CHAPTER 7

Binary Variables and Logistic Regression

7.1 Probability distributions

In this chapter we consider generalized linear models in which the outcome variables are measured on a binary scale. For example, the responses may be alive or dead, or present or absent. *Success* and *failure* are used as generic terms of the two categories.

First, we define the **binary random variable**

$$Z = \begin{cases} 1 \text{ if the outcome is a success} \\ 0 \text{ if the outcome is a failure} \end{cases}$$

with probabilities $\Pr(Z = 1) = \pi$ and $\Pr(Z = 0) = 1 - \pi$, which is the Bernoulli distribution $\mathrm{B}(\pi)$. If there are n such random variables $Z_1, \ldots, Z_n$, which are independent with $\Pr(Z_j = 1) = \pi_j$, then their joint probability is

$$\prod_{j=1}^{n} \pi_j^{z_j} (1-\pi_j)^{1-z_j} = \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right], \quad (7.1)$$

which is a member of the exponential family (see equation (3.3)).

Next, for the case where the π_j's are all equal, we can define

$$Y = \sum_{j=1}^{n} Z_j$$

so that Y is the number of successes in n "trials". The random variable Y has the distribution $\mathrm{Bin}(n, \pi)$:

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n. \quad (7.2)$$

Finally, we consider the general case of N independent random variables $Y_1, Y_2, \ldots, Y_N$ corresponding to the numbers of successes in N different subgroups or strata (Table 7.1). If $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, the log-likelihood function is

$$\begin{aligned} & l(\pi_1, \ldots, \pi_N; y_1, \ldots, y_N) \\ &= \sum_{i=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]. \quad (7.3) \end{aligned}$$

Table 7.1 *Frequencies for N Binomial distributions.*

	Subgroups			
	1	2	...	N
Successes	Y_1	Y_2	...	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Totals	n_1	n_2	...	n_N

7.2 Generalized linear models

We want to describe the proportion of successes, $P_i = Y_i/n_i$, in each subgroup in terms of factor levels and other explanatory variables which characterize the subgroup. As $\mathrm{E}(Y_i) = n_i\pi_i$ and so $\mathrm{E}(P_i) = \pi_i$, we model the probabilities π_i as

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

where $\mathbf{x}_i$ is a vector of explanatory variables (dummy variables for factor levels and measured values for covariates), $\boldsymbol{\beta}$ is a vector of parameters and g is a link function.

The simplest case is the **linear model**

$$\pi = \mathbf{x}^T\boldsymbol{\beta}.$$

This is used in some practical applications, but it has the disadvantage that although π is a probability, the fitted values $\mathbf{x}^T\mathbf{b}$ may be less than zero or greater than one.

To ensure that π is restricted to the interval [0,1] it is often modelled using a cumulative probability distribution

$$\pi = \int_{-\infty}^{t} f(s)ds,$$

where $f(s) \geqslant 0$ and $\int_{-\infty}^{\infty} f(s)ds = 1$. The probability density function $f(s)$ is called the **tolerance distribution**. Some commonly used examples are considered in Section 7.3.

7.3 Dose response models

Historically, one of the first uses of regression-like models for Binomial data was for bioassay results (Finney 1973). Responses were the proportions or percentages of "successes"; for example, the proportion of experimental animals killed by various dose levels of a toxic substance. Such data are sometimes called **quantal responses**. The aim is to describe the probability of "success", π, as a function of the dose, x; for example, $g(\pi) = \beta_1 + \beta_2 x$.

If the tolerance distribution $f(s)$ is the uniform distribution on the interval

$[c_1, c_2]$

$$f(s) = \begin{cases} \dfrac{1}{c_2 - c_1} & \text{if } c_1 \leqslant s \leqslant c_2 \\ 0 & \text{otherwise} \end{cases},$$

then π is cumulative

$$\pi = \int_{c_1}^{x} f(s)ds = \frac{x - c_1}{c_2 - c_1} \qquad \text{for } c_1 \leqslant x \leqslant c_2$$

(see Figure 7.1). This equation has the form $\pi = \beta_1 + \beta_2 x$, where

$$\beta_1 = \frac{-c_1}{c_2 - c_1} \text{ and } \beta_2 = \frac{1}{c_2 - c_1}.$$

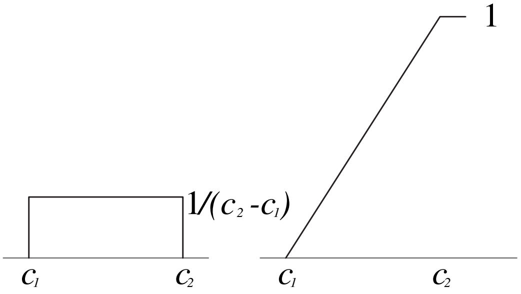

Figure 7.1 *Uniform distribution: $f(s)$ and π.*

This **linear model** is equivalent to using the identity function as the link function g and imposing conditions on x, β_1 and β_2 corresponding to $c_1 \leq x \leq c_2$. These extra conditions mean that the standard methods for estimating β_1 and β_2 for generalized linear models cannot be directly applied. In practice, this model is not widely used.

One of the original models used for bioassay data is called the **probit model**. The Normal distribution is used as the tolerance distribution (see Figure 7.2).

$$\begin{aligned} \pi &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right), \end{aligned}$$

where Φ denotes the cumulative probability function for the standard Normal distribution $\mathrm{N}(0, 1)$. Thus,

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 x$$

where $\beta_1 = -\mu/\sigma$ and $\beta_2 = 1/\sigma$ and the link function g is the inverse cumulative Normal probability function Φ^{-1}. Probit models are used in several areas of biological and social sciences in which there are natural interpretations of the model; for example, $x = \mu$ is called the **median lethal dose** LD(50) because it corresponds to the dose that can be expected to kill half of the animals.

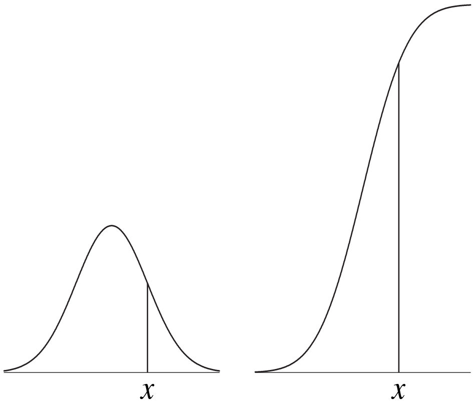

Figure 7.2 *Normal distribution: $f(s)$ and π.*

Another model that gives numerical results very much like those from the probit model, but which computationally is somewhat easier, is the **logistic** or **logit model**. The tolerance distribution is

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \exp(\beta_1 + \beta_2 s)]^2},$$

so

$$\pi = \int_{-\infty}^{x} f(s) ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)}.$$

This gives the link function

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x.$$

The term $\log[\pi/(1-\pi)]$ is sometimes called the **logit function** and it has a natural interpretation as the logarithm of odds (see Exercise 7.2). The logistic model is widely used for Binomial data and is implemented in many statistical programs. The shapes of the functions $f(s)$ and $\pi(x)$ are similar to those for the probit model (Figure 7.2) except in the tails of the distributions (see Cox and Snell 1989).

Several other models are also used for dose response data. For example, if the **extreme value distribution**

$$f(s) = \beta_2 \exp\left[(\beta_1 + \beta_2 s) - \exp\left(\beta_1 + \beta_2 s\right)\right]$$

is used as the tolerance distribution, then

$$\pi = 1 - \exp\left[-\exp\left(\beta_1 + \beta_2 x\right)\right],$$

and so $\log[-\log(1-\pi)] = \beta_1 + \beta_2 x$. This link, $\log[-\log(1-\pi)]$, is called the **complementary log-log function**. The model is similar to the logistic and probit models for values of π near 0.5 but differs from them for π near 0 or 1. These models are illustrated in the following example.

Table 7.2 *Beetle mortality data.*

Dose, x_i ($\log_{10}\text{CS}_2\text{mgl}^{-1}$)	Number of beetles, n_i	Number killed, y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

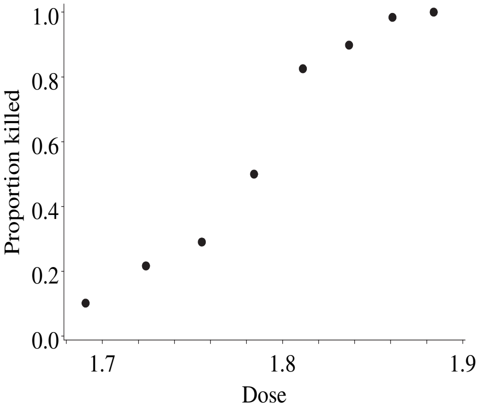

Figure 7.3 *Beetle mortality data from Table 7.2: proportion killed, $p_i = y_i/n_i$, plotted against dose, x_i ($\log_{10} CS_2 mgl^{-1}$).*

7.3.1 Example: Beetle mortality

Table 7.2 shows numbers of beetles dead after five hours exposure to gaseous carbon disulphide at various concentrations (data from Bliss 1935). Figure 7.3 shows the proportions $p_i = y_i/n_i$ plotted against dose x_i (actually x_i is the logarithm of the quantity of carbon disulphide). We begin by fitting the logistic model

$$\pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}$$

so

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_1 + \beta_2 x_i$$

and

$$\log(1 - \pi_i) = -\log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right].$$

Therefore from equation (7.3) the log-likelihood function is

$$l = \sum_{i=1}^{N}\left[y_i(\beta_1 + \beta_2 x_i) - n_i \log\left[1 + \exp(\beta_1 + \beta_2 x_i)\right] + \log\left(\begin{array}{c} n_i \\ y_i \end{array}\right)\right],$$

and the scores with respect to β_1 and β_2 are

$$\begin{aligned} U_1 &= \frac{\partial l}{\partial \beta_1} = \sum\left\{y_i - n_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} = \sum(y_i - n_i\pi_i) \\ U_2 &= \frac{\partial l}{\partial \beta_2} = \sum\left\{y_i x_i - n_i x_i\left[\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}\right]\right\} \\ &= \sum x_i(y_i - n_i\pi_i). \end{aligned}$$

Similarly the information matrix is

$$\mathfrak{I} = \left[\begin{array}{cc} \sum n_i\pi_i(1 - \pi_i) & \sum n_i x_i \pi_i(1 - \pi_i) \\ \\ \sum n_i x_i \pi_i(1 - \pi_i) & \sum n_i x_i^2 \pi_i(1 - \pi_i) \end{array}\right].$$

Maximum likelihood estimates are obtained by solving the iterative equation

$$\mathfrak{I}^{(m-1)}\mathbf{b}^m = \mathfrak{I}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)}$$

(from (4.22)) where the superscript (m) indicates the mth approximation and $\mathbf{b}$ is the vector of estimates. Starting with $b_1^{(0)} = 0$ and $b_2^{(0)} = 0$, successive approximations are shown in Table 7.3. The estimates converge by the sixth iteration. The table also shows the increase in values of the log-likelihood function (7.3), omitting the constant term $\log\binom{n_i}{y_i}$. The fitted values are $\widehat{y}_i = n_i\widehat{\pi}_i$ calculated at each stage (initially $\widehat{\pi}_i = 0.5$ for all i).

For the final approximation, the estimated variance–covariance matrix for $\mathbf{b}$, $\left[\mathfrak{I}(\mathbf{b})^{-1}\right]$, is shown at the bottom of Table 7.3 together with the deviance

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n - y_i}{n - \widehat{y}_i}\right)\right]$$

(from Section 5.6.1).

The estimates and their standard errors are

$$\begin{aligned} & b_1 = -60.72, \quad \text{standard error} = \sqrt{26.840} = 5.18 \\ \text{and} \quad & b_2 = 34.27, \quad \text{standard error} = \sqrt{8.481} = 2.91. \end{aligned}$$

Table 7.3 *Fitting a linear logistic model to the beetle mortality data.*

	Initial	Approximation		
	estimate	First	Second	Sixth
β_1	0	−37.856	−53.853	−60.717
β_2	0	21.337	30.384	34.270
log-likelihood	−333.404	−200.010	−187.274	−186.235

Observations		Fitted values			
y_1	6	29.5	8.505	4.543	3.458
y_2	13	30.0	15.366	11.254	9.842
y_3	18	31.0	24.808	23.058	22.451
y_4	28	28.0	30.983	32.947	33.898
y_5	52	31.5	43.362	48.197	50.096
y_6	53	29.5	46.741	51.705	53.291
y_7	61	31.0	53.595	58.061	59.222
y_8	60	30.0	54.734	58.036	58.743

$$[\mathfrak{I}(\mathbf{b})]^{-1} = \begin{bmatrix} 26.840 & -15.082 \\ -15.082 & 8.481 \end{bmatrix}, \quad D = 11.23$$

If the model is a good fit of the data, the deviance should approximately have the distribution $\chi^2(6)$ because there are $N = 8$ covariate patterns (i.e., different values of x_i) and $p = 2$ parameters. But the calculated value of D is almost twice the "expected" value of 6 and is almost as large as the upper 5% point of the $\chi^2(6)$ distribution, which is 12.59. This suggests that the model does not fit particularly well.

Statistical software for fitting generalized linear models for dichotomous responses often differs between the case when the data are grouped as counts of successes y and failures $n-y$ in n trials with the same covariate pattern and the case when the data are binary (0 or 1) responses (see later example with data in Table 7.8). For Stata the logistic regression model for the grouped data for beetle mortality in Table 7.2 can be fitted using the following command

Stata code (logistic regression)

```
.glm y x, family(binomial n) link(logit)
```

The estimated variance–covariance matrix can be obtained by selecting the option to display the negative Hessian matrix using

Stata code (logistic regression)

```
.glm y x, family(binomial n) link(logit) hessian
```

Evaluated for the final estimates, this matrix is given as $\begin{bmatrix} 58.484 & 104.011 \\ 104.011 & 185.094 \end{bmatrix}$, which can be inverted to obtain $\begin{bmatrix} 26.8315 & -15.0775 \\ -15.0775 & 8.4779 \end{bmatrix}$, which is the same as in Table 7.3 (except for rounding effects). The value for the log-likelihood shown by Stata does not include the term $\sum_{i=1}^{N} \log \binom{n_i}{y_i}$ in (7.3). For the beetle mortality data the value of this term in -167.5203 so, compared with the value of -186.235 in Table 7.3, the log-likelihood value shown by Stata is $-186.235-$ $(-167.5203) = -18.715$. Fitted values can be obtained after the model is fitted using the command

Stata code (fitted values)

```
.predict fit, mu
```

To use R to fit generalized linear models to grouped dichotomous data, it is necessary to construct a response matrix with two columns, y and $(n - y)$, as shown below for the beetle mortality data using the S-PLUS lists denoted by c.

R code (data entry and manipulation)

```
>y=c(6,13,18,28,52,53,61,60)
>n=c(59,60,62,56,63,59,62,60)
>x=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839)
>n_y=n-y
>beetle.mat=cbind(y,n_y)
```

The logistic regression model can then be fitted using the command

R code (logistic regression)

```
>res.glm1=glm(beetle.mat~x, family=binomial(link="logit"))
```

The fitted values obtained from

R code (fitted values)

```
>fitted.values(res.glm1)
```

are estimated proportions of deaths in each group and so the fitted values for **y** need to be calculated as follows:

R code (fitted values)

```
>fit_p=c(fitted.values(res.glm1))
>fit_y=n*fit_p
```

Several alternative models can be fitted to the beetle mortality data. The results are shown in Table 7.4. Among these models the extreme value model appears to fit the data best. For Stata the relevant commands are

Stata code (probit model)

```
.glm y x, family(binomial n) link(probit)
```

Table 7.4 *Comparison of observed numbers killed with fitted values obtained from various dose-response models for the beetle mortality data. Deviance statistics are also given.*

Observed value of Y	Logistic model	Probit model	Extreme value model
6	3.46	3.36	5.59
13	9.84	10.72	11.28
18	22.45	23.48	20.95
28	33.90	33.82	30.37
52	50.10	49.62	47.78
53	53.29	53.32	54.14
61	59.22	59.66	61.11
60	58.74	59.23	59.95
D	11.23	10.12	3.45
$b_1(s.e.)$	−60.72(5.18)	−34.94(2.64)	−39.57(3.23)
$b_2(s.e.)$	34.27(2.91)	19.73(1.48)	22.04(1.79)

and

Stata code (extreme value model)

```
.glm y x, family(binomial n) link(cloglog)
```

For R they are

R code (probit model)

```
>res.glm2=glm(beetle.mat~x, family=binomial(link="probit"))
```

and

R code (extreme value model)

```
>res.glm3=glm(beetle.mat~x, family=binomial(link="cloglog"))
```

7.4 General logistic regression model

The simple linear logistic model $\log[\pi_i/(1-\pi_i)] = \beta_1 + \beta_2 x_i$ used in Example 7.3.1 is a special case of the general logistic regression model

$$\text{logit } \pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta},$$

where $\mathbf{x}_i$ is a vector of continuous measurements corresponding to covariates and dummy variables corresponding to factor levels and $\boldsymbol{\beta}$ is the parameter vector. This model is very widely used for analyzing data involving binary or Binomial responses and several explanatory variables. It provides a powerful technique analogous to multiple regression and ANOVA for continuous responses.

Maximum likelihood estimates of the parameters $\boldsymbol{\beta}$, and consequently of the probabilities $\pi_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$, are obtained by maximizing the log-likelihood function

$$l(\boldsymbol{\pi};\mathbf{y}) = \sum_{i=1}^{N}\left[y_i \log \pi_i + (n_i - y_i)\log(1-\pi_i) + \log \tbinom{n_i}{y_i}\right] \tag{7.4}$$

using the methods described in Chapter 4.

The estimation process is essentially the same whether the data are grouped as frequencies for each **covariate pattern** (i.e., observations with the same values of all the explanatory variables) or each observation is coded 0 or 1 and its covariate pattern is listed separately. If the data can be grouped, the response Y_i, the number of "successes" for covariate pattern i, may be modelled by the Binomial distribution. If each observation has a different covariate pattern, then $n_i = 1$ and the response Y_i is binary.

The deviance, derived in Section 5.6.1, is

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)\right]. \tag{7.5}$$

This has the form

$$D = 2\sum o \log \frac{o}{e}$$

where o denotes the observed "successes" y_i and "failures" $(n_i - y_i)$ from the cells of Table 7.1 and e denotes the corresponding estimated expected frequencies or fitted values $\widehat{y}_i = n_i\widehat{\pi}_i$ and $(n_i - \widehat{y}_i) = (n_i - n_i\widehat{\pi}_i)$. Summation is over all $2 \times N$ cells of the table.

Notice that D does not involve any nuisance parameters (like σ^2 for Normal response data), so goodness of fit can be assessed and hypotheses can be tested directly using the approximation

$$D \sim \chi^2(N-p),$$

where p is the number of parameters estimated and N the number of covariate patterns.

The estimation methods and sampling distributions used for inference depend on asymptotic results. For small studies or situations where there are few observations for each covariate pattern, the asymptotic results may be poor approximations. However software, such as StatXact and LogXact, has been developed using "exact" methods so that the methods described in this chapter can be used even when sample sizes are small.

Table 7.5 *Embryogenic anther data.*

Storage condition		Centrifuging force (g) 40	 150	 350
Control	y_{1k}	55	52	57
	n_{1k}	102	99	108
Treatment	y_{2k}	55	50	50
	n_{2k}	76	81	90

7.4.1 Example: Embryogenic anthers

The data in Table 7.5, cited by Wood (1978), are taken from Sangwan-Norrell (1977). They are numbers y_{jk} of embryogenic anthers of the plant species *Datura innoxia* Mill. obtained when numbers n_{jk} of anthers were prepared under several different conditions. There is one qualitative factor with two levels, a treatment consisting of storage at 3°C for 48 hours or a control storage condition, and one continuous explanatory variable represented by three values of centrifuging force. We will compare the treatment and control effects on the proportions after adjustment (if necessary) for centrifuging force.

The proportions $p_{jk} = y_{jk}/n_{jk}$ in the control and treatment groups are plotted against x_k, the logarithm of the centrifuging force, in Figure 7.4. The response proportions appear to be higher in the treatment group than in the control group, and at least for the treated group, the response decreases with centrifuging force.

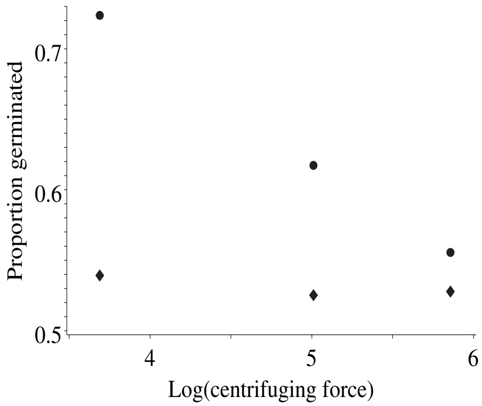

Figure 7.4 *Anther data from Table 7.5: proportion that germinated $p_{jk} = y_{jk}/n_{jk}$ plotted against* $\log_e$*(centrifuging force); dots represent the treatment condition and diamonds represent the control condition.*

We will compare three logistic models for π_{jk}, the probability of the anthers being embryogenic, where $j = 1$ for the control group and $j = 2$ for the treatment group and $x_1 = \log_e 40 = 3.689, x_2 = \log_e 150 = 5.011$, and $x_3 = \log_e 350 = 5.858$.

Model 1: logit $\pi_{jk} = \alpha_j + \beta_j x_k$ (i.e., different intercepts and slopes);
Model 2: logit $\pi_{jk} = \alpha_j + \beta x_k$ (i.e., different intercepts but the same slope);
Model 3: logit $\pi_{jk} = \alpha + \beta x_k$ (i.e., same intercept and slope).

These models were fitted by the method of maximum likelihood. The results are summarized in Table 7.6. To test the null hypothesis that the slope is the same for the treatment and control groups, we use $D_2 - D_1 = 2.591$. From the tables for the $\chi^2(1)$ distribution, the significance level is between 0.1 and 0.2, and so we could conclude that the data provide little evidence against the null hypothesis of equal slopes. On the other hand, the power of this test is very low and both Figure 7.4 and the estimates for Model 1 suggest that although the slope for the control group may be zero, the slope for the treatment group is negative. Comparison of the deviances from Models 2 and 3 gives a test for equality of the control and treatment effects after a common adjustment for centrifuging force: $D_3 - D_2 = 0.491$, which is consistent with the hypothesis that the storage effects are not different. The observed proportions and the corresponding fitted values for Models 1, 2 and 3 are shown in Table 7.7. Obviously, Model 1 fits the data very well but this is hardly surprising since four parameters have been used to describe six data points—such "over-fitting" is not recommended!

Table 7.6 *Maximum likelihood estimates and deviances for logistic models for the embryogenic anther data (standard errors of estimates in brackets).*

Model 1	Model 2	Model 3
$a_1 = 0.234(0.628)$ $a_2 - a_1 = 1.977(0.998)$ $b_1 = -0.023(0.127)$ $b_2 - b_1 = -0.319(0.199)$	$a_1 = 0.877(0.487)$ $a_2 - a_1 = 0.407(0.175)$ $b = -0.155(0.097)$	$a = 1.021(0.481)$ $b = -0.148(0.096)$
$D_1 = 0.028$	$D_2 = 2.619$	$D_3 = 8.092$

These results can be reproduced using Stata. If the control and treatment groups are recoded to 0 and 1, respectively (in a variable called newstor $= j - 1$), and an interaction term is created by multiplying this variable and the $\mathbf{x}$ vector, then the models can be fitted using the following commands:

Stata code (logistic models)

```
.glm y newstor x interaction, family(binomial n) link(logit)
.glm y newstor x, family(binomial n) link(logit)
.glm y x, family(binomial n) link(logit)
```

For R it is necessary to set up the response matrix with columns of values of $\mathbf{y}$ and $(\mathbf{n}-\mathbf{y})$ and then the commands are similar to those for Stata. For example, for Model 1 the interaction term can be explicitly used and it includes the main effects

R code (logistic model)

```
>res.glm3=glm(anther.mat~newstor*x,family=binomial(link="logit"))
```

Table 7.7 *Observed and expected frequencies for the embryogenic anther data for various models.*

Storage	Covariate	Observed	Expected frequencies		
condition	value	frequency	Model 1	Model 2	Model 3
Control	x_1	55	54.82	58.75	62.91
	x_2	52	52.47	52.03	56.40
	x_3	57	56.72	53.22	58.18
Treatment	x_1	55	54.83	51.01	46.88
	x_2	50	50.43	50.59	46.14
	x_3	50	49.74	53.40	48.49

7.5 Goodness of fit statistics

Instead of using maximum likelihood estimation we could estimate the parameters by minimizing the weighted sum of squares

$$S_w = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}$$

since $\mathrm{E}(Y_i) = n_i\pi_i$ and $\mathrm{var}(Y_i) = n_i\pi_i(1-\pi_i)$.

This is equivalent to minimizing the **Pearson chi-squared statistic**

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

where o represents the observed frequencies in Table 7.1, e represents the expected frequencies and summation is over all $2 \times N$ cells of the table. The reason is that

$$\begin{aligned} X^2 &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)} \\ &= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_w. \end{aligned}$$

When X^2 is evaluated at the estimated expected frequencies, the statistic is

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} \tag{7.6}$$

which is asymptotically equivalent to the deviances in (7.5),

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

The proof of the relationship between X^2 and D uses the Taylor series expansion of $s\log(s/t)$ about $s = t$, namely,

$$s\log\frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \dots \quad .$$

Thus,

$$\begin{aligned} D &= 2\sum_{i=1}^{N}\{(y_i - n_i\widehat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i} + [(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)] \\ &\quad + \frac{1}{2}\frac{[(n_i - y_i) - (n_i - n_i\widehat{\pi}_i)]^2}{n_i - n_i\widehat{\pi}_i} + \dots\} \\ &\cong \sum_{i=1}^{N}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} = X^2. \end{aligned}$$

The asymptotic distribution of D, under the hypothesis that the model is correct, is $D \sim \chi^2(N-p)$, therefore, approximately $X^2 \sim \chi^2(N-p)$. The choice between D and X^2 depends on the adequacy of the approximation to the $\chi^2(N-p)$ distribution. There is some evidence to suggest that X^2 is often better than D because D is unduly influenced by very small frequencies (Cressie and Read 1989). Both the approximations are likely to be poor, however, if the expected frequencies are too small (e.g., less than 1).

In particular, if each observation has a different covariate pattern so y_i is zero or one, then neither D nor X^2 provides a useful measure of fit. This can happen if the explanatory variables are continuous, for example. The most commonly used approach in this situation is due to Hosmer and Lemeshow (1980). Their idea was to group observations into categories on the basis of their predicted probabilities. Typically about 10 groups are used with approximately equal numbers of observations in each group. The observed numbers of successes and failures in each of the g groups are summarized as shown in Table 7.1. Then the Pearson chi-squared statistic for a $g \times 2$ contingency table is calculated and used as a measure of fit. We denote this **Hosmer–Lemeshow statistic** by X^2_{HL}. The sampling distribution of X^2_{HL} has been found by simulation to be approximately $\chi^2(g-2)$. The use of this statistic is illustrated in the example in Section 7.8.

Sometimes the log-likelihood function for the fitted model is compared with the log-likelihood function for a minimal model, in which the values π_i are all equal (in contrast to the saturated model which is used to define the deviance).

Under the **minimal model** $\widetilde{\pi} = (\Sigma y_i)\,/\,(\Sigma n_i)$. Let $\widehat{\pi}_i$ denote the estimated probability for Y_i under the model of interest (so the fitted value is $\widehat{y}_i = n_i\widehat{\pi}_i$). The statistic is defined by

$$C = 2\left[l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right) - l\left(\widetilde{\boldsymbol{\pi}};\mathbf{y}\right)\right],$$

where l is the log-likelihood function given by (7.4). Thus,

$$C = 2\sum\left[y_i \log\left(\frac{\widehat{y}_i}{n\widetilde{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - \widehat{y}_i}{n_i - n_i\widetilde{\pi}_i}\right)\right].$$

From the results in Section 5.2, the approximate sampling distribution for C is $\chi^2(p-1)$ if all the p parameters except the intercept term β_1 are zero (see Exercise 7.4). Otherwise C will have a non-central distribution. Thus C is a test statistic for the hypothesis that none of the explanatory variables is needed for a parsimonious model. C is sometimes called the **likelihood ratio chi-squared statistic**.

By analogy with R^2 for multiple linear regression (see Section 6.3.2) another statistic sometimes used is

$$\text{pseudo}R^2 = \frac{l\left(\widetilde{\boldsymbol{\pi}};\mathbf{y}\right) - l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right)}{l\left(\widetilde{\boldsymbol{\pi}};\mathbf{y}\right)},$$

which represents the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared with the minimal model. This statistic is produced by some statistical programs as a measure of goodness of fit. As for R^2, the sampling distribution of pseudo R^2 is not readily determined (so p-values cannot be obtained), and it increases as more parameters are added to the model. Therefore, various modifications of pseudo R^2 are used to adjust for the number of parameters (see, for example, Liao and McGee 2003). For logistic regression R^2-type measures often appear alarmingly small even when other measures suggest that the model fits the data well. The reason is that pseudo R^2 is a measure of the predictability of individual outcomes Y_i rather than the predictability of all the event rates (Mittlbock and Heinzl 2001).

The **Akaike information criterion** AIC and the Schwartz or **Bayesian information criterion** BIC are other goodness of fit statistics based on the log-likelihood function with adjustment for the number of parameters estimated and for the amount of data. These statistics are usually defined as follows:

$$AIC = -2l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right) + 2p \tag{7.7}$$

$$BIC = -2l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right) + 2p \times \ln(\text{number of observations}),$$

where p is the number of parameters estimated. The statistical software R uses this definition of AIC, for example. However, the versions used by Stata are somewhat different:

$$AIC_{Stata} = (-2l\left(\widehat{\boldsymbol{\pi}};\mathbf{y}\right) + 2p)/N,$$

where N is the number of subgroups in Table 7.1, and

$$BIC_{Stata} = D - (N - p) \ln \sum_{i=1}^{N} n_i,$$

where D is the deviance for the model, $(N - p)$ is the corresponding degrees of freedom and $\sum_{i=1}^{N} n_i$ is the total number of observations.

Note that the statistics (except for pseudo R^2) discussed in this section summarize how well a particular model fits the data. So a small value of the statistic and, hence, a large p-value, indicates that the model fits well. These statistics are not usually appropriate for testing hypotheses about the parameters of nested models, but they can be particularly useful for comparing models that are not nested.

For the logistic model for the beetle mortality example (Section 7.3.1), the log-likelihood for the model with no explanatory variable is $l(\widetilde{\boldsymbol{\pi}}; \mathbf{y}) = (-167.5203 - 155.2002) = -322.7205$, while $l(\widehat{\boldsymbol{\pi}}; \mathbf{y}) = (-167.5203 - 18.7151) = -186.2354$, so the statistic $C = 2 \times (-186.2354 - (-322.7205)) = 272.970$ with one degree of freedom, indicating that the slope parameter β_1 is definitely needed! The pseudo R^2 value is $2 \times (-322.72051 - (-186.23539))/2 \times (-322.72051) = 0.4229$ indicating reasonable but not excellent fit. The usual value for $AIC = -2 \times (-18.7151) + 2 \times 2 = 41.430$—this is the value given by R, for example. Stata gives $AIC_{Stata} = 41.430/8 = 5.179$ and $BIC_{Stata} = 11.2322 - (8 - 2) \times \ln 481 = -25.823$. All these measures show marked improvements when the extreme value model is fitted compared with the logistic model.

7.6 Residuals

For logistic regression there are two main forms of residuals corresponding to the goodness of fit measures D and X^2. If there are m different covariate patterns, then m residuals can be calculated. Let Y_k denote the number of successes, n_k the number of trials and $\widehat{\pi}_k$ the estimated probability of success for the kth covariate pattern.

The **Pearson**, or **chi-squared**, **residual** is

$$X_k = \frac{(y_k - n_k \widehat{\pi}_k)}{\sqrt{n_k \widehat{\pi}_k (1 - \widehat{\pi}_k)}} \quad , k = 1, \ldots, m. \tag{7.8}$$

From (7.6), $\sum_{k=1}^{m} X_k^2 = X^2$, the Pearson chi-squared goodness of fit statistic. The **standardized Pearson residuals** are

$$r_{Pk} = \frac{X_k}{\sqrt{1 - h_k}},$$

where h_k is the leverage, which is obtained from the hat matrix (see Section 6.2.6).

Deviance residuals can be defined similarly,

$$d_k = \text{sign}(y_k - n_k\widehat{\pi}_k)\left\{2\left[y_k \log\left(\frac{y_k}{n_k\widehat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\widehat{\pi}_k}\right)\right]\right\}^{1/2} \tag{7.9}$$

where the term $\text{sign}(y_k - n_k\widehat{\pi}_k)$ ensures that d_k has the same sign as X_k. From equation (7.5), $\sum_{k=1}^{m} d_k^2 = D$, the deviance. Also standardized deviance residuals are defined by

$$r_{Dk} = \frac{d_k}{\sqrt{1 - h_k}}.$$

Pearson and deviance residuals can be used for checking the adequacy of a model, as described in Section 2.3.4. For example, they should be plotted against each continuous explanatory variable in the model to check if the assumption of linearity is appropriate and against other possible explanatory variables not included in the model. They should be plotted in the order of the measurements, if applicable, to check for serial correlation. Normal probability plots can also be used because the standardized residuals should have, approximately, the standard Normal distribution N(0, 1), provided the numbers of observations for each covariate pattern are not too small.

If the data are binary, or if n_k is small for most covariate patterns, then there are few distinct values of the residuals and the plots may be relatively uninformative. In this case, it may be necessary to rely on the aggregated goodness of fit statistics X^2 and D and other diagnostics (see Section 7.7).

For more details about the use of residuals for Binomial and binary data see Chapter 5 of Collett (2003a), for example.

7.7 Other diagnostics

By analogy with the statistics used to detect influential observations in multiple linear regression, the statistics delta-beta, delta-chi-squared and delta-deviance are also available for logistic regression (see Section 6.2.7).

For binary or Binomial data there are additional issues to consider. The first is to check the choice of the link function. Brown (1982) developed a test for the logit link which is implemented in some software. The approach suggested by Aranda-Ordaz (1981) is to consider a more general family of link functions

$$g(\pi, \alpha) = \log\left[\frac{(1-\pi)^{-\alpha} - 1}{\alpha}\right].$$

If $\alpha = 1$, then $g(\pi) = \log[\pi/(1-\pi)]$, the logit link. As $\alpha \to 0$, then $g(\pi) \to \log[-\log(1-\pi)]$, the complementary log-log link. In principle, an optimal value of α can be estimated from the data, but the process requires several steps. In the absence of suitable software to identify the best link function, it is advisable to experiment with several alternative links.

The second issue in assessing the adequacy of models for binary or Binomial data is **overdispersion**. Observations Y_i may have observed variance greater than Binomial variance $n_i\pi_i(1-\pi_i)$, or equivalently $\text{var}(\widehat{\pi}_i)$ may be greater than $\pi_i(1-\pi_i)/n_i$. There is an indicator of this problem if the deviance D is much greater than the expected value of $N-p$. This could be due to inadequate specification of the model (e.g., relevant explanatory variables have been omitted or the link function is incorrect) or to a more complex structure (see Exercise 7.5). One approach is to include an extra parameter ϕ in the model so that $\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$. This is implemented in various ways in statistical software. For example, in R there is an option in `glm` to specify a quasibinomial distribution instead of a Binomial distribution. Another possible explanation for overdispersion is that the Y_i's are not independent. For example if the binary responses counted by Y_i are not independent, the effective number of trials n', will be less than n so that $\text{var}(\widehat{\pi}_i) = \pi_i(1-\pi_i)/n'_i > \pi_i(1-\pi_i)/n_i$. Methods for modelling correlated data are outlined in Chapter 11. For a detailed discussion of overdispersion for Binomial data, see, for example, Collett (2003a, Chapter 6).

7.8 Example: Senility and WAIS

A sample of elderly people was given a psychiatric examination to determine whether symptoms of senility were present. Other measurements taken at the same time included the score on a subset of the Wechsler Adult Intelligent Scale (WAIS). The data are shown in Table 7.8. The data in Table 7.8 are binary although some people have the same WAIS scores and so there are $m = 17$ different covariate patterns (see Table 7.9). Let Y_i denote the number of people with symptoms among n_i people with the ith covariate pattern. The logistic regression model

Table 7.8 *Symptoms of senility (s=1 if symptoms are present and s=0 otherwise) and WAIS scores (x) for N=54 people.*

x	s	x	s	x	s	x	s	x	s
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

Table 7.9 *Covariate patterns and responses, estimated probabilities ($\widehat{\pi}$), Pearson residuals (X) and deviance residuals (d) for senility and WAIS.*

x	y	n	$\widehat{\pi}$	X	d
4	1	2	0.751	−0.826	−0.766
5	0	1	0.687	0.675	0.866
6	1	2	0.614	−0.330	−0.326
7	1	3	0.535	0.458	0.464
8	0	2	0.454	1.551	1.777
9	4	6	0.376	−0.214	−0.216
10	5	6	0.303	−0.728	−0.771
11	5	6	0.240	−0.419	−0.436
12	2	2	0.186	−0.675	−0.906
13	5	6	0.142	0.176	0.172
14	5	7	0.107	1.535	1.306
15	3	3	0.080	−0.509	−0.705
16	4	4	0.059	−0.500	−0.696
17	1	1	0.043	−0.213	−0.297
18	1	1	0.032	−0.181	−0.254
19	1	1	0.023	−0.154	−0.216
20	1	1	0.017	−0.131	−0.184
Sum	40	54			
		Sum of squares		8.084*	9.418*

* Sums of squares differ slightly from the goodness of fit statistics X^2 and D mentioned in the text due to rounding errors.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 x_i; \quad Y_i \sim \text{Bin}(n_i, \pi_i) \quad i = 1, \ldots, m,$$

was fitted with the following results:

$b_1 = 2.404$, standard error $(b_1) = 1.192$;
$b_2 = -0.3235$, standard error $(b_2) = 0.1140$;
$X^2 = \sum X_i^2 = 8.083$ and $D = \sum d_i^2 = 9.419$.

As there are $m = 17$ covariate patterns (different values of x, in this example) and $p = 2$ parameters, X^2 and D can be compared with $\chi^2(15)$ (by these criteria the model appears to fit well).

Figure 7.5 shows the observed relative frequencies y_i/n_i for each covariate pattern and the fitted probabilities $\widehat{\pi}_i$ plotted against WAIS score, x (for $i = 1, \ldots, m$). The model appears to fit better for higher values of x.

Table 7.9 shows the covariate patterns, estimates $\widehat{\pi}_i$ and the corresponding chi-squared and deviance residuals calculated using equations (7.8) and (7.9), respectively.

The residuals and associated residual plots (not shown) do not suggest that

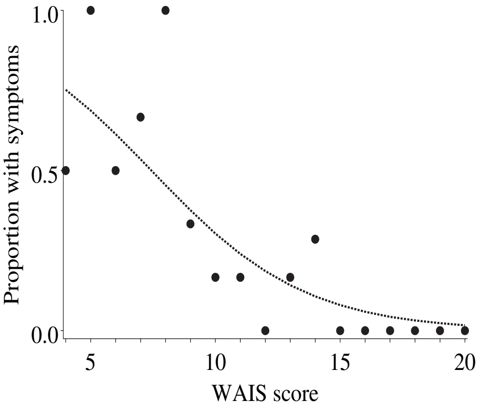

Figure 7.5 *Relationship between presence of symptoms and WAIS score from data in Tables 7.8 and 7.9; circles represent observed proportions and dotted line represents estimated probabilities.*

there are any unusual observations but the small numbers of observations for each covariate value make the residuals difficult to assess. The Hosmer–Lemeshow approach provides some simplification; Table 7.10 shows the data in categories defined by grouping values of $\widehat{\pi}_i$ so that the total numbers of observations per category are approximately equal. For this illustration, $g =$ 3 categories were chosen. The expected frequencies are obtained from the values in Table 7.9; there are $\sum n_i\widehat{\pi}_i$ with symptoms and $\sum n_i\left(1-\widehat{\pi}_i\right)$ without symptoms for each category. The Hosmer–Lemeshow statistic X^2_{HL} is obtained by calculating $X^2 = \Sigma\left[(o-e)^2/e\right]$, where the observed frequencies, o, and expected frequencies, e, are given in Table 7.10 and summation is over all 6 cells of the table; $X^2_{HL} = 1.15$, which is not significant when compared with the $\chi^2(1)$ distribution.

For the minimal model, without x, the maximum value of the log-likelihood function is $l(\widetilde{\boldsymbol{\pi}},\mathbf{y}) = -30.9032$. For the model with x, the corresponding value is $l(\widehat{\boldsymbol{\pi}},\mathbf{y}) = -25.5087$. Therefore, from Section 7.5, $C = 10.789$, which is highly significant compared with $\chi^2(1)$, showing that the slope parameter is non-zero. Also pseudo $R^2 = 0.17$ which suggests the model does not predict the outcomes for individuals particularly well even though the residuals for all the covariate patterns are small and the other summary measures suggest the fit is good.

These data illustrate the differences between fitting the model to binary (ungrouped) and Binomial (grouped) data. The relevant Stata commands are

Stata code (binary model)

```
.glm s x, family(binomial 1) link(logit)
```

and

Table 7.10 *Hosmer–Lemeshow test for data in Table 7.9: observed frequencies (o) and expected frequencies (e) for numbers of people with or without symptoms, grouped by values of $\widehat{\pi}$.*

Values of $\widehat{\pi}$		≤ 0.107	0.108–0.303	> 0.303
Corresponding values of x		14–20	10–13	4–9
Number of people	o	2	3	9
with symptoms	e	1.335	4.479	8.186
Number of people	o	16	17	7
without symptoms	e	16.665	15.521	7.814
Total number of people		18	20	16

Stata code (Binomial model)

```
.gen n=1
.collapse (sum) n s, by(x)
.glm s x, family(binomial n) link(logit)
```

For R the function for the ungrouped data is

R code (binary model)

```
>res.glm=glm(s~x, family=binomial(link="logit"))
```

For grouped data a matrix of columns of “successes” and “failures” has to be constructed, then the function is

R code (Binomial model)

```
>res.glm=glm(waisgrp.mat~x, family=binomial(link="logit"))
```

For either form of the data the values of the estimates and their standard errors are the same, but the measures of goodness of fit differ as shown in Table 7.11. In this table M_0 refers to a model with only a constant (i.e., no effect of WAIS scores x), M_1 refers to the model with a constant and an effect of x, and M_S refers to the saturated model with a parameter for every observation. The statistics pseudo R^2 and AIC can be interpreted to indicate that the model M_1 is better able to predict the group outcomes (i.e., event rates) than to predict individual outcomes. However, the differences caused by the form in which the data are analyzed indicate that caution is needed when assessing adequacy of logistic regression models using these measures.

7.9 Exercises

7.1 The number of deaths from leukemia and other cancers among survivors of the Hiroshima atom bomb are shown in Table 7.12, classified by the

Table 7.11 *Measures of goodness of fit for models for data in Table 7.9 obtained using ungrouped and grouped observations.*

	Ungrouped	Grouped
Number of observations	54	17
M_0 log-likelihood	−30.90316	−17.29040
M_1 log-likelihood	−25.50869	−11.89593
M_S log-likelihood	0.0	−7.18645
M_0 deviance	61.80632	20.20791
M_1 deviance	51.01738	9.14897
$M_0 - M_1$ deviance	10.7889	10.7889
M_1 pseudo R^2	0.1746	0.3120
M_1 AIC	55.0173	27.7919
M_1 AIC_{Stata}	1.01884	1.6348

radiation dose received. The data refer to deaths during the period 1950–1959 among survivors who were aged 25 to 64 years in 1950 (from data set 13 of Cox and Snell 1981, attributed to Otake 1979).

(a) Obtain a suitable model to describe the dose–response relationship between radiation and the proportional cancer mortality rates for leukemia.

(b) Examine how well the model describes the data.

(c) Interpret the results.

Table 7.12 *Deaths from leukemia and other cancers classified by radiation dose received from the Hiroshima atomic bomb.*

	Radiation dose (rads)					
Deaths	0	1–9	10–49	50–99	100–199	200+
Leukemia	13	5	5	3	4	18
Other cancers	378	200	151	47	31	33
Total cancers	391	205	156	50	35	51

7.2 **Odds ratios**. Consider a 2×2 contingency table from a prospective study in which people who were or were not exposed to some pollutant are followed up and, after several years, categorized according to the presence or absence of a disease. Table 7.13 shows the probabilities for each cell. The odds of disease for either exposure group is $O_i = \pi_i/(1-\pi_i)$, for

$i = 1, 2$, and so the odds ratio

$$\phi = \frac{O_1}{O_2} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

is a measure of the relative likelihood of disease for the exposed and not exposed groups.

Table 7.13 *2×2 table for a prospective study of exposure and disease outcome.*

	Diseased	Not diseased
Exposed	π_1	$1-\pi_1$
Not exposed	π_2	$1-\pi_2$

(a) For the simple logistic model $\pi_i = e^{\beta_i}/(1+e^{\beta_i})$, show that if there is no difference between the exposed and not exposed groups (i.e., $\beta_1 = \beta_2$), then $\phi = 1$.

(b) Consider J 2×2 tables like Table 7.13, one for each level x_j of a factor, such as age group, with $j = 1, \ldots, J$. For the logistic model

$$\pi_{ij} = \frac{\exp(\alpha_i + \beta_i x_j)}{1 + \exp(\alpha_i + \beta_i x_j)}, \qquad i = 1, 2, \quad j = 1, \ldots, J.$$

Show that $\log \phi$ is constant over all tables if $\beta_1 = \beta_2$ (McKinlay 1978).

7.3 Tables 7.14 and 7.15 show the survival 50 years after graduation of men and women who graduated each year from 1938 to 1947 from various Faculties of the University of Adelaide (data compiled by J.A. Keats). The columns labelled S contain the number of graduates who survived and the columns labelled T contain the total number of graduates. There were insufficient women graduates from the Faculties of Medicine and Engineering to warrant analysis.

(a) Are the proportions of graduates who survived for 50 years after graduation the same all years of graduation?

(b) Are the proportions of male graduates who survived for 50 years after graduation the same for all Faculties?

(c) Are the proportions of female graduates who survived for 50 years after graduation the same for Arts and Science?

(d) Is the difference between men and women in the proportion of graduates who survived for 50 years after graduation the same for Arts and Science?

7.4 Let $l(\mathbf{b}_{\min})$ denote the maximum value of the log-likelihood function for the minimal model with linear predictor $\mathbf{x}^T\boldsymbol{\beta} = \beta_1$, and let $l(\mathbf{b})$ be the corresponding value for a more general model $\mathbf{x}^T\boldsymbol{\beta} = \beta_1 + \beta_2 x_1 + \ldots + \beta_p x_{p-1}$.

Table 7.14 *Fifty years survival for men after graduation from the University of Adelaide.*

Year	Faculty							
of	Medicine		Arts		Science		Engineering	
graduation	S	T	S	T	S	T	S	T
1938	18	22	16	30	9	14	10	16
1939	16	23	13	22	9	12	7	11
1940	7	17	11	25	12	19	12	15
1941	12	25	12	14	12	15	8	9
1942	24	50	8	12	20	28	5	7
1943	16	21	11	20	16	21	1	2
1944	22	32	4	10	25	31	16	22
1945	12	14	4	12	32	38	19	25
1946	22	34			4	5		
1947	28	37	13	23	25	31	25	35
Total	177	275	92	168	164	214	100	139

Table 7.15 *Fifty years survival for women after graduation from the University of Adelaide.*

Year	Faculty			
of	Arts		Science	
graduation	S	T	S	T
1938	14	19	1	1
1939	11	16	4	4
1940	15	18	6	7
1941	15	21	3	3
1942	8	9	4	4
1943	13	13	8	9
1944	18	22	5	5
1945	18	22	16	17
1946	1	1	1	1
1947	13	16	10	10
Total	126	157	58	61

(a) Show that the likelihood ratio chi-squared statistic is

$$C = 2\left[l(\mathbf{b}) - l(\mathbf{b}_{\min})\right] = D_0 - D_1,$$

where D_0 is the deviance for the minimal model and D_1 is the deviance for the more general model.

(b) Deduce that if $\beta_2 = \ldots = \beta_p = 0$, then C has the central chi-squared distribution with $(p-1)$ degrees of freedom.

7.5 Let Y_i be the number of successes in n_i trials with

$$Y_i \sim \text{Bin}(n_i, \pi_i),$$

where the probabilities π_i have a Beta distribution

$$\pi_i \sim \text{Be}(\alpha, \beta).$$

The probability density function for the beta distribution is $f(x;\alpha,\beta) = x^{\alpha-1}(1-x)^{(\beta-1)}/B(\alpha,\beta)$ for x in [0, 1], $\alpha > 0, \beta > 0$ and the beta function $B(\alpha,\beta)$ defining the normalizing constant required to ensure that $\int_0^1 f(x;\alpha,\beta)dx = 1$. It can be shown that $\text{E}(X) = \alpha/(\alpha+\beta)$ and $\text{var}(X) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$. Let $\theta = \alpha/(\alpha+\beta)$, and hence, show that

(a) $\text{E}(\pi_i) = \theta$

(b) $\text{var}(\pi_i) = \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)$

(c) $\text{E}(Y_i) = n_i\theta$

(d) $\text{var}(Y_i) = n_i\theta(1-\theta)[1+(n_i-1)\phi]$ so that $\text{var}(Y_i)$ is larger than the Binomial variance (unless $n_i = 1$ or $\phi = 0$).