Introduction to Generalized Linear Modelling

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

August 15, 2011

Contents

Preface

Preliminary statement. When I first wrote my lecture notes for the Part II course, Sarah Shea–Simonds very kindly typed the core notes in TeX, and I added to them bit by bit, again in TeX. However, my style was still rather like a telegram, partly as I was trying to save on paper. Now that I am retired, I have time to retype the notes in LaTeX. I have tried to make the style rather more 'flowing', and have included more various graphs, exercises, Tripos questions and solutions. This editing process is quite enjoyable but rather slow. I'll put the revisions on my webpage from time to time, and of course would appreciate comments and suggestions. Special thanks are due to Professor Yuri Suhov for his comments and suggestions.

There are already several excellent books on this topic. For example McCullagh and Nelder(1989) have written the classic research monograph, and Aitkin et al. (1989) have an invaluable introduction to the pioneering software GLIM. Although I was very glad to learn a great deal by using GLIM, that particular software was superseded some years ago by excellent and powerful languages such as S-Plus and R.

Students will naturally gain a much deeper understanding of the theory by putting it into practice on real (if small) datasets. An excellent text book to help them to do this in Splus and/or R is the one by Venables and Ripley (2002), particularly their Chapters 6 and 7.

Dobson (1990) has written a very full and clear introduction, which is not linked to any one particular software package. Agresti (2002) in a very clearly written text with many interesting data-sets, introduces Generalized Linear Modelling with particular reference to categorical data analysis.

The notes presented here are designed as a SHORT course for mathematically able students, typically third-year undergraduates at a UK university, studying for a degree in mathematics or mathematics with statistics. The text is designed to cover a total of about 20 student contact hours, of which 10 hours would be lectures, 6 hours would be computer practicals, and the remaining 4 are classes or small-group tutorials doing the problem sheets, for which the solutions are available at the end of the book. It is assumed that the students have already had an introductory course on statistics. While my notes are not dependent on any one particular statistical software, I wrote 'worksheets' to serve as computer practicals to introduce the students to (S-plus or) R. These worksheets (now extended somewhat) may be seen on `http://www.statslab.cam.ac.uk/~pat/redwsheets.pdf`

Both the practical sessions and the problem sheets are designed to challenge the students and deepen their understanding of the material of the course. These notes do not have a separate section as an introduction to R and its properties. My experience of computer practicals with students is that they learn to use R or S-plus quite fast by the 'plunge-in'

method (as if being taught to swim). Of course this is now aided by the very full on-line help system available in R and S-plus.
R.W.M.Wedderburn, who took the Cambridge Diploma in Mathematical Statistics in 1968-9, having graduated from Trinity Hall, was with J.A.Nelder, the originator of Generalized Linear Modelling. Nelder and Wedderburn published the first paper on the topic in 1972, while working as statisticians at the AFRC Rothamsted Institute of Arable Crops Research (as it is now called). Robert Wedderburn died tragically young, aged only 28. But his original ideas were extensively developed, both in terms of mathematical theory, particularly by McCullagh and Nelder, and computational methods, so that now every major statistical package, eg SAS, Genstat, R, S-plus has a generalized linear modelling (glm) component.

Chapter 1

The asymptotic likelihood theory needed for glm

1.1 Set-up and notation

Take $x_1, \ldots, x_n$ a r.s. (*random sample*) from the pdf (*probability density function*) $f(x|\theta)$. Define

$$\exp[L_n(\theta)] = \prod_1^n f(x_i|\theta)$$

as the likelihood function of θ, given the data x. Then

$$L_n(\theta) = \sum_1^n \log f(x_i|\theta)$$

is the loglikelihood function.
Note: $\{\log f(X_i|\theta)\}$ form a set of i.i.d. (*independent and identically distributed*) random variables. (The capital letter X_i denotes a random variable.)

1.2 The two key results of this chapter

Preamble. Suppose $\hat{\theta}_n$ maximises $L_n(\theta)$, that is $\hat{\theta}_n$ is the m.l.e. (*maximum likelihood estimator*) of θ. How good is $\hat{\theta}_n$ as an estimator of the unknown parameter(s) θ as the sample size $n \to \infty$? Clearly we hope that, in some sense,

$$\boxed{\hat{\theta} \to \theta \text{ as } n \to \infty}.$$

Write $x = (x_1, ..., x_n)$. Then we know that for $t(x)$ an unbiased estimator of θ , and θ a scalar parameter,

$$var(t(X)) \geq 1 \Big/ \mathbb{E}\left(\frac{-\partial^2}{\partial\theta^2} L_n(\theta)\right) \equiv v_{\text{CRLB}}(\theta).$$

This is the Cramèr Rao lower bound (CRLB) for the variance of an unbiased estimator of θ. There is a corresponding matrix inequality if t, θ are vectors.
Result 1. For θ real,

$$\hat{\theta}_n \overset{\text{approx}}{\sim} N\big(\theta, v_{\text{CRLB}}(\theta)\big) \quad \text{for } n \text{ large.}$$

The vector version of this result, which is of great practical use, is the following.
For θ a k-dimensional parameter,

$$\hat{\theta}_n \overset{\text{approx}}{\sim} N_k\big(\theta, \Sigma_n(\theta)\big) \quad \text{for large } n.$$

This says that $\hat{\theta}_n$, which is a random vector, by virtue of its dependence on $X_1, \ldots, X_n$, is asymptotically k-variate normal, with mean vector $= \theta$ (which of course is the true parameter value) and covariance matrix is $\Sigma_n(\theta)$, where $\Sigma_n(\theta)$ is given by

$$\begin{aligned}&\big(\Sigma_n(\theta)\big)^{-1} \quad \text{has as its } (i,j)^{\text{th}}\text{element}\\ &\mathbb{E}\left(\frac{-\partial^2}{\partial\theta_i\partial\theta_j} L_n(\theta)\right).\end{aligned}$$

Thus you can see, at least for the scalar version, that the asymptotic variance of $\hat{\theta}_n$ is indeed the CRLB.
Remarks on Result 1:

(0) One obvious consequence of Result 1 is that any component of $\hat{\theta}_n$, e.g. $(\hat{\theta}_n)_1$ is asymptotically Normal.

(1) We have omitted any mention of the necessary regularity conditions. This omission is appropriate for the robust 'coal-face' approach of this course. But we stress here that k must be **fixed** (and finite).

(2) $\Sigma_n(\theta)$, since it depends on θ, is generally unknown. However, to use this result, for example in constructing a confidence interval for a component of θ, we may replace

$$\Sigma_n(\theta) \quad \text{by} \quad \Sigma_n(\hat{\theta}),$$

i.e. replace

$$\mathbb{E}\frac{-\partial^2}{\partial\theta_i\partial\theta_j} L_n(\theta)$$

by its value at $\theta = \hat{\theta}$. In fact, we can often replace it by

$$\frac{-\partial^2}{\partial\theta_i\partial\theta_j} L_n(\theta) \quad \text{evaluated at } \theta = \hat{\theta}.$$

In some cases it may turn out that two of these three quantities, or even all three quantities, are the same thing.
Result 2. Suppose we wish to test

$$\begin{aligned}H_0 &: \ \theta \in \omega \quad \text{against}\\ H_1 &: \ \theta \in \Omega\end{aligned}$$

where $\omega \subset \Omega$, and ω is of lower dimension than Ω. Now the Neyman–Pearson lemma tells us that the most powerful size α test of

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta = \theta_1$$

is of the form : reject H_0 in favour of H_1 if

$$exp(L_n(\theta_1))/exp(L_n(\theta_0)) > \text{a constant}$$

where the constant is chosen to arrange that

$$P(\text{reject } H_0 \mid H_0 \text{ true}) = \alpha.$$

Leading on from the ideas of the Neyman–Pearson lemma, it is natural to consider as test statistic the ratio of maximised likelihoods, defined as

$$R_n \equiv \max_{\theta\in\Omega}(\exp L_n(\theta)) \Big/ \max_{\theta\in\omega}(\exp L_n(\theta))$$

where we reject $\theta \in \omega$ if and only if the above ratio is too large. But how large is 'too large'?
We want, if possible, to control the SIZE of the test, say to arrange that

$$P(\text{reject } \ \omega \mid \theta) \le \alpha$$

for all $\theta \in \omega$, where we might choose $\alpha = .05$ (for a 5% significance test). We *may* be able to find the *exact* distribution of the ratio R_n, for any $\theta \in \omega$, and hence achieve this. But in general this is an impossible task, so in practice we need to appeal to
Result 2: Wilks' Theorem.
For large n, if ω true,

$$2\log R_n \overset{\text{approx}}{\sim} \chi^2_p \quad \text{where } p = \dim(\Omega) - \dim(\omega).$$

i.e. $2\log R_n$ is approximately distributed as chi-squared, with p degrees of freedom (df). Hence for a test of ω having approximate size α, we reject ω if $2\log R_n > c$, where c is found from tables as

$$Pr(U > c) = \alpha, \text{ where } U \sim \chi^2_p.$$

1.3 The maximum likelihood estimator (mle)

Write $\hat{\theta}_n(X)$ as the value of θ that maximises

$$L_n(\theta) = \sum_1^n \log f(X_i \mid \theta)$$

or $\hat{\theta}_n$ for short; it is a r.v. through its dependence on the sample X. Usually we are able to find $\hat{\theta}_n$ as follows: $\hat{\theta}_n$ is the solution of

$$\frac{\partial}{\partial\theta_j} L_n(\theta) = 0, \quad 1 \le j \le k$$

(θ being assumed to be of dimension k, say). These equations are conventionally called the *likelihood equations*.
Warning

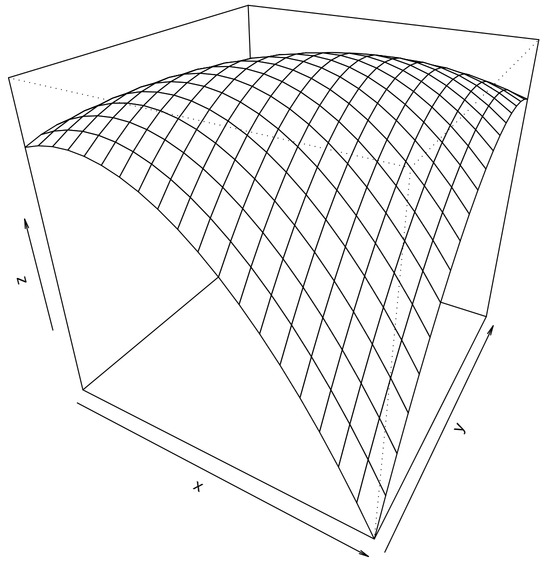

Figure 1.1: A perspective plot of a concave function

(a) As usual in maximising any function, we have to take care to check that these equations do indeed correspond to the maximum, rather than just a local maximum, or perhaps a minimum or even a saddlepoint. So, we must check that **minus the matrix of 2nd derivatives is positive-definite** to ensure that the log-likelihood surface is CONCAVE. As an example, we show the perspective plot and the contour plot of a particular concave function in Figure 1.1 and Figure 1.2. (This particular function is actually a constant $-(x^2 - 2\rho xy + y^2)/2(1 - \rho^2)$ with $\rho = .7$, computed as the log of the corresponding bivariate normal density function. It thus has a unique stationary point. In this particular case this point is at $x = 0, y = 0$, which is **the maximum of the function.**

(b) We may need to use iterative techniques to solve them for a wide class of problems.

1.4 Basic properties of the maximum likelihood estimator (mle)

(a) We use the **Factorisation Theorem** to relate the mle to sufficient statistics. Suppose $t(x)$ is a sufficient statistic for θ. Then

$$\prod_{1}^{n} f(x_i \mid \theta) = g(t(x), \theta)h(x)$$

say. Thus $\hat{\theta}(x)$ depends on x only through $t(x)$, the sufficient statistic.
But in general, $\hat{\theta}(x)$ itself is not necessarily sufficient for θ.

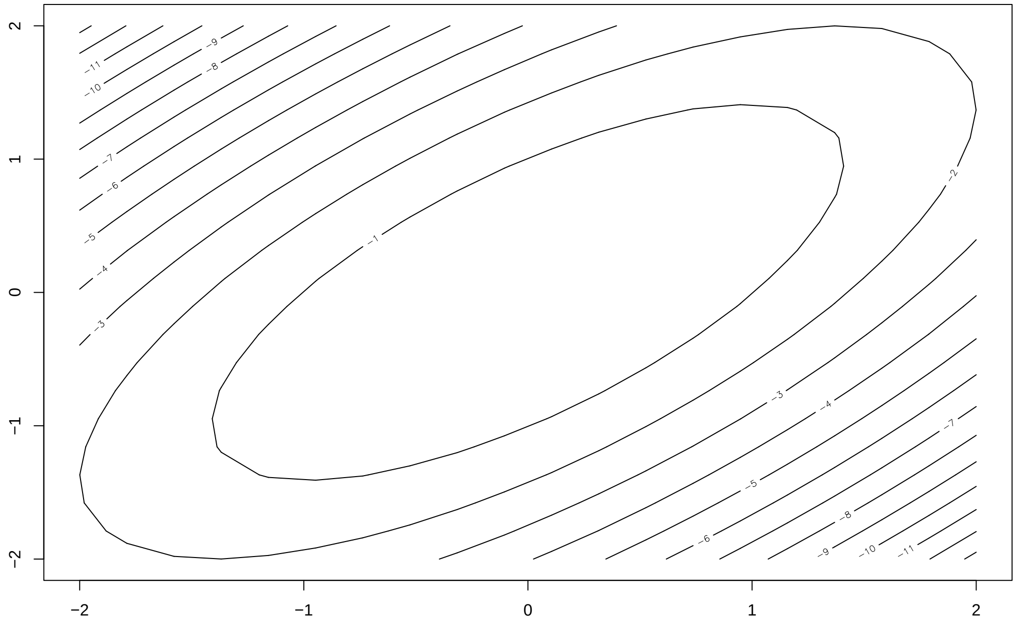

Figure 1.2: The contour plot of the same concave function

Example. Take $x_1, \ldots, x_n$ a r.s. from $f(x \mid \theta)$, the pdf of $N(\mu, \sigma^2)$. Thus

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \quad \text{and}$$

$$t(x) = (\bar{x}, \Sigma(x_i - \bar{x})^2) \quad \text{is sufficient for } \theta.$$

Show that
$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$$

and hence $\hat{\theta}$ depends on x only through $t(x)$.
Proof The log-likelihood, $\ell(\mu, \sigma^2)$ is a constant +

$$-(n/2)\log(\sigma^2) - \Sigma(x_i - \mu)^2/2\sigma^2.$$

Write $\bar{x} = \Sigma x_i/n$ and rewrite

$$\Sigma(x_i - \mu)^2 = \Sigma((x_i - \bar{x}) - (\mu - \bar{x}))^2 = \Sigma(x_i - \bar{x})^2 + n(\mu - \bar{x})^2.$$

Now find $\partial\ell/\partial\mu, \partial\ell/\partial\sigma^2$ and set this vector to $(0, 0)$. This gives

$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \Sigma(x_i - \bar{x})^2/n.$$

You should also compute the matrix of second-derivatives of $-\ell$ at this point, and show that it is positive-definite, in order to show that the stationary point is indeed the maximum of ℓ.
Thus $(\hat{\mu}, \hat{\sigma}^2)$ depends on the data x only through the sufficient statistic $t(x) = (\bar{x}, \Sigma(x_i - \bar{x})^2)$.

(b) Suppose θ is scalar, and there exists $t(x)$ an unbiased estimator of θ, and var $\big(t(X)\big)$ attains the CRLB. Then $t(x)$ is the mle of θ.

Proof. First we prove the CRLB, by way of useful review. Consider the random variables A, B, defined by

$$A = t(X), B = \frac{\partial L_n(\theta)}{\partial \theta}.$$

Then we know that from the Cauchy-Schwarz inequality that

$$(cov(A, B))^2 \leq var(A)var(B)$$

with $=$ if and only if B is a linear function of A.
But we can easily show, as follows, that for this particular A, B,

$$\mathbb{E}(B) = 0 \text{ and } cov(A, B) = 1.$$

Firstly,

$$B = \frac{\partial L_n}{\partial \theta} = \frac{\partial \log f(x \mid \theta)}{\partial \theta}$$

where x is the whole sample.
(**) Thus

$$\begin{aligned}\mathbb{E}(B) = \mathbb{E}(\frac{\partial L_n}{\partial \theta}) &= \int_x \frac{\partial L_n}{\partial \theta} f(x|\theta)dx \\ &= \int_x \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} f(x|\theta)dx \\ &= \frac{\partial}{\partial \theta} \int_x f(x|\theta)dx = \frac{\partial}{\partial \theta} 1 = 0.\end{aligned}$$

Here we made use of the fact that that $f(x|\theta)$ is a pdf, and so will integrate over x to 1 for all θ. Hence we see that

$$\begin{aligned}\text{cov}\left(t(X), \frac{\partial L_n}{\partial \theta}\right) &= \mathbb{E}\left(t(X)\frac{\partial L_n}{\partial \theta}\right) = \int t(x)\frac{\partial}{\partial \theta} f(x \mid \theta)dx \\ &= \frac{\partial}{\partial \theta} \int t(x) f(x \mid \theta)dx = \frac{\partial}{\partial \theta}\theta = 1\end{aligned}$$

since t is known to be an unbiased estimator of θ.
Thus

$$\text{var}\,(t(X)) \geq 1 \Big/ \mathbb{E}\left(\frac{\partial L_n}{\partial \theta}\right)^2 = v_n(\theta) \quad \text{say,}$$

$$\text{with } = \text{ if and only if } \frac{\partial L_n}{\partial \theta} = a(\theta)(t(X) - \theta) + b(\theta) \text{ say.}$$

[But, taking $\mathbb{E}$ of this equation, we see that $b(\theta) = 0$.] Thus, if $t(X)$ is unbiased with variance attaining the CRLB, then

$$\frac{\partial L_n}{\partial \theta} = a(\theta)(t(x) - \theta),$$

and so

$$\mathbb{E}\left(\frac{\partial L_n}{\partial \theta}\right)^2 = (a(\theta))^2\, v_n(\theta),$$

i.e. $1/v_n(\theta) = (a(\theta))^2 v_n(\theta)$, hence $v_n(\theta) = [a(\theta)]^{-1}$ We know that $a(\theta) > 0$, since $\mathrm{cov}\left(t(X), \frac{\partial L_n}{\partial \theta}\right) = 1)$.
Thus if $t(x)$ is an unbiased estimator of θ, and its variance attains the CRLB, then

$$\frac{\partial L_n}{\partial \theta} = [v_n(\theta)]^{-1}(t(x) - \theta) \quad \text{where} \quad v_n(\theta) > 0,$$

and so $L_n(\theta)$ has a unique *maximum*, at its stationary point, $\hat{\theta} = t(x)$.
Exercise (1) Using $\int_x f(x \mid \theta)dx = 1$ for all θ, show

$$\mathbb{E}\left(\frac{\partial L_n}{\partial \theta}\right)^2 = \mathbb{E}\left(\frac{-\partial^2}{\partial \theta^2}\, L_n\right).$$

Exercise (2) Take

$$f(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

where $x_i = 0$ or 1, that is $x_1, \ldots, x_n$ is a r.s. from $Bi(1, \theta)$. Show that

$$\frac{\partial L_n}{\partial \theta} = \frac{n}{\theta(1-\theta)}\,(\bar{x} - \theta),$$

and hence $\hat{\theta}_n = \bar{x}$. Show *directly* that $\mathbb{E}(\hat{\theta}_n) = \theta$, $\mathrm{var}\,(\hat{\theta}) = \theta(1-\theta)/n$ and use the CLT (*Central Limit Theorem*) to show that, for large n,

$$\hat{\theta}_n \overset{approx}{\sim} N\left(\theta, \frac{\theta(1-\theta)}{n}\right).$$

1.5 Outline Proof of Result 1

i.e. that

$$\hat{\theta}_n \overset{approx}{\sim} N\left(\theta, 1 \Big/ \mathbb{E}\left(\frac{\partial L_n}{\partial \theta}\right)^2\right) \quad \text{for large } n.$$

Proof. For clarity we will suppose that θ_0 is the *true* value of the parameter θ. We know that $\hat{\theta}_n$ maximises $L_n(\theta) = \sum_1^n \log f(X_j \mid \theta) = \sum_{j=1}^n S_j(\theta)$ say. We assume that we are dealing, exclusively, with the totally straightforward case where

$$\hat{\theta}_n \quad \text{is the solution of} \quad \frac{\partial L_n}{\partial \theta}(\theta) = 0.$$

$*$ Now

$$\left.\frac{\partial}{\partial \theta}\, L_n(\theta)\right|_{\hat{\theta}_n} \simeq \frac{\partial}{\partial \theta}\, L_n(\theta)|_{\theta_0} + (\hat{\theta}_n - \theta_0)\frac{\partial^2}{\partial \theta^2}\, L_n(\theta)|_{\theta_0}$$

assuming the remainder is negligible. The left hand side of $* = 0$, by definition of $\hat{\theta}_n$. Hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \simeq \left\{ \frac{1}{\sqrt{n}} \sum_1^n \frac{\partial S_j}{\partial \theta}\bigg|_{\theta_0} \right\} \Big/ \left\{ -\frac{1}{n} \sum_1^n \frac{\partial^2 S_j}{\partial \theta^2}\bigg|_{\theta_0} \right\}.$$

Write

$$U_j = \frac{\partial}{\partial \theta} \log f(X_j \mid \theta)\bigg|_{\theta_0} \quad \text{(this is a r.v.)}.$$

Now, as already proved, $\mathbb{E}_{\theta_0}(U_j) = 0$. Furthermore,

$$\begin{aligned} \mathrm{var}_{\theta_0}(U_j) &= \mathbb{E}_{\theta_0}(U_j^2) = \int \left(\frac{\partial}{\partial \theta} \log f(x_j|\theta) \right)^2 f(x_j|\theta) dx_j \\ &\qquad \text{evaluated at } \theta = \theta_0 \\ &= \int \left(\frac{-\partial^2}{\partial \theta^2} \log f(x_j|\theta) \right) f(x_j|\theta) dx_j \\ &\qquad \text{evaluated at } \theta = \theta_0 \end{aligned}$$

Write $\mathrm{var}_{\theta_0}(U_j) = i(\theta_0)$. Hence $\frac{1}{\sqrt{n}} \sum_1^n U_j$ has mean 0, variance $i(\theta_0)$. Thus, by the Central Limit Theorem (CLT), the distribution of $\frac{1}{\sqrt{n}} \sum U_j$ tends to the distribution of $N(0, i(\theta_0))$. But, for large n, we may use the Strong Law of Large Numbers (SLLN) to show that

$$\frac{-1}{n} \sum_1^n \frac{\partial^2 S_j}{\partial \theta^2}\bigg|_{\theta=\theta_0} \simeq \frac{-1}{n} \sum_1^n \mathbb{E}\left(\frac{\partial^2 S_j}{\partial \theta^2} \right)\bigg|_{\theta=\theta_0} = i(\theta_0).$$

Hence, for large n, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ has approximately the same distribution as $Z/i(\theta_0)$, where $Z \sim N(0, i(\theta_0))$, i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \quad \text{is approximately} \quad N(0, 1/i(\theta_0)).$$

The statistician's way of writing this is,

$$\text{for large } n, \qquad \hat{\theta}_n \overset{approx}{\sim} N\left(\theta_0, \frac{1}{n i(\theta_0)} \right).$$

Comments

(i) The basic steps used in the above are the Taylor series expansion about θ_0, and the applications of the CLT, and of the SLLN.

(ii) The result generalises immediately to vector θ, giving

$$\hat{\theta}_n \overset{approx}{\sim} N\left(\theta_0, \frac{1}{n} \big(i(\theta_0) \big)^{-1} \right),$$

the matrix $i(\theta_0)$ having $(i,j)^{\text{th}}$ element

$$\mathbb{E}\left(\frac{-\partial^2}{\partial \theta_i \partial \theta_j} \log f(X_1 \mid \theta) \right)\bigg|_{\theta_0}.$$

(iii) The result also generalises to the case where $X_1, \ldots, X_n$ are independent but *not identically* distributed. For example, we may take X_i independent $Po(\mu_i)$, ie Poisson with mean μ_i) where $\log \mu_i = \beta^T z_i$, and z_i is a given covariate and β is the unknown parameter of interest. Hence

Thus
$$f(x_i \mid \beta) \propto e^{-\mu_i} \mu_i^{x_i}$$

giving
$$\log f(x_i \mid \beta) = -\exp(\beta^T z_i) + (z_i^T \beta) x_i + \text{constant}.$$

Define
$$S_j(\beta) = \frac{\partial}{\partial \beta} \log f(x_j \mid \beta),$$

so that $\mathbb{E}\big(S_j(\beta)\big) = 0$. Then it can be shown, by applying a suitable variant of the CLT to $S_1(\beta), \ldots, S_n(\beta)$, that if

$$\hat{\beta} \quad \text{is the solution of} \quad \frac{\partial L_n}{\partial \beta}(\beta) = 0,$$

then, for large n, $\hat{\beta}$ is approximately normal, with mean vector β, and covariance matrix $\left(\mathbb{E}\Big(\frac{-\partial^2 L_n}{\partial \beta \partial \beta^T}\Big)\right)^{-1}$.

The asymptotic normality of the mle, for n independent observations, is used repeatedly in our application of glm.

1.6 Result 2: Wilks' Theorem

We state it again (slightly differently): let

$$x_1, \ldots, x_n \text{ be a r.s. from } f(x \mid \theta), \quad \theta \in \Theta \quad \text{where } \Theta \subset \mathbb{R}^r.$$

Procedure. To test $H_0 : \theta \in \omega$ against $H_1 : \theta \in \Omega$ where $\omega \subset \Omega \subset \Theta$, and ω, Ω, Θ are given sets, we reject ω in favour of Ω if and only if

$$2 \log R_n \equiv 2 \left[\max_{\theta \in \Omega} L_n(\theta) - \max_{\theta \in \omega} L_n(\theta) \right]$$

is too large, and we find the appropriate critical value by using
the asymptotic result: For large n, if ω true,

$$2 \log R_n \overset{\text{approx}}{\sim} \chi_p^2$$

where $p = \dim \Omega - \dim \omega$. (As for the mle, this result also holds for the more general case where $X_1, \ldots, X_n$ are independent, but not identically distributed).
We *prove* this very important theorem only for the following special case:
$\omega = \{\theta = \theta_0\}$, i.e. ω a point, hence of dimension 0, and $\Theta = \Omega$, assumed to be of dimension r.
Thus $p = r$.
(Even an outline proof of the theorem, in the case of general ω, Ω, takes several pages:

see for example Cox and Hinkley(1974).)
In the special case of ω a point,

$$2\log R_n = 2 \quad \big[L_n(\hat{\theta}_n) - L_n(\theta_0)\big],$$

where $\hat{\theta}_n$ maximises $L_n(\theta)$ subject to $\theta \in \Theta$, i.e. is the usual mle. Thus

$$L_n(\theta_0) \simeq L_n(\hat{\theta}_n) + (\theta_0 - \hat{\theta}_n)^T a(\hat{\theta}_n) + \tfrac{1}{2}(\theta_0 - \hat{\theta}_n)^T b(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)$$

where

$$\begin{aligned} a(\hat{\theta}_n) \quad &= \text{vector of first derivatives of } L_n(\theta) \text{ at } \hat{\theta}_n \\ b(\hat{\theta}_n) \quad &= \text{matrix of second derivatives of } L_n(\theta) \text{ at } \hat{\theta}_n. \end{aligned}$$

By definition of $\hat{\theta}_n$ as the mle, $a(\hat{\theta}_n) = 0$ (subject to the usual regularity conditions) and

$$-b(\hat{\theta}_n) \simeq \left(\mathbb{E}\left(\frac{-\partial^2 L_n}{\partial\theta_i\partial\theta_j}\right)\right)_{\text{at } \theta_0} = ni(\theta_0)$$

$$\text{giving} \qquad 2\big(L_n(\theta_0) - L_n(\hat{\theta}_n)\big) \simeq -(\theta_0 - \hat{\theta}_n)^T\big(ni(\theta_0)\big)(\theta_0 - \hat{\theta}_n)$$

$$\text{i.e.} \qquad 2\log R_n = 2\big(L_n(\hat{\theta}_n) - L_n(\theta_0)\big) \simeq (\hat{\theta}_n - \theta_0)^T\big(ni(\theta_0)\big)(\hat{\theta}_n - \theta_0).$$

But, if $\theta = \theta_0$, we know that

$$(\hat{\theta}_n - \theta_0) \overset{\text{approx}}{\sim} N\Big(0, \big(ni(\theta_0)\big)^{-1}\Big).$$

Hence, for $\theta \in \omega$,

$$2\log R_n \overset{\text{approx}}{\sim} \chi^2_p,$$

For this last step we have made use of the following lemma.
Lemma. If

$$\begin{array}{ll} \underset{r\times 1}{Z} \sim N_r(0,\Sigma), & \text{then} \quad Z^T\Sigma^{-1}Z \sim \chi^2_r \\ & \text{(provided that } \Sigma \text{ is of full rank).} \end{array}$$

Proof. By definition, $\Sigma = \mathbb{E}(ZZ^T)$, the covariance matrix of Z. For any fixed $r \times r$ matrix L,

$$LZ \sim N_r(0, L\Sigma L^T).$$

[recall that $\mathbb{E}(LZ) = L\mathbb{E}(Z) = 0, \quad \mathbb{E}(LZ)(LZ)^T = L[\mathbb{E}(ZZ^T)]L^T \quad .$]
But, Σ is an $r \times r$ positive-definite matrix, so we may choose L real, non-singular, such that $L\Sigma L^T = I_r$, the identity matrix, i.e. $\Sigma = L^{-1}(L^{-1})^T$.
Then $LZ \sim N_r(0, I_r)$, so that $(LZ)_1, \ldots, (LZ)_r$ are $NID(0,1)$ r.v.s. So, by definition of χ^2_r, the sum of squares of these $\sim \chi^2_r$. But this sum of squares is just

$$\begin{aligned} (LZ)^T(LZ), \quad &\text{i.e. } Z^T L^T LZ \\ &\text{i.e } Z^T\Sigma^{-1}Z. \end{aligned}$$

Hence $Z^T\Sigma^{-1}Z \sim \chi^2_r$ as required.
You can thus prove that it has mean r, variance $2r$.

1.7 Exponential Family Distributions

If the pdf of a single observation Y may be written in the form

$$f(y \mid \theta) = a(\theta)b(y)\exp\big(\tau(y)\pi(\theta)\big) \text{ for } y \in E$$

where E, the sample space, is free of θ, and $a(\cdot)$ is such that

$$\int_{y \in E} f(y \mid \theta)dy = 1,$$

we say that Y has an exponential family distribution. In this case, if $y_1, \ldots, y_n$ is the r.s. from $f(y \mid \theta)$, the likelihood of the sample is

$$f(y_1, \ldots, y_n \mid \theta) = \big(a(\theta)\big)^n b(y_1), \ldots, b(y_n)\exp\big(\pi(\theta)\sum_1^n \tau(y_i)\big)$$

and so $\sum_1^n \tau(y_i) \equiv t(y)$ is a sufficient statistic for θ. If, for $y \in E$,

$$f(y \mid \pi) = a(\pi)b(y)\exp\big(\tau(y)\pi\big), \quad \int_{y \in E} f(y \mid \pi)dy = 1,$$

we say that Y has an exponential family distribution,with **natural** parameter π.
The k-parameter generalisation of this is

$$f(y \mid \pi_1, \ldots, \pi_k) = a(\pi)b(y)\exp\big(\sum_1^k \pi_i\tau_i(y)\big),$$

in which case $(\pi_1, \ldots, \pi_k)$ are the natural parameters, and by writing down

$$\prod_1^n f(y_j \mid \pi),$$

you will see that

$$\big(t_1 \equiv \sum_1^n \tau_1(y_j), \ldots, t_k \equiv \sum_1^n \tau_k(y_j)\big)$$

is a set of sufficient statistics for $(\pi_1, \ldots, \pi_k)$.
Exponential families have many nice properties. Several well-known distributions, for example normal (ie Gaussian), Poisson and binomial, are of exponential family form. Here is one nice property.

1.8 Maximum likelihood estimation and exponential families

Assume $f(y \mid \pi)$ is as defined above, with π a scalar parameter. Then, if $y_1, \ldots, y_n$ is a random sample from $f(y \mid \pi)$, we see that

$$L_n(\pi) = n\log a(\pi) + \pi t(y) + \text{constant, where } t(y) \equiv \sum_1^n \tau(y_i).$$

(**) Hence
$$\frac{\partial L_n}{\partial \pi} = \frac{na'(\pi)}{a(\pi)} + t(y).$$

But $\big(a(\pi)\big)^{-1} = \int_{y\in E} b(y)e^{\pi\tau(y)}dy$ since $f(y \mid \pi)$ is a pdf. Differentiate with respect to π.

$(*)$Thus $\dfrac{-a'}{a^2} = \int \tau(y)b(y)e^{\pi\tau(y)}dy$

so $\dfrac{-a'}{a} = \int a(\pi)\tau(y)b(y)e^{\pi\tau(y)}dy = \mathbb{E}(\tau(Y)).$

Further, from (**)
$$\frac{\partial^2 L}{\partial \pi^2} = n\frac{\partial}{\partial \pi}\left(\frac{a'(\pi)}{a(\pi)}\right)$$
$$= n\left[\frac{a''}{a} - \left(\frac{a'}{a}\right)^2\right].$$

But, differentiating (*) gives
$$\frac{-a''}{a^2} + \frac{2(a')^2}{a^3} = \int \big(\tau(y)\big)^2 b(y)e^{\pi\tau(y)}dy$$
so
$$\frac{-a''}{a} + \frac{2(a')^2}{a^2} = \mathbb{E}\big(\tau(Y)\big)^2$$
giving
$$\frac{-a''}{a} + \left(\frac{a'}{a}\right)^2 = \mathrm{var}\big(\tau(Y)\big).$$

Hence for all π
$$\frac{\partial^2 L}{\partial \pi^2} = -n\mathrm{var}\big(\tau(Y)\big) < 0.$$

Hence, if $\hat{\pi}$ is a solution of $\frac{\partial L}{\partial \pi} = 0$, it is *the* maximum of $L(\pi)$. Furthermore, we may rewrite
$$\left.\frac{\partial L}{\partial \pi}\right|_{\hat{\pi}} = 0$$
as
$$t(y) = \mathbb{E}\big(t(Y)\big)\Big|_{\pi=\hat{\pi}}$$
that is, at the mle, the observed and expected values of $t(y)$ agree exactly.
The multiparameter version of this result, which is proved similarly, is the following :

If
$$f(y_i \mid \pi) = a(\pi)b(y_i)\exp\left(\sum_1^k \pi_j\tau_j(y_i)\right)$$

is the pdf of Y_i, where π is now a $k-$dimensional vector, then

$$\left(\frac{-\partial^2 L_n}{\partial \pi_j \partial \pi_{j'}}\right)$$

is a positive-definite matrix, i.e. $L_n(\pi)$ is a CONCAVE function of π. This nice property of the shape of the loglikelihood function makes parameter estimation for exponential families relatively straightforward.

Chapter 2

The Generalised Linear Model

2.1 Introduction to glm

Our methods are suitable for the following types of statistical problem, all of which have n independent observations, and some regression structure:
(i) *The usual linear regression model*

$$Y_i \sim NID(\mu_i, \sigma^2), 1 \leq i \leq n$$

where $\mu_i = \beta^T x_i$ and x_i a given covariate of dimension p, and β, σ^2 are both unknown. For example, $\mu_i = \beta_1 + \beta_2 x_i$, where x_i is scalar, and so β is of dimension 2, and we might want to estimate β_2, β_1, to test $\beta_2 = 0$, and so on.
(ii) *Poisson regression*

$$Y_i \text{ independent } Po(\mu_i), \ \log \mu_i = \beta^T x_i, 1 \leq i \leq n$$

(note that $\mu_i > 0$, by definition).
More generally, we might suppose that

$$g(\mu_i) = \beta^T x_i,$$

where $g(\cdot)$ is a known function, β is an unknown vector, and x_i is a known covariate vector.
(iii) *Binomial regression*

$$Y_i \text{ independent } Bi(r_i, \pi_i)$$

where π_i depends on x_i, a known covariate, for $1 \leq i \leq n$. For example, in a pharmaceutical experiment, we may have data (Y_i, r_i, x_i) where r_i =number of patients given a dose x_i of a new drug, and Y_i =number of these giving *positive* response to this drug (e.g. cured).
Suppose that we observe that Y_i/r_i tends to increase with x_i and we want to model this relationship,
For example we may wish to find the x which will give $\mathbb{E}(Y/r) = .90$, that is the dose which gives a 90% cure rate. Additionally, we may seek to compare the performance of this drug with a well-established drug. We might find that a simple plot of Y/r against dose for each of the old and the new drugs suggests that the old drug is better than the new at low doses, but the new drug better than the old at higher doses.

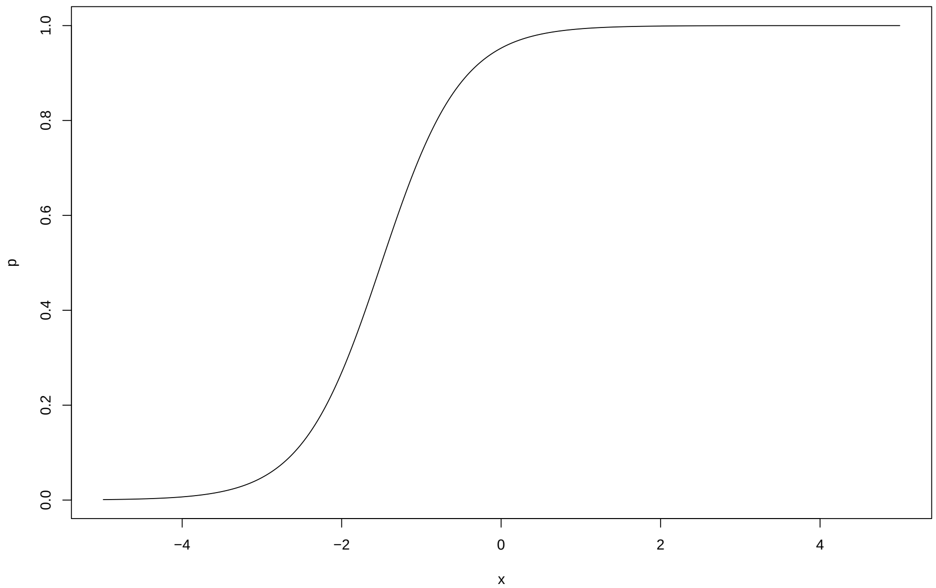

Figure 2.1: An example of a logistic function

Thus we seek a model in which π_i is a function of x_i, but we must take account of the constraint $0 < \pi_i < 1$. This means that $\pi_i = \beta_1 + \beta_2 x_i$ is not a suitable model, but

$$\log \frac{\pi_i}{1-\pi_i} = \beta_1 + \beta_2 x_i$$

a **logistic** model, often works well. Thus we take

$$g\big(\mathbb{E}(Y_i/r_i)\big) = \text{a linear function of } x_i,\ 1 \le i \le n$$

where

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

is the 'link function', so-called because it links the expected value of the response variable Y_i to the explanatory covariates x_i.
Verify that this particular choice of $g(\ )$ gives

$$\pi_i = \big(\exp(\beta_1 + \beta_2 x_i)\big) \Big/ \big(1 + \exp(\beta_1 + \beta_2 x_i)\big)$$

so that $\pi_i \uparrow$ as $x_i \uparrow$ for $\beta_2 > 0$).
An example of such a function, here with $p = \pi$ and $\beta_1 = 3, \beta_2 = 2$ is given in Figure 2.1.
(iv) *Contingency tables* (a less obvious application of glm).
For example, suppose $(N_{ij}) \sim Mn\big(n; (p_{ij})\big)$, where $\sum\sum p_{ij} = 1$ ie (N_{ij}) has the multinomial distribution, with parameters $n, (p_{ij})$.
For example, N_{ij} might be number of people of ethnic group i voting for political party j in a sample of size n, for $1 \le i \le I, 1 \le j \le J$.
Suppose that the problem of interest is to test $H_0 : p_{ij} = \alpha_i \beta_j$ for all (i, j), where $(\alpha_i), (\beta_j)$ unknown and $\sum \alpha_i = \sum \beta_j = 1$,

that is, to test the hypothesis that ethnic group and party are independent.
Note that $E(N_{ij}) = np_{ij}$, so that

$$\log \mathbb{E}(N_{ij}/n) = \log p_{ij},$$

and thus, under H_0,

$$\log p_{ij} = \log \alpha_i + \log \beta_j,$$

equivalently

$$\log \mathbb{E}(N_{ij}) = \text{const} + a_i + b_j \text{ for some } a, b.$$

Thus, in terms of $\log \mathbb{E}(N_{ij})$, testing H_0 is equivalent to testing a hypothesis which is *linear in the unknown parameters.*

All of the above problems fall within the same general class, and we can exploit this fact to do the following:
(a) We use the same algorithm to evaluate the maximum likelihood estimates of the parameters, and their (asymptotic) standard errors.
From now on we use the abbreviation **se** to denote standard error. The se is the square root of the **estimated variance**.
(b) We test the adequacy of our models, usually by Wilks' theorem.

2.2 Exponential families revisited

We will need to be able to work with the case where $Y_1, \ldots, Y_n$ are independent but not identically distributed, so we study the following general form for the distribution of $Y_1, \ldots, Y_n$.
Here we use standard glm notation, see for example Aitkin et al., p. 322.
Take $Y_1, \ldots, Y_n$ independent and assume that Y_i has pdf

$$f(y_i \mid \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\phi}\right] \times \exp\, c(y_i, \phi).$$

Thus

$$\log f(y_i \mid \theta_i) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi).$$

Assume further that $E(Y_i) = \mu_i$ (we will see that μ_i is a function of θ_i only), and that there exists a known function $g(\cdot)$ such that

$$g(\mu_i) = \beta^T x_i$$

where x_i is known, and β is unknown.
Our problem, in general, is the estimation of β. This naturally includes finding the se of the estimator. The parameter ϕ, which in general is also unknown, is called the *scale* parameter.
First we use simple calculus to find expressions for the mean and variance of Y. For

convenience we drop the suffix i for this Lemma.
Lemma 1. If Y has pdf

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

then for all θ, ϕ,

$$\mathbb{E}(Y) = b'(\theta), \mathrm{var}(Y) = \phi b''(\theta).$$

Proof.

$$\log f(y \mid \theta, \phi) = \big(y\theta - b(\theta)\big)/\phi + c(y, \phi).$$

Hence

$$\frac{\partial}{\partial\theta} \log f(y \mid \theta, \phi) = \big(y - b'(\theta)\big)/\phi,$$

and so

$$\frac{\partial^2}{\partial\theta^2} \log f(y \mid \theta, \phi) = -b''(\theta)/\phi.$$

But for all θ, ϕ

$$\int_y f(y \mid \theta, \phi) dy = 1.$$

Thus, asssuming that we can interchange $\int_y$ and $\frac{\partial}{\partial\theta}$, we see that

$$\mathbb{E}\left(\frac{\partial}{\partial\theta} \log f(Y \mid \theta, \phi)\right) = 0$$

thus $\mathbb{E}(Y) = b'(\theta)$. Similarly,

$$0 = \int \frac{\partial^2}{\partial\theta^2} f(y \mid \theta, \phi) dy = \int \left\{\left(\frac{\partial^2}{\partial\theta^2} \log f\right) f + \left(\frac{\partial}{\partial\theta} \log f\right)^2 f\right\} dy$$

giving

$$0 = \mathbb{E}\left(\frac{\partial^2}{\partial\theta^2} \log f\right) + \mathbb{E}\left(\frac{\partial}{\partial\theta} \log f\right)^2$$

giving

$$\mathbb{E}\left(\frac{Y - b'(\theta)}{\phi}\right)^2 = \frac{b''(\theta)}{\phi}$$

i.e.

$$\mathrm{var}(Y) = \phi b''(\theta).$$

Hence, returning to data $y_1, \ldots, y_n$, we see that the loglikelihood function is, say,

$$\ell(\beta) = \sum_{i=1}^{n} \big(y_i\theta_i - b(\theta_i)\big)/\phi + \sum_{i=1}^{n} c(y_i, \phi).$$

(This is in fact $\ell(\beta, \phi)$, but for the present we suppress ϕ.) Thus

$$\frac{\partial\ell}{\partial\beta} \equiv s(\beta) \text{ (say) } = \sum_{i=1}^{n} \frac{\big(y_i - b'(\theta_i)\big)}{\phi} \frac{\partial\theta_i}{\partial\beta}$$

where we have used the chain rule of differentiation, viz.

$$\frac{\partial}{\partial \beta}(\cdot) = \frac{\partial}{\partial \theta_i}(\cdot)\,\frac{\partial \theta_i}{\partial \beta} \text{ for each } i.$$

But $g(\mu_i) = \beta^T x_i$, and so we see that, because $\mu_i = b'(\theta_i)$,

$$g\big(b'(\theta_i)\big) = \beta^T x_i,$$

hence, on taking $\frac{\partial}{\partial \beta}$, we see that

$$g'\big(b'(\theta_i)\big)b''(\theta_i)\frac{\partial \theta_i}{\partial \beta} = x_i$$

that is

$$g'(\mu_i)b''(\theta_i)\frac{\partial \theta_i}{\partial \beta} = x_i.$$

$$\frac{\partial \ell}{\partial \beta} = s(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{\phi g'(\mu_i)b''(\theta_i)}$$

$$\frac{\partial \ell}{\partial \beta} = s(\beta) = \sum_{1}^{n} \frac{(y_i - \mu_i)}{g'(\mu_i)V_i}\, x_i$$

where $V_i = \text{var}(Y_i) = \phi b''(\theta_i)$; see Lemma 1.
The vector $s(\beta)$ is called the **score vector** for the sample, and $\hat{\beta}$ is found as the solution of $\frac{\partial \ell}{\partial \beta} = 0$, i.e. $s(\beta) = 0$.
In general this set of equations needs to be solved iteratively, so we will need $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$,the matrix of second derivatives of the loglikelihood. In fact glm works with $\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)$: to find this we use
Lemma 2.

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right) = -\mathbb{E}\left(\frac{\partial \ell}{\partial \beta}\,\frac{\partial \ell}{\partial \beta^T}\right).$$

Proof. Write $\ell(\beta) = \log f(y \mid \beta, \phi)$. Then for all β (and all ϕ)

$$\int_y f(y \mid \beta) dy = 1$$

Thus

$$\frac{\partial}{\partial \beta}\int_y f(y \mid \beta) dy = 0$$

$$\mathbb{E}\left(\frac{\partial}{\partial \beta}\,\ell(\beta)\right) = 0 \text{ (a vector)}$$

and

$$\frac{\partial^2}{\partial \beta \partial \beta^T}\int_y f(y \mid \beta) dy = 0 \text{ (a matrix).}$$

But

$$\int \frac{\partial^2}{\partial \beta \partial \beta^T}\, f(y \mid \beta) dy = \mathbb{E}\left(\frac{\partial^2}{\partial \beta \partial \beta^T}\,\log f(y \mid \beta)\right) + \mathbb{E}\left(\frac{\partial}{\partial \beta}\,\ell(\beta)\,\frac{\partial}{\partial \beta^T}\,\ell(\beta)\right)$$

hence

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = -\mathbb{E}\left(\frac{\partial \ell}{\partial\beta}\,\frac{\partial \ell}{\partial\beta^T}\right).$$

This concludes the proof of Lemma 2.
We may apply this Lemma to obtain a simple expression for the expected value of the matrix of second derivatives. Now

$$\frac{\partial \ell}{\partial\beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{g'(\mu_i)V_i}$$

and $\mathbb{E}(y_i - \mu_i) = 0$, and $y_1, \ldots, y_n$ are independent. Hence

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = -\mathbb{E}\left(\sum_1^n \frac{(y_i-\mu_i)^2}{\left(g'(\mu_i)V_i\right)^2}\, x_i x_i^T\right)$$

$$= -\sum_1^n \frac{V_i}{\left(g'(\mu_i)\right)^2 V_i^2}\, x_i x_i^T$$

$$= -\sum_1^n w_i x_i x_i^T \;\text{ say, } w_i \equiv 1/\left(V_i\left(g'(\mu_i)\right)^2\right).$$

We write W as the diagonal matrix

$$\begin{pmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & w_{n-1} & 0 \\ 0 & 0 & \cdots & 0 & w_n \end{pmatrix}$$

and thus we see

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = -X^T W X \qquad \text{.................}\textbf{Expectation}.$$

where X is the $n \times p$ matrix defined by

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}.$$

Hence we can say that if $\hat\beta$ is the solution of $s(\beta) = 0$, then $\hat\beta$ is asymptotically normal, with mean β and covariance matrix having as inverse

$$-\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = X^T W X.$$

2.3 Reminder: The Newton-Raphson algorithm

This is how we solve
$$\frac{\partial \ell(\beta)}{\partial \beta} = 0.$$
Take β_0 as the 'starting value'. Expanding about $\beta = \beta_0$, we note that
$$\left.\frac{\partial \ell(\beta)}{\partial \beta}\right|_{\beta_1} \simeq \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0} + \left.\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right|_{\beta_0} (\beta_1 - \beta_0).$$
Set the left hand side $= 0$ (because we seek $\hat\beta$ such that $\frac{\partial \ell}{\partial \beta} = 0$ at $\beta = \hat\beta$).
Then find β_1 from β_0 by
$$0 = \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0} + \left.\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)\right|_{\beta_0} (\beta_1 - \beta_0) \quad \text{............}\textbf{Iteration}.$$
giving β_1 as a linear function of β_0.
Now find β_2 from β_1 by
$$0 = \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_1} + \left.\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)\right|_{\beta_1} (\beta_2 - \beta_1)$$
giving β_2 as linear function of β_1, and so on.
This process gives $\beta_\nu \to \hat\beta$. Convergence for glm examples is usually remarkably quick: in practice we stop the iteration when $\ell(\beta_\nu)$ and $\ell(\beta_{\nu-1})$ are sufficiently close, and this may only require 4 or 5 iterations. (But note that some extreme configurations of data, for example zero frequencies in binomial regression, may have the effect that the loglikelihood function does not have a finite maximum. In this case the glm algorithm should report the failure to converge, and may give strangely large parameter estimates with very large standard errors.)
In the glm algorithm the matrix $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$ is replaced in **Iteration** by its expectation, from **Expectation**.
The inverse covariance matrix
$$-\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)$$
of $\hat\beta$ is estimated by replacing β by $\hat\beta$. In addition, ϕ is replaced by $\hat\phi$, but in any case $\phi = 1$ for the binomial and Poisson distributions. The estimation of ϕ for the normal distribution will be discussed further below.
Example 1. $Y_i \sim NID(\beta^T x_i, \sigma^2)$, $1 \le i \le n$. Take the special case $\beta^T x_i = \beta x_i$, i.e. linear regression through the origin. Thus
$$f(y_i \mid \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \; \exp -\frac{1}{2\sigma^2}(y_i - \beta x_i)^2$$
giving
$$\log f(y_i \mid \beta) = +\frac{1}{\sigma^2}\left(\beta y_i x_i - \frac{\beta^2}{2} x_i^2\right) - \frac{y_i^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$
which is of the form
$$\big(y_i\theta_i - b(\theta_i)\big)/\phi + c(y_i, \phi)$$

with

$$b'(\theta_i) = \mu_i = \beta x_i,\ g(\mu_i) = \mu_i,\ \phi = \sigma^2$$

and

$$\theta_i = \beta x_i,\ b(\theta_i) = \theta_i^2/2.$$

[Hence $b''(\theta_i) = 1$, $\text{var}(Y_i) = \phi b''(\theta_i)$: *check*.]
In this case, it is trivial to show directly that $\hat{\beta} = \sum x_i Y_i / \sum x_i^2$.
What does the glm algorithm do? If we substitute in

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

and

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta^2}\right) = -\sum w_i x_i^2, \text{ where } w_i^{-1} = V_i\big(g'(\mu_i)\big)^2$$

we see that here

$$\frac{\partial \ell}{\partial \beta} = \sum (y_i - \beta x_i)x_i/\sigma^2$$

and

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta^2}\right) = -\sum x_i^2/\sigma^2$$

so the glm iteration evaluates β_1 from β_0 by

$$0 = \frac{\sum (y_i - \beta_0 x_i)x_i}{\sigma^2} - (\beta_1 - \beta_0)\frac{\sum x_i^2}{\sigma^2}$$

(thus the precise choice of β_0 is irrelevant), giving

$$\beta_1 = \sum x_i y_i / \sum x_i^2 = \hat{\beta}.$$

Hence only one iteration is needed to attain the mle. (One iteration will always be enough to maximise a quadratic loglikelihood function.)
Furthermore, from the fact that $\hat{\beta} = \sum x_i Y_i / \sum x_i^2$, where Y_i are independent, each with variance σ^2, it is easy to see directly that the exact distribution of $\hat{\beta}$ is normal, mean β, and $\text{var}(\hat{\beta}) = \sigma^2 / \sum x_i^2$, (agreeing, of course, with the asymptotic distribution). The general glm formula gives us

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta^2}\right) = -\sum w_i x_i^2 = -\sum\ x_i^2/\sigma^2,$$

and hence the general glm formula gives us

$$\text{var}(\hat{\beta}) \simeq \sigma^2 / \sum x_i^2$$

(consistent with the above exact variance, of course).
Example. Repeat the above, but now taking

$$Y_i \sim NID(\beta_1 + \beta_2 x_i, \sigma^2)\ 1 \le i \le n$$

i.e. the usual linear regression, with $\sum x_i = 0$ (without loss of generality). Thus now you are maximising a function of 2 parameters, so you will you need to find $\frac{\partial \ell}{\partial \beta_1}$, $\frac{\partial \ell}{\partial \beta_2}$, and so on.
You should find, again, that the glm algorithm needs only one iteration to reach the well-known mle

$$\hat{\beta}_1 = \bar{y},\ \hat{\beta}_2 = \sum x_i y_i / \sum x_i^2,$$

regardless of the position of the starting point (β_{10}, β_{20}).

Example 2. Assume that

$$Y_i \text{ independent } Bi(1, \mu_i),\ 1 \leq i \leq n$$

and

$$\log\bigl(\mu_i/(1-\mu_i)\bigr) = \beta x_i \text{ say,}$$

ie

$$g(\mu_i) = \beta x_i,$$

thus defining $g(\cdot)$ as the link function. Then

$$P(Y_i = y_i \mid \mu_i) = f(y_i \mid \mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$$

giving

$$\log f(y_i \mid \mu_i) = y_i \log \frac{\mu_i}{1-\mu_i} + \log(1-\mu_i)$$

which we can rewrite in the general glm form as

$$\log f(y_i \mid \mu_i) = \bigl(y_i\theta_i - b(\theta_i)\bigr)/\phi \quad \text{where } \phi = 1$$

and

$$\theta_i = \log\bigl(\mu_i/(1-\mu_i)\bigr),\ b(\theta_i) = -\log(1-\mu_i).$$

Thus

$$\mu_i = e^{\theta_i}/(1+e^{\theta_i}),\ b(\theta_i) = +\log(1+e^{\theta_i})$$

giving

$$b'(\theta_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}},\ b''(\theta_i) = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = \mu_i(1-\mu_i)$$

all of which, of course, agrees with what we already know, that $Y_i \sim Bi(1, \mu_i)$ implies that $\mathbb{E}(Y_i) = \mu_i,\ \text{var}(Y_i) = \mu_i(1-\mu_i)$.
Furthermore,

$$\ell(\beta) = \sum y_i \beta x_i - \sum \log(1+e^{\beta x_i})$$

(remembering that $g(\mu) = \log\bigl(\mu/(1-\mu)\bigr)$). Hence

$$\frac{\partial \ell}{\partial \beta} = \sum x_i y_i - \sum x_i \frac{e^{\beta x_i}}{1+e^{\beta x_i}}.$$

So we can see at once that the only way to solve $\frac{\partial \ell}{\partial \beta} = 0$ is by iteration. Now

$$\frac{\partial \ell}{\partial \beta} = \sum x_i y_i - \sum x_i \left(1 - \frac{1}{1 + e^{\beta x_i}}\right)$$

Thus

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\sum x_i^2 \frac{e^{\beta x_i}}{(1 + e^{\beta x_i})^2} = \mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta^2}\right)$$

i.e.

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta^2}\right) = -\sum w_i x_i^2,\ w_i = \frac{1}{V_i\big(g'(\mu_i)\big)^2}$$

where

$$V_i = \mu_i(1 - \mu_i),\ g(\mu_i) = \log\big(\mu_i/(1 - \mu_i)\big).$$

You should now check this.
This time, to compute $\hat{\beta}$, we find β_1 from β_0 by

$$0 = \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0} + \left.\left(\frac{\partial^2 \ell}{\partial \beta^2}\right)\right|_{\beta_0} (\beta_1 - \beta_0)$$

and so on. This process converges to $\hat{\beta}$, where

$$\hat{\beta} \overset{\text{approx}}{\sim} N\big(\beta, v_n(\beta)\big)$$

where $v_n(\beta) = 1/\sum w_i x_i^2$,

$$w_i = \frac{e^{\beta x_i}}{(1 + e^{\beta x_i})^2}$$

which may be estimated by replacing β by $\hat{\beta}$.
Exercise. Repeat the above with $Y_i \sim Po(\mu_i)$, $\log \mu_i = \beta x_i$, i.e. the Poisson distribution and the log link function. (You will find this gives $\phi = 1$ again.)

2.4 The Canonical Link functions

In general in glm models, $\mathbb{E}(Y_i) = \mu_i$, $g(\mu_i) = \beta^T x_i$ and the matrix $\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$ may be different from the matrix $\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)$. But for a given exponential family $f(\cdot)$, there is a 'canonical link function' $g(\cdot)$ such that these two matrices are the same.
If $g(\cdot)$ is such that we can write the loglikelihood $\ell(\beta)$ as

$$\ell(\beta) = \big(\sum_1^p \beta_\nu t_\nu(y) - \psi(\beta)\big)/\phi + \text{constant}$$

where $\psi(\beta)$ is free of y [and $t_1(y), \ldots, t_p(y)$ are of course the sufficient statistics], then $g(\cdot)$ is said to be the canonical link function. In this case

$$\frac{\partial \ell}{\partial \beta} = \frac{[t(y) - \frac{\partial \psi}{\partial \beta}]}{\phi}$$

and

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = -\frac{1}{\phi} \frac{\partial^2 \psi}{\partial \beta \partial \beta^T}$$

which is not a random variable. Hence

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right) = \frac{\partial^2 \ell}{\partial \beta \partial \beta^T} \text{ for all } y.$$

Verify: If $Y_i \sim Po(\mu_i)$, $g(\mu_i) = \beta_1 + \beta_2 x_i$, then $g(\mu) = \log \mu$ is a canonical link function. What are $(t_1(y), t_2(y))$ in this case?
Exercise (1) Take $Y_i \sim Bi(1, \mu_i)$, thus $\mu_i \in [0, 1]$. Take as link $g(\mu_i) = \Phi^{-1}(\mu_i)$, the *probit* link, where Φ is the distribution function of $N(0, 1)$. (Take $g(\mu_i) = \beta x_i$.) Show this is *not* the canonical link function.
Exercise (2) Suppose, for simplicity, that β is of dimension 1, and the loglikelihood

$$\ell(\beta) = \big(\beta t(y) - \psi(\beta)\big)/\phi.$$

Prove that

$$\text{var } t(Y) = \phi\left(\frac{\partial^2 \psi}{\partial \beta^2}\right).$$

and hence that

$$\frac{\partial^2 \ell}{\partial \beta^2} < 0 \text{ for all } \beta.$$

Hence any stationary point of $\ell(\beta)$ is *the* unique maximum of β. Generalise this result to the case of vector β.
Solution.

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\phi}(t(y) - \frac{\partial \psi}{\partial \beta})$$

and we know that this has expectation 0. Similarly

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\frac{1}{\phi}\frac{\partial^2 \psi}{\partial \beta^2}$$

and this is free of the random quantity Y. Further, we can quote the general result that

$$\mathbb{E}(\frac{\partial \ell}{\partial \beta}\frac{\partial \ell}{\partial \beta}) = \mathbb{E}(-\frac{\partial^2 \ell}{\partial \beta^2}).$$

In this case this expectation is just

$$(-\frac{\partial^2 \ell}{\partial \beta^2}),$$

since we know that this is free of Y. Hence

$$var(t(Y)) = \phi\frac{\partial^2 \psi}{\partial \beta^2}$$

and we know that because this expression is a variance, it must be > 0. Thus

$$-\frac{\partial^2 \ell}{\partial \beta^2} > 0,$$

hence $\ell(\beta)$ is a strictly concave function. Thus if it has a stationary point, ie a solution of $\partial\ell/\partial\beta = 0$, then this point must be the unique maximum of $\ell(\beta)$.
The generalization of this result to the matrix version uses the facts that

$$\mathbb{E}(\frac{\partial \ell}{\partial \beta}) = 0,$$

and the covariance matrix of the vector $\frac{\partial \ell}{\partial \beta}$ is

$$\mathbb{E}(-\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}).$$

Now apply the fact that a covariance matrix must be positive definite, and hence show that $\ell(\beta)$ is a strictly concave function of the vector β.

2.5 Testing hypotheses about β, and a measure of the goodness of fit

Returning to our original glm model, with loglikelihood for observations $Y_1, \ldots, Y_n$ as

$$\ell(\beta, \phi) = \sum_1^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} \qquad \textbf{(glm)}$$

with $\mathbb{E}(Y_i) = \mu_i$, $g(\mu_i) = \beta^T x_i$, where x_i given, we proceed to work out ways of testing hypotheses about the components of β.
(i) If, for example, we just want to test $\beta_1 = 0$ where

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

then we can find $\hat{\beta}_1$, and $se(\hat{\beta}_1)$ its standard error, and refer $|\hat{\beta}_1|/se(\hat{\beta}_1)$ to $N(0,1)$. We reject $\beta_1 = 0$ if this is too large. The quantity $se(\hat{\beta}_1)$ is of course obtained as the square root of the $(1,1)^{\text{th}}$ element of the inverse of the matrix

$$\left. \frac{-\partial^2 \ell}{\partial \beta \partial \beta^T} \right|_{\hat{\beta}}.$$

So here we are using the asymptotic normality of the mle $\hat{\beta}$, together with the formula for its asymptotic covariance matrix.
(ii) If we want to test $\beta = 0$, we can use the fact that, asymptotically, $\hat{\beta} \sim N\big(\beta, V(\beta)\big)$, say. Hence

$$(\hat{\beta} - \beta)^T \big(V(\hat{\beta})\big)^{-1} (\hat{\beta} - \beta) \sim \chi^2_p,$$

approximately, so that to test $\beta = 0$, just refer $\hat{\beta}^T \big(V(\hat{\beta})\big)^{-1} \hat{\beta}$ to χ^2_p.
Similarly we could find an approximate $(1-\alpha)$-confidence region for β by observing that, with c defined in the obvious way from the χ^2_p distribution,

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

giving an ellipsoidal confidence region for β centred on $\hat{\beta}$. This procedure can be adapted, in an obvious way, to give a $(1-\alpha)$-confidence region for, say, $\binom{\beta_1}{\beta_2}$.
(iii) But we are more likely to want to test hypotheses about (vector) components of β; for example with

$$Y_i \sim NID(\mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$$

we may wish to test $\binom{\beta_2}{\beta_3} = \binom{0}{0}$, or, if

$$Y_{ij} \sim Po(\mu_{ij}),\ 1 \le i \le r, 1 \le j \le s,$$

with

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij},\ 1 \le i \le r, 1 \le j \le s,$$

we may wish to test $\gamma_{ij} = 0$ for all i, j.
In general, with $\ell(\beta)$ as in **(glm)** above, suppose that we wish to test $\beta \in \omega_c$ (the 'current model') against $\beta \in \omega_f$ (the 'full model'), where $\omega_c \subset \omega_f$ (and ω_c, ω_f are linear hypotheses). Assume that ϕ is known. Define $S(\omega_c, \omega_f) = 2(L_f - L_c)$, where L_f, L_c are loglikelihoods maximised on ω_f, ω_c respectively. Then

$$S(\omega_c, \omega_f) = 2 \sum \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right] / \phi$$

where $\hat{\theta}_i$ = mle under ω_c, $\tilde{\theta}_i$ = mle under ω_f. Define $D(\omega_c, \omega_f) = \phi S(\omega_c, \omega_f)$. Then $D(\omega_c, \omega_f)$ is termed the deviance of ω_c relative to ω_f ,
and $S(\omega_c, \omega_f)$ is termed the scaled deviance of ω_c relative to ω_f.

2.6 Distribution of the scaled deviance

If ω_c is true, then

$$S(\omega_c, \omega_f) \overset{\text{approx}}{\sim} \chi^2_{t_1 - t_2},\ \text{where } t_1 = \dim(\omega_f),\ \text{and } t_2 = \dim(\omega_c).$$

This result is *exact* for normal distributions with $g(\mu) = \mu$.
A practical difficulty, and how to solve it. In practice, for normal distributions, ϕ is generally unknown (for binomial and Poisson, $\phi = 1$). In this case we replace ϕ by its *estimate* under the full model, and for the normal distribution we would then use the F distribution for our test of ω_c against ω_f.
This is discussed in greater detail (but still without a complete proof) below.

A highly important special case of a generalised linear model is that of the linear model with normal errors. This model, and its analysis, have been extensively studied, and there are many excellent text-books devoted to this one subject, demonstrating it to be both useful and beautiful. In this brief text, we introduce the reader to this topic in the next Chapter.

2.7 From recent Mathematical Tripos questions

Mathematical Tripos Part IIA, 1997 4/13
This is the 'Essay' question for 1997, designed to take the well-prepared candidate about 40 minutes.
Suppose that $Y_1, ..., Y_n$ are independent random variables, and that Y_i has probability density function

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

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

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

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

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

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

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

Solution
This is 'the calculus at the heart of glm': see your lecture notes for the full story. (Incidentally, this makes it a rather easy question for the diligent candidate.)
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 glm will solve, by iteration, the simultaneous equations

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

Mathematical Tripos 2000 Part IIA 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

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

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

with $x_1, ..., x_n$ known covariate values, each of dimension 1. Your solution should include a statement of the large-sample distribution of $\hat{\beta}$.

SOLUTION

(i)
a) Since

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

it follows that the likelihood function is $\exp \ell(\beta)$, and so by the Factorisation Theorem, $(t_1(Y), ..., t_p(Y))$ is sufficient for the vector β.
b) Now we know that in general

$$\mathbb{E}\left(\frac{\partial \ell}{\partial \beta}\right) = 0.$$

Here, we see that

$$\frac{\partial \ell}{\partial \beta} = (1/\phi)\big(t(y) - \frac{\partial \psi}{\partial \beta}\big).$$

Hence,

$$\mathbb{E}(t_\nu(Y)) = \frac{\partial \psi}{\partial \beta_\nu}$$

as required.
ii) We also know that in general

$$\mathbb{E}\left(-\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right) = \mathbb{E}\left(\frac{\partial \ell}{\partial \beta}\frac{\partial \ell}{\partial \beta^T}\right).$$

Here

$$-\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = \frac{1}{\phi}\left(\frac{\partial^2 \psi}{\partial \beta \partial \beta^T}\right),$$

which in fact is free of Y, the random vector. Hence

$$\frac{1}{\phi^2} cov(t(Y)) = \frac{1}{\phi}\frac{\partial^2 \psi}{\partial \beta \partial \beta^T},$$

giving the equation

$$cov(t(Y)) = \phi \frac{\partial^2 \psi}{\partial \beta \partial \beta^T}.$$

Since this is a covariance matrix, and we are told that $\phi > 0$, it follows that

$$-\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}$$

is a positive-definite matrix. Thus the function $\ell(\beta)$ is a concave function. This has the useful practical consequence that if we can find a solution of $\frac{\partial \ell}{\partial \beta} = 0$, we know it must be the unique maximum of the function $\ell(\beta)$.
(Much of this is already familiar to students from the lecture notes given above.)
The binary logistic regression example requires us to compute the term $\ell(\beta)$, and its first and second derivative. We must also state that for large n, the approximate distribution of $\hat{\beta}$ is $N(\beta, v_n(\beta))$, say, where

$$1/v_n(\beta) = \mathbb{E}\big(-\frac{\partial^2 \ell}{\partial \beta^2}\big).$$

(This is relatively easy to work out as we are told that β is scalar.)

Chapter 3

Regression for normal errors

3.1 Basic set-up and distributional results

Assume that

$$Y_i \sim NID(\beta^T x_i, \sigma^2)$$

where each of $\beta, x_1, \ldots, x_n$ is of dimension p. Assume further that $x_1, \ldots, x_n$ are linearly independent. Then we may rewrite our model in the following matrix form

$$Y \sim N_n(X\beta, \sigma^2 I_n)$$

X being called the 'design' matrix, which has rank p, since $x_1, \ldots, x_n$ are linearly independent.
We now write our model in the form $\mathbb{E}(Y) \in \omega_f$ where ω_f is the linear subspace of vectors of the form $X\beta$, for some β. For the present, we will assume that σ^2 is known. First we find the maximum likelihood estimator of β for $\mathbb{E}(Y) \in \omega_f$. In this case

$$f(y \mid \beta, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp -\frac{1}{2} \sum_1^n (y_i - \beta^T x_i)^2/\sigma^2,$$

equivalently,

$$f(y \mid \beta, \sigma^2) \propto \frac{1}{(\sigma^n)} \exp -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta).$$

Thus $\tilde{\beta}$, the mle of β under ω_f, minimises $(y - X\beta)^T(y - X\beta)$. Now

$$\frac{\partial}{\partial\beta}(y - X\beta)^T(y - X\beta) = \frac{\partial}{\partial\beta}(y^T y - 2(X^T y)^T\beta + \beta^T(X^T X)\beta)$$

$$= -2(X^T y) + 2(X^T X)\beta.$$

The matrix of second derivatives is easily seen to be $2(X^T X)$. For any vector u, $u^T(X^T X)u = (Xu)^T(Xu) \geq 0$, with equality if and only if $u = 0$, thus $(X^T X)$ is is a positive definite matrix.
Thus we can say that the quadratic form $(y - X\beta)^T(y - X\beta)$ attains its minimum at its stationary point, which in this case is given by

$$(X^T y) = (X^T X)\tilde{\beta}.$$

Further, since X is of rank p, so also is $X^T X$, and thus this matrix has a unique inverse $(X^T X)^{-1}$. Hence the mle of β, which is also the least squares estimator, is given by

$$\tilde{\beta} = (X^T X)^{-1} X^T y,$$

We now obtain the expression for the corresponding fitted value of y under ω_f. This is $\tilde{y}$, where

$$\tilde{y} = X\tilde{\beta} = X(X^T X)^{-1} X^T y.$$

We rewrite this equation as

$$\tilde{y} = P_f y,$$

thus defining P_f as $X(X^T X)^{-1} X^T$. A very important property of P_f is that it is a **projection matrix**. This means that it satisfies $P_f = P_f^T$ and $P_f P_f = P_f$; it is easy to check these two equations. You may also check that since X is of rank p, then so also is P_f. Furthermore, for any n-dimensional vector v say, $v^T P_f v = (P_f v)^T (P_f v) \geq 0$, hence P_f is a positive-semidefinite matrix.

Exercises.

i) Show that $\tilde{\beta}$ has the distribution $N(\beta, \sigma^2 (X^T X)^{-1})$.

This follows easily by writing $\tilde{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon)$, where $\epsilon \sim N(0, \sigma^2 I)$. Now recall that the distribution of a linear transform of a multivariate normal is again multivariate normal. (See Appendix 1.)

ii) Verify that

$$\max_{\omega_f} f(y \mid \beta, \sigma^2) = \frac{\text{const}}{\sigma^n} \exp -(y - \tilde{y})^T (y - \tilde{y})/2\sigma^2.$$

In a typical statistical modelling situation we are only interested in fitting the model $\omega_f : \mathbb{E}(Y) = X\beta$ for some β as our 'baseline' model. We will almost certainly want to compare ω_f, the 'full' model, with say ω_c, a 'constrained' model, where ω_c is a given linear subspace of ω_f. If we can 'explain' our data y with a simpler model (ie one using fewer parameters) then generally we will gain in two respects. Not only do we find that to interpret the simpler model gives us more insight than trying to interpret an unnecessarily complicated model, but also we will find that estimation for fewer parameters, loosely speaking, will be more accurate than the corresponding estimation in the larger model. In this sense, it pays to 'declutter' our statistical model whenever possible. We do this within the formal framework of linear models, building on the results given above.

First we must introduce suitable notation.

Partition X, β as $(X_1 \vdots X_2)$, $\binom{\beta_1}{\beta_2}$ respectively, so that $X\beta = X_1\beta_1 + X_2\beta_2$. Assume that β_1, β_2 are of dimensions p_1, p_2 respectively, and that X_1, X_2 are of ranks p_1, p_2 respectively, with $p_1 + p_2 = p$. Then, suppose we wish to test the hypothesis $H_0 : \beta_2 = 0$.

We see that H_0 can be rewritten as $H_0 : \mathbb{E}(Y) \in \omega_c$ where ω_c is a linear subspace of ω_f, which is R^p. Now we know from our derivation of $\tilde{\beta}$ given above, that the least squares estimator of β_1 under ω_c is say $\hat{\beta}_1$, where

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y.$$

Further, we can see that the 'fitted' value of y under ω_c must be

$$\hat{y} = P_c y,$$

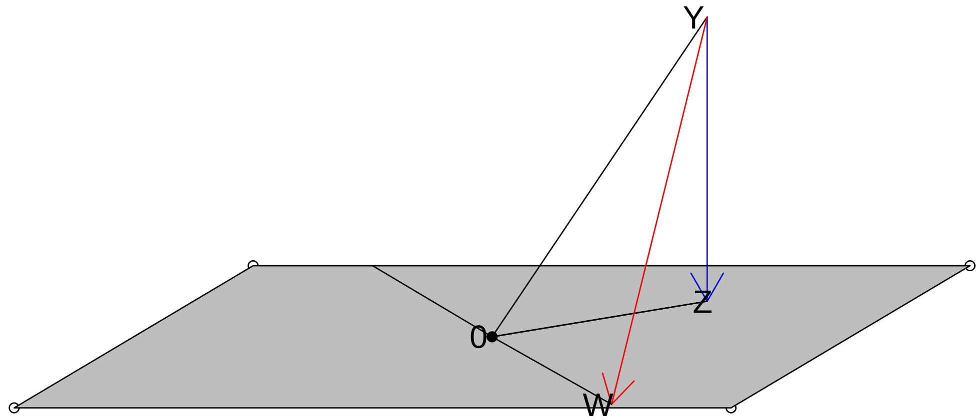

Figure 3.1: Projecting the vector Y down to the space ω_f, (blue arrow), and also to the subspace ω_c (red arrow).

where P_c is the projection of y onto ω_c. It is then easy to check that

$$\max_{\omega_c} f(y \mid \beta, \sigma^2) = \frac{\text{const}}{\sigma^n} \exp -(y - \hat{y})^T (y - \hat{y})/2\sigma^2.$$

Now since $P_c = X_1(X_1^T X_1)^{-1} X_1^T$, it is a projection matrix of rank p_1. You can verify that the scaled deviance is

$$S(\omega_c, \omega_f) = \left[-(y - \tilde{y})^T (y - \tilde{y}) + (y - \hat{y})^T (y - \hat{y}) \right] / \sigma^2.$$

We can illustrate this for ourselves with a sketch in which Y is of dimension 3, ω_f is a plane, ω_c is a line within ω_f,and all of Y, ω_f, ω_c pass through the point 0. (A vector subspace always passes through the origin.) This is shown in Figure 3.1 with $Z = P_f Y$ and $W = P_c Y$. Observe from your picture that

$$Q_c \equiv (y - \hat{y})^T (y - \hat{y}) \geq (y - \tilde{y})^T (y - \tilde{y}) \equiv Q_f,$$

Q_c, Q_f being the deviances in fitting ω_c, ω_f respectively.
The quantities Q_c, Q_f are very important. Here we are dealing with the **normal linear**

model, and Q_c, Q_f are usually referred to as the **residual sums of squares** fitting ω_c, ω_f respectively. We rewrite the equation

$$S(\omega_c, \omega_f) = \big[(y - P_c y)^T (y - P_c y) - (y - P_f y)^T (y - P_f y)\big]/\sigma^2$$

as

$$S = [y^T y - y^T P_c y - y^T y + y^T P_f y]/\sigma^2$$

giving (using $P_c^T P_c = P_c$, etc.)

$$S = y^T (P_f - P_c) y/\sigma^2.$$

But

$$(P_f - P_c)(P_f - P_c) = P_f - 2P_c + P_c = P_f - P_c \ ,$$

since $P_f P_c = P_c P_f = P_c$, (using the fact that $\omega_c \subset \omega_f$). Hence

$$S = y^T P y/\sigma^2$$

where P is the projection matrix, $P_f - P_c$. At this point we quote, without proof: If $y \sim N(\mu, \sigma^2 I)$ and $P\mu = 0$, then $y^T P y/\sigma^2 \sim \chi^2_r$, where $r = \text{rank } P$. Check that if $\mathbb{E}(Y) \in \omega_c$, then

$$(P_f - P_c)\mathbb{E}(Y) = 0.$$

Hence, to test $\mu \in \omega_c$ against $\mu \in \omega_f$, we refer

$$y^T P y/\sigma^2 \text{ to } \chi^2_r,$$

i.e. we refer [residual ss under ω_c- residual ss under $\omega_f]/\sigma^2$ to χ^2_r, where $r = \dim \omega_f - \dim \omega_c$. But in practice this result is not directly useful, because

$$\boxed{\sigma^2 \text{ is unknown}}.$$

It is not difficult for you to show that under ω_f, the mle of σ^2 is

$$(y - P_f)^T (y - P_f)/n.$$

This mle is not quite what we use for testing hypotheses about β. However, consideration of this mle leads us on to the following very important theorem.
Theorem. Suppose $Y \sim N(\mu, \sigma^2 I)$, where $\mu \in$ the linear subspace ω_f. Suppose the linear subspace $\omega_c \subset \omega_f$. Let

$$\tilde{\mu} = P_f Y, \ \hat{\mu} = P_c Y$$

where P_f is the projection onto ω_f, P_c the projection onto ω_c. As defined before, we take

$$Q_f = (Y - \tilde{\mu})^T (Y - \tilde{\mu})$$

and

$$Q_c = (Y - \hat{\mu})^T (Y - \hat{\mu}),$$

as the residual sums of squares fitting ω_f, and fitting ω_c respectively, so that by definition $Q_c \geq Q_f$. Then

$$Q_f/\sigma^2 \sim \chi^2_{df}$$

and

$$(Q_c - Q_f)/\sigma^2 \sim \chi^2_r \ \text{(noncentral)},$$

and these two random variables are **independent**.
The second term is *central* χ^2_r if and only if $\mu \in \omega_c$. Here

$$df = \text{degrees of freedom of } Q_f = n - \dim(\omega_f) = n - p$$

and

$$r = \dim(\omega_f) - \dim(\omega_c) = p_2.$$

Corollaries
(1) $\mathbb{E}(Q_f/df) = \sigma^2$
so that if $\mu \in \omega_f$, then Q_f/df (the 'mean deviance') is an unbiased estimate of σ^2.
(2) To test $\mu \in \omega_c$ against $\mu \in \omega_f$, we use

$$\frac{(Q_c - Q_f)/r}{Q_f/df}$$

which we refer to $F_{r,df}$, rejecting ω_c if this ratio is too large.
Exercise. Show that under ω_c, $P_fY - P_cY$ is normal, with mean 0 and covariance matrix

$$cov(P_fY) - cov(P_cY)$$

(this is a positive-definite matrix).

3.2 Proof of results about distributions of quadratic forms

To prove the theorem you do need to 'recall' some algebraic results. Omit this section if you do not have the necessary background in matrix algebra.
Proof of the above Theorem
We start by showing that

$$Q_f/\sigma^2 \sim \chi^2_{df}.$$

Here's how to proceed.
Recall that under ω_f, the random vector Y may be written $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I)$. Now we require the distribution of the quadratic form $Y^T(I - P_f)Y$, where $(I - P_f)$ is clearly a projection matrix, of rank $n - p = df$ in our previous notation. We note that $Y = X\beta + \epsilon$, and $P_fY = X\beta + P_f\epsilon$. Thus

$$Y^T(I - P_f)Y = \epsilon^T(I - P_f)\epsilon.$$

Now if B is any n by n projection matrix of rank df say, then it has eigen-values $\lambda_1, \ldots, \lambda_n$ say, and $\lambda_1 = 1, \ldots, \lambda_{df} = 1$, with the remaining λ_i's as zero. (It is easy to check that any eigen value of B is either 1 or 0, and then we need to recall the fact that the rank of a matrix is the number of its eigen-values that are non-zero.)

Let $u_1, u_2, \ldots, u_n$ be the eigen vectors corresponding to $\lambda_1 = 1, \ldots, \lambda_n$ respectively, so that $u_i^T u_i = 1, u_i^T u_j = 0$ for i, j distinct. We may write $B = \sum_1^n \lambda_i u_i u_i^T$. In this case we see, by taking $B = I - P_f$, that

$$(I - P_f) = \sum_1^{df} u_i u_i^T$$

and so

$$\epsilon^T (I - P_f)\epsilon = \sum_1^{df} Z_i^2$$

where $Z_i = u_i^T \epsilon$. It is thus easily checked that $Z_1, \ldots, Z_{df}$ are $NID(0, \sigma^2)$ random variables, and so we see (from the definition of the χ^2 distribution) that

$$Y^T (I - P_f) Y/\sigma^2 \sim \chi^2_{df},$$

which is the required result. We remind the reader that $Y^T(I-P_f)Y$ is called the 'residual sum of squares under ω_f'.
Observe that $(P_f - P_c)Y$ and $(I - P_f)Y$ are independent, since $(P_f - P_c)(I - P_f)^T$ is an $n \times n$ matrix with every element 0. (remember that $P_f P_c = P_c$). Hence the quadratic forms

$$((P_f - P_c)Y)^T (P_f - P_c)Y \text{ and } ((I - P_f)Y)^T (I - P_f)Y$$

are also independent, in other words

$$Y^T (P_f - P_c)Y \text{ and } Y^T (I - P_f)Y$$

are independent. It only remains to show that

$$Y^T (P_f - P_c)Y/\sigma^2 \sim \text{ non} - \text{central } \chi^2_r,$$

with parameter of non-centrality which vanishes if and only if $\mathbb{E}(Y) \in \omega_c$.
(This proof is very similar to the one given above for the distribution of $Y^T(I - P_f)Y$.)
Observe that $(P_f - P_c)$ is a projection matrix of rank $r = p_2$: put $P = (P_f - P_c)$ for brevity.
Thus we may write

$$P = \sum_1^r v_i v_i^T$$

say, where $v_1, \ldots, v_r$ are the eigen-vectors of P corresponding to the its r eigen-values $1, \ldots, 1$. As usual $v_i^T v_j = 1$ for $i = j$, $v_i^T v_j = 0$ for $i \neq j$.
Hence, the quadratic form

$$Y^T P Y = \sum_1^r (v_i^T Y)^2 = \sum_1^r W_i^2$$

say, where $W_i = v_i^T Y$. Now $W_1, \ldots, W_r$ are NID random variables, each with variance σ^2. Further, let $\mathbb{E}(Y) = \mu$. Then we know from our model that $\mu \in \omega_f$. Furthermore, $\mathbb{E}(W_i) = v_i^T \mu$. You can therefore check that the sum of squares of $\mathbb{E}(W_i)$ is exactly $\mu^T P \mu$. This vanishes if and only if $P\mu = 0$, that is $P_c \mu = P_f \mu$. But we know that $\mu \in \omega_f$, hence it follows that $P_f \mu = \mu$, so that $P\mu = 0$ if and only if $P_c\mu = \mu$, in other words if and only if $\mu \in \omega_c$. Thus we have proved that

$$Y^T(P_f - P_c)Y/\sigma^2 \sim \text{non} - \text{central } \chi^2_r,$$

with parameter of non-centrality which vanishes if and only if $\mathbb{E}(Y) \in \omega_c$.

3.3 Inference about β when σ^2 is unknown.

We have already shown that under ω_f,

$$\tilde{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1}).$$

In order to cope with the problem of unknown σ^2, we need the following
Theorem
Under ω_f, $\tilde{\beta}$ is distributed independently of the residual sum of squares $Y^T(I - P_f)Y/\sigma^2$.
Proof

$$\tilde{\beta} = (X^TX)^{-1}X^TY, \text{ and } (I - P_f)Y = (I - X(X^TX)^{-1}X^T)Y.$$

Now it is straightforward to show that the $p \times n$ matrix $(X^TX)^{-1}X^T(I - P_f)$ has every element 0. Thus $\tilde{\beta}$ is independent of $(I - P_f)Y$, and hence it is also independent of the quadratic form $((I - P_f)Y)^T(I - P_f)Y = Y^T(I - P_f)Y$, which of course is our required result.
This Theorem enables us to use the Student's t-distribution, for example to construct confidence interval for a component of β.
An important special case is that of **simple linear regression** of y on the single covariate x, for which the model may be written as

$$\omega_f : y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \ 1 \le i \le n.$$

You may check that in terms of our previous notation, the design matrix $X = (1_n \vdots x)$, where we have used 1_n as the n-dimensional vector with every element 1. Hence

$$X^TX = \begin{pmatrix} 1_n^T 1_n & 1_n^T x \\ x^T 1_n & x^T x \end{pmatrix}$$

which is a 2×2 matrix having determinant $\Delta = n\sum(x_i - \bar{x})^2$. Thus

$$(X^TX)^{-1} = \begin{pmatrix} S_{xx}/\Delta & -S_x/\Delta \\ -S_x/\Delta & n/\Delta \end{pmatrix}$$

where $S_{xx} = \sum x_i^2, S_x = \sum x_i$. Hence, applying the equation $\tilde{\beta} = (X^TX)^{-1}X^Ty$, we see for example that

$$\tilde{\beta}_2 = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sum(x_i - \bar{x})^2.$$

This is normally distributed, with mean β_2 and variance $\sigma^2/\sum(x_i - \bar{x})^2$, and is independent of the residual sum of squares, which is say $s^2(n-2)$: this of course has the $\sigma^2\chi^2_{df}$ distribution, where $df = n - 2$. Hence we can say that the ratio

$$\frac{(\tilde{\beta}_2 - \beta_2)}{s\sqrt{(\sum(x_i - \bar{x})^2}}$$

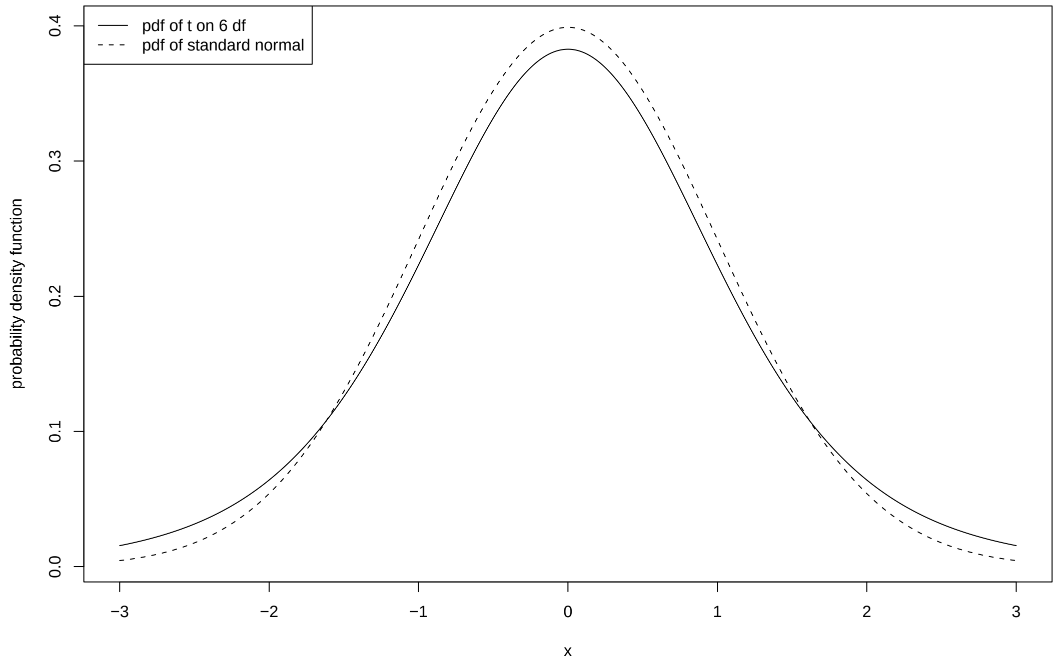

Figure 3.2: The pdf of t_6

has the t-distribution with $n-2$ degrees of freedom. Of course all of this derivation could be carried out directly, without recourse to matrices and the general results. The intention in working it out here from the matrix results is that of helpful illustration. Figure 3.2 shows you the probability density function of a t_6 distribution, together with that of the $N(0,1)$ distribution for comparison. It is intuitively sensible that the former is more 'diffuse' than the latter: if we do not know σ^2 we will tend to get wider confidence intervals for our parameter β_2 than if we do know σ^2.

3.4 Analyses of variance, and the definition of a factor

Example 1. One-way Analysis of Variance (anova)
Another important special case of our general linear model is used in the following simple experimental set-up.
Suppose that we are comparing t different treatments. We take as the model for the data (y_{ij})

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

for $1 \le i \le t, 1 \le j \le n_i$. For example, in an agricultural experiment, y_{ij} might be the **yield** in the jth observation on the ith type of fertiliser. We will assume that the design of the experiment is such that $\epsilon_{ij} \sim NID(0, \sigma^2)$, where y_{ij} = response of j^{th} observation on i^{th} treatment. It is important to realise that our model may be rewritten in the form

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

provided that we 'string out' the observations (y_{ij}) into the n-dimensional vector Y, where $n = \sum n_i$, β is the vector with elements $\mu, \theta_1, \ldots, \theta_t$ and so forth. (We **could** spell out the matrix X if we wished, but there is no particular merit to that exercise.)
We want to test whether all the treaments are the same in their effect on y, or not. So in this case our full hypothesis ω_f is $\mathbb{E}(y_{ij}) = \mu + \theta_i$ for all i, j, and the 'constrained' hypothesis of interest ω_c is $\mathbb{E}(y_{ij}) = \mu$ for all i, j, i.e. ω_c is the hypothesis that there is no difference between treatments.
It is easy to check that the residual sum of squares (i.e. the deviance) fitting ω_c is $\sum\sum(y_{ij} - \bar{y})^2 \equiv Q_c$.
When we come to fitting ω_f, we need to pause for thought about the number of **independent** parameters that we can fit.
First note that 'treatments' is a **factor** here: we wish to fit (θ_i) and not $(\theta \times i)$. This will necessitate a **factor declaration** in any glm package. Omitting such a declaration would have serious and unwanted consequences. This is one of many instances in computational statistics where we learn by making mistakes.
Next, to fit ω_f, we must tackle the problem of lack of **parameter identifiability** in our model. Since, for example,

$$\mathbb{E}(y_{ij}) = \mu + \theta_i = (\mu + 10) + (\theta_i - 10),$$

we see that $\mu, (\theta_i)$ and $(\mu + 10), (\theta_i - 10)$ give identical models for the data. We resolve this difficulty by imposing a linear constraint on the parameters (θ_i). The particular constraint chosen has no statistical interpretation: it is merely a device to enable us to obtain a unique solution to the likelihood equations.
The standard glm constraint is $\theta_1 = 0$. This is known as the 'corner-point' constraint. This actually means that we can write our model ω_f in the form

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

with X of full rank, by taking the components of the t-dimensional vector β as $\mu, \theta_2, \ldots, \theta_t$, for example. Equivalently, we could impose the constraint $\sum n_i\theta_i = 0$. In any case, we still get the same fitted values, which you can check are say,

$$\tilde{y}_{ij} = \sum_j y_{ij}/n_i \equiv \bar{y}_i$$

and the same deviance

$$\sum_{i,j}(y_{ij} - \tilde{y}_{ij})^2 \equiv Q_f.$$

This gives the **Analysis of Variance** Table 3.1. This is traditionally set out as a sum, to show how Q_c is partitioned into its components S_T, Q_f, with the corresponding degrees of freedom partitioned accordingly. You should verify that $Q_f \equiv Q_c - \left[\sum \bar{y}_i^2 n_i - cf\right] = Q_c - S_T$, where cf = correction factor = $\left(\sum\sum y_{ij}\right)^2/n$. Then, to test ω_c, we refer

$$\frac{S_T/(t-1)}{Q_f/(n-t)} \text{ to } F_{t-1,n-t}.$$

Here $S_T = Q_c - Q_f$ = (deviance under ω_c - deviance under ω_f).
N.B. In using glm, you don't need to know the formulae for Q_c, Q_f etc, since glm works

Due to	sum of squares	degrees of freedom
treatments	$S_T = \sum_i \bar{y}_i^2 n_i - cf$	$t-1$
residual ss	Q_f	$n-t$
Total ss	$Q_c = \sum\sum(y_{ij} - \bar{y})^2$	$n-1$

Table 3.1: A simple Analysis of Variance table

them out for you. You just need to know how to use Q_c, Q_f, S_T etc. to construct an Anova, and hence apply our Theorem which gives the distribution of Q_c, Q_f to construct the F test of ω_c against ω_f.
Of course, because Anovas are of such everyday practical importance, many statistical packages, eg SAS, Genstat, S-plus will have a single directive (eg aov() in R or S-plus) which will set up the whole Anova table in one fell swoop. Furthermore, they will generally use a more efficient way of computing the sums of squares than the glm method that we use here, which takes no account of any special properties of the design matrix X. But it's good for you at this stage to have to think about exactly how this table is constructed from differences in residual sums of squares.

Example 2. Two-way Anova.
Consider the dataset given in Table 3.2. This dataset was published in The Independent,

driver	surgeon	barrister	MP	country
86	85	82	86	Denmark
75	83	75	79	Netherlands
77	70	70	68	France
61	70	66	75	UK
67	66	64	67	Belgium
56	65	69	67	Spain
52	67	65	63	Portugal
57	55	59	64	W.Germany
47	58	60	62	Luxembourg
52	56	61	58	Greece
54	56	55	59	Italy
43	51	50	61	Ireland

Table 3.2: The percentage having equal confidence in both sexes, for 4 professions and 12 countries

on October 8, 1992, under the headline "Irish and Italians are the 'sexists of Europe' ", and it shows the percentage of people in each of 12 countries, having equal confidence in the ability of males and females, to carry out any one of 4 professions. (Here 'driver' here means 'bus or train driver'). Clearly there are differences between the 4 professions, and also between the 12 countries. Figure 3.3 shows how the mean percentage depends on the profession, and also on the country. Note that the overall mean percentage is 64.46, and that the means for Luxembourg and Greece coincide.

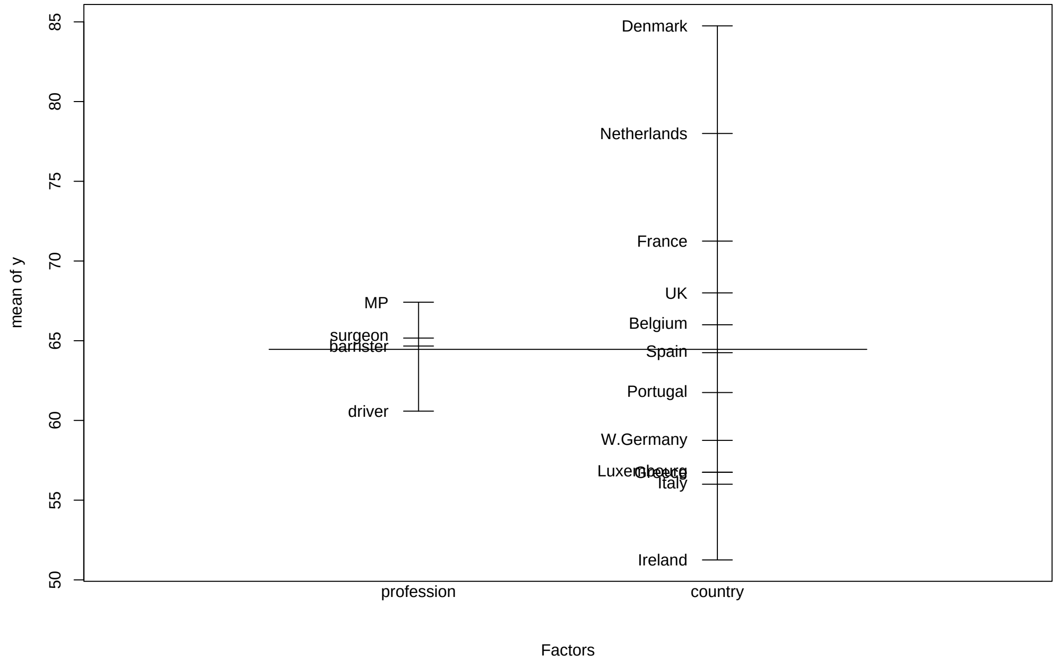

Figure 3.3: A 'factor' plot for the 2-way design

We now work out whether these differences are significant.
Here we have two factors country and profession, having $I = 12, J = 4$ levels respectively, and we have $u = 1$ observations on each of the IJ combinations of factor levels. We take as our model for the responses (y_{ijk})

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \text{ where } \epsilon_{ijk} \sim NID(0, \sigma^2)$$

with $1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq u$. Our 'full' hypothesis is

$$\omega_f : \mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j$$

and we might want to test any or all of the following linear hypotheses:
$\omega_0 : \alpha = 0,\ \beta = 0$, ie no difference between countries, and no difference between professions,
$\omega_1 : \alpha = 0$, ie no difference between countries,
$\omega_2 : \beta = 0.$, ie no difference between professions.
Observe that

$$\omega_0 \subset \omega_1 \subset \omega_f,\ \omega_0 \subset \omega_2 \subset \omega_f.$$

Once again we need to impose constraints on the parameters to ensure identifiability. For the exercises below, it is algebraically convenient to impose the **symmetric** constraints

$$\Sigma\alpha_i = \Sigma\beta_j = 0$$

rather than the default glm constraints

$$\alpha_1 = \beta_1 = 0.$$

Of course, if
$$\mu + \alpha_i + \beta_j = m + a_i + b_j$$
for all i, j, where
$$\Sigma\alpha_i = \Sigma\beta_j = 0, \text{ and } a_1 = b_1 = 0,$$
then you can easily work out the relationships between the two sets of parameters $\mu, (\alpha_i), (\beta_j)$ and $m, (a_i), (b_j)$.

Exercises

(i) Show that the residual ss fitting ω_0 is $\sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2$.

(ii) Show (from the one-way Anova) that the residual ss fitting ω_1 is
$$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{+j+})^2.$$
Note that $\omega_1 : \mathbb{E}(y_{ijk}) = \mu + \beta_j$ (we define $\bar{y}_{+j+} = \sum_i \sum_k y_{ijk}/Iu$).

(iii) Show that $\tilde{y}_{ijk}$, the fitted value under ω_f, may be written as
$$\tilde{y}_{ijk} = \bar{y}_{i++} + \bar{y}_{+j+} - \bar{y}_{+++}, 1 \le i \le I, 1 \le j \le J, 1 \le k \le u.$$
(Hint on solution: we seek to minimise
$$\sum_i \sum_j \sum_k (y_{ijk} - \mu - \alpha_i - \beta_j)^2$$
subject to the constraints
$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0.$$
Thus we take as our Lagrangian
$$\sum_i \sum_j \sum_k (y_{ijk} - \mu - \alpha_i - \beta_j)^2 + \theta \sum_i \alpha_i + \phi \sum_j \beta_j$$
and find the partial derivatives with respect to μ, α_i, β_j in turn, set each of these to 0, and evaluate the Lagrange multipliers θ, ϕ by applying the constraints $\sum_i \alpha_i = 0, \sum_j \beta_j = 0$. This is an example where it is much more efficient to find the formulae for the least squares estimators **directly**, rather than by writing the model in the $y = X\beta + \epsilon$ form and working out what $(X^T X)^{-1}$ must be.)

The residual ss fitting $\omega_f = \sum \sum \sum (y_{ijk} - \tilde{y}_{ijk})^2$.

Show that
$$\begin{aligned} &\text{the residual ss fitting } \omega_2 - \text{ the residual fitting } \omega_f \\ &= \text{the residual ss fitting } \omega_0 - \text{the residual ss fitting } \omega_1. \end{aligned}$$
For the dataset given in Table 3.2, you can fit the linear models $\omega_2, \omega_f, \omega_0, \omega_1$ in turn. You can thus compute the resulting residual sums of squares (deviances). These are, in the corresponding order,

fitting country only, residual sum of squares = 848.00(df = 36)

fitting country + profession, residual sum of squares = 556.250(df = 33)

fitting a constant, residual sum of squares = 5091.917(df = 47) and

fitting profession only, residual sum of squares = 4800.167(df = 44).

Because of the **balance** of the design with respect to the two factors in question, these four deviances obey the linear equation given above, that is

$$848.00 - 556.250 = 5091.917 - 4800.167 (= 291.75, df = 3).$$

This leads us to our next important definition.

3.5 Orthogonality in the Linear Model

Definition of parameter orthogonality for a linear model

Suppose, we have the usual general linear model

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

and the matrix X and the vector β are partitioned as before, so that

$$X\beta = (X_1 X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

where β_1, β_2 are of dimensions p_1, p_2 respectively, and $p_1 + p_2 = p$, where p is the dimension of β. Then β_1, β_2 are said to be mutually orthogonal sets of parameters if and only if

$$\begin{array}{cccc} X_1^T & X_2 & = & 0 \\ \swarrow & \swarrow & & \searrow \\ p_1 \times n & n \times p_2 & & p_1 \times p_2 \end{array}$$

that is, the first p_1 columns of the $n \times p$ matrix X are orthogonal to the last p_2 columns. It is not always easy to check this condition directly. You may well find that an easier way to check that the parameters β_1, β_2 are mutually orthogonal is to apply the Lemma O1.

Lemma O1. β_1, β_2 are orthogonal if and only if

$$\hat{\beta}_1 \equiv \tilde{\beta}_1$$

(nb: this is an an identity in Y), where

$\hat{\beta}_1$ = lse of β_1 in fitting $Y = X_1\beta_1 + \epsilon$ (i.e. assuming $\beta_2 = 0$)

$\tilde{\beta} = \binom{\tilde{\beta}_1}{\tilde{\beta}_2}$ = lse of β in fitting $Y = X\beta + \epsilon$ (i.e. the full model).

Here we use the abbreviation **lse** to denote **Least Squares Estimator**.

Proof. We have already seen that $\tilde{\beta}$ is the solution of

$$X^T X \tilde{\beta} = X^T Y;$$

similarly $\hat{\beta}_1$ is the solution of

$$X_1^T X_1 \hat{\beta}_1 = X_1^T Y.$$

The result follows from writing $X^T X$ as

$$\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} \begin{pmatrix} X_1 & X_2 \end{pmatrix} = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix}.$$

Orthogonality between sets of parameters has an important consequence for residual sums of squares, as shown in Lemmas O2 and O3.

Lemma O2. If β_1, β_2 are orthogonal, then
(residual ss fitting $E(Y) = X_1\beta_1$)$-$ (residual ss fitting $\mathbb{E}(Y) = X_1\beta_1 + X_2\beta_2$)
=(residual ss fitting $\mathbb{E}(Y) = 0$)$-$ (residual ss fitting $\mathbb{E}(Y) = X_2\beta_2$).

Proof.
The residual ss fitting ($\mathbb{E}(Y) = X_1\beta_1$)) is $(Y - X_1\hat{\beta}_1)^T(Y - X_1\hat{\beta}_1)$.
Further,
(residual ss fitting $\mathbb{E}(Y) = X\beta) = (Y - X\tilde{\beta})^T(Y - X\tilde{\beta})$,
and
(residual ss fitting $\mathbb{E}(Y) = 0) = Y^TY$.
Lastly,
(residual ss fitting $\mathbb{E}(Y) = X_2\beta_2) = (Y - X_2\hat{\beta}_2)^T(Y - X_2\hat{\beta}_2)$.
The result follows from writing X^TX as a partitioned matrix, and then using the fact that $X_1^TX_2 = 0$.
Lemma 03. Suppose β_1, β_2 are orthogonal, then we may write
'sum of squares due to β' as
'sum of squares due to β_1' + 'sum of squares due to β_2'
ie

$$Y^TX(X^TX)^{-1}X^TY = Y^TX_1(X_1^TX_1)^{-1}X_1^TY + Y^TX_2(X_2^TX_2)^{-1}X_2^TY$$

equivalently

$$Y^TPY = Y^TP_1Y + Y^TP_2Y,$$

where P_1P_2 is the $n \times n$ matrix with every element 0. Hence Y^TP_1Y, Y^TP_2Y are independent quadratic forms, and each is independent of $Y^T(I - P)Y$. We have already found their distribution.
Proof.
This is a straightforward exercise. Note that $P = P_1 + P_2$.

If β_1, β_2 are not mutually orthogonal, then good software will remind you of this fact by mentioning a phrase such as 'terms added sequentially' in the layout of the Analysis of Variance. The numerical result of an anova will **depend on the order in which the model terms are written in the model**.
Apply Lemma O1 to answer the following questions, in which the errors ϵ_i may be assumed to have the usual distribution.

Exercise 1. In the model

$$Y_i = \beta_1 + \beta_2x_i + \epsilon_i$$

with $1 \leq i \leq n$, show that the parameters β_1, β_2 are mutually orthogonal if and only if $\Sigma x_i = 0$.

Exercise 2. In the model

$$Y_{ij} = \mu + \theta_i + \epsilon_{ij}, 1 \le j \le u, 1 \le i \le t,$$

show that if we impose the constraint $\Sigma\theta_i = 0$, then μ is orthogonal to the set (θ_i).

In practice we are never interested in fitting the hypothesis $E(Y) = 0$, but we are interested in fitting the model

$$E(Y) = \mu 1_n$$

as our 'baseline' model (1_n here denoting the n-dimensional unit vector). For this reason we need the following.

Extension of the definition of orthogonality

Suppose we rewrite our linear model as

$$Y = \mu 1_n + X\beta + \epsilon,$$

where μ, β are unknown, and $\dim(\beta) = p$.
Let us now partition X, β as

$$X = (1_n \vdots X_1 \vdots X_2),\ \ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

respectively, where β_1, β_2 are of dimensions p_1, p_2, and $p_1 + p_2 = p$. Thus we may rewrite our model as say

$$y_i = \mu + \beta_1^T x_{1i} + \beta_2^T x_{2i} + \epsilon_i.$$

Then μ, β_1, β_2 are mutually orthogonal sets of parameters if

$$1_n^T X_1 = 0,\ \ 1_n^T X_2 = 0,\ \ X_1^T X_2 = 0.$$

Consider the linear hypotheses

$$\omega_0 : E(Y) = \mu 1_n,\ \ \omega_1 : E(Y) = \mu 1_n + X_2\beta_2$$

and

$$\omega_2 : E(Y) = \mu 1_n + X_1\beta_1,\ \ \omega_f : E(Y) = \mu 1_n + X\beta.$$

Then, as in Lemma 02, we can show that if μ, β_1, β_2 are mutually orthogonal, then

$$\begin{aligned} &\text{residual ss fitting } \omega_2 - \text{residual ss fitting } \omega_f, \\ = &\text{residual ss fitting } \omega_0 - \text{residual ss fitting } \omega_1. \end{aligned}$$

The proof is left as an exercise:
note that the residual ss fitting ω_0 is $(Y - \mu^* 1_n)^T (Y - \mu^* 1_n)$
where $\mu^* = \sum Y_i/n = \bar{Y}$.

For the dataset given in Table 3.2, with the model

$$\omega_f : \mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j,$$

Due to	degrees of freedom	sum of squares	Mean square	F value	p-value
country	11	4243.9	385.8	22.8885	2.438e-12
profession	3	291.7	97.2	5.7694	0.002752
Residuals	33	556.3	16.9		
Total	47	5091.9			

Table 3.3: A two-way Analysis of Variance table, for a balanced experiment

and identifiability constraints $\sum \alpha_i = 0, \sum \beta_j = 0$, the three sets of parameters $\mu, (\alpha_i), (\beta_j)$ are mutually orthogonal. Because of this fact we can write the corresponding Analysis of Variance as Table 3.3. We say that 'the sum of squares due to (country, profession)' has been partitioned into its orthogonal components 'sum of squares due to country' and 'sum of squares due to profession'. The $F-$ values in the anova show us that there are clearly significant differences between the 12 countries, and also between the 4 professions, in their effect on y. (You will have noticed the ridiculously accurate $p-$ value of $2.438e-12$ which is given in the standard computer output. It is nothing to get excited about: for our purposes we only need know that it is $< .0001$, for example.)
You should now be able to extend the definition to orthogonality between any number of sets of parameters.
Exercise 1. In the model

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \epsilon_i$$

for $i = 1, \ldots, n$ show that the parameters $\beta_1, \beta_2, \beta_3$ are mutually orthogonal if and only if $\Sigma x_i = \Sigma z_i = \Sigma x_i z_i = 0$.
Solution
Just write the design matrix X as $(1_n \vdots x \vdots z)$, then you will see that for mutual orthogonality of the parameters $\beta_1, \beta_2, \beta_3$ we require the 3×3 matrix $X^T X$ to be a diagonal matrix: this gives the required result.
Exercise 2. In the model for the response Y to factors A, B say

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

with $k = 1, \ldots, u, i = 1, \ldots, I, j = 1, \ldots, J$, and constraints $\Sigma\alpha_i = \Sigma\beta_j = 0$, show that $\mu, (\alpha_i), (\beta_j)$ are mutually orthogonal sets of parameters.
Solution Let H_0 be the hypothesis $H_0 : \mathbb{E}(Y_{ijk}) = \mu$ for $k = 1, \ldots, u, i = 1, \ldots, I, j = 1, \ldots, J$. You may check that $\sum_i \sum_j \sum_k ((Y_{ijk} - \mu)^2$ is minimised with respect to μ by $\mu = \bar{Y}$, the mean value of (Y_{ijk}).
Now let H_1 be the hypothesis $H_1 : \mathbb{E}(Y_{ijk}) = \mu + \alpha_i$ for $k = 1, \ldots, u, i = 1, \ldots, I, j = 1, \ldots, J$. Take $\sum_i \alpha_i = 0$. It is easily checked that $\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i)^2$ is minimised with respect to μ and (α_i) subject to $\sum_i \alpha_i = 0$ by

$$\mu = \bar{Y}, \alpha_i = \bar{Y}_{i..} - \bar{Y},$$

where $\bar{Y}_{i..}$ is the mean $\sum_j \sum_k Y_{ijk}/(Ju)$. Hence our least squares estimator of μ under H_0, as a function of (Y_{ijk}), is identical to our least squares estimator of μ under H_1. Thus, by Lemma O1, we see that μ is orthogonal to the set of parameters (α_i).
Let H_2 be the hypothesis $H_1 : \mathbb{E}(Y_{ijk}) = \mu + \beta_j$ for $k = 1, \ldots, u, i = 1, \ldots, I, j = 1, \ldots, J$.

Take $\sum \beta_j = 0$, then it is easily seen by symmetry that the residual sum of squares is minimised with respect to μ, β_j such that $\sum_j \beta_j = 0$ by

$$\mu = \bar{Y}, \beta_j = \bar{Y}_{.j.} - \bar{Y}.$$

where $\bar{Y}_{.j.}$ is the mean $\sum_i \sum_k Y_{ijk}/(Iu)$.
Once again our least squares estimator of μ is the same function of (Y_{ijk}) as for H_0, and so μ is orthogonal to the set of parameters (β_j).
Finally, take H_2 as the hypothesis $H_2 : \mathbb{E}(Y_{ijk}) = \mu + \alpha_i + \beta_j$ for $k = 1, \ldots, u, i = 1, \ldots, I, j = 1, \ldots, J$, with $\sum_i \alpha_i = 0, \sum \beta_j = 0$. You will see that now the residual sum of squares is minimised subject to the constraints by the functions of (Y_{ijk}) which are (respectively) identical to those given above, namely

$$\mu = \bar{Y}, \alpha_i = \bar{Y}_{i..} - \bar{Y}, \beta_j = \bar{Y}_{.j.} - \bar{Y}.$$

Thus $\mu, (\alpha_i)(\beta_j)$ are mutually orthogonal sets of parameters.
Note, this argument was nice and straightforward because we had equal numbers of observations, say u, for each (i, j) combination. If instead we had had the basic model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},\ 1 \le i \le I,\ 1 \le j \le J,\ 1 \le k \le u_{ij}$$

then in general we do not have this nice orthogonality property. Thus for example our estimates of (α_i) if (β_j) is included in the model will in general be different from those of (α_i) if (β_j) is not included in the model.
Exercise 3. The model in Ex. 2 above assumes that the effects of the two factors are **additive**. We may want to check for the presence of an **interaction** between A, B, using the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

with $i = 1, \ldots, I,\ j = 1, \ldots, J, k = 1, \ldots, u$.
Show that with the constraints on $(\alpha_i), (\beta_j)$ as above, and also with the constraints $\Sigma_j \gamma_{ij} = 0$ for each $i, \Sigma_i \gamma_{ij} = 0$ for each j, then the sets of parameters

$$\mu, (\alpha_i), (\beta_j), (\gamma_{ij})$$

are mutually orthogonal.
Solution Simply minimise

$$\sum_i \sum_j \sum_k (Y_{ijk} - \mu - \alpha_i - \beta_j - \gamma_{ij})^2$$

subject to all the constraints on the three sets of parameters $(\alpha_i), (\beta_j), (\gamma_{ij})$ and you will find that the least squares estimators of $\mu, (\alpha_i), (\beta_j)$ are functions of (Y_{ijk}) identical to those given above, and the least squares estimator of γ_{ij} is

$$\gamma_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}\ .$$

We emphasize that both the two orthogonality results given as Exercise 2, 3 above depend on the fact that the experiment is 'balanced', that is we have equal numbers of observations for each (i, j) combination.

3.6 Interaction between two factors: the interpretation

What does it mean to say that there is an **interaction** between two factors?
If an interaction γ is present, then the effect of one factor, say A, on the response Y depends on the level of the second factor, say B. This is best illustrated by an example. Consider a (fictitious) psychological experiment where A is the noise level , say 'quiet' or 'loud', here $0, 1$ respectively, and B is the gender, female or male, of the subject. Let Y be the response, which is a score in answer to the question 'How much does this noise distress you?' and suppose we have the data given in Table 3.4. (Incidentally, this dataset is slightly 'unbalanced', that is the number of observations for each of the four factor combinations are not quite the same.)
Then if for example males perceived a much larger difference between 'quiet' and 'loud' than the corresponding difference perceived by the females, we say that there is an interaction between A and B.
An interaction between two factors is almost always most easily explained by drawing a graph, and for the current example this is shown in Figure 3.4, which shows the mean value of Y_{ijk} against j, for each level of i.
Recall that our model is

	Y	noise	gender
1	22.0	0	female
2	23.7	0	female
3	21.5	0	female
4	23.0	1	female
5	23.0	1	female
6	22.7	1	female
7	15.0	0	male
8	15.2	0	male
9	15.3	0	male
10	14.7	0	male
11	19.0	1	male
12	19.3	1	male
13	20.7	1	male

Table 3.4: A dataset invented to show an interaction between the factors noise and gender

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

with $i = 1, 2$ for quiet, loud, respectively, $j = 1, 2$ for female, male respectively, and $k = 1, \ldots, n_{ij}$. Let us take the corner point constraints for the parameters, which as we have noted will be the default constraints in R. These constraints are

$$\alpha_1 = 0, \beta_1 = 0, \gamma_{1j} = 0 \text{ for all } j, \gamma_{i1} = 0 \text{ for all } i.$$

You may check that with these constraints, for the dataset given above, $\hat{\gamma}_{22} = 4.1167(.7973)$, corresponding to a $t-$value of 5.163 which is clearly significant when referred to t_9.

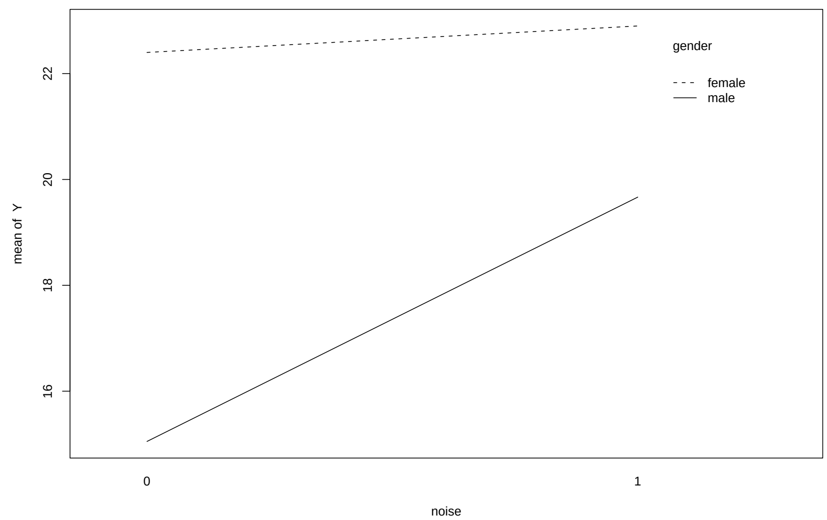

Figure 3.4: A graph showing an interaction between the factors noise and gender

3.7 Collinearity

For convenience we restate our original model

$$Y_i \sim NID(\beta^T x_i, \sigma^2),\ i = 1, \ldots, n$$

or equivalently

$$Y \sim N(X\beta, \sigma^2 I_n).$$

We know that the Least Squares Equations are

$$X^T X \hat{\beta} = X^T Y.$$

The $p \times p$ matrix $X^T X$ is non-singular if and only if X is of full rank. If X is of less than full rank, then there is an infinity of possible solutions to the Least Squares Equations. (Of course, this is just another way of saying that the matrix $X^T X$ does not possess an inverse.) The columns of X are then said to be **collinear**, in other words, they are linearly dependent.
We have already seen that for certain models, for example

$$E(Y_{ij}) = \mu + \theta_i$$

a constraint is needed on the parameters to ensure identifiability, and hence to find a unique solution to the Least Squares Equations. In the case of factor levels, this constraint will be automatically imposed for us by the glm package. Typically, this is $\theta_1 = 0$, etc.
What happens if we, perhaps by accident, try to fit a model $E(Y) = X\beta$ where X is not of full rank, and where the problem is not automatically 'fixed up' for us by the glm imposing its own constraints? For example, what happens if we ask the glm to fit

$$E(Y_i) = \mu + \beta_1 x_i + \beta_2 z_i + \beta_3 w_i,$$

where (for good or bad reasons) we have arranged that $w_i = 6\ x_i + 7\ z_i$, say? Hence, we have certainly arranged that the design matrix X is of less than full rank. A sophisticated glm package will report this to us right away, with some phrase involving 'singular': this enables us, if we so wish, to reduce the set of covariates to get a design matrix of full rank. However, with almost all glm packages, we could just press on and insist on our original choice of covariates. In this case, the glm package would work out for us that not all the parameters **can** be estimated, and would consequently report in the list of parameter estimates that some are **aliased**. In the example above, β_3 would be reported as aliased, since once the first 3 parameters are estimated, β_3 cannot be estimated. Thus the glm package will set β_3 to zero.

Exercise 1. In the model

$$E(Y_{ij}) = \mu + \theta_i + \beta x_i,$$

where $j = 1, \ldots, u$ and $i = 1, \ldots, I$ and (x_i) are given covariates, show that not all the parameters $(\theta_i), \beta$ can be estimated. Experiment with this model with a small set of fictitious data and your favourite glm package.
Exercise 2. Algebraically, we can see that given points $(x_i, Y_i),\ i = 1, \ldots, n$ where (x_i) is scalar, then we should be able to find a polynomial of degree $(n-1)$ which will give a perfect fit:

$$Y_i = \beta_0 + \beta_1 x_i + ... + \beta_{n-1} x_i^{n-1}.$$

In practice this approach is not useful and is not even numerically feasible, as the following experiment will show you. Try generating a random sample of n points ($n = 30$ say) (x_i) from the rectangular distribution on $[0, 1]$, and generate an independent random sample of n points (Y_i). What happens when you fit a straight line, a quadratic, a cubic. . . and so on for the dependence of Y on x? You should find that when you get up to a polynomial of degree more than about 6, the matrix $X^T X$ becomes effectively singular, so that the coefficients of x^7 and so on may be reported as 'aliased'.

3.8 From recent Part II Mathematical Tripos questions

Mathematical Tripos, Part IIA 1997 1/12
i) *This is the 'easy' part of the question*
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) is currently omitted, as it was a GLIM-dependent analysis of a dataset.
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 rank(X) = 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 $R_\Omega,\ R_\omega$ 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 Ω, $R_\Omega/\sigma^2 \sim {\chi^2}_{n-p}$
and on ω, $(R_\omega - R_\Omega)/\sigma^2 \sim {\chi^2}_{p_2}$
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.

Mathematical Tripos, Part IIA, 1998, 4/14 *This is the Paper 4 'Essay' question, designed to take about 40 minutes by the well-prepared candidate.*
Write an essay on fitting the model

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

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

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

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

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

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

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, Part IIA, 1999 4/14
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 $0 \le h_i \le 1$, 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,

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

SOLUTION.
This is another 'Essay' question, and the solution is in effect contained in the Lecture Notes.

Mathematical Tripos, Part IIA, 2000 1/13
(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?

SOLUTION.
(i) is the easy 'bookwork' part, and we have no need to repeat its solution.
(ii) The total number of observations is 44, so we see that the residual sum of squares fitting the null model $H_{null} : \mathbb{E}(log(n_{ij})) = \mu$ is 43.45 with 43 degrees of freedom. This means that the 'ss due to regression' in fitting the given model H_0 is only $(43.45 - 37.89)$

with 2 df (since the difference in dimension between H_0 and H_{null} is 2). So the first thing that we see about the model H_0 is that it is a pretty poor fit, we could find

$$R^2 = (43.45 - 37.89)/43.45.$$

For this point there is no need to do a formal calculation or a formal test (for which, in any case, students in the examination would not have the wherewithal.)
The model H_0 clearly corresponds to 2 parallel lines, each with slope β. The first one, corresponding to First Division clubs, has intercept μ, and the second has intercept $\mu + \theta_2$. But, since we are given the corresponding estimates and their se's, we can do a quick 'by eye' test for the signifiance of the 3 parameters in H_0. For example, formally we could refer $\hat{\beta}/se(\hat{\beta})$ to the t_{41} distribution. However, in practice we simply note that $\hat{\beta}/se(\hat{\beta}) = .8863/.3647 > 2$, so that we will clearly reject the hypothesis that $\beta = 0$. (Here we note that t_{41} will be very like $N(0, 1)$.) Similarly, it appears that μ, θ_2 can probably be taken as zero, so the graph of 2 parallel lines with non-zero intercepts may possibly be adequately replaced by a single line with zero intercept. This remark anticipates the final question 'What linear model would you try next' to which the answer is to try

$$\mathbb{E}(log(n_{ij}) = \beta log(a_{ij}).$$

The response to the question 'How would you check the assumptions about the distribution of (ϵ_{ij}) ?' is intended to be brief remarks, with sketch graphs, about plots such as the residuals against the fitted values, and the qqplot of the residuals as an approximate 'normality' check.

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

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

Solution
i) is all straightforward bookwork, with $H = X(X^TX)^{-1}X^T$ as usual.
ii) Since H is idempotent, and of rank p, it is easily seen that the eigen-values of H are 1, repeated p times, and 0, repeated $n-p$ times. Further, since $trace(H)$ is simply the sum of the eigen values of H, it follows that

$$\sum_i H_{ii} = p,$$

as reqired. For the given special case, we have $X = (1 \; x \; z)$, which is an $n \times 3$ matrix. Thus

$$X^TX = \begin{pmatrix} 1^T1 & 0 & 0 \\ 0 & x^Tx & x^Tz \\ 0 & x^Tz & z^Tz \end{pmatrix}.$$

Find the elements a, b, c by inverting the 2×2 matrix

$$\begin{pmatrix} x^Tx & x^Tz \\ x^Tz & z^Tz \end{pmatrix}.$$

Hence $a = z^Tz/\Delta, b = -x^Tz/\Delta, c = x^Tx/\Delta$, where as usual $\Delta = (x^Tx)(z^Tz) - (x^Tz)^2$. Multiplying out, we derive H in the given form.
Finally, note that since $\hat{Y} = HY$, it follows that $var\hat{Y}_i = H_{ii}\sigma^2$, so for the final expression we need only note that the given expression for H does indeed imply that its ith diagonal element is

$$\sigma^2\left(\frac{1}{n} + a{x_i}^2 + 2bx_iz_i + cz_i^2\right), \text{ for } 1 \le i \le n.$$

Your eyes do not deceive you: the 3rd year Cambridge undergraduates are indeed being asked to invert a 2×2 matrix! (The old skills are the best...)

Mathematical Tripos, Part IIA, 2002 4/14
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
```

Solution

The first part of the question as standard bookwork that we have now seen several times, so here we only give the solution to the 'numbers' part of the question.

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.

Chapter 4

Regression for binomial data

4.1 Basic notation and distributional results

Suppose tha the random variables R_i are independent $Bi(n_i, p_i)$, $1 \leq i \leq k$ and $(r_1, \ldots, r_k)$ are the corresponding observed values. Our general hypothesis is

$$\omega_f : 0 \leq p_i \leq 1,\ 1 \leq i \leq k,$$

and under ω_f,

$$\text{loglikelihood}\,(p) = \sum [r_i \log p_i + (n_i - r_i)\log(1 - p_i)] + \text{constant}$$

which, as you can check, is maximised with respect to $p \in \omega_f$ by $p_i = r_i/n_i$.
Define $logit(p) = log(p/(1-p))$: we will work with this particular link function here. (Later you may wish to try other choices for the link function.) We wish to fit

$$\omega_c : logit(p_i) = \beta^T x_i,\ 1 \leq i \leq k$$

where x_i are given covariates of dimension p, β is of dimension p and $p < k$. Under ω_c, as you can check,

$$\text{loglikelihood } = \ell(\beta) = \beta^T \sum r_i x_i - \sum n_i \log(1 + e^{\beta^T x_i}) + \text{const.}$$

since

$$p_i = e^{\beta^T x_i}/(1 + e^{\beta^T x_i}) = p_i(\beta).$$

Thus $\ell(\beta)$ is maximised by $\hat{\beta}$, the solution to

$$\sum r_i x_i = \sum n_i x_i \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}\ .$$

Put $e_i = n_i p_i(\hat{\beta})$, the 'expected values' under ω_c. Verify that

$$2 \times [\text{loglikelihood maximised under } \omega_f - \text{loglikelihood maximised under } \omega_c]$$

$$= 2 \sum \left(r_i \log \frac{r_i}{e_i} + (n_i - r_i) \log \frac{(n_i - r_i)}{(n_i - e_i)} \right) \equiv D,\ \text{say}.$$

To test ω_c against ω_f, we refer D to χ^2_{k-p}, rejecting ω_c if D is too big, so for a good fit we should find $D \leq k - p$. Assuming that ω_c fits well, we may wish to go on to test, say, $\omega_1 : \beta_2 = 0$, where

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

and β_1, β_2 are of dimensions p_1, p_2 respectively. So under ω_1, $\log\big(p_i/(1-p_i)\big) = \beta_1^T x_{1i}$, say.
Let D_1 be the deviance of ω_1, defined as in $(*)$ $[e_i = e_i(\beta_1^*)]$. By definition $D_1 > D$, and, by Wilks' theorem, to test ω_1 against ω_c we refer $D_1 - D$ to $\chi^2_{p_2}$, rejecting ω_1 in favour of ω_c if this difference is too large. Most versions of glm prints $D_1 - D$ as 'increase in deviance', with the corresponding increase in degrees of freedom (p_2).
Note. At the stage of fitting ω_c we get (from glm), $\hat\beta$ and $se(\hat\beta_j)$ for $j = 1, \ldots, p$. The standard errors come from the square root of the diagonal elements of the matrix

$$\left[-\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right)\right]^{-1} \quad .$$

Since $\hat\beta$ is asymptotically normal,with mean β, we can, for example, test $\beta_p = 0$ by referring $\big(\hat\beta_p/\mathrm{se}(\hat\beta_p)\big)$ to $N(0,1)$.

4.2 An example from criminology, and some exercises

Here is an example of binomial logistic regression with 3 2-level explanatory factors.
Farrington and Morris of the Cambridge University Institute of Criminology collected data from Cambridge City Magistrates' Court on 391 different persons sentenced for Theft Act Offences between January and July 1979.
Leaving aside the 85 persons convicted for *burglary*, there were 120 people for *shoplifting* and 186 convicted for *other theft acts*. (The burglary offences are not considered further here.) The types of sentence were sorted according as to whether they were 'lenient' or 'severe', and those convicted were sorted into men and women, showing that 153 out of 203 men were given a 'lenient' sentence, compared with 89 out of 103 as the corresponding figure for the women. These bald summary statistics suggest that men are being treated more harshly than women, but of course, there's more to this than first meets the eye. A more detailed examination of these 306 individuals allowed the individuals to be classified also by *Previous convictions* (none/one or more), and *Offence type* (shoplifting only/other). For those convicted of shoplifting only, the numbers given lenient sentences were

$$24/25, 17/23, 48/51, 15/21$$

these being given in the order

$$m, m, f, f$$

for gender, and

$$n, p, n, p,$$

for n = Having no previous conviction, and p = Having one or more previous convictions. For those convicted of some other offence, the corresponding figures are

$$52/61, 60/94, 22/24, 4/7.$$

Let y_{ijk} be the number given a lenient sentence, and let tot_{ijk} be the corresponding total, for $i, j, k = 1, 2$. We take $i = 1, 2$ for gender = male, female, $j = 1, 2$ for Previous convictions = none or some, and $k = 1, 2$ for Offence type = shoplifting or other. We assume that y_{ijk} are independent, $Bi(tot_{ijk}, p_{ijk})$. Then, using binomial logistic regression, it can be shown that the model

$$logit(p_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

with the usual constraints $\alpha_1 = \beta_1 = \gamma_1 = 0$ fits well: its deviance is 1.5565 which is well below the expectation of a χ^2_4 variable. The estimates of $\mu, \alpha_2, \beta_2, \gamma_2$ with their corresponding se's are

$$2.627(.4376), .009485(.3954), -1.522(.3361), -.6044(.3662)$$

respectively. Comparing the ratio (.009485/.3954) with $N(0, 1)$ suggests to us that the parameter α_2 can be dropped from the model. In other words, whether or not an individual is given a Lenient sentence is not affected by gender. Removing the term α_2 from the model causes the deviance to increase by only .001 for an increase of 1 df: the resulting model has deviance 1.5571, which may be referred to χ^2_5. The estimates of μ, β_2, γ_2 for this reduced model are 2.634(.3461), −1.524(.3261), −.6082(.3301) respectively, showing that, as we might expect, the odds in favour of getting a Lenient sentence are reduced if there is one or more previous conviction, and reduced if the offence type is other than shoplifting. More specifically, if there is one or more previous conviction, then the odds are reduced by a factor of about $(1/4.6) = \exp -1.524$: if the offence type is other than shoplifting then the odds of getting a Lenient sentence are reduced by a factor of about (1/1.8).

Exercise 1. Explore what happens to the above model if you allow an interaction between previous conviction and offence type, i.e. if you try the model

$$logit(p_{ijk}) = \mu + \beta_j + \gamma_k + \delta_{jk}.$$

Solution

If we put in the interaction term, we find that now the residual deviance is 0.96203 on 4 df, and $\hat{\delta}_{22} = 0.5513(.7249)$. This means (from comparing 0.5513/.7249 with $N(0, 1)$) that the interaction term is non-significant. The advantage of this conclusion is that we do not have to try to interpret an interaction to our client!

Exercise 2. Try the above exercise with the link functions
$g(p) = \Phi^{-1}(p)$, the probit
$g(p) = \log(-\log(1 - p))$, the complementary $\log - \log$.

Solution

Take the model

$$g(p_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k.$$

The deviance for this model with the logit link function is 1.5565, with 4 df. With $g(p) = \Phi^{-1}(p)$ we find that the residual deviance is 1.3228 on 4 df; with $g(p) = \log(-\log(1-p))$ we

find that the residual deviance is 1.0702 on 4 df. The small discrepancies between the three residual deviances are unimportant, and I prefer the logit link for ease of interpretation. For each of the three link functions we reach the same conclusion: that the term α_i can be dropped from the model.
Exercise 3. Warning: in the case of BINARY data, ie when $n_i = 1$ for all i, we cannot use the deviance to assess the fit of the model (the asymptotics go wrong). Show that if $n_i = 1$ for all i, so that r_i only has $0, 1$ as possible values, then the maximum value of the log-likelihood under ω_f is always 0.
Solution
The loglikelihood is say $\sum_i \ell_i(p_i)$, where $\ell_i(p_i) = r_i log(p_i) + (n_i - r_i) log(1 - p_i)$, which we wish to maximise with respect to p_i, for $0 \leq p_i \leq 1$. Now $n_i = 1$, and so if $r_i = 1$, we must maximise $log(p_i)$ in $0 \leq p_i \leq 1$: clearly this is attained at $p_i = 1$, and then $\ell_i(p_i) = 0$, But if $r_i = 0$, we also see that $\ell_i(p_i)$ has maximum value 0. Hence, under ω_f, the maximum value of the log-likelihood is always 0.

4.3 From recent Mathematical Tripos questions

1997 paper 2/11
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 R analysis, with the corresponding (slightly reduced) output. *(in fact, in 1997 I set this as a GLIM example, but here it is recast in R)*. Discuss its interpretation carefully.

```
Yes  <-  c(14,32,11,12)
No <- c(93,81,52, 43)
tot <- Yes + No
Race <- gl(2, 2, length=4, labels=c("White", "Black"))
Drug_use <- gl(2,1, length=4, labels=c("Yes", "No"))
first.glm <- glm(Yes/tot ~ Race + Drug_Use, binomial, weights=tot)
summary(first.glm)
.....................
```

```
Coefficients:
            Estimate Std. Error
(Intercept) -1.73755    0.24038
RaceBlack   -0.05548    0.28861
Drug_useNo   0.71946    0.27898

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom

Number of Fisher Scoring iterations: 4
```

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

$$Yes_{ij} \sim Bi(tot_{ij}, \pi_{ij}),\ 1 \le i, j \le 2$$

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 glm 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, (0.7195/.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.
Four iterations were required to fit this model, and the 'null deviance' was the deviance obtained in fitting the model $\pi_{ij} =$ constant. Since this was 8.3499 on 3 df, the null model was obviously a poor fit.
Note that we could have tried

```
glm(Yes/tot ~ Race* Drug_Use, binomial, weights=tot)
```

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.

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.)
1998 PAPER 1/13
The numerical parts of this question has been edited somewhat. It has been recast from GLIM into R, but is essentially asking the same as in the original version of the question. The R output is given in slightly reduced form.
(i) Suppose $Y_1, ..., Y_n$ are independent observations, with

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

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

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

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

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

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

(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, for which the corresponding output has been slightly reduced. (You need not prove any of the relevant theorems needed for your discussion, but should quote them carefully.)

```
drugdata <-  read.table("data", header=T)
attach(drugdata)
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
  sex2        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
  psych2      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) Firstly we have an easy little bit on 'the calculus at the heart of glm'.
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) Now to the numerical example. The model that we are fitting is $r_i \sim$ independent $Bi(n_i, \pi_i)$, for $1 \leq i \leq 20$, 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).

Chapter 5

Poisson regression and contingency tables

5.1 Loglinear regression for the early UK AIDS data

The total number of reported new cases per month of AIDS in the UK up to November 1985 are listed in Table 5.1 (data taken from A.M. Sykes 1986). Thes are the data for 36

0	0	3	0	1	1	1	2	2	4	2	8	0	3	4	5	2	2
2	5	4	3	15	12	7	14	6	10	14	8	19	10	7	20	10	19

Table 5.1: Early UK AIDS data

consecutive months, and should be read across the Table.
Let us take as our model for Y_i the number of new cases reported in the i^{th} month, the following:
Y_i are independent Poisson with mean μ_i, $1 \le i \le 36$. Thus the 'full' model is

$$\omega_f : \mu_i \ge 0, 1 \le i \le 36.$$

If we plot Y_i against i, we observe that Y_i increases (more or less) as i increases. So let us try to model this by a simple loglinear relationship . Thus the 'constrained' model is

$$\omega_c : \log \mu_i = \alpha + \beta i, \ 1 \le i \le 36,$$

giving

$$\mu_i = \exp(\alpha + \beta i), \text{and } \ell(\alpha, \beta) = \sum \log(e^{-\mu_i} \mu_i^{y_i})$$

hence

$$\ell(\alpha, \beta) = -\sum \exp(\alpha + \beta i) + \sum y_i(\alpha + \beta i).$$

Hence we can find the mle's of α, β as the solution of

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

and we can find the se's of these estimators in the usual way, from the matrix of the second derivatives of ℓ.

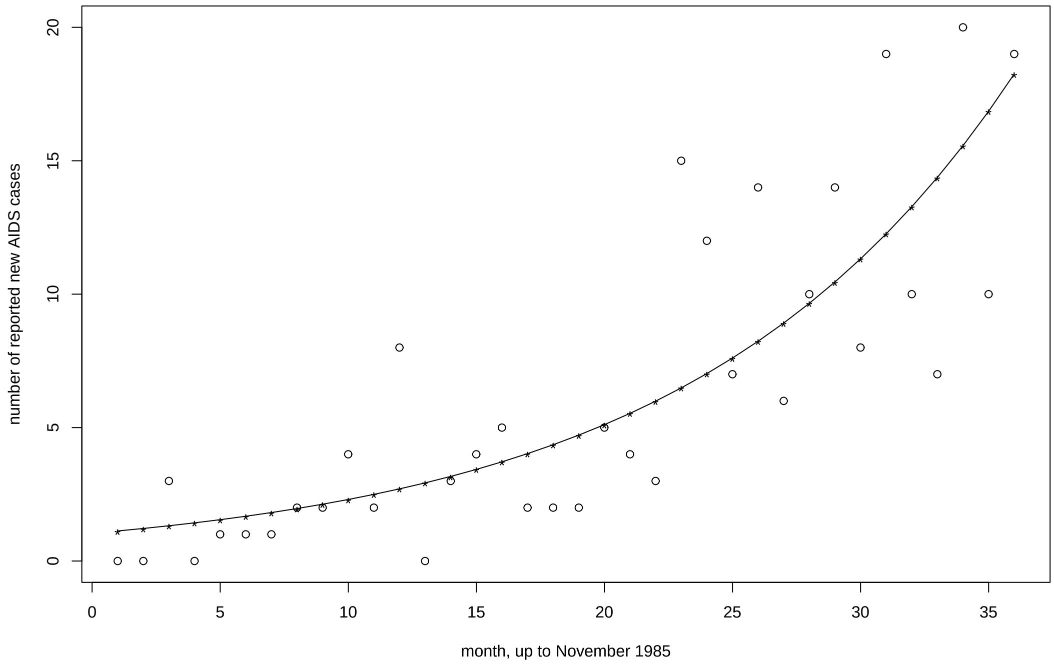

Figure 5.1: Poisson regression for early AIDS data

This fitting is easily achieved in glm using the Poisson "family" with the log-link function, which of course is the canonical link function for this distribution.
You can check that $\hat{\alpha} = 0.03966(0.21200), \hat{\beta} = 0.7957(0.00771)$. Figure 5.1 shows the original data, together with the fitted values under the Poisson model, and the exponential curve $Y = exp(\hat{\alpha} + \hat{\beta} i)$.
To test $\beta = 0$, we refer $\hat{\beta}/se(\hat{\beta}) = (.07957/.007709)$ to $N(0,1)$, or refer 127.8, the increase in deviance when i is droppped from the model to χ^2_1. These two tests are asymptotically equivalent.
Note that the fit of ω_c is not very good: the deviance of 62.36 is large compared with χ^2_{34}. The approximation to the χ^2 distribution cannot be expected to be very good here since many of the e_i, the **fitted values under the null hypothesis** ω_c, are very small. We could improve the approximation by combining some of the cells to give a smaller number of cells overall, but with each of (e_i) greater than or equal to 5.

5.2 Two useful general results

Consider the general model $Y_i \sim Po(\mu_i), 1 \leq i \leq n$ with ω_f as the hypothesis $\mu_i > 0$, and ω_c as the hypothesis $log(\mu_i) = \alpha + \beta^T x_i, 1 \leq i \leq n$, where $x_1, \ldots, x_n$ are given covariate vectors, and α, β are unknown, of dimensions $1, p$ respectively.. It is then easy to show that the deviance for testing ω_c against ω_f is

$$2 \sum y_i \log \frac{y_i}{e_i}, \text{ where}$$

$e_i = e_i(\hat{\alpha}, \hat{\beta}) = \exp(\hat{\alpha} + \hat{\beta} x_i)$.
This deviance is approximately distributed as χ^2_{n-p-1}, if ω_c true, provided that (e_i) is not too small.
By writing $y_i = e_i + \Delta_i$, so that $\sum \Delta_i = 0$, and expanding $\log(1 + (\Delta_i/e_i))$ we can show that the deviance

$$2 \sum y_i log(y_i/e_i)$$

is approximately

$$2 \sum (e_i + \Delta_i)\left(\frac{\Delta_i}{e_i} - \frac{1}{2}\frac{\Delta_i^2}{e_i^2} + \cdots\right).$$

Collecting up the terms, recalling that $\sum \Delta_i = 0$, and neglecting terms of order higher than Δ_i^2 shows us that the the deviance is approximately equal to

$$\sum (y_i - e_i)^2/e_i.$$

This latter expression is called **Pearson's** χ^2.
For the current example the deviance and Pearson's χ^2 are $62.36,\ 62.03$ respectively, and $n - p - 1 = 34$.
Example 2. Accidents 1978–81, for traffic *into* Cambridge
The data are given in Table 5.2.

Let us take as our model for (Y_{ij}), the number of accidents,

	Time of day	Accidents	Estimated traffic volume
Trumpington Road	07.00–09.30	11	2206
Trumpington Road	09.30–15.00	9	3276
Trumpington Road	15.00–18.30	4	1999
Mill Road	07.00–09.30	4	1399
Mill Road	09.30–15.00	20	2276
Mill Road	15.00–18.30	4	1417

Table 5.2: Cambridge traffic data from 1978-81

$$Y_{ij} \sim \text{independent } Po(\mu_{ij})$$

for Road i, and Time of day j, where $i = 1, 2,\ j = 1, 2, 3$.
We might reasonably expect the number of accidents to depend on the traffic *volume*, so we look for a model

$$\mu_{ij} \propto a_i b_j \times v_{ij}^{\gamma}$$

that is

$$\log \mu_{ij} = \text{constant} + \log a_i + \log b_j + \gamma \log v_{ij}.$$

This then enables us to estimate a, b, γ and test $a = 1$ etc. Written more obviously as a glm, this is :

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma \log v_{ij}$$

say, where $i = 1, 2$, $j = 1, 2, 3$, and $\alpha_1 = 0, \beta_1 = 0$ for identifiability.
Hence $\alpha_2 = 0$ if and only if the two roads are equally risky, β_2 represents the difference between time 2 and time 1, and β_3 represents the difference between time 3 and time 1.
The estimate of α_2 compared with its se, ie the ratio (6.123/2.671), shows that Mill Road is more dangerous than Trumpington Road. The model seems to fit well (its deviance is 1.88, which is non-significant when referred to χ^2_1). The 1st and 3rd Times of Day are about as dangerous as each other, and each is quite a lot more dangerous than the 2nd Time of Day. (The estimates of β_2, β_3 are respectively $-6.075(2.972)$, $.04858(.5673)$.)
The accident rate has a strong dependence on the traffic volume, as we would expect: the estimate of γ is 15.42(6.885). We take a further look at how the rate depends on the Road and on the Time of Day, by dropping the corresponding parameters from the model, in turn, and assessing, from the relevant χ^2 distributions, whether or not the resultant increases in deviance are significant. For example, dropping the Road term gives an increase in deviance of 5.709, which is significant compared with χ^2_1, so we put it back into the model. Similarly, dropping Time of Day from the model gives an increase in deviance of 5.701, which is significant compared with χ^2_2, so we put this term back into the model.
But you can check that the model can be simplified by combining the 1st and 3rd Times of Day, so that we have a new 2-level factor (with levels 'rush-hour' and 'non-rush-hour' say). The resulting model fits well: its deviance of 1.8896 is low compared with χ^2_2.
Question: Predict the number of accidents on Mill Road between 0700 and 0930 for traffic flow 2000. [Warning: You get a weird answer. It turns out that the question being asked is a silly one: can you see why?]
Example 3. *The Independent*, October 18, 1995, under the headline "So when should a minister resign?", gave the almost all the following data for the periods when the Prime Ministers were, respectively, Attlee, Churchill, Eden, Macmillan, Douglas-Home, Wilson, Heath, Wilson, Callaghan, Thatcher, Major, Blair. (Happily for me in my ceaseless quest for data, I was able to add the data for the years 1997-2005 following a particular resignation in 2005, but I am still missing the figures for the period 1995–1997.)
In the years
1945–51, 51–55, 55–57, 57–63, 63–64, 64–70, 70–74, 74–76, 76–79, 79–90, 90–95, 97–2005
when the Governments were, respectively,

$$\text{lab, con, con, con, con, lab, con, lab, lab, con, con, lab}$$

(where 'lab' = Labour, and 'con' = Conservative), the total number of ministerial resignations were

$$7, 1, 2, 7, 1, 5, 6, 5, 4, 14, 11, 12.$$

(These resignations occurred for one or more of the following reasons: Sex scandal, Financial scandal, Failure, Political principle, or Public criticism.)
We can fit a Poisson model to Y_i, the number of resignations, taking account of the type of Government (a 2-level factor) and the length in years of that Government. Thus, our model is

$$\log(\mathbb{E}(Y_i)) = \mu + \alpha_j + \gamma\, logyears_i$$

where $j = 1, 2$ for con, lab respectively, and *logyears* is defined as $\log(years)$. We have taken 'years' as

$$6, 4, 2, 6, 1, 6, 4, 2, 3, 11, 5, 8 :$$

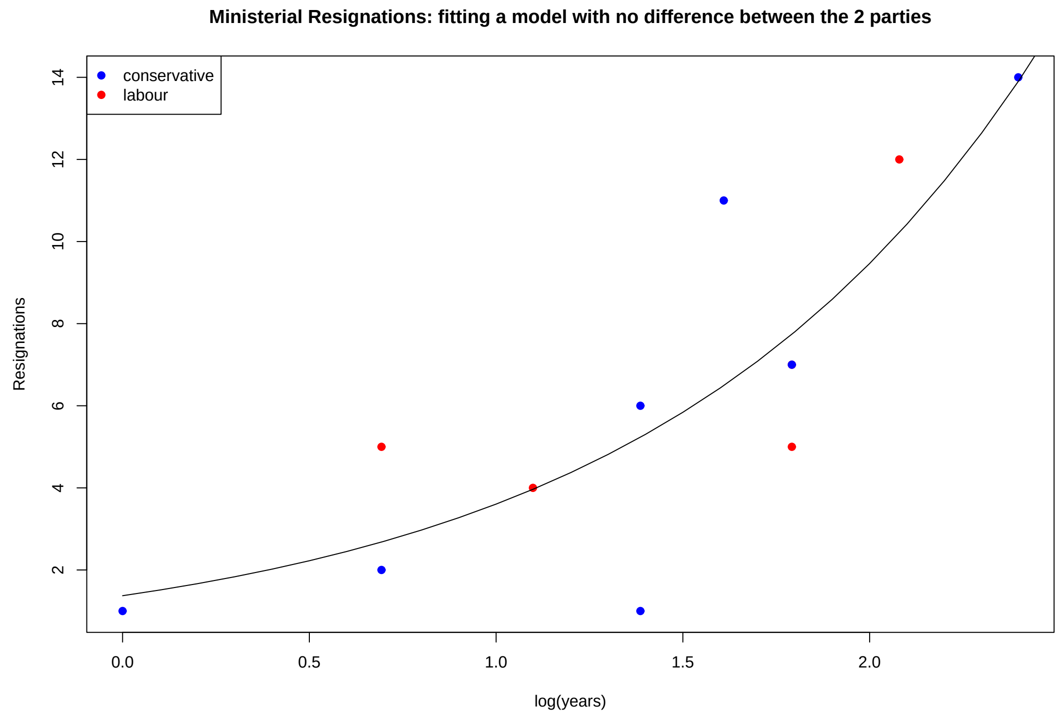

Figure 5.2: Ministerial Resignations, against log('time at risk')

this clearly introduces some error due to rounding, but the exact dates of the respective Governments are not given. This model fits surprisingly well: the deviance is 11.276 (with 9 df). Note that although Labour is very slightly worse than Conservative, the effect of political party is non-significant: $\alpha_1 = 0$ as usual, and $\hat{\alpha}_2 = 0.03541(.23271)$). Each party is as bad/good as the other.
The coefficient $\hat{\gamma}$ is .96636(.22258). For a Poisson process this coefficient would be **exactly** one. We could force the glm to fit the model with γ set to one by declaring *logyears* as an **offset** when fitting the glm. The resulting model then has deviance 11.299 (df = 10).
Figure 5.2 shows Y_i plotted against $logyears_i$, together with the fitted curve from the model which ignores the effect of political party, that is

$$log(\mathbb{E}(Y_i)) = \mu + \gamma \ logyears_i.$$

In this case $\hat{\mu} = .3168(.3993), \hat{\gamma} = .9654(.2219)$ and the residual deviance is 11.299 on $10df$. Conservative resignations are shown on the graph as blue points and Labour ones are shown as red points.

Example 4.

Observe that if S_i is distributed as $Bi(r_i, p_i)$ where r_i is large and p_i is small, then S_i is approximately Poisson, mean μ_i, where

$$log(\mu_i) = log(r_i) + log(p_i).$$

In this case, binomial logistic regression of the observed values (s_i) on explanatory variables (x_i), say, will give extremely similar results, for example in terms of deviances and parameter estimates, to those obtained by the Poisson regression of (s_i) on (x_i), with the usual log-link function, and an **offset** of $(log(r_i))$.

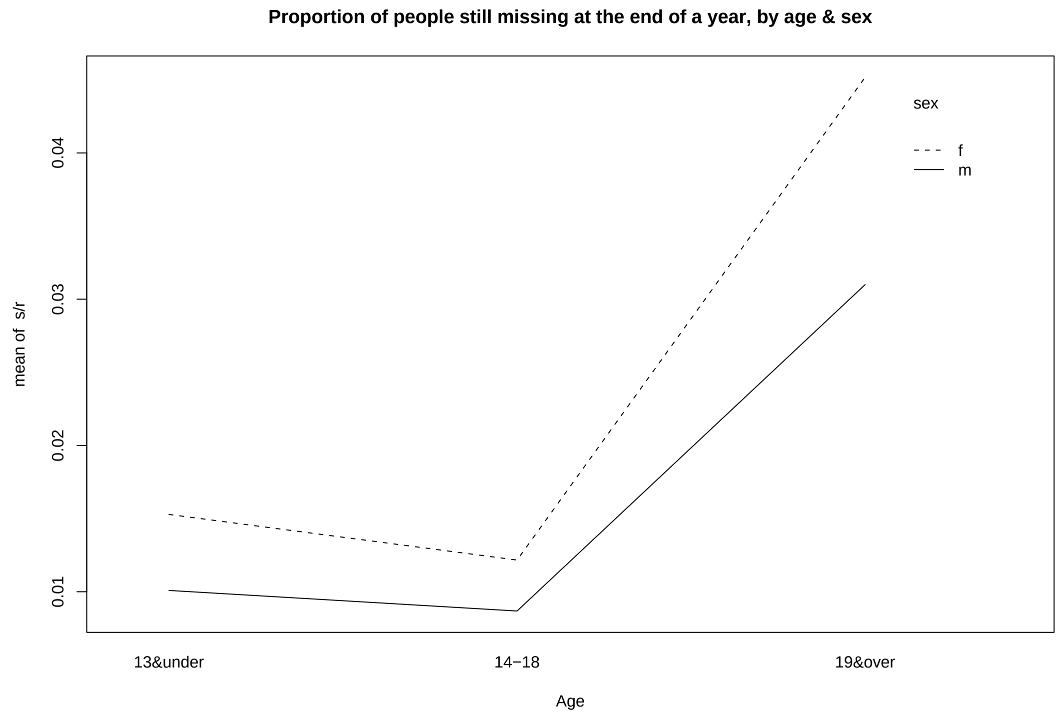

Figure 5.3: Proportion still missing at the end of the year against Age in years

Try both binomial and Poisson regression on the following data-set, which appeared in *The Independent*, March 8, 1994, under the headline 'Thousands of people who disappear without trace'.

$$s/r = 33/3271, 63/7257, 157/5065 \text{ for males}$$

$$s/r = 38/2486, 108/8877, 159/3520 \text{ for females.}$$

Here, using figures from the Metropolitan police,
r = the number reported missing during the year ending March 1993,
and
s = the number still missing at the end of that year.
and the 3 binomial proportions correspond respectively to ages 13 years and under, 14 to 18 years, 19 years and over.
Questions of interest are whether a simple model fits these data, whether the age and/or sex effects are significant, and how to interpret the statistical conclusions to the layman.
Figure 5.3 shows how the proportion s/r still missing at the end of the year depends on age and sex.

5.3 Contingency tables

Example. The *Daily Telegraph* (28/10/88), under the headline 'Executives seen as Drink Drive threat', presented the data in Table 5.3 from the police breath-test operations at Royal Ascot and at Henley Regatta: So at Ascot, 1.1% of those tested are arrested,

	Arrested	Not arrested	Tested
Ascot	24	2210	2234
Henley	5	680	685
Total	29	2890	2919

Table 5.3: A simple 2×2 table

compared with just 0.7% at Henley. Is the percentage at Ascot significantly different from that at Henley? To look at this problem using glm techniques, we first present a reminder of the notation for

<u>The multinomial distribution</u>

Assume $(N_{ij}) \sim Mn\big(n, (p_{ij})\big)$ where n is fixed ($= 2919$ in our example), and where $p_{ij} = P$(an individual is in row i, column j) of the two by two table. Thus with data (n_{ij}),

$$p(n \mid p) = n! \prod \prod \frac{p_{ij}^{n_{ij}}}{n_{ij}!},$$

where $\sum \sum p_{ij} = 1$.
We wish to test

$$H_0 : p_{ij} = p_{i+} p_{+j} \text{ for all } i, j,$$

i.e. for this example, we wish to test whether or not you are arrested is independent of whether you are at Ascot or Henley.
Verify, for the 2×2 table, H_0 is equivalent to

$$p_{11}/p_{1+} = p_{21}/p_{2+}.$$

For our example this is equivalent to the statement
Probability(arrested, given tested at Ascot)= Probability(arrested, given tested at Henley).
Verify that in general H_0 is equivalent to

$$\log p_{ij} = \text{constant} + \alpha_i + \beta_j \text{ for some } \alpha, \beta.$$

where the constant is such that $\sum \sum p_{ij} = 1$.
Now, there is no multinomial 'error' function in glm. The following Lemma shows that for testing independence in a 2-way contingency table we can use the *Poisson* error function as a 'surrogate'.

<u>Using the Poisson error function in glm for the multinomial distribution</u>

The Poisson 'trick' for a 2-way contingency table

Consider the $r \times c$ contingency table $\{y_{ij}\}$. Thus y_{ij} = number of people in row i, column j, $1 \leq i \leq r, 1 \leq j \leq c$. Assume that the sampling is such that $(Y_{ij}) \sim Mn\big(n, (p_{ij})\big)$ multinomial parameters $n, (p_{ij})$. Then

$$p\big((y_{ij}) \mid (p_{ij})\big) = n! \prod \prod \big(p_{ij}^{y_{ij}} / y_{ij}!\big),$$

and, to test

$$H_0 : p_{ij} = \alpha_i \beta_j \text{ for some } \alpha, \beta$$

such that ($\sum\sum \alpha_i\beta_j = 1$), against

$$H : p_{ij} \geq 0, \text{and } \sum\sum p_{ij} = 1,$$

we maximise $L(p) = \sum\sum y_{ij} \log p_{ij}$ on each of H, H_0 respectively. This gives maxima

$$\sum\sum y_{ij} \log(y_{ij}/n), \ \sum\sum y_{ij} \log(e_{ij}/n)$$

where e_{ij} is the expected frequency under H_0, so $e_{ij} = y_{i+}y_{+j}/n$. We know that we can apply Wilks' theorem to *reject* H_0 if and only if $D = 2\sum\sum y_{ij} \log(y_{ij}/e_{ij})$ is too BIG compared with χ^2_f where $f = (r-1)(c-1)$.
But how can we make use of the Poisson error function in glm to compute this deviance function?
Here's the trick: suppose now that $Y_{ij} \sim$ indep $Po(\mu_{ij})$. Consider testing

$$HP_0 : \log \mu_{ij} = \alpha'_i + \beta'_j$$

for some α', β', for all i, j, against

$$HP : \log \mu_{ij} \text{ any real number.}$$

Now

$$\text{the loglikelihood} = L(\mu) = -\sum\sum \mu_{ij} + \sum\sum y_{ij} \log \mu_{ij} + \text{constant.}$$

You will find that $L(\mu)$ is maximised under HP by

$$\hat{\mu}_{ij} = y_{ij} \text{ for all } i, j.$$

You will also find that $L(\mu)$ is maximised under HP_0 by

$$\mu^*_{ij} = y_{i+}y_{+j}/y_{++} = e_{ij} \text{ say.}$$

Applying Wilks' theorem we see that we **reject** HP_0 in favour of HP if and only if DP is too big compared with χ^2_f, where

$$2L(\hat{\mu}) - 2L(\mu^*) = DP$$

and so

$$DP/2 = -\sum\sum \hat{\mu}_{ij} + \sum\sum y_{ij} \log \hat{\mu}_{ij} + \sum\sum \mu^*_{ij} - \sum\sum y_{ij} \log \mu^*_{ij}.$$

But, as you can check, $\sum\sum \hat{\mu}_{ij} = \sum\sum \mu^*_{ij}$ for all (y_{ij}). Hence we have the following *identity*:

$$DP = D \equiv 2\sum\sum y_{ij} \log(y_{ij}/e_{ij}).$$

So we can compute the appropriate deviance for testing independence for the multinomial model by pretending that (y_{ij}) are observations on independent Poisson random variables. This is a special case of the following

General result, relating Poisson and multinomial loglinear models.
We assume that we are given

$$(Y_i) \sim Mn\big(n,(p_i)\big),\ Y_1+\cdots+Y_k=n,\ p_1+\cdots+p_k=1,$$

and given covariates $x_1,\ldots,x_k$. Let (y_i) be the corresponding observed values. We wish to test the null hypothesis

$$H_0:\log p_i=\mu+\beta^T x_i, 1\le i\le k,\ \text{for some } \beta$$

where β is of dimension p, and where μ is such that $\sum p_i=1$, against the more general hypothesis

$$H: p_i\ge 0,\ \text{and}\ \sum p_i=1.$$

Then the deviance for testing H_0 against H may be computed as if (y_i) were observations on independent $Po(\mu_i)$ random variables, and as if we are testing

$$HP_0:\log(\mu_i)=\mu'+\beta^T x_i$$

against

$$HP:\log(\mu_i)=\ \text{any real numbers.}$$

Reminder: In proving this general result we make use of the following
Lemma for exponential families in which $t(y)$ is the vector of sufficient statistics.
Suppose that the pdf of the sample y is

$$f(y\mid\beta)=a(y)b(\beta)\exp\big(\beta^T t(y)\big)$$

where $\int f(y\mid\beta)dy=1$. Then at the mle of β, say $\hat\beta$, the observed and expected values of $t(y)$ agree exactly. This is proved by observing that

$$L(\beta)=\log b(\beta)+\beta^T t(y),\ E(\frac{\partial L}{\partial\beta})=0,$$

and $\hat\beta$ is the solution of $\frac{\partial L}{\partial\beta}=0$.
Proof of the General Result
With (y_i) as observations from the $Mn\big(n,(p_i)\big)$ distribution, we see that the loglikelihood is, say,

$$L(p)=\sum y_i\log p_i+\text{constant}.$$

Under H_0, $p_i\propto\exp(\beta^T x_i)$, so

$$p_i=\big(\exp(\beta^T x_i)\big)/\sum\exp(\beta^T x_j),$$

Thus the loglikelihood is

$$L(p)=\sum y_i\big(\beta^T x_i-\log\sum\exp(\beta^T x_j)\big)+\text{constant}$$

Hence

$$L\big(p(\beta)\big)=\beta^T(\sum y_i x_i)-y_+\log\big(\sum\exp(\beta^T x_j)\big)+\text{constant}.$$

This in turn is maximised with respect to β by β such that

$$* \qquad \frac{\partial L}{\partial \beta} = 0, \text{ie} \quad \sum y_i x_i = y_+ \left(\sum_j x_j \exp(\beta^T x_j) \right) \Big/ \sum_j \exp(\beta^T x_j),$$

$**$ giving $e_i = np_i^*$ as 'fitted values' under H_0, $p_i^* \propto \exp(\hat{\beta}^T x_i)$, $\hat{\beta}$ being the solution of $*$.
It follows from $p_1^* + \cdots + p_k^* = 1$ that $\sum_1^k e_i = n$. Thus $D = 2 \sum y_i \log(y_i/e_i)$.
But if, on the other hand, we assume (y_i) are observations on independent $Po(\mu_i)$, and we test

$$HP_0 : \log \mu_i = \mu' + \beta^T x_i, \ 1 \le i \le k$$

where $(\dim HP_0 = p + 1)$ against

$$HP : \log \mu_i \text{ any real number}$$

where $(\dim HP = k)$. we find that

$$\text{the loglikelihood} = L(\mu) = -\sum \mu_i + \sum y_i \log \mu_i + \text{constant}$$

So, under HP_0,

$$L(\mu) = L(\mu', \beta) = -\sum \exp(\mu' + \beta^T x_i) + \sum y_i(\mu' + \beta^T x_i) + \text{constant}$$

giving

$$\frac{\partial L}{\partial \mu'}(\mu', \beta) = 0 \text{ thus } \sum \exp(\mu' + \beta^T x_i) = \sum y_i$$

and

$$\frac{\partial L}{\partial \beta}(\mu', \beta) = 0 \text{ thus } \sum x_i \exp(\mu' + \beta^T x_i) = \sum y_i x_i.$$

Hence

$$e^{\hat{\mu}'} = \sum_i y_i \Big/ \sum_j \exp(\hat{\beta}^T x_j)$$

and $\hat{\beta}$ is the solution of

$$\sum y_i x_i = y_+ \frac{\sum x_i \exp(\hat{\beta}^T x_i)}{\sum \exp(\hat{\beta}^T x_j)},$$

i.e. $\hat{\beta}$ is as in $*$.
Further, the sufficient statistics are $(\sum y_i, \sum x_i y_i)$ for (μ', β). So at the mle, the observed and expected values of $\sum Y_i$ agree exactly, and we find

$$max_{HP} L(\mu) - max_{HP_0} L(\mu) = -\sum \hat{\mu}_i + \sum y_i \log \hat{\mu}_i + \sum \mu_i^* - \sum y_i \log \mu_i^*$$

so that $\hat{\mu}_i$ = mle ofμ_i under HP, hence $\hat{\mu}_i = y_i$ where μ_i^* = mle of μ_i under HP_0, hence $\sum \mu_i^* = y_+$,
and $\mu_i^* = e_i$ with e_i as in $**$.
Hence $\sum \hat{\mu}_i = \sum \mu_i^*$. Hence
D (multinomial deviance)$= 2 \sum y_i \log(y_i/e_i) \equiv DP$ (Poisson deviance)$= 2 \sum y_i \log(\hat{\mu}_i/e_i)$.

Exercise 1. With (y_i) distributed as Multinomial, with parameters $n, (p_i)$ and with $log(p_i) = \beta^T x_i +$ constant, as above, show that the asymptotic covariance matrix of $\hat{\beta}$ may be written as the inverse of the matrix

$$n[\Sigma p_j x_j {x_j}^T - \Sigma(p_j x_j)\Sigma(p_j {x_j}^T)]$$

and verify directly that this is a positive-definite matrix.
Reminder: A is a positive-definite matrix if and only if $u^T A u \geq 0$ for any vector u, with $u^T A u = 0$ implying that $u = 0$.

Exercise 2. Let x_1, z_1 be vectors of dimension q , and let x_2, z_2 be vectors of dimension p.
Take a_{11} a $q \times q$ matrix, a_{12} a $q \times p$ matrix, a_{21} a $p \times q$ matrix, and a_{22} a $p \times p$ matrix. Solve the simultaneous equations

$$a_{11}x_1 + a_{12}x_2 = z_1$$

$$a_{21}x_1 + a_{22}x_2 = z_2$$

for x_2 in terms of z_1, z_2
This enables you to discover the form of the inverse of the partitioned matrix a, where

$$a = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

Now use this result with $q = 1$ to find the asymptotic covariance matrix of $\hat{\beta}$, given (y_i) observations on independent Poisson variables, mean μ_i, where

$$\log(\mu_i) = \mu' + \beta^T x_i.$$

Compare the result with the answer to Exercise 1.
Solution
This time the log-likelihood is $\ell(\mu', \beta)$ say, where

$$\ell(\mu', \beta) = \mu' \sum y_i + \beta^T \sum x_i y_i - \sum \exp(\mu' + \beta^T x_i).$$

Now form

$$\begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu'^2} & \frac{\partial^2 \ell}{\partial \mu' \partial \beta^T} \\ \frac{\partial^2 \ell}{\partial \mu' \partial \beta} & \frac{\partial \ell^2}{\partial \beta \partial \beta^T} \end{pmatrix}$$

You need to invert this, to show that the asymptotic covariance matrix of $\hat{\beta}$ is of the same form as that given in Exercise 1, provided we write

$$\mu_i = exp(\mu' + \beta^T x_i), \ n = \sum \mu_i, \text{ and } p_i = \mu_i/n.$$

Exercise 3. Let y_i be observations on independent Poisson, mean μ_i, as above, with

$$\log(\mu_i) = \mu' + \beta^T x_i \ .$$

Let $L(\mu', \beta)$ be the corresponding log-likelihood. Derive an expression for the **profile log likelihood** $L(\beta)$, which is defined as the function $L(\mu', \beta)$, maximised with respect to

μ'. Show that this profile log-likelihood function is the identical to a constant + the log-likelihood function for the multinomial distribution, with the usual log-linear model(i.e. $log(p_i) = \beta^T x_i$+ constant).
Profile log-likelihood functions, in general, are an ingenious device for 'eliminating' nuisance parameters, in this case μ'. But they are not the only way of eliminating such parameters: the Bayesian method would be to integrate out the corresponding nuisance parameters using the appropriate probability density function, derived from the joint prior density of the whole set of parameters.

Multi-way contingency tables: for enthusiasts only
Given several discrete-valued random variables, say $A, B, C, \ldots$, there are many different sorts of independence between the variables that are possible. This makes analysis of multi-way contingency tables interesting and complex. Fortunately, the relationship between the variety of types of independence and log-linear models fits naturally within the glm framework. We will once again make use of the relationship between the Poisson and the multinomial in the context of log-linear models. An example with only 3 variables, say A, B and C, serves to illustrate the methods used in tables of dimension higher than 2. Suppose A, B, C correspond respectively to the rows, columns and layers of the 3-way table. Let

$$p_{ijk} = P(A = i, B = j, C = k) \text{ for } i = 1, \ldots, r, j = 1, \ldots, c, k = 1, \ldots, \ell$$

so that $\Sigma p_{ijk} = 1$, and let (n_{ijk}) be the corresponding observed frequencies, assumed to be observations from a multinomial distribution, parameters $n, (p_{ijk})$. For example, we might have data from a random sample of 454 people eligible to vote in the next UK election. Each individual in the sample has told us the answer to questions A, B, C, where
A=voting intention (Labour, Conservative, Other)
B= employment status (employed, unemployed, student, pensioner)
C=place of residence (urban, rural).
Let us suppose that the (fictitious) resulting 3-way table is Table 5.4. There are 8 different

place of residence	urban	urban	urban	rural	rural	rural
voting intention	Lab	Cons	Other	Lab	Cons	Other
employed	50	40	13	31	40	9
unemployed	40	7	5	60	5	5
student	14	9	16	32	7	11
pensioner	10	14	6	3	25	2

Table 5.4: A three-way contingency table

loglinear hypotheses corresponding to types of independence between A, B, C that we now consider. Assume in all of these that the parameters given are such that $\Sigma p_{ijk} = 1$.
We now enumerate the possible loglinear hypotheses.
H_0 : For some α, β, γ, $p_{ijk} = \alpha_i \beta_j \gamma_k$ for all i, j, k,
thus H_0 corresponds to A, B, C independent.
$H_1 : p_{ijk} = \alpha_i \beta_{jk}$ for all i, j, k, for some α, β,

thus H_1 corresponds to A independent of (B, C).
(Likewise, we could consider the hypothesis : B independent of (A, C),
and the hypothesis : C independent of (A, B).)
$H_2 : p_{ijk} = \beta_{ij}\gamma_{ik}$ for all i, j, k, for some β, γ.
You may check that H_2 is equivalent to

$$P(B = j, C = k|A = i) = P(B = j|A = i)P(C = k|A = i) \text{ for all } i, j, k.$$

Thus H_2 corresponds to the hypothesis that, for each i, conditional on A= i, the variables B,C are independent. In this case we say that 'B, C are independent, conditional on A'.
(Likewise, we can define 2 similar hypotheses by interchanging A,B,C):
$H_3 : p_{ijk} = \alpha_{jk}\beta_{ik}\gamma_{ij}$ for all i, j, k, for some α, β, γ.
This final hypothesis, which is symmetric in A, B, C, cannot be given an interpretation in terms of conditional probability. We say that H_3 corresponds to 'no 3-way interaction' between A, B, C. In other words, the interaction between any 2 factors, say A, B for a given level of the 3rd factor, say $C = k$, is the same for all k. Written formally, this is that for each i, j,

$$\frac{(p_{ijk}p_{rck})}{(p_{ick}p_{rjk})}$$

is the same for all k .
The 8 hypotheses are easily seen to be related to one another: you may check that

$$H_0 \subset H_1 \subset H_3, \; H_0 \subset H_2 \subset H_3, \text{ and } H_1 \cap H_2 = H_0.$$

All of the 8 hypotheses above may be written as loglinear hypotheses and hence tested within the glm framework with the Poisson distribution and log link function (the default for the Poisson). For example, we may rewrite H_2 as

$$log(p_{ijk}) = \phi_{ij} + \psi_{ik}$$

for some ϕ, ψ which, in the glm notation for interactions between factors, corresponds to the model

$$A * B + A * C \text{ or equivalently } A * (B + C).$$

All of the 8 hypotheses, except H_3 (the hypothesis of no 3-way interaction), can be represented by a **graph** joining (or not joining) the 3 vertices A, B and C. For example, H_0 corresponds to the graph in which there are no links between the 3 points A, B, C. The null hypothesis H_2, in which B, C are conditionally independent given the level of A, is represented by a graph in which there is no direct link from B to C: there are just the links AB, AC, as in Figure 5.4.
Exercise 1. Show that in the same notation, H_0, H_1, H_3 correspond respectively to

$$A + B + C, \quad A + B * C, \quad (B * C + A * B + A * C)$$

Exercise 2. The data in Table 5.4 above were partly invented to show a 3-way interaction between the factors A, B, C: we might expect that the relationship between voting intention and employment status would not be the same for the Urban voters as for the Non-urban ones. Using the notation above, and your glm package, show that the residual deviance for

$$(A + B + C) * (A + B + C) \text{ is } 15.242(6df)$$

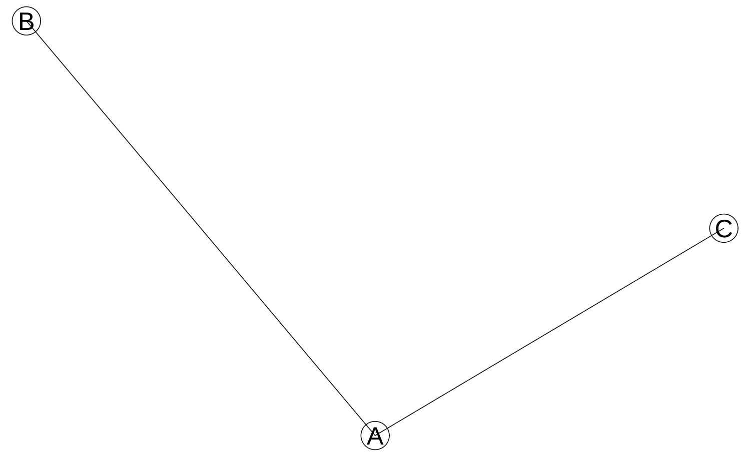

Figure 5.4: The variables B and C are conditionally independent, given the value of the variable A

$$(A + B) * C \text{ is } 122.07(12df)$$
$$(A * B) + C \text{ is } 27.144(11df)$$
$$A + B + C \text{ is } 132.3(17df).$$

Of course, since H_3 failed to fit the data, it was in fact obvious that none of the stronger hypotheses could fit the data.

Exercise 3. Consider the $2 \times 2 \times 2$ Table 5.5. Show that the deviance for fitting the

	C=1	C=1	C=2	C=2
	A=1	A=2	A=1	A=2
B =1	17	23	36	50
B =2	29	14	59	24

Table 5.5: A 3-way table showing no 3-way interaction

model $A * B + B * C + A * C$ is $.12362, 1df$.
By comparing the parameter estimates for this model with their se's, find the simplest model that fits the 3-way table, and interpret it by an independence statement.

The relation between binomial logistic regression and loglinear models in a multi-way contingency table

In a multi-way contingency table, it may not be appropriate to treat the variables, say $A, B, C, \ldots$ symmetrically. For example it may be more natural to treat
A as a **response** variable, and

$B, C, \ldots$ as **explanatory** variables.
In particular, if the number of levels of A is 2, for example corresponding to yes, no, then it may make the analysis easier to interpret if we do a binomial logistic regression of A on the factors $B, C, \ldots$.
Is such an analysis essentially different from a loglinear analysis? We can see from the following considerations that there must be certain exact correspondences between the two approaches. To be specific, take the case where (Y_{ijk}) is multinomial, parameters $n, (p_{ijk})$ and suppose $i = 1, 2$. Write y_{+jk} as $y_{1jk} + y_{2jk}$. Then $Y_{1jk}|y_{+jk}$ are independent Binomial variables, parameters y_{+jk}, θ_{jk} where

$$\theta_{jk} = p_{1jk}/p_{+jk}.$$

So, for example, the model $A*B+B*C+C*A$ for (p_{ijk}) can be shown to be equivalent to the model

$$\text{logit}(\theta_{jk}) = \beta_j + \gamma_k.$$

Exercise. Use the data from the $2 \times 2 \times 2$ table above, with A as the response variable, so that you use the binomial proportions $17/40, 29/43, 36/86, 59/83$ as the responses corresponding to factors (B, C) as $(1,1), (2,1), (1,2), (2,2)$. Show that the deviance and the fitted frequencies for the model $B+C$ are **exactly** the same as those for $A*B+B*C+A*C$ with data (y_{ijk}) and the Poisson model, as above. Check algebraically that this must be so.

Simpson's Paradox (also known as **Yule's Paradox**)
We only have space in these notes for a brief discussion of the fascinating ramifications of multi-way contingency tables. But we will just issue the following WARNING. We have already seen that for 3-way tables, there are several different varieties of independence.
It may be misleading to collapse a multi-way table over (possibly important) categories. For example, suppose that the 2×2 table on (Henley/Ascot) and (Arrested/Not arrested) was in fact derived from the $2 \times 2 \times 2$ table 5.6. Hence although the overall arrest rate at

	Ascot	Ascot	Henley	Henley
	Arrested	Not arrested	Arrested	Not arrested
Men	23	2	3	340
Women	1	2208	2	340

Table 5.6: Illustration of Simpson/Yule paradox

Ascot is not significantly different from that at Henley, there is a clear difference between the Arrest rate for men at Ascot and the Arrest rate for men at Henley.
For example, the deviance for testing independence on the marginal 2-way table (Ascot/Henley) $\times$ (Arrested/Not arrested) is 0.6773, which is non-significant when compared to χ^2_1, suggesting that the arrest rate at Ascot (.011) is not significantly different from that (.007) at Henley.
Now you see that things are quite complex, because of course the way in which any two of the factors depend on each other depends strongly on the level of the third factor; we deliberately invented a data-set with a strong 3-way interaction. You can see from the full

3-way table that the arrest rate is *independent* of gender for Henley although the arrest rate strongly depends on gender for Ascot.
The 2×2 Table 5.7 for Henley gives a deviance of 0.19990 with 1 df, while the 2×2

	Arrested	Not arrested
Men	3	340
Women	2	340

Table 5.7: The Henley sub-table for Simpson/Yule paradox

Table 5.8 for Ascot gives a deviance of 234.0 also with 1 df. Of course, it is scarcely

	Arrested	Not arrested
Men	23	2
Women	1	2208

Table 5.8: The Ascot sub-table for Simpson/Yule paradox

necessary to find the exact numerical values of the deviances to understand about the 3-factor interaction: we include them here merely for completeness.
Finally, you may like to check that the deviance for testing the null hypothesis of no $3-$way interaction in the $2 \times 2 \times 2$ table is 29.5, with 1 df. (This is an example where some of the frequencies are very small, so distributional approximations will not work well.)

5.4 From recent Mathematical Tripos questions

Mathematical Tripos Part IIA, 1998 2/ 11
(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
special2    0.6622  0.2984
```

SOLUTION

(i) Here is the 'familiar' easy bit of the question.
We have $f(y_i|\mu_i) \propto e^{-\mu_i} \ {\mu_i}^{y_i}$

from which we see that the loglikelihood is

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

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

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

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.

Mathematical Tripos Part IIA, 1999, 2/12
(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?

Solution
i) This you have seen before, so we don't repeat it here.
ii) The log-likehood (+ a constant) is easily seen to be

$$\ell(\beta_0,\beta_1) = \beta_0 \sum y_i + \beta_1 \sum x_i y_i - \sum \exp(\beta_0 + \beta_1 x_i),$$

hence

$$\frac{\partial \ell}{\partial \beta_0} = \sum y_i - \sum \mu_i,$$
$$\frac{\partial \ell}{\partial \beta_1} = \sum x_i y_i - \sum x_i \mu_i.$$

Here $\mu_i = \exp(\beta_0 + \beta_1 x_i)$. Thus $(\hat\beta_0, \hat\beta_1)$ is found as the solution to $\frac{\partial \ell}{\partial \beta_0} = 0, \frac{\partial \ell}{\partial \beta_1} = 0$ (and these equations can only be solved by iteration). Further, we know that for large n the asymptotic distribution of $\hat\beta$ is $N(\beta, v(\beta))$, where the 2×2 covariance matrix $v(\beta)$ is the inverse of

$$\begin{pmatrix} -\frac{\partial^2 \ell}{\partial \beta_0^2} & -\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ -\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & -\frac{\partial^2 \ell}{\partial \beta_1^2} \end{pmatrix}.$$

Working out the inverse, and picking out the $(2,2)th$ term, shows that for large n,

$$var(\hat\beta_1) \approxeq \frac{\sum \mu_i}{\Delta}$$

where $\Delta = (\sum \mu_i)(\sum x_i^2 \mu_i) - (\sum x_i \mu_i)^2$, the determinant.
Finally, if $n = 35$, and the deviance fitting ω is 29.3, then we refer 29.3 to χ^2 with 33 degrees of freedom. Since this has expectation $= 33$, we conclude that ω fits well.
This last part is very easy.

Mathematical Tripos Part IIA, 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

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

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

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

Solution

(a)

$$Y_i \sim Po(\mu_i t_i)$$

implies that for the observation y_i, the log-likelihood is say $\ell(\mu_i) = -\mu_i t_i + y_i \log(\mu_i) +$ a constant. Hence, differentiating with respect to μ_i shows that $\hat{\mu}_i = y_i/t_i$.

(b) Now take

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

thus the log-likelihood is say

$$\ell(\beta) = -\sum t_i \exp \beta^T x_i + \sum y_i \beta^T x_i.$$

Differentiate with respect to β to show that $\hat{\beta}$ is the solution of

$$\sum x_i t_i \exp \beta^T x_i = \sum x_i y_i.$$

Further

$$\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = -\sum x_i x_i^T t_i \mu_i,$$

where $\mu_i = \exp \beta^T x_i$ for each i. The iterative solution for $\hat{\beta}$ follows the usual Newton-Raphson scheme, which you should describe, and the large-sample distribution of $\hat{\beta}$ is $N(\beta, v(\beta))$, where

$$(v(\beta))^{-1} = \sum x_i x_i^T t_i \mu_i.$$

(c) The deviance for assessing the fit of ω is computed as $2(\ell(\hat{\mu}) - \ell(\hat{\beta}))$, which has the approximate distribution of a χ^2 with df $n - p$ if ω is true.
Since $\mathbb{E}\chi^2_{n-p} = n - p$, we see that if the computed value of the deviance is 27.3, and

$n-p=33$, then ω is a good fit. (Note that $var(\chi^2_{n-p})=2(n-p)$.)
(d) If we now change the fitted model (and the link function) to

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

we must be aware that only solutions for which $\beta^T x_i > 0,\ 1 \leq i \leq n$ will make sense. This time, when we find the first derivative of $\ell(\beta)$ and set it to 0, we obtain the equations

$$\sum t_i x_i = \sum y_i x_i/(\beta^T x_i).$$

While it is perfectly possible to solve these equations by iteration, answers for which $\beta^T x_i \leq 0$ will not make sense, statistically, and glm would give us an error message.
There is opportunity for a small 'bonus' mark here, since use of the identity link in this context would be unfamiliar to most candidates.

Chapter 6

Appendix 1: The Multivariate Normal Distribution.

We say that the k-dimensional random vector Y is multivariate normal, parameters μ, Σ if the probability density function of Y is

$$f(y|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp{-}(y-\mu)^T\Sigma^{-1}(y-\mu)/2$$

for all real $y_1, \ldots, y_k$. We write this as

$$Y \sim N_k(\mu, \Sigma).$$

Observe that

$$\int f(y|\mu,\Sigma)dy = 1, \text{ for all } \mu, \Sigma.$$

Furthermore, it is easily verified that Y has characteristic function $\psi(t)$ say, where

$$\psi(t) = \mathbb{E}(exp(it^TY)) = \int exp(it^Ty) f(y|\mu,\Sigma)dy$$

so that

$$\psi(t) = exp(i\mu^Tt - t^T\Sigma t/2).$$

By differentiating the characteristic function, it may be shown that

$$\mathbb{E}(Y) = \mu \ , \mathbb{E}(Y-\mu)(Y-\mu)^T = \Sigma$$

and hence

$$\mathbb{E}(Y_i) = \mu_i, \operatorname{cov}(Y_i, Y_j) = \Sigma_{ij}.$$

Σ is a symmetric non-negative definite matrix: thus its eigen-values are all real and greater than or equal to zero.
If A is any $p \times k$ constant matrix, and $Z = AY$, then Z is also multivariate normal, with

$$Z \sim N_p(A\mu, A\Sigma\ A^T).$$

Hence, for example, $Y_1 \sim N_1(\mu_1, \Sigma_{11})$.

Chapter 7

Appendix 2: Regression diagnostics for the Normal Model

<u>**Residuals and leverages**</u> Take $y_i = \beta^T x_i + \epsilon_i$, $1 \le i \le n$, $\epsilon_i \sim NID(0, \sigma^2)$. Equivalently,

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

where, as usual, we assume that Y, ϵ are vectors of dimension n, X is a $n \times p$ matrix of rank p, and β is an unknown vector of dimension p.
We compute the lse $\hat{\beta}$ as $(X^T X)^{-1} X^T Y$ and, using $\epsilon \sim N(0, \sigma^2 I)$, we can say that

$$\hat{\beta} \sim N(\beta, \sigma^2 \big(X^T X)^{-1}\big)$$

independently of the usual residual sum of squares $R(\hat{\beta})$, whose distribution is given by

$$\frac{R(\hat{\beta})}{\sigma^2} = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{\sigma^2} \sim \chi^2_{n-p}.$$

This fundamental distributional result is used, for example, to test $\beta_2 = 0$, by using $\hat{\beta}_2$, se $(\hat{\beta}_2)$. The construction of all of our hypothesis tests and confidence regions will depend on the assumption

$$\epsilon_i \sim NID(0, \sigma^2)$$

so we need some way of checking this: this is what *qq plots* do.
Define $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$, the fitted value $\hat{Y} \equiv HY$ say where H is the 'hat matrix'. Then the residual vector is $\hat{\epsilon} = Y - \hat{Y}$, observed–fitted.
Then $\hat{\epsilon} = X\beta + \epsilon - H(X\beta + \epsilon) = (I - H)\epsilon$ (*check*). Hence

$$\hat{\epsilon} \sim N\big(0, \sigma^2 (I - H)(I - H)^T\big)$$

but $H = H^T, HH = H$, so

$$\hat{\epsilon} \sim N\big(0, \sigma^2 (I - H)\big).$$

Let $h_i = H_{ii}$ for $1 \le i \le n$; then

$$\hat{\epsilon}_i \sim N\big(0, \sigma^2 (1 - h_i)\big).$$

We define

$$\eta_i = \hat{\epsilon}_i \Big/ \sqrt{1 - h_i}$$

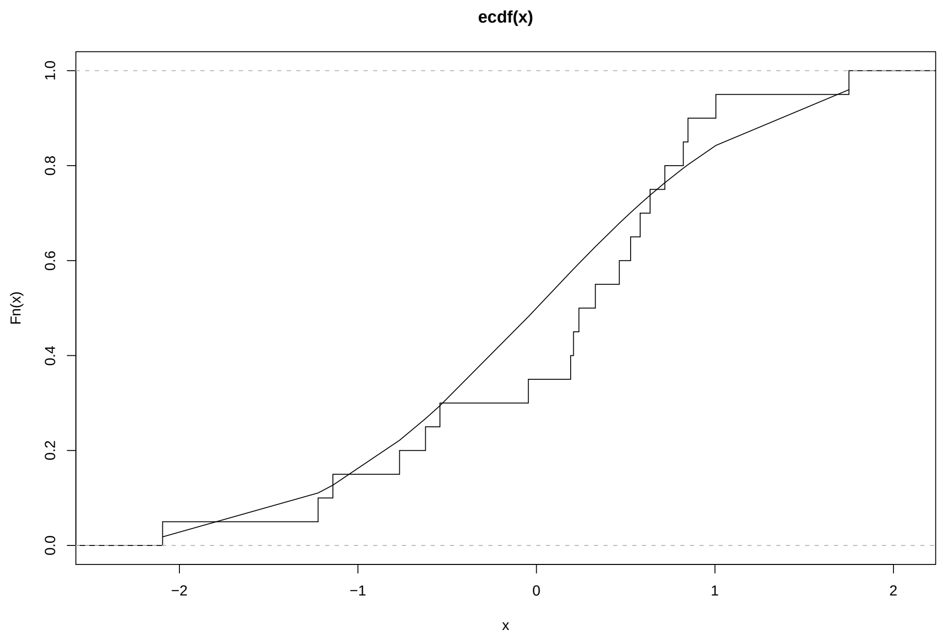

Figure 7.1: The ECDF of a random sample of 20 points from N(0,1), and the normal distribution function

as the *standardised* residuals. We do a visual check of whether $\eta_1, \ldots, \eta_n$ forms a random sample from $N(0, \sigma^2)$ as follows.
We need a new **definition**: the Empirical Cumulative Distribution Function (ECDF) of $(\eta_1, \ldots, \eta_n)$ is defined as $F_n(x)$, where for each x,

$$F_n(x) = \frac{\text{number out of } (\eta_1, \ldots, \eta_n) \leq x}{n}.$$

Hence $F_n(x) \uparrow$ as $x \uparrow$, and for large n, we should find

$$F_n(x) \simeq \Phi(x/\sigma)$$

which is the distribution function of $N(0, \sigma^2)$.
We could sketch $F_n(x)$ against x, and see if it resembles a $\Phi(x/\sigma)$ for some σ. This is hard to do. So instead we sketch $\Phi^{-1}\big(F_n(x)\big)$ to see if it looks like x/σ for some σ, i.e. a straight line through origin. This is what a qq plot does for you. Filliben's coefficient measures the closeness to a straight line. (The Weisberg-Bingham test is also useful.)
Here is a very simple example. A random sample of 20 points from the $N(0, 1)$ distribution has smallest value -2.094, largest value 1.751. The corresponding ECDF plot ($F_n(x)$ against x) is given in Figure 7.1, together with the Cumulative Distribution function of the $N(0, 1)$ distribution. This plot is followed by Figure 7.2 which shows $\Phi^{-1}\big(F_n(x)\big)$ against x, together with the plot of the straight line $y = x$.

Leverages.

Note: $\hat{Y} = HY$, $H = X(X^TX)^{-1}X^T$, giving

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \text{ say, where } h_{ii} = h_i.$$

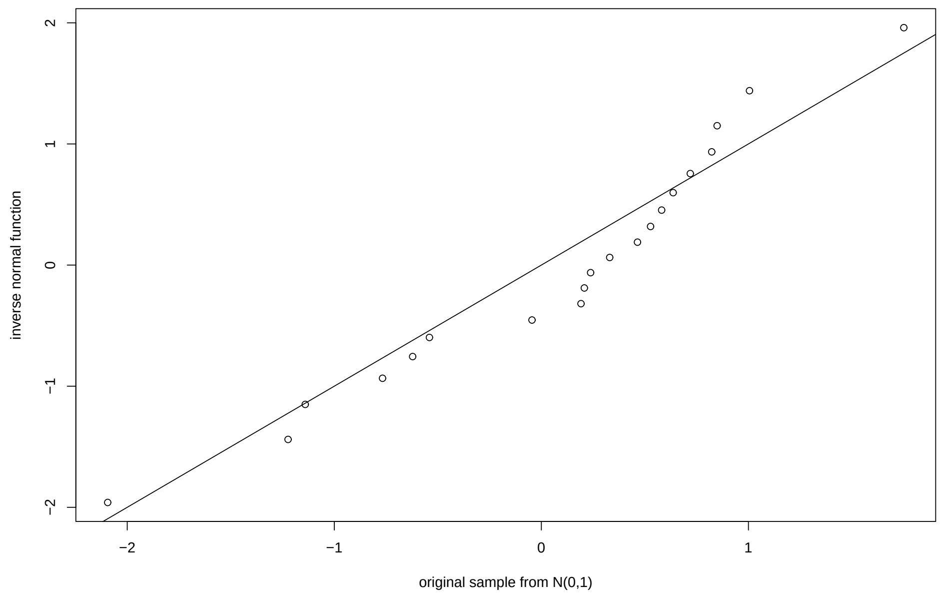

Figure 7.2: The same plots, transformed by the inverse normal distribution function

Since $\hat{\epsilon} = (I - H)\epsilon$, we can see that

$$\text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i), \quad \text{hence } h_i \leq 1.$$

Further, H is a positive-semidefinite matrix, so that $h_i \geq 0$.
The larger h_i is, the closer $\hat{y}_i$ will be to y_i. We say that x_i has high 'leverage' in the regression if h_i large relative to the other $h's$. Note that whatever the $n \times p$ matrix X, we can say that the resulting matrix H has eigen values 1 (exactly p times) and 0 (exactly $n - p$ times). This follows from the fact that H is idempotent of rank p. Hence

$$\sum_1^n h_i = trace(H) = \text{sum of the eigen values of } H = \text{rank}(H) = p.$$

A point x_i for which $h_i > 2p/n$ is said to be a 'high leverage' point. Leverages are also referred to as 'influence values' in some packages.
Exercise 1. Suppose

$$X = (a_1 \vdots \ldots \vdots a_p)$$

where $a_i^T a_j = 1$ for $i \neq j$, and $a_i^T a_j = 1$ for $i = j$.
Then show

$$h_i = a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2, \ 1 \leq i \leq n,$$

(so verify $\sum_1^n h_i = p$).
Exercise 2. Most modern regression software will give you qq plots and leverage plots: note that leverages depend only on the covariate values $(x_1, \ldots, x_n)$. Some regression software will also give **Cook's distances**: these measure the influence of a particular

data point (x_i, y_i) on the estimate of β. Specifically, let $\hat{\beta}_{(i)}$ be the lse of β obtained from the data-set $(x_1, y_1), \ldots, (x_n, y_n)$ with (x_i, y_i) omitted. Thus, using an obvious notation,

$$X_{(i)}^T X_{(i)} \hat{\beta}_{(i)} = X_{(i)}^T y_{(i)}.$$

The Cook's distance of (x_i, y_i) is defined as

$$D_i = \frac{d_i^T (X^T X) d_i}{ps^2}$$

where

$$d_i = \hat{\beta}_{(i)} - \hat{\beta},$$

and s^2 is the usual estimator of σ^2.
These are scaled so that a value of $D_i > 1$ corresponds to a point of high influence.
Note that

$$X_{(i)}^T X_{(i)} = X^T X - x_i x_i^T.$$

and given any non-singular symmetric matrix A and vector b, of the same dimension, we may write

$$(A - bb^T)^{-1} = A^{-1} - A^{-1}b(1 - b^T A^{-1} b)^{-1} b^T A^{-1}.$$

Hence show that if $\hat{y}_{(i)}$ is defined as $x_i^T \hat{\beta}_{(i)}$ then

$$\hat{y}_{(i)} = (\hat{y}_i - h_i y_i)/(1 - h_i)$$

where $h_i = x_i^T (X^T X)^{-1} x_i$, the leverage of x_i as defined previously.
We have briefly described some regression diagnostics for the important special case of the normal linear model. You will find that the more sophisticated glm packages also give regression diagnostics corresponding to those that we have described for any glm model, for example Poisson or binomial. It is a matter of good statistical practice to use these diagnostics, which are usually just 'educated eyeball tests', ie quick graphical checks.

RESUMÉ. The important things you need for this course are

(i) How to find $\frac{\partial}{\partial\beta}, \frac{\partial^2}{\partial\beta\partial\beta^T}$, for example of $L(\beta)$), the log-likelihood function.

(ii) How to find $\mathbb{E}(Y)$ and $\mathrm{cov}(Y)$.

(iii) Basic properties of the normal, Poisson and binomial distributions.

(iv) The asymptotic distribution of $\hat{\theta}$ (the mle), and how to apply of Wilks' theorem ($\sim \chi^2_p$).

(v) Time in front of the computer console, studying the glm directives, trying out different things, interpreting the glm output, and learning from your mistakes, whether they be trivial or serious.

Chapter 8

REFERENCES

Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley.
Aitkin, M., Anderson, D.,Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM.* Oxford: Oxford University Press.
Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman & Hall.
Dobson, A.J. (2001) *An introduction to Generalized Linear Models*. Second Edition. London: Chapman & Hall.
McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. Second Edition. London: Chapman & Hall.
Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*.
Fourth Edition. New York: Springer-Verlag.

Chapter 9

R code for the graphs

You may want to see the R code for drawing the Figures in this document. Here it is. (Some is very quick and easy, others less so.)

```
#Figures 1 and 2
x = seq(-2,2, length=20) ; y = x ; rho =.7
bivnd= function(x,y){
exp(-(x*x - 2*rho*x*y + y*y)/(2*(1-rho*rho)))
}
z = x%*% t(y)
for (i in 1:20){
for (j in 1:20){
 z[i,j] = bivnd(x[i], y[j])
}
}
 # title("A perspective plot of a concave function z")


z = x%*% t(y)
for (i in 1:20){
for (j in 1:20){
z[i,j] = log(bivnd(x[i], y[j]))
}
}

postscript("concave2.ps")
contour(x,y,z)
dev.off()


postscript("concave1.ps")
persp(x,y,z, theta=30, phi =30)
dev.off()
############################################
#Figure 3
```

```
x= (1:1000)/100 - 5

 beta1 = 3 ; beta2 =2
lp = beta1 + beta2*x
p = 1/(1 + exp(-lp))
plot(x,p, type="l")
postscript("logistic.ps")
plot(x,p, type="l")
dev.off()

###########################################
#Figure 4 (my chef d'oeuvre)
 x = c(1,2,-1,-2)
  y = c(-2, 2, 2, -2)
y = y/10
plot(y~x, xlim=c(-2.5, 2.5) , ylim=c(-1,1), xlab="", ylab="", axes= FALSE)
 points(0,0, pch=19)
 polygon(x,y, col="gray")
 x0 = 0 ; y0= 0 ; x1 = .9 ; y1 = .9
 x2 = .9 ; y2 = .1
 arrows(x0,y0, x1,y1, length=0)
 segments(x0,y0, x2,y2)
 x3 = .5 ; y3 = -.19
 points(0,0, pch=19)
 arrows(x1,y1, x2,y2, col="blue")
  segments(x0,y0, x2,y2)
 arrows(x1,y1, x3, y3, col="red")
 segments(x0,y0,x3,y3)
 x4 = -.5 ; y4 =.2
 segments(x0,y0, x4, y4)
points.lab = c("O   ", "Y   ", "Z ", "W   ")
x = c(x0,x1,x2,x3)
y= c(y0,y1,y2,y3)
points(x,y, type="n")
text(x,y, points.lab, cex=1.5)
# Now repeat all the above to put the outcome into a .ps file
 x = c(1,2,-1,-2)
  y = c(-2, 2, 2, -2)
y = y/10
postscript("projectionplot.ps")
plot(y~x, xlim=c(-2.5, 2.5) , ylim=c(-1,1), xlab="", ylab="", axes= FALSE)
 points(0,0, pch=19)
 polygon(x,y, col="gray")
 x0 = 0 ; y0= 0 ; x1 = .9 ; y1 = .9
 x2 = .9 ; y2 = .1
 arrows(x0,y0, x1,y1, length=0)
 segments(x0,y0, x2,y2)
```

```
 x3 = .5 ; y3 = -.19
 points(0,0, pch=19)
 arrows(x1,y1, x2,y2, col="blue")
  segments(x0,y0, x2,y2)
 arrows(x1,y1, x3, y3, col="red")
 segments(x0,y0,x3,y3)
 x4 = -.5 ; y4 =.2
 segments(x0,y0, x4, y4)
points.lab = c("0   ", "Y   ", "Z ", "W   ")
x = c(x0,x1,x2,x3)
y= c(y0,y1,y2,y3)
points(x,y, type="n")
text(x,y, points.lab, cex=1.5)
dev.off()
#######################################################
#Figure 5
x = (-300:300)/100

y = dnorm(x)
z = dt(x, df=6)
matplot(x, cbind(z,y) , type="l", ylab=
"probability density function",lty=c(1,2),col=1)
legend("topleft", legend=c("pdf of t on 6 df",
"pdf of standard normal"), lty=c(1,2),col=1)
postscript("t-dn.ps")
matplot(x, cbind(z,y) , type="l", ylab="probability density function",
lty=c(1,2),col=1)
legend("topleft", legend=c("pdf of t on 6 df",
"pdf of standard normal"), lty=c(1,2),col=1)
dev.off()
#######################################################
#Figure 6


y=
c(86,85,82,86,
75,83,75,79,
77,70,70,68,
61,70,66,75,
67,66,64,67,
56,65,69,67,
52,67,65,63,
57,55,59,64,
47,58,60,62,
52,56,61,58,
54,56,55,59,
43,51,50,61)
```

```
Profession=
c("driver","surgeon","barrister","MP")

Country=
c("Denmark","Netherlands","France","UK","Belgium", "Spain",
"Portugal","W.Germany","Luxembourg","Greece","Italy","Ireland")

country=gl(12,4,length=48,,labels=Country)
profession=gl(4,1,length=48,labels=Profession)
plot.design(y~profession+country)
postscript("ws4fplot.ps")
plot.design(y~profession+country)
dev.off()

#######################################################
#Figure 7
int.data = read.table("interaction.data", header=T)
attach(int.data)
int.data
noise = factor(noise)
# first.lm = lm(Y ~ noise*gender)
# summary(first.lm)
# anova(first.lm)
interaction.plot(noise, gender, Y)
postscript("interaction.ps")
interaction.plot(noise, gender, Y)
dev.off()

#######################################################
#Figure 8

y = scan("aidsdataforFigure8")
i = 1:36
aids.reg = glm(y~i, poisson)
plot(i,y, xlab="month, up to November 1985", ylab=
"number of reported new AIDS cases")
# aids.reg = glm(y~i, poisson)
fv = aids.reg$fitted.values
points(i,fv, pch="*")
lines(i,fv)
summary(aids.reg)
y
postscript("AIDS.ps")
plot(i,y,xlab="month, up to November 1985",
 ylab="number of reported new AIDS cases")
points(i,fv, pch="*")
```

```
lines(i,fv)
dev.off()

#######################################################
#Figure 9

Resignations = read.table("ResignationsFigure9data", header=T)
attach(Resignations)
plot(Res ~ log(years), pch=19, col=c(4,2) [Gov], ylab= "Resignations")
title("Ministerial Resignations:
 fitting a model with no difference between the 2 parties")
legend("topleft", legend=c("conservative","labour"), col=c(4,2), pch=19)
# next.glm= glm(Res ~ Gov + offset(log(years)), poisson); summary(next.glm)
last.glm = glm(Res ~log(years),poisson); summary(last.glm)
l <- (0:25)/10
fv <- exp(0.3168 + 0.9654*l)
lines(l,fv)
postscript("MinResignations.ps")
plot(Res ~ log(years), pch=19, col=c(4,2) [Gov], ylab= "Resignations")
title("Ministerial Resignations: fitting a model with no difference
 between the 2 parties")
legend("topleft", legend=c("conservative","labour"), col=c(4,2), pch=19)
lines(l,fv)
dev.off()

#######################################################
#Figure 10

# ‘Thousands of people who disappear without trace’
s =c(33,63,157,38,108,159)
r=c(3271,7256,5065,2486,8877,3520)
sex = gl(2,3,length=6, labels=c("male","female"))

age=gl(3,1,length=6, labels=c("13&under","14-18","19&over"))
# bin.add = glm(s/r ~ sex + age, binomial, weights=r); summary(bin.add)
interaction.plot(age, sex,s/r, type="l")
title("Proportion of people still missing at the end of a year, by age & sex")
postscript("ws8.ps")
interaction.plot(age, sex,s/r, type="l")
title("Proportion of people still missing at the end of a year, by age & sex")
dev.off()


#######################################################
#Figure 11

# conditional independence
```

```
library(graphics)
x = c(2,0,4) ; y = c(0,4,2); var.names = c("A","B","C")
postscript("conditionalind.ps")
plot(x,y, pch=1, cex=3,axes="F", xlab="", ylab="")
text(x,y,var.names,cex= 1)
arrows(x[2],y[2],x[1],y[1], length=0)
arrows(x[3],y[3],x[1],y[1], length=0)
dev.off()

########################################################
#Figures 12 and 13


  par(mfrow=c(1,2)) # for onscreen graphics
set.seed(1.3) # to ensure the same random sample each time
x = rnorm(20)
F20 = ecdf(x)
X = sort(x)
X
plot(F20, verticals=TRUE, do.p= FALSE)
 lines(X, pnorm(X))
 # title(" ecdf(x)")

y = qqnorm(X, plot.it=FALSE)
plot(y$y, y$x,xlab="original sample from N(0,1)", ylab="inverse normal function")
abline(0,1)
postscript("ecdf.ps")
plot(F20, verticals=TRUE, do.p= FALSE)
lines(X, pnorm(X))
title(" ecdf(x)")
dev.off()

postscript("transform.ps")
plot(y$y, y$x,xlab="original sample from N(0,1)", ylab="inverse normal function")
abline(0,1)
dev.off()
##########################################################
And finally, here are the 3 datasets, which you will need to
 arrange in 3 separate files.

interaction.data
 Y noise gender
22  0     female
23.7  0   female
21.5 0    female
23   1    female
23   1    female
```

```
22.7 1    female
15   0    male
15.2  0   male
15.3  0   male
14.7  0   male
19    1   male
19.3  1   male
20.7  1   male


aidsdataforFigure8
0 0 3 0 1 1 1 2 2 4 2 8 0 3 4 5 2 2 2 5
4 3 15 12 7 14 6 10 14 8 19 10 7 20 10 19


ResignationsFigure9data
epoch Gov Res years
45-51 lab  7  6
51-55 con  1  4
55-57 con  2  2
57-63 con  7  6
63-64 con  1  1
64-70 lab  5  6
70-74 con  6  4
74-76 lab  5  2
76-79 lab  4  3
79-90 con 14 11
90-95 con 11  5
97-05 lab 12  8
```