MAS31001/MAS61001/MAS61004
Generalized Linear Models

Kostas Triantafyllopoulos
University of Sheffield

Module Information

1. This module is taught by Dr Kostas Triantafyllopoulos. The module has a Blackboard page, and all course materials will be available there. There is a more detailed information about the structure of the module in Blackboard. Below is a brief summary.

2. You can use the discussion boards on Blackboard for asking questions, and you can post questions anonymously.

3. There are three sets of non-assessed exercises, which will be available in due course on Blackboard. Deadlines for these will be given on the discussion board. You can ask for help with these exercises at any time.

4. There is a project, worth 30% of the assessment for MAS61001 and MAS61004. More details and deadlines will be posted on Blackboard.

5. There are exercises ("Tasks ") distributed throughout the notes. You should attempt these as we reach them during the semester. Solutions of these will be provided and for some of the tasks I have made videos with clarifications. These will be available in Blackboard.

6. You can contact me and arrange to see me, if you want to discuss anything about the module.

7. The module will cover Chapters 1-6. Chapter 7 is left here, if someone is interested in Contigency tables.

8. References. The following are some references on generalized linear models and I have provided links you can legally download the books for free.

   (a) Faraway, J.J. (2006) Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman-Hall.
   `https://englianhu.files.wordpress.com/2016/01/`
   `faraway-extending-the-linear-model-with-r-e28093-2006.pdf`

   (b) Dobson, A.J. (2001) An Introduction to Generalized Linear Models. 2nd edition, Chapman-Hall.
   `https://reneues.files.wordpress.com/2010/01/an-introduction-to-generalized`
   `pdf`

   (c) Lindsey, J.K. (1997) Applying Generalized Linear Models. Springer.
   `https://reneues.files.wordpress.com/2010/01/lindsey-j-k-applying-generali`
   `pdf`

   (d) McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd edition, Chapman-Hall.
   `http://www.utstat.toronto.edu/~brunner/oldclass/2201s11/readings/`
   `glmbook.pdf`

Contents

Chapter 1

Brief review of the linear model

This chapter gives a short review of the basic theory for linear models. Technical details and examples of data analysis are not provided. The main purpose of this chapter is:

- to give a short reminder of the linear model and to link this module with the linear model studied in other modules;
- to obtain an appreciation of the assumptions made in linear models. Knowledge of these assumptions will help motivate the *generalised linear models* in Chapters 2-7. We shall need to replace those assumptions by new more general ones to enable us to develop the theory and application of generalised linear modelling.

You should not spend a lot of time on this chapter. If you feel there are parts which you are not sure about you should ask me or refer to modules covering linear models.

1.1 The linear model

Suppose we want to see how a response variable y relates to several explanatory variables $x_1, \ldots, x_p$. We have data

$$(x_{i1}, \ldots, x_{ip}, y_i), \quad i = 1, \ldots, n$$

Linear models have two parts: a `linear predictor` and a `random error`. The linear predictor formalises the idea that the response is a linear combination of terms involving the explanatory variables. The error term allows for variation in the response for identical values of the explanatory variables. With the variables above we could propose the statistical (linear) model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1.1}$$

to explain how the response depends on the explanatory variables (analogous to the simple regression model). This is an example of **multiple regression**. $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is the linear predictor, ϵ_i is the error term.

It is convenient to express such models using vectors and matrices. We can always collect the observed values of the variable y into a vector $y = (y_1, \ldots, y_n)^T$ and we can similarly define $x = (x_1, \ldots, x_n)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. If we also define the design matrix X to be the $n \times (p+1)$ matrix, which has columns $(1, , \ldots, 1)^T$ and $(x_{11}, \ldots, x_{n1})^T$, … $(x_{1p}, \ldots, x_{np})^T$, then (1.1) can be written as

$$y = X\beta + \epsilon. \tag{1.2}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the parameter vector.

1.2 Estimating model parameters using the least squares method

1.2.1 Fitted values

Suppose we are performing simple linear regression so that we are estimating an intercept β_0 and gradient term β_1. Suppose we have estimates of these two parameters $\hat{\beta}_0$ and $\hat{\beta}_1$. Then the fitted value for observation i is $\hat{\beta}_0 + \hat{\beta}_1 x_i$. Suppose we have an estimate $\hat{\beta}$ of the parameter vector β. We define the fitted values to be $X\hat{\beta}$ (this gives the vector of fitted values).

1.2.2 Residuals

Residuals play a key role in linear models. We have seen that the linear model has several parameters (the β_i contained in the vector of parameters β). We need to use our observed data to estimate these parameters. The parameters that we choose are those that minimize the sums of squares of the residuals.

The residual for observation i (usually denoted as e_i) is defined to be $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ (i.e. observed value - fitted value). The vector of residuals will be

$$\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = e = y - X\hat{\beta}$$

and the sum of squares of these residuals is nicely expressed as

$$S_r = S(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = e^T e = (y - X\hat{\beta})^T (y - X\hat{\beta}). \tag{1.3}$$

This sum is known as **residual sum of squares** and plays an important role in the the analysis of linear models.

The least squares estimator is that $\hat{\beta}$ which minimises $S(\hat{\beta})$. Intuitively, a value of $\hat{\beta}$ that makes the residuals small "fits" the data well. So the least squares estimator is generally described as giving the best fit.

1.2.3 LS estimators

Theorem 1 *Assume that X has full rank. Then the least squares (LS) estimator of β is*

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{1.4}$$

1.3 Gauss-Markov conditions

So far we have not imposed any statistical assumptions on model (1.2). Although the LS estimator $\hat{\beta}$ is still valid with no further assumptions, for the full development of the linear model we need to impose distributional assumptions on $\epsilon = y - X\beta$. These assumptions, known as Gauss-Markov conditions, can be expressed as

$$E(\epsilon) = 0 \quad \text{and} \quad \text{Var}(\epsilon) = \sigma^2 I, \tag{1.5}$$

and I denotes the $n \times n$ identity matrix.

This effectively says that the mean of ϵ_i is zero, the variance of ϵ_i is σ^2 and that ϵ_i is uncorrelated with ϵ_j for $i \neq j$, with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$. We further assume that the distribution of ϵ_i is normal, written as $\epsilon_i \sim N(0, \sigma^2)$. The above assumptions can be compactly expressed by

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

where here $N(\cdot)$ denotes the n-variate normal distribution, having zero vector mean and covariance matrix $\sigma^2 I$.

1.4 Maximum likelihood estimators

The full version of our model, in which we include the assumption of normality, simply says that $y \sim N(X\beta, \sigma^2 I)$. The likelihood therefore comes directly from the definition of the multivariate normal density.

$$\begin{aligned} L(\beta, \sigma^2; y) &= f(y|\beta, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right) \\ &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right) \end{aligned}$$

In this derivation we have used the fact that $|\sigma^2 I_n| = (\sigma^2)^n$.

To maximise this likelihood with respect to β, we obviously must minimise $(y - X\beta)^T(y - X\beta)$. But we already know the result of this. It yields the least squares estimates, so the MLE of β is the same $\hat{\beta}$ that we have already obtained using both least squares and the Gauss-Markov theorem.

After we have done this first stage maximisation of the likelihood, we have

$$L(\hat{\beta}, \sigma^2; y) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta})\right).$$

and we now need to maximise this with respect to σ^2. This is easily done (by differentiating the log-likelihood with respect to σ^2 and setting the result equal to zero - see review exercises 2) and produces the MLE, $n^{-1}(y - X\hat{\beta})^T(y - X\hat{\beta})$.

1.5 Linear models assumptions

Recall from Section 1.3 the Gauss-Markov conditions. These enable us to derive the LS estimator of $\hat{\beta}$. In addition to the Gauss-Markov conditions, the distribution of the error term ϵ is usually specified as Gaussian (or normal distribution), or $\epsilon \sim N(0, \sigma^2 I)$. This is a typical requirement to derive confidence intervals, maximum likelihood likelihood estimators and hypothesis testing. However, the normality assumption has important implications, which are highlighted next and which will motivate the development of *Geeneralised linear modelling.*

From the Gauss-Markov conditions we can obtain

$$y \mid \beta, \sigma^2 \sim N(X\beta, \sigma^2 I).$$

This has the following implications:

1. Given X, β and σ^2, the distribution of y is assumed to be Gaussian (or normal distribution).

2. Given X, β and σ^2 the variance of y is equal to $\sigma^2 I$. This means that y_i and y_j are independent (so that their covariance is zero) and that $\text{Var}(y_i) = \text{Var}(y_j) = \sigma^2$, for any $i \neq j$. Hence, $y_1, \ldots, y_n$ are independent and they have the same variance.

The following observations are important assumptions we are making when applying the linear model. Departures from these assumptions may well invalidate our analysis. Such departures may include:

A The distribution of our data y_i may not be Gaussian. Many examples of such data can be thought of such as data generated from a skew distribution such as log-normal, or gamma. In particular, if data are generated from a discrete distribution (Poisson, binomial), then the above Gaussian distribution seems completely inappropriate.

B The data $y_1, \ldots, y_n$ may not be independent, for example y_{i+1} may depend on y_i. This is the situation in time series analysis where i represents time (say the average value of temperature collected today depends on the relevant value collected yesterday). You will study time series analysis in MAS372 / MAS61005, but in the context of this module one could consider to introduce a non-zero covariance matrix C to replace I, so that $\text{Var}(y \mid X, \beta, \sigma^2) = \sigma^2 C$.

C The data $y_1, \ldots, y_n$ may not have the same variance (heteroskedastic). For example in finance, the variance of financial returns – also known as volatility – is known to change over time. In most applications if the length of the data n is long enough, it is expected that $\text{Var}(y_i) \neq \text{Var}(y_j)$, for some $i \neq j$.

Generalized linear models work around model-modifications and generalisations to cater for point (A) above. We will no-longer restrict ourselves to data generated from a normal distribution. However, we shall still consider that our data are independent (point B), although we will study in Chapter 6 a class of models which relax the independence assumption. Point (C) is partly dealt with by the response distribution assumed, e.g. if the response follow a Poisson distribution, then its mean and variance are the same; if the mean λ_i depends on i, then the variance does the same (see e.g. Chapter 5).

Chapter 2

Logistic regression

2.1 Introduction

This chapter and those following are concerned with extensions of the basic linear model briefly discussed in Chapter 1. The resulting models are called *Generalized Linear Models* (GLMs). (They are to be distinguished from *general linear models*, meaning standard normal-theory linear models.)

Generalized linear models (GLMs) provide a widely useful, extension of the usual Normal linear model. The extensions allow, for example, data about proportions or data about counts to be modelled directly as Binomial or Poisson, respectively – there is no need for transformations of the data or for assuming approximate Normality. In addition, the use of a link function between the linear predictor and the mean (so the model is non-linear) can ensure that estimated means are always within bounds (proportions in $[0, 1]$, counts non-negative). A simple device means that Multinomial data can be regarded as Poisson data, and hence that many models for the analysis of contingency tables also fall within the GLM framework. In this chapter we introduce the main ideas of GLMs by considering *Binary data* for motivation; we are able to discuss and motivate GLMs for this special and important case of data. Before we start, below we give a reminder of the basic distribution assumptions of the Normal linear model, which we wish to extend; see also Section 1.5.

2.1.1 Standard linear model assumptions

The standard Normal-theory linear model is usually expressed as $Y \sim N(X\beta, \sigma^2 I)$. However, for developments here, we separate the modelling of the mean from assumptions about the distribution of Y. To this end, let

$E(Y_i) = \mu_i$ and let x_i be the vector of explanatory variables for observation i, i.e. x_i^T is the i^{th} row of the design matrix X. Then the model can be written as:

(i-G) $E(Y_i) = \mu_i = x_i^T \beta = \sum_j \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \ldots$, for the ith observation
(thus the mean is linear in β);

(ii-G) $\text{var}(Y_i) = \sigma^2$ (the variance is constant);

(iii-G) the Y_i are independent and Normal.

2.1.2 Why we need to extend linear models theory

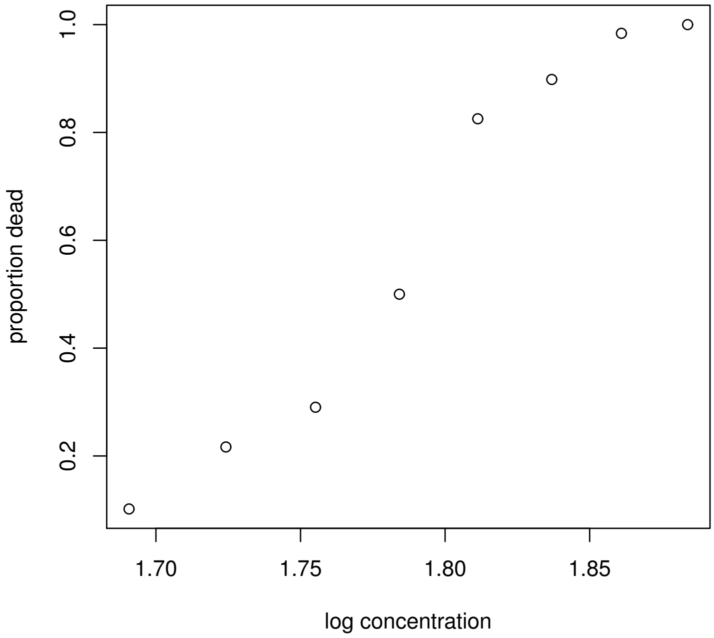

Figure 2.1: Proportion of beetle deaths by log concentration

If we have independent observations, as in the linear models case, in which situations do we need to extend the linear model framework?

The following data, from Dobson (1990, Example 8.1, pp. 108–111; 2002, §7.3.1, pp. 119–121), give the number of beetles exposed (n_i) and killed

(d_i) after 5 hours exposure to gaseous carbon disulphide (CS_2) at various concentrations (x_i in $\log_{10} CS_2$ mg/l).

x_i	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
n_i	59	60	62	56	63	59	62	60
d_i	6	13	18	28	52	53	61	60

The data is in the 'beetle' dataframe on the .RData workspace on Blackboard. Figure 2.1 shows a plot of the proportion of beetles dead by log concentration of CS_2. The number of beetles used in the experiment at each log dose is fixed and it is reasonable to assume that, at a given dose, each beetle dies with a certain probability, independently of the other beetles.If we want to predict the proportion of deaths at a log concentration of 1.77 we need to construct a model that relates the proportion of deaths to the log concentration. One way to do this would be to try to fit a linear model. Looking at the plot there is clear curvature and it is not of a simple form. We may be able to find regressor variables (functions of the log concentration) that enable us to fit a linear model reasonably well but this seems a rather ad hoc way of doing it. A more important issue is that the response varies only between 0 and 1, using a linear model we could not easily ensure this is the case in the model. For example, using a linear model for the beetles data, the proportion of deaths at a log concentration of 1.1 would almost certainly be negative. Generalized linear modeling (GLM) offers a flexible way of fitting such models.

In linear models we must have a response variable that is normally distributed. If we have a binary response then clearly we can't use linear model theory. GLM theory copes perfectly well with a binary response and indeed this is probably the most common application of it (certainly in the medical literature where it is used extensively for risk modelling in case control studies).

Binary data occur when there are just two possible responses, such as dead or alive. The usual generic terms are 'failure' and 'success'. These can be coded so that the variables Y_i take only the values 0 or 1, respectively. The probabilities are given by:

$$P(Y_i = 1) = \mu_i \quad \text{and} \quad P(Y_i = 0) = 1 - \mu_i.$$

Then $E(Y_i) = \mu_i$, with $0 \leq \mu_i \leq 1$. Clearly we cannot set $E(Y_i) = \mu_i = x_i^T \beta$, as we would in Normal linear models, as the support of μ_i is not the real line $(-\infty, +\infty)$. Thus we need to map μ_i to $(-\infty, +\infty)$. One possibility is to set

$$\log \frac{\mu_i}{1 - \mu_i} = x_i^T \beta,$$

which is valid as it maps μ_i to the real line. In GLMs we shall write

$$\mathrm{logit}(\mu_i) = \log \frac{\mu_i}{1-\mu_i} = x_i^T\beta$$

or

$$\mu_i = h(x_i^T\beta) = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}$$

so that $0 \leq \mu_i \leq 1$, for any x_i^T and β. This mapping (referred to as *link function* or just link) is known as *logit* and is one popular choice for binary data. There are other options of link functions, which will be explored later on.

We revisit the beetle data described on page 12. With the notation defined earlier the number of deaths (D_i) at log dose x_i can be therefore be modeled as $D_i \sim Bi(n_i, \mu_i)$. It follows that $E(D_i) = n_i\mu_i$ and $\mathrm{var}(D_i) = n_i\mu_i(1-\mu_i)$. Here n_i is the number of beetles exposed at a certain dose and μ_i is the probability of death. How does the probability of death μ_i relate to the log dose x_i? In a GLM we allow a general relationship of the form $\mu_i = h(\beta_0 + \beta_1 x_i)$ where we can choose the inverse link function (h) that best fits. The choice of h is one of the decisions to be made when fitting a GLM.

Instead of considering the number of beetles dying we can deal with the proportion of beetles dying: $Y_i = D_i/n_i$. Then $E(Y_i) = E(D_i)/n_i = \mu_i = h(\beta_0 + \beta_1 x_i)$, $\mathrm{var}(Y_i) = 1/n_i^2 \mathrm{var}(D_i) = \mu_i(1-\mu_i)/n_i$. If we do this the we have a GLM with:

i) inverse link h

ii) $E(Y_i) = \mu_i = h(\beta_0 + \beta_1 x_i)$

iii) $\mathrm{var}(Y_i) = \mu_i(1-\mu_i)/n_i$

Key point The variance of Y_i is allowed to depend on the mean value μ_i (which in turn depends on the linear predictor). This is not allowed in linear models but because we have widened the choice of probability distributions for Y_i (beyond the Normal distribution), they can now have variances that depend on their expected values in some well-defined way.

Link function

In this example, the mean of the response μ_i is a probability. To be sure that $\mu_i = h(\beta_0 + \beta_1 x)$ gives a probability whatever value x takes, the function h must take values in $[0, 1]$ and, to have an inverse, it must be monotone, so that (assuming, without any loss of generality, it is increasing) mathematically h is in fact a distribution function. If $h(u) = \Phi(u)$, where Φ is the distribution of the standard Normal, this gives rise to the **probit** link (and $\Phi^{-1}(u)$ is called the probit of u); if $h(u) = e^u/(1+e^u)$, then $g(u) = \log(u/(1-u))$, which is often called logit(u), leading to the **logit** link, which is already discussed above. Note that $0 < h(u) < 1 \quad \forall u \in \mathbb{R}$ as required.

Frequently, as in a toxicity trial of the beatles data (see page 12), there are n_i replicates for the same x_i giving rise to n_i binary observations Y_{ij}, say. Then summing the observations with the same x_i, $S_i = \sum_j Y_{ij}$ gives the number of successes, which, providing that the assumptions of independence and constant probability are satisfied, contain all information from the experiment about μ. (Technically, S is a sufficient statistic for μ.) Hence the variables S_i, whose distributions are $Bi(n_i, \mu_i)$, are usually analysed instead of the original Y_{ij}. In the context of generalized linear modelling it is more natural to deal with $Y_i = S_i/n_i = \sum_j Y_{ij}/n_i$, (the observed proportion of successes). This has $E(Y_i) = \mu_i \in [0,1]$.

EDA for Logistic regression

Jeremy Oakley has written some notes on exploratory data analysis (EDA) for logistic regression.

The notes are available here http://www.jeremy-oakley.staff.shef.ac.uk/mas61004/EDAtutorial/eda-for-logistic-regression.html

2.2 Binomial likelihood

For the binomial case and the logit link function discussed above (e.g. the beetle data) in which $y_i \sim \text{Bin}(n_i, \mu_i)$ and with m independent observations, the likelihood is

$$\begin{aligned}
L(\mu_i, y_i; x_i) &= \prod_{i=1}^{m} \binom{n_i}{y_i} \mu_i^{y_i}(1-\mu_i)^{n_i-y_i} \\
L(\beta, y_i; x_i) &= \prod_{i=1}^{m} \binom{n_i}{y_i} \left(\frac{\exp(x_i^T\boldsymbol{\beta})}{1+\exp(x_i^T\boldsymbol{\beta})}\right)^{y_i} \left(\frac{1}{1+\exp(x_i^T\boldsymbol{\beta})}\right)^{n_i-y_i} \\
&= \prod_{i=1}^{m} \binom{n_i}{y_i} (\exp(x_i^T\boldsymbol{\beta}))^{y_i}(1+\exp(x_i^T\boldsymbol{\beta}))^{-n_i}
\end{aligned}$$

and the log-likelihood as

$$l(\beta, y_i; x_i) = \sum_{i=1}^{m} \left[y_i x_i^T\boldsymbol{\beta} - n_i log(1+\exp(x_i^T\boldsymbol{\beta}))\right] + \boldsymbol{constant}$$

This log-likelihood can be maximised with respect to β, hence provide MLE estimators for β. For simplicity consider the beetle data above so that

$x_i^T\beta = \beta_0 + x_i\beta_1$, where x_i here is the covariate (concentrations to which the beetles are exposed to). The two partial derivatives of $l(\beta, y_i; x_i)$ with respect to β_0 and β_1 are

$$\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{m}\left[y_i - \frac{n_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]$$

and

$$\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{m}\left[y_i x_i - \frac{n_i x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]$$

Equalising these two to zero we obtain the equations

$$\sum_{i=1}^{m} \frac{n_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \sum_{i=1}^{m} y_i$$
$$\sum_{i=1}^{m} \frac{n_i x_i \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \sum_{i=1}^{m} y_i x_i,$$

which can be solved to provide the MLE's $\hat{\beta}_0$ and $\hat{\beta}_1$.

2.3 Fitting a GLM

This section describes how we can fit a GLM in R and we apply it using the Beetles data.

The R function —glm— fits generalized linear models. Basic usage is

```
glm(response variable ~ explanatory variables,
family = family.name, data=...)
```

where —family.name— is the name of the distribution for the response and —data— is the data frame being used. The response data must be appropriate for the distribution being fitted. For a Gamma model, for example, observations must be non-negative. For a —binomial— model (logistic regression) there are several ways to fit the model (using proportions with weights or numbers of successes and failures - see chapter 4 for more details). The formula

```
response variable ~ explanatory variables
```

is a model formula of the same type as used by other linear modelling functions such as —lm— and —lme—. The usual formula syntax applies.

The default link function used by —glm— is the canonical link function for the specified distribution (see section 3.2.5). A different link would be specified by an argument within the —family— object: for example

```
    glm(response ~ explanatory variables, family = family.name(link = link.name),
    data=...)
```

Here `link.name` is the name of the link being used. The —glm— function produces an object of class —glm— (as —lm— and —lme— produce objects of class —lm— and —lme— respectively) from which information about the fit may be extracted as usual.

We illustrate these commands by fitting a GLM to the beetle data. The beetles data is a data frame containing the following:

```
> beetle
    conc number dead propn.dead
1 1.6907     59    6  0.1016949
2 1.7242     60   13  0.2166667
3 1.7552     62   18  0.2903226
4 1.7842     56   28  0.5000000
5 1.8113     63   52  0.8253968
6 1.8369     59   53  0.8983051
7 1.8610     62   61  0.9838710
8 1.8839     60   60  1.0000000
```

For the beetles data we have a binomial distribution for the response. If we choose to use the probit link function then we would use the command

```
     glm(propn.dead ~ conc, family = binomial(link=probit),
     weights=number, data=beetle)
```

This yields the output

```
> first.beetle.glm=glm(propn.dead~conc,family=binomial(link=probit),
+ weights=number,data=beetle)
> first.beetle.glm

Call:glm(formula=propn.dead~conc,family= binomial(link=probit),
+ data=beetle,weights=number)

Coefficients:
(Intercept)         conc
```

```
     -34.94        19.73

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance: 10.12        AIC: 40.32
```

Parameter estimates
As for linear models we obtain estimates of the parameters in the linear predictor; $\beta_0 + \beta_1 x = -34.94 + 19.73x$. We will talk much more about deviance later but it is used in model building (deciding which terms go in the linear predictor), a bit like using the `anova` command for nested linear models. As for linear models, the summary command gives us a little more information:

```
> summary(first.beetle.glm)

Call:
glm(formula = propn.dead  ~ conc, family = binomial(link = probit),
    data = beetle, weights = number)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
conc          19.728      1.487   13.27   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  10.120  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4
```

Fitted values
For now we shall just look at the fitted values.

```
> round(first.beetle.glm$fitted,digits=3)
```

```
    1     2     3     4     5     6     7     8
0.057 0.179 0.379 0.604 0.788 0.904 0.962 0.987
```

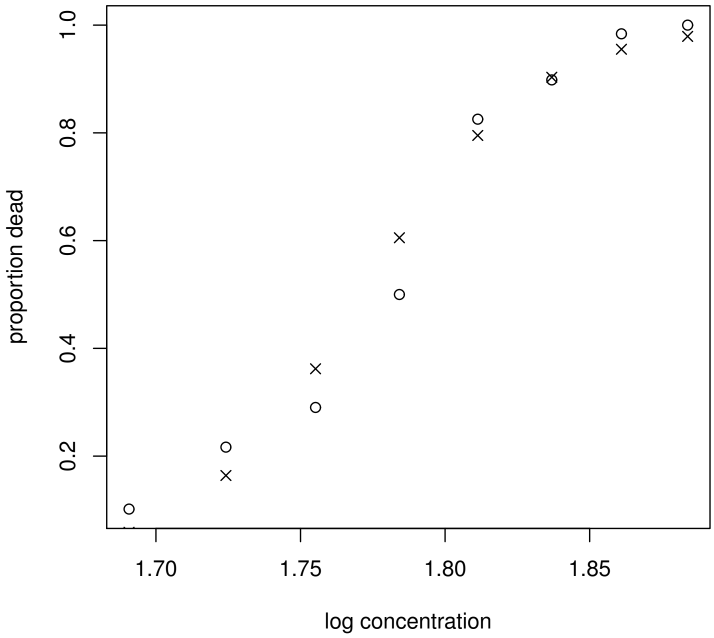

Figure 2.2: Observed (circles) and fitted (crosses) values for the beetle data

Figure 2.2 shows the fitted values along with the observed values. Circles are observed, crosses are fitted. The fit looks to be reasonably good.

2.3.1 Calculating fitted values

To further clarify what is being fitted, consider the 4th observation. The observed proportion of dead beetles is 0.5 (look at the data frame shown above). The fitted value is 0.604 to 3 d.p. How is this value obtained? To answer this we need the relationship stated above $E(Y_i) = \mu_i = h(\beta_0 + \beta_1 x_i)$. Remember that our response is $Y_i = D_i/n_i$ where $D_i \sim Bin(n_i, \mu_i)$ and $E(Y_i) = \mu_i$. $\hat{\mu}_i$ (the fitted value) is $h(\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$. Remember that h is the inverse link function (in this case we used the probit inverse link, $h = \Phi$) so

$$\hat{\mu}_i = h(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \Phi(\widehat{\beta}_0 + \widehat{\beta}_1 x_i) = \Phi(-34.935 + 19.728 x_i).$$

So

$\hat{\mu}_4 = \Phi(34.935+19.728x_4) = \Phi(-34.935+19.728\times1.7842) = \Phi(0.2637) = 0.604$

confirming the fitted value that `R` produces. Note that the Φ function is `pnorm` in `R`:

```
> pnorm(-34.935 + 19.728*1.7842)
[1] 0.6039935
```

Key Points:

- We use `R` to calculate the parameter estimates.
- We use the parameter estimates to calculate the value of the linear predictor for an observation.
- In linear models this is the expected (fitted) value.
- In a GLM, we apply the inverse link function to the linear predictor to get the fitted value. So we have an extra stage (applying the inverse link function) in the calculation for the fitted values in GLMs compared with linear models.

Binomial modeling like this is discussed further in Chapter 4.

Exercise:

- For the beetle mortality data, fit a binomial GLM with the logit link function;
- Using `R`, find the fitted values;
- Using the logit link, verify the result for the 6th fitted value.
- The complementary log-log link is a link function for a response that is a proportion. The link function g is $g(\mu_i) = log(-log(1-\mu_i)) = \eta_i$
- Derive h, the inverse link function
- This link is fitted to the beetle data with the following output

```
> cloglog<-glm(propn.dead~conc,binomial(cloglog),
+ weights=number,data=beetle)
> cloglog
Coefficients:
(Intercept)         conc
     -39.57        22.04
```

- Show that the fitted value for a log concentration of 1.7242 is 0.188

Chapter 3

Generalized linear models (GLMs): basic facts

3.1 Introduction

In this chapter, the general theory of the models and tests is given. Subsequent chapters show how the models can be used in particular areas – for proportions in Chapters 2 and 4 (logistic regression), non-negative data in Chapter 5 (Poisson regression), and contingency tables in Chapters 7. It will probably be best to read this chapter in conjunction with Chapters 2, 4 and 5 rather than try to understand everything here on a first pass. At various points in this chapter you are referred to related examples in later chapters. You are strongly advised to look carefully at these examples.

3.1.1 Standard linear model assumptions

The standard Normal-theory linear model is usually expressed as $Y \sim N(X\beta, \sigma^2 I)$. However, for developments here, we separate the modelling of the mean from assumptions about the distribution of Y. To this end, let $E(Y_i) = \mu_i$ and let x_i be the vector of explanatory variables for observation i, i.e. x_i^T is the i^{th} row of the design matrix X. Then the model can be written as:

(i-G) $E(Y_i) = \mu_i = x_i^T\beta = \sum_j \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \dots$, for the ith observation
(thus the mean is linear in β);

(ii-G) $\text{var}(Y_i) = \sigma^2$ (the variance is constant);

(iii-G) the Y_i are independent and Normal.

3.1.2 Generalized linear model assumptions

For the generalized linear model (GLM) the assumptions above are generalized as follows.

(i-G) An assumption about the **mean**.
Let $\eta_i = x_i^T\beta$ for some (vector) parameter β; η is called the **linear predictor**.
Let g be a known monotonic function and let h be its inverse, neither depending on i, and suppose that the mean $\mu_i = E(Y_i)$ is represented as

$$\mu_i = h(\eta_i) = h\left(x_i^T\beta\right), \quad \text{and so} \;\; g(\mu_i) = \eta_i = x_i^T\beta.$$

Here g is called the **link (function)**, and h the **inverse link**. Evidently g links the mean of the response to the linear predictor and hence the explanatory variables.
(This provides a particular class of means that are non-linear in the parameters – except, of course, when g is the identity)

(ii-G) An assumption about the **variance**:

$$\text{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i),$$

where w_i are known **weights**, the (scalar) parameter ϕ is called the **dispersion parameter** or **scale parameter** and the function V is called the **variance function**. ϕ will be different depending on the distribution used for the response Y_i. It is known for the Binomial and Poisson cases but has to estimated for the Normal and Gamma cases. However, because it directly affects the variance of the response Y_i, it can, even for the Binomial and Poisson distributions be allowed to be a free parameter that has to be estimated. This allows for what is called over-dispersion and is discussed further in section 3.9. Note that V is just a function of the mean that occurs in the variance; it is not itself the variance of anything. Different forms of function V are used for different distributions.
Key point: The response is now allowed to have a non-constant variance. The form of this non-constant variance depends on the distribution assumed for the response. Note that the variance does not depend on the choice of link function.

(iii-G) Assumptions about the distribution.
The Y_i are independent.

The pdf (probability 'density' function) for Y_i has a special form (explored and explained in §3.2). The density (with variable y and parameters θ_i and ϕ) can be written as:

$$f_i(y : \theta_i, \phi) = \exp\left\{ w_i \frac{y\theta_i - b(\theta_i)}{\phi} + c(y, \phi) \right\}. \tag{3.1}$$

Here 'density' is in quotes because some examples are discrete and then 'density' should be interpreted as 'probability function'.

3.2 Distributional properties in GLMs

Throughout this section (i.e. §3.2) Y is a scalar, not a vector.

3.2.1 Allowed distributions in GLMs

For GLMs, the pdf (in scalar y) of a single observation has the form

$$f(y : \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi/w} + c(y, \phi) \right\}, \tag{3.2}$$

where both θ and ϕ are parameters. This general form involves parameters θ and ϕ, which are related to the mean and variance of Y. We see how next, in §3.2.2 and §3.2.3; the result is that

$$E(Y) = b'(\theta) \quad \text{and} \quad \text{var}(Y) = \frac{\phi}{w} b''(\theta).$$

Various common and important distributions are included in the family (3.2): Normal, Poisson, Binomial and Gamma, with the appropriate correspondences recorded in §3.2.4.

In the usual formulation where there are several independent observations $Y_i, i = 1, 2, \ldots$, the values of θ and w are allowed to depend on i, giving θ_i and w_i, while the value of ϕ and the functions b, c are assumed not to depend on i. Thus Y_i is taken from the density

$$\exp\left\{ w_i \frac{y\theta_i - b(\theta_i)}{\phi} + c(y, \phi) \right\}, \tag{3.3}$$

as given in equation (3.1).

When ϕ is known, so that there is only one parameter, θ, (3.2) has the form of equation (3.5) below with $q = 1$ and so is from the exponential family, but it may not be from the exponential family when ϕ is unknown.
Unfortunately, there is no standard notation for the density in equation (3.2).

Background on exponential families

The **exponential family** is a class of distributions useful for modelling many kinds of data and whose members enjoy particularly tractable properties. The family includes the Normal, Binomial, Poisson, Gamma and Multinomial distributions.

A distribution with a one-dimensional parameter θ belongs to an exponential family if its pdf is of the form

$$f(y:\theta) = \exp(a(y)b(\theta) + c(y) + d(\theta)) \tag{3.4}$$

for suitably smooth functions $a - d$.
More generally for a vector parameter $\boldsymbol{\theta}$ a distribution belongs to a q-parameter exponential family if its pdf is of the form

$$f(y:\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{q} a_j(y)b_j(\boldsymbol{\theta}) + c(y) + d(\boldsymbol{\theta})\right\} \tag{3.5}$$

for $q \geq 1$ and suitable functions a_j, b_j, c and d.

3.2.2 Background on the Score Statistic

Let $l(\theta, \psi, y)$ $(= \log f(y:\theta,\psi))$ be the log-likelihood corresponding to a pdf f (not necessarily of the form (3.2); the results below are more general). The **score statistic** is defined as:

$$\frac{\partial l(y, \theta, \psi)}{\partial \theta}. \tag{3.6}$$

Result. For any pdf (provided it is regular in θ in a sense indicated in the proof) the score statistic, (3.6), has mean 0 and variance

$$-E\left(\frac{\partial^2 l}{\partial \theta^2}\right).$$

when θ, ψ are the parameter values in the distribution from which Y was generated.

Proof. If $f = f(y:\theta,\psi)$ is any pdf (which depends on parameters θ, ψ), then

$$\int f(y:\theta,\psi)dy = 1,$$

so that, differentiating both sides, and assuming that everything is sufficiently well-behaved that the differentiation and integration can be interchanged (which is the regularity needed)

$$0 = \frac{\partial 1}{\partial \theta} = \frac{\partial}{\partial \theta}\int f dy = \int \frac{\partial f}{\partial \theta} dy = \int \frac{\partial f(y,\theta,\psi)}{\partial \theta} dy.$$

This can be rewritten as

$$0 = \int \frac{1}{f}\frac{\partial f}{\partial \theta} f dy = \int \frac{\partial \log f}{\partial \theta} f dy = \int \frac{\partial l}{\partial \theta} f dy,$$

which is exactly

$$0 = E\left(\frac{\partial l}{\partial \theta}\right),$$

(i.e. the expected value of the score statistic is 0).

Differentiating again, assuming that differentiation and integration can be interchanged, and using $\partial f/\partial\theta = f\partial l/\partial\theta$,

$$\begin{aligned}
0 = \frac{\partial}{\partial \theta}\left(\int \frac{\partial l}{\partial \theta} f dy\right) = \int \frac{\partial}{\partial \theta}\left(\frac{\partial l}{\partial \theta} f\right) dy &= \int \left(\frac{\partial^2 l}{\partial \theta^2} f + \frac{\partial l}{\partial \theta}\frac{\partial f}{\partial \theta}\right) dy \\
&= \int \left(\frac{\partial^2 l}{\partial \theta^2} f + \frac{\partial l}{\partial \theta}\frac{\partial l}{\partial \theta} f\right) dy \\
&= E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\left(\frac{\partial l}{\partial \theta}\right)^2\right).
\end{aligned}$$

The first part of the result established that $\partial l/\partial\theta$ has mean zero, hence the second term here must therefore be its variance and so, because the two terms sum to zero, the variance is also given by minus the first term, which is the required result. (Also, the second term, and hence minus the first term, is the expected **information** on θ.

The same proof works for multivariate Y and vector θ.

3.2.3 Mean and variance for GLM distributions

Now we apply the results on the score statistic to a single observation from the density given by equation (3.2), for which

$$l(\theta, \phi, y) = \frac{y\theta - b(\theta)}{\phi/w} + c(y, \phi). \tag{3.7}$$

Differentiating $l(\theta, \phi, y)$,

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{\phi/w} \quad \text{and} \quad \frac{\partial^2 l}{\partial \theta^2} = \frac{-b''(\theta)}{\phi/w}.$$

The first of these is the score statistic, which has mean zero, and so

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = E\left(\frac{y - b'(\theta)}{\phi/w}\right),$$

which implies

$$\mu = E(Y) = b'(\theta). \tag{3.8}$$

This is a simple formula for the mean, μ, making μ and θ functions of each other. However, the form of the density in (3.7) would lose much of its simplicity if written in terms of μ instead of θ.

Note that $\partial^2 l/\partial\theta^2$ does not depend on Y and so is unchanged by taking its expectation (over Y). Now the result on the variance of score statistic gives, after a little calculation,

$$\mathrm{var}(Y) = \frac{\phi}{w} b''(\theta). \tag{3.9}$$

This shows that var(Y) is a product of a term depending on θ and another depending on ϕ. In the language of (ii-G) in §3.1.2, $b''(\theta)$, when written in terms of μ, gives the variance function, $V(\mu)$.

Task 1 *Verify formula (3.9) for var(Y).*

3.2.4 Common GLM distributions

The table below summarizes characteristics of the most important GLM distributions.

	ϕ	w	$b(\theta)$	$c(y,\phi)$	$\mu = b'(\theta)$	$b''(\theta)$
Normal $Y \sim N(\theta,\phi)$	ϕ	1	$\theta^2/2$	$-(y^2/\phi + \log(2\pi\phi))/2$	θ	1
Poisson $Y \sim Po(e^\theta)$	1	1	e^θ	$-\log(y!)$	e^θ	μ
Binomial: $nY \sim Bi(n, e^\theta/(1+e^\theta))$	1	n	$\log(1+e^\theta)$	$\log\binom{n}{ny}$	$e^\theta/(1+e^\theta)$	$\mu(1-\mu)$
Gamma $Y \sim Ga(\nu,\lambda)$ †	ϕ	1	$-\log(-\theta)$	$\nu\log\nu + (\nu-1)\log y$ $-\log\Gamma(\nu)$	$-1/\theta$	μ^2

†pdf $f(y) = \lambda^\nu y^{\nu-1} e^{-\lambda y}/\Gamma(\nu)$ where $\theta = -\lambda/\nu$ and $\phi = 1/\nu$.

As illustrated in the beetles example (see pp. 11-12, §2.1.2), for the Binomial, $Y_i = S_i/n_i$, the observed proportion of successes, is taken as the response variable.

Task 2 *Use equations (3.8) and (3.9) to verify the entries in the table above.*

3.2.5 The Canonical Link

In a GLM, by (iii-G) in §3.1.2, Y_i has the density function

$$f(y:\theta_i,\phi)=\exp\left\{\frac{y\theta_i-b(\theta_i)}{\phi/w_i}+c(y,\phi)\right\},$$

where $E(Y_i)=\mu_i=h(x_i^T\beta)$. We have also shown in equation (3.8) that $E(Y_i)=b'(\theta_i)$, and so, equating the two expressions for $E(Y_i)$, $b'(\theta_i)=h(x_i^T\beta)$, i.e.

$$x_i^T\beta=h^{-1}b'(\theta_i)=gb'(\theta_i) \tag{3.10}$$

The link function g for which $\theta_i=x_i^T\beta$ (that is, for which $g=(b')^{-1}$) is called the **canonical link**. In the Binomial case, the canonical inverse link is $h=b'(\theta_i)=e^{\theta_i}/(1+e^{\theta_i})$. Rearranging yields the canonical link function (g) as

$$g(\mu_i)=\log(\mu_i/(1-\mu_i))$$

commonly called the `logit` link. For the Normal, Poisson and Gamma, the canonical links are the identity, log and reciprocal, respectively.

Task 3 *Verify the canonical links for the Normal, Binomial, Poisson and Gamma.*

3.3 Estimation of parameters - general theory

In a linear model, parameter estimation is easy. If X is the design matrix, y the vector of response values, and σ^2 the variance of the error term, then the least squares estimator of β is

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

and the covariance matrix is

$$\sigma(X^TX)^{-1}.$$

These are easily calculated and only require specification of X and y. Things get more complicated in a generalised linear model. In a GLM the log-likelihood $l=l(\theta,\phi,y)$ is given by

$$l=\sum_i\left\{\frac{y_i\theta_i-b(\theta_i)}{\phi/w_i}+c(y_i,\phi)\right\}, \tag{3.11}$$

where the θ_i are functions of the unknown β, see equation (3.10), and the other unknown parameter is ϕ. Hence, for maximum likelihood (ML) estimation, one equation is $\partial l/\partial\beta = 0$, and the other is $\partial l/\partial\phi = 0$. Because neither ϕ nor c depends on θ or, as a consequence of this, on β,

$$\frac{\partial l}{\partial \beta} = 0 \quad \Longrightarrow \quad \frac{\partial}{\partial \beta}\sum_i \{w_i\,[y_i\theta_i - b(\theta_i)]\} = 0.$$

These equations involve β only, not ϕ. They are solved to provide an estimate of β, which will be independent of ϕ or the estimate of ϕ when it is unknown. When ϕ is unknown it is usually estimated by a moment method rather than by maximum likelihood — see §3.8.

In virtually all cases except the Normal linear model the equation for estimating β is non-linear, and has to be solved numerically, by iteration; the most popular method is a version of **Iteratively Reweighted Least Squares** (weighted least squares in which the weights depend on β, and hence, along with β, are iteratively estimated). The non-linear part in $\partial l/\partial\beta$ is $b'(\theta_i)$ (see section 3.2.4) which is usually non-linear in θ_i.

3.3.1 Covariance matrix for the MLEs

When ϕ is known, the asymptotic variance-covariance matrix for $\mathrm{var}(\widehat{\beta})$ can be obtained as the inverse of the information matrix on β; note that the information matrix is also called the Hessian matrix. Here the observed information matrix has (j,k)th entry

$$\begin{aligned} -\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} &= -\frac{\partial^2}{\partial\beta_j\partial\beta_k}\left(\sum_i\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i}\right\}\right) \\ &= -\frac{1}{\phi}\frac{\partial^2}{\partial\beta_j\partial\beta_k}\left(\sum_i \{w_i\,[y_i\theta_i - b(\theta_i)]\}\right). \end{aligned}$$

The same formulae are usually used when ϕ is unknown, but now with ϕ replaced by its estimator. Estimation of ϕ is discussed briefly in §3.8.

Task 4 *For the canonical link (see, in particular, equation (3.10)) show that*

$$\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} = \frac{1}{\phi}\sum_i \left\{-x_{ij}w_i h'(x_i^T\beta)x_{ik}\right\}.$$

Standardizing (subtract the hypothesized mean and divide by the e.s.e.) gives the **Wald test statistic**, which allows simple tests on a parameter or on a linear combination of parameters $c^T\beta$ (using $\mathrm{var}(c^T\widehat{\beta}) = c^T[\mathrm{var}(\widehat{\beta})]c$).

The usual Wald test that $\beta_j = 0$ uses $t = \widehat{\beta}_j/\text{e.s.e.}(\widehat{\beta}_j)$. The asymptotic theory leads to a $N(0,1)$ distribution, but using a t_{n-p} distribution may be better (the residual degrees of freedom are $n-p$ if the regression matrix X, with i^{th} row x_i, has n rows and is of rank p). For an illustration see Example 4.3.1(10).

3.4 Estimation of parameters - Iteratively reweighted least squares

The objective is to obtain estimates of the parameters in the GLM along with the covariance matrix. The process, known as iteratively reweighted least squares proceeds as follows:

1 Specify an initial vector of parameters $b^{(1)} = (\beta_0, \beta_1, ..., \beta_p)^T$;

2 Specify a weight matrix W that depends on the current parameter estimates;

3 Specify a vector z that depends on the current parameter estimates and response values;

4 Calculate recursively a new vector of parameter estimates $b^{(m)}$ using the relationship $b^{(m)} = (X^TWX)^{-1}X^TWz$ where W and z are calculated using $b^{(m-1)}$, the previous iteration values of $(\beta_0, \beta_1, ..., \beta_p)^T$;

5 Continue until the parameters value converge within a given tolerance.

The matrix W is a diagonal matrix with diagonal elements w_{ii} given by

$$w_{ii} = \frac{1}{\text{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

and the vector z has elements given by

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$$

Remember that η_i is the linear predictor. We illustrate this by applying it to the beetle data with a binomial distribution for the response and a logit link function. We first do the analysis in R to see what we should be getting.

```
second.beetle.glm=glm(propn.dead˜conc,family=binomial(link=logit),
weights=number,data=beetle)
```

```
> summary(second.beetle.glm)

Call:
glm(formula = propn.dead ~ conc, family = binomial(link = logit),
    data = beetle, weights = number)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
conc          34.270      2.912   11.77   <2e-16 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

The covariance matrix is

```
> vcov(second.beetle.glm)
            (Intercept)       conc
(Intercept)    26.83966 -15.082090
conc          -15.08209   8.480525
```

So for the linear predictor $\beta_0 + \beta_1 x_i$, the estimated value of β_0 is -60.72 with standard error $\sqrt{(26.84)} = 5.18$ and the estimated value of β_1 is 34.27 with standard error $\sqrt{(8.481)} = 2.91$; the covariance between the estimates is -15.082 to 3d.p. The last line of the `summary` command gives the number of iterations used before the estimates converged between iterations (4 in this example). We will show how to verify the parameter estimates and standard errors using iteratively reweighted least squares. For the logit link, we know that the relationship between $E(Y_i) = \mu_i$ and the linear predictor η_i is $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ so

$$\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = \frac{(1+e^{\eta_i})(e^{\eta_i}) - (e^{\eta_i})^2}{(1+e^{\eta_i})^2} = \frac{e^{\eta_i}}{(1+e^{\eta_i})^2}$$

by the quotient rule. We know that var$(Y_i) = \frac{\mu_i(1-\mu_i)}{n_i} = \frac{e^{\eta_i}}{n_i(1+e^{\eta_i})^2}$ so that applying

$$w_{ii} = \frac{1}{\text{var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

yields

$$w_{ii} = \frac{n_i e^{\eta_i}}{(1+e^{\eta_i})^2}$$

These are the diagonal elements in the weight matrix (where $\eta_i = x_i^T \beta$). For the z_i we have that $\eta_i = log\left(\frac{\mu_i}{1-\mu_i}\right)$ so that

$$\left(\frac{\partial \eta_i}{\partial \mu_i} \right) = \left(\frac{1-\mu_i}{\mu_i} \right) \left(\frac{1-\mu_i+\mu_i}{(1-\mu_i)^2} \right) = \frac{1}{\mu_i(1-\mu_i)}$$

by the chain rule. So applying $z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$ yields

$$z_i = \eta_i + \frac{(y_i - \mu_i)}{\mu_i(1-\mu_i)}$$

where $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$. Now that we have calculated the diagonal elements of the matrix W and the vector z we can use the iterative procedure to estimate the iterative reweighted least squares estimates of the parameters.

Taking initial values of β_0 and β_1 as -60 and 35, we obtain

$$w_{11} = \frac{59 \times e^{(-60+35\times1.6907)}}{\left(1+e^{(-60+35\times1.6907)}\right)^2} = 12.497$$

If we repeat this for the other seven weights we get the weight matrix

$$W = \begin{pmatrix} 12.497 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 14.557 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9.648 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.106 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.977 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.786 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.361 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.158 \end{pmatrix}$$

We next calculate z. This is easier if we calculate the expected values (μ values) first.

$$\mu_1 = \frac{e^{\eta_1}}{1+e^{\eta_1}} = \frac{e^{(-60+35\times1.6907)}}{\left(1+e^{(-60+35\times1.6907)}\right)} = 0.3045974$$

Then

$$z_1 = \eta_1 + \frac{(y_1 - \mu_1)}{\mu_1(1-\mu_1)} = -60 + 35 \times 1.6907 + \frac{(0.1017 - 0.3046)}{0.3046 \times (1 - 0.3046)} = -1.7834$$

It we repeat for the other z values we get

$$z = \begin{pmatrix} -1.783 \\ -1.175 \\ -1.889 \\ -3.287 \\ -1.134 \\ -2.331 \\ 3.369 \\ 6.939 \end{pmatrix}$$

With the design matrix X given by

$$X = \begin{pmatrix} 1 & 1.6907 \\ 1 & 1.7242 \\ 1 & 1.7552 \\ 1 & 1.7842 \\ 1 & 1.8113 \\ 1 & 1.8369 \\ 1 & 1.8610 \\ 1 & 1.8839 \end{pmatrix}$$

we get that

$$(X^T W X) = \begin{pmatrix} 44.08992 & 76.4819 \\ 76.48190 & 132.7406 \end{pmatrix}$$

and

$$X^T W z = \begin{pmatrix} -72.87857 \\ -126.34615 \end{pmatrix}$$

so that the next estimates of the parameters are

$$(X^T W X)^{-1} X^T W z = \begin{pmatrix} 44.08992 & 76.4819 \\ 76.48190 & 132.7406 \end{pmatrix}^{-1} \begin{pmatrix} -72.87857 \\ -126.34615 \end{pmatrix} = \begin{pmatrix} -3.535201 \\ 1.085069 \end{pmatrix}$$

If we let $b^{(j)}$ be the estimated parameter vector at iteration j, then our initial values give $b^{(1)} = (-60, 35)^T$ and the next iteration provides $b^{(2)} = (-3.535, 1.085)^T$. If we continue doing this we find that

$b^{(1)} = (-60, 35)^T$
$b^{(2)} = (-3.535, 1.085)^T$

$$
\begin{aligned}
b^{(3)} &= (-63.17875, 36.00085)^T \\
b^{(4)} &= (-55.9871, 31.56474)^T \\
\vdots \quad & \quad \vdots \\
b^{(7)} &= (-60.71745, 34.27033)^T \\
b^{(8)} &= (-60.71745, 34.27033)^T
\end{aligned}
$$

so that we get convergence after 7 iterations. `R` finds a smarter way to do it in just 4 iterations. Our parameter estimates obtained by iteratively reweighted least squares are $\hat{\beta}_0 = -60.71$ and $\hat{\beta}_1 = 34.27$ in agreement with the `R` output above.

Using the converged estimates of the parameters, the covariance matrix is given by

$$(X^T W X)^{-1} = \begin{pmatrix} 26.840 & -15.082 \\ -15.082 & 8.481 \end{pmatrix}$$

again in agreement with the `vcov` output in `R`. Note that these estimates are maximum likelihood estimates. Standard theory states that the sampling distribution of the maximum likelihodd estimates is asymptotically normal so we can find confidence intervals for the parameters in the usual way. For example, for the beetle data using the logit link, a 95% confidence interval for β_1 is $34.27 \pm 1.96\sqrt{8.48} = (28.6, 40.0)$. Note that because this is a numerical maximization technique, care must be taken to avoid obtaining a local rather than a global mle.

3.5 Scaled deviance and (residual) deviance

3.5.1 Definitions

The model contains parameters θ_i, or alternatively μ_i, for each i, where $\mu_i = b'(\theta_i)$; but the important point is that the n values $\{\mu_i\}$ are expressed as functions of the parameter β with, say, p components, via the inverse link function — $\mu_i = h(\eta_i) = h(x_i^T \beta)$. The likelihood of the sample is a function of the parameters β and ϕ, but it may be thought of instead as a function of μ and ϕ, where μ is the vector of length n with components μ_i. The likelihood will be denoted by $L(\mu, \phi, y)$ and the log-likelihood by $l(\mu, \phi, y)$ — but the true parameter remains β, and the components of the vector μ remain constrained by the relationships $\mu_i = h(\eta_i) = h(x_i^T \beta)$.

The maximum value of L (given ϕ) is then $L(\widehat{\mu}, \phi, y)$, where $\widehat{\mu} = h(x_i^T \widehat{\beta})$ and derives from the ML estimate $\widehat{\beta}$ discussed in §3.3; the corresponding $\widehat{\theta}_i$ make up the vector $\widehat{\theta}$, resulting from equation (3.8).

The model that allows the mean of each observation to be a separate parameter is called the **full** or **saturated** or **maximal** model. Denote the means for this model by $\mu^\diamond$, with $\theta^\diamond$ the corresponding θ. Estimation is easy. The maximum possible value of L as the components of $\mu^\diamond$ vary without any restrictions (or, essentially equivalently, when there is an element of β for each observation, so that $\mu_i^\diamond = h(\eta_i) = h(\beta_i)$) occurs when $\mu_i^\diamond = y_i$, hence $\widehat{\mu_i^\diamond} = y_i$ for all i. Thus, for the saturated model the estimate of the means is $\widehat{\mu^\diamond} = y$, which is independent of the particular link being used. Denote the corresponding estimate of $\theta^\diamond$ by $\widehat{\theta^\diamond}$, given by $\widehat{\mu^\diamond} = b'(\widehat{\theta^\diamond})$.

Task 5 *Show that, for the maximal model, y_i is the maximum likelihood estimate of μ_i. (Hint: the likelihood equation for θ_i is $y_i = b'(\theta_i)$, and we already know that $b'(\theta_i) = \mu_i$.)*

Suppose a proposed model ($\mu_i = h(x_i^T\beta)$) has corresponding estimated means $\widehat{\mu}$ and associated $\widehat{\theta}$, linked via equation (3.8). The **scaled deviance** $S(y,\widehat{\mu})$ for that model is defined as twice the difference in the log-likelihoods of the maximal model and that model:

$$\begin{aligned}
S(y,\widehat{\mu}) &= -2\left\{ l(\widehat{\mu},\phi,y) - l(\widehat{\mu^\diamond},\phi,y)\right\} \\
&= 2\sum_i \frac{y_i(\widehat{\theta_i^\diamond} - \widehat{\theta}_i) - b(\widehat{\theta_i^\diamond}) + b(\widehat{\theta}_i)}{\phi/w_i} \\
&= \frac{2}{\phi}\sum_i w_i\left[y_i(\widehat{\theta_i^\diamond} - \widehat{\theta}_i) - b(\widehat{\theta_i^\diamond}) + b(\widehat{\theta}_i)\right].
\end{aligned}$$

Hence, $\phi S(y,\widehat{\mu})$ does not depend on ϕ and so is a function only of the data (including the w_i). Then the **deviance** (also called the **residual deviance**) for the model with estimated means $\widehat{\mu}$ is defined by

$$D(y,\widehat{\mu}) = \phi S(y,\widehat{\mu}) = 2\sum_i w_i\left[y_i(\widehat{\theta_i^\diamond} - \widehat{\theta}_i) - b(\widehat{\theta_i^\diamond}) + b(\widehat{\theta}_i)\right]. \tag{3.12}$$

Both S and D are non-negative, and the maximal model has $S = D = 0$.

In R output, D is called the residual deviance and deviance is used for the difference between the residual deviances of two models where one is nested in the other; see §3.6.2.

N.B. Some authors use G or G^2 for D. Unfortunately, Dobson uses the terms deviance and scaled deviance the opposite way round to that above. The definitions above are a bit different from those actually used in R. Whenever you see these terms used you will have to check out the author's preferred definition. Probably, in most other contexts, deviance is used for twice the difference in log-likelihood, which here is called the scaled deviance.

3.5.2 Testing model fit using deviance (approximately)

Under rather strong assumptions (and if the model is correct) $S(y, \widehat{\mu})$, has, asymptotically, a χ^2_{n-p} distribution, for reasons indicated in §3.6.1 – but this may not be a very good approximation. When ϕ is unknown – in particular for Normal and Gamma models – an estimate of ϕ is needed (see §3.8). See section 4.3.1 for an example.

3.5.3 Testing model fit using pseudo R^2

For linear models we can use R^2 as a measure of fit (with the caveat that it cannot decrease as more parameters are added). Remember that for a linear model R^2 is the proportion of the variation of the response that is explained by the model (the variation here being measured by the sums of squares of the response). In GLMs there is a version of this called `pseudo` R^2. The pseudo R^2 is given by

$$\frac{l(\tilde{\mu}, \phi; y) - l(\hat{\mu}, \phi; y)}{l(\tilde{\mu}, \phi; y)}$$

where $l(\tilde{\mu}, \phi; y)$ is the log-likelihood of the minimal model (in which the linear predictor contains just a constant term) and $l(\hat{\mu}, \phi; y)$ is the log-likelihood of the model under consideration. So the pseudo R^2 represents the proportional improvement in the log-likelihood due to the model under consideration compared to the minimal model.

`R` produces an AIC value for each model and we can use this to get the likelihood. Remember that

$$AIC = -2l + 2p$$

where l is the log-likelihood and p is the number of parameters. We can rearrange this to get $l = p - 1/2AIC$. We can therefore use the AIC output from `R` for the minimal model and any given model to estimate the pseudo R^2 value. See 4.3.3 for an example.

3.5.4 Deviances for common distributions

For the common four distributions we mainly use, the (residual) deviances, $D(y, \widehat{\mu})$, are as follows.

$$\begin{array}{rl}
\text{Normal} & \sum (y_i - \widehat{\mu}_i)^2 \\
\text{Poisson} & 2 \sum \left\{ y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i) \right\} \\
\text{Binomial} & -2 \sum_i n_i \left\{ y_i \log\left(\frac{\widehat{\mu}_i}{y_i}\right) + (1 - y_i) \log\left(\frac{1 - \widehat{\mu}_i}{1 - y_i}\right) \right\} \\
\text{Gamma} & -2 \sum \left\{ \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - \left(\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right) \right\}
\end{array}$$

For Poisson and Gamma, $\sum (y_i - \widehat{\mu}_i) = 0$ under a power or log link when the model has a constant term, and so the deviance usually simplifies accordingly.

Task 6 *Confirm these using the table in §3.2.4 and equation (3.12).*

In all four cases the form of $D(y, \widehat{\mu})$ supports its use as a measure of fit since it can be interpreted as summarising how near the fitted values $\widehat{\mu}$ are to the observed values y.

3.6 Comparing nested models

The **Pearson X^2 statistic** is often used as a measure of fit. It could therefore be used to compare the fit of two models but D is usually preferred when looking at changes between models as it is exactly additive when a hypothesis is partitioned into orthogonal components.

3.6.1 Applying GLRT to GLM

Suppose two models for the mean (i.e. linear predictors η^* and η°) are to be compared, and that in Model 1 $\eta^* = X\beta$, and in Model 2 $\eta^\circ = X\beta + Z\gamma$ — so that the first is a special case of the second (i.e is nested within the second). Then, when ϕ is known, η^* can be tested against η° (i.e. testing the null hypothesis $\gamma = 0$) by using the **Generalized Likelihood Ratio**

Test (GLRT), with the test statistic

$$\begin{aligned}-2\log(L^*/L^\circ) = -2(l^* - l^\circ) &= -2\left(l(\widehat{\mu^*}, \phi, y) - l(\widehat{\mu^\circ}, \phi, y)\right) \\ &= S(y, \widehat{\mu^*}) - S(y, \widehat{\mu^\circ}) \\ &= \frac{D(y, \widehat{\mu^*}) - D(y, \widehat{\mu^\circ})}{\phi}.\end{aligned}$$

Note that, if ϕ is unknown, the true Generalised Likelihood Ratio Test would use maximum likelihood estimates of ϕ under both Model 1 (in l^*) and Model 2 (in l°). Instead, in this context, it is common to compute $\widehat{\phi}$ from Model 2 and then treat it as known.

Under the null hypothesis that $\gamma = 0$ (i.e. that the model η^* is appropriate) and under suitable assumptions of regularity, the distribution of the test statistic

$$\frac{D(y, \widehat{\mu^*}) - D(y, \widehat{\mu^\circ})}{\phi}$$

(which is just the difference of scaled deviances) is asymptotically χ^2_r when γ has r linearly independent components (i.e. rank$(Z) = r$). If we are just testing whether adding an extra explanatory variable improves the fit then $r = 1$; this is the test we usually use.

The claim that $S(y, \widehat{\mu})$, has, asymptotically, a χ^2_{n-p} distribution, made in section 3.5.2, is just a particular case of this result, with Model 1 giving $\widehat{\mu}$, based on p parameters and Model 2 being the saturated model, with n parameters.

Of course all of this theory presumes that Model 2 and the associated distributional assumptions are suitable. Hence, before doing any testing based on the theory above one should first check that Model 2 is adequate with residual plots (§3.7) and any other devices that are appropriate.

3.6.2 Model building - Analysis of Deviance

On the basis of the theory just discussed it is clear that approximate tests of **nested** models can be made by using the deviance. Comparing nested GLMs using analysis of deviance is a very similar process to using F-tests for nested linear models. A convenient way of presenting these is in an Analysis of Deviance table, which is similar to an Analysis of Variance table, and reduces to it in the Normal linear case. The general structure splits the variation for the simpler model (model 1), measured by its deviance, into two components, one representing the improvement of the more complicated model (model 2) over the simpler one and the other representing the variation left after fitting the more complicated model.

Source	Deviance	d.f.
Model 2 after fitting Model 1	$D(y,\widehat{\mu^*}) - D(y,\widehat{\mu^\circ})$	$r = p^\circ - p^*$
Model 2	$D(y,\widehat{\mu^\circ})$	$n - p^\circ$
Model 1	$D(y,\widehat{\mu^*})$	$n - p^*$

where $\text{rank}([X,Z]) = p^\circ$, $\text{rank}(X) = p^*$, and so $\text{rank}(Z) = p^\circ - p^* = r$.

Typically an Analysis of Deviance involves a series of nested models, not just two. In the table this amounts to decomposing the Model 2 term further into 'Model 3' and 'Model 3 after fitting Model 2', in just the same way, and so on. Also, when extended tables of this form are constructed, the simplest model considered, and used as 'Model 1', is usually the one where all the means are assumed the same; this is called the **null** or **minimal** model and its deviance is called the **null deviance**. In the null model all the entries in η^* are the same; $\eta_i^* = \beta_0$ for all i. If $W = \sum w_i$, the estimate of μ in the minimal model is easily seen to be $W^{-1}\sum_i w_i y_i$, and does not depend on the link used.

For an example of using analysis of deviance in model building, see section 4.3.1.

Task 7 *Use the results in §3.5.4 to obtain simple expressions for the null deviances for Poisson (weights must be one) and Gamma (assume weights are one).*

You can see examples of these extended Analysis of Deviance Tables, in various formats, produced by packages. In R, $D(y,\widehat{\mu^\circ})$ is called a residual deviance and differences between these, $D(y,\widehat{\mu^*}) - D(y,\widehat{\mu^\circ})$, are called deviances.

3.6.3 Analysis of deviance in R

To compare two nested models in R use the `anova` command with the a `test="Chi"` option. So `anova(model1,model2,test="Chi")` where model1 is nested in model2, will produce an analysis of deviance table like that in section 3.6.2. See section 4.3.1 for an example.

3.6.4 hypothesis tests on a single parameter

The Wald $t-$test can sometimes be badly misleading. Hence, although comparing estimates with their estimated standard errors can be a useful guide in model simplification, the change in deviance test is preferred when actually performing a test on a single parameter.

3.6.5 Remember that p-values are only approximate

Note that, apart from the Normal linear model, even if all the assumptions hold precisely, the deviance tests and the Wald $t-$tests for $r = 1$ in §3.3 are based on asymptotic theory, and hence are approximate in practice. How close the null distribution is to the assumed one depends on many things, not just the size of n. Thus all the tests should be viewed as guidelines only. It is not realistic, for example with a 1 df χ^2-test ($\chi^2_{1,0.95} = 3.84$), to say that 3.85 shows a significant effect ($p < 0.05$), and 3.83 does not show one ($p > 0.05$); in both cases all you know is that $p \approx 0.05$.

3.7 Residuals

There are a number of different types of residuals which can be defined for generalized linear models; in the Normal linear model they all reduce to variants of the basic (raw) least squares residual $e_i = y_i - \widehat{\mu_i}$. Here two are defined. It is not clear which type of residual is more useful in different situations and for different purposes, and the default may differ in different packages.

3.7.1 Distribution of residuals

Clearly when the response variable is not normally distributed, we wouldn't expect the residuals to be normally distributed - but in some cases they might be, although often only very approximately. Consider the case of a response that has a binomial distribution. We would expect the errors to be binomially distributed - but we know that provided n is large and p not too small, that a $bin(n, p)$ distribution is not too different from a $N(np, np(1-p))$ distribution (using the Central Limit Theorem). So for GLMs, it might at first seem reasonable to assume that, providing certain conditions are met, we might expect the residuals to be roughly normally distributed. But there is also the added problem that we have to estimate the expectation. Bearing all this in mind we shouldn't put too much emphasis on a QQ plot or histogram. These are often omitted from the list of diagnostic checks. As we do for linear models, we could perhaps check for autocorrelation (correlation between successive observations), outliers, non-linearity and homoscedasticity (constant variance). It may at first seem counter-intuitive to expect homoscedasticity for Pearson residuals in a GLM as by definition the variance of the response is allowed to depend on the expected value. However the Pearson residuals are scaled by the variance so the variance should be approximately constant for all values (the variance of Y_i is dictated by the assumed distribution of Y_i) . Also, be aware that for a Poisson distribution when counts are small, there may well be patterns in the residuals resulting from this aspect of the data, which should not be construed as indicating model inadequacy.

3.7.2 Pearson residuals

$$e_{P,i} = \sqrt{w_i}\frac{y_i - \widehat{\mu}_i}{\sqrt{V\left(\widehat{\mu}_i\right)}}$$

Note that $y_i - \widehat{\mu}_i$ has been divided by the estimate of var$(Y_i)/\phi$, not by its estimated scaled variance.

The **Pearson X^2 statistic** (discussed as a possible measure of fit was briefly discussed in section 3.5.2). It is related to the Pearson residuals by the relationship

$$X^2 = X^2(y, \widehat{\mu}) = \sum e^2_{P,i},$$

Note that **Pearson X^2 statistic** is asymptotically equivalent to the deviance for the model (D).

Task 8 *For the Poisson distribution, verify that D and X^2 are asymptotically equivalent. (Hint: use that for small $|x - y|$, $x\log(x/y) \approx (x - y) + (x - y)^2/(2y)$.)*

Note that $N(\mu_i, V(\mu_i)/w_i)$ is the usual large n approximate distribution for Y_i under the Binomial and Poisson (where $\phi = 1$). Estimating μ_i and $V(\mu_i)$, and standardizing, gives $e_{P,i}$ for these two cases. Even with this large n approximation we see that the Pearson residuals will at best have only an approximate $N(0, 1)$ distribution (if n is large) since we have to estimate μ_i and $V(\mu_i)$.

Task 9 *Obtain the form of the Pearson residual for the Binomial, Poisson and Gamma. (These are given in Chapters 4 and 5)*

3.7.3 Deviance residuals

The deviance arises from the difference in log-likelihoods. In a log-likelihood for independent observations each observation causes the addition of a non-negative term. Hence the deviance has the form $D = \sum d_i$, where d_i arises from the ith observation. In section 3.5.4 the residual deviances for some common distributions are defined. The d_i defined in this section is the contribution of the ith observation to this residual deviance. So referring to section 3.5.4, if Y_i has a binomial distribution then

$$d_i = -2 \times n_i \left\{ y_i \log \left(\frac{\widehat{\mu}_i}{y_i} \right) + (1 - y_i) \log \left(\frac{1 - \widehat{\mu}_i}{1 - y_i} \right) \right\}$$

the above form for d_i is only true for the binomial distribution. For other distributions you need to consult section 3.5.4 for the specific form.

The ith deviance residual is defined in terms of d_i as

$$e_{D,i} = \operatorname{sgn}(y_i - \widehat{\mu}_i) \times \sqrt{d_i},$$

where $\operatorname{sgn}(x) = 1$ if $x > 0, -1$ if $x < 0$. Note that $\sum e_{D,i}^2 = D$.

Note: If ϕ is not 1, the Pearson or deviance residuals may be scaled by $\sqrt{\phi}$ or its estimate to give **scaled Pearson residuals** or **scaled deviance residuals**.

See section 4.3.4 for an example of how to calculate Pearson and deviance residuals for a binomial response.

3.8 Estimating the scale parameter

A common ('moment') estimate of the scale parameter ϕ is:

$$\frac{X^2(y, \widehat{\mu})}{n - p} \quad \text{or} \quad \frac{D(y, \widehat{\mu})}{n - p},$$

where n is the number of observations and p is the number of parameters (including the intercept) in the model. The estimator of ϕ used in `R` is the X^2 one. Note that in the binomial case, n is the number of proportions observed, not the total number of observations these proportions are based on; see 4.4.1 for an example. This is not a maximum likelihood estimate, but then neither is the usual estimate of the error variance in Normal linear model theory.

3.9 Quasi-likelihood

If a GLM has ϕ fixed, as for the Binomial and Poisson (which have $\phi = 1$), a simple extension, which may be appropriate for some data sets, is to assume the variance has the form for that distribution but scaled by ϕ with ϕ unknown. The model can be fitted in the usual way for the originating distribution, and then ϕ can be estimated (see §3.8).

For the Binomial and Poisson, this allows a general model for **over-dispersion** (or conceivably under-dispersion) to be fitted. These models can also be used for data restricted to $[0, 1]$ (Binomial) or $[0, \infty)$ (Poisson) even if the underlying variable is not integer valued.

Since the only assumptions are about the mean $E(Y)$ and the variance $\text{var}(Y)$, the estimation cannot be a maximum likelihood procedure, and the generalized likelihood ratio test theory does not apply. However, an extension, called **quasi-likelihood** theory, shows that asymptotically, the generalized likelihood ratio test described in §3.6.1 applies.

An illustration is given in Example 4.4.1.

Chapter 4

Logistic regression revisited

4.1 Introduction

Logistic regression was discussed in Chapter 2. In that chapter we motivated the logistic regression model and we saw how we can use the `glm` function in `R` to fit a GLM with binomial response to the beetle data (see the toxicity example on page 12). Then we went on to describe the theory of generalised linear models in Chapter 3 with the emphasis placed on model building and model performance using the Deviance. In this chapter we apply this theory of Chapter 3 to logistic regression, in order to further discuss model building and model testing for this important class of models. You should read this chapter in conjunction with Chapter 3 (for the theory) and Chapter 2 (for the basic model fitting for logistic regression).

4.2 Links for binomial responses

The standard logistic regression model considers that observations y_i are generated from a binomial distribution

$$Y_i \sim Bin(n_i, \mu_i),$$

with size n_i and probability μ_i. The link function $g(\cdot)$ maps μ_i to the linear predictor, or

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

There are three commonly used link functions (applied to μ_i, the expected proportion of successes) – the first two were noted on page 13.

$$\begin{array}{rl} \text{Logit} & \text{logit}(\mu_i) = \log\{\mu_i/(1-\mu_i)\} = x_i^T\beta \\ \text{Probit} & \text{probit}(\mu_i) = \Phi^{-1}(\mu_i) = x_i^T\beta \\ \text{Complementary log-log} & \text{cloglog}(\mu_i) = \log(-\log(1-\mu_i)) = x_i^T\beta \end{array}$$

Task 10 *How does a GLM with identity link relate to a linear model? Under which situations does it make sense to fit a linear model instead of a GLM with identity link?*

For a single explanatory variable x, a rough check on the link, g, and linearity of the predictor in x can be obtained by plotting $g(y_i^*)$ against x_i, where, to avoid problems if $y_i = 0$ or $y_i = 1$, $y_i^* = (n_i y_i + 1/2)/(n_i + 1) = (s_i + 1/2)/(n_i + 1)$. Using a less appropriate link can lead to needing extra polynomial terms in x to account for the extra curvilinearity – see §4.3.1.

4.2.1 Choosing a link function for the beetle data of chapter 2

The beetle data used extensively in chapter 3 (repeated in the table below) reports the number of beetles exposed (n_i) and killed (s_i) after 5 hours exposure to gaseous carbon disulphide (CS_2) at various concentrations (x_i in $\log_{10}CS_2$ mg/l).

Note the shape of the top left plot in Figure 4.1. It has the shape of a distribution function: tending to zero at $-\infty$ and tending to 1 at $+\infty$ which is a common shape for data that naturally give rise to proportions. How can we determine which link function to use to model the data? We can check that plotting $g(y_i)$ against x_i gives a straight line. Clearly, the form that $g(y_i)$ takes depends on the link function being considered. For the logit link we plot x_i against

$$log\left(\frac{y_i}{1-y_i}\right) = log\left(\frac{s_i/n_i}{1-s_i/n_i}\right) = log\left(\frac{s_i}{n_i - s_i}\right)$$

(or use $y_i^* = (n_i y_i + 1/2)/(n_i + 1) = (s_i + 1/2)/(n_i + 1)$ if $y_i = 0$ or $y_i = 1$) Note that $log\left(\frac{y_i}{1-y_i}\right)$ is termed the empirical logit. For the probit and cloglog links we plot x_i against $\phi^{-1}(y_i^*)$ and $log(-log(1-y_i^*))$ respectively (see Figure 4.1).All three plots look fairly linear though there is possibly some curvature in the logit plot.

More formally we could assess the fit using the scaled deviance (or residual deviance since $\phi = 1$ for the binomial). The scaled deviance for the beetle data (obtained from the `summary` command in `R` - see section 2.3 for the probit output), fitting an intercept and linear term, for the probit, logit and cloglog links are:

- probit link, $S = 10.12$ on 6 df $(n = 8, p = 2)$ (1-pchisq(10.12,6)=0.12)
- logit link, $S = 11.23$ on 6 df (1-pchisq(11.23,6)=0.08)
- cloglog link $S = 3.45$ on 6 df (1-pchisq(3.45,6)=0.751)

All links provide reasonable fits (at the 5% level) where the null is that the model is 'adequate'. The cloglog links has the lowest scaled deviance of the three but inspection of residuals should also be done before making any definitive conclusions.

x_i	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
n_i	59	60	62	56	63	59	62	60
s_i	6	13	18	28	52	53	61	60

4.3 Model building, assessing model fit and residuals

The deviance is given by (§3.5.4)

$$D = -2 \sum_i n_i \left\{ y_i \log \left(\frac{\widehat{\mu}_i}{y_i} \right) + (1 - y_i) \log \left(\frac{1 - \widehat{\mu}_i}{1 - y_i} \right) \right\}.$$

The null deviance (constant mean) does not depend on the link.

The Pearson residual (see §3.7.2) is

$$e_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i \left(1 - \widehat{\mu}_i\right) / n_i}}.$$

Some calculation shows that, with $s_i = n_i y_i$,

$$X^2 = \sum e_{P,i}^2 = \sum \left\{ \frac{[s_i - n_i \widehat{\mu}_i]^2}{n_i \widehat{\mu}_i} + \frac{[(n_i - s_i) - n_i(1 - \widehat{\mu}_i)]^2}{n_i(1 - \widehat{\mu}_i)} \right\}.$$

Thus, given the 'expected values', E (i.e. the estimated means), X^2 is the usual sum of $(O - E)^2/E$, with 'observed value', O, over the $2 \times n$ table with one row for 'success' and one for 'failure' — in the i^{th} row, the observed number of successes in s_i and the observed number of failures is $n_i - s_i$.

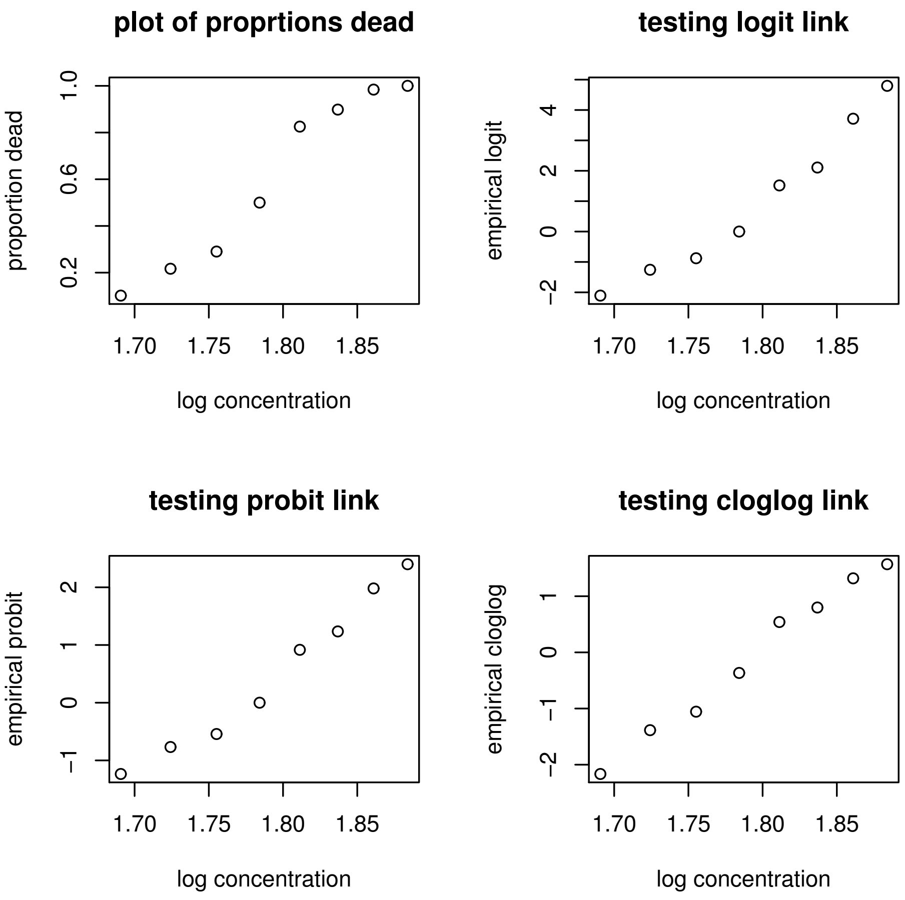

Figure 4.1: Proportion of beetle deaths by log concentration

Similarly, D is the sum over the $2 \times n$ table of $2 \times O \times \log(O/E)$.

As mentioned in §3.6, marginal Wald t-ratios on coefficients can be seriously misleading and this can happen for Binomial data. Thus the test based on the difference in (residual) deviance is preferred.

Note that to fit data as Binomial, it is necessary to specify both the response y_i and the weights n_i (or some other function of these).

Task 11 *Confirm that*

$$\sum e_{P,i}^2 = \sum \left\{ \frac{[s_i - n_i\widehat{\mu}_i]^2}{n_i\widehat{\mu}_i} + \frac{[(n_i - s_i) - n_i(1-\widehat{\mu}_i)]^2}{n_i(1-\widehat{\mu}_i)} \right\}.$$

Hint: use $\dfrac{1}{p(1-p)} = \dfrac{1}{p} + \dfrac{1}{1-p}$.

4.3.1 Model building for the beetle data

Consider again the beetle mortality data of Example 4.2.1. This data set will be used to demonstrate some of the model-fitting and testing theory of Chapter 3. In this example we use the logit link function. Using the other link functions appears as an exercise in Exercise 2.

1. **Fitting the minimal model**

```
null.glm<-glm(propn.dead~1,binomial(logit),weights=number)
```

The null deviance ($\eta = \beta_0$) is 284.2 (on 7 df). Looking at the left hand plot in Figure 4.2, the proportion of deaths clearly depends on log concentration.

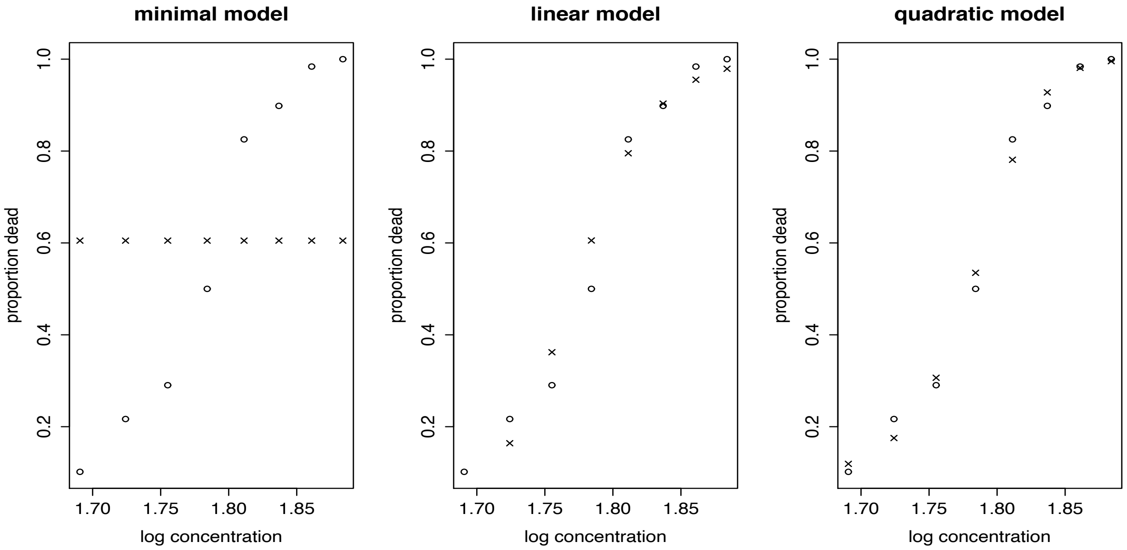

Figure 4.2: Fitted and observed values for various models applied to the beetle data: circles=observed values, crosses=fitted values

2. **Adding a log concentration term to the model**

```
linear.glm<-glm(propn.dead~1+conc,binomial(logit),weights=number)
```

$\eta = \beta_0 + \beta_1 x$ gives a (residual) deviance of 11.232 on 6 df. The Pearson X^2 value (see §3.7.2) is 10.026 (roughly comparable to the (residual) deviance). This is obtained from my by using the command

```
sum(resid(linear.glm,"pearson")^2)
```

3. The change in deviance from the null model to the linear model is $284.2 - 11.232 = 272.97$ on 1 df. $\chi^2_{1,0.95} = 3.84$ so overwhelming evidence to reject the null hypothesis $H_0 : \beta_1 = 0$. This is analysis of deviance for nested models described in section 3.6.2.

4. Since $\chi^2_{6,0.95} = 12.59$, the fit appears acceptable (11.232 is less than 12.59). This is using the (residual) deviance to assess model fit – see the end of §3.5.1.

   Looking at the middle plot in Figure 4.2, we can see considerable improvement in the fit compared to the minimal model.

5. The left hand figure in Figure 4.3 shows the Pearson residuals against fitted value for the model with a linear term for log concentration. It suggests considerable curvilinearity and it is sensible to consider adding a quadratic term into the linear predictor.

6. **Adding a quadratic log concentration term to the model**

   ```
   quad.glm<-glm(propn.dead~1+conc+I(conc^2),binomial(logit),weights=number)
   ```

   Fitting a quadratic in x ($\eta = \beta_0 + \beta_1 x + \beta_2 x^2$) gives a (residual) deviance of 3.195 on 5 df .

7. This again is a significant improvement since the change in deviance is 8.037 ($\chi^2_{1,0.95} = 3.84$, $\chi^2_{1,0.99} = 6.64$) so there is strong evidence to reject the null hypothesis $H_0 : \beta_2 = 0$. This is analysis of deviance for nested models described in section 3.6.2. Also since $\chi^2_{5,0.95} = 11.07$, the (residual) deviance is well below this so the fit is good. This is a further use of (residual) deviance to assess model fit - see the end of §3.5.1.

   The right hand plot in Figure 4.2 shows the fitted values for the quadratic model - showing a considerable improvement in the fitted values compared to the model with just a linear term. The right hand figure in Figure 4.3 shows the Pearson residuals versus fitted values for the quadratic model. The curvature in the residuals has gone.

8. Finally, the (residual) deviance is so low that there is now essentially no scope for further improvement by adding another parameter (the cubic term, for example): the (residual) deviance from the quadratic model is less than $\chi^2_{1,0.95} = 3.84$ so no change in (residual) deviance can be less than 3.84.

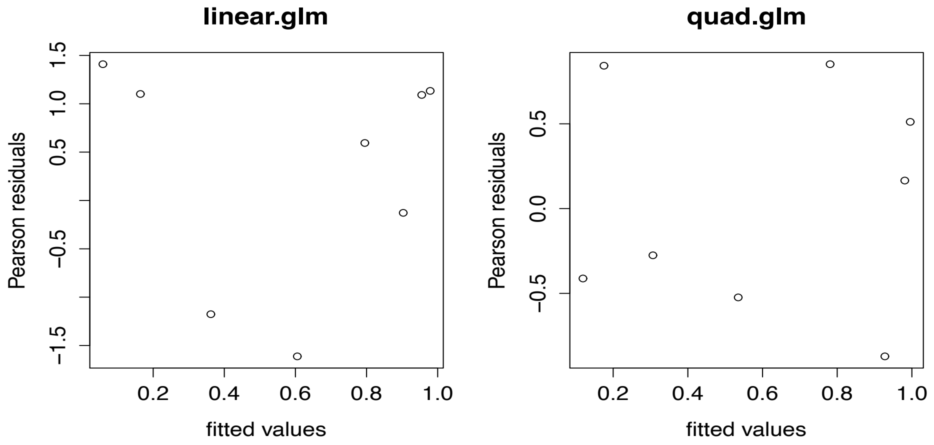

Figure 4.3: Pearson residuals versus fitted values for linear.glm and quad.glm

9. Hence, using the logit link we get a good fit - residual deviance is low and residual plots are satisfactory.
10. As an example of the Wald t-test (§3.3 and §3.6.4), consider the quadratic fit (using a logit link). Obtaining the estimates for the parameters in the linear predictor using the `summary` command we have that $\widehat{\beta_2} = 156.41$, with e.s.e.$(\widehat{\beta_2}) = 57.853$, so the t-value for testing $\beta_2 = 0$ is 2.703. This gives a p-value of 0.00687. This comes from $2 \times (1 - \Phi(2.703)) = 0.007$ using Neave's tables; or `2*pnorm(-2.703)` in `R`.

4.3.2 Analysis of deviance for the beetle data

As noted in section 3.6.2, we can perform an analysis of deviance in `R` using the `anova` command. Using `anova` on a single GLM gives the residual deviances for all the univariate models nested within the GLM under consideration. So for the quadratic model we would obtain the deviances for the null, linear and quadratic models:

```
> anova(quad.glm)
Analysis of Deviance Table
Model: binomial, link: logit
Response: propn.dead
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
```

```
NULL                          7    284.202
conc        1  272.970        6     11.232
I(conc^2)   1    8.037        5      3.195
```

We can also use the anova command to compare the residual deviance of two GLMs.

```
> anova(linear.glm,quad.glm,test="Chi")
Analysis of Deviance Table

Model 1: propn.dead ~ 1 + conc
Model 2: propn.dead ~ 1 + conc + I(conc^2)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         6    11.2322
2         5     3.1949  1   8.0373  0.004582 **
```

reproducing the results in point 10 above.

4.3.3 Calculating pseudo R^2 for the beetle quadratic model

We can get the AIC values using the '$aic' attribute as follows:

```
> quad.glm$aic
[1] 35.39294
```

There are 3 parameters for the quadratic model. This gives a log-likelihood of $3 - 1/2 \times 35.39 = -14.695$. So the pseudo R^2 for the quadratic model is $(-155.2 - (-14.70))/(-155.2) = 0.91$ which means that there is a 91% increase in the log-likelihood for the quadratic model compared to the minimal model (where the linear predictor contains just an intercept term).

4.3.4 Calculating residuals for the beetle quadratic model

In section 4.3 it was stated that the value of the Pearson residual with a binomial response is

$$e_{P,i} = \frac{y_i - \widehat{\mu_i}}{\sqrt{\widehat{\mu_i}\left(1 - \widehat{\mu_i}\right)/n_i}},$$

where μ_i is the ith fitted value. We can get these from R using

```
> round(resid(quad.glm,"pearson"),digits=3)
     1      2      3      4      5      6      7      8
-0.412  0.843 -0.275 -0.524  0.852 -0.872  0.165  0.511
```

Other residuals are available. We can check these manually using the fitted values from R.

```
> y=beetle$propn.dead
> fitted=quad.glm$fitted
> n=beetle$number
> round((y-fitted)/sqrt(fitted*(1-fitted)/n),digits=3)
     1      2      3      4      5      6      7      8
-0.412  0.843 -0.275 -0.524  0.852 -0.872  0.165  0.511
```

In section 3.5.4 the contribution of the ith observation to the residual deviance for the binomial distribution is

$$-2\sum_i n_i \left\{ y_i \log\left(\frac{\widehat{\mu_i}}{y_i}\right) + (1-y_i)\log\left(\frac{1-\widehat{\mu_i}}{1-y_i}\right)\right\}.$$

So the d_i are obtained by

```
> round(sqrt(-2*n*(y*log(fitted/y)+(1-y)*log((1-fitted)/(1-y+0.000001)))),
digits=3)
    1     2     3     4     5     6     7     8
0.422 0.819 0.277 0.523 0.875 0.825 0.170 0.722
```

We then makes those d_i negative for which $y_i < \mu_i$. These can be checked in R using

```
> round(resid(quad.glm,"deviance"),digits=3)
     1      2      3      4      5      6      7      8
-0.422  0.819 -0.277 -0.523  0.875 -0.825  0.170  0.722
```

Task 12 *Using the logit fitted line in x in Example 4.3.1, $\widehat{\eta} = -60.72 + 34.27x$, show by direct substitution that $\widehat{\mu_1} \approx 0.059$, $e_{P,1} \approx 1.41$, $e_{D,1} \approx 1.28$.*

4.3.5 Example 2 of model building: plant anthers

The data below, from Dobson (1990, pp. 113–115, Example 8.2; 2002, §7.4.1, pp. 123–125), are the number of embryogenic anthers of a plant species prepared (n_{ij}) and obtained (s_{ij}) under 2 different storage conditions and 3 values of centrifuging force (x_j). Interest is in comparing the effects of the storage conditions, one of which is a control, and the other is a new treatment (3°C for 48 hours). The table entries are (n_{ij}, s_{ij}), $i = 1, 2$, $j = 1, 2, 3$. The data is in the 'anthers' dataframe on the .RData workspace on MOLE.

	Centrifuging force (g)		
	40	150	350
Control	(102, 55)	(99, 52)	(108, 57)
Treatment	(76, 55)	(81, 50)	(90, 50)

Let M represent an indicator variable which is zero for the control group and 1 for the treatment group.

1. The plot of $y_i = s_i/n_i$ against $z_i = \log(x_i)$ (log force) suggests differences between the treatments, with no strong dependence on z (or x), but possibly an interaction between treatment and z (different slopes). The plot against x is very similar.
2. Plotting $g(y_i^*)$ against z or x shows no real difference between the various links. Consider the logit link.
3. The minimal deviance (Model 1: $\eta = \alpha$) is 10.452 (on 5 df), which is not large ($\chi^2_{5,0.95} = 11.07$).
4. Fitting two groups with no dependence on z (Model 2: $\eta_{ij} = \alpha + \alpha_1 M$) gives a deviance of 5.173 (on 4 df), showing that there is strong evidence to reject $H_0 : \alpha_i = 0$ (deviance change from Model 1 to Model 2 is 5.3 on 1 df).
5. Fitting two parallel lines (Model 3: $\eta_{ij} = \alpha + \alpha_1 M + \beta z_j$, gives a deviance of 2.619 (on 3 df), showing that there is not significant evidence to reject $H_0 : \beta = 0$ (deviance change from Model 2 to Model 3 is 2.55 on 1 df).
6. Fitting two non-parallel lines (Model 4: $\eta_{ij} = \alpha + \alpha_1 M + \beta z_j + \beta_1 M z_j$) gives a deviance of 0.028 (on 2 df) showing there is not significant evidence to reject the null hypothesis $H_0 : \beta_1 = 0$ (deviance change from Model 2 to Model 4 is 5.145 on 2 df - where $\chi^2_{2,0.95} = 5.99$). Note that in `R` this is obtained using `qchisq(0.95,2)`.
7. Using other links gives similar results which is not surprising as all sample proportions are between 0.52 and 0.73 and the different links are all similar for this range of y values.
8. Thus, the fits provide evidence of differences between the treatments, but not much evidence for a change with force.

Note that all the model comparisons using the change in (residual) deviance only compare nested models. Comparing non-nested models using this method is not valid.

Task 13 *How do the conclusions in example 4.3.5 change if the analysis is performed using x_i instead of $z_i = log x_i$ as an explanatory variable?*

4.4 Over-dispersion

If the trials aggregated to form y_i are, in fact, not homogeneous – i.e. if the success probabilities differ even though the x_i are the same – the distribution of S_i is no longer Binomial, and instead of

$$\text{var}(Y_i) = \frac{\mu_i(1-\mu_i)}{n_i}$$

a possible assumption which can be fitted using quasi-likelihood (see §3.9) is

$$\text{var}(Y_i) = \phi\frac{\mu_i(1-\mu_i)}{n_i},$$

with, usually, $\phi > 1$, which is over-dispersion. A similar situation may arise if the observations are not independent. The extra parameter, ϕ is estimated in the manner suggested in §3.8.

Also, several models for over-dispersion, e.g. Beta-Binomial, can be fitted directly. Note that an apparent need for assuming $\phi > 1$ can be due to the omission of necessary covariates. It has been suggested that over-dispersion should not be assumed unless $\hat{\phi} > 2$.

4.4.1 An example of the use of over-dispersion: grain beetle deaths

D.J. Finney (1971, Probit Analysis, Example 8, pp. 72–76) considers the following data from an experiment on the toxicity of ethylene oxide (in mg/100 ml: x is $\log_{10}$ of this) to grain beetles, the number killed (s_i) out of n_i was found 1 hour after exposure. The data is in the 'grain' dataframe on the .RData workspace on MOLE.

x_i	.394	.391	.362	.322	.314	.260	.225	.199	.167	.033
n_i	30	30	31	30	26	27	31	30	31	24
s_i	23	30	29	22	23	7	12	17	10	0

The circles in the left plot of Figure 4.4 are the observed proportion of grain beetle deaths by the log of ethylene oxide concentration. The plot looks reasonably linear but we should not fit a linear model to the data because a log concentration of 0.5 looks like it would yield an estimated probability exceeding 1. The link functions for the binomial response ensures that the response remains in the $[0, 1]$ interval.

In parts 1-4 below we fit a standard GLM with probit link function.

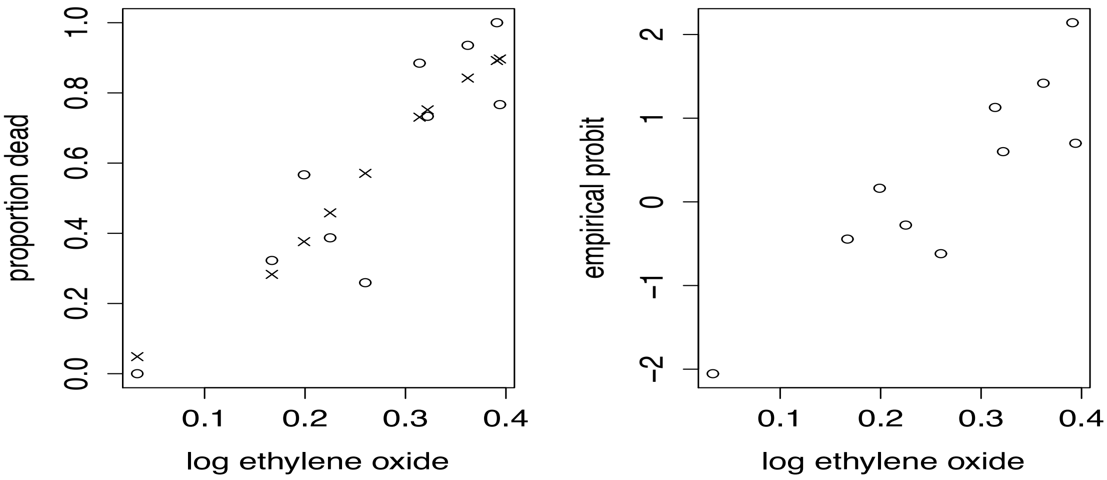

Figure 4.4: Grain beetle data plots. Left plot: fitted values (crosses) and observed values (circles) for the model with a linear term. Right plot: empirical probit plot for model with a linear term.

1. As usual, consider the proportion of deaths; $y_i = s_i/n_i$; $y_i^* = (s_i + 0.5)/(n_i + 1)$.

2. The null deviance ($\eta = \beta_0$) is 123.4 (on 9 df).

3. The right hand plot in Figure 4.4 shows a a probit plot: $g(y_i^*) = \Phi^{-1}(y_i^*)$ against x_i. It suggests a line predictor ($\eta = \beta_0 + \beta_1 x$).

4. The linear predictor $\eta = \beta_0 + \beta_1 x$ (using a probit link) has a deviance (D) of 29.44 and $X^2 = 28.95$ on 8 df. The change in deviance (94 on 1 df) is clearly very high showing an improvement of the model with a linear term over the null model (overwhelming evidence to reject $H_0 : \beta_1 = 0$). But the actual deviance for the model is high (29.44) providing evidence against this Binomial model ($\chi^2_{8,0.9995} = 27.87$). Other links give very similar results. There is no evidence of systematic curvilinearity, and no other possible explanatory variables were available. Note that the X^2 value is obtained as the sum of the squares of the Pearson residuals; see section 3.7.2. The crosses in the left hand plot of Figure 4.4 represent the fitted values for this model.

5. This lack of fit with no apparent departure from linearity and very high Pearson residuals suggests fitting an over-dispersed Binomial model

estimating ϕ as $D(y, \widehat{\mu})/(n-p) = 28.95/8 = 3.62$ (this estimate of ϕ can be obtained directly using the quasinomial argument).

6. Standard errors for parameter estimates are then multiplied by $\sqrt{\widehat{\phi}} = \sqrt{3.62} \approx 1.90$.

7. We can do the usual model fit comparing $\eta = \beta_0$ to $\eta = \beta_0 + \beta_1 x$ taking into account this estimate of $\widehat{\phi}$. The change in scaled deviance is $94/3.62 = 26$ on 1 df which is still very high showing an improvement of the model with a linear term over the null model. The scaled deviance for this model is now 29.44/3.62=8.13, a much better fit than before ($\chi^2_{7,0.95} = 14.1$). Using the `summary` command we see that $\widehat{\beta_1} = 8.64$, with e.s.e.$(\widehat{\beta_1}) = 0.980 \times 1.90 = 1.86$, so the Wald test statistic is $t = \widehat{\beta_1}/\text{e.s.e.}(\widehat{\beta_1}) = 4.7$. The slope is highly significant.

Note that the fitted values and hence residuals don't change in the over-dispersed model but our model fits the data much better after scaling by the scale parameter ϕ. Care must be exercised in doing this as it possible to over fit to the data. In this case the value of ϕ is convincingly large and there is good evidence for a over-dispersed model. If you fit an over-dispersed model and ϕ is less than 2 and your data set is small then you should be wary of fitting an over-dispersed model. Of course the degrees of freedom is reduced to account for estimation of this extra parameter and this penalizes the inclusion of the extra parameter. Note that the command to perform the over-dispersed binomial is

```
quasi.grain.glm<-glm(adj.prop.dead~1+x,family=quasibinomial,weights=n)
```

Task 14 *Reproduce this analysis, in particular obtain the estimate of ϕ directly from* R *using the quasinomial argument.*

4.5 Odds and odds ratios

If 'success' has probability μ, then the **odds** of (or in favour of, or on) 'success' are

$$\lambda = \frac{\mu}{1-\mu}, \quad 0 \leq \lambda \leq \infty, \quad \text{and so} \quad \mu = \frac{\lambda}{1+\lambda}.$$

The **log-odds** are then

$$\log \lambda = \log \mu - \log(1-\mu) = \text{logit}(\mu), \quad -\infty \leq \log \lambda \leq \infty.$$

Interest if often in the risk of some event in individuals exposed to a risk factor compared to the risk in those not exposed to the risk factor, otherwise known as the odds ratios. In medical studies the event of interest is often death or having some disease.

If Y_i is binary`{0,1}` with $\mu_i = p(Y_i = 1)$, then the odds of $Y_i = 1$ are given by

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = \frac{\mu_i}{1 - \mu_i}$$

But the logit link gives

$$log\left(\frac{\mu_i}{1 - \mu_i}\right) = \alpha + \beta x_i.$$

So it follows that

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha + \beta x_i}$$

So the odds of $Y_i = 1$ are given by $e^{\alpha+\beta x_i}$ This relationship allows us to calculate the quantity of interest, the odds ratio, from parameter estimates from logistic regression. As the name suggests, the odds ratio is just the ratio of two odds at different values of x_i. We now consider several different cases according to the form that X_i takes, factor or continuous variable. We illustrate the different cases with some data relating saturated fat intake, gender and age to coronary heart disease (CHD). The data set consists of 120 individuals collected as part of a case-control study. The data is in the 'CHD' dataframe on the .RData workspace on MOLE. The variables available are

- *Status* - whether the individual has CHD or not;
- *Sat.fat* - whether the individual has a high fat diet (coded as 2), a medium fat diet (coded as 1) or a low fat diet (coded as 0);
- *Gender* - female (coded as 0) or male (coded as 1);
- *Age* - the age of the individual in years.

The response variable (*Status*) is binary (affected by CHD or not).

4.5.1 X_i is a factor with 2 levels

Assume initially that X_i has only two levels `{0,1}`.

Then when $X_i = 1$ the odds of $Y_i = 1$ are $\frac{p(Y_i=1)}{1-p(Y_i=1)} = e^{\alpha+\beta}$ and when

$X_i = 0$ the odds of $Y_i = 1$ are $\frac{p(Y_i=1)}{1-p(Y_i=1)} = e^{\alpha}$. So the ratio of the odds of $Y_i = 1$ when $X_i = 1$ compared to the odds of $Y_i = 1$ when $X_i = 0$ is $\frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$.
This is the odds ratio of $Y_i = 1$ when $X_i = 1$ compared to when $X_i = 0$. So, if X_i is a factor with 2 levels the odds ratio of $Y_i = 1$ is e^{β} where β is the coefficient of X_i in the logistic regression analysis.

Fitting a logistic model (using the logit link) we get:

```
> gender.glm<-glm(status~gender,family=binomial,data=CHD.data)
> summary(gender.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.03509    0.26495  -0.132   0.8946
gender       0.80056    0.37875   2.114   0.0345 *
```

The linear predictor is $\beta_0 + \beta_1 x_i$ where x_i is 1 for males and 0 for females. The odds of CHD ($Y_i = 1$) for females ($x_i = 0$) is e^{β_0}. The odds of CHD ($Y_i = 1$) for males ($x_i = 1$) is $e^{\beta_0+\beta_1}$. So the estimated odds ratio for being affected by CHD in males compared to females is $e^{\widehat{\beta_1}} = e^{0.801} = 2.23$ to 2d.p. This is generally interpreted as meaning that, based on this study, the risk of CHD for males is 2.2 higher than the risk of CHD for females.

4.5.2 X_i is a factor with 3 levels

If we code the three factors as (0,1,2) then when $X_i = 2$

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha+2\beta}$$

and when $X_i = 1$

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha+\beta}$$

and when $X_i = 0$

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha}$$

So the odds ratio of $Y_i = 1$ comparing $X_i = 1$ to $X_i = 0$ is e^{β} and the odds ratio of $Y_i = 1$ comparing $X_1 = 2$ to $X_i = 0$ is $e^{2\beta} = \left(e^{\beta}\right)^2$. So with X_i coded in this way, the odds ratios behave in a multiplicative manner.

```
> satfat.glm<-glm(status~sat.fat,family=binomial,data=CHD.data)
> summary(satfat.glm)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3623     0.2970  -1.220  0.22251
sat.fat       0.8181     0.2644   3.095  0.00197 **
```

The linear predictor is $\beta_0 + \beta_1 x_i$ where X_i is 2 for a high fat diet, 1 for a medium fat diet and 0 for a low fat diet. The estimated odds ratio of $Y_i = 1$ comparing

- $X_i = 1$ to $X_i = 0$ is $e^{\widehat{\beta_2}} = e^{0.818} = 2.7$;
- $X_i = 2$ to $X_i = 1$ is $e^{\widehat{\beta_2}} = e^{0.818} = 2.7$;
- $X_i = 2$ to $X_i = 0$ is $e^{\widehat{2\beta_2}} = \left(e^{0.818}\right)^2 = 5.13$ to 2d.p.

This coding of the variable assumes that the risk of CHD is multiplicative across group levels. With this kind of variable coding, the lowest level (0 by convention) is taken to be the baseline level and odds ratios are usually calculated comparing the odds in other levels to the baseline. It is possible to allow a non-multiplicative model of disease risk. This requires setting up dummy variables so that each level of the factor has a different parameter.

4.5.3 X_i is continuous

When $X_i = t + 1$

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha + \beta(t+1)}$$

and when $X_i = t$

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha + \beta t}$$

So e^{β} is the odds ratio of $Y_i = 1$ comparing $X_i = t + 1$ to $X_i = t$ So e^{β} is the odds ratio of $Y_i = 1$ for a unit increase in X_i

```
> age.glm<-glm(status~age,family=binomial,data=CHD.data)
> summary(age.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.20096    1.11289  -1.978   0.0480 *
age          0.05459    0.02340   2.333   0.0196 *
```

So the estimated odds ratio of $Y_i = 1$ comparing $X_i = t + 1$ to $X_i = t$ is $e^{0.055} = 1.06$. For every year older the risk of CHD increases by 6%. For a 10 year difference in age, the estimated odds ratio is $e^{(10\times 0.055)} = \left(e^{0.055}\right)^{10} = (1.06)^{10} = 1.80$; an 80% increase in CHD risk for a 10 year difference in age.

4.5.4 Multiple explanatory variables

Suppose we have k variables $X_1, X_2, ..., X_k$. How do we interpret the parameters of the logistic regression model? Suppose that $X_1 = x_1 + 1, X_2 = x_2, ..., X_k = x_k$ Then the odds are

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha+\beta_1(x_1+1)+\beta_2 x_2+...+\beta_k x_k}$$

Suppose that $X_1 = x_1, X_2 = x_2, ..., X_k = x_k$ Then the odds are

$$\frac{p(Y_i = 1)}{1 - p(Y_i = 1)} = e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}$$

Then e^{β_1} is the odds ratio of $Y_i = 1$ for a unit increase in X_1 holding all other values in the model fixed. This is often called 'adjusting' for another variable and is common in epidemiological studies. The variable you wish to adjust for is simply included as an explanatory variable in the linear predictor.

Suppose we want to calculate the odds ratio of $Y_i = 1$ comparing males to females but adjusting for age. We would use the estimates from:

```
> both.glm<-glm(status~gender+age,family=binomial,data=CHD.data)
> summary(both.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.59535    1.13704  -2.283   0.0225 *
gender       0.80587    0.38835   2.075   0.0380 *
age          0.05436    0.02347   2.317   0.0205 *
```

So the estimated odds ratio of CHD comparing males to females but adjusting for age is $e^{0.806} = 2.24$; hardly any change from the unadjusted odds ratio of 2.23 in the previous example but in some cases adjusting can lead to very different estimates.

4.5.5 Confidence intervals for odds ratios

Using the output above we can calculate a 95% confidence interval for the **log** odds ratio of CHD comparing males to females but adjusting for age as

$0.806 \pm 1.96 \times 0.388$ (estimate ± 1.96 standard errors) which gives (0.05,1.57). This gives a 95% CI for the odds ratio as (1.05,4.79).

Chapter 5

Poisson regression and models for count data

5.1 Introduction

The Poisson distribution is often appropriate for count data when a Binomial is not (because there is no upper bound, Bernoulli trials are inappropriate, etc). Its relationship to the Poisson process also makes it a natural first choice when modelling counts of events, including rare events. It has a further indirect use in connection with contingency tables, to be explored in Chapter 7. Here we review direct Poisson regression and similar regression models for non-negative data, not necessarily discrete, in which the variance depends on the mean. Non-negative and positively skew data, for which these models could be suitable, arise in many applications, such as lifetimes, survival times, recovery times, counts and lengths. Some other models for such data are briefly described.

5.2 Modelling counts using the Poisson distribution

The Poisson GL model for non-negative integer data is

$$Y_i \sim Po(\mu_i) \;\text{ with } \mu_i = h(x_i^T\beta).$$

The canonical link function is the log link (which ensures $\mu_i > 0$)

$$\log \mu_i = x_i^T\beta \;\text{ i.e. }\; \mu_i = \exp(x_i^T\beta) = \prod_j \exp(x_{ij}\beta_j)$$

giving a multiplicative structure for μ_i (for example, a term for age times one for height).

For a single explanatory variable x, a rough check on this link and linearity can be obtained by plotting $\log(y_i + 1/2)$ (the $1/2$ is to avoid problems if $y_i = 0$) against x_i. Other links can be used as long as non-negativity of the means is ensured.

The Deviance is, as given in §3.5.4,

$$D = 2\sum\left\{y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right\}$$

with the second sum usually 0, and the Pearson residual is

$$e_{P,i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i},}$$

with

$$X^2 = \sum \frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$$

Thus D and X^2 are, as in §4.3, the sums of $2\times O\times\log(O/E)$ and $(O-E)^2/E$, respectively.

This is a GLM with $V(\mu) = \mu$, and all the results in Chapter 3 apply. Since $\phi = 1$ is known, the (residual) deviance is, for a correct distribution and model, also asymptotically χ^2_{n-p}, although in practice the χ^2_{n-p} distribution may not be a good approximation. (It should be good when all the μ_i are fairly large.)

5.2.1 Example: AIDS deaths over time

Dobson (1990, Example 3.3 and Exercise 4.1) gives the data below for the numbers of deaths from AIDS in Australia per quarter from the first quarter of 1983 to the second quarter of 1986 (numbered 1 to 14). The data is in the 'AIDS' dataframe on the .RData workspace on MOLE.

y_i	0	1	2	3	1	4	9	18	23	31	20	25	37	45
x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The data are shown in the left plot of Figure 5.1.

1. The plot of $\log(y_i + 1/2)$ against x_i shown in the right of Figure 5.1 suggests a straight line (with a hint of downwards curvature).
2. Fitting a Poisson with log link and a line predictor on x produces a fit

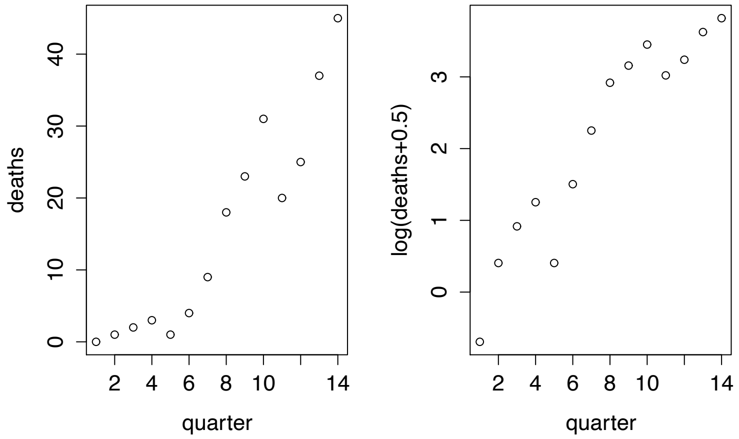

Figure 5.1: AIDS data plots- left: aids deaths per quarter; right: log(aids deaths) per quarter

which is a bit poor ($D = 29.65$ on 12 df, $\chi^2_{12,0.95} = 21.03$).

3. Adding a quadratic term (x^2) is an improvement (difference 13.01 on 1 df), and also adequate ($D = 16.37$ on 11 df).
4. A line predictor including a $\log(x)$ term has $D = 17.09$ on 12 df, which appears adequate. Adding a quadratic term (in $\log x$) does not improve the fit ($D = 16.64$ on 11 df).
5. Thus possible simple models are a line in $\log x$ or a quadratic in x, but there are reservations about both.

Task 15 *Verify the analysis in Example 5.2.1.*

5.3 Adjusting for exposure: offsets

Suppose y_i is the number of events in time interval t_i. Then, modelling events by Poisson process, with possibly different rates in different intervals t_i, it is natural to suppose that y_i is the observed value of a random variable Y_i for which

$$Y_i \sim Po(\mu_i) = Po(\lambda_i t_i),$$

where λ_i the rate per unit time in the t_i interval. Suppose we are interested in how the rates λ_i depend on covariates x;

$$\log \mu_i = \log(\lambda_i t_i) = \log \lambda_i + \log t_i = x_i^T \beta + \log t_i.$$

Here $\log t_i$ looks like a covariate, but with a known coefficient (equal to 1). Note that β models how *rates* vary, which is usually what really matters, rather than how *means* vary.

Of course one could include $\log t_i$ as a regular explanatory variable and see whether its regression coefficient is near 1 before using it as an offset. This is a check on the model, not a way of indicating that an offset should be used; the desire to use a variable as an offset arises from the modelling, not from looking at the data.

In general when data on counts arise from different degrees of exposure it is natural, as above, to want to model and to draw conclusions about rates per unit exposure. For example, suppose the data are on death counts taken over different large populations, perhaps different health authorities, and some other explanatory variables. Then, the logarithm of the population size would be used as an offset, since it is death rates per person (or, more usually, per 1000 people) that are of interest.

An offset for a Poisson regression model is specified in `R` by means of the `offset` function. The following example illustrates.

5.3.1 Example: Smoking and heart disease

Dobson (2002, §9.2.1, pp. 154–7) gives the following (historically important) data on deaths from coronary heart disease after 10 years among British male doctors (from Doll et al.). Age groups are coded as 1 for ages 35–44; 2 for 45–54; 3 for 55–64; 4 for 65–74; and 5 for 75–84. Exposure is measure in person-years; a person who lived throughout the period would contribute 10 person-years, but someone who died, either of heart disease or of something else, would contribute fewer. The data is in the 'smoking' dataframe on the .RData workspace on Blackboard.

Age	Smokers		Non-smokers	
Group	Deaths	Person-years	Deaths	Person-years
1	32	52407	2	18790
2	104	43248	12	10673
3	206	28612	28	5710
4	186	12663	28	2585
5	102	5317	31	1462

1. The age group codes $(1, 2, \ldots)$ are easily translated to a numerical variable 'mid-point of age group' giving 40, 50, 60, 70 and 80 years.

2. Over the age groups, the rounded death rates (deaths per 100,000 person years) are 61, 240, 720, 1469 and 1918 for smokers, and 11,

112, 490, 1083 and 2120 for non-smokers. Death rates for smokers are higher except for age group 5. Their ratios are 5.5, 2.1, 1.5, 1.4 and 0.90. The death rates for the two groups are shown in the top left plot of Figure 5.2.

3. The plot of log death rate against age suggests that there are differences between the two groups, and that, with a log link, straight lines will not be adequate. The curves appear to be approaching a maximum or asymptote as age increases suggesting quadratic terms may be needed. The log death rates for the two groups are shown in the top right plot of Figure 5.2.

4. The numbers of deaths are counts, so it is natural to think of a Poisson model. However, the number of deaths will be influenced by the number of people at risk. In other words, person-years must be taken into account; it is really death rates per person-year of exposure that we wish to model. Accordingly a Poisson model with log link and linear predictor

   $$\eta = \log(\text{person-years}) + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2$$

   was fitted, where x_1 is 1 for smokers, 0 for non-smokers; x_2 is the age category (as a covariate). This fits separate quadratics in age for the two groups, but with the same term in x_2^2.

   This model has fixed coefficient 1 for log(person-years), and so an offset is needed. See below for the command to include it.

5. The fit is extremely good, with $D = 1.635$ on 5 df, and with all terms appearing necessary. The fitted values for this model are shown in the bottom left plot of Figure 5.2. The fitted coefficients are -19.70, 2.36, 0.36, -0.020, -0.031, respectively. Thus for non-smokers, $\widehat{\eta}$ is $-19.7 + 0.36x_2 - 0.02x_2^2$, and for smokers it is $-17.34 + 0.33x_2 - 0.02x_2^2$. The residuals for this model are shown in the bottom right right plot of Figure 5.2.

Task 16 *Verify the analysis in Example 5.3.1, using the coomand*

```
glm(deaths~offset(log(person.years))+smoke*age+I(age^2),poisson,data=smoking)
```

5.4 Non-negative data with variance ∝ mean

Using quasi-likelihood, the model in which, for $Y_i > 0$, $\text{var}(Y_i) = \phi \times \mu_i$, can be fitted for any non-negative data by fitting the Poisson distribution and then (if unknown) estimating ϕ. The Y variable does not need to be a count for this to be acceptable.

Task 17 *Compare the output from fitting a Poisson with log link and a line predictor on x to the data in Example 5.2.1 with that obtained using the the log link and assuming that the variance is proportional to the mean.*

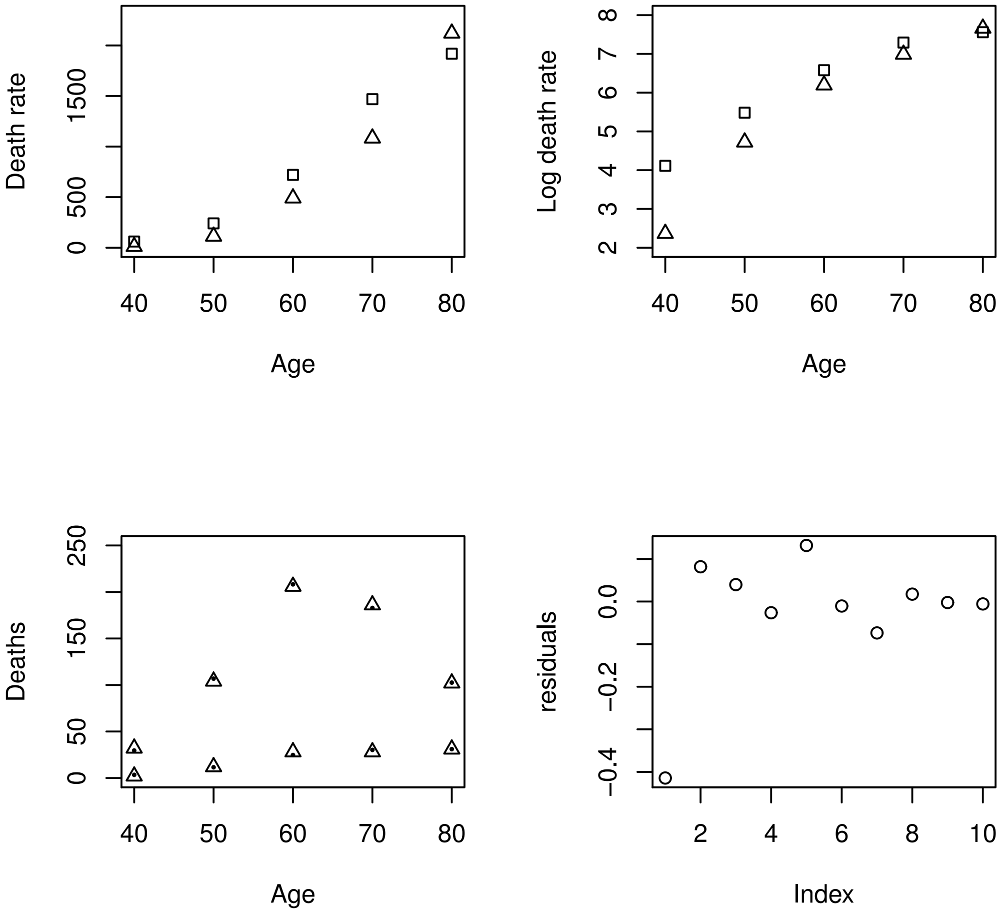

Figure 5.2: Plots for the smoking data - in the top-row plots, squares are smokers, triangles are non-smokers - in the bottom left plot, triangles are observed values, dots are fitted values

Chapter 6

Random effects models

6.1 Introduction

In many studies we collect a number of observations for each individual. For example in a drug development study for each individual i taking part in a trial we may collect data at three periods of time during the trial. This is known as *repeated measures* and is typical in *longitudinal studies.* The repeated measures are likely to be similar for a given individual and so they introduce correlation to the data. Often when our data is correlated it is due to repeated measures. In a linear or generalised linear model there are components we are interested in to estimate (these are usually called *fixed effects*, and there are components which can account for the variation and correlation (called *random effects*). A model may include fixed effects and random effects and sometimes such a model is called *mixed model.* The generalised linear models we have described in Chapters 1-5 include fixed effects only. In this chapter we discuss models which include both fixed and random effects. We start in Section 6.2 by considering linear models with a Normally distributed response and in Section 6.3 we extend this to generalised linear models.

6.2 Random effects in (Normal) linear models

This section introduces random effects models for Normally distributed data and discusses statistical inference. To fix notation, suppose that we collect observations y_{it}, for individual $i = 1, 2, \ldots, n$ and for repeated measures $t = 1, 2, \ldots, T_i$. Frequently we shall have fixed length of the repeated measures $T_i = T$, but this is not need to be assumed. We use the index t for time, as in longitudinal studies repeated measures are collected over time. However, we shall not consider any time-specific features of the index (we can use j instead of t, but using t is better to distinguish i from t). We shall also assume familiarity with linear models estimation, e.g. least squares estimation. Before we discuss the general model formulation (Section 6.2.3) we discuss two special cases which introduce random effects models.

Consider the `sleepstudy` data, source Belenky et al. (2003),[1] available in the `lme4` package in `R`. A sleep deprivation study was conducted to assess sleep habits of individuals. The average reaction time for each individual (referred to as subject in the study) was recorded, for a number of tests taken during the day after the sleep. The data were collected over 10 days: on day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night.

	Reaction time	Day	Subject
1	249.5600	0	308
2	258.7047	1	308
3	250.8006	2	308
4	321.4398	3	308
5	356.8519	4	308
6	414.6901	5	308
7	382.2038	6	308
8	290.1486	7	308
9	430.5853	8	308
10	466.3535	9	308
11	222.7339	0	309
12	205.2658	1	309
13	202.9778	2	309
14	204.7070	3	309
15	207.7161	4	309

Here we have y_{it} as the average reaction time, $i = 1, 2, \ldots, 18$ and $t = 1, 2, \ldots, 10$, totalling $nT = 18 \times 10 = 180$ observations. The table above tabulates the first 15 rows of the data.

[1] Belenky, G., Wesensten, N.J., Thorne, D.R., Thomas, M.L., Sing, H.C., Redmond, D.P., Russo, M.B. and Balkin, T.J. (2003) Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. Journal of Sleep Research 12, 1–12.

6.2.1 Random intercept model

Consider observations y_{it}, for individual $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T_i$, for some fixed n and T_i. Consider a linear model

$$y_{it} = \tau_i + x_{it}^T \gamma + \epsilon_{it},$$

where x_{it} is a $p \times 1$ column vector of covariates, γ a $p \times 1$ vector of coefficients, and ϵ_{it} an error term, with $\epsilon_{it} \sim N(0, \sigma_\epsilon^2 I)$. It is also assumed that $\tau_i \sim N(\tau, \sigma^2)$ and that τ_i is independent of ϵ_{it}, for any i and t. This model assumes that the intercept τ_i is individual-specific and its mean is equal to a constant intercept τ. The variance σ^2 allows for variation of the intercept around τ.

As we said before random effects models can account for correlated data around the repeated measures. In this case we can see that the correlation coefficient of y_{it} and y_{is}, for $i \neq s$ is

$$\rho_{ts} = \frac{\mathrm{cov}(y_{it}, y_{is})}{\sqrt{\mathrm{var}(y_{it})\mathrm{var}(y_{is})}} = \frac{\sigma^2}{\sqrt{(\sigma^2 + \sigma_\epsilon^2)^2}} = \frac{\sigma^2}{\sigma^2 + \sigma_\epsilon^2}, \tag{6.1}$$

since the covariance of y_{it} and y_{is} is

$$\mathrm{cov}(y_{it}, y_{is}) = \mathrm{cov}(\tau_i, \tau_i) + \mathrm{cov}(\epsilon_{it}, \epsilon_{is}) = \mathrm{var}(\tau_i) = \sigma^2.$$

We can see that if $\sigma^2 = 0$, then $\tau_i = \tau$ and the intercept τ is not random and can be included in the fixed effects γ. Likewise when σ^2 is small compared to σ_ϵ^2 the correlation introduced by the repeated measures is small. On the other hand if σ^2 is large there is more variability between individuals i due to the repeated measures.

Model (6.1) can be written in matrix form as

$$\begin{aligned} y_i &= \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{pmatrix} = \begin{pmatrix} 1 & x_{i1}^T \\ 1 & x_{i2}^T \\ \vdots & \vdots \\ 1 & x_{iT_i}^T \end{pmatrix} \begin{pmatrix} \tau \\ \gamma \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (\tau_i - \tau) + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT_i} \end{pmatrix} \\ &= Z_i \beta + W_i b_i + \epsilon_i \end{aligned}$$

where the fixed effects $\beta^T = (\tau, \gamma^T)$ includes γ and τ and the random effects $b_i = \tau_i - \tau \sim N(0, \sigma^2)$, with respective matrices

$$Z_i = \begin{pmatrix} 1 & x_{i1}^T \\ 1 & x_{i2}^T \\ \vdots & \vdots \\ 1 & x_{iT_i}^T \end{pmatrix} \quad \text{and} \quad W_i = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

6.2.2 Random slope model

The model with individual-specific intercept of the previous section can be extended to accommodate an individual-specific slope as well. Consider observations y_{it}, for individual $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T_i$, for some fixed n and T_i. Consider a linear model

$$y_{it} = \tau_i + x_{it}^T \gamma_i + \epsilon_{it},$$

where τ_i, x_{it}, ϵ_{it} are defined in Section 6.2.1 and γ_i is a vector of coefficients, which is individual-specific. The random effects are put in a vector β_i with

$$\beta_i = \begin{pmatrix} \tau_i \\ \gamma_i \end{pmatrix} \qquad \beta_i \sim N(\beta, Q) \equiv N\left[\begin{pmatrix} \tau \\ \gamma \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & Q_\gamma \end{pmatrix} \right],$$

for σ^2 the variance of τ_i and Q_γ the covariance matrix of γ_i. Here β is the fixed effects vector and β_i the random effects vector.

This model can be written as

$$\begin{aligned} y_i &= \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{pmatrix} = \begin{pmatrix} 1 & x_{i1}^T \\ 1 & x_{i2}^T \\ \vdots & \vdots \\ 1 & x_{iT_i}^T \end{pmatrix} \begin{pmatrix} \tau \\ \gamma \end{pmatrix} + \begin{pmatrix} 1 & x_{i1}^T \\ 1 & x_{i2}^T \\ \vdots & \vdots \\ 1 & x_{iT_i}^T \end{pmatrix} \begin{pmatrix} \tau_i - \tau \\ \gamma_i - \gamma \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iT_i} \end{pmatrix} \\ &= Z_i \beta + W_i b_i + \epsilon_i \end{aligned}$$

For this model we can show that the correlation of y_{it} and y_{is} is

$$\operatorname{cor}(y_{it}, y_{is}) = \frac{\sigma^2 + x_{it}^T Q_\gamma x_{is}}{\sqrt{\left(x_{it}^T Q_\gamma x_{it} + \sigma^2 + \sigma_\epsilon^2\right)\left(x_{is}^T Q_\gamma x_{is} + \sigma^2 + \sigma_\epsilon^2\right)}}.$$

6.2.3 General model formulation

From the models considered in Sections 6.2.1 and 6.2.2 we can motivate the general model form of the random effects linear model. With the notation of y_{it} of Section 6.2 consider the model

$$y_i = Z_i \beta + W_i b_i + \epsilon_i, \tag{6.2}$$

where Z_i is a design matrix relevant to the fixed effects β, W_i is a matrix of weights relevant to the random effects b_i and ϵ_i is the column vector comprising of the error terms $\epsilon_{i1}, \ldots, \epsilon_{iT_i}$. It is further assumed that

$$b_i \sim N(0, Q), \tag{6.3}$$

for some covariance matrix Q and that b_i and ϵ_i are independent. Matrices Z_i and W_i will be known according to the particular structure or form of the design as in the random intercept and random slope models of Sections 6.2.1 and 6.2.2. Model (6.2)-(6.3) is known as the *random effects* linear model for Normal data; sometimes it is referred to *mixed effects* linear model, because it includes both fixed and random effects. This model can be written as

$$y_i = Z_i\beta + \epsilon_i^*, \tag{6.4}$$

where b_i is embedded in ϵ_i^*, so that

$$\epsilon_i^* = W_i b_i + \epsilon_i \sim N(0, W_i Q W_i^T + \sigma_\epsilon^2 I).$$

Hence model (6.2) can be seen as a linear model (6.4), with errors ϵ_i^*, which are heteroscedastic (variances depend on i and have correlated observations (as $W_i Q W_i^T + \sigma_\epsilon I$ is not proportional to the identity matrix).

6.2.4 Statistical inference

Consider that data y_i are generated from model (6.2)-(6.3). For a realised set of data $y = (y_1, y_2, \ldots, y_n) = (y_{11}, y_{12}, \ldots, y_{nT_n})$ we wish to estimate the fixed effects β and the random effects b_i.

Known variances

First consider the case of known variances which are usually put in notation $\theta = (Q, \sigma_\epsilon^2)$. Write $V_i = W_i Q W_i^T + \sigma_\epsilon^2 I$ the covariance matrix of ϵ_i^* (see equation (6.4)). If θ is known, then V_i is known too.

First we deal with fixed effects β. Write $V_i^{-1/2}$ the square root of the matrix V_i^{-1}: this can be any suitable definition e.g. being the symmetric square root or based on the Choleksi decomposition). From the equivalent form of model (6.4) and as V_i is known, we can write this model as

$$\begin{aligned} V_i^{-1/2} y_i &= V_i^{-1/2} Z_i \beta + V_i^{-1/2} \epsilon_i^* \\ y_i^* &= Z_i^* \beta + \epsilon_i^{**}, \quad \epsilon_i^{**} \sim N(0, I), \end{aligned} \tag{6.5}$$

where $y_i^* = V_i^{-1/2} y_i$, $Z_i^* = V_i^{-1/2} Z_i$ and $\epsilon_i^{**} = V_i^{-1/2}\epsilon_i^*$. It is easy to verify that $\epsilon_i^{**} \sim N(0, I)$. Now, model (6.5) is a standard linear model and the LS estimator of β is

$$\begin{aligned} \hat{\beta} &= \left(\sum_{i=1}^n Z_i^{*T} Z_i^* \right)^{-1} \sum_{i=1}^n Z_i^{*T} y_i^* \\ &= \left(\sum_{i=1}^n Z_i^T V_i^{-1} Z_i \right)^{-1} \sum_{i=1}^n Z_i^T V_i^{-1} y_i. \end{aligned} \tag{6.6}$$

Equation (6.6) is known as generalised least squares estimation for the linear model. Here the random effects provide a means of determining the correlation structure on the covariance matrix V_i.

Moving on to the estimation of b_i, note that unconditionally of the data, $b_i \sim N(0, Q)$, where Q is known. It does not make sense to take the expectation of b_i as the estimate of b_i, as this is zero. We shall calculate the conditional distribution of b_i, given the data y and then take the expectation of this distribution as a point estimate of b_i.

Below we give a reminder from standard multivariate Normal distribution theory. Let X and Y be random vectors, each following a Normal distribution, $X \sim N(m_X, V_X)$, $Y \sim N(m_Y, V_Y)$ and covariance $\text{cov}(X, Y) = C_{XY}$, for some known $m_X, m_Y, V_X, V_Y, C_{XY}$. Write the joint distribution of X and Y as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} m_X \\ m_Y \end{pmatrix}, \begin{pmatrix} V_X & C_{XY} \\ C_{XY}^T & V_Y \end{pmatrix}\right].$$

Then the conditional distribution of X, given $Y = y$ is

$$X \mid Y = y \sim N(m_X + C_{XY}V_Y^{-1}(y - m_Y), V_X - C_{XY}V_y^{-1}C_{XY}^T) \qquad (6.7)$$

for some observed value y of Y.

Coming back now to our situation the joint distribution of b_i and y_i (using $\beta = \hat{\beta}$) is

$$\begin{pmatrix} b_i \\ y_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ Z_i\hat{\beta} \end{pmatrix}, \begin{pmatrix} Q & C \\ C^T & V_i \end{pmatrix}\right],$$

where the covariance C is

$$C = \text{cov}(b_i, y_i) = \text{cov}(b_i, W_i b_i) = \text{var}(b_i)W_i^T = QW_i^T.$$

then from equation (6.7) the conditional expectation of b_i given y_i is

$$\hat{b}_i = E(b_i \mid y_i) = QW_i^T V_i^{-1}(y_i - Z_i\hat{\beta}).$$

We remark that $\hat{b}_i$ depends on observation y_i only from the random effects and might depend on y_j $j \neq i$ only via $\hat{\beta}$.

Unknown variances

In practice, $\theta = (Q, \sigma^2)$ and so V_i will not be known. In this case we are not able to perform the first step in equation (6.6) as y_i^* depends on V_i. In this case we work with model (6.4). We can form the log-likelihood function as

$$\begin{aligned} l(\beta, \theta; y) &= \log \prod_{i=1}^{n} f(y_i \mid \beta, \theta) \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\log|V_i| + (y_i - Z_i\beta)^T V_i^{-1}(y_i - Z_i\beta)\right] + c, \quad (6.8) \end{aligned}$$

where c is a constant (not involving β or V_i) and we have used $y_i \mid \theta \sim N(Z_i\beta, V_i)$. Notice that we can estimate β and the variances Q and σ^2, but we cannot estimate b_i as this is embedded into ϵ_i^* in model (6.4). This is not an issue, as the reason we have introduced the random effects is to explain the variability of the data and so b_i does not represent an actual quantity we wish to estimate its effects.

There are many possibilities for maximisation of (6.8) with respect to β and θ. Some of these possibilities include maximisation based on Newton-Raphson method and Fisher scores and the expectation maximisation (EM) algorithm; these are described in detail in Chapter 7 of Fahrmeir and Tutz (2001)[2]. An alternative to maximum likelihood estimation is restricted maximum likelihood estimation (REML), which is the default estimation method in the `lmer` function in `R`; for details see below.

Sleep-study data revisited

In this section we revisit the `sleepstudy` data described in Section 6.2. To fit model (6.2) in `R` we need to use the command `lmer` in the package `lme4`. The general structure of the command is

```
lmer( response ~ covariate1 + ( covariate2 | variable ), data)
```

where `response` is the response of the data, `covariate1` is the specification for the fixed effects, and the notation `( covariate2 | variable )` indicates the random effects, for some `covariate2`. Note that if we want to specify b_i with $W_i = 1$, then we can write `( 1 | variable )` in the random effects section.

For our data we shall fit the model:

$$y_i = \beta_{0,D} + x_{it}\beta_{1,D} + b_{0,i} + x_{it}b_{1,i} + \epsilon_i,$$

where $\beta_{0,D}$ and $\beta_{1,D}$ are the fixed effects for Day (intercept and slope) and $b_{0,i}$ and $b_{1,i}$ the random effects (intercept and slope) for subject. This model is fitted in `R` using the command and giving the ouput

```
> fit <- lmer(Reaction ~ Days + (Days | Subject), data=sleepstudy)
>
> summary(fit)
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
```

[2] Fahrmeir, L. and Tutz, G. (2001) Multivariate Statistical Modelling Based on Generalised Linear Models. 2nd edition, Springer

```
   Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9536 -0.4634  0.0231  0.4634  5.1793

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.10   24.741
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825  36.838
Days          10.467      1.546   6.771

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

The output gives the variances of the random effects, 612.1 for the intercept $b_{0,i}$ and 35.07 for the slope $b_{1,i}$ and the estimates of the fixed effects, $\hat{\beta}_{0,D} = 251.405$ and $\hat{\beta}_{1,D} = 10.467$. Note that estimates of b_i are not provided because in model (6.4) b_i is embedded into ϵ_t^* and only the variances Q and σ^2 can be estimated. An application of a standard fixed effects linear model (without the Subject being included in the random effects) reveals the benefit of the random effects linear model

```
> fm0 <- lm(Reaction ~ Days, data=sleepstudy)
> summary(fm0)

Call:
lm(formula = Reaction ~ Days, data = sleepstudy)

Residuals:
     Min       1Q   Median       3Q      Max
-110.848  -27.483    1.546   26.142  139.953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  251.405      6.610  38.033  < 2e-16 ***
Days          10.467      1.238   8.454 9.89e-15 ***
---

Residual standard error: 47.71 on 178 degrees of freedom
Multiple R-squared:  0.2865,Adjusted R-squared:  0.2825
F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

Here we notice that the estimates of the intercept and the slope in the linear model are exactly equal to the estimates of the fixed effects part in the random effects model. Note that the residuals of the mixed effects model (or random effects model) are the scaled residuals, while the residuals in the fixed effects only model are the non-standardised residuals. If we standardise the residuals of the fixed effects model we will find they are comparable of that of the mixed effects model.

6.3 Random effects in generalised linear models

6.3.1 Model description

Consider now that data y_{it} are generated from the exponential family of distributions (3.1) so that the mean response function $\mu_{it} = E(y_{it})$ is mapped via the link function to the linear predictor η_{it} with the usual link

$$g(\mu_{it}) = \eta_{it} \quad \text{or} \quad \mu_{it} = g^{-1}(\eta_{it}) = h(\eta_{it}).$$

In the formulation of the GLM of Chapter 3 the linear predictor includes only fixed effects, i.e. $\eta_{it} = Z_i\beta$. Motivated from the discussion of Section 6.2 for Normal random effects linear models, we extend the linear predictor assumption in the GLM to incorporate random effects

$$\eta_{it} = Z_{it}\beta + W_{it}b_i, \tag{6.9}$$

where β are the fixed effects and b_i are the random effects. The design matrix Z_{it} and the weights matrix W_{it} depend on the specific model structure. This model is called *random effects generalised linear model* and extends the generalised linear model of Chapter 3 to incorporate random effects and extends the random effects linear model of Section 6.2 to account for non-Gaussian (non-Normal) data.

The random intercept and random slope models of Sections 6.2.1 and 6.2.2 can be readily transferred to generalised linear setting, e.g. for the random intercept we have a GLM set-up with linear predictor

$$\eta_{it} = \tau_i + x_{it}^T\gamma,$$

with τ_i being the random intercept with $\tau_i \sim N(\tau, \sigma^2)$. The fixed effects are τ and γ and the random effects are $b_i = \tau_i - \tau$, so that $b_i \sim N(0, \sigma^2)$. Here we have that the covariance matrix Q of the random effects is scalar and equal to σ^2; see also Section 6.2.1.

Statistical inference is more complex compared to random effects linear models and generalised linear models with fixed effects only. For known variances Q, one popular approach is posterior mode estimation. For unknown Q perhaps the most efficient estimation is via the EM algorithm, but REML is also a popular choice. Details of these approaches are reported in Chapter 7 of Fahrmeir and Tutz (2001)[3].

6.3.2 Example

In this section we briefly discuss the logistic regression model with the inclusion of random effects. Similar model formulations apply for the Poisson regression models.

Binary logistic model

Consider y_{it} is generated from a binary logistic model, with probability

$$\mu_{it} = E(y_{it} \mid b_i) = P(y_{it} = 1 \mid b_i) \tag{6.10}$$

where μ_{it} is linked to the linear predictor η_{it} using the logit transformation

$$\mu_{it} = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})},$$

In this model the linear predictor is

$$\eta_{it} = z_{it}^T \beta + b_i, \quad b_i \sim N(0, Q),$$

for some design vector z_{it} (which may include the unit for a fixed effects intercept) and $Q \geq 0$. The fixed effects are described by β and the random effects by b_i.

In order to fit generalised random effects linear models in `R` we use the function `glmer` available in the package `lme4`. The syntax of this function is similar to the syntax of the `glm` and `lmer` function and is given below

```
glmer(response ~ covariate1 + ( covariate2 | variable), data, family)
```

[3] Fahrmeir, L. and Tutz, G. (2001) Multivariate Statistical Modelling Based on Generalised Linear Models, 2nd edition. Springer.

where `response` is the response, `covariate1` and `covariate2` indicate suitable covariates and `variable` is the random effects and `family` indicates the the exponential family of distributions, which is set as in the `glm` function (see Section 2.3).

We consider the `ohio` data set available in the `R` in the package `geepack`. In a study that aim to assess air-pollution effects of children at Ohio, binary observations $0, 1$ are recorded of wheeze status 1: is associated with difficulty in breathing and 0 not. According to Wikipedia: Wheezing is a high-pitched whistling sound made while breathing. It's often associated with difficulty breathing. In the `ohio` data this is the variable `resp`.The data set includes 3 more variables: `id` a numeric vector for subject id, `age`: the age of each child coded -2, -1,0,1,2 and 0 corresponding to 9 years old, `smoke` an indicator variable indicating whether the mother of each child was a smoker or not at year 7. The data are described in Zeger et al. (1988)[4] and Fitzmaurice and Laird (1993)[5]

We start by fitting a GLM with fixed effects only:

```
> fit1 <- glm(resp ~ age + smoke, data=ohio, family=binomial)
> summary(fit1)

Call:
glm(formula = resp ~ age + smoke, family = binomial, data = ohio)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6685  -0.5909  -0.5608  -0.5045   2.0613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.88373    0.08384 -22.467   <2e-16 ***
age         -0.11341    0.05408  -2.097   0.0360 *
smoke        0.27214    0.12347   2.204   0.0275 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1829.1  on 2147  degrees of freedom
Residual deviance: 1819.9  on 2145  degrees of freedom
```

[4] Zeger, S.L., Liang, K-Y.and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach

[5] Fitzmaurice, G.M. and Laird, N.M. (1993) A likelihood-based method for analyzing longitudinal binary responses, Biometrika 80: 141–151.

This model includes `age` and `smoke` as covariates. The interaction `age:smoke` was not included as the difference of the deviance of the nested models is $1819.9 - 1819.5 = 0.4$, with the quantile of the $\chi^2_1 = 3.84$. Hence, the null hypothesis H_0: coefficient of interaction is zero cannot be rejected.

Now we fit the random effects GLM

```
> fit2 <- glmer(resp ~ age + smoke + (1 | id), data=ohio,
+ family=binomial)
> summary(fit3)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: resp ~ age + smoke + (1 | id)
   Data: ohio

     AIC      BIC   logLik deviance df.resid
  1597.9   1620.6   -794.9   1589.9     2144

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.4027 -0.1802 -0.1577 -0.1321  2.5176

Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 5.49     2.343
Number of obs: 2148, groups:  id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.37395    0.27498 -12.270   <2e-16 ***
age         -0.17676    0.06797  -2.601   0.0093 **
smoke        0.41478    0.28704   1.445   0.1485
```

We note that the fixed effects estimates (intercept, age and smoke) changed in the random effects model compared to the respective estimates in the fixed effects GLM. We also observe that the variance of the random effects b_i is estimate as $\hat{Q} = 5.49$. Although, we cannot formally compare the fixed effects with the random effects model (because they use different estimation methods), if the fixed effects model is a better fit we would expect $\hat{Q} \approx 0$ (or at least $\hat{Q}$ to be small) in the random effects model, because $Q = 0$ reduces the random effects model to the fixed effects model ($b_i = 0$). In this case $\hat{Q} = 5.49$ is quite large and hence we think the random effects is better.

Chapter 7

Two-way contingency tables

7.1 Types of 2-way tables - response/controlled variables

Data often arise in the form of counts, **cross-classified** by various factors (often referred to as categorical variables). The table of cross-classified counts is called a **contingency table**. An important distinction in the analysis of contingency table data is between **response variables** and **controlled variables**. This distinction depends on what totals are known in advance of collecting any data, i.e. the method of sampling as illustrated by the following 2 examples, referred to as Case(a) and Case(b) for the rest of this chapter.

7.1.1 Case(a): Skin cancer (melanoma) data - 2 response variables

Dobson (1990, Example 9.1; 2002, Example 9.3.1) gives the following table from a cross-sectional study of malignant melanoma. The sample size was 400, and site and type (labelled T_1 to T_4 here, T_1: Hutchinson's melanotic freckle; T_2: Superficial spreading melanoma; T_3: Nodular; T_4: Indeterminate) were recorded. Both tumour type and site are response variables because none of the row or column totals were fixed in advance of the data collection. The data is in the 'Mela' dataframe on the .RData workspace on MOLE.

		Site		
Tumour Type	Head and Neck	Trunk	Extremities	Total
T_1	22	2	10	34
T_2	16	54	115	185

T_3	19	33	73	125
T_4	11	17	28	56
Total:	68	106	226	400

7.1.2 Case(b): Flu vaccine data - 1 response and 1 controlled variable

Dobson (1990, Example 9.2; 2002, Example 9.3.2) gives the following table from a prospective controlled trial on a new influenza vaccine. Patients were randomly assigned to the two groups (Placebo, Vaccine), and the response (levels of an antibody found in the blood six weeks after vaccination) was determined. Antibody level is the response and vaccine group is a controlled variable (with totals fixed by experimental design). Note that a large response is good. The data is in the 'vaccine' dataframe on the .RData workspace on MOLE.

	Response			
	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35
Total:	31	26	16	73

7.2 Notation for two-way tables

Denote the row factor by A, with levels i, where $1 \leq i \leq I$, and the column factor by B, with levels j, where $1 \leq j \leq J$; let the observations be y_{ij} and suppose they are the observed values of random variables Y_{ij}. We use the notation $\sum_{ij} y_{ij}$ for $\sum_i \sum_j y_{ij}$ and a . in the subscript of a variable means sum over that subscript so that

- $y_{i.}$ means $\sum_j y_{ij}$
- $y_{..}$ means $\sum_i \sum_j y_{ij}$

Note that where it might cause confusion with this notation, I have omitted full stops at the ends of sentences. Conventions for the observed values and probabilities are given in the tables below. In case(b) we always assume that the rows represent the levels of the controlled variable.

	col 1	col 2	...	col J	Total
row 1	y_{11}	y_{12}	...	y_{1J}	$y_{1.}$
row 2	y_{21}	y_{22}	...	y_{2J}	$y_{2.}$
	.	.	.	.	
	.	.	.	.	
	.	.	.	.	
row I	y_{I1}	y_{I2}	...	y_{IJ}	$y_{I.}$
Total	$y_{.1}$	$y_{.2}$	...	$y_{.J}$	$y_{..}$

	col 1	col 2	...	col J	Total
row 1	π_{11}	π_{12}	...	π_{1J}	$\pi_{1.}$
row 2	π_{21}	π_{22}	...	π_{2J}	$\pi_{2.}$
	.	.	.	.	
	.	.	.	.	
	.	.	.	.	
row I	π_{I1}	π_{I2}	...	π_{IJ}	$\pi_{I.}$
					$\pi_{..}$

7.2.1 Association, Independence and Homogeneity

Usually the interest is in whether the response variables are associated (in which case the response variables are treated symmetrically, as in correlation).

In case(a), the skin cancer data with 2 response variables, the probabilities of interest,π_{ij}, are the joint probabilities $\pi_{ij} = P(A = i, B = j)$ where $\pi_{..} = 1$. Only the total sample size $n = y_{..}$ is fixed. Independence implies $P(A = i, B = j) = P(A = i) \times P(B = j)$ for all i and j, where $\pi_i = P(A = i)$

and $\pi_j = P(B = j)$ are the marginal probabilities of row i and column j.

In case (b), the flu vaccine data with 1 response and 1 controlled variable, the probabilities of interest,π_{ij}, are conditional probabilities: $\pi_{ij} = P(B = j|A = i)$. It follows that $\pi_{i.} = 1$ for $1 \leq i \leq I$. The row totals $y_{i.}$ are fixed, not random. The interest is in whether the probability distribution of the response (antibody level) is the same in each level of the controlled variable (drug group). If it doesn't depend on i then we can write ($\pi_{ij} = \pi_j$). This is known as **homogeneity**.

The analysis here is similar to that which you may have encountered previously in testing for independence or homogeneity, except that the deviance is used here rather than Pearson's X^2, and the present approach allows a richer set of models.

7.3 Distributions for two-way tables

Natural models for the Y_{ij} in cases (a) and (b) are described below. A third model, labelled (c), which at first sight appears unnatural since the total sample size is random, is also given for use in §7.4 and later. We next need the Multinomial distribution: $Mn(n, (\pi_i))$ where $\sum_i \pi_i = 1$. The probability function of the Multinomial distribution is

$$P(y_1, y_2, \ldots, y_r) = \begin{cases} \frac{n!}{y_1!y_2!\ldots y_r!} \prod_i \pi_i^{y_i} & \text{if } \sum_i y_i = n \\ 0 & \text{otherwise.} \end{cases}$$

7.3.1 Case (a): two response variables.

The only fixed quantity is $n = y_{..}$, and the IJ observations have a multinomial distribution

$$\{Y_{ij}\} \sim \mathcal{M}(n, \{\pi_{ij}\}).$$

The likelihood is

$$n! \prod_{i,j} \frac{\pi_{ij}^{y_{ij}}}{y_{ij}!}$$

7.3.2 Case (b): one response variable.

Now the distribution of $\{Y_{ij}\}$ is the **product multinomial**: a product of multinomials, one for each row,

$$\{Y_{ij}\} \sim \mathcal{M}(n_i, \{\pi_{ij}\}), \text{ independent, for } \ i = 1, \ldots, I$$

so that the likelihood is

$$\prod_i \left(n_i! \prod_j \frac{\pi_{ij}^{y_{ij}}}{y_{ij}!} \right)$$

where $n_i = y_{i\cdot}$, the row totals, are fixed.

7.3.3 Case (c): independent Poissons (no fixed margins).

If the Y_{ij} are independent Poisson variables, with mean μ_{ij} then the likelihood is

$$\prod_{i,j} \left(e^{-\mu_{ij}} \frac{(\mu_{ij})^{y_{ij}}}{y_{ij}!} \right).$$

This seems unnatural as the Y_i are not independent (since the total number of observations is always fixed). However this case is needed since it forms the basis of log-linear modelling.

7.3.4 Expected values

It is more convenient in modelling to work with expected values (of the Y_{ij}) rather than probabilities: $E(Y_{ij}) = \mu_{ij}$. So in case (a) $\mu_{ij} = n\pi_{ij}$, and in case (b) $\mu_{ij} = n_i\pi_{ij}$ where for (a) $\sum_{ij} \pi_{ij} = 1$, and for case (b) $\sum_j \pi_{ij} = 1$

7.4 GLMs and two-way contingency tables

As discussed in section 7.3, the distributions that arise in connection with contingency tables are multinomial or product-multinomial, which cannot be expressed directly as a GLM (because the observations are dependent). However, it is possible to analyse such data by using GLMs, by using a Poisson distribution and a log link and including all the controlled variables and their interactions in the linear predictor. The justification for this, which applies to higher-way tables too, is developed in the document 'Treat Product-Multinomial as Poisson' on MOLE. A Poisson model of this kind with a linear predictor and a log link is often called a **log-linear model**.

7.4.1 Natural hypotheses are log-linear models

For case (a), the hypothesis of independence is that the joint probability can be written as $\pi_{ij} = \pi_i\pi_j$, so that, in terms of expected values, $\mu_{ij} = n\pi_i\pi_j$ Taking logs gives

$$\eta_{ij} = \log \mu_{ij} = \mu + \alpha_i + \beta_j \quad \text{(i.e. } A + B\text{)}$$

where $\mu = \log n$, $\alpha_i = \log \pi_i$, $\beta_j = \log \pi_j$

Similarly for case (b): the hypothesis of homogeneity is that the conditional probability $\pi_{ij} = P(B = j|A = i)$ can be written $\pi_{ij} = \pi_j$ ($P(B = j|A = i)$ doesn't depend on the row i, so $\mu_{ij} = n_i \pi_j$, and again

$$\eta_{ij} = \log \mu_{ij} = \mu + \alpha_i + \beta_j \quad \text{(i.e. } A + B\text{)}$$

where now $\mu + \alpha_i = \log n_i$ and $\beta_j = \log \pi_j$

In both cases, the hypothesis of interest implies additivity ($A + B$) for the log mean. Moreover, in both cases the saturated model (unrestricted π_{ij}) can be written

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \quad \text{(i.e. } A * B\text{)} \,.$$

In this case the $(\alpha\beta)_{ij}$ terms are referred to as interaction terms, and are zero under the hypotheses of independence or homogeneity.

So in both cases we wish to see whether the model $A + B$ is adequate. If it is, then the data will be deemed consistent with independence in case (a) or homogeneity in case (b).

Key point reiterated: Multinomial and product-multinomial data may be analysed as though they were independent Poisson data, with a log-link, provided terms corresponding to the fixed margins (including all interactions) are included in the model. Both the estimators and the deviance will be correct.

7.4.2 Poisson log-linear modelling for two-way tables

In all cases the Y_{ij} are modelled as independent Poisson with the log link.

Case(a)
Here there are no controlled variables. The linear predictor including just the intercept is the **minimal model**. The log-linear models that are allowed, along with the corresponding estimates of $\hat{\pi}_{ij}$ (derived in the next section) are:

- $A * B$: the non-independence saturated model, $\hat{\pi}_{ij} = y_{ij}/n$ ($D = X^2 = 0$, df $= 0$)
- $A + B$: independence (probabilities the same in each row, $\hat{\pi}_{ij} = y_{i.}y_{.j}/(n^2)$
- A : independence and equal probabilities for each column $\hat{\pi}_{ij} = y_{i.}/(nJ)$
- B : independence and equal probabilities for each row $\hat{\pi}_{ij} = y_{.j}/(nI)$

- 1 : independence and equal probabilities for all rows and columns $\hat{\pi}_{ij} = 1/(IJ)$

Case(b)
Here row (labelled as factor A and indexed by i) is the controlled variable. So the linear predictor including just A (and the intercept) is the **minimal model**. Now the only (log-linear) models that are allowed are:

- $A * B$: the non-homogeneity, saturated model, $\hat{\pi}_{ij} = y_{ij}/n_i$ ($D = X^2 = 0$, df $= 0$)
- $A + B$: homogeneity (probabilities the same in each row, $\hat{\pi}_{ij} = y_{.j}/n$)
- A : homogeneity and equal probabilities for each column ($\hat{\pi}_{ij} = 1/J$)

7.4.3 Maximum likelihood estimation for π_{ij} in case(a)

The mles for π_{ij} in section 7.4.2 were not justified. Here we derive one of them. Consider the linear predictor containing just A (and the intercept) for case(a). Since $\mu_{ij} = n\pi_{ij}$ we have that $log(\mu_{ij}) = log(n) + log(\pi_{ij})$. If A (indexed by i) is the only term in the linear predictor then π_{ij} can only depend on i, i.e. $\pi_{ij} = \pi_i$

Since we model $Y_{11}, ..., Y_{IJ}$ as **independent** Poisson random variables with mean μ_{ij} if follows that $Y_{ij} \sim Po(\mu_{ij}) \sim Po(n\pi_{ij})$. The likelihood (L) is therefore

$$\begin{aligned}
L &= \prod_{i=1}^{I}\prod_{j=1}^{J}\left\{\frac{\exp(-n\pi_i)[n\pi_i]^{y_{ij}}}{y_{ij}!}\right\} \\
l = log(L) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\{-n\pi_i + y_{ij}log(n\pi_i) - log(y_{ij}!)\} \\
\frac{\partial l}{d\pi_i} &= \sum_{j=1}^{J}\left\{-n + \frac{y_{ij}}{\pi_i}\right\} \\
\frac{\partial l}{d\pi_i} &= 0 \Rightarrow \hat{\pi}_i = \frac{y_{i.}}{nJ}
\end{aligned}$$

7.5 Interaction plots and examination of residuals

Just as with ordinary linear models, it is possible to investigate interactions by means of suitable plots. If there is no interaction, the independence and homogeneity models have $\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$, so that if $\log \mu_{ij}$ is plotted against j for different i, the graphs will be parallel; exactly the same will be true for plots of $\log \mu_{ij}$ against i for different j. If there is interaction, they will not be parallel. However, if $I > 2$ or $J > 2$, additivity may hold over some of the rows and/or columns, even if not over all.

The best estimate of $\log \mu_{ij}$ (under the saturated model) is $\log(y_{ij})$, so that it is graphs of $\log(y_{ij})$, or $\log(y_{ij} + 3/8)$, that are used, and these may show the cause of the interaction — for example that one level of A behaves quite differently from the others. In `R`, `interaction.plot` produces these graphs.

Another approach is through the examination of residuals. Large (absolute) values of the deviance or Pearson residuals $e_{D,ij}$ or $e_{P,ij}$ can show why a hypothesis is not acceptable, since their square makes a large contribution to D or X^2, and perhaps suggest sub-tables over which the hypothesis may hold. Note that the deviance is exactly additive when a hypothesis is partitioned into orthogonal components (mentioned in §3.7.2); this is illustrated in Example 7.6.3.

7.6 Analysis of the skin cancer data (case(a)) using log-linear models

From now on we will refer to modelling multinomial or product-multinomial responses using a poisson response with a log link as log-linear modelling (making sure that all controlled variables are included in the linear predictor). In this section we look at fitted values for various log-linear models applied to the skin cancer data. We show how to enter the data into `R` and how to use the methods we have developed in earlier chapters to choose a suitable model linking the counts to the potential explanantory variables.

7.6.1 Fitted values for the skin cancer data (case(a))

With two response variables, the general form of μ_{ij} is $\mu_{ij} = n\pi_{ij}$
According to which terms are entered into the linear predictor, π_{ij} changes accordingly.

Key Point
Essentially when a variable is included in the linear predictor, it fixes the marginal totals of that variable. So in the skin cancer data, if tumour type (row variable) is included in the linear predictor then the marginal (row) totals of the fitted values must be the same as the marginal (row) totals of the observed values.

The fitted values for this two response case are given below using the results of section 7.4.2.

- $A * B$: no restrictions on π_{ij}; $\hat{\mu}_{ij} = y_{ij}$

	Site			
Tumour Type	Head and Neck	Trunk	Extremities	Total
T_1	22	2	10	34
T_2	16	54	115	185
T_3	19	33	73	125
T_4	11	17	28	56
Total:	68	106	226	400

- $A + B$: independence; $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = y_{i.}y_{.j}/n$

	Site			
Tumour Type	Head and Neck	Trunk	Extremities	Total
T_1	5.8	9.0	19.2	34
T_2	31.5	49.0	104.5	185
T_3	21.3	33.1	70.6	125
T_4	9.5	14.8	31.7	56
Total:	68	106	226	400

- A : independence and the same probability for each column category: $\hat{\mu}_{ij} = n\hat{\pi}_{ij} = y_{i.}/J$

	Site			
Tumour Type	Head and Neck	Trunk	Extremities	Total
T_1	11.33	11.33	11.33	34
T_2	61.67	61.67	61.67	185
T_3	41.67	41.67	41.67	125
T_4	18.67	18.67	18.67	56
Total:	133.33	133.33	133.33	400

7.6.2 Log-linear modelling in `R`

To read the data into `R`, the skin cancer data should be presented in the following way

```
count type site
22 1 1
2 1 2
```

```
.    .    .
.    .    .
28  4  3
```

Remembering that the explanatory variables are factors we can perform the analysis in `R` using commands of the form `glm(count~factor(type)*factor(site),poisson(log))`. This command fits the saturated/maiximal model.

7.6.3 Skin cancer data (case(a)) revisited

Consider the skin cancer data.

1. The test of independence based on the log-linear model $A + B$ has $D = 51.795$ on 6 df.
2. The usual Pearson-χ^2 test for independence gives $X^2 = 65.813$ on 6 df. (Use the `chisq.test` function in `R` applied to the table of frequencies `matrix(Mela$number,nrow=4)`.
3. Each test provides overwhelming evidence of dependence. There are now several ways to examine the data to see whether the dependence can be described fairly simply.
4. Examine the row or column proportions. Note the different pattern for T_1 and for 'Head and Neck'.
5. Examine the residuals from the independence model; plotting them against one factor and labelling by the other again shows that T_1 and 'Head and Neck' look unusual. It is seen that the largest (in modulus) residual is $\mathrm{e}_{P,11} = 6.75$ (Pearson) or $\mathrm{e}_{D,11} = 5.14$ (deviance); squaring these shows that this cell makes a very large contribution to X^2 and D.
6. The interaction plot of $\log(y_{ij})$ for tumour type and tumour site again shows that the $(1, 1)$ cell looks like the source of the dependence.
7. In summary so far, all of these show that the first row and the first column differ from the others, and the $(1, 1)$ cell has a much larger value than expected under independence ($y_{11} = 22$, $\hat{\mu}_{11} = 5.78$). A few other residuals in the first row or first column are larger than 2 in modulus.
8. Removing Type T_1, the test of independence on the remaining 3×3 table has $D = 6.509$ ($X^2 = 6.562$) on 4 df, showing that independence is acceptable. Similarly, removing Head & Neck, the test of independence on the remaining 4×2 table has $D = 2.165$ ($X^2 = 2.025$) on 3 df, showing that independence is acceptable.

9. The difference of the first row or first column from the rest can be confirmed by comparing these with the pooled values from the rest. For rows, pooling gives the 2×3 table

	H&N	Tr	Ex
T_1	22	2	10
T_2,T_3,T_4	46	104	216

The test of independence on this table has $D = 45.286$ ($X^2 = 60.532$) on 2 df, showing strong evidence against independence. Note that this deviance (45.286) added to the deviance from the test for independence on types T_2, T_3, T_4 alone (6.509) exactly gives the deviance for all 4 rows (51.795) (the corresponding contrasts are orthogonal).

10. Similarly, comparing column 1 with the pooled frequencies from columns 2 and 3 gives a 4×2 table with $D = 49.630$ ($X^2 = 64.548$) on 3 df, showing strong evidence against independence. Again, this D (49.630) added to the D from the test on columns 2 and 3 (2.165) exactly gives D for all 3 columns (51.795).

11. To see if it is just the $(1,1)$ cell which is causing the association, a log-linear model can be fitted which is additive in the factors, but includes a term for the $(1,1)$ cell — an indicator variable for that cell (that is, treats it as an outlier). This has $D = 8.002$ ($X^2 = 7.882$) on 5 df, showing that the evidence against independence is attributable to the high count in the $(1,1)$ cell.

12. Overall, there is overwhelming evidence of association between tumour type and site; there is no reason to doubt that types T_2, T_3 and T_4 have a similar distribution over the body, but the distribution of T_1 over the body is different, with a higher concentration on Head and Neck. (Note that, even in sampling situation (a) conclusions may be easiest to understand when phrased as statements about conditional distributions.)

7.7 Flu vaccine data (case(b)) revisited

1. The minimal model A has $D = 23.68$ on 4 df. $\chi^2_{4,0.95} = 9.49$ so not a good fit.
2. The homogeneity model (A+B) has $D = 18.643$ on 2 df.
3. $\Delta D = 5.04$ on 2df so not much of an improvement ($\chi^2_{2,0.95} = 5.99$)
4. This analysis strongly suggests that groups differ in their response. The observed row proportions show that the vaccine has a much better (larger) response.

7.7.1 Fitted values for the $A + B$ model for the flu data (case(b))

$\hat{\pi}_{11} = y_{.1}/n = 31/73$ and $\hat{\mu}_{11} = n_1\hat{\pi}_{11} = 38 \times 31/73 = 16.14$
The table of estimated values for π_{ij} and μ_{ij} is

	Small	Moderate	Large	Total
Placebo	$\hat{\pi}_{11} = 31/73$	$\hat{\pi}_{12} = 26/73$	$\hat{\pi}_{13} = 16/73$	1
Vaccine	$\hat{\pi}_{21} = 31/73$	$\hat{\pi}_{22} = 26/73$	$\hat{\pi}_{23} = 16/73$	1
Placebo	$\hat{\mu}_{11} = 16.14$	$\hat{\mu}_{12} = 13.53$	$\hat{\mu}_{13} = 8.33$	38
Vaccine	$\hat{\mu}_{21} = 14.86$	$\hat{\mu}_{22} = 12.47$	$\hat{\mu}_{23} = 7.67$	35
Total	$y_{.1} = 31$	$y_{.2} = 26$	$y_{.3} = 16$	