty payable". The 'experts' rating' column contains entries 3, 4, 5 for lagers actually tasted (5 being the most liked and 3 the least liked), and an entry 0 for lagers not actually tasted. Reading the data provided on file, use

```
lm()
```

to answer the following questions:
(a) Does the price depend on o.g., percent alcohol, and calories per half pint, and if so, how?
(b) Is the price of those lagers tasted significantly different from the price of those not tasted?

8. In a study on the possible relationship between the psychological well-being of mothers and that of their children, the psychiatrist Prof I. Goodyer collected data, some of which is summarised below. Table 3.3 shows, for the 200 children in the study, how many in each of eight categories were 'cases', i.e. anxious or

Price	og	percent	cal	rating
17.5	1031	3.1	76	3
20.5	1032	3.2	79	3
22.5	1031	3.3	76	3
22.5	1032	3.3	78	0
18.5	1035	3.4	83	4
22.5	1033	3.5	78	3
23	1031	3.6	76	3
22	1033	3.6	82	3
19.5	1033	3.6	81	0
18.5	1033	3.6	81	0
22.5	1036	3.7	88	0
22	1034	3.7	83	3
24	1036	3.8	87	5
24	1036	3.8	87	0
18.5	1037	3.8	91	4
19.5	1038	3.9	91	5
25	1041	4.0	100	3
26	1036	4.0	84	3
20	1037	4.0	90	5
27	1037	4.0	91	5
22.5	1037	4.1	89	3
20.5	1038	4.1	92	0
43.5	1045	4.7	110	4
27.5	1045	4.8	109	4
29	1046	4.8	110	4
26.5	1047	4.9	116	3
29	1046	4.9	112	0
24.5	1047	4.9	116	0
18.5	1034	3.5	81	3
31	1045	5.0	109	0
32	1047	5.0	117	3
24	1046	5.0	111	4
29	1046	5.1	111	3
33	1046	5.1	110	0
29	1048	5.2	119	0
33.5	1050	5.4	121	0
26	1051	5.5	125	5
43	1058	6.0	146	0
31.5	1079	8.9	197	0
31.5	1081	8.9	204	0

Table 3.2: Lager data table

depressed, and the total number in each category. Those who are not cases are 'controls', assumed well. The eight categories are defined by three binary factors:
(a) 'rmq', which corresponds to a particular measure of the mother's psychological well-being;
(b) 'mcr', which indicates whether or not the mother has good 'confiding' relations with other adults;
(c) 'events', which indicates whether or not the child has experienced recent stressful life events in the 12 months prior to the study.
In each case a value of 1 for the factor corresponds to its status being 'good' or 'normal', and a value of 2 corresponds to its being 'poor'.
Assuming that the number of cases in a given category is binomial, given the total in that category, find a model relating the probability that a child is a case rather than a control to the three factors given. Discuss carefully how to interpret your best-fitting model, and its estimates, to the psychiatrist.
Table 3.4 shows the result of separating the cases into two categories, 'anxious' or 'depressed' (defined to be mutually exclusive), and the eight further categories are defined by the same three factors as before. Does the particular diagnosis of a case (i.e. anxious rather than depressed) depend at all on any of the factors?
[This illustrates the use of binomial regression.]

Case	Total	rmq	mcr	events
19	81	1	1	1
5	9	2	1	1
1	4	1	2	1
4	4	2	2	1
38	66	1	1	2
14	15	2	1	2
12	13	1	2	2
7	8	2	2	2

Table 3.3: Prof I.Goodyer's first dataset

anxious	depressed	Case	rmq	mcr	events
13	6	19	1	1	1
5	0	5	2	1	1
0	1	1	1	2	1
2	2	4	2	2	1
28	10	38	1	1	2
7	7	14	2	1	2
9	3	12	1	2	2
4	3	7	2	2	2

Table 3.4: Prof I.Goodyer's second dataset

9. Agresti (1990) *Categorical Data Analysis*, p. 377, gives the Table 3.5 below, relating mother's education to father's education for a sample of eminent black Americans (defined as persons having a biographical sketch in the publication *Who's Who Among Black Americans*). Here, for education,

Mother	Father=1	Father=2	Father=3	Father=4
1	81	3	9	11
2	14	8	9	6
3	43	7	43	18
4	21	6	24	87

Table 3.5: Eminent Black Americans: educational levels of Mothers and Fathers

1 = 8th grade or less, 2 = Part High School, 3 = High School, 4 = College.
Let $p_{ij} = Pr$(mother's education is i, father's education is j), $1 \leq i, j \leq 4$. Consider the model

$$\omega : p_{ij} = \theta\phi_i + (1-\theta)\alpha_i\beta_j, \quad i = j$$

$$p_{ij} = (1-\theta)\alpha_i\beta_j, \quad i \neq j.$$

where $\Sigma\phi_i = 1, \Sigma\alpha_i = 1, \Sigma\beta_j = 1$, and $0 < \theta < 1$.
Can you interpret ω to a sociologist?
Show that, under ω,

$$\log p_{ij} = a_i + b_j, \; i \neq j,$$

for suitably defined a_i, b_j.
With n, the cell frequencies, declared as Poisson variables, with the default link function, and with the two factors M.Ed and F.Ed each with the 4 given values, glm finds that
$n \sim M.Ed + F.Ed$ has deviance 159.25, with 9 df.
Why should you expect this deviance to be so large?
But if we omit the 4 diagonal entries of the table, and fit
$n \sim M.Ed + F.Ed$, we find that the deviance is 4.6199, with 5 df. How do you interpret this?
Find the fitted frequencies for this latter model.

10. You see below the results of using glm to analyse data from Agresti(1996, p247) on tennis matches between 5 top women tennis players (1989-90). Let (r_{ij}) be the number of wins of player i against player j, and let n_{ij} be the total number of matches of i against j, for $1 \leq i < j \leq 5$. Thus we have 10 observations, which we will assume are independent binomial, with $\mathbb{E}(r_{ij}) = n_{ij}p_{ij}$.
The model we will fit is

$$\log(p_{ij}/(1-p_{ij})) = \alpha_i - \alpha_j, \text{ with } \alpha_5 = 0.$$

equivalently

$$p_{ij} = \pi_i/(\pi_i + \pi_j)$$

where $\pi_i = e^{\alpha_i}$.
The data can be read in (read.table(“ ”, header=T)) as the table

```
wins tot sel graf saba navr sanc
2     5   1   -1   0    0     0
1     1   1    0   -1   0     0
3     6   1    0   0    -1    0
2     2   1    0   0    0     -1
6     9   0    1   -1   0     0
3     3   0    1   0    -1    0
7     8   0    1   0    0     -1
1     3   0    0   1    -1    0
3     5   0    0   1     0   -1
3     4   0    0   0     1   -1
```

Thus for example, the first row of numbers tells us that ‘sel’ played ‘graf’ for a total of 5 matches, and ‘sel’ won 2 of these.
The result of

```
glm(wins/tot~ sel+graf+saba+navr-1, binomial, weights=tot)
```

is deviance 4.6493, with 6 df.
The parameter estimates are

```
  sel = 1.533(0.7870),
 graf = 1.933(0.6783),
 saba = 0.731(0.6770),
 navr = 1.087( 0.7236),
 with sanc =0 by assumption.
```

(i) Why do we impose a constraint on $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$?
(ii) Can we confidently say that Graf is better than Sanchez?
(iii) Can we confidently say that Graf is better than Seles?
(iv) What is your estimate of the probability that Sabatini beats Sanchez, in a single match? (Answer: 0.6750)
(v) Table 3.6 gives corresponding data for 5 top men tennis players during 1989–90, taken from Agresti (1996, p255). Analyse them, fitting a model of the same form as above.

11. The purpose of this example is to introduce you to the topic of *overdispersion* in the context of the Poisson distribution.
Suppose now that the observations $Y_1, ..., Y_n$ are independent, with
$\mathbb{E}(Y_i) = \mu_i$, and $var(Y_i) = \phi\mu_i$, and $log(\mu_i) = \beta x_i$,
for some unknown ϕ and unknown scalar parameter β.
Let β_0 be the true value of this unknown parameter.
Our aim is to estimate β, but ϕ is an unknown ‘dispersion’ parameter. Clearly

			Loser		
Winner	Edberg	Lendl	Agassi	Sampras	Becker
Edberg	–	5	3	2	4
Lendl	4	–	3	1	2
Agassi	2	0	–	1	3
Sampras	0	1	2	–	0
Becker	6	4	2	1	–

Table 3.6: The 5 top men tennis players in 1989-90

$\phi > 1$ will correspond to over-dispersion relative to the Poisson. In the absence of knowledge of ϕ, we choose our estimator $\hat{\beta}$ to maximise the function $l_p(\beta)$, where

$$l_p(\beta) = -\Sigma\mu_i + \beta\Sigma x_i y_i + \text{constant}.$$

(Thus $l_p()$ is in general not the 'correct' loglikelihood function: we work out below whether this is a serious problem.)
By expanding

$$\frac{\partial l_p(\beta)}{\partial\beta}$$

evaluated at $\hat{\beta}$, about β_0, show that $(\hat{\beta}-\beta_0)$ is approximately equal to $(I(\beta_0))^{-1}U(\beta_0)$, where

$$U(\beta) = \frac{\partial l_p(\beta)}{\partial\beta},$$

and

$$I(\beta) = \Sigma x_i^2 \exp\beta x_i,$$

and hence show that, approximately,

$$\mathbb{E}(\hat{\beta}) = \beta_0, \text{ and } var(\hat{\beta}) = \phi(I(\beta_0))^{-1}.$$

Thus if $\phi > 1$, the true variance of the $\hat{\beta}$ will be greater than the value given by software which assumes the Poisson distribution.
R allows you to make this simple correction for overdispersion, by

```
summary(glm(y~x, poisson), dispersion =3.21))
```

where, for example, the 'dispersion' ϕ has been estimated as (deviance/df), here 3.21. (A better estimate for ϕ may be obtained by replacing the deviance in this formula by the chi-squared statistic, $\Sigma(o-e)^2/e$, following the standard notation.) Experiment with this.
Generalise your result to the case where β, x_i are vectors of dimension p.

3.2 Solutions to Example Sheet 3

1. Clearly $log(f(t_i|\alpha,\beta)) = -\nu log(\mu_i) - t_i/\mu_i +$ constant .

Hence loglikelihood is

$$L(\alpha,\beta) = \Sigma\nu log(\alpha+\beta x_i) - \alpha\Sigma t_i - \beta\Sigma x_i t_i + constant$$

where Σ runs from $i=1$ to $i=n$.
Hence, by the factorisation theorem, (a_1,a_2) is sufficient for (α,β).
Take $x_i = 0$ for i=1,...,m, and x_i=1 for $i=m+1,...,n$.
Then

$$L(\alpha,\beta) = \nu mlog(\alpha) + \nu(n-m)log(\alpha+\beta) - \alpha(T_1+T_2) - \beta T_2 + constant$$

where $T_1 = t_1 + ... + t_m, T_2 = t_{m+1} + ... + t_n$.
To estimate β, solve $\partial L/\partial\alpha = 0, \partial L/\ \partial\beta = 0$ for $(\hat{\alpha},\hat{\beta})$.
We know that the asymptotic distribution of $(\hat{\alpha},\hat{\beta})$ is bivariate normal, with mean (α,β), and covariance matrix V, say, where $-V^{-1}$ is the expectation of the matrix of 2nd partial derivatives of L.
Now

$$\partial L/\partial\alpha = (\nu m/\alpha) + \nu(n-m)/(\alpha+\beta) - (T_1+T_2)$$

$$\partial L/\partial\beta = \nu(n-m)/(\alpha+\beta) + T_2$$

Hence, after evaluating the matrix of 2nd derivatives of L, (whose elements turn out not to be random variables), we see that V^{-1} is of the form

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

(where you will find the values of a,b,c).
Hence, inverting V shows us that $\hat{\beta}$ is approximately $N(\beta, a/(ac-b^2))$.
Thus the approximate confidence interval for β will be $\hat{\beta} \pm 2\sqrt{(}a/(ac-b^2))$.

2. With the given model, we see that the mle's of (α,β) minimise

$$\Sigma\Sigma(y_{ij} - \alpha - \beta x_{ij})^2$$

where in $\Sigma\Sigma,\ j=1,...,n_i, i=1,...,t$.
But $x_{ij} = x_i$, for all i,j so that the mle's of (α,β) minimise

$$\Sigma\Sigma(y_{ij} - \alpha - \beta x_i)^2$$

giving $\hat{\beta} = S_{xy}/Sxx$ where $S_{xy} = \Sigma\Sigma(y_{ij} - \bar{y})x_i$ and S_{xx} is defined similarly.
Now it is true that

$$\bar{y}_i = \alpha + \beta x_i + \eta_i \text{ say}$$

where η_i is dist'd as $NID(0,\sigma^2/n_i)$.
If we choose (α,β) to minimise $\Sigma(\bar{y}_i - \alpha - \beta x_i)^2$, we will obtain less accurate estimators of (α,β) than $(\hat{\alpha},\hat{\beta})$. You can check that the estimators thus obtained

will still be unbiased, but will have larger variances.
However, if we choose (α, β) to minimise

$$\Sigma(\bar{y}_i - \alpha - \beta x_i)^2 n_i$$

then we will obtain $(\hat{\alpha}, \hat{\beta})$ as above, ie the correct estimators. So using only the reduced data set $(x_i, \bar{y}_i, n_i)$ in this way will be fine, as long as σ^2 is known.
If σ^2 is unknown, as is almost always the case in practice, then we will get a more accurate estimator of it by using the ORIGINAL data set(and estimating it as usual by $(residual\ ss)/df)$ than by using the corresponding expression when we have "condensed" the data down to $(x_i, \bar{y}_i, n_i)$, namely

$$\frac{1}{(t-2)}\Sigma(\bar{y}_i - \hat{\alpha} - \hat{\beta} x_i)^2 n_i.$$

Furthermore, if we only have the "condensed" dataset, we are much less likely to be able to do an adequate job of checking the validity of the original linear model with constant error variance (Sketch some simple examples with $t = 3$).

3. Write Y_{ij} for the (i, j)th binary observation. We will assume that Y_{ij} are distributed independently as $Bi(1, p(x_i))$ for $j = 1, ..., n_i, i = 1, ..., t$.
Thus the loglikelihood of the data is

$$L(\alpha, \beta) = \Sigma\Sigma[y_{ij} logp(x_i) + (1 - y_{ij})log(1 - p(x_i))]$$

where $log(p(x_i)/(1 - p(x_i)) = logit(p(x_i)) = \alpha + \beta x_i$. Hence

$$L(\alpha, \beta) = \Sigma(y_{i+} log(p(x_i)) + (n_i - y_{i+})log(1 - p(x_i)).$$

Clearly, the loglikelihood is the same whether we enter the data as
$(0, x_1), (1, x_1), (0, x_1), ...$ ie 15 separate cases,
or as $(1, 3, x_1), ...$ ie 4 separate cases.
Thus, since $(\hat{\alpha}, \hat{\beta})$, and its asymptotic covariance matrix, are obtained purely from the loglikelihood function, we will get the same answers for these whichever of the 2 ways we choose to enter the data.
(But try a simple numerical example. Why do you get different expressions for the deviances and their df's ?)

4.a) Minimising $(Y - X\beta)^T(Y - X\beta)$ in β gives

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

Hence, if X is of full rank, $\hat{\beta} = (X^T X)^{-1} X^T Y$ so that
b) $\hat{Y} = X\hat{\beta} = HY$ say, where
$H = X(X^T X)^{-1} X^T$.
It is easy to check that $H = H^T$ and $HH = H$,(ie H is a projection matrix).

c) $e = Y - X\hat{\beta} = (I - H)Y = (I - H)\epsilon$.
Now, ϵ dist'd as $N(0, \sigma^2 I)$. Thus, using $(I - H)(I - H)^T = (I - H)$ we see that e is distributed as $N(0, \sigma^2(I - H))$
giving e_i as $N(0, \sigma^2(1 - h_i))$ as required.
d) Suppose u is an eigen vector of H corresponding to eigen value λ .
Then $Hu = \lambda u$, thus $HHu = \lambda Hu$ thus $Hu = \lambda Hu$ thus
$Hu = 0$ or $\lambda = 1$, so that $\lambda = 0$ *or* 1.
Hence $h_1 + ... + h_n = trace(H) = \lambda_1 + \lambda_2 + ... + \lambda_n$ where (λ_i) are the eigen values of H. Hence

$$(h_1 + ... + h_n) = rank(H) = rank(X) = p.$$

e) Clearly, $var(e_i) = (1 - h_i)\sigma^2$, hence $h_i \leq 1$.
Further, $\hat{Y} = HY = H(X\beta + \epsilon)$,
hence $var(\hat{Y_i}) = h_i\sigma^2$ hence $h_i \geq 0$.
For the case of simple linear regression, the first column of X is 1 (the vector of 1's) and the second column of X is say x', where ${x_i}' = x_i - \bar{x}$.
Hence $X^T X$ is $diag(n, S_{xx})$ where $S_{xx} = \Sigma(x_i - \bar{x})^2$.
Evaluating H from this shows that

$$h_i = 1/n + (x_i - \bar{x})^2/S_{xx}$$

f) We see from e) above that x_i distant from $\bar{x}$ will give rise to relatively large h_i.(Sketch graphs of various possible configurations to get a feel for (h_i)) .

5. The solution to this is in the lecture notes.
The practical importance is of course that there is no need for a *multinomial* distribution within glm(): so long as we are fitting log-linear models we can use the Poisson distribution and log-linear models, provided that we include the 'intercept' term.

6. This example arose from some psychiatry data provided by Professor I.Goodyer of Cambridge University. The 4 factors were
$S = 1, 2$ for girls,boys
$D = 1, 2$ for depression no or yes
$A = 1, 2$ for anxiety symptoms no or yes
$B = 1, 2$ for behavioural symptoms no or yes.
Thus the table is in fact rather *sparse* for the large sample theory to be realistic, but we give this analysis as an illustrative example of the glm( ) modelling.

$$Sat: \quad log(p_{ijkl}) = \mu + \alpha_i + ... + (\alpha\beta)_{ij} + ... + (\alpha\beta\gamma)_{ijk} + ... + (\alpha\beta\gamma\delta)_{ijkl}$$

is the saturated model (we assume the usual glm() constraints $\alpha_1 = 0$ etc.
a) fits the saturated model (ie 2^4 parameters for 2^4 observations) so that we expect to get deviance 0 with df = 0.

However a quick glance at the parameter estimates and their se's for this model will usually suggest to us which of the high-order interactions(if any) can be dropped. Our object is to fit the simplest possible model which is consistent with the data, and of course we want to interpret this model to the scientist who provided the data.
b) fits Sat with

$$H_0 : (\alpha\beta\gamma\delta)_{ijkl} = 0$$

ie no 4th order interaction. Refer the increase in deviance (2.72) to χ^2_1 to test H_0 (applying Wilks' theorem). The result is non-significant, so we accept H_0.
c) fits Sat with H_0 and $H_1 : (\alpha\gamma\delta)_{ikl} = 0, (\beta\gamma\delta)_{jkl} = 0$.
Refer the resulting increase in deviance (ie 3.42−2.72) to χ^2_2 to test H_1 assuming H_0 true. You find that you accept H_1.
d) fits

$$H_2 : log(p_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\beta\delta)_{jl}$$

so we have dropped yet more parameters in moving from c) to d). The increase in deviance is $(8.48 - 3.42)$ which is non-significant when referred to χ^2_5 so dropping these extra parameters was permissible.
Furthermore, we can assess the fit of the model H_2 by referring 8.48 to χ^2_8. In fact 8.48 is almost the same as $\mathbb{E}(\chi^2_8)$, so you see that H_2 is a good fit. You could check that no further parameters can safely be dropped.
The final model is therefore $p_{ijkl} = a_{ij}b_{jk}c_{jl}$ for some a, b, c. We will rewrite this in a more enlightening way as

$$Pr(A, B, D|S) = Pr(A|D)Pr(B|D)Pr(D|S).$$

(We choose to write it in this asymmetric form this since S is clearly not a 'response' variable.) We can express this in words as
'D depends directly on S, but B and A only depend on S **through** D. Furthermore, conditional on the level of D, the variables B and A are independent.'
This is a simple example of a *graphical* model of conditional independence. Draw a graph in which the 4 points S, D, B, A are connected only by the arcs SD, DB, DA, as in Figure 3.1.
Finally, the model in which S, D, B, A are independent corresponds to a graph with no links at all between the 4 points, and in glm terms it is

$$H_I : log(p_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l.$$

7. a) As with most regression problems, it is best to START by doing some simple plots: in this case all possible pairwise plots of the 4 variable concerned.This reveals that price is positively related to each of og, percent and calories, but also, as a student doing this problem in an examination wrote,
"as any experienced lager-drinker knows", the 3 variables og, percent and calories are all measuring more or less the same thing, so we could not reasonably

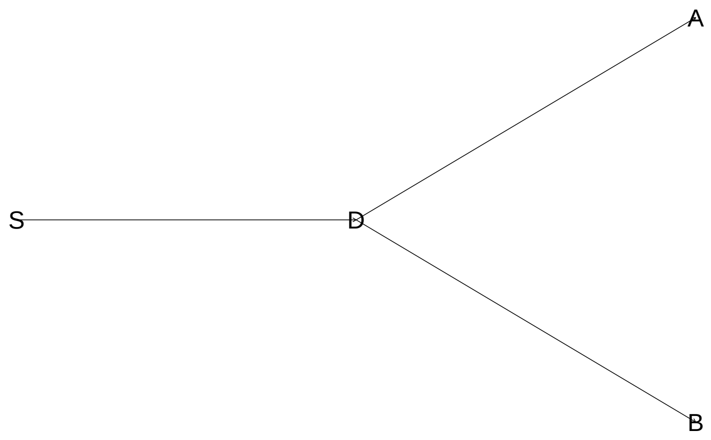

Figure 3.1: The graph showing the interdependence of the S, D, A, B data

expect 2 or 3 of them to give a much better prediction of price than just one of them.(Were teetotal students at a disadvantage in this examination?)
Linear regression of price on these 3 "explanatory" variables bears out this observation. In fact, og is the best single predictor. Inspection of the residuals reveals that some lagers are priced at a ridiculously high level, but perhaps they are connoisseurs' lagers, ie not aimed at those interested only in the alcohol content.
b) This question is asking the student to set up a 2-level factor (tasted/not tasted) and to do the regression of price on og, percent, calories and this new 2-level factor.
The SPlus6 output and corresponding graph are given at the end of these solutions. (You could use R to the same effect.)

8. Setting up the 3 factors rmq, mcr, events, each with 2 levels (level 2 being the "bad" one) with binomial regression of case on these 3, and the logistic link function, and "total" as the binomial denominator, gives

$$\log(P(case)/P(control))$$

$$= -1.248(.253) + 1.716(.514)rmq(2) + 1.257(.561)mcr(2) + 1.620(.330)events(2)$$

with deviance $5.053, df = 4$.
Thus this model fits well, and none of the 3 factors can be dropped from the model. Either compare each estimate with its se, or compare the increase in

deviance with χ^2_1 when each factor is dropped from the model in turn.
Note: having rmq at level 2 rather than level 1 increases the odds in favour of being a case by a factor of exp(1.716). Having all of rmq, mcr, events at the "bad" level rather than the "good"level increases the odds of being a case by a factor of exp(1.716 + 1.257 + 1.620). We could use the covariance matrix of these estimates to find the corresponding confidence interval for this odds ratio.
Table 2 is analysed similarly.

9. Let M be the Mother's educational level and F the Father's. Under ω,
$(M, F) = (M_1, F_1)$ with probability θ, where $M_1 = F_1$ and $P(M_1 = i) = \phi_i$.
$(M, F) = (M_2, F_2)$ with probability $1 - \theta$, where M_2, F_2 are independent, with $P(M_2 = i) = \alpha_i, P(F_2 = j) = \beta_j$,
ie, with probability θ, the mother's education MUST be identical to the father's, and with probability $1 - \theta$, they are independent (in which case there is still a chance that they are identical).
It is easily seen that ω is equivalent to the model
$\log p_{ij} = a_i + b_j$ for $i \neq j$.
If we set up ma, pa as the 4-level factors corresponding respectively to mother's, father's educational level, then using the Poisson family and log link function the glm of n(the frequency) on $(ma+pa)$, gives a deviance of 159.25,which is clearly hugely significant compared with χ^2_9. This tests the hypothesis of *independence* of the factors ma, pa, so common-sense suggests that this hypothesis is unlikely to hold: in any case, a glance at the 4×4 contingency table shows that the diagonal entries are much too large for the independence hypothesis to be plausible.
Now omit the entries of the table for which $i = j$ (use

```
  subset= (i!=j)
```

in glm() command). Now the deviance is much the same as its expected value, showing that ω is a good fit. From the table of "fitted values" we could find $\hat{\theta}$, and so on.
This model is what sociologists call a "mover-stayer" model; it is also an example of a "quasi-independence" model for a contingency table.

10. These data on women's and men's tennis matches are taken from "An introduction to categorical data analysis " by Alan Agresti (1996) and their analysis is discussed in Agresti's Chapter 9.
i) The model being fitted is:
w_{ij} (the number of wins) is dist'd as independent $Bi(t_{ij}, p_{ij})$ for$1 \leq i < j \leq 5$
with $logit(p_{12}) = \mu + sel - graf, logit(p_{13}) = \mu + sel - sab$ etc
but the import of the -1 term in the glm statement is that we set μ= 0.
Hence our model is $logit(p_{ij}) = \alpha_i - \alpha_j$ for $1 \leq i < j \leq 5$.
But note that with this model we could replace eg, α_5 by $\alpha_5 + 13$, and then

replace α_4 by $\alpha_4 + 13$ and so on, without changing the formula for $logit(p_{ij})$. So we impose a constraint, without loss of generality $\alpha_5 = 0$, to ensure **parameter identifiability** .
(If we include a term + sanc in our fitting, glm( ) will estimate sel, graf, saba, navr as given, but will obligingly tell us, in effect, that sanc is *aliased*, meaning that it cannot be estimated if the previous 4 parameters are already in the model.)
This is an example of the Bradley-Terry model for paired comparisons.
ii) Can we confidently say that Graf is better than Sanchez?
YES, because the model fits well (refer its deviance of 4.65 to χ^2_6) and $\hat{\alpha}_2/se(\hat{\alpha}_2)$ = 2.854: refer this to $N(0,1)$.
iii) Can we confidently say that Graf is better than Seles?
Now $\hat{\alpha}_2 - \hat{\alpha}_1 = 1.933 - 1.533$.
and the estimated variance of $(\hat{\alpha}_2 - \hat{\alpha}_1) = (.6687)^2$
(Use the parameter estimate correlation matrix)
Referring $(\hat{\alpha}_2 - \hat{\alpha}_1)/.6687 = 0.5982$ to $N(0,1)$, we see that the difference is not significant.
(this small correction made March 17, 2014, thanks to Dr R.Shah.)
iv) Our estimate of the probability that Sabatini beats Sanchez, in one match, is

$$e^{\cdot 7308}/(1 + e^{\cdot 7308}) = 1/(1 + e^{-.7308}) = .675$$

and note that using $.7308 \pm 1.96 \times .6764$
we could attach a confidence interval to this probability
ie $1/(1 + exp(-.7308 \pm 1.96 \times .6764))$.
v) The dataset for the men's tennis doesn't fit the Bradley-Terry model so well.
Note that our underlying model contains the assumption
w_{ij} distributed independently as $Bi(t_{ij}, p_{ij})$.
This may not be reasonable. For example, if Graf beats Sanchez in the first match, then this result may affect the probability that Graf beats Sanchez in their next match, and so on. Further, we should perhaps take account of the surface on which the match is played (grass, clay, etc). But we do not have the data to be able to take account of such factors in our modelling.

11. Take β_0 as the true value of β. The usual Taylor series expansion of the first derivative of $l_p(\beta)$ at $\hat{\beta}$ about β_0 shows that, to first order,

$$0 = U(\beta_0) - (\hat{\beta} - \beta_0)I(\beta_0)$$

which gives the required approximation for $(\hat{\beta} - \beta_0)$.
Now, as usual, $\mathbb{E}(U(\beta_0)) = 0$, giving us the required expression for $\mathbb{E}(\hat{\beta})$. Further, by noting that

$$var(\hat{\beta}) = var(\Sigma x_i Y_i)/(I(\beta_0))^2$$

and substituting $var(Y_i) = \phi\mu_i$, we obtain the required expression for $var(\hat{\beta})$.

....

For number 7 (lager data), Figure 3.2 gives the corresponding 'pairs-plot'. Here is the corresponding code and output.

```
>lager <- read.table("lager", header=T)
>summary(lager)
>pairs(lager)
>attach(lager)
> first.lm <- lm(Price ~ og + percent + cal)
>summary(first.lm,cor=F)
Call: lm(formula = Price ~ og + percent + cal)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.953 -3.217 -0.2989 1.827 16.53

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1430.4215  1487.7819    -0.9614     0.3427
         og     1.4393     1.4849     0.9693     0.3389
    percent    -2.4402     5.0846    -0.4799     0.6342
        cal    -0.3201     0.5135    -0.6233     0.5370

Residual standard error: 4.856 on 36 degrees of freedom
Multiple R-Squared: 0.4136
F-statistic: 8.465 on 3 and 36 degrees of freedom,
 the p-value is 0.0002186

>summary(first.lm,cor=T)

Call: lm(formula = Price ~ og + percent + cal)
Residuals:
    Min     1Q  Median    3Q   Max
 -6.953 -3.217 -0.2989 1.827 16.53

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1430.4215  1487.7819    -0.9614     0.3427
         og     1.4393     1.4849     0.9693     0.3389
    percent    -2.4402     5.0846    -0.4799     0.6342
        cal    -0.3201     0.5135    -0.6233     0.5370

Residual standard error: 4.856 on 36 degrees of freedom
Multiple R-Squared: 0.4136
F-statistic: 8.465 on 3 and 36 degrees of freedom,
 the p-value is 0.0002186

Correlation of Coefficients:
        (Intercept)      og percent
     og -1.0000
percent  0.4782     -0.4781
    cal  0.9217     -0.9218  0.1035

>next.lm <- lm(Price ~ og); summary(next.lm)
Call: lm(formula = Price ~ og)
Residuals:
    Min     1Q  Median    3Q   Max
 -7.617 -3.135 -0.3271 2.045 16.89

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -336.5331   71.2697    -4.7220    0.0000
         og    0.3475    0.0684     5.0802    0.0000

Residual standard error: 4.763 on 38 degrees of freedom
Multiple R-Squared: 0.4045
F-statistic: 25.81 on 1 and 38 degrees of freedom,
 the p-value is 1.033e-05
>rat <- (rating>0)*1 ; Rat <- factor(rat)
>third.lm <- lm(Price ~ og + Rat); summary(third.lm)
Call: lm(formula = Price ~ og + Rat)
Residuals:
    Min     1Q Median    3Q   Max
 -7.609 -3.343  -0.22 1.934 17.05

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept) -330.0376   78.3167    -4.2141    0.0002
         og    0.3415    0.0748     4.5667    0.0001
        Rat   -0.3635    1.7006    -0.2138    0.8319

Residual standard error: 4.824 on 37 degrees of freedom
Multiple R-Squared: 0.4052
F-statistic: 12.6 on 2 and 37 degrees of freedom,
 the p-value is 6.696e-05

>last.lm <- lm(Price ~ Rat); summary(last.lm)
Call: lm(formula = Price ~ Rat)
Residuals:
    Min     1Q Median    3Q   Max
 -9.067 -4.405  -0.78 3.273 19.22
```

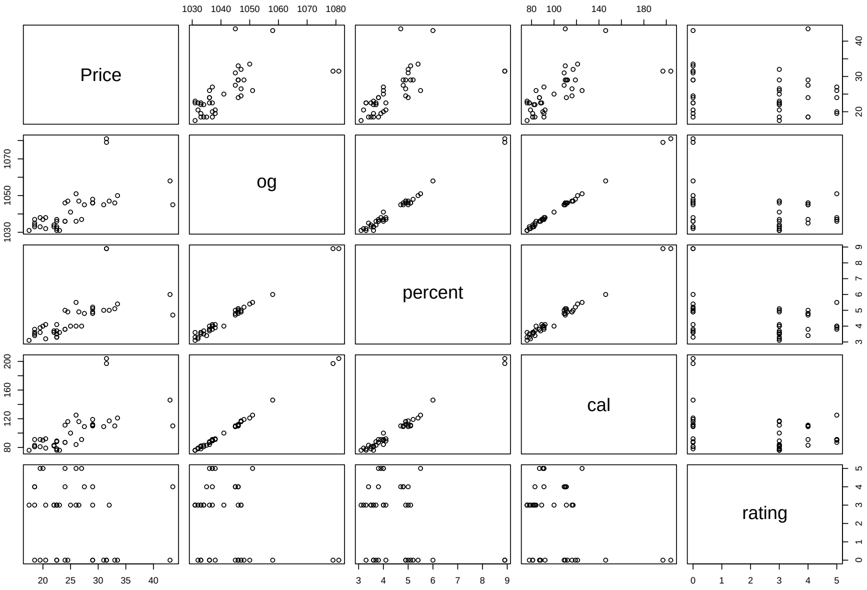

Figure 3.2: The pairs plot for the lager data

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  27.5667   1.5370    17.9359   0.0000
        Rat  -3.2867   1.9441    -1.6906   0.0991

Residual standard error: 5.953 on 38 degrees of freedom
Multiple R-Squared: 0.06995
F-statistic: 2.858 on 1 and 38 degrees of freedom,
 the p-value is 0.09911
>tapply(Price,Rat,mean)
        0     1
 27.56667 24.28
```