MTH6134 Statistical Modelling II
Notes

Autumn 2023

[1]

1 Introduction

1.1 Topics to be covered

In the course Statistical Modelling I, regression models were studied in which there was a **dependent variable**, Y, and $p-1$ **explanatory variables**, $X_1, \ldots, X_{p-1}$. The variable Y was continuous quantitative and the Xs were quantitative. The focus of the course was the **general linear model** given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the vector of responses, $\mathbf{X}$ is the $n \times p$ **design matrix**, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ is the parameter vector and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is the error vector. Note that the number of columns p of $\mathbf{X}$ is greater than the number of explanatory variables $p-1$, since the matrix has an additional column in which all elements are equal to one. It was usually assumed that $\boldsymbol{\epsilon}$ is multivariate normal with zero mean vector and covariance matrix $\sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ denotes the identity matrix of order n.

In this course, we consider the case where the Ys have error distribution models other than a normal distribution. We first review the normal linear model from the perspective of the likelihood. We then meet the exponential family of distributions and introduce generalised linear models, a flexible generalisation of ordinary linear regression. It will then be shown how these models may be fitted to binary response data and count data, leading to logistic regression and Poisson regression, respectively. Finally, we show how to model survival data. Throughout the course, the statistical computing package R will be used to illustrate the main ideas.

1.2 Examples of generalised linear models

Each of the examples below can be formulated in the context of a generalised linear model.

Example 1.1 Binary response data.
Suppose that $Y_i \sim \text{Bin}(r_i, \pi_i)$ for $i = 1, 2, \ldots, n$, all independent, where π_i depends on a known covariate x_i. For example, in a clinical trial, r_i may be the number of patients given

[1] Lecture notes provided by S Coad (formerly of QMUL).

a dose x_i of a new drug and Y_i is the number of these giving a positive response. In this case, we can consider the logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 x_i.$$

Since $E(Y_i/r_i) = \pi_i$, as we will see in Chapter 4, this means that we are using a logit link function in a generalised linear model.

Example 1.2 Count data.
Suppose that $Y_i \sim \text{Poisson}(\mu_i)$ for $i = 1, 2, \ldots, n$, all independent, where μ_i depends on a known covariate x_i, For example, Y_i may be the number of AIDS deaths in month x_i. In this case, we can consider the Poisson regression model

$$\log(\mu_i) = \beta_0 + \beta_1 x_i.$$

Since $E(Y_i) = \mu_i$, as we will see in Chapter 5, this means that we are using a log link in a generalised linear model.

Example 1.3 Normal response data.
Suppose that $Y_i \sim \text{N}(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent, where μ_i depends on a known covariate x_i. For example, Y_i may be the level of carbon dioxide at Mauna Loa, an extinct volcano in Hawaii, in year x_i. In this case, we can consider the linear regression model

$$\mu_i = \beta_0 + \beta_1 x_i.$$

Since $E(Y_i) = \mu_i$, as we will see in Chapter 2, this means that we are using an identity link in a generalised linear model.

2 Normal Linear Model

2.1 Likelihood

Suppose that the random variables $Y_1, \ldots, Y_n$ have a joint distribution which is specified, except for the unknown parameters $\theta_1, \ldots, \theta_p$. As we will see later, in the context of the normal linear model, these parameters will consist of the regression coefficients and the error variance. In Statistical Modelling I, the method of least squares was used to estimate the unknown parameters. Here, an alternative method of estimation is introduced, which is more general and has better properties.

Definition 2.1 *For discrete data $y_1, \ldots, y_n$, the* **likelihood**, *$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n)$, is the* **joint probability mass function** *of the data, that is,*

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = P(Y_1 = y_1, \ldots, Y_n = y_n).$$

For continuous data $y_1, \ldots, y_n$, the likelihood, $L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n)$, is the **joint probability density function** *of the data, that is,*

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n).$$

If the Y values are **independent**, then for discrete distributions we have

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = \prod_{i=1}^{n} P(Y_i = y_i);$$

and for the continuous distributions we have

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

For all of the models in this course, the Y values will be assumed independent.

Example 2.2 Simple linear regression.
Suppose that $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent. Then the likelihood is

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}. \end{aligned}$$

For notational convenience, we will sometimes write the data as a vector, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, and write the parameters as a vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$. So we write the likelihood as $L(\boldsymbol{\theta}; \mathbf{y})$.

2.2 Method of maximum likelihood

The maximum likelihood estimates of $\theta_1, \ldots, \theta_p$ are the values of $\theta_1, \ldots, \theta_p$ which maximise the likelihood $L(\boldsymbol{\theta}; \mathbf{y})$. The maximum likelihood estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top$, where

$$\hat{\theta}_j = \hat{\theta}_j(\mathbf{Y}), \quad j = 1, 2, \ldots, p,$$

We usually find the maximum likelihood estimates by maximising the **log-likelihood** $\ell(\boldsymbol{\theta}; \mathbf{y}) = \log L(\boldsymbol{\theta}; \mathbf{y})$ and then solving the **likelihood equations**

$$\frac{\partial \ell}{\partial \theta_j} = 0, \quad j = 1, 2, \ldots, p.$$

Example 2.3 Simple linear regression revisited.
The log-likelihood is

$$\ell(\beta_0, \beta_1, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Thus, we have

$$\frac{\partial \ell}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial \ell}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$$

and

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Setting the first two derivatives to zero, we obtain

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

and

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0.$$

Now, the first of these equations yields

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Substituting this equation into the previous one, we have

$$\sum_{i=1}^{n} x_i y_i - (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,$$

which may be rearranged to give

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Finally, setting the third derivative to zero, we obtain

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Note that the maximum likelihood estimates of β_0 and β_1 are the same as the least squares estimates that were obtained in Statistical Modelling I. It was shown there that the least squares estimators of β_0 and β_1 are unbiased. However, since $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/\sigma^2 \sim \chi^2_{n-2}$, $\hat{\sigma}^2$ is a biased estimate of σ^2: an unbiased one is $s^2 = n\hat{\sigma}^2/(n-2)$.

Example 2.4 Manatee data set.
Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Below are data on powerboat registrations (x), in thousands, and the number of manatees killed by boats (y) in Florida in the years 1977 to 1987.

Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
x	447	460	481	498	513	512	526	559	585	614	645
y	13	21	24	16	24	20	15	34	33	33	39

The data give $\sum_{i=1}^{11} x_i = 5,840$, $\sum_{i=1}^{11} y_i = 272$, $\sum_{i=1}^{11} x_i y_i = 149,153$ and $\sum_{i=1}^{11} x_i^2 = 3,140,490$. So we have

$$\hat{\beta}_1 = \frac{149,153 - 5,840 \times 272/11}{3,140,490 - 5,840^2/11} = 0.1187$$

and

$$\hat{\beta}_0 = \frac{272}{11} - 0.1187 \times \frac{5,840}{11} = -38.29.$$

It follows that the least squares regression line is

$$\hat{y}_i = \hat{\mu}_i = -38.29 + 0.1187 x_i.$$

2.3 Asymptotic distribution of the maximum likelihood estimator

The following result, stated without proof, gives the **asymptotic distribution** of $\hat{\boldsymbol{\theta}}$.

Theorem 2.1 *Under fairly general conditions, for large n, $\hat{\boldsymbol{\theta}} \sim \mathrm{N}_p(\boldsymbol{\theta}, V^{-1})$, where V is the $p \times p$ (expected)* **Fisher information matrix** *with (i,j)-th element*

$$v_{ij} = E\left\{-\frac{\partial^2 \ell(\boldsymbol{\theta}; \mathbf{Y})}{\partial\theta_i \partial\theta_j}\right\}$$

for $i, j = 1, 2, \ldots, p$.

A consequence of this result is that every component of $\hat{\boldsymbol{\theta}}$ is asymptotically normal, that is, for large n, $\hat{\theta}_j \sim \mathrm{N}(\theta_j, v^{jj})$ for $j = 1, 2, \ldots, p$, where v^{jj} denotes the jth diagonal element of V^{-1}.

Example 2.5 Simple linear regression revisited. In this case, the Fisher information matrix (FIM) V is 3×3. For the diagonal entries of this matrix, we have

$$\frac{\partial^2 \ell}{\partial \beta_0^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 \ell}{\partial \beta_1^2} = -\frac{\sum_{i=1}^n x_i^2}{\sigma^2} \text{ and } \frac{\partial^2 \ell}{\partial \sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Off diagonal entries of the FIM require computing crossed derivatives

$$\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} = -\frac{\sum_{i=1}^n x_i}{\sigma^2}, \quad \frac{\partial^2 \ell}{\partial \beta_0 \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$

and

$$\frac{\partial^2 \ell}{\partial \beta_1 \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i).$$

After reversing signs and taking expectations, we have $v_{11} = n/\sigma^2$, $v_{22} = \sum_{i=1}^n x_i^2/\sigma^2$, $v_{33} = n/(2\sigma^4)$, $v_{12} = \sum_{i=1}^n x_i/\sigma^2$ and $v_{13} = v_{23} = 0$. Note that, formally speaking, for the expectation we require the notation Y_i rather than y_i. Collecting these results, we have

$$V = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{\sum_{i=1}^n x_i}{\sigma^2} & 0 \\ \frac{\sum_{i=1}^n x_i}{\sigma^2} & \frac{\sum_{i=1}^n x_i^2}{\sigma^2} & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Hence, we have

$$V^{-1} = \begin{pmatrix} \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}} & -\frac{\sigma^2 \overline{x}}{S_{xx}} & 0 \\ -\frac{\sigma^2 \overline{x}}{S_{xx}} & \frac{\sigma^2}{S_{xx}} & 0 \\ 0 & 0 & \frac{2\sigma^4}{n} \end{pmatrix},$$

where $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$. This shows that, for large n, $\hat{\beta}_0 \sim \mathrm{N}\{\beta_0, \sigma^2 \sum_{i=1}^n x_i^2/(nS_{xx})\}$, $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2/S_{xx})$ and $\hat{\sigma}^2 \sim \mathrm{N}(\sigma^2, 2\sigma^4/n)$. Note that, from Statistical Modelling I, the distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are exact.

2.4 Multiple linear regression

So far, we have only applied the likelihood theory to the simple linear regression model in which there is a single explanatory variable. Let us now consider multiple linear regression, where there are $p-1$ explanatory variables. This means that $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent, where

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{p-1} x_{p-1,i} = \mathbf{x}_i^\top \boldsymbol{\beta}$$

and $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{p-1,i})^\top$.

The likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\right\}. \end{aligned}$$

Now, since

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{X}$ is the $n \times p$ design matrix with ith row $\mathbf{x}_i^\top$, we may write

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}.$$

So the log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

We have the following derivatives of the log-likelihood

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{and} \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Assume that $\text{rank}(\mathbf{X}) = p$, so that the $p \times p$ matrix $\mathbf{X}^\top\mathbf{X}$ is non-singular. Then, setting the above derivatives to zero, we obtain the maximum likelihood estimates

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

Again, the maximum likelihood estimate of $\boldsymbol{\beta}$ is the same as the least squares estimate that was obtained in Statistical Modelling I. An unbiased estimate of σ^2 is $s^2 = n\hat{\sigma}^2/(n-p)$.

We now compute the Fisher information matrix (FIM). We have diagonal terms

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} \quad \text{and} \quad \frac{\partial^2 \ell}{\partial \sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

as well as cross derivative term

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \sigma^2} = -\frac{1}{\sigma^4}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

After sign reversal and taking expectations, the Fisher information matrix is

$$V = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \mathbf{0} \\ \mathbf{0}^\top & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Hence, we have

$$V^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}^\top & \frac{2\sigma^4}{n} \end{pmatrix}.$$

This shows that, for large n, $\hat{\boldsymbol{\beta}} \sim \text{N}_p\{\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\}$ and $\hat{\sigma}^2 \sim \text{N}(\sigma^2, 2\sigma^4/n)$. Note that, from Statistical Modelling I, the distribution of $\hat{\boldsymbol{\beta}}$ is exact.

2.5 Generalised likelihood ratio tests

We assume that $Y_1, \ldots, Y_n$ have a joint distribution which depends on the unknown parameters $\theta_1, \ldots, \theta_p$. The set of all possible values of $\boldsymbol{\theta}$ is called the **parameter space** Ω.

A hypothesis restricts $\boldsymbol{\theta}$ to lie in ω, where $\omega \subset \Omega$. If we wish to test whether $\boldsymbol{\theta} \in \omega$, then we test the **null hypothesis** $H_0 : \boldsymbol{\theta} \in \omega$ against the **alternative hypothesis** $H_1 : \boldsymbol{\theta} \in \Omega\backslash\omega$. If ω is a single point, the hypothesis is **simple**. Otherwise, the hypothesis is **composite**. The set of all possible values $\mathbf{y}$ of the random vector $\mathbf{Y}$ is called the **sample space** S.

Example 2.6 Simple linear regression revisited. Suppose that $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent, and that we wish to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Then $\Omega = \mathbb{R}^2 \times \mathbb{R}^+$, $\omega = \mathbb{R} \times \{0\} \times \mathbb{R}^+$ and $S = \mathbb{R}^n$.

The **critical region** is the subset $R \subseteq S$ such that we reject H_0 if and only if $\mathbf{y} \in R$. A general approach to finding R is now presented.

Definition 2.7 *The generalised likelihood ratio test has critical region* $R = \{\mathbf{y} : \Lambda(\mathbf{y}) < a_\alpha\}$, *where*

$$\Lambda(\mathbf{y}) = \frac{\max_{\boldsymbol{\theta} \in \omega} L(\boldsymbol{\theta}; \mathbf{y})}{\max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}; \mathbf{y})}$$

is the **generalised likelihood ratio** *and* a_α *is a constant chosen to give a test with significance level* α. *Clearly,* $0 \leq \Lambda(\mathbf{y}) \leq 1$, *since* $\omega \subset \Omega$.

Let $\hat{\boldsymbol{\theta}}_0$ be the value of $\boldsymbol{\theta}$ which maximises the likelihood $L(\boldsymbol{\theta}; \mathbf{y})$ in ω. Then we may write

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\boldsymbol{\theta}}_0; \mathbf{y})}{L(\hat{\boldsymbol{\theta}}; \mathbf{y})},$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$. We call $\hat{\boldsymbol{\theta}}_0$ the **restricted maximum likelihood estimate** of $\boldsymbol{\theta}$ under H_0.

Example 2.8 Simple linear regression: testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Recall that the model is $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i = 1, 2, \ldots, n$, all independent with likelihood

$$L(\beta_0, \beta_1, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\}.$$

For the development below, we require maximum likelihood estimates of the model parameters β_0, β_1 and σ^2, computed under two cases: unrestricted and restricted to H_0. Unrestricted maximum likelihood estimates are $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2/n$, where $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$; restricted maximum likelihood estimates $\hat{\beta}_{00} = \overline{y}$, $\hat{\beta}_{10} = 0$ and $\hat{\sigma}_0^2 = \sum_{i=1}^n (y_i - \overline{y})^2/n$. The generalised likelihood ratio is

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\beta}_{00}, \hat{\beta}_{10}, \hat{\sigma}_0^2; \mathbf{y})}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2; \mathbf{y})} = \frac{(2\pi\hat{\sigma}_0^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\hat{\sigma}_0^2} \sum_{i=1}^n (y_i - \hat{\beta}_{00} - \hat{\beta}_{10} x_i)^2 \right\}}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \right\}} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{\frac{n}{2}}.$$

The likelihood ratio $L(\bar{y}, 0, \sum_{i=1}^n (y_i - \bar{y})^2)/L(\bar{y} - \hat{\beta}_1 \bar{x}, s_{xy}/s_{xx}, \sum_{i=1}^n (y_i - \bar{y})^2) = (\hat{\sigma}^2/\hat{\sigma}_0^2)^{\frac{n}{2}}$ can be written using $\hat{\sigma}^2 = SS_E/n$ and $\hat{\sigma}_0^2 = SS_{TC}/n$, where SS_E is the sum of squares of the error and SS_{TC} is the total corrected sum of squares. The development is as follows

$$\left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} \right)^{\frac{n}{2}} = \left\{ \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2} \right\}^{\frac{n}{2}} = \left[\frac{\sum_{i=1}^n \{(y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})\}^2}{\sum_{i=1}^n (y_i - \overline{y})^2} \right]^{\frac{n}{2}} = \left[\frac{SS_E}{SS_{TC}} \right]^{\frac{n}{2}}.$$

Since the critical region is $R = \{\mathbf{y} : \Lambda(\mathbf{y}) < a_\alpha\}$, we reject H_0 if $\left[\frac{SS_E}{SS_{TC}} \right]^{\frac{n}{2}} < a_\alpha$. We next transform the previous inequality into an inequality involving the known F ratio MS_R/MS_E of linear regression. We use the fact that the total corrected sum of squares satisfies $SS_{TC} =$

$SS_E + SS_R$, where SS_R is the sum of squares of the regression. The starting point is transforming the inequality above to

$$\frac{SS_E}{SS_E + SS_R} < a_\alpha^{2/n},$$

and further manipulation (details left as exercise) leads into the inequality with F ratio

$$\frac{MS_R}{MS_E} = \frac{SS_R/1}{SS_E/(n-2)} > (n-2)\left(a_\alpha^{-2/n} - 1\right).$$

Under H_0, the distribution of MS_R/MS_E above is well known from regression theory. It is the ratio of two chi-square distributions divided by their degrees of freedom, hence $MS_R/MS_E \sim F_{1,n-2}$. If $F_{1,n-2,\alpha}$ is the upper α quantile of $F_{1,n-2}$ distribution, we reject H_0 if

$$\frac{MS_R}{MS_E} > F_{1,n-2,\alpha}.$$

We manipulated the likelihood ratio $\Lambda(\mathbf{y})$ and its critical region to show equivalence with the F test for $H_0 : \beta_1 = 0$ in regression. It can be shown that all the standard tests in normal-theory problems are generalised likelihood ratio tests.

Example 2.9 Manatee example revisited. The test statistic is $MS_R/MS_E = 24.723$. As $F_{1,n-2,0.05} = 5.117$, we reject $H_0 : \beta_1 = 0$ and conclude that $\beta_1 \neq 0$.

Equivalently, by inversion of the above expression for MS_R/MS_E, we retrieve $\Lambda(\mathbf{y}) = 6.744 \times 10^{-8}$, and the critical region are all those values smaller than 0.003598 (using the same inversion). The conclusion remains unchanged and we reject H_0.

2.6 Wilks' theorem

In more complex cases, we cannot obtain the exact distribution of the generalised likelihood ratio, $\Lambda(\mathbf{Y})$. Instead, we use the following result, stated without proof, which gives the asymptotic distribution of $-2\log\{\Lambda(\mathbf{Y})\}$ under H_0.

Theorem 2.2 Wilks' theorem.
Suppose that $Y_1, \ldots, Y_n$ have a joint distribution depending on the parameters $\theta_1, \ldots, \theta_p$ and consider testing $H_0 : \boldsymbol{\theta} \in \omega$ against $H_1 : \boldsymbol{\theta} \in \Omega \backslash \omega$, where $\dim(\omega) = p_0$. Then, under regularity conditions, when H_0 is true and n is large, $-2\log\{\Lambda(\mathbf{Y})\} \sim \chi^2_s$, where $s = p - p_0$ is the number of **constraints** *imposed by H_0.*

By the above result, for large n, the critical region for a test with approximate significance level α is $R = \{\mathbf{y} : -2\log\{\Lambda(\mathbf{y})\} > \chi^2_{s,\alpha}\}$. Note that p_0 is the number of unknown parameters under H_0.

Example 2.10 Simple linear regression. For testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, we have $\theta = (\beta_0, \beta_1, \sigma^2)$ so $p = 3$ and $p_0 = 2$. The asymptotic distribution of $-2\log\{\Lambda(\mathbf{y})\}$ is χ^2_1 as $s = p - p_0 = 3 - 2 = 1$. In this case, the exact distribution of $\Lambda(\mathbf{Y})$ can be obtained.

As we will see in Chapter 3, the above result provides the basis for the analysis of deviance, which is used to assess the fit of a generalised linear model and to compare alternative models. In the normal case considered in Statistical Modelling I, the analysis of deviance reduces to the analysis of variance, where the distributions of the test statistics are exact.

3 Generalised Linear Models

3.1 Exponential families

There is a class of distributions which includes the normal, Poisson, binomial, gamma, chi-squared, exponential and others.

Definition 3.1 *The random variable Y with parameters $\theta_1, \ldots, \theta_p$ has a distribution in the* **exponential family** *if its range does not depend on the parameters and its probability mass function or probability density function can be written in the form*

$$f_Y(y; \boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^{p} a_j(y) b_j(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) + d(y)\right\}.$$

Note that, for the one-parameter exponential family, this reduces to

$$f_Y(y; \theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\}.$$

If $a(y) = y$, then the distribution is in **canonical form** and $b(\theta)$ is called the **natural parameter**. All of the distributions that we consider in this course will be in canonical form.

Example 3.2 Poisson distribution.
Suppose that $Y \sim \text{Poisson}(\mu)$. Then we may write

$$\begin{aligned} f_Y(y; \mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \exp\{y \log \mu - \mu - \log(y!)\}. \end{aligned}$$

Thus, we have $a(y) = y$, $b(\mu) = \log \mu$, $c(\mu) = -\mu$ and $d(y) = -\log(y!)$. It follows that the distribution is in canonical form and $\log \mu$ is the natural parameter.

The following result gives the mean and variance of $a(Y)$.

Lemma 3.3 *Suppose that Y has a distribution in the one-parameter exponential family. Then*

$$E\{a(Y)\} = -\frac{c'(\theta)}{b'(\theta)}$$

and

$$\text{Var}\{a(Y)\} = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3}.$$

Proof We prove the result in the continuous case only. The proof in the discrete case is similar. By definition, we know that

$$\int_S \exp\{a(y)b(\theta) + c(\theta) + d(y)\} dy = 1,$$

where S is the sample space. Differentiating both sides with respect to θ gives

$$\int_S \{a(y)b'(\theta) + c'(\theta)\} f_Y(y; \theta) dy = 0,$$

which implies that

$$E\{a(Y)b'(\theta) + c'(\theta)\} = 0.$$

It follows that $E\{a(Y)\} = -c'(\theta)/b'(\theta)$. Continuing, differentiating the penultimate equation with respect to θ yields

$$\int_S \left[\{a(y)b'(\theta) + c'(\theta)\}^2 + \{a(y)b''(\theta) + c''(\theta)\}\right] f_Y(y;\theta)dy = 0,$$

which implies that

$$E\left[\{a(Y)b'(\theta) + c'(\theta)\}^2 + \{a(Y)b''(\theta) + c''(\theta)\}\right] = 0.$$

Since $E\{a(Y)\} = -c'(\theta)/b'(\theta)$, this becomes

$$\{b'(\theta)\}^2 E\left[\{a(Y)\}^2\right] - \{c'(\theta)\}^2 - \frac{b''(\theta)c'(\theta)}{b'(\theta)} + c''(\theta) = 0,$$

so that

$$E\left[\{a(Y)\}^2\right] = \frac{\{c'(\theta)\}^2}{\{b'(\theta)\}^2} + \frac{b''(\theta)c'(\theta)}{\{b'(\theta)\}^3} - \frac{c''(\theta)}{\{b'(\theta)\}^2}.$$

Hence, we obtain

$$\begin{aligned}
\mathrm{Var}\{a(Y)\} &= E\left[\{a(Y)\}^2\right] - [E\{a(Y)\}]^2 \\
&= \frac{\{c'(\theta)\}^2}{\{b'(\theta)\}^2} + \frac{b''(\theta)c'(\theta)}{\{b'(\theta)\}^3} - \frac{c''(\theta)}{\{b'(\theta)\}^2} - \frac{\{c'(\theta)\}^2}{\{b'(\theta)\}^2} \\
&= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{\{b'(\theta)\}^3},
\end{aligned}$$

which completes the proof. □

Example 3.4 Poisson example revisited.
Suppose that $Y \sim \text{Poisson}(\mu)$. Then we know that $b(\mu) = \log\mu$ and $c(\mu) = -\mu$. It follows that $b'(\mu) = 1/\mu$, $b''(\mu) = -1/\mu^2$, $c'(\mu) = -1$ and $c''(\mu) = 0$. So we have

$$E(Y) = -\frac{-1}{1/\mu} = \mu$$

and

$$\mathrm{Var}(Y) = \frac{(-1/\mu^2 \times -1) - (0 \times 1/\mu)}{1/\mu^3} = \mu,$$

which confirms what we know.

Now suppose that the random variables $Y_1, \ldots, Y_n$ are independent with distributions from the **same** subfamily of the exponential family with parameter θ. Then the likelihood is

$$\begin{aligned}
L(\theta;\mathbf{y}) &= \prod_{i=1}^{n} \exp\{a(y_i)b(\theta) + c(\theta) + d(y_i)\} \\
&= \exp\left\{\sum_{i=1}^{n} a(y_i)b(\theta) + nc(\theta) + \sum_{i=1}^{n} d(y_i)\right\}.
\end{aligned}$$

As we will see in the next section, in a generalised linear model, the distribution of each Y_i is in canonical form and depends on a single parameter θ_i.

3.2 The generalised linear model

The idea of a generalised linear model was introduced to unify many statistical methods involving linear combinations of parameters.

Assume that the random variables $Y_1, \ldots, Y_n$ are independent with distributions from the same subfamily of the exponential family with the following properties. The distribution of each Y_i is in canonical form and depends on a single parameter θ_i, with the θ_i not necessarily being the same. Then the likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}) &= \prod_{i=1}^{n} \exp\{y_i b(\theta_i) + c(\theta_i) + d(y_i)\} \\ &= \exp\left\{\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right\}, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$.

The θs themselves are not usually of direct interest, since there is one for each observation. For a generalised linear model, we consider a smaller set of parameters $\beta_1, \ldots, \beta_p$, where $p < n$. We suppose that

$$g(\mu_i) = \boldsymbol{\beta}^\top \mathbf{x}_i,$$

where $\mu_i = E(Y_i)$, g is a monotonic differentiable function called the **link function**, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. In what follows, we write $\eta_i = g(\mu_i)$.

In normal linear models, the link function is the identity. This makes sense because μ_i and $\boldsymbol{\beta}^\top \mathbf{x}_i$ can take any values on the real line. However, if we are dealing with counts, where the Poisson distribution is appropriate, then $\mu_i > 0$. In this case, if we take $\eta_i = \log(\mu_i)$, which implies that $\mu_i = e^{\eta_i} > 0$, then this would seem to be a better choice of link function than taking $\eta_i = \mu_i$, the identity link. For the binomial distribution, if we consider the observations as proportions, then they must lie between 0 and 1. Therefore, we look for link functions that map $(0, 1)$ onto the real line. Three commonly used link functions are the logit link

$$\eta = \log\left(\frac{\pi}{1-\pi}\right),$$

the probit link

$$\eta = \Phi^{-1}(\pi),$$

where Φ denotes the standard normal distribution function, and the complementary log-log link

$$\eta = \log\{-\log(1-\pi)\}.$$

3.3 Fitting the model

We now show how to obtain the maximum likelihood estimates of the parameters $\beta_1, \ldots, \beta_p$ in a generalised linear model.

The log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i).$$

We know that

$$E(Y_i) = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}$$

from Lemma 3.3 and

$$g(\mu_i) = \boldsymbol{\beta}^\top \mathbf{x}_i = \sum_{j=1}^{p} x_{ij}\beta_j = \eta_i,$$

where g is a monotonic differentiable function. Also, again from Lemma 3.3,

$$\mathrm{Var}(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{\{b'(\theta_i)\}^3}.$$

Now, we have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j},$$

where

$$\ell_i = y_i b(\theta_i) + c(\theta_i) + d(y_i).$$

By the chain rule, we can write

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}.$$

We consider each of these partial derivatives in turn. First, we have

$$\frac{\partial \ell_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i).$$

Next, we see that

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{-b'(\theta_i)c''(\theta_i) + c'(\theta_i)b''(\theta_i)}{\{b'(\theta_i)\}^2} = b'(\theta_i)\mathrm{Var}(Y_i).$$

Finally, again by the chain rule, we have

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}\frac{\partial \mu_i}{\partial \eta_i}.$$

Hence, we obtain

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}.$$

It follows that

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}.$$

To obtain the maximum likelihood estimates of $\beta_1, \ldots, \beta_p$, we need to solve the likelihood equations $\partial \ell / \partial \beta_j = 0$ for $j = 1, 2, \ldots, p$. In general, these are non-linear and they need to be solved numerically by iteration.

One approach is to use the **Newton-Raphson method**. Here, the mth approximation $\boldsymbol{\beta}^{(m)}$ is given by

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} - \left\{\mathbf{H}^{(m-1)}\right\}^{-1} \left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m-1)}},$$

where $\mathbf{H}^{(m-1)}$ is the **Hessian matrix** of ℓ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}$, which has (j,k)th element

$$h_{jk}^{(m-1)} = \left.\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m-1)}}$$

for $j, k = 1, 2, \ldots, p$. An alternative procedure, which is sometimes simpler, is **Fisher's method of scoring**. It involves replacing the Hessian matrix by its expected value, which is minus the Fisher information matrix V. The following result, stated without proof, shows how V may be calculated from $\partial \ell / \partial \beta_j$ for $j = 1, 2, \ldots, p$.

Theorem 3.1 *Suppose that the random variables $Y_1, \ldots, Y_n$ have distributions depending on the parameters $\beta_1, \ldots, \beta_p$ and that their ranges do not depend on the parameters. Then*

$$E\left\{-\frac{\partial^2 \ell(\boldsymbol{\beta}; \mathbf{Y})}{\partial \beta_j \partial \beta_k}\right\} = E\left\{\frac{\partial \ell(\boldsymbol{\beta}; \mathbf{Y})}{\partial \beta_j} \frac{\partial \ell(\boldsymbol{\beta}; \mathbf{Y})}{\partial \beta_k}\right\}$$

for $j, k = 1, 2, \ldots, p$.

Using the above result, we can write

$$\begin{aligned} E\left(-\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) &= E\left(\frac{\partial \ell_i}{\partial \beta_j} \frac{\partial \ell_i}{\partial \beta_k}\right) \\ &= E\left\{\frac{(Y_i - \mu_i) x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(Y_i - \mu_i) x_{ik}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right\} \\ &= \frac{x_{ij} x_{ik}}{\{\mathrm{Var}(Y_i)\}^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 E\left\{(Y_i - \mu_i)^2\right\} \\ &= \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \end{aligned}$$

Therefore, the (j, k)th element of the Fisher information matrix is

$$v_{jk} = E\left(-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

Hence, the Fisher information matrix is

$$V = \mathbf{X}^\top \mathbf{W} \mathbf{X},$$

where $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ and

$$w_i = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2.$$

We call w_i the **iterative weight** for the ith observation. By Theorem 2.1, for large n, $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p\{\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}\}$. Note that, in the normal linear model case, we have $\partial \eta_i / \partial \mu_i = 1$ and $\mathrm{Var}(Y_i) = \sigma^2$, so that $w_i = 1/\sigma^2$. It follows that $V = \mathbf{X}^\top \mathbf{X} / \sigma^2$, as in Section 2.4.

For Fisher's method of scoring, the mth approximation is given by

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \left\{V^{(m-1)}\right\}^{-1} \left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}},$$

where $V^{(m-1)}$ is the Fisher information matrix evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}$. Multiplying both sides of the above equation by $V^{(m-1)}$ gives

$$V^{(m-1)} \boldsymbol{\beta}^{(m)} = V^{(m-1)} \boldsymbol{\beta}^{(m-1)} + \left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(m-1)}}.$$

Now, the right-hand side of this equation is a vector with jth component

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \beta_k^{(m-1)} + \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\mathrm{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i},$$

evaluated at $\boldsymbol{\beta}^{(m-1)}$. Thus, we can write

$$V^{(m-1)}\boldsymbol{\beta}^{(m-1)} + \left.\frac{\partial \ell}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m-1)}} = \mathbf{X}^\top \mathbf{W}\mathbf{z},$$

where the vector $\mathbf{z}$ has ith component

$$z_i = \sum_{k=1}^{p} x_{ik}\beta_k^{(m-1)} + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} = \{\boldsymbol{\beta}^{(m-1)}\}^\top \mathbf{x}_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i},$$

with μ_i and $\partial\eta_i/\partial\mu_i$ evaluated at $\boldsymbol{\beta}^{(m-1)}$. We call z_i the **working dependent variate** for the ith observation. Hence, the iterative equation for Fisher's method of scoring can be written as

$$\mathbf{X}^\top \mathbf{W}\mathbf{X}\boldsymbol{\beta}^{(m)} = \mathbf{X}^\top \mathbf{W}\mathbf{z},$$

which is equivalent to

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}\mathbf{z}.$$

This has a similar form to the estimates for a normal linear model obtained by least squares, except that the above equation has to be solved iteratively because $\mathbf{z}$ and $\mathbf{W}$ usually depend on $\boldsymbol{\beta}$. Note that, in the normal linear model case, we have $z_i = y_i$, so that the maximum likelihood estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$, as in Section 2.4.

For generalised linear models, the maximum likelihood estimates are obtained by an **iteratively reweighted least squares** procedure. It begins by using some initial approximation $\boldsymbol{\beta}^{(0)}$ to evaluate $\mathbf{z}$ and $\mathbf{W}$. Then the iterative equation is solved to give $\boldsymbol{\beta}^{(1)}$, which, in turn, is used to obtain better approximations for $\mathbf{z}$ and $\mathbf{W}$, and so on until adequate convergence is achieved. When the difference between successive approximations $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\beta}^{(m-1)}$ is sufficiently small, $\boldsymbol{\beta}^{(m)}$ is taken as the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

Example 3.5 Beetle data set.
A certain number of beetles (r) are exposed to various concentrations of gaseous carbon disulphide, in milligrammes per litre, for five hours and the number of beetles killed (y) is recorded. The dose (x) is the base 10 logarithm of the concentration. Below are the data.

x	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
r	59	60	62	56	63	59	62	60
y	6	13	18	28	52	53	61	60

Let Y_i denote the number of beetles killed out of the r_i exposed to dose x_i. Then it is assumed that $Y_i \sim \text{Bin}(r_i, \pi_i)$ for $i = 1, 2, \ldots, 8$, all independent, where

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

Thus, we have $\mu_i = E(Y_i) = r_i\pi_i$, $\text{Var}(Y_i) = r_i\pi_i(1-\pi_i)$ and

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mu_i}{r_i - \mu_i}\right).$$

It follows that

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{r_i}{\mu_i(r_i - \mu_i)} = \frac{1}{r_i\pi_i(1-\pi_i)}$$

and $w_i = r_i\pi_i(1-\pi_i)$. After four iterations of Fisher's method of scoring, the maximum likelihood estimates of β_0 and β_1 are $\hat{\beta}_0 = -60.717$ and $\hat{\beta}_1 = 34.270$, and their respective standard errors are $\sqrt{\hat{v}^{11}} = 5.181$ and $\sqrt{\hat{v}^{22}} = 2.912$. Thus, the fitted logistic regression model is

$$\hat{\pi}_i = \frac{\exp(-60.717 + 34.270x_i)}{1 + \exp(-60.717 + 34.270x_i)}.$$

Also, the value of the test statistic for testing $H_0 : \beta_1 = 0$ is

$$z = \frac{\hat{\beta}_1}{\sqrt{\hat{v}^{22}}} = 11.77.$$

Since the p-value is $P < 0.001$, there is overwhelming evidence that $\beta_1 \neq 0$.

3.4 Assessing the fit of a model

The adequacy of a model is defined relative to a **maximal model**, which has the same number of parameters as observations.

The maximal model involves the parameter vector $\boldsymbol{\beta}_{\max} = (\beta_1, \ldots, \beta_n)^\top$. We compare this with another model specified by the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, where $p < n$. A measure of goodness of fit is provided by the generalised likelihood ratio

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\boldsymbol{\beta}}; \mathbf{y})}{L(\hat{\boldsymbol{\beta}}_{\max}; \mathbf{y})},$$

where $\hat{\boldsymbol{\beta}}_{\max}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_{\max}$. Small values of $\Lambda(\mathbf{y})$ provide evidence that the model with parameter vector $\boldsymbol{\beta}$ is a poor fit to the data. By Wilks' theorem, for large n, if this model is a good fit, then $D = -2\log\{\Lambda(\mathbf{Y})\} \sim \chi^2_{n-p}$. Thus, the critical region for a test with approximate significance level α is $R = \{\mathbf{y} : -2\log\{\Lambda(\mathbf{y})\} > \chi^2_{n-p,\alpha}\}$. We call the statistic D the **deviance**.

In general, we may wish to compare a model with parameter vector $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_q)^\top$ with one with parameter vector $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_p)^\top$, where $q < p < n$. Let D_0 and D_1 denote their respective deviances. Then, again by Wilks' theorem, for large n, if both models fit the data well, we have $D_0 \sim \chi^2_{n-q}$ and $D_1 \sim \chi^2_{n-p}$, and it can be shown that $D_0 - D_1 \sim \chi^2_{p-q}$. If $D_0 - D_1$ is small, then we would choose the model with q parameters, since it is simpler. However, if $D_0 - D_1$ is large, we would choose the one with p parameters. Note that, although the χ^2 approximation is not very accurate for the deviance, it is a much better approximation for the difference in deviances.

Example 3.6 Beetle example revisited.
In this case, we are fitting a logistic regression model with $p = 2$ parameters and the maximal model has $n = 8$ parameters. The data give $D = 11.232$. Since $\chi^2_{6,0.1} = 10.64$ and $\chi^2_{6,0.05} = 12.59$, the p-value is $0.05 < P < 0.1$, and so there is weak evidence that the logistic regression model does not fit the data particularly well. Note that the deviances for the probit model and the extreme value model are $D = 10.12$ and $D = 3.4464$, respectively. The latter model clearly provides the best description of the data.

Returning to the logistic regression model, suppose that we wish to compare a one-parameter model with just an intercept, that is, the **null model**, with the two-parameter one. Then we have $q = 1$, $D_0 = 284.202$ and $D_1 = 11.232$, so that $D_0 - D_1 = 272.970$. Since $\chi^2_{1,0.001} = 10.83$, the p-value is $P < 0.001$, and so there is overwhelming evidence that the two-parameter model provides a better description of the data than the null model.

3.5 Inspecting and checking models

There are a number of different residuals that we can calculate in order to help us judge the fit of a generalised linear model.

We define the **Pearson residual** by

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},$$

where $V(\hat{\mu}_i)$ is the variance of Y_i in terms of the fitted values $\hat{\mu}_i$. Note that the sum of squares of these residuals gives Pearson's goodness-of-fit test statistic X^2. A large value of the Pearson residual means that observation is making a large contribution to X^2. For example, the Pearson residual for the Poisson distribution is

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}},$$

whereas, for the binomial disribution, it is

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{\mu}_i/r_i)}}.$$

The Pearson residuals do not have unit variance.

We define the **deviance residual** by

$$e_i^D = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where d_i is the term in the deviance corresponding to the ith observation, and $\text{sgn}(y_i - \hat{\mu}_i) = +1$ if $y_i > \hat{\mu}_i$, -1 if $y_i < \hat{\mu}_i$ and 0 if $y_i = \hat{\mu}_i$. The deviance is the sum of squares of the deviance residuals. For example, the deviance residual for the Poisson distribution is

$$e_i^D = \text{sgn}(y_i - \hat{\mu}_i)\left[2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right\}\right]^{\frac{1}{2}},$$

whereas, for the binomial distribution, it is

$$e_i^D = \text{sgn}(y_i - \hat{\mu}_i)\left[2\left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (r_i - y_i)\log\left(\frac{r_i - y_i}{r_i - \hat{\mu}_i}\right)\right\}\right]^{\frac{1}{2}}.$$

The distribution of the deviance residuals is skew and their variance is not one.

We define the **Anscombe residual** by

$$e_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}},$$

where the transformation

$$A(x) = \int \frac{1}{\sqrt[3]{V(x)}} dx$$

is chosen to make the distribution of the residuals as normal as possible. It follows that we can check the distributional assumptions by ordering the Anscombe residuals and plotting them against the expected normal quantiles. For example, the Anscombe residual for the Poisson distribution is

$$e_i^A = \frac{3(y_i^{\frac{2}{3}} - \hat{\mu}_i^{\frac{2}{3}})}{2\hat{\mu}_i^{\frac{1}{6}}},$$

whereas, for the binomial distribution, it has a complicated expression.

4 Binary Data

4.1 Modelling binary response probabilities

We now study in more detail generalised linear models in which the response variables are measured on a binary scale.

Suppose that $Y_i \sim \text{Bin}(r_i, \pi_i)$ for $i = 1, 2, \ldots, n$, all independent, with

$$g(\pi_i) = \boldsymbol{\beta}^\top \mathbf{x}_i,$$

where g is the link function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\top$. Then the likelihood function is

$$L(\boldsymbol{\pi}; \mathbf{y}) = \prod_{i=1}^{n} \binom{r_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{r_i - y_i}.$$

So the log-likelihood is

$$\ell(\boldsymbol{\pi}; \mathbf{y}) = \sum_{i=1}^{n} \log \binom{r_i}{y_i} + \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (r_i - y_i) \log(1 - \pi_i).$$

Recall from Section 3.3 that the maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ are obtained using an iteratively reweighted least squares procedure.

4.2 Logistic regression

In Section 3.4, we introduced the deviance as a measure of goodness of fit of a model. We now show how it can be calculated for a logistic regression model.

The deviance can be written as

$$D = -2\{\ell(\hat{\boldsymbol{\pi}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\pi}}_{\max}; \mathbf{y})\},$$

where $\hat{\boldsymbol{\pi}}$ is the maximum likelihood estimate of $\boldsymbol{\pi}$ in the model and $\hat{\boldsymbol{\pi}}_{\max}$ is the corresponding estimate in the maximal model. Now, we know that

$$\hat{\pi}_i = g^{-1}\left(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i\right),$$

where g^{-1} is the inverse of g. However, for the maximal model, we have

$$\frac{\partial \ell}{\partial \pi_i} = \frac{y_i}{\pi_i} - \frac{r_i - y_i}{1 - \pi_i}.$$

Setting this derivative to zero, we obtain

$$y_i(1 - \hat{\pi}_{i,\max}) - (r_i - y_i)\hat{\pi}_{i,\max} = 0,$$

which yields the maximum likelihood estimate $\hat{\pi}_{i,\max} = y_i / r_i$. Hence, we have

$$\ell(\hat{\boldsymbol{\pi}}_{\max}; \mathbf{y}) = \sum_{i=1}^{n} \log \binom{r_i}{y_i} + \sum_{i=1}^{n} y_i \log(\hat{\pi}_{i,\max}) + \sum_{i=1}^{n} (r_i - y_i) \log(1 - \hat{\pi}_{i,\max}).$$

It follows that

$$\begin{aligned} D &= -2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{\hat{\pi}_i}{\hat{\pi}_{i,\max}}\right) + (r_i - y_i)\log\left(\frac{1-\hat{\pi}_i}{1-\hat{\pi}_{i,\max}}\right)\right\} \\ &= 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{r_i\hat{\pi}_i}\right) + (r_i - y_i)\log\left\{\frac{r_i - y_i}{r_i(1-\hat{\pi}_i)}\right\}\right]. \end{aligned}$$

Note that $D = \sum_{i=1}^{n} d_i$, where d_i is the contribution to the deviance from the ith observation, since $\hat{\mu}_i = r_i\hat{\pi}_i$. By Wilks' theorem, for large n, if the model is a good fit, then $D \sim \chi^2_{n-p}$.

Example 4.1 Rat data set.
The effects of the dose of poison (x), in milligrammes, and the method of delivery (w) on the probability of survival were examined in a study of rats. The two methods of delivery were as a solid with food or as a liquid in water. For each combination of dose and method of delivery, a certain number of rats (r) were used and the number who survived (y) is recorded. Below are the data.

x	1.3	1.6	2.0	2.5	3.0	3.5	1.3	1.6	2.0	2.5	3.0	3.5
w	1	1	1	1	1	1	2	2	2	2	2	2
r	30	30	20	15	10	5	30	30	20	15	10	5
y	28	23	11	5	2	0	29	25	14	8	3	1

Let Y_{jk} denote the number of rats who survived out of the r_{jk} exposed to dose x_k and method of delivery j. Then it is initially assumed that $Y_{jk} \sim \text{Bin}(r_{jk}, \pi_{jk})$ for $j = 1, 2$ and $k = 1, 2, \ldots, 6$, all independent, where

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta_j x_k.$$

Thus, we are allowing a different intercept and slope for each method of delivery. In terms of our notation in Section 4.1, we have $\mathbf{x}_{1k} = (1, 0, x_k, 0)^\top$, $\mathbf{x}_{2k} = (0, 1, 0, x_k)^\top$ and $\boldsymbol{\beta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)^\top$. So we are fitting a logistic regression model with $p = 4$ parameters. On the other hand, the maximal model has $n = 12$ parameters. The data give $D = 3.7217$. Since $\chi^2_{8,0.1} = 13.36$, the p-value is $P > 0.1$, and so there is no evidence that this model does not fit the data well.

Next, we assume that the slope is the same for the two methods of delivery, so that

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta x_k.$$

This means that the regression lines are parallel. Again, in terms of our notation in Section 4.1, we have $\mathbf{x}_{1k} = (1, 0, x_k)^\top$, $\mathbf{x}_{2k} = (0, 1, x_k)^\top$ and $\boldsymbol{\beta} = (\alpha_1, \alpha_2, \beta)^\top$. So we are fitting a three-parameter model. The data give $D = 4.1682$. Thus, the difference in the deviances is 0.4465 on one degree of freedom. Clearly, the p-value is $P > 0.1$, and so there is no evidence that the regression lines are not parallel, that is, that $\beta_1 \neq \beta_2$.

Finally, we assume that both the intercept and the slope are the same for the two methods of delivery, so that

$$\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \beta x_k.$$

This means that there is no difference between the methods of delivery. Again, in terms of our notation in Section 4.1, we have $\mathbf{x}_k = (1, x_k)^\top$ and $\boldsymbol{\beta} = (\alpha, \beta)^\top$. So we are fitting a

two-parameter model. The data give $D = 7.7833$. Thus, the difference in the deviances for this model and the previous one is 3.6151 on one degree of freedom. Since $\chi^2_{1,0.1} = 2.71$ and $\chi^2_{1,0.05} = 3.84$, the p-value is $0.05 < P < 0.1$, and so there is weak evidence of a difference between the two methods of delivery, that is, that $\alpha_1 \neq \alpha_2$.

On the basis of the above, we choose the second model. After four iterations of Fisher's method of scoring, the maximum likelihood estimates of α_1, α_2 and β are $\hat{\alpha}_1 = 4.737$, $\hat{\alpha}_2 = 5.400$ and $\hat{\beta} = -2.148$, and their respective standard errors are $\sqrt{\hat{v}^{11}} = 0.651$, $\sqrt{\hat{v}^{22}} = 0.353$ and $\sqrt{\hat{v}^{33}} = 0.312$. Thus, the fitted logistic regression model is

$$\hat{\pi}_{1k} = \frac{\exp(4.737 - 2.148x_k)}{1 + \exp(4.737 - 2.148x_k)}$$

for those rats given the poison in food and

$$\hat{\pi}_{2k} = \frac{\exp(5.400 - 2.148x_k)}{1 + \exp(5.400 - 2.148x_k)}$$

for those given the poison in water. We can also find approximate confidence intervals for the different parameters. For example, an approximate 95% confidence interval for β is

$$\hat{\beta} \pm 1.96 \times \sqrt{\hat{v}^{33}} = -2.148 \pm 1.96 \times 0.312 = -2.148 \pm 0.612$$

or $(-2.760, -1.536)$.

4.3 Pearson's goodness-of-fit test statistic

There is an alternative measure of goodness of fit which is asymptotically equivalent to the deviance.

Pearson's goodness-of-fit test statistic is defined by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

where $\hat{\mu}_i = r_i\hat{\pi}_i$ and $V(\hat{\mu}_i) = \hat{\mu}_i(1 - \hat{\mu}_i/r_i)$. It follows that

$$\begin{aligned} X^2 &= \sum_{i=1}^{n} \frac{(y_i - r_i\hat{\pi}_i)^2}{r_i\hat{\pi}_i(1 - \hat{\pi}_i)} \\ &= \sum_{i=1}^{n} \frac{(y_i - r_i\hat{\pi}_i)^2}{r_i\hat{\pi}_i} + \sum_{i=1}^{n} \frac{\{r_i - y_i - r_i(1 - \hat{\pi}_i)\}^2}{r_i(1 - \hat{\pi}_i)}. \end{aligned}$$

Thus, the test statistic has the form

$$X^2 = \sum \frac{(o - e)^2}{e},$$

where o denotes the observed frequencies y_i and $r_i - y_i$, e denotes the corresponding estimated expected frequencies $r_i\hat{\pi}_i$ and $r_i(1 - \hat{\pi}_i)$, and the sum is over all cells in a $2 \times n$ table.

Now consider the deviance. Then we know that

$$D = 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{r_i\hat{\pi}_i}\right) + (r_i - y_i) \log\left\{\frac{r_i - y_i}{r_i(1 - \hat{\pi}_i)}\right\}\right].$$

So this statistic has the form

$$D = 2\sum o \log\left(\frac{o}{e}\right).$$

To see why the two statistics are asymptotically equivalent, we can use the Taylor series expansion of $f(u) = u\log(u/v)$ about the point $u = v$. We obtain

$$\begin{aligned} f(u) &= f(v) + (u-v)f'(v) + \frac{(u-v)^2}{2}f''(v) + \ldots \\ &= v\log 1 + (u-v)\left\{1 + \log\left(\frac{u}{v}\right)\right\}\Big|_{u=v} + \frac{(u-v)^2}{2}\left(\frac{1}{u}\right)\Big|_{u=v} + \ldots \\ &= u - v + \frac{(u-v)^2}{2v} + \ldots. \end{aligned}$$

Thus, the deviance can be written as

$$\begin{aligned} D &= 2\sum\left\{o - e + \frac{(o-e)^2}{2e} + \ldots\right\} \\ &\simeq \sum\frac{(o-e)^2}{e} = X^2, \end{aligned}$$

since $\sum o = \sum e$. Consequently, for large n, if the model is a good fit, then $X^2 \sim \chi^2_{n-p}$. The χ^2 approximation is often better for X^2 than D because the latter is unduly influenced by very small frequencies.

Example 4.2 Rat example revisited.
For the three logistic regression models that we fitted, the values of D were 3.7217, 4.1682 and 7.7833. In comparison, the values of X^2 are 3.3172, 3.7374 and 7.3369, which are quite similar. The conclusions about the individual models are the same as before.

4.4 Overdispersion

In some cases, there may be greater variability in the data than would be expected under the assumed model. This phenomenon is called **overdispersion**.

One approach is to assume that $\mathrm{Var}(Y_i) = \psi V(\mu_i)$, where $\psi > 0$ is an unknown dispersion parameter. Although maximum likelihood estimation could be employed to estimate ψ, it is more common to use the estimate

$$\hat{\psi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Note that $\hat{\psi}$ has a similar form to s^2 in Section 2.4. The above estimate is much simpler to compute, and, in some cases, offers greater numerical stability than the maximum likelihood estimate.

5 Count Data

5.1 Poisson regression

This chapter is concerned with the analysis of data in which the response and explanatory variables are all categorical.

Suppose that $Y_i \sim \text{Poisson}(\mu_i)$ for $i = 1, 2, \ldots, n$, all independent, with

$$g(\mu_i) = \boldsymbol{\beta}^T \mathbf{x}_i,$$

where g is the link function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. Then the likelihood function is

$$L(\boldsymbol{\mu}; \mathbf{y}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

So the log-likelihood is

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log(y_i!).$$

The maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ are obtained using the iteratively reweighted least squares procedure in Section 3.3.

The deviance can be written as

$$D = -2\{\ell(\hat{\boldsymbol{\mu}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_{\max}; \mathbf{y})\},$$

where $\hat{\boldsymbol{\mu}}$ is the maximum likelihood estimate of $\boldsymbol{\mu}$ in the model and $\hat{\boldsymbol{\mu}}_{\max}$ is the corresponding estimate in the maximal model. Now, we know that

$$\hat{\mu}_i = g^{-1}\left(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i\right),$$

where g^{-1} is the inverse of g. However, for the maximal model, we have

$$\frac{\partial \ell}{\partial \mu_i} = \frac{y_i}{\mu_i} - 1.$$

Setting this derivative to zero yields the maximum likelihood estimate $\hat{\mu}_{i,\max} = y_i$. Hence, we have

$$\ell(\hat{\boldsymbol{\mu}}_{\max}; \mathbf{y}) = \sum_{i=1}^{n} y_i \log(\hat{\mu}_{i,\max}) - \sum_{i=1}^{n} \hat{\mu}_{i,\max} - \sum_{i=1}^{n} \log(y_i!).$$

It follows that

$$\begin{aligned} D &= -2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\hat{\mu}_i}{\hat{\mu}_{i,\max}}\right) - \hat{\mu}_i + \hat{\mu}_{i,\max} \right\} \\ &= 2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i \right\}. \end{aligned}$$

Note that $D = \sum_{i=1}^{n} d_i$, where d_i is the contribution to the deviance from the ith observation. By Wilks' theorem, for large n, if the model is a good fit, then $D \sim \chi^2_{n-p}$.

Example 5.1 Count data set.
The data below are counts (y) observed at various values of a covariate (x).

y	2	3	6	7	8	9	10	12	15
x	-1	-1	0	0	0	0	1	1	1

Let Y_i denote the count for covariate x_i. Then it is assumed that $Y_i \sim \text{Poisson}(\mu_i)$ for $i = 1, 2, \ldots, 9$, all independent, where

$$\log(\mu_i) = \beta_0 + \beta_1 x_i.$$

Thus, we have $\text{Var}(Y_i) = \mu_i$ and $\eta_i = \log(\mu_i)$. It follows that $\partial\eta_i/\partial\mu_i = 1/\mu_i$ and $w_i = \mu_i$. After four iterations of Fisher's method of scoring, the maximum likelihood estimates of β_0 and β_1 are $\hat{\beta}_0 = 1.889$ and $\hat{\beta}_1 = 0.670$, and their respective standard errors are $\sqrt{\hat{v}^{11}} = 0.142$ and $\sqrt{\hat{v}^{22}} = 0.179$. Thus, the fitted Poisson regression model is

$$\hat{\mu}_i = e^{1.889+0.670x_i}.$$

In this case, we are fitting a model with $p = 2$ parameters and the maximal model has $n = 9$ parameters. The data give $D = 2.9387$. Since $\chi^2_{7,0.1} = 12.02$, the p-value is $P > 0.1$, and so there is no evidence that this model does not fit the data well.

5.2 Models for contingency tables

Data often consist of counts or frequencies in the cells of a **contingency table** formed by the cross-classification of response and explanatory variables.

Consider a two-dimensional table in which variable A has J categories and variable B has K categories. Let Y_{jk} denote the frequency in cell (j, k), and let $Y_{j\cdot} = \sum_{k=1}^{K} Y_{jk}$ and $Y_{\cdot k} = \sum_{j=1}^{J} Y_{jk}$ denote the row and column totals. Then we have a table as follows:

	B_1	B_2	$\cdots$	B_K	Total
A_1	Y_{11}	Y_{12}	$\cdots$	Y_{1K}	$Y_{1\cdot}$
A_2	Y_{21}	Y_{22}	$\cdots$	Y_{2K}	$Y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_J	Y_{J1}	Y_{J2}	$\cdots$	Y_{JK}	$Y_{J\cdot}$
Total	$Y_{\cdot 1}$	$Y_{\cdot 2}$	$\cdots$	$Y_{\cdot K}$	$Y_{\cdot\cdot}$

Note that the overall total is $N = Y_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk}$.

The simplest model is obtained by assuming that $Y_{jk} \sim \text{Poisson}(\mu_{jk})$ for $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$, all independent. In this case, the joint probability mass function of the Ys is

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}) &= \prod_{j=1}^{J}\prod_{k=1}^{K} P(Y_{jk} = y_{jk}) \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\mu_{jk}^{y_{jk}} e^{-\mu_{jk}}}{y_{jk}!}, \end{aligned}$$

where $\mathbf{y} = (y_{11}, \ldots, y_{JK})^\top$.

In practice, there are usually constraints on the Ys, such as that the overall total N is fixed. Now, we know that $N \sim \text{Poisson}(\mu_{..})$, where $\mu_{..} = \sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{jk}$. Thus, the probability mass function of N is

$$P(N=n) = \frac{\mu_{..}^n e^{-\mu_{..}}}{n!}.$$

It follows that the joint probability mass function of the Ys conditional on $N=n$ is

$$\begin{aligned} P(\mathbf{Y}=\mathbf{y}|N=n) &= \frac{P(\mathbf{Y}=\mathbf{y}, N=n)}{P(N=n)} \\ &= \frac{P(\mathbf{Y}=\mathbf{y})}{P(N=n)} \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\mu_{jk}^{y_{jk}}e^{-\mu_{jk}}}{y_{jk}!}\frac{n!}{\mu_{..}^n e^{-\mu_{..}}} \\ &= n!\prod_{j=1}^{J}\prod_{k=1}^{K}\left(\frac{\mu_{jk}}{\mu_{..}}\right)^{y_{jk}}\frac{1}{y_{jk}!}. \end{aligned}$$

Let $\theta_{jk} = \mu_{jk}/\mu_{..}$. Then we have

$$P(\mathbf{Y}=\mathbf{y}|N=n) = n!\prod_{j=1}^{J}\prod_{k=1}^{K}\frac{\theta_{jk}^{y_{jk}}}{y_{jk}!},$$

where $0 \leq \theta_{jk} \leq 1$ and $\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{jk} = 1$. This is a **multinomial distribution**. The quantity θ_{jk} is the probability that an individual is in cell (j,k).

Another possibility is that the row totals are fixed. Here, the joint probability mass function of the Ys for each row is multinomial and it is assumed that the rows are independent. Thus, the joint probability mass function of the Ys conditional on $Y_{j.} = y_{j.}$ for $j = 1, 2, \ldots, J$ is

$$P(\mathbf{Y}=\mathbf{y}|Y_{j.}=y_{j.}, j=1,2,\ldots,J) = \prod_{j=1}^{J}y_{j.}!\prod_{k=1}^{K}\frac{\theta_{jk}^{y_{jk}}}{y_{jk}!},$$

where $\theta_{jk} = \mu_{jk}/\mu_{j.}$ and $\sum_{k=1}^{K}\theta_{jk} = 1$ for each j. This is called a **product multinomial distribution**.

5.3 Log-linear models

For contingency tables, the usual hypotheses can all be formulated as multiplicative models for the expected cell frequencies. This suggests that the logarithm is the natural link function between the expected cell frequencies and a linear combination of the parameters.

First, consider the multinomial distribution. Then, if the row and column variables are **independent**, we have $\theta_{jk} = \theta_{j.}\theta_{.k}$, where $\theta_{j.}$ and $\theta_{.k}$ represent the respective probabilities that an individual is in row j and column k, and $\sum_{j=1}^{J}\theta_{j.} = \sum_{k=1}^{K}\theta_{.k} = 1$. It follows that, under the **independence hypothesis**, the expected frequency for cell (j,k) is

$$E(Y_{jk}|N=n) = n\theta_{j.}\theta_{.k}.$$

Thus, we can write

$$\begin{aligned} \eta_{jk} = \log\{E(Y_{jk}|N=n)\} &= \log n + \log(\theta_{j.}) + \log(\theta_{.k}) \\ &= \mu + \alpha_j + \beta_k, \end{aligned}$$

where μ is the overall effect, α_j is the effect of row j and β_k is the effect of column k. This is an example of a **log-linear model**. The corresponding maximal model $E(Y_{jk}|N = n) = n\theta_{jk}$ can be written as

$$\eta_{jk} = \log\{E(Y_{jk}|N = n)\} = \mu + \alpha_j + \beta_k + \gamma_{jk},$$

where γ_{jk} is the interaction effect of row j and column k. So the independence hypothesis $\theta_{jk} = \theta_{j.}\theta_{.k}$ for all j and k is equivalent to the **no interaction hypothesis** $\gamma_{jk} = 0$ for all j and k. Note that, if γ_{jk} is included in the model, then the lower-order terms α_j and β_k are also included. We say that the models are **hierarchical**.

Now, for the first log-linear model, the constraints $\sum_{j=1}^{J} \theta_{j.} = \sum_{k=1}^{K} \theta_{.k} = 1$ imply a set of nonlinear constraints on the parameters α_j and β_k. This means that the model is **overparametrised**, that is, there are more parameters than are really needed. To avoid complications due to the presence of nonlinear constraints, we choose the parametrisation

$$\mu = \log n + \frac{1}{J}\sum_{j=1}^{J} \log(\theta_{j.}) + \frac{1}{K}\sum_{k=1}^{K} \log(\theta_{.k}),$$

$$\alpha_j = \log(\theta_{j.}) - \frac{1}{J}\sum_{j=1}^{J} \log(\theta_{j.})$$

and

$$\beta_j = \log(\theta_{.k}) - \frac{1}{K}\sum_{k=1}^{K} \log(\theta_{.k}),$$

so that there are two linear constraints $\sum_{j=1}^{J} \alpha_j = 0$ and $\sum_{k=1}^{K} \beta_k = 0$. Alternatively, we can take $\alpha_1 = 0$ and $\beta_1 = 0$, which means that the first level of each variable is the reference level. This is the parametrisation used by R. In either case, the model has $J + K - 1$ parameters. Neither the fitted values nor the deviance are affected by the choice of parametrisation.

Similarly, for the maximal model, there are the additional constraints $\sum_{j=1}^{J} \gamma_{jk} = 0$ and $\sum_{k=1}^{K} \gamma_{jk} = 0$. Alternatively, we can take $\gamma_{1k} = 0$ and $\gamma_{j1} = 0$, which is the parametrisation used by R. In either case, the model has

$$1 + (J - 1) + (K - 1) + (J - 1)(K - 1) = JK$$

parameters, which is the number of observations.

Next, consider the product multinomial distribution with fixed row totals $y_{j.}$. Then, if the cell probabilities are the same in each row, we have $\theta_{jk} = \theta_{.k}$ for all j. It follows that, under the **homogeneity hypothesis**, the expected frequency for cell (j, k) is

$$E(Y_{jk}|Y_{j.} = y_{j.}) = y_{j.}\theta_{.k},$$

where $\sum_{k=1}^{K} \theta_{.k} = 1$. Thus, we can write

$$\begin{aligned} \eta_{jk} = \log\{E(Y_{jk}|Y_{j.} = y_{j.})\} &= \log(y_{j.}) + \log(\theta_{.k}) \\ &= \mu + \alpha_j + \beta_k. \end{aligned}$$

So we have a log-linear model. The corresponding maximal model $E(Y_{jk}|Y_{j.} = y_{j.}) = y_{j.}\theta_{jk}$ can be written as

$$\eta_{jk} = \log\{E(Y_{jk}|Y_{j.} = y_{j.})\} = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

So the homogeneity hypothesis $\theta_{jk} = \theta_{.k}$ for all j and k is equivalent to the no interaction hypothesis $\gamma_{jk} = 0$ for all j and k. For both log-linear models, there are the same constraints as before.

For the multinomial and product multinomial distributions, certain quantities are fixed and all terms relating to these must always be included in the model. Consequently, in each case, there is a **minimal model**, which has the minimum number of terms. Since the overall total N is fixed for the multinomial distribution, we must include μ in the model. On the other hand, the row totals $Y_{j.}$ are fixed for the product multinomial distribution, and so we must include $\mu + \alpha_j$. By **Birch's conditions**, the maximum likelihood estimates are the same for both models.

5.4 Fitting the models

We now show how to fit a log-linear model using maximum likelihood estimation when the overall total N is fixed.

Let $\boldsymbol{\theta} = (\theta_{11}, \ldots, \theta_{JK})^\top$. Then the likelihood is

$$L(\boldsymbol{\theta}; \mathbf{y}) = n! \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!}.$$

So the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \log(n!) + \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \log(\theta_{jk}) - \sum_{j=1}^{J} \sum_{k=1}^{K} \log(y_{jk}!).$$

For the maximal model, the maximum likelihood estimates are obtained by maximising the log-likelihood subject to the constraint $\sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{jk} = 1$. This can be achieved using a Lagrange multiplier ξ, that is, finding θ_{jk} and ξ to maximise

$$t(\boldsymbol{\theta}, \xi; \mathbf{y}) = \log(n!) + \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \log(\theta_{jk}) - \sum_{j=1}^{J} \sum_{k=1}^{K} \log(y_{jk}!) - \xi \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{jk} - 1 \right).$$

Thus, we have

$$\frac{\partial t}{\partial \theta_{jk}} = \frac{y_{jk}}{\theta_{jk}} - \xi$$

and

$$\frac{\partial t}{\partial \xi} = 1 - \sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{jk}.$$

Setting these derivatives to zero yields the maximum likelihood estimates $\hat{\theta}_{jk,\max} = y_{jk}/n$ and $\hat{\xi} = n$. Under the independence hypothesis $\theta_{jk} = \theta_{j.}\theta_{.k}$ for all j and k, the log-likelihood is

$$\ell(\boldsymbol{\theta}; \mathbf{y}) = \log(n!) + \sum_{j=1}^{J} y_{j.} \log(\theta_{j.}) + \sum_{k=1}^{K} y_{.k} \log(\theta_{.k}) - \sum_{j=1}^{J} \sum_{k=1}^{K} \log(y_{jk}!).$$

This time, the maximum likelihood estimates are obtained by maximising the log-likelihood subject to the constraints $\sum_{j=1}^{J} \theta_{j.} = \sum_{k=1}^{K} \theta_{.k} = 1$. Using Lagrange multipliers ξ_1 and ξ_2,

we find $\theta_{j.}$, $\theta_{.k}$, ξ_1 and ξ_2 to maximise

$$\begin{aligned} t(\boldsymbol{\theta}, \xi_1, \xi_2; \mathbf{y}) &= \log(n!) + \sum_{j=1}^{J} y_{j.} \log(\theta_{j.}) + \sum_{k=1}^{K} y_{.k} \log(\theta_{.k}) - \sum_{j=1}^{J} \sum_{k=1}^{K} \log(y_{jk}!) \\ &\quad -\xi_1 \left(\sum_{j=1}^{J} \theta_{j.} - 1 \right) - \xi_2 \left(\sum_{k=1}^{K} \theta_{.k} - 1 \right). \end{aligned}$$

Thus, we have

$$\frac{\partial t}{\partial \theta_{j.}} = \frac{y_{j.}}{\theta_{j.}} - \xi_1, \quad \frac{\partial t}{\partial \theta_{.k}} = \frac{y_{.k}}{\theta_{.k}} - \xi_2,$$

$$\frac{\partial t}{\partial \xi_1} = 1 - \sum_{j=1}^{J} \theta_{j.}$$

and

$$\frac{\partial t}{\partial \xi_2} = 1 - \sum_{k=1}^{K} \theta_{.k}.$$

Setting these derivatives to zero yields the maximum likelihood estimates $\hat{\theta}_{j.} = y_{j.}/n$, $\hat{\theta}_{.k} = y_{.k}/n$ and $\hat{\xi}_1 = \hat{\xi}_2 = n$. It follows that, under the independence hypothesis, the expected frequency for cell (j, k) is

$$e_{jk} = n\hat{\theta}_{j.}\hat{\theta}_{.k} = \frac{y_{j.}y_{.k}}{n}.$$

By substituting the estimates $\hat{\theta}_{j.}$ and $\hat{\theta}_{.k}$ into the equations for μ, α_j and β_k in Section 5.3, we obtain $\hat{\mu}$, $\hat{\alpha}_j$ and $\hat{\beta}_k$.

The deviance is

$$\begin{aligned} D &= -2\{\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \ell(\hat{\boldsymbol{\theta}}_{\max}; \mathbf{y})\} \\ &= -2 \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\hat{\theta}_{jk,\max}} \right) \\ &= 2 \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \log \left(\frac{y_{jk}}{e_{jk}} \right). \end{aligned}$$

By Wilks' theorem, for large n, if the model is a good fit, then $D \sim \chi^2_{(J-1)(K-1)}$. Pearson's goodness-of-fit test statistic is

$$X^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(y_{jk} - e_{jk})^2}{e_{jk}}.$$

By the argument used in Section 4.3, the two statistics are asymptotically equivalent. Consequently, for large n, if the model is a good fit, then $X^2 \sim \chi^2_{(J-1)(K-1)}$. Note that X^2 is more commonly used for contingency table data than D.

Example 5.2 Cancer data set.
In a cross-sectional study of bone cancer, the type of cancer and the site were recorded for 300 patients. The contingency table below shows the number of patients (y) with each combination of type of cancer and site.

	Site				
Type	Head	Arms	Body	Legs	Total
I	21	13	42	8	84
II	10	26	20	35	91
III	30	34	32	29	125
Total	61	73	94	72	300

The null hypothesis is that type of cancer and site are independent. Let Y_{jk} denote the number of patients classified in row j and column k. Then it is assumed that the Y_{jk} have a multinomial distribution with parameters n and θ_{jk} for $j = 1, 2, 3$ and $k = 1, 2, 3, 4$, where $n = 300$ and θ_{jk} is the probability that a patient is classified in row j and column k. The data give $D = 38.869$. Since $\chi^2_{6,0.001} = 22.46$, the p-value is $P < 0.001$, and so there is very strong evidence that type of cancer is not independent of site. In comparison, we have $X^2 = 37.928$, which leads to the same conclusion.

6 Survival Data

6.1 Survivor and hazard functions

In many applications, we are interested in the study of the lifetimes of individuals, such as components in engineering and patients in medicine.

Suppose that the length of the life $T > 0$ of an individual has probability density function $f(t)$ and distribution function $F(t)$. Then the **survivor function** is defined by

$$S(t) = P(T > t) = 1 - F(t).$$

The **hazard function** is the conditional probability density function of T given survival up to time t. It is given by

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \frac{1}{P(T > t)} \\ &= \frac{f(t)}{1 - F(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

The function defined by

$$H(t) = \int_0^t h(u) du$$

is called the **integrated hazard function**. Now, we can write

$$h(t) = -\frac{d}{dt} \log\{S(t)\}.$$

Since $S(0) = 1$, it follows that

$$S(t) = \exp\{-H(t)\}$$

and

$$f(t) = h(t) \exp\{-H(t)\}.$$

So we can find the probability density function or the survivor function given the hazard function. In fact, we can find the other two functions from any one, that is, each is an equivalent representation of the lifetime distribution.

Example 6.1 Exponential distribution.
Suppose that $T \sim \text{Exp}(\lambda)$. Then we have

$$S(t) = \int_t^\infty \lambda e^{-\lambda u} du = e^{-\lambda t}$$

and

$$h(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$

Thus, we see that the exponential distribution has a constant hazard function. This is often not a realistic assumption in practice because it represents a situation where there is no aging. The probability of an individual dying in a time interval $(t_0, t_0 + t]$ does not depend on the starting point t_0.

6.2 Exponential regression

For survival data, explanatory variables are usually taken account of by fitting a model to the hazard function.

Suppose that $T_i \sim \text{Exp}(\lambda_i)$ for $i = 1, 2, \ldots, n$, all independent, with

$$g(\lambda_i) = \boldsymbol{\beta}^\top \mathbf{x}_i,$$

where g is the link function, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\top$. Then the likelihood is

$$L(\boldsymbol{\lambda}; \mathbf{t}) = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i t_i},$$

where $\mathbf{t} = (t_1, \ldots, t_n)^\top$. So the log-likelihood is

$$\ell(\boldsymbol{\lambda}; \mathbf{t}) = \sum_{i=1}^{n} \log(\lambda_i) - \sum_{i=1}^{n} \lambda_i t_i.$$

The maximum likelihood estimates of $\beta_1, \ldots, \beta_p$ are again obtained using the iteratively reweighted least squares procedure in Section 3.3.

The deviance can be written as

$$D = -2\{\ell(\hat{\boldsymbol{\lambda}}; \mathbf{t}) - \ell(\hat{\boldsymbol{\lambda}}_{\max}; \mathbf{t})\},$$

where $\hat{\boldsymbol{\lambda}}$ is the maximum likelihood estimate of $\boldsymbol{\lambda}$ in the model and $\hat{\boldsymbol{\lambda}}_{\max}$ is the corresponding estimate in the maximal model. Now, we know that

$$\hat{\lambda}_i = g^{-1}\left(\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i\right),$$

where g^{-1} is the inverse of g. However, for the maximal model, we have

$$\frac{\partial \ell}{\partial \lambda_i} = \frac{1}{\lambda_i} - t_i.$$

Setting this derivative to zero yields the maximum likelihood estimate $\hat{\lambda}_{i,\max} = 1/t_i$. Hence, we have

$$\ell(\hat{\boldsymbol{\lambda}}_{\max}; \mathbf{t}) = \sum_{i=1}^{n} \log(\hat{\lambda}_{i,\max}) - n.$$

It follows that

$$\begin{aligned} D &= -2\sum_{i=1}^{n} \left\{ \log\left(\frac{\hat{\lambda}_i}{\hat{\lambda}_{i,\max}}\right) - \hat{\lambda}_i t_i + n \right\} \\ &= 2\sum_{i=1}^{n} \{-\log(\hat{\lambda}_i t_i) - n + \hat{\lambda}_i t_i\}. \end{aligned}$$

Note that $D = \sum_{i=1}^{n} d_i$, where d_i is the contribution to the deviance from the ith observation. By Wilks' theorem, for large n, if the model is a good fit, then $D \sim \chi^2_{n-p}$.

We can check whether the survival times have an exponential distribution. Recall that $\lambda_i T_i \sim \text{Exp}(1)$ for $i = 1, 2, \ldots, n$, all independent. Let $V_i = \hat{\lambda}_i T_i$. Then the V_i have an approximate standard exponential distribution, although they are not independent. Now, the survivor function of this distribution is $S(v) = e^{-v}$. A probability plot can be formed by plotting the empirical survivor function $\hat{S}(v)$ of the v_i against e^{-v}. In fact, plotting $-\log\{\hat{S}(v)\}$ against v or $\log[-\log\{\hat{S}(v)\}]$ against $\log v$ is advisable, as the nature of the departure from exponentiality may show up.

6.3 Censoring

A survival time is **censored** if the exact time is not known, such as when the event of interest has not occurred yet. This event could be the failure of a component or the death of a patient.

Censoring might occur in practice because a patient withdraws from the study or a patient is lost to follow-up during the study period. When censoring has occurred, the likelihood changes. Suppose that, for individual i, we have the data (t_i, δ_i), where $\delta_i = 1$ if $T_i = t_i$ and $\delta_i = 0$ if $T_i > t_i$. We call δ_i the **censoring variable**, since it indicates whether the survival time is uncensored or not. It follows that the likelihood is

$$\begin{aligned} L(\boldsymbol{\lambda}; \mathbf{t}) &= \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \{S(t_i)\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \frac{f(t_i)}{S(t_i)} \right\}^{\delta_i} S(t_i) \\ &= \prod_{i=1}^{n} \{h(t_i)\}^{\delta_i} S(t_i) \\ &= \prod_{i=1}^{n} \lambda_i^{\delta_i} e^{-\lambda_i t_i}. \end{aligned}$$

Thus, the log-likelihood is

$$\ell(\boldsymbol{\lambda}; \mathbf{t}) = \sum_{i=1}^{n} \delta_i \log(\lambda_i) - \sum_{i=1}^{n} \lambda_i t_i.$$

Similar calculations to those in Section 6.2 lead to an expression for the deviance in this case and Wilks' theorem can also be applied.