Computational Statistics and Statistical Modelling

These notes were used as the basis for the 2005 Part II course 'Statistical Modelling', which was a 24-hour course, of which approximately one third consisted of practical sessions. The course is essentially an Introduction to Generalized Linear Models.

(January 2008) An expanded version of these notes, including more examples, graphs, tables etc may be seen at

http://www.statslab.cam.ac.uk/ pat/All.ps
This version evolves, slowly.

TABLE OF CONTENTS

Appendix 2. Regression diagnostics for the Normal model: residuals and leverages.

ACKNOWLEDGEMENTS

Many students, both final-year undergraduates and first-year graduates, have made helpful corrections and suggestions to these notes. I should particularly like to thank Dr B.D.M.Tom for his careful proof-reading.

There are 4 Examples Sheets, with outline solutions: the first contains some quite easy revision exercises. In addition you should work carefully through the R-worksheets provided in the booklets, preferably writing in your own words a 1- or 2-page summary of the analysis achieved by the worksheet.

Chapter 1: Introduction

There are already several excellent books on this topic. For example McCullagh and Nelder(1989) have written a classic research monograph, and Aitkin et al. (1989) have an invaluable introduction to GLIM. Dobson (1990) has written a very full and clear introduction, which is not linked to any one particular software package. Agresti (1996) in a very clearly written text with many interesting data-sets, introduces Generalized Linear Modelling with particular reference to categorical data analysis.

These notes 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.

The computer practicals serve to introduce the students to (S-plus or) R, and both the practical sessions and the 4 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 on R and its properties. The author's 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 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, Glim4 has a generalized linear modelling (glm) component.

Chapter 2: The asymptotic likelihood theory needed for glm

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.)

The two key results of this chapter

Preamble. Suppose $\hat{\theta}_n$ maximises $L_n(\theta)$, i.e. $\hat{\theta}_n$ 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,

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

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

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

This is the Cramèr Rao lower bound (CRLB) for the variance of an unbiased estimator of θ (and there is a corresponding matrix inequality if t, θ are vectors).

Result 1. For θ real,

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

The vector version of this result, which is of great practical use, is:

For θ a k-dimensional parameter,

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

that is, $\hat{\theta}_n$, 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 is the true parameter value of course, and covariance matrix $\Sigma_n(\theta)$, where $\Sigma_n(\theta)$ is given by

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

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

Notes:

(0) Hence any component of $\hat{\theta}_n$, e.g. $(\hat{\theta}_n)_1$ is asymptotic 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. However, we will 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

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

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

$$H_0 : \theta \in \omega$$

against

$$H_1 : \theta \in \Omega$$

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

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

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

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

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 X). Usually we are able to find $\hat{\theta}_n$ as follows: $\hat{\theta}_n$ is the solution of

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

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

Warning

(a) As usual in maximising any function, we have to take care to check that these equations do indeed correspond to the maximum (not just a local maximum, not a saddlepoint, and so on). So, check that

minus the matrix of 2nd derivatives is positive-definite

to ensure that the log-likelihood surface is CONCAVE.

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

Basic properties of the 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 $\hat{\theta}(x)$ itself is not necessarily sufficient for θ).

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

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

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

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 $t(x)$.

(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. Consider the r.v.s

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

(∗) Now $$\left\{\text{cov}\left[t(X), \frac{\partial}{\partial\theta}\, L_n(\theta)\right]\right\}^2 \leq \text{var}\big(t(X)\big)\ \text{var}\left(\frac{\partial L_n}{\partial\theta}\right)$$

with $=$ if and only if $\dfrac{\partial L_n}{\partial\theta}$ is a linear function of $t(X)$.

But $$\frac{\partial L_n}{\partial\theta} = \frac{\partial}{\partial\theta}\,\log f(x \mid \theta) \qquad x \text{ being the whole sample.}$$

(∗∗) Thus
$$\begin{aligned}
\mathbb{E}\left(\tfrac{\partial L_n}{\partial\theta}\right) &= \textstyle\int_x \tfrac{\partial L_n}{\partial\theta}\, f(x \mid \theta)dx \\
&= \textstyle\int_x \tfrac{1}{f(x|\theta)}\, \left[\tfrac{\partial}{\partial\theta}\, f(x \mid \theta)\right]\, f(x \mid \theta)dx \\
&= \tfrac{\partial}{\partial\theta}\, \textstyle\int_x f(x \mid \theta)dx \qquad \text{(under suitable regularity conditions)} \\
&= \tfrac{\partial}{\partial\theta}\, 1 = 0 \qquad \text{(remembering that } f(x \mid \theta) \text{ is a pdf).}
\end{aligned}$$

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

t being an unbiased estimator.

Thus (∗) can be rewritten as

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

with $=$ if and only if $\dfrac{\partial L_n}{\partial\theta} = a(\theta)(t(X) - \theta) + b(\theta)$ 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 $\text{cov}\left(t(X), \frac{\partial L_n}{\partial \theta}\right) = 1$).

Thus if $t(x)$ unbiased, and its variance attains the CRLB, then

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

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 $\mathbb{E}(\hat{\theta}_n) = \theta$, $\text{var}\,(\hat{\theta}) = \theta(1-\theta)/n$ and use the CLT (*Central Limit Theorem*) to show that, for large n,

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

<u>Outline Proof of Result 1</u> 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 (you may disagree!) we will refer to θ_0 as 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

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

assuming the remainder is negligible, and 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 proved on page 6, $\mathbb{E}_{\theta_0}(U_j) = 0$. Furthermore,

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

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 $\frac{1}{\sqrt{n}} \sum U_j \to$ 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,

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

Comments

(i) The basic steps used in the above are Taylor series expansion about θ_0, CLT, and 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 $(i,j)^{\text{th}}$ el.

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

(iii) The result also generalises to the case where $X_1, \ldots, X_n$ are independent but *not identically* distributed, e.g.

X_i independent $Po(\mu_i)$ (Poisson with mean μ_i)
with $\log\mu_i = \beta^T z_i$, z_i given covariate,
β the unknown parameter of interest.

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.

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,

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

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,

$$2\log R_n = 2 \quad \big[L_n(\hat\theta_n) - L_n(\theta_0)\big],$$

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

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

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

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

giving $$2\big(L_n(\theta_0) - L_n(\hat\theta_n)\big) \simeq -(\theta_0 - \hat\theta_n)^T \big(ni(\theta_0)\big)(\theta_0 - \hat\theta_n)$$

i.e. $$2\log R_n = 2\big(L_n(\hat\theta_n) - L_n(\theta_0)\big) \simeq (\hat\theta_n - \theta_0)^T \big(ni(\theta_0)\big)(\hat\theta_n - \theta_0).$$

But, if $\theta = \theta_0$,

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

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

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

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

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

Hence $Z^T\Sigma^{-1}Z \sim \chi^2_r$ as required (and hence has mean r, variance $2r$: prove this).

Exponential Family Distributions

If

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

is the pdf of Y 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 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, e.g. normal, Poisson, binomial, are of exponential family form. Here is one nice property.

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 $\quad \dfrac{\partial L_n}{\partial \pi} = \dfrac{n a'(\pi)}{a(\pi)} + t(y).$

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

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

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

Further, from $(**)$,

$$\begin{aligned} \frac{\partial^2 L}{\partial \pi^2} &= n \frac{\partial}{\partial \pi} \left(\frac{a'(\pi)}{a(\pi)} \right) \\ &= n \left[\frac{a''}{a} - \left(\frac{a'}{a} \right)^2 \right]. \end{aligned}$$

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 $\qquad \dfrac{-a''}{a} + \dfrac{2(a')^2}{a^2} = \mathbb{E}\big(\tau(Y)\big)^2$

giving $\qquad \dfrac{-a''}{a} + \left(\dfrac{a'}{a} \right)^2 = \operatorname{var}\big(\tau(Y)\big).$

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

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 $\qquad t(y) = \mathbb{E}\big(t(Y)\big) \Big|_{\pi = \hat{\pi}}$

i.e. 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 estimation for exponential families relatively straightforward.

Chapter 3: Introduction to glm: Generalised Linear Models

Our methods are suitable for the following types of statistical problem (all having 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 scalar, and so β of dimension 2, (and we might want to estimate β_2, β_1, to test $\beta_2 = 0$, and so on).

(ii) *Poisson regression*

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

(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 unknown, and x_i is a known covariate.

(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

r_i = number of patients given a dose x_i of a new drug
Y_i = number of these giving *positive* response to this drug (e.g. cured).

We observe that Y_i/r_i tends to increase with x_i and we want to model this relationship,

e.g. to find the x which will give $\mathbb{E}(Y/r) = .90$ (i.e. the dose which gives a 90% cure rate)

e.g. 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.

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

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

often works well. Thus we take

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

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 $\beta_2 > 0$).

(iv) *Contingency tables* (a less obvious application of glm).

e.g. $(N_{ij}) \sim Mn\big(n; (p_{ij})\big)$, $\sum\sum p_{ij} = 1$
multinomial, parameters $n, (p_{ij})$.

e.g. N_{ij} = number of people of ethnic group i voting for political party j
in a sample of size n, $\quad 1 \le i \le I, 1 \le j \le J$.

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 $\sum \alpha_i = \sum \beta_j = 1$,
i.e. to test the hypothesis that ethnic group and party are independent.

Note that $\mathbb{E}(N_{ij}) = np_{ij}$,

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

and under H_0 $\qquad \log p_{ij} = \log \alpha_i + \log \beta_j$

equivalently $\qquad \log \mathbb{E}(N_{ij}) = \text{const} + a_i + b_j \quad$ 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 mle's 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 (by Wilks' theorem, usually).

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 $\mathbb{E}(Y_i) = \mu_i$ (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.

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

Proof.

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

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

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

But for all θ, ϕ $$\int_y f(y \mid \theta, \phi) dy = 1$$

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

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

i.e. $$\mathbb{E}\left(\frac{Y - b'(\theta)}{\phi}\right)^2 = \frac{b''(\theta)}{\phi} \quad \text{from } *,$$

i.e. $\text{var}(Y) = \phi b''(\theta)$.

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

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

$$\uparrow$$

(actually $\ell(\beta, \phi)$)

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

where we have used the chain rule, viz.

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right) &= -\mathbb{E}\left(\sum_1^n \frac{(y_i - \mu_i)^2}{\big(g'(\mu_i) V_i\big)^2}\, x_i x_i^T\right) \\
&= -\sum_1^n \frac{V_i}{\big(g'(\mu_i)\big)^2 V_i^2}\, x_i x_i^T \\
&= -\sum_1^n w_i x_i x_i^T \text{ say}, \quad w_i \equiv 1/\left(V_i \big(g'(\mu_i)\big)^2\right).
\end{aligned}$$

We write W as the diagonal matrix

$$\begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$$

and thus we see

$$\mathbb{E}\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right) = -X^T \underset{n\times n}{W} \underset{n\times p}{X} \qquad \text{.....Ex.}$$

where

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

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

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

Reminder: The Newton-Raphson algorithm

To solve

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

Take β_0 as 'starting value'. 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

$$0 = \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_0} + \left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)\Bigg|_{\beta_0} (\beta_1 - \beta_0) \quad \text{.....}\mathbf{It}.$$

giving β_1 as a linear function of β_0.

Find β_2 from β_1 by

$$0 = \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta_1} + \left(\frac{\partial^2 \ell}{\partial \beta \partial \beta^T}\right)\Bigg|_{\beta_1} (\beta_2 - \beta_1)$$

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 **It.** by its expectation, from **Ex**.

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

$$b'(\theta_i) = \mu_i = \beta x_i, \quad g(\mu_i) = \mu_i, \quad \phi = \sigma^2$$
$$\theta_i = \beta x_i, \quad b(\theta_i) = \tfrac{1}{2}\theta_i^2.$$

[Hence $b''(\theta_i) = 1$, $\mathrm{var}(Y_i) = \phi b''(\theta_i)$: *check*.] In this case, it is trivial to show directly that $\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}$.

What does the glm algorithm do? If we substitute in

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

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 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 $\hat{\beta}$ is normal, mean β, and $\mathrm{var}(\hat{\beta}) = \sigma^2 / \sum x_i^2$ (The result here *exact*, not asymptotic only). The general glm formula gives us

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

and hence the general glm formula gives us

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

(consistent with the above exact variance, of course).

Example. Repeat the above, but now taking

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

i.e. the usual linear regression, with $\sum x_i = 0$ (without loss of generality) (so now 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

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

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

Example 2. Assume that

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

and

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

i.e.

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

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

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

$$\theta_i = \log\big(\mu_i/(1-\mu_i)\big), \quad b(\theta_i) = -\log(1-\mu_i).$$

Thus

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

all of which, of course, agrees with what we already know, that

$$Y_i \sim Bi(1, \mu_i) \Rightarrow \mathbb{E}(Y_i) = \mu_i, \quad \text{var}(Y_i) = \mu_i(1-\mu_i).$$

Furthermore,

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

(remembering that $g(\mu) = \log\big(\mu/(1-\mu)\big)$)

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

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

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

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

i.e.

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

where

$$\left.\begin{aligned} V_i &= \mu_i(1-\mu_i) \\ g(\mu_i) &= \log\big(\mu_i/(1-\mu_i)\big) \end{aligned}\right\}$$ **check**

This time, to compute $\hat{\beta}$, we find β_1 from β_0 by

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

and so on, and this converges to $\hat{\beta}$, where

$$\hat{\beta} \overset{\text{approx}}{\sim} N\big(\beta, v_n(\beta)\big)$$
$$v_n(\beta) = 1/\sum w_i x_i^2$$

where

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

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.)

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

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

and

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^T} = -\frac{1}{\phi} \frac{\partial^2 \psi}{\partial \beta \partial \beta^T} \quad \text{which is not a random variable.}$$

Hence

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

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

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

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

Testing hypotheses about β

and

A measure of the goodness of fit: the scaled deviance

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

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

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

$$\beta_1 = 0 \quad \text{where} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

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

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

so, 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

$$\begin{aligned} Y_{ij} &\sim Po(\mu_{ij}), \quad 1 \leq i \leq r, 1 \leq j \leq s, \\ \log \mu_{ij} &= \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \quad 1 \leq i \leq r, 1 \leq j \leq s, \end{aligned}$$

we may wish to test $(\alpha\beta)_{ij} = 0$ for all i, j.

In general, with $\ell(\beta)$ as in (glm) on p. 23, 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

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

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.

Distribution of the scaled deviance. If ω_c is true,

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

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.

Chapter 4: Regression for normal errors

Assume that

$$Y_i \sim NID(\beta^T x_i, \sigma^2) \qquad \dim \beta = p.$$

We may rewrite this assumption as

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

X being called the 'design' matrix, assumed here to have rank p.

Partition X, β as $(X_1 \vdots X_2), \binom{\beta_1}{\beta_2}$ respectively, so that $X\beta = X_1\beta_1 + X_2\beta_2$.Then, suppose we wish to test $H_0 : \beta_2 = 0$. Hence we can see that H_0 can be rewritten as $H_0 : \beta \in \omega_c$ where ω_c is a linear subspace of ω_f, which is $\mathbb{R}^p$. Now,

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

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 $(y-X\beta)^T(y-X\beta)$. Taking $\frac{\partial}{\partial\beta}\ldots = 0$ gives

$$(X^TX)\tilde{\beta} = X^Ty$$

and hence $$\tilde{\beta} = (X^TX)^{-1}X^Ty$$

and hence $$X\tilde{\beta} = X(X^TX)^{-1}X^Ty$$

which we rewrite as $$\tilde{y} = X\tilde{\beta} = P_fy$$

where

$\tilde{y}$ = the fitted values under the full model
$P_f = X(X^TX)^{-1}X^T$.

Check that P_f is a **projection matrix**. This means that it satisfies $P_f = P_f^T$ and $P_fP_f = P_f$. Hence

$$\max_{\omega_f} f(y \mid \beta, \sigma^2) = \frac{\text{const}}{\sigma^n} \exp -\tfrac{1}{2}(y-\tilde{y})^T(y-\tilde{y})/\sigma^2.$$

Here $\tilde{y}$ is the projection of y onto the subspace ω_f. Similarly,

$$\max_{\omega_c} f(y \mid \beta, \sigma^2) = \frac{\text{const}}{\sigma^n} \exp -\tfrac{1}{2}(y-\hat{y})^T(y-\hat{y})/\sigma^2,$$

where $\hat{y}$ is the projection of y onto ω_c (also called the fitted values under ω_c) so that $\hat{y} = P_cy$, where $P_c = X_1(X_1^TX_1)^{-1}X_1^T$. Hence the scaled deviance is

$$S(\omega_c, \omega_f) = 2\big[-\tfrac{1}{2}(y-\tilde{y})^T(y-\tilde{y}) + \tfrac{1}{2}(y-\hat{y})^T(y-\hat{y})\big]/\sigma^2.$$

Try illustrating this for yourself 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.)

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.

Hence $$S(\omega_c, \omega_f) = \big[(y-P_cy)^T(y-P_cy) - (y-P_fy)^T(y-P_fy)\big]/\sigma^2$$

giving $$S = [y^Ty - y^TP_cy - y^Ty + y^TP_fy]/\sigma^2$$

(using $P_c^TP_c = P_c$, etc.)

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

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 $r =$ 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

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

i.e. we refer [residual ss under ω_c– residual ss under ω_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

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

We overcome this difficulty by the following theorem, which we quote, WITHOUT PROOF.

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

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

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

$$\left.\begin{array}{ll} Q_f = (Y - \tilde{\mu})^T (Y - \tilde{\mu}), & \text{the residual ss fitting } \omega_f \\ Q_c = (Y - \hat{\mu})^T (Y - \hat{\mu}), & \text{the residual ss fitting } \omega_c \end{array}\right\}$$

(so by definition $Q_c \geq Q_f$). Then

$$\left.\begin{array}{rl} & Q_f/\sigma^2 \sim \chi^2_{df} \\ \text{and} \qquad & (Q_c - Q_f)/\sigma^2 \sim \chi^2_r \quad \text{(noncentral),} \end{array}\right\} \quad \text{independent}$$

and this second term is *central* χ^2_r if and only if $\mu \in \omega_c$. Here

$$\begin{aligned} df &= \text{degrees of freedom of } Q_f = n - \dim(\omega_f) \\ r &= \dim(\omega_f) - \dim(\omega_c). \end{aligned}$$

Corollaries

(1) $\quad \mathbb{E}(Q_f/df) = \sigma^2$

so $\mu \in \omega_f \Rightarrow Q_f/df$ (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.

Example 0. The distribution of the least-squares estimator.
Show that under ω_f, $\tilde{\beta}$ has the $N(\beta, V\sigma^2)$ distribution, where $V = (X^T X)^{-1}$.

Example 1. One-way Analysis of Variance (anova) 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 $1 \leq i \leq t, 1 \leq j \leq n_i$, and we assume that $\epsilon_{ij} \sim NID(0, \sigma^2)$, where y_{ij} = response of j^{th} observation on i^{th} treatment. So
ω_f is $\mathbb{E}(y_{ij}) = \mu + \theta_i$ for all i, j, and
ω_c is $\mathbb{E}(y_{ij}) = \mu$ for all i, j,
i.e. ω_c is no difference between treatments.
The residual ss (i.e. deviance) fitting ω_c is $\sum\sum(y_{ij} - \bar{y})^2 \equiv Q_c$.

Note that 'treatments' is a **factor** here: we wish to fit (θ_i) and not (θi). This will necessitate a **factor declaration** in any glm package. Omitting such a declaration would have serious and unwanted consequences: be consoled that this is one of many instances in computational statistics where we learn by making mistakes.

To fit ω_f, we must first 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 $\theta_1 = 0$. 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

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

and the same deviance

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

This gives the **Analysis of Variance**

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

(Here 'treatments' really means 'Due to differences between treatments')

where $$Q_f \text{ (check)} \equiv Q_c - \left[\sum \bar{y}_i^2 n_i - cf\right] = Q_c - S_T$$

$$cf = \text{correction factor} = \left(\sum\sum y_{ij}\right)^2/n.$$

Thus, to test ω_c, refer

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

Here $S_T = Q_c - Q_f$ = difference in deviances.

N.B. In using glm, you don't need to know the formulae for Q_c, Q_f etc, since glm works 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 to do F tests.

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 Splus) 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.

Suppose we have two factors having I, J levels respectively, and we have u observations on each of the IJ combinations of factor levels .We take as our model for the responses (y_{ijk})

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

with $1 \le i \le I, 1 \le j \le J, 1 \le k \le u$. For example, y_{ijk} is the k^{th} observation on the i^{th} country, j^{th} profession, and $u = 1$ in your practical worksheet. We might want to test

$$\omega_0 : \alpha = 0, \quad \beta = 0$$
$$\omega_1 : \alpha = 0$$
$$\omega_2 : \beta = 0$$

Here ω_f is $\mathbb{E}(y_{ijk}) = \mu + \alpha_i + \beta_j$. Thus

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

Exercises.

Note: 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

$$\Sigma\alpha_i = \Sigma\beta_j = 0, \quad \text{and} \quad a_1 = b_1 = 0,$$

then you can easily work out the relationships between the two sets of parameters $\mu, (\alpha_i), (\beta_j)$ and $m, (a_i), (b_j)$.

(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

$$\tilde{y}_{ijk} = \bar{y}_{i++} + \bar{y}_{+j+} - \bar{y}_{+++}$$

and hence the residual ss fitting $\omega_f = \sum\sum\sum (y_{ijk} - \tilde{y}_{ijk})^2$. Show that

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

In your practical worksheet on the Two-way Anova, you will see that the residual ss fitting $\omega_2, \omega_f, \omega_0, \omega_1$ correspond respectively to the deviances fitting

country only,
country + occupation,
a constant, and
occupation only.

Because of the **balance** of the design with respect to the two factors in question, these four deviances obey the linear equation given above. This leads us to our next important definition.

Definition of parameter orthogonality for a linear model

Suppose, as on p. 25, we have

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

and

$$X\beta = (X_1 \vdots X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{where } p_1 + p_2 = p.$$

Then β_1, β_2 are said to be mutually orthogonal sets of parameters if and only if

$$\underset{p_1 \times n}{X_1^T} \quad \underset{n \times p_2}{X_2} \quad = \quad \underset{p_1 \times p_2}{0}$$

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 01.

Lemma 01. β_1, β_2 are orthogonal if and only if $\hat{\beta}_1 \equiv \tilde{\beta}_1$ (an identity in Y), where

$$\begin{aligned} \hat{\beta}_1 &= \text{lse of } \beta_1 \text{ in fitting} \quad Y = X_1\beta_1 + \epsilon \quad \text{(i.e. assuming } \beta_2 = 0) \\ \tilde{\beta} = \tbinom{\tilde{\beta}_1}{\tilde{\beta}_2} &= \text{lse of } \beta \text{ in fitting} \quad Y = X\beta + \epsilon \quad \text{(i.e. the full model).} \end{aligned}$$

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

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

Orthogonality between sets of parameters has an important consequence for residual sums of squares, as shown in Lemma 02.

Lemma 02. If β_1, β_2 are orthogonal, then

$$\begin{aligned} &(\text{residual ss fitting } \mathbb{E}(Y) = X_1\beta_1) - (\text{residual ss fitting } \mathbb{E}(Y) = X_1\beta_1 + X_2\beta_2) \\ =&(\text{residual ss fitting } \mathbb{E}(Y) = 0) - (\text{residual ss fitting } \mathbb{E}(Y) = X_2\beta_2). \end{aligned}$$

Proof.

$$(\text{residual ss fitting } \mathbb{E}(Y) = X_1\beta_1) = (Y - X_1\hat{\beta}_1)^T(Y - X_1\hat{\beta}_1).$$

Further

$$(\text{residual ss fitting } \mathbb{E}(Y) = X\beta) = (Y - X\tilde{\beta})^T(Y - X\tilde{\beta}),$$

and

$$(\text{residual ss fitting } \mathbb{E}(Y) = 0) = Y^T Y.$$

Lastly,

$$(\text{residual ss fitting } \mathbb{E}(Y) = X_2\beta_2) = (Y - X_2\hat{\beta}_2)^T(Y - X_2\hat{\beta}_2).$$

The result follows from writing $X^T X$ as a partitioned matrix, and then using the fact that $X_1^T X_2 = 0$.

Apply Lemma 01 to answer the following questions, in which the errors ϵ_i may be assumed to have the usual distribution.

Exercise 1. In the model

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

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

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

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 $\mathbb{E}(Y) = 0$, but we are interested in fitting the model

$$\mathbb{E}(Y) = \mu 1_n$$

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

$$Y = X\beta + \epsilon \quad \text{and} \quad \dim(\beta) = p = 1 + p_1 + p_2,$$

which we rewrite as

$$Y = \mu 1_n + X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where β_1, β_2 are of dimensions p_1, p_2 respectively.

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

and $$y_i = \mu + \beta_1^T x_{1i} + \beta_2^T x_{2i} + \epsilon_i, \quad \text{say.}$$

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

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

Consider the linear hypotheses

$$\left.\begin{array}{l} \omega_0 : \mathbb{E}(Y) = \mu 1_n \\ \omega_1 : \mathbb{E}(Y) = \mu 1_n + X_2\beta_2 \\ \omega_2 : \mathbb{E}(Y) = \mu 1_n + X_1\beta_1 \\ \omega_f : \mathbb{E}(Y) = \mu 1_n + X\beta \end{array}\right\}$$

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

$$\begin{array}{r} \text{residual ss fitting } \omega_2 \text{ – residual ss fitting } \omega_f, \\ = \text{residual ss fitting } \omega_0 \text{ – residual ss fitting } \omega_1. \end{array}$$

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 $\mu^* = \sum Y_i/n = \bar{Y}$.

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 $i = 1, \ldots, n$ with $\Sigma x_i = \Sigma z_i = 0$, show that the parameters $\beta_1, \beta_2, \beta_3$ are mutually orthogonal if and only if $\Sigma x_i z_i = 0$.

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.

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

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

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.

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. For example, in a psychological experiment where A is the noise level , say 'quiet' or 'loud', and B is the sex of the subject, (male or female) then if 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,

eg of the fitted value of Y_{ijk} against j, for each level of i.

The standard glm constraints for α, β, γ are not the symmetric ones given above, but the 'corner point' ones

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

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 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

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

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 $\mathbb{E}(Y) = X\beta$ 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

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

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 'non-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

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

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:

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

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'.

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

$$\omega_f : 0 \le p_i \le 1, \quad 1 \le i \le k,$$

and under ω_f,

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

which is maximised with respect to $p \in \omega_f$ by $p_i = r_i/n_i$, [*check*].

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

$$\omega_c : logit(p_i) = \beta^T x_i, \quad 1 \le i \le k$$
$$\downarrow$$
$$p \times 1$$

where x_i are given covariates, $p < k$, say. Under ω_c,

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

$$\big(\text{since } p_i = e^{\beta^T x_i}/(1 + e^{\beta^T x_i}) = p_i(\beta)\big),$$

so $\ell(\beta)$ 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

$$2\times \text{ [loglikelihood maximised under } \omega_f - \text{ loglikelihood maximised under } \omega_c]$$
$$(*) \quad = 2 \sum \left(r_i \log \frac{r_i}{e_i} + (n_i - r_i) \log \frac{(n_i - r_i)}{(n_i - e_i)} \right) \equiv D, \quad \text{say.}$$

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 $D \le k - p$). Assuming that ω_c fits well, we may wish to go on to test, say, $\omega_1 : \beta_2 = 0$, where

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

where 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. 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

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

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

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 = No previous conviction, and p = Previous conviction.

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

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

Exercise 2. Try the above exercise with the link functions

$$g(p) = \Phi^{-1}(p), \quad \text{the probit}$$
$$g(p) = \log(-\log(1-p)), \quad \text{the complementary log} - \log.$$

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.

Chapter 6: Poisson regression and contingency tables

Example 1. The total number of reported new cases per month of AIDS in the UK up to November 1985 are listed below (data taken from A.M. Sykes 1986).

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

(data for 36 consecutive months – reading across)

Let us take as our model for Y_i the number of new cases reported in the i^{th} month,

$$Y_i \text{ independent} \qquad \text{Poisson with mean } \mu_i,\ 1 \le i \le 36.$$

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

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

giving

$$\begin{aligned} \mu_i &= \exp(\alpha + \beta i) \\ \ell(\alpha, \beta) &= \sum \log(e^{-\mu_i} \mu_i^{y_i}) \qquad \text{with } \mu_i = \mu_i(\alpha, \beta) \\ &= -\sum \exp(\alpha + \beta i) + \sum y_i(\alpha + \beta i). \end{aligned}$$

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

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

and we can find the se's of these estimators in the usual way. This 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. 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 .

Verify: The deviance for testing ω_c against ω_f is

$$2 \sum_1^{36} y_i \log \frac{y_i}{e_i}, \quad \text{where}$$

$e_i = e_i(\hat{\alpha}, \hat{\beta}) = \exp(\hat{\alpha} + \hat{\beta} i)$.

This deviance is approximately distributed as χ^2_{34}, if ω_c true, provided that (e_i) is not too small.

A useful general result. By writing $y_i = e_i + \Delta_i$, so that $\Sigma \Delta_i = 0$, and expanding $\log(1 + (\Delta_i/e_i))$ show that the deviance

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

is approximately

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

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.

Example 2. Accidents 1978–81, for traffic *into* Cambridge

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

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

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

for Road i, and Time of day j.

We might reasonably expect the number of accidents to depend on the traffic *volume*, so we look for a model

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

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

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 (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 following data for the periods when the Prime Ministers were, respectively, Attlee, Churchill, Eden, Macmillan, Douglas-Home, Wilson, Heath, Wilson, Callaghan, Thatcher, Major. In the years

1945–51, 51–55, 55–57, 57–63, 63–64, 64–70, 70–74, 74–76, 76–79, 79–90, 90–

when the Governments were, respectively,

lab,con,con,con,con,lab,con,lab,lab,con,con

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

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

(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: 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 10.898 (with 8 df). Note that the effect of political party is non-significant ($\hat{\alpha}_2 = -.04607(.2710)$.)

The coefficient $\hat{\gamma}$ is .9188 (se = .2344). For a Poisson process this coefficient would be **exactly** one. We can 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 11.016 (df = 9).

Example 4. Observe that if R_i is distributed as $Bi(n_i, p_i)$ where n_i is large and p_i is small, then R_i is approximately Poisson, mean μ_i, where

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

In this case, binomial logistic regression of the observed values (r_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 (r_i) on (x_i), with the usual log-link function, and an **offset** of $(log(n_i))$.

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'.

$$r/n = 33/3271, 63/7257, 157/5065 \quad \text{for males}$$
$$r/n = 38/2486, 108/8877, 159/3520 \quad \text{for females.}$$

Here, using figures from the Metropolitan police,

n = the number reported missing during the year ending March 1993, and
r = 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.

Contingency tables

Example. The *Daily Telegraph* (28/10/88), under the headline 'Executives seen as Drink Drive threat', presented the following data from breath-test operations at Royal Ascot and at Henley Regatta:

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

So at Ascot, 1.1% of those tested are arrested, compared with 0.7% at Henley.

The multinomial distribution

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

$$p(n \mid p) = n! \prod\prod \frac{p_{ij}^{n_{ij}}}{n_{ij}!}, \quad \sum\sum p_{ij} = 1.$$

We wish to test

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

i.e. (for this example) whether or not you are arrested is independent of whether you are at Ascot or Henley.

[Verify, for this example, H_0 is equivalent to

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

i.e. $P(\text{arrested}|\text{Ascot}) = P(\text{arrested}|\text{Henley})$.]

Verify H_0 is equivalent to

$$* \qquad \log p_{ij} = \text{const} + \alpha_i + \beta_j \quad \text{for some } \alpha, \beta.$$

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'.

Using the Poisson error function in glm for the multinomial distribution

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, $1 \le i \le r, 1 \le j \le c$. 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

$$H_0 : p_{ij} = \alpha_i \beta_j \text{ for some } \alpha, \beta (\textstyle\sum\sum \alpha_i \beta_j = 1) \text{ against}$$
$$H : p_{ij} \ge 0, \textstyle\sum\sum p_{ij} = 1,$$

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

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

where e_{ij} = expected frequency under H_0, so $e_{ij} = y_{i+}y_{+j}/n$. We 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 $\big($where $f = (r-1)(c-1)\big)$.

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

$$HP_0 : \log \mu_{ij} = \alpha'_i + \beta'_j \text{ for some } \alpha', \beta', \text{ for all } i, j, \text{ against}$$
$$HP : \log \mu_{ij} \text{ any real number.}$$

Now

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

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

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

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

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

say, and 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

$$2L(\hat{\mu}) - 2L(\mu^*) =$$
$$DP = 2[-\sum\sum \hat{\mu}_{ij} + \sum\sum y_{ij} \log \hat{\mu}_{ij} + \sum\sum \mu^*_{ij} - \sum\sum y_{ij} \log \mu^*_{ij}].$$

But $\sum\sum \hat{\mu}_{ij} = \sum\sum \mu^*_{ij}$ (*check*). 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 r.v.s. This is a special case of the following

General result, relating Poisson and multinomial loglinear models.

We assume that we are given

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

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

$$\begin{aligned} H_0 &: \log p_i = \mu + \beta^T x_i, 1 \le i \le k \text{ for some } \beta \text{ (of dim } p) \\ &\quad \text{(where } \mu \text{ is such that } \textstyle\sum p_i = 1) \text{ against} \\ H &: p_i \ge 0, \textstyle\sum p_i = 1. \end{aligned}$$

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 that we are testing

$$\begin{aligned} HP_0 &: log(\mu_i) = \mu' + \beta^T x_i \text{ against} \\ HP &: \log(\mu_i) = \text{ any real numbers.} \end{aligned}$$

Reminder: In proving this general result we make use of the following **Lemma**.

Suppose that the pdf of 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

$$L(\beta) = \log b(\beta) + \beta^T t(y), \ \frac{\partial L}{\partial \beta} = \ldots$$ (see p. 12 for completion)

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

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

which is maximised with respect to β by

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

$**$ giving $e_i = n p_i^*$ as 'fitted values' under H_0, $p_i^* \propto \exp(\hat\beta^T x_i)$, $\hat\beta$ being 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

$$\begin{array}{lll} & HP_0 : \log \mu_i = \mu' + \beta^T x_i, \quad 1 \le i \le k & (\dim HP_0 = p+1) \\ \text{against} & HP \ : \log \mu_i \quad \text{anything} & (\dim HP = k) \end{array}$$

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

So, under HP_0,

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

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

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

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

where $\hat{\mu}_i = \text{mle of } \mu_i \text{ under } HP$, hence $\hat{\mu}_i = y_i$

$\mu_i^* = \text{mle of } \mu_i \text{ under } HP_0$, hence $\sum \mu_i^* = y_+$

and $\mu_i^* = e_i \quad \text{with } e_i \text{ as in } **.$

Hence $\sum \hat{\mu}_i = \sum \mu_i^*$. Hence

$$D \text{ (multinomial deviance)} = 2 \sum y_i \log(y_i/e_i)$$
$$\equiv DP \text{ (Poisson deviance)} = 2 \sum y_i \log(\hat{\mu}_i/e_i).$$

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.

Exercise 2. Let x_1, z_1 be scalars and let x_2, z_2 be p-dimensional vectors. Take a_{11} scalar, $a_{12} = a_{21}^T$ vectors, 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 (hence discovering the form of the inverse of a partitioned matrix).

Now use this result 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.

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,. . ., 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

$$p_{ijk} = P(A = i, B = j, C = k) \quad \text{for} \quad i = 1, \ldots, r, j = 1, \ldots, c, k = 1, \ldots, \ell$$

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 intentions (Labour,Conservative,Other)
B = employment status (employed,unemployed,student,pensioner)
C = place of residence (urban,non-urban)

Let us suppose that the (fictitious) resulting 3-way table is

	C = urban			C = non-urban		
	A =			A=		
B =	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

There are 8 different 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

$$P(B = j, C = k|A = i) = P(B = j|A = i)P(C = k|A = i) \quad \text{for all } i, j, k.$$

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 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 and 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

$$H_0 \subset H_1 \subset H_3, \text{ and } H_0 \subset H_2 \subset H_3 \quad \text{and} \quad H_1 \cap H_2 = H_0.$$

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

$$A * B + A * C \quad \text{or equivalently} \quad A * (B + C)$$

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 the example 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 deviance for

$$\begin{aligned} (A + B + C) * (A + B + C) &\quad \text{is 15.242 (6 df)} \\ (A + B) * C &\quad \text{is 122.07 (12 df)} \\ (A * B) + C &\quad \text{is 27.144 (11 df)} \\ A + B + C &\quad \text{is 132.3 (17 df).} \end{aligned}$$

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

	C = 1		C = 2	
	A = 1	A = 2	A = 1	A = 2
B = 1	17	23	36	50
B = 2	29	14	59	24

Show that the deviance for fitting the model $A*B+B*C+A*C$ is .12362, 1 df.

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,... symmetrically. For example it may be more natural to treat

A as a **response** variable, and
B,C,... 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,....

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 $2\times2\times2$ 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:

	Arrested	Not arrested
Ascot	24 ⟨ 23 men / 1 woman	2210 ⟨ 2 men / 2208 women
Henley	5 ⟨ 3 men / 2 women	680 ⟨ 340 men / 340 women

Hence although the overall arrest rate at 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

3	340
2	340

gives a deviance of 0.19990, while the 2×2 table

23	2
1	2208

gives a deviance of 234.0. Of course, it is scarcely necessary to find the exact numerical values of the deviances to understand about the 3-factor interaction: we include them here merely for completeness.

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

$$\psi(t) = E(exp(it^TY)) = \int exp(it^Ty)f(y|\mu,\Sigma)dy$$

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

$$\mathbb{E}(Y_i) = \mu_i, \text{cov}(Y_i, Y_j) = \Sigma_{ij}.$$

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, $Y_1 \sim N_1(\mu_1, \Sigma_{11})$.

Appendix 2: Regression diagnostics for the Normal Model

Residuals and leverages

Take $y_i = \beta^T x_i + \epsilon_i$, $1 \le i \le n$, $\epsilon_i \sim NID(0, \sigma^2)$.

Equivalently,
$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\,\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

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

$$\left.\begin{aligned} \hat{\beta} &\sim N(\beta, \sigma^2 (X^T X)^{-1}) \\ \text{and} \qquad \frac{R(\hat{\beta})}{\sigma^2} &= \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{\sigma^2} \sim \chi^2_{n-p} \end{aligned}\right\} \text{ independent,}$$

and hence test, e.g., $\beta_2 = 0$, by using $\hat{\beta}_2$, se $(\hat{\beta}_2)$ etc.

All our hypothesis tests will depend on the assumption

$$\epsilon_i \sim NID(0, \sigma^2)$$

so we need some way of checking this: this is what *qqplots* do.

Define
$$\begin{aligned} \hat{Y} = X\hat{\beta} &= X(X^T X)^{-1} X^T Y, && \text{fitted value} \\ &\equiv HY \quad \text{say}, && H \text{ 'hat matrix'} \\ \text{residual} \quad \hat{\epsilon} &= Y - \hat{Y}, && \text{observed–fitted.} \end{aligned}$$

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 $h_i = H_{ii}$; then

$$\hat{\epsilon}_i \sim N\big(0, \sigma^2 (1 - h_i)\big).$$

We define

$$\eta_i = \hat{\epsilon}_i \Big/ \sqrt{1 - h_i}$$

as the *standardised* residuals. We do a visual check of whether $\eta_1, \ldots, \eta_n$ forms a r.s. from $N(0, \sigma^2)$ as follows.

What is the sample distribution function of $(\eta_1, \ldots, \eta_n)$? It is defined as

$$F_n(x) \text{ say} = \frac{\text{no. out of } (\eta_1, \ldots, \eta_n) \le x}{n}.$$

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 qqplot does for you. Filliben's coefficient measures the closeness to a straight line. (The Weisberg-Bingham test is also useful.)

Leverages. Note: $\hat{Y} = HY$, $H = X(X^TX)^{-1}X^T$, giving

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j \qquad \text{say, } h_{ii} = h_i.$$

Now, because $\quad \text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_i), \quad$ we can see $h_i \leq 1$

and because $\quad H \geq 0, \quad$ we can see $h_i \geq 0$.

The larger h_i is, the closer $\hat{y}_i$ will be to y_i. We say that $\mathbf{x}_i$ has high 'leverage' if h_i large. Note

$$\text{rank}(H) = p, \quad HH = H \Rightarrow \sum_{1}^{n} h_i = tr(H) = \text{rank}(H) = p.$$

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

$$\underset{n\times p}{X} = (a_1 \vdots \ldots \vdots a_p) \quad \text{where} \quad \overset{\text{orthogonal columns}}{\swarrow} \quad a_i^T a_j = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases}$$

Then show

$$h_i = a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2, \quad 1 \leq i \leq n,$$

(so verify $\sum_1^n h_i = p$).

Exercise 2. Most modern regression software will give you qqplots 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 quick graphical checks.

RESUMÉ. The important things you need for this course are

(i) How to find $\dfrac{\partial}{\partial\beta}$, $\dfrac{\partial^2}{\partial\beta\partial\beta^T}$ (e.g. of $L(\beta)$).

(ii) How to find $\mathbb{E}(Y)$ and $\text{cov}(Y)$.

(iii) Basic properties of normal, Poisson and binomial.

(iv) Asymptotic distribution of $\hat{\theta}$ (mle).
Application 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.