DEPARTMENT OF STATISTICS

Paper STATS 330

ADVANCED STATISTICAL MODELLING

Chapter 5 of Course notes by Arden Miller, Alan Lee, Chris Triggs and Ross Ihaka

February 1995, revised September 2002

Contents

Chapter 5

Categorical Responses

5.1 Introduction

In this chapter we deal with models that are appropriate for categorical responses. First, we consider logistic regression models where the response is a binary variable that indicates whether a particular event has occurred or not. Then we consider the analysis of contingency tables where the response can be considered to be a count To accomplish this, we need to make some adaptations to the techniques that were covered in earlier chapters.

5.2 Logistic regression analysis

Example 1. The data in Table 5.2 are taken from a book by Hosmer and Lemeshow (*Applied Logistic Regression*, Wiley (1989)) and were gathered by examining 100 randomly selected patients and recording the presence or absence of coronary heart disease (`chd`, absent = 0 present = 1) and the patient's age (`age`). Our aim is to use the patients age to predict the probability that a patient has coronary heart disease. It may be tempting to fit an ordinary regression model using `chd` as the response and `age` as an explanatory variable but a moments thought should convince us this is not a good idea. First, a key assumption of ordinary regression is that the response has a Normal distribution. Clearly since `chd` can only take values 0 or 1 its distribution is not even close to Normal – a binomial distribution would be a much more suitable model. Second as we want to predict a probability we would like a model that produces predictions that are between 0 and 1. There is no guarantee that this will be the case if we use ordinary regression.

We will use a technique known as logistic regression to create a suitable model. Let π represent the probability a patient will have CHD. The logistic model that relates π to the age of the patient has the form:

$$\pi = \frac{\exp(\beta_0 + \beta_1 \texttt{age})}{1 + \exp(\beta_0 + \beta_1 \texttt{age})}. \tag{5.1}$$

Notice that this model ensures that $0 < \pi < 1$, since (i) $\exp(\beta_0 + \beta_1\texttt{age}) > 0$ and (ii) $1 + \exp(\beta_0 + \beta_1\texttt{age}) > \exp(\beta_0 + \beta_1\texttt{age})$.

In general, a logistic model that uses a single numeric explanatory variable has a sigmoidal (S or reverse S) shape. Examples are shown in Figure 5.1. The values of β_0 and β_1 determine the characteristics of the logistic model:

Table 5.1: CHD and AGE for 100 patients

age	chd	age	chd	age	chd	age	chd	age	chd	age	chd
20	0	23	0	24	0	25	0	25	1	26	0
26	0	28	0	28	0	29	0	30	0	30	0
30	0	30	0	30	0	31	1	32	0	32	0
33	0	33	0	34	0	34	0	34	1	34	0
34	0	35	0	35	0	36	1	36	0	36	0
37	0	37	1	37	0	38	0	38	0	39	0
39	1	40	0	40	1	41	0	41	0	42	0
42	0	42	0	41	1	43	0	43	0	43	1
44	0	44	0	44	1	44	1	45	0	45	1
46	0	46	1	47	0	47	0	47	1	48	0
48	1	48	1	49	0	49	0	49	1	50	0
50	0	50	1	52	0	52	1	53	1	53	1
54	1	55	0	55	1	55	1	56	1	56	1
56	1	57	0	57	0	57	1	57	1	57	1
57	1	57	1	58	1	58	1	59	1	59	1
60	0	60	0	61	1	62	1	62	1	63	1
64	0	64	1	65	1	69	1				

1. The sign of β_1 determines whether π increases ($\beta_1 > 0$) or decreases ($\beta_1 < 0$) as the value of the explanatory variable increases.

2. The absolute value of β_1 determines how sensitive π is to changes in the explanatory variable – the larger the value of $\|\beta_1\|$ the more sensitive π is to changes in the value of the explanatory variable. The curves in Figure 5.1 would become steeper if $\|\beta_1\|$ was increased and more gradual if $\|\beta_1\|$ was decreased.

3. The value of β_0 determines the value of π when the explanatory variable is 0. Thus the value of β_0 controls the location of the logistic curve on the x-axis. If we were to change β_0 (but hold β_1 constant) then the curves in Figure 5.1 would stay the same shape but would be shifted left or right along the x-axis.

5.2.1 Fitting the model

For ordinary regression models, least squares is the standard method used to estimate the model parameters. Least squares works well when the response is Normal and has constant variance but it is not so good for fitting logistic regression models. Instead, we will use a method known as *maximum likelihood*. The idea behind maximum likelihood estimation is to select the values for the parameters that are most compatible with the observed data. Put another way, the fitted parameters are those values that make the observed data most likely to occur. Note that if the response is Normal, then the maximum likelihood method gives the same estimated coefficients as least squares.

Consider the CHD data. We assume that for each patient our response variable `chd` has a binary distribution (binomial with $n = 1$). Further we assume that π_i, $\Pr(\texttt{chd} = 1)$ for patient i, is related to the age of patient i in the manner described by equation 5.1. Thus

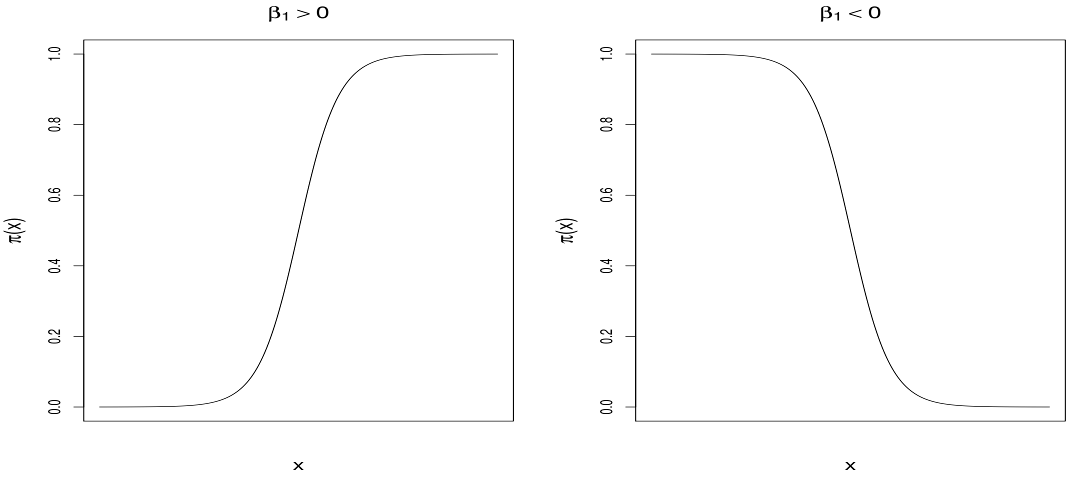

Figure 5.1: Shape of the logistic curve $\pi(x) = \exp(\beta_0 + \beta_1 x)/1 + \exp(\beta_0 + \beta_1 x)$. (i) $\beta_1 > 0$, (ii) $\beta_1 < 0$.

we have the following probability function for $\texttt{chd}_i$:

$$\begin{aligned} \Pr(\texttt{chd}_i = 1) &= \tfrac{\exp(\beta_0+\beta_1 \texttt{age}_i)}{1+\exp(\beta_0+\beta_1 \texttt{age}_i)}, \\ \Pr(\texttt{chd}_i = 0) &= 1 - \tfrac{\exp(\beta_0+\beta_1 \texttt{age}_i)}{1+\exp(\beta_0+\beta_1 \texttt{age}_i)} = \tfrac{1}{1+\exp(\beta_0+\beta_1 \texttt{age}_i)}. \end{aligned} \tag{5.2}$$

Provided that we assume that the observations are independent, we can calculate the probability of getting the observed set of values for the $\texttt{chd}_i$'s by multiplying the individual probabilities:

$$\Pr(\texttt{chd}_1 = 0,\, \texttt{chd}_2 = 0, \ldots, \texttt{chd}_{100} = 1) = \Pr(\texttt{chd}_1 = 0)\times\Pr(\texttt{chd}_2 = 0)\times\ldots\times\Pr(\texttt{chd}_{100} = 1).$$

Thus by plugging in the appropriate expression from (5.2) for each term we can get an expression for the probability of getting the observed data that is a function of β_0 and β_1. This expression is called the *likelihood function* and is denoted by $L(\beta_0, \beta_1)$. The *maximum likelihood estimates* of β_0 and β_1 are the values of β_0 and β_1 that maximise $L(\beta_0, \beta_1)$. For our current example $L(\beta_0, \beta_1)$ is quite a complicated expression and it would be very tedious to find the maximum likelihood estimates (MLE's) by hand. Of course, computer programmes exist that use numerical algorithms to find the MLE's and we will use one of these.

Example 2. To fit the logistic model to our CHD data, we can use the `glm` function in R. GLM stands for "generalised linear model" and designates a class of statistical models that includes both the logistic model and the ordinary (Normal) regression model. GLM's are discussed in more detail in Section 5.4.

The `glm` function works in much the same manner as the `lm` function. Suppose we have entered the data from Table 5.1 into a data frame in R named `chd.df`. Then we can produce a "glm object" that contains the fitted logistic regression model by typing

```
> chd.glm<-glm(chd~age,family=binomial,data=chd.df)
```

Notice that the `family=binomial` option is included in this command to indicate that the response variable is assumed to have a binomial distribution. We can examine the fitted model using the `summary` function just as we did for regression objects produced using `lm`.

```
> summary(chd.glm)

Call:
glm(formula = chd ~ age, family = binomial, data = chd.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9686  -0.8480  -0.4607   0.8262   2.2794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.2784     1.1296  -4.673 2.97e-06 ***
age           0.1103     0.0240   4.596 4.30e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.68  on 98  degrees of freedom
AIC: 111.68

Number of Fisher Scoring iterations: 3
```

This summary is similar to what would be produced for a lm object but there are several differences. For now, just note that the MLE's for the two parameters are in the `Estimate` column in the section headed `Coefficients` (the other parts of this output will be discussed later).

The fitted logistic regression model can be written in a number of different forms. We will consider three: the "logistic form", the "odds form", and the "logit form". Keep in mind that these three forms are simply different ways of expressing the same model.

The logistic form of the fitted model is obtaining by putting the estimates $\widehat{\beta}_0 = -5.2784$ and $\widehat{\beta}_1 = 0.1103$ into (5.1):

$$\widehat{\pi} = \frac{\exp(-5.2784 + 0.1103\ \texttt{age})}{1 + \exp(-5.2784 + 0.1103\ \texttt{age})}.$$

This model relates the probability a subject has CHD, π, to the subject's age. The fitted model is superimposed on a plot of `chd` versus `age` in Figure 5.2. Note that as age increases the proportion of subjects with chd ($\texttt{chd} = 1$) increases and that this trend is captured by the fitted logistic model. To predict π for a specific value of age we can simply plug the value into this formula or we can use the `predict` function in R. For example to get the estimated probability of CHD for $\texttt{age} = 45$:

```
> predict(chd.glm,data.frame(age=45),type="response")
[1] 0.4221367
```

The second way we can write the logistic model is as an expression for the odds of `chd`. Recall that the odds an event E occurs is the probability E occurs divided by the probability E doesn't occur: $\text{odds(E)} = \Pr(\text{E})/\left(1 - \Pr(\text{E})\right)$. Thus if $\text{odds(E)} = 2$ this means that the probability E occurs is twice the probability that E does not occur which implies

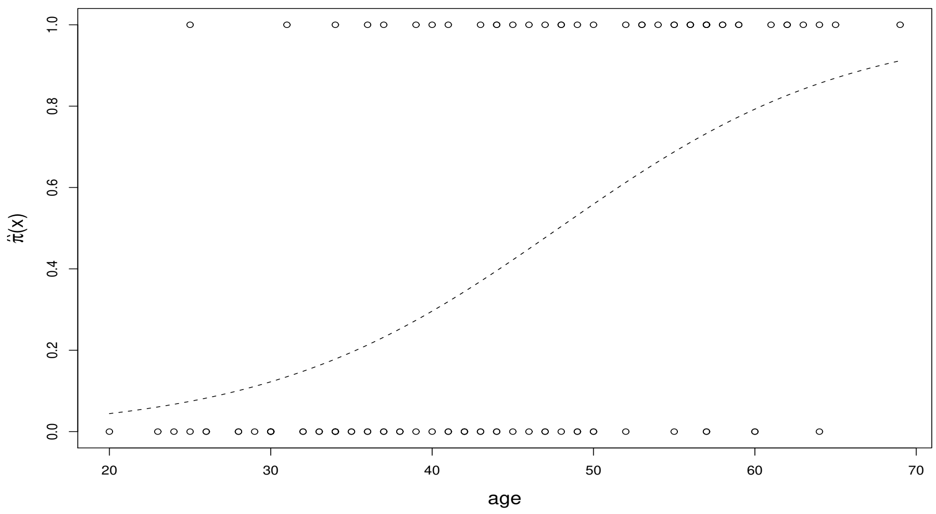

Figure 5.2: Fitted logistic regression model for the CHD example.

$\Pr(\text{E}) = 2/3$. The odds of CHD can be estimated by $\widehat{\text{odds}}(\texttt{chd}) = \widehat{\pi}/(1-\widehat{\pi})$. Using the expression for $\widehat{\pi}$ from our fitted model and a bit of algebraic manipulation gives:

$$\widehat{\text{odds}}(\texttt{chd}) = \exp(-5.2784 + 0.1103\texttt{age}).$$

One advantage of this form of the logistic regression model is that it allows the coefficient of age to be interpreted in a convenient manner. Suppose we fix age at a specified value $\texttt{age}_o$ and consider the effect on $\widehat{\text{odds}}(\texttt{chd})$ of increasing age by 1 year. For $\texttt{age} = \texttt{age}_o$ we can write:

$$\begin{aligned}\widehat{\text{odds}}(\texttt{chd}|\texttt{age} = \texttt{age}_o) &= \exp(-5.2784 + 0.1103\,\texttt{age}_o) \\ &= \exp(-5.2784) \times \exp(0.1103\,\texttt{age}_o).\end{aligned}$$

For $\texttt{age} = \texttt{age}_o + 1$ we have:

$$\begin{aligned}\widehat{\text{odds}}(\texttt{chd}|\texttt{age} = \texttt{age}_o + 1) &= \exp(-5.2784 + 0.1103\,(1 + \texttt{age}_o)) \\ &= \exp(-5.2784) \times \exp(0.1103) \times \exp(0.1103\,\texttt{age}_o). \\ &= \exp(0.1103) \times \widehat{\text{odds}}(\texttt{chd}|\texttt{age} = \texttt{age}_o).\end{aligned}$$

Thus for each increase of 1 year in age, the estimated odds of CHD is multiplied by $\exp(0.1103) = 1.117$.

The third way we can write the logistic regression function is in the *logit* form. The logit function is simply the log-odds:

$$\text{logit}(\text{E}) = \log \frac{\Pr(\text{E})}{1 - \Pr(\text{E})}.$$

Thus by taking the log of each side of the odds form of the fitted model we get the logit form:

$$\widehat{\text{logit}}(\texttt{chd}) = -5.2784 + 0.1103\texttt{age}.$$

Note that now the right hand side of the model is simply a linear function of `age`. Thus for each increase in `age` of one unit $\widehat{\text{logit}}$(`chd`) increases by 0.1103.

We now consider statistical inference for the regression coefficients. The estimates of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have sampling distributions that are asymptotically Normal and thus we can use the Normal distribution to produce (approximate) confidence intervals and perform hypothesis tests. The section headed `Coefficients` from the output generated by the `summary` function provides us with the estimated values of the parameters, the standard errors of the estimates, a test statistic for testing $\beta_j = 0$, and the p-value for this test. Thus the line for `age` indicates that the maximum likelihood estimate of β_1 is $\widehat{\beta}_1 = 0.1103$ and that the standard error of this estimate is $s.e.(\widehat{\beta}_1) = 0.0240$. Further it indicates that we can test $H_0{:}\beta_1 = 0$ using the statistic $z = \widehat{\beta}_1 / s.e.(\widehat{\beta}_1) = 4.596$ which produces a p-value of `4.30e-06` (very strong evidence against H_0). The given p-value represents $2 \times \Pr(Z \geq 4.596)$ where $Z \sim N(0,1)$. A 95% confidence interval for β_1 can be calculated as

$$\begin{array}{rcl} \hat{\beta}_1 & \pm & s.e.(\hat{\beta}_1) \times 1.960 \\ 0.1103 & \pm & 0.0240 \times 1.960 \\ 0.1103 & \pm & 0.0470. \end{array}$$

since $\Pr(-1.96 \leq Z \leq 1.96) = .95$.

5.2.2 Multiple logistic regression

The logistic regression model can easily be extended to situations where there is more than one explanatory variable. As before we wish to model the probability that an event E occurs which we will denote by π. However now π will depend on the values of of several explanatory variables $X_1, \ldots, X_k$. The multiple logistic regression model is:

$$\pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}.$$

This model can also be written in the odds form,

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k),$$

or in the logit form,

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

The parameters are estimated using the maximum likelihood method. As before the likelihood function, L, uses the assumed model to express the probability of getting the observed data as function of the unknown parameters β_0, β_1, $\ldots$, β_k. The values of the parameters that maximize L (the MLE's) can be found using the `glm` function in R.

Example 3. The data in Table 5.1 were collected on 83 children undergoing corrective spinal surgery. The objective was to identify risk factors for kyphosis (flexing of the spine) following surgery. The variable `Kyphosis` is a binary response with 1 indicating the presence of kyphosis and 0 the absence of kyphosis.

The risk factors studied were age in months (`Age`), the starting vertebrae level of the surgery (`Start`), and the number of vertebrae involved (`Number`). The data are in a data frame called `kypho.data`.

The first step is to plot the variables:

Table 5.2: Data for Example 5.

Age	Start	Number	Kyphosis	Age	Start	Number	Kyphosis
71	5	3	0	158	14	3	0
128	5	4	1	2	1	5	0
1	15	4	0	1	16	2	0
61	17	2	0	37	16	3	0
113	16	2	0	59	12	6	1
82	14	5	1	148	16	3	0
18	2	5	0	1	12	4	0
243	8	8	0	168	18	3	0
1	16	3	0	78	15	6	0
175	13	5	0	80	16	5	0
27	9	4	0	22	16	2	0
105	5	6	1	96	12	3	1
131	3	2	0	15	2	7	1
9	13	5	0	12	2	14	1
8	6	3	0	100	14	3	0
4	16	3	0	151	16	2	0
31	16	3	0	125	11	2	0
130	13	5	0	112	16	3	0
140	1	5	0	93	16	3	0
1	9	3	0	52	6	5	1
20	9	6	0	91	12	5	1
73	1	5	1	35	13	3	0
143	3	9	0	61	1	4	0
97	16	3	0	139	10	3	1
136	15	4	0	131	13	5	0
121	3	3	1	177	14	2	0
68	10	5	0	9	17	2	0
139	6	10	1	2	17	2	0
140	15	4	0	72	15	5	0
2	13	3	0	120	8	3	1
51	9	7	0	102	13	3	0
130	1	4	1	114	8	7	1
81	1	4	0	118	16	3	0
118	16	4	0	17	10	4	0
195	17	2	0	159	13	4	0
18	11	4	0	15	16	5	0
158	14	5	0	127	12	4	0
87	16	4	0	206	10	4	0
11	15	3	0	178	15	4	0
157	13	3	1	26	13	7	0
120	13	2	0	42	6	7	1
36	13	4	0				

```
> pairs(kypho.data)
```

The plot of the response versus age, shown in Figure 5.3, indicates that as age increases the

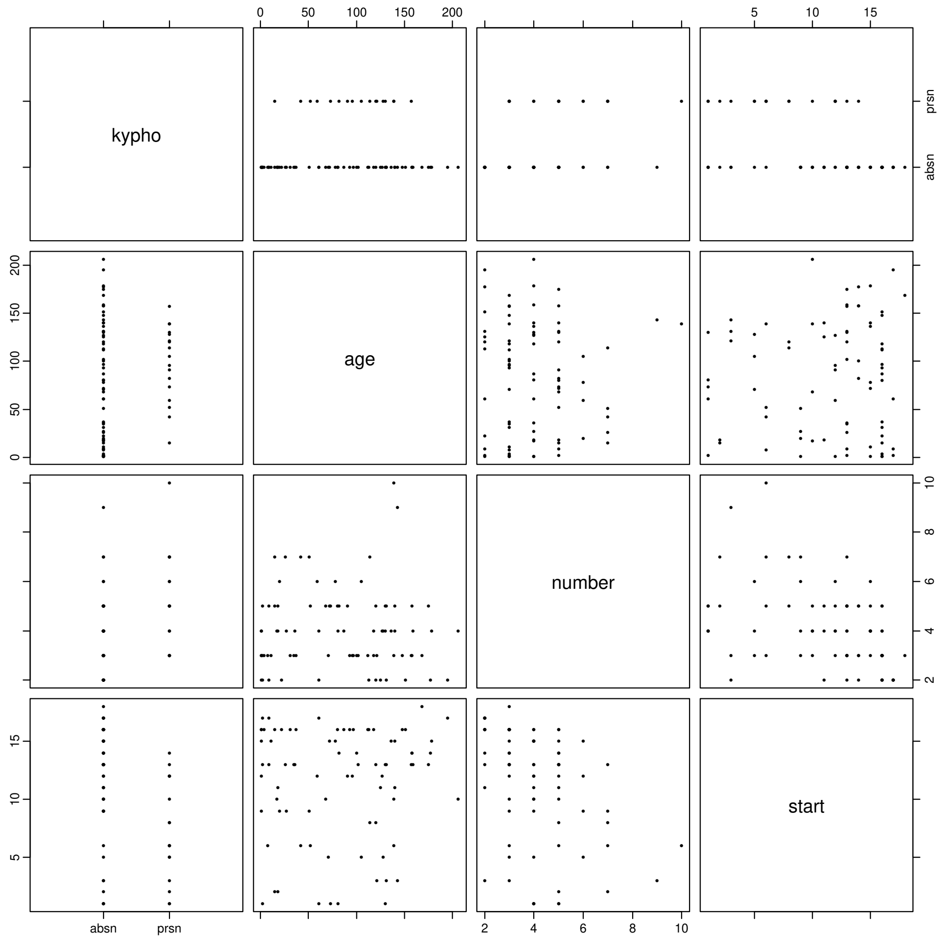

Figure 5.3: Scatterplots for the kyphosis data.

risk first increases and then decreases. This suggests that a quadratic term in age should be used to model this aspect of the data. Accordingly we fit a logistic regression of the form

$$\text{logit}\ \texttt{Kyphosis} = \beta_0 + \beta_1\,\texttt{Age} + \beta_2\,\texttt{Age}^2 + \beta_3\,\texttt{Number} + \beta_4\,\texttt{Start}$$

to the data, using the command:

```
> kypho.glm<-glm(Kyphosis~Age+I(Age^2)+Number+Start, family=binomial,
                                                      data=kypho.data)
> summary(kypho.glm)

Call:
glm(formula = Kyphosis ~ Age + I(Age^2) + Number + Start,
```

```
                                   family = binomial, data = kypho.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.23572  -0.51241  -0.24509  -0.06109   2.35494

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.3834531  2.0478366  -2.141   0.0323 *
Age          0.0816390  0.0343840   2.374   0.0176 *
I(Age^2)    -0.0003965  0.0001897  -2.090   0.0366 *
Number       0.4268603  0.2361167   1.808   0.0706 .
Start       -0.2038411  0.0706232  -2.886   0.0039 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234  on 80  degrees of freedom
Residual deviance: 54.428  on 76  degrees of freedom
AIC: 64.428
Number of Fisher Scoring iterations: 5
```

There is evidence that $\texttt{Age}^2$ is required in the model, since the p-value is small (0.0366). The coefficient of $\texttt{Age}^2$ is negative, indicating that the chance of kyphosis first increases with age and the decreases. The variable `Start` seems important (p-value $= 0.0039$), but the contribution of the variable `Number` is less certain (p-value $= 0.0706$). The coefficient of `Start` is negative which indicates that the higher the value of `Start` the smaller the probability of kyphosis. We should always be cautious when interpreting the p-values produced by `summary`. The Normal approximation used is not very reliable and thus this p-values should be viewed as rough indications of the strength of evidence against the coefficient being 0. In section 5.2.4 we will introduce a more reliable method of judging the significance of regressors.

5.2.3 Analysis of grouped data using logistic regression

Suppose that we have a data set containing n cases and k explanatory variables and that some of the cases have identical sets of values for the explanatory variables. We will say that such cases have the same "covariate vector" - for each case the covariate vector is simply a vector that contains the levels of the explanatory variables for that case. If there is a good deal of duplication of covariate vectors among the cases then it is often more convenient to record and analyse the data by grouping cases that have identical covariate vectors together.

Suppose that there are m distinct covariate vectors in the data set and label these as $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\ldots \boldsymbol{x}_m$. For each $\boldsymbol{x}_i$, let n_i represent the number of cases and suppose that s_i of these that are successes ($Y = 1$) and that $n_i - s_i$ are failures ($Y = 0$). For each $\boldsymbol{x}_i$, we can regard the n_i cases having this covariate vector as a sequence of Bernoulli trials, of which s_i are successes. Under these conditions s_i has a $\mathrm{Bin}(n_i, \pi_i)$ distribution, where $\pi_i = \Pr[Y = 1|\boldsymbol{x}_i]$ i.e. the probability that an individual case having covariate vector $\boldsymbol{x}_i$ will be a "success" and have $Y = 1$. As usual, the logistic model assumes that $\mathrm{logit}(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

For each of m possible covariate vectors we have two observations, the number of cases n_i that have the covariate vector $\boldsymbol{x}_i$, and the number s_i out of n_i that are successes. We fit the model using the *proportion* of successes as the response, and the *number* of trials as a "weight" that is specified as in weighted least squares. We illustrate the procedure using some data from an industrial experiment.

Example 4. The data in Table 5.2 comes from Cox and Snell, "The Analysis of Binary Data", p. 11, and consists of observations on the "unreadiness for rolling" of metal ingots prepared with different soaking times and different heating times. For each combination of heating and soaking times (except one) the total number of ingots examined and the number "not ready for rolling" are given. The first number in each pair is s_i, the second n_i. Thus $0, 10$ signifies 0 not ready out of 10.

Table 5.3: Data for Example 6

	Heating time			
Soaking time	7	14	27	51
1.0	0,10	0,31	1,56	3,13
1.7	0,17	0,43	4,44	0,1
2.2	0,7	2,33	0,21	0,1
2.8	0,12	0,31	1,22	0,0
4.0	0,9	0,19	1,16	0,1

The data are first entered into a data frame `ingots`, with variables `heat`, `soak`, `notready` and `total`:

```
> ingots
  heat soak notready total
1    7  1.0        0    10
2   14  1.0        0    31
3   27  1.0        1    56
4   51  1.0        3    13
5    7  1.7        0    17
6   14  1.7        0    43
7   27  1.7        4    44
8   51  1.7        0     1
9    7  2.2        0     7
10  14  2.2        2    33
11  27  2.2        0    21
12  51  2.2        0     1
13   7  2.8        0    12
14  14  2.8        0    31
15  27  2.8        1    22
16  51  2.8        0     0
17   7  4.0        0     9
18  14  4.0        0    19
19  27  4.0        1    16
20  51  4.0        0     1
```

Note that no observations were made for soaking time 2.8 and heating time 51. We would get the same output from R if this line were deleted. We designate $\pi = \Pr(\texttt{notready} = 1)$ and fit the model

$$\text{logit (notready)} = \beta_0 + \beta_1 \texttt{heat} + \beta_2 \texttt{soak}$$

by typing

```
> ingots.glm<-glm(notready/total~heat+soak,weight=total,
+                              family=binomial,data=ingots.data)
> summary(ingots.glm)

Call:
glm(formula = notready/total ~ heat + soak, family = binomial,
    data = ingots.data, weights = total)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.28311  -0.78184  -0.50514  -0.09702   1.71922

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.55915    1.11826  -4.971 6.65e-07 ***
heat         0.08203    0.02372   3.459 0.000542 ***
soak         0.05677    0.33089   0.172 0.863776
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.395  on 18  degrees of freedom
Residual deviance: 13.753  on 16  degrees of freedom
AIC: 34.08

Number of Fisher Scoring iterations: 4
```

There is no evidence that soaking time affects the probability of being not ready, since the p-value for the hypothesis $\beta_2 = 0$ is 0.8638. However, there is very strong evidence that heating time affects this probability since the p-value for $\beta_1 = 0$ is 0.0005. As the coefficient for `heat` is positive Increasing the heating time increases the probability of being not ready.

5.2.4 Deviances

At the bottom of the output produced when the `summary` command is applied to `glm` objects are two lines that report the "Null Deviance" and the Residual Deviance". A deviance is a measure of how well an estimated model fits the data – the Null Deviance indicates how well a model that just contains an intercept term fits the data and the Residual Deviance indicates how well the specified model fits the data. Note that deviances are always ≥ 0 and that the larger the deviance the worse the fit (a deviance of 0 indicates a perfect fit). For GLM's the deviance fills the role that the residual sums of squares plays for ordinary regression models. The way that the deviance of a model is calculated will be illustrated using the following example.

Example 5. The larvae of the tobacco budworm, *Heliothis virescens*, are responsible for much damage to cotton crops in the United States, and Central and Southern America. As a result of intensive cropping practices and the misuse of pesticides, particularly synthetic pyrethroids, the insect has become an important crop pest. Many studies on the resistance of the larvae to pyrethroids have been conducted, but the object of the experiment by Holloway (1989) was to examine levels of resistance in the adult moth to the pyrethroid trans-cypermethrin.

In the experiment batches of pyrethroid resistant moths of each sex were exposed to a range of doses of cypermethrin two days after emergence from pupation. The number of moths which were either knocked down (movement of the moth uncoordinated) or dead (the moth is unable to move and does not respond when poked with a blunt instrument) were recorded 72 hours after treatment. Reference: Holloway, J.W. (1989), A comparison of the toxicity of the pyrethroid trans-cypermethrin, with and without the synergist piperonyl butoxide, to adult moths from two strains of *Heliothis virescens.* University of Reading, Ph.D. thesis, Department of Pure and Applied Zoology.

The problem is to assess the effect of increasing dose of cypermethrin on toxicity. The data are shown in Table 5.4.

Table 5.4: Data from the tobacco budworm toxicity experiment.

Sex of moth	Dose (mg) of cypermethrin	Number affected out of 20
Male	1.0	1
	2.0	4
	4.0	9
	8.0	13
	16.0	18
	32.0	20
Female	1.0	0
	2.0	2
	4.0	6
	8.0	10
	16.0	12
	32.0	16

Clearly we are dealing with grouped data. There are $m = 12$ distinct covariate vectors: (Male, 1.0), (Male, 2.0), (Male, 4.0), (Male, 8.0), (Male, 16.0), (Male, 32.0), (Female, 1.0), (Female, 2.0), (Female, 4.0), (Female, 8.0), (Female, 16.0) and (Female, 32.0). We will label these covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_{12}$. There are 20 cases (moths) observed for each covariate pattern, ($n_i = 20$ for $i = 1, 2, \ldots, 12$) and the total number of cases is $n = 20 \times 12 = 240$.

Now assume that the responses are independent, and that the probability that a moth will be "knocked down" ($Y = 1$) depends only on the covariates. Then s_i is an observation from a binomial distribution, $\mathrm{Bin}(n_i, \pi_i)$, where π_i is the probability of $Y = 1$ for a case having covariate vector $\boldsymbol{x}_i$. Thus the probability of observing $Y = 1$ s_i times out of n_i cases is

$$\binom{n_i}{s_i} \pi_i^{s_i} (1 - \pi_i)^{n_i - s_i}.$$

The likelihood function is produced by multiplying the probabilities for the m covariate patterns together:

$$L = \prod_{i=1}^{m} \binom{n_i}{s_i} \pi_i^{s_i} (1 - \pi_i)^{n_i - s_i}. \tag{5.3}$$

Before we can define the deviance for a model, we must first introduce the concept of a *maximal model.* The maximal model is defined as the model that gives the best possible fit to the data. In other words, it is the model that gives the highest possible value of the likelihood function. For binary response data, specifying the maximal model is equivalent to specifying the set of π_i's that will maximize equation 5.3 when no restrictions are placed on the π_i's. It is not difficult to show (using calculus) that this is accomplished by setting $\hat{\pi}_i = s_i/n_i$. Plugging these values into equation 5.3 gives the highest possible value for the likelihood function which we will denote as $\mathrm{L}_{\mathrm{max}}$. This value serves as a benchmark to which we will compare other models.

The maximal model represents the most complicated model we can fit for the given set of covariate vectors. Usually, it is desirable to find a simpler model that can describe the relationship between the $\boldsymbol{x}_i$'s and the π_i's. The simplest model that we might consider is the *null model* which has $\hat{\pi}_i =$ constant for all i. Note that this model severely restricts our choices for the $\hat{\pi}_i$'s (they must all be the same). Under this restriction the maximum value of the likelihood function is obtained by setting $\hat{\pi}_i = s/n$ for all i where $s = \sum s_i$. We will call this value $\mathrm{L}_{\mathrm{null}}$. The deviance for the null model is calculated as

$$\text{null model deviance} = 2\log \mathrm{L}_{\mathrm{max}} - 2\log \mathrm{L}_{\mathrm{null}}.$$

The model that we are really interested in is the logistic model. For this model the $\hat{\pi}_i$'s must satisfy:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \mathtt{sex}_i + \hat{\beta}_2 \mathtt{dose}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \mathtt{sex}_i + \hat{\beta}_2 \mathtt{dose}_i)}.$$

The fitted coefficients (the $\hat{\beta}$'s) for this model are selected so that the resulting $\hat{\pi}_i$'s make the value of L as large as possible (under this model). Let $\mathrm{L}_{\mathrm{mod}}$ represent this value. Then the deviance for the logistic model is

$$\text{logistic model deviance} = 2\log \mathrm{L}_{\mathrm{max}} - 2\log \mathrm{L}_{\mathrm{mod}}.$$

Table 5.5: Data from the tobacco budworm toxicity experiment.

						$\hat{\pi}_i$'s	
i	sex	dose	s_i	n_i	maximal	logistic	null
1	0	1	1	20	0.05	0.27	0.46
2	0	2	4	20	0.20	0.30	0.46
3	0	4	9	20	0.45	0.37	0.46
4	0	8	13	20	0.65	0.53	0.46
5	0	16	18	20	0.90	0.80	0.46
6	0	32	20	20	1.00	0.98	0.46
7	1	1	0	20	0.00	0.12	0.46
8	1	2	2	20	0.10	0.14	0.46
9	1	4	6	20	0.30	0.18	0.46
10	1	8	10	20	0.50	0.30	0.46
11	1	16	12	20	0.60	0.60	0.46
12	1	32	16	20	0.80	0.95	0.46

The fitted probabilities for each covariate pattern in the budworm data set are given in Table 5.5 for the maximal, the logistic, and the null models. Plugging these values into (5.3) gives

$$\begin{aligned} 2\log \mathrm{L}_{\mathrm{max}} &= -30.110, \\ 2\log \mathrm{L}_{\mathrm{mod}} &= -58.078, \\ 2\log \mathrm{L}_{\mathrm{null}} &= -154.986. \end{aligned}$$

Now the deviances for the logistic model and the null model can be calculated:

$$\begin{aligned}\text{logistic model deviance} &= -30.1104-(-58.0784)=27.968,\\ \text{null model deviance} &= -30.1104-(-154.986)=124.876.\end{aligned}$$

Of course, in practice, we never go to the bother of doing all these calculations ourselves since they are easily obtained using the `summary` function in R.

```
>  bugs.glm<-glm(s/n~sex+dose,family=binomial,weight=n,data=budworm.df)
> summary(bugs.glm)

Call:
glm(formula = s/n ~ sex + dose, family = binomial, data = budworm.df,
    weights = n)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5567  -1.3326   0.3384   1.1254   1.8838

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1661     0.2615  -4.459 8.24e-06 ***
sex          -0.9686     0.3295  -2.939  0.00329 **
dose          0.1600     0.0234   6.835 8.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:  27.968  on  9  degrees of freedom
AIC: 64.078
```

Clearly the deviance for the logistic model (Residual deviance) is much smaller than that for the null model (Null deviance). This indicates that the logistic model gives a much better fit to the data than does the null model. But how do we tell when the deviance for the logistic model is small enough to indicate that it provides a reasonable fit to the data? Suppose the logistic model is the true model. Then, provided m is not too big and the n_i's are large, the asymptotic distribution of the deviance is χ^2_{m-k-1} – although this approximation is often not very accurate. We can perform a test of the hypothesis "the logistic model is reasonable" by comparing the model deviance to a χ^2_{m-k-1} distribution. Since large values of the deviance are evidence *against* the logistic model being correct, a suitable p-value is calculated by finding the area under the χ^2_{m-k-1} curve to the *right* of the value of the deviance. Given the questionable nature of the approximation, we should treat the p-value as a rough indication of how well the the model fits the data.

For our example the residual deviance is 27.968 and $m-k-1=12-2-1=9$ (note this is the degrees of freedom for the residual deviance given on the output). In R the p-value can be found using

```
> 1-pchisq(27.968,9)
[1] 0.0009656815
```

This test gives very strong evidence against the hypothesis that the fitted logistic model is adequate and thus we should explore alternative models. It is known from previous experience that a logistic model using log dose rather than dose often fits mortality data from pesticide experiments well. So we will try fitting the model

$$\text{logit } \pi = \beta_0 + \beta_1 \, \texttt{sex} + \beta_2 \log(\texttt{dose}).$$

The output for this model is:

```
> logbugs.glm<-glm(s/n ~ sex + log(dose), family = binomial,
                                                  weight=n, data=budworm.df)
> summary(logbugs.glm)

Call:
glm(formula = s/n ~ sex + log(dose), family = binomial, data = budworm.df,
    weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.10540  -0.65343  -0.02225   0.48471   1.42945

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3724     0.3854  -6.156 7.46e-10 ***
sex          -1.1007     0.3557  -3.094  0.00197 **
log(dose)     1.5353     0.1890   8.123 4.54e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:   6.757  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 3
```

The residual deviance is now much smaller (6.757) indicating a better fit. If we test the hypothesis that this model is adequate, we get a large p-value :

```
>  1-pchisq(6.757,9)
[1] 0.6624024
```

Thus there is no evidence against the hypothesis and we conclude that the model using log(dose) is adequate. Note that there may be other models that are adequate and that may provide an even better fit to the data.

We can assess the fit graphically by plotting the logits of the observed proportions against the log dose, with the fitted lines for males and females added. Since $\texttt{sex} = 0$ for males and $\texttt{sex} = 1$ for females, the fitted lines are:

$$\begin{aligned} \text{males: logit}(y) &= -2.37 + 1.54 \log(\texttt{dose}) \\ \text{females: logit}(y) &= -3.47 + 1.54 \log(\texttt{dose}) \end{aligned}$$

To draw the plot, using M's for males and F's for females, and draw in the fitted lines in R:

```
> plot(log(dose), log((s+0.5)/(n-s+0.5)),type="n")
> text( log(dose), log((s+0.5)/(n-s+0.5)), ifelse(sex==0,"M","F"))
> abline(-2.372412,1.535336,lty=1)
> abline(-2.372412-1.100743,1.535336,lty=2)
> legend(0,3,c("M","F"),lty=c(1,2))
```

Note that when calculating the logits of the observed proportions, we have added a "fudge factor" of 0.5 to avoid trying to take the log of zero. The result, shown in Figure 5.4, indicates that the model fits quite well.

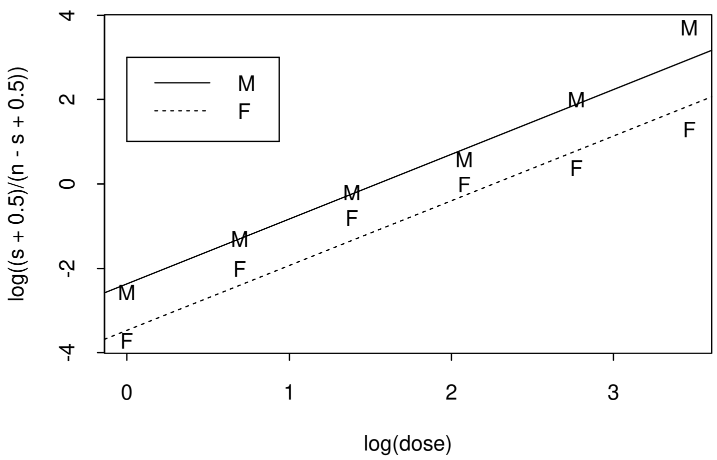

Figure 5.4: Fitted lines for the budworm data.

Example 6. For the ingot example (pp. 13-15), we can test the adequacy of the logistic model by following the procedure above. The residual deviance for the fitted model was 13.753 on 16 degrees of freedom. Thus our test statistic is $\chi_o^2 = 13.753$ which we compare to a χ_{16}^2 distribution:

```
> 1-pchisq(13.753,16)
[1] 0.617109
```

The p-value 0.6171 is quite large, so there is no evidence that the fitted logistic model is inadequate.

The "sparse" case

What happens when $n_i = 1$ for most or all of the distinct covariate vectors? Then the conditions for the asymptotics do not hold, and so the assumption that the deviance has approximately a χ^2 distribution is not valid. However, the deviance is still useful as it allows us to compare models as we will see below.

Comparing models

Suppose that we have a logistic model, and we want to see if we can drop a subset of variables from the model. In other words, we want to see if a submodel of the original logistic model is adequate. Let $\mathrm{L_{sub}}$ and $\mathrm{L_{full}}$ represent the maximum values of the likelihood function under the submodel and under the full model respectively. Standard likelihood theory indicates that if the submodel is adequate, the difference $2\log \mathrm{L_{full}} - 2\log \mathrm{L_{sub}}$ will have a distribution that is approximately χ^2_d where d is the number of variables dropped. This difference can be expressed as a difference of deviances:

$$\begin{aligned} 2\log \mathrm{L_{full}} - 2\log \mathrm{L_{sub}} &= (2\log \mathrm{L_{max}} - 2\log \mathrm{L_{sub}}) - (2\log \mathrm{L_{max}} - 2\log \mathrm{L_{full}}) \\ &= \text{deviance of submodel} \; - \; \text{deviance of full model}. \end{aligned}$$

This difference represents the increase in the deviance when we drop the d terms from the model. While this difference will always be positive, if the increase is small then dropping the extra terms will not increase the deviance by very much and so the variables can be dropped in the interests of getting a simpler model. The difference in the deviance has approximately a χ^2_d distribution if the dropped variables are not needed in the model. Thus we can calculate a p-value to test the hypothesis that the dropped variables are not needed by comparing this difference in deviances to a χ^2_d distribution. A small p-value provides evidence *against* the submodel - i.e. a small p-value indicates that we should not drop all d of the variables from the model.

Unlike the approximation to the deviance itself, the χ^2 approximation to this *difference* between deviances is usually quite accurate. In particular it will be good provided that m (the number of distinct covariate vectors) is large. Note that we don't need the n_i's to be large.

Example 7. For the kyphosis example (pp. 10-13), suppose we want to test whether we should drop both of the variables `start` and `number`. First we fit the submodel which only uses `age` and `age`^{2} as regressors:

```
> sub.glm<-glm(Kyphosis~ Age+I(Age^2),family=binomial,data=kypho.data)
```

We can use the `deviance` command in R to extract the residual deviance from our `glm` object:

```
> deviance(sub.glm)
[1] 72.73858
```

Thus the submodel deviance is 72.73858, so the difference between the submodel and full model deviances is $72.739 - 54.428 = 18.311$. To get the p-value we use a χ^2 distribution with $d = 2$ degrees of freedom:

```
> 1-pchisq(18.311,2)
[1] 0.0001056372
```

Thus the test is highly significant which indicates that variables `Start` and `Number` should **not** be dropped from the model.

The `anova` command in R can be used to perform the above test for us as follows:

```
> anova(sub.glm,kypho.glm,test="Chisq")
Analysis of Deviance Table
```

```
Model 1: Kyphosis ~ Age + I(Age^2)
Model 2: Kyphosis ~ Age + I(Age^2) + Number + Start
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        78     72.739
2        76     54.428  2   18.311 0.0001056
```

Note that we simply enter the objects containing the submodel and the full model and use the `test="Chisq"` option to indicate we want to use a χ^2 reference distribution.

Dropping regressors sequentially

The above procedure can be used to investigate dropping variables one at a time if the submodel is created by dropping one variable from the full model. This approach provides a more reliable method of assessing the importance of regressors than the p-values provided by the `summary` command. The `anova` function provides a convenient way to examine the changes in the deviance as terms are added to the model one at a time. For the kyphosis example, the following table can be produced using `anova`:

```
> anova(kypho.glm,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: Kyphosis
Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        80     83.234
Age       1    1.302        79     81.932     0.254
I(Age^2)  1    9.194        78     72.739     0.002
Number    1    8.876        77     63.863     0.003
Start     1    9.435        76     54.428     0.002
```

Each line in this table represents an application of the submodel testing procedure and considers adding the variable listed for that line to the model that contains all variables listed above that line in the table. Thus this table summarises the following series of tests:

line	submodel	full model
Age	null	Age
I(Age^2)	Age	Age+I(Age^2)
Number	Age+I(Age^2)	Age+I(Age^2)+Number
Start	Age+I(Age^2)+Number	Age+I(Age^2)+Number+Start

In interpreting this table we should start at the bottom and work our way up. The line for `Start` indicates there is strong evidence that `Start` should be retained in the model. Given this result the line for `Number` isn't of much relevance since it considers removing `Number` from the model that does not contain `Start` (this line would be relevant had we decided to remove `Start`).

The order that terms are added in the output from `anova` is determined by the order used when specifying the model in `glm`. Thus we can investigate adding the terms in a different order as follows:

```
> kypho.glm2<-glm(Kyphosis~ Age+I(Age^2)+Start+Number,
                            family=binomial,data=kypho.data)
> anova(kypho.glm2,test="Chisq")
```

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: Kyphosis
Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                        80     83.234
Age       1    1.302        79     81.932     0.254
I(Age^2)  1    9.194        78     72.739     0.002
Start     1   14.324        77     58.414 0.0001539
Number    1    3.986        76     54.428     0.046
```

The last line for this table tests whether `Number` should be dropped from the model that contains all the variables and gives a p-value of 0.046. Compare this to the line for `Number` from the output from `Summary` on page 13 – this line is testing the same hypothesis but gives a p-value of 0.0706. The reason for this discrepancy is that the two tests are based on different approximations: the test used in `summary` assumes that the estimated coefficient has a Normal distribution whereas the test used in `anova` assumes that the change in deviance has a χ^2 distribution. Usually the two tests will agree fairly well but sometimes there is a discrepancy. In such cases, the χ^2 test is more reliable.

Testing the significance of the regression

A special case of testing a submodel is the problem of testing the overall significance of the regression – are any of the explanatory variables useful in explaining the response? For this test, the null model is used as the submodel and the fitted model is used as the full model. Thus the test statistic is the difference in deviance between the null model and the fitted model. The reference distribution is χ^2_k where k is the number of regressors in the fitted model.

Example 8. For the kyphosis example, consider the model that contains `Age`, `Age`^{2}, `Start`, and `Number` as regressors (see page 13) and suppose we want to test the hypothesis that the coefficients for all four regressors are zero. In this case, the deviance of the fitted model is 54.428, and that of the null model is 83.234. The difference is $83.234 - 54.428 = 28.806$. This value is compared to a χ^2 distribution with 4 degrees of freedom – note that this value is the difference between the degrees of freedom for the null deviance and the residual deviance.

```
> 1-pchisq(28.806,4)
[1] 8.559766e-06
```

The p-value is very small indicating strong evidence against the hypothesis that none of the regressors are needed in the model.

5.2.5 Diagnostics for logistic regression models

For logistic regression models we require diagnostic procedures to check the validity of our fitted model. These procedures are often analogous to the procedures we used for ordinary regression models. We start by introducing two types of residuals which are required for a number of our diagnostics.

1. Pearson residuals. These are defined by

$$r_i = \frac{s_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

   The Pearson r_i represents the difference between the observed number of successes and the predicted number of successes for the ith covariate pattern divided by the standard error of the predicted number of successes. Thus Pearson residuals are similar to standardised residuals for the ordinary linear regression model.

2. Deviance residuals. Let $\hat{\pi}_i$ and $\tilde{\pi}_i$ be the estimated probabilities of success for the ith covariate pattern using the logistic model and the maximal model respectively. Then it can be shown that the deviance of the logistic model can be written deviance $= \sum_{i=1}^{m} d_i^2$ where

$$d_i = \pm\left\{-2\left(y_i \log\left(\frac{\hat{\pi}_i}{\tilde{\pi}_i}\right) + (n_i - s_i)\log\left(\frac{1-\hat{\pi}_i}{1-\tilde{\pi}_i}\right)\right)\right\}^{\frac{1}{2}}.$$

   The sign of d_i is selected to be the same as the sign of r_i. The d_i's are called *deviance residuals.*

Note that the Pearson residuals are only useful for grouped data, while the deviance residuals are useful for ungrouped as well as grouped data.

Outlier detection

Large values of $|r_i|$ or $|d_i|$ denote covariate patterns that are poorly fitted by the model. More precisely a large $|r_i|$ indicates a covariate patterns where there is a large discrepancy between the observed and the predicted number of successes and a large value of $|d_i|$ indicates an observation that makes an usually large contribution to the residual deviance of the model.

Both types of residual are calculated as part of the "glm object", and are extracted with the `residuals` function. For the ingot data model (pp. 14-15), we can extract the residuals as follows:

```
> d.resids<-residuals(ingots.glm,type="deviance")
> p.resids<-residuals(ingots.glm,type="pearson")
```

Simple index plots of the residuals are useful for outlier detection. We can use the following commands in R to get an index plots of each type of residual.

```
> plot(p.resids,type="n",main="Pearson residuals")
> text(p.resids)
> abline(h=0,lty=3)
> plot(d.resids,type="n",main="Deviance residuals")
> text(d.resids)
> abline(h=0,lty=3)
```

These plots indicate that the 7th and 10th covariate patterns have large Pearson residuals and large deviance residuals. We can identify the cases by printing out the corresponding rows of the data frame:

```
ingots[c(7,10),]
   heat soak notready total
```

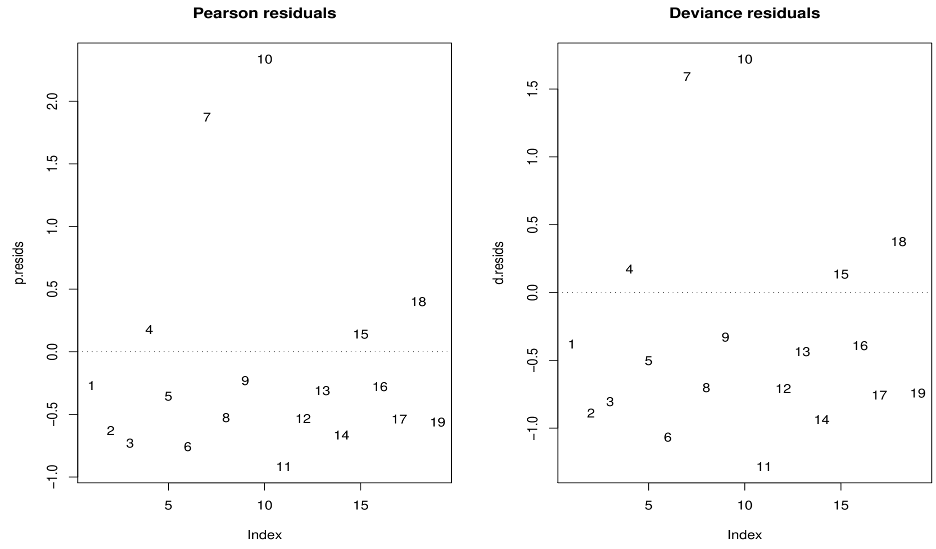

Figure 5.5: Plots of residuals for the ingot data.

7	27	1.7	4	44
10	14	2.2	2	33

In both cases $r_i > 0$ which indicates the observed number of successes are larger than what could reasonably be expected under the fitted model.

Non-linear regression surfaces

Recall that the logistic regression model assumes that logit(Y) is a linear function of the regressors:

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

We can check this assumption by (i) plotting the Pearson residuals versus the linear predictors $(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p)$ and (ii) plotting the Pearson residuals versus each of the numerical explanatory variable. In each case, we want to observe a patternless horizontal band of points. Clear evidence of curvature indicates that the linearity assumption is not valid.

To illustrate these plots consider the first model we tried for the budworm data (page 18): logit$(Y) = \beta_0 + \beta_1$ `sex` $+ \beta_2$ `dose`. Recall that we found evidence that this model was not adequate. Lets see if we can find evidence of non-linearity using the above plots. The following commands can be used to produce these plots in R:

```
> l.pred<-bugs.glm$linear.predictors
> p.res<-residuals(bugs.glm,type="pearson")
> plot(l.pred,p.res,xlab="linear predictors", ylab="Pearson residuals",
+      main="Pearson residuals vs linear predictors")
> lines(lowess(l.pred,p.res),lty=3)
> plot(budworm.df$dose,p.res,xlab="dose",ylab="Pearson residuals",
+      main="Pearson residuals vs dose")
```

```
> lines(lowess(budworm.data$dose,p.res),lty=3)
```

The plots produced in this manner are given in Figure 5.6. These plots both give clear evidence that the linearity assumption is not valid for the fitted model.

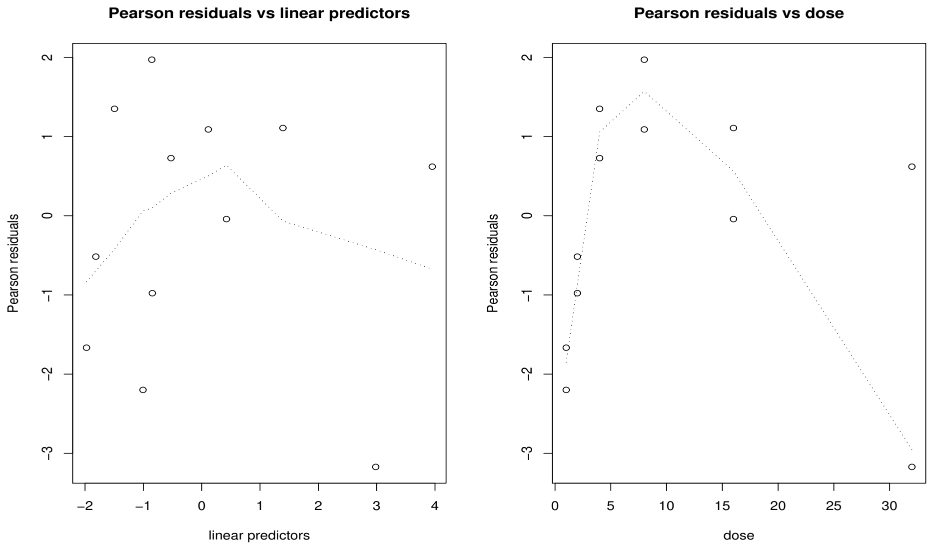

Figure 5.6: Plots of Pearson residuals for the budworm data.

Detecting high leverage and influential points

The diagnostics used to detect high leverage and influential points for logistic regression models are, for the most part, similar to those used for ordinary regression models. To illustrate these diagnostics, we will consider the fitted model we used for the budworm data from page 19 – note that this is the model that used log (dose) as a predictor. To evaluate the leverage of points we use the diagonal elements of the "hat" matrix as was done for ordinary regression models. These can be obtained as follows:

```
> hats<-lm.influence(logbugs.glm)$hat
> hats
 [1] 0.2320029 0.2906448 0.2919340 0.2807781 0.2459772 0.1733597 0.1282480
 [8] 0.1994008 0.2564510 0.2896237 0.3199622 0.2916174
```

To evaluate the overall influence of observations on the fitted model we will consider two diagnostics: (i) Cook's distance and (ii) deviance changes. The definition of Cook's distance for logistic regression models is analogous to the definition used for ordinary regression models and it is interpreted in the same manner – an unusually large value indicates an influential point. To calculate the exact values of Cook's distance for logistic regression models is computationally expensive and thus the following approximation is usually used:

$$C_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1-h_{ii})^2}$$

where r_i is the ith Pearson residual, p is the number of regressors and h_{ii} is the leverage for the ith point. For our current example the Cook's distances can be calculated as follows:

```
> p.res<-residuals(logbugs.glm,type="pearson")
> cds<-(p.res^2*hats)/(2*(1-hats)^2)
> cds
          1           2           3           4           5           6
0.062822072 0.005639510 0.002704108 0.050153073 0.038405042 0.132962650
          7           8           9          10          11          12
0.052347157 0.012597009 0.246038542 0.113644900 0.240402695 0.201680823
```

The second diagnostic we used to evaluate the impact of observations on the response is the "deviance changes". This is a "leave-one-out" diagnostic that indicates how much the residual deviance would change if that observation was deleted. Unusually large values of deviance changes indicate an influential point. The deviance changes are computationally expensive to calculate exactly and so are usually approximated by:

$$\Delta D_i = d_i^2 + \frac{r_i^2 \; h_{ii}}{1 - h_{ii}}$$

where d_i is the ith deviance residual. In R they can be calculated as follows:

```
> d.res<-residuals(logbugs.glm,type="deviance")
> dcs<-d.res^2 + p.res^2 *(hats/(1-hats))
> dcs
         1          2          3          4          5          6          7
0.46710821 0.02782310 0.01310170 0.25239632 0.24935365 2.26313779 1.31317447
         8          9         10         11         12
0.09643678 1.33189072 0.55361642 0.99300264 0.90571315
```

Index plots of leverage, Cook's distance and deviance changes for the budworm model are given in Figure 5.7. In these plots we are simply looking for observations that stand out as being substantially larger than the others. For the current example, none of the observations have large enough values of any of these diagnostics to be of real concern.

Changes in the coefficients are measured by quantities similar to those in ordinary linear regression, and in fact the function `influence.measures` described in Chapter 2 for the calculation of influence statistics computes these when given a "glm object" as its argument.

Example 9. The data in Table 5.3 were obtained in a study of the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin (Finney, Biometrika, 1947, p. 320). The nature of the measurement process was such that only the occurrence or non-occurrence of vaso-constriction could be reliably measured. Three subjects were involved in the study: the first contributed 9 responses, the second contributed 8 responses, and the third contributed 22 responses. It was decided to use log(Volume) and log(Rate) as the regressors for the logistic model.

```
>  vaso.glm<-glm(Response~log(Volume)+log(Rate),
                          data=vaso,family="binomial")
> summary(vaso.glm)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.875      1.306  -2.200  0.02777 *
log(Volume)    5.179      1.843   2.810  0.00496 **
log(Rate)      4.561      1.818   2.509  0.01211 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

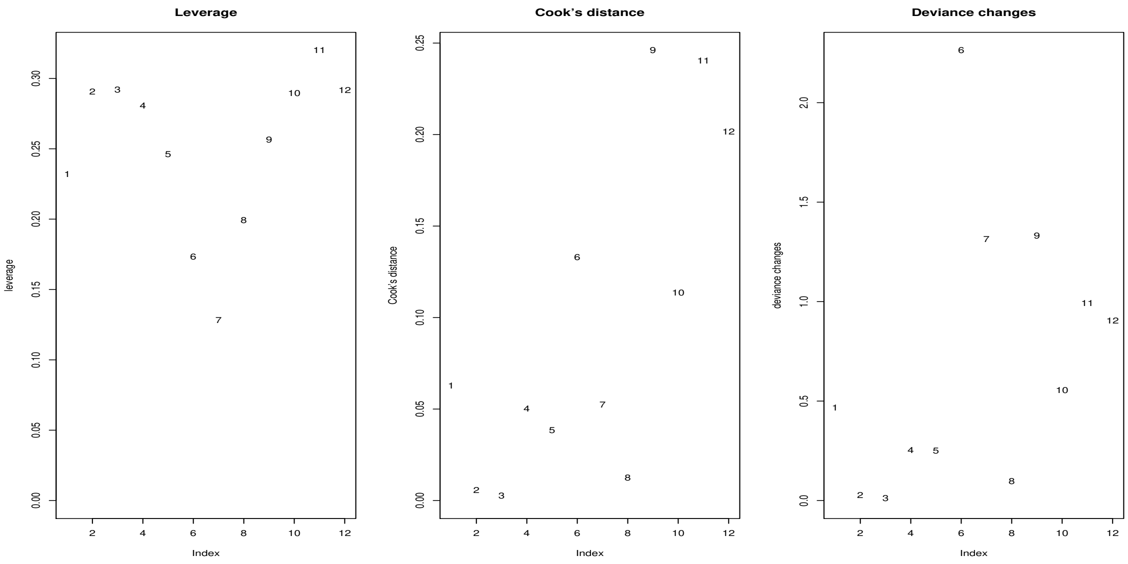

Figure 5.7: Leverage, Cook's distance and deviance changes for the budworm model.

```
    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.227  on 36  degrees of freedom
```

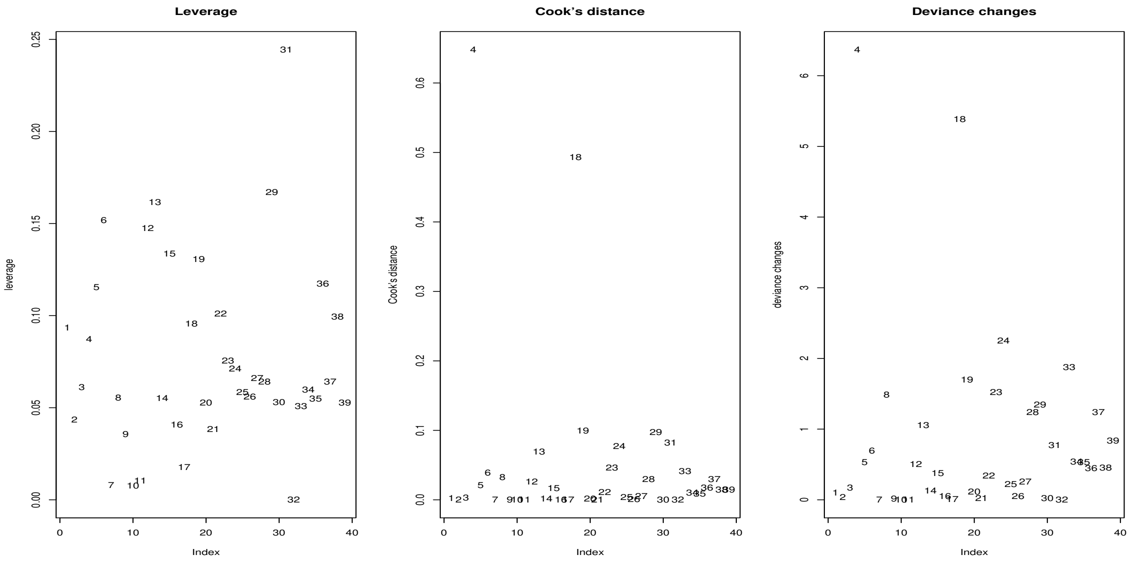

Figure 5.8: Influence plots for the vaso-constriction data.

Index plots of leverage, Cook's distance, and deviance changes are shown in Figure 5.8. These plots show that two observations, the 4th and 18th, have a considerable influence on the fit. The effect of deleting these points can be determined by refitting:

```
> new.glm<-glm(Response~log(Volume)+log(Rate),
                       data=vaso[-c(4,18),],family="binomial")
> summary(new.glm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -24.58      13.80  -1.781   0.0750 .
log(Volume)    39.54      22.89   1.728   0.0840 .
log(Rate)      31.93      17.49   1.826   0.0678 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 51.266  on 36  degrees of freedom
Residual deviance:  7.361  on 34  degrees of freedom
AIC: 13.361
```

Note the large changes in the coefficients. Although the p-values generated by `summary` are borderline for both variables, the χ^2 tests indicate both variables are required:

```
> anova(new.glm,test="Chisq")
Analysis of Deviance Table
            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                           36     51.266
log(Volume)  1    8.764        35     42.502     0.003
log(Rate)    1   35.141        34      7.361 3.067e-09

>  new2.glm<-glm(Response~log(Rate)+log(Volume),
                data=vaso[-c(4,18),],family="binomial")
> anova(new2.glm,test="Chisq")
Analysis of Deviance Table
            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                           36     51.266
log(Rate)    1    5.204        35     46.062     0.023
log(Volume)  1   38.701        34      7.361 4.939e-10
```

5.2.6 Binary anova

The logistic regression method can be easily extended to situations where all or some l of the explanatory variables are factors. The terms “binary ANOVA” (all regressors are factors) and “binary ANCOVA” (some of the regressors are factors) are sometimes used to denote these situations. Note that we have already considered a case of binary ANCOVA (example 5). We now consider two examples of binary ANOVA.

Example 10. A study was conducted on the reproduction of plum trees by taking cuttings from older trees. Half the cuttings were planted immediately while the other half were bedded in sand until spring when they were planted. Two lengths of cuttings were used: long (12 cm) and short (6cm). A total of 240 cuttings were taken for each of the 4 combinations of planting time and cutting length and the number of cuttings that survived in each situation was recorded:

Length of cutting	Time of planting	Number surviving (out of 240)
short	at once	107
	in spring	31
long	at once	156
	in spring	84

This is a simple example of a two-way binary ANOVA. There are two explanatory variables each of which is a factor. Possible models for this data are:

1. The null model
2. Planting time used as a factor
3. Cutting length used as a factor
4. Both planting time and cutting length used as factors (no interaction)
5. Both planting time and cutting length used as factors and the interaction is included

We want to find the simplest model that can adequately explain the data. To do this we start with the most complicated model (5) and investigate eliminating terms sequentially.

```
> len<-c("short","short","long","long")
> time<-c("at once","spring","at once","spring")
> survive<-c(107,31,156,84)
> time<-C(factor(time),treatment)
> len<-C(factor(len),treatment)
> plum.glm<-glm(survive/240~time*len,
                   family=binomial,weights=rep(240,4))
> anova(plum.glm,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: survive/240
Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         3    151.019
time      1   97.579         2     53.440 5.175e-23
len       1   51.147         1      2.294 8.572e-13
time:len  1    2.294         0  4.655e-14     0.130
```

This table suggests that the interaction term is not needed in the model. Given that we drop the `time:len` interaction then we can use this table to consider dropping `len` as well. The small p-value for the `len` line indicates very strong evidence that `len` should not be dropped from the model. Note that since `len` is retained the line for `time` from this table is not relevant. To test whether `time` should be kept in the model we need to consider the table that adds the variables in the other order:

```
> plum2.glm<-glm(survive/240~len+time,family=binomial,weights=rep(240,4))
> anova(plum2.glm,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: survive/240
Terms added sequentially (first to last)
```

```
     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    3    151.019
len   1   45.837        2    105.182 1.285e-11
time  1  102.889        1      2.294 3.545e-24
```

Thus we conclude that both `len` and `time` are needed in the model but not the `time:len` interaction. We can get the fitted probabilities for the four combinations of `len` and `time` by using the `predict` command:

```
> predict(plum2.glm,type="response")
        1         2         3         4
0.4245994 0.1504006 0.6712339 0.3287661
```

Thus our model predicts that the following probabilities of survival:

length	time	$\hat{\pi}$
short	at once	0.42
	in spring	0.15
long	at once	0.67
	in spring	0.33

We conclude that survival probabilities are higher for: (i) cuttings planted at once compared to those planted in the spring and (ii) long cuttings compared to short cuttings.

Example 11. Table 5.6 contains the results of a series of experiments to compare the effects of x-rays and beta-rays on the mitotic rates in grasshopper neuroblasts. For each experiment, embryos from the same egg were divided into three groups, one serving as a control, and the other two being exposed to physically equivalent doses of x-rays and beta-rays. After irradiation, approximately equal numbers of cells in each of the 12 experiment $\times$ treatment combinations were examined and the number of cells passing through mid-mitosis was noted. We thus have a binary ANOVA, with the cells being the experimental units (cases), the binary response being 1 if the cell has passed mid-mitosis and zero otherwise. There are two factors, the experiment with 4 levels (1 through 4) and the treatment with three levels (Control, X-ray, Beta-ray).

For our analysis we will read the data into R, define the factors, create a data frame, and fit models using `glm`. We type

```
> # Enter data
> yes<-c(12,3,4,14,5,6,9,5,2,17,5,7)
> no<-c(10,15,12,3,10,9,11,17,17,2,14,13)
> total<-yes+no

> # Create factors
> experiment<-factor(rep(1:4,c(3,3,3,3)))
> treat<-rep(c("Control","X-ray","Beta-ray"),4)
> # We want the levels in the order they occur
> treat<-factor(treat,levels=unique(treat))

> # Make a data frame
> grass<-data.frame(experiment,treat,yes,total)
> grass
   experiment    treat yes total
1           1  Control  12    22
```

Table 5.6: Grasshopper data.

Experiment	Mid Mitosis	Control	X-ray	Beta-ray
1	Yes	12	3	4
	No	10	15	12
2	Yes	14	5	6
	No	3	10	9
3	Yes	9	5	2
	No	11	17	17
4	Yes	17	5	7
	No	2	14	13

```
2           1    X-ray   3    18
3           1 Beta-ray   4    16
4           2  Control  14    17
5           2    X-ray   5    15
6           2 Beta-ray   6    15
7           3  Control   9    20
8           3    X-ray   5    22
9           3 Beta-ray   2    19
10          4  Control  17    19
11          4    X-ray   5    19
12          4 Beta-ray   7    20

# Fit the model using both factors plus their interaction
> grass.glm<-glm(yes/total~experiment*treat,weight=total,
                               family="binomial", data=grass)
> anova(grass.glm,test="Chisq")
Analysis of Deviance Table
Model: binomial, link: logit
Response: yes/total
Terms added sequentially (first to last)

                 Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                11     54.838
experiment        3   11.662         8     43.175     0.009
treat             2   38.212         6      4.963 5.039e-09
experiment:treat  6    4.963         0  4.276e-16     0.549
```

There is no evidence that the interaction is needed in the model. So we drop the interaction and consider the line for `treat`. This test gives extremely strong evidence that `treat` is needed in the model. For this example, the `experiment` factor should be considered as a blocking variable. That is we are not really interested in the effect of `experiment` but we want to take this effect into account when we compare levels of `treat`.

As with ordinary ANOVA, an interaction plot is a useful graphic display to complement the test for no interaction. The only difference is that we want to plot the logit values for each cell rather than the cell means. The *R* function `interaction.plot` can be used, but we need to supply an extra argument to plot the logits of the cell proportions rather than the proportions. (For binary data, proportions and means are the same.)

```
> attach(grass)
```

```
> logit<-function(x){log(x/(1-x))}
> interaction.plot(experiment,treatment,yes/total,fun=logit)
```

The plot is shown in Figure 5.9. As in ordinary interaction plots, parallel lines (approxi-

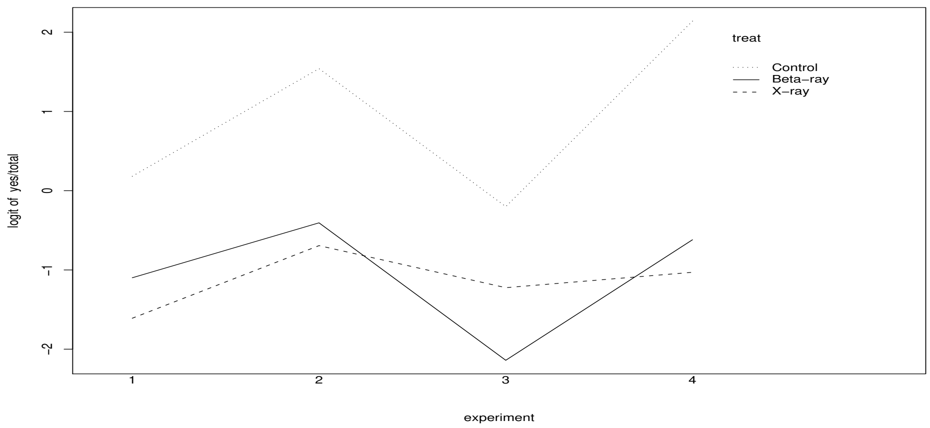

Figure 5.9: Interaction plot for the grasshopper data.

mately) indicate no interaction. Apart from a rather high value for the logit of the X-ray proportion in Experiment 3, the profiles are quite parallel, indicating the absence of interaction. Thus the effect of changing from one treatment to another is the same for all four experiments, although there are significant differences between experiments.

The output from `dummy.coef` can be useful for comparing levels for the different factors.

```
grass2.glm<-glm(yes/total ~ experiment + treat,
          family = "binomial", data= grass, weights = total)
> dummy.coef(grass2.glm)
Full coefficients are

(Intercept):      0.366342
experiment:              1          2          3          4
                 0.0000000  1.0393635 -0.2951794  0.9551598
treat:             Control      X-ray   Beta-ray
                  0.000000  -1.957590  -1.848509
```

For this example, we will concentrate on `treat` since `experiment` can be considered as a blocking factor. That is we are not really interested in the effect of `experiment` but we want to take this effect into account when we compare levels of `treat`. In comparing the levels of `treat`, it is most convenient to consider the impact on the odds of mitosis. The differences between the entries for Control, Beta-ray, and X-ray from `dummy.coef` represent the differences in the estimated logits (log odds). These differences can be converted to multiplicative factors for comparing the odds for different levels by applying the exponential function. For example, according to the difference in log odds of mitosis between the control group and the X-ray group is $0 - (-1.96) = 1.96$. Thus the odds of mitosis for the control group are estimated to be $\exp(1.96) = 7.10$ times the odds of mitosis for the X-ray group. In a similar manner, our model estimates that the odds for the control group are $\exp(1.85) =$

6.36 times the odds for the Beta-ray group, and that the odds for the Beta-ray group are exp(.109) = 1.12 times the odds for the X-ray group.

5.3 Contingency tables

A contingency table is convenient way of displaying data that results from the classification of a number of subjects or cases into different categories. We will illustrate the analysis of contingency table data using three examples.

Example 12. In our first example, Table 5.7 contains data on U.S. deaths by falling (1970 data). Each death is classified by a single factor, the month of occurrence. The table records the number of deaths that occurred during each month in 1970.

Table 5.7: U.S. deaths by falling, 1970.

Month	Number of falls	Month	Number of falls
Jan	1688	July	1406
Feb	1407	Aug	1446
Mar	1370	Sept	1322
Apr	1309	Oct	1363
May	1341	Nov	1410
June	1388	Dec	1526

Example 13. For our second example, consider the data in Table 5.8, where 655 students chosen at random from the student body at Auckland University are cross-classified according to their degree course and socio-economic status (SES). This is a two-way table since students are classified by two criteria (factors). Each combination of SES and degree creates a cell in the table. As there are 7 categories of degree and 6 categories of SES the total number of cells is $7 \times 6 = 42$. The table records the number of students that fall into each cell.

Table 5.8: University of Auckland students classified by degree and SES.

	Degree enrolled for						
SES	Arts	Science	Law	Engineering	Commerce	Medicine	Other
1	76	28	38	28	17	23	27
2	44	31	25	17	24	9	14
3	37	14	8	20	16	4	12
4	38	12	9	19	15	2	110
5	4	55	0	3	1	1	1
6	9	4	3	4	2	1	0

Example 14. For our third example of a contingency table, consider the data in Table 5.9, where 123 patients suffering from diabetes are classified on the basis of three criteria. Each

Table 5.9: Diabetes patients classified by three criteria.

Family history of diabetes		Yes		No	
Dependent on insulin injections		Yes	No	Yes	No
Age at onset	< 45	6	1	16	2
	≥ 45	6	36	8	48

criteria divides the patients into 2 categories. Thus we have a three-way table with a total of $2 \times 2 \times 2 = 8$ cells.

In general a k-way contingency table consists of cases or subjects that have been cross-classified using k different criteria. The total number of cells in the table, denote this as m, is simply the product of the numbers of categories for each of the criteria. For contingency table data our analysis involves comparing the probabilities that an observation occurs in the different cells. For k-way tables where $k > 1$, we are typically interested in investigating the relationships between the k criteria. The models we use differ from regression models in that none of the k factors is singled out as a response – in effect the cell counts are the response.

5.3.1 Analysis of one-way tables

Consider the death by falling data (example 12) and suppose we wish to answer the question: Is the frequency of death by falling related to month? Our strategy will be to identify a suitable model for the data and then use this model to answer the question. We will consider two different models that could be used for the data in Table 5.7: (i) a multinomial sampling model, and (ii) a Poisson sampling model.

Multinomial sampling model

Suppose that the probability that an individual chosen at random is placed in the ith cell is π_i, for $i = 1, \ldots, m$. We assume that the categories are defined such that each case is classified into exactly one cell. This implies that $\pi_1 + \cdots + \pi_m = 1$. Let $Y_1,\ Y_2,\ \ldots Y_m$ be random variables denoting the number of cases that are classified into each of the cells. Provided that the sample represents a small fraction of the total population, the observed set of counts $(y_1, y_2, \ldots y_m)$ can be considered as an observation from a multinomial distribution. The multinomial distribution extends the binomial distribution to situations where there are more than 2 possible outcomes. The multinomial probability function is:

$$\Pr(Y_1 = y_1,\ Y_2 = y_2, \ldots Y_m = y_m) = \frac{n!}{y_1! y_2! \ldots y_m!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_m^{y_m}. \qquad (5.4)$$

We want to test the hypothesis that $\pi_1 = \pi_2 = \ldots = \pi_{12} = \frac{1}{12}$ – i.e. someone who has died from falling is equally likely to have done so in any given month. To do this, we will consider whether a multinomial model with $\pi_i = 1/12$ for $i = 1, \ldots, 12$ provides reasonable fit to our data.

The test we will use is based on the deviance for this model. The likelihood function for the multinomial model is the probability function from equation 5.4, regarded as a function of the π_i's. To find the deviance we need to find the maximum value of the likelihood function both under the maximal model and under the hypothesized model:

$$\text{deviance} = 2\log \mathrm{L}_{\max} - 2\log \mathrm{L}_{\mathrm{mod}}.$$

Let $\tilde{\pi}_i$, $i = 1, \ldots, m$, be the values of the probabilities for the maximal model (these are the values that maximise equation 5.4 with the only restriction being $\sum_i \tilde{\pi}_i = 1$) and let $\hat{\pi}_i$'s be the probabilities that maximise equation 5.4 under the constraints imposed by the hypothesised model. Then the deviance can (with a bit of algebraic manipulation) be written as:

$$\text{deviance} = 2\sum_{i=1}^{m} y_i \log \tilde{\pi}_i - 2\sum_{i=1}^{m} y_i \log \hat{\pi}_i.$$

For the multinomial distribution, the values of the $\tilde{\pi}_i$'s are always given by the observed proportions: $\tilde{\pi}_i = y_i/n$. Thus to get $\mathrm{L}_{\max}$ we substitute these into equation 5.4. For this example, the hypothesised model completely specifies the $\hat{\pi}_i$'s and so to get $\mathrm{L}_{\mathrm{mod}}$ we substitute $\pi_i = 1/12$ for all i into equation 5.4. Thus the expression for the deviance becomes:

$$\text{deviance} = 2\sum_{i=1}^{12} y_i \log\left(\frac{y_i}{n}\right) - 2\sum_{i=1}^{12} y_i \log\left(\frac{1}{12}\right).$$

This can be evaluated using R as follows:

```
> y<-c(1688,1407,1370,1309,1341,1388,1406,1446,1322,1363,1410,1526)
> n<-sum(y)
> maximal<-2*sum(y*log(y/n))
> model<-2*sum(y*log(1/12))
> maximal-model
[1] 81.09515
```

Under the hypothesised model, the deviance has a χ^2 distribution with $m-1-c$ degrees of freedom, where c is the number of parameters that must be estimated under the hypothesised model. In our case $c = 0$ since the hypothesised model completely specifies the $\hat{\pi}_i$'s. Thus the degrees of freedom for the χ^2 reference distribution is $12 - 1 - 0 = 11$ and the p-value is calculated by $\Pr(\chi^2_{11} \geq 81.0951)$:

```
> 1-pchisq(81.09515,11)
[1] 9.05831e-13
```

The p-value is very small indicating very strong evidence against the hypothesised model. That is we have very strong evidence against $\pi_i = 1/12$ for all i.

Poisson sampling model

A second way we could look at the death by falling data is to use a Poisson sampling model. In this case, we consider the number of deaths by falling for month i is an observation from a Poisson distribution with mean λ_i. To answer our question, "Is the frequency of deaths by falling related to month?", we test the hypothesis that $\lambda_1 = \lambda_2 = \ldots = \lambda_{12}$. The Poisson probability function is:

$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

Thus if we assume that the observations are independent, then the likelihood function is:

$$\Pr(Y_1 = y_1,\ Y_2 = y_2, \ldots Y_m = y_m) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}. \tag{5.5}$$

Again we will use a test that is based on the deviance of the hypothesised model. To calculate this deviance we will need to determine the values of $\mathrm{L}_{\max}$ and $\mathrm{L}_{\mathrm{mod}}$ for the Poisson distribution. The maximal model corresponds to the set of values for the λ_i's that maximise equation 5.5 when no restrictions are placed on the λ_i's. We will denote these values as $\tilde{\lambda}_i$ for $i = 1, \ldots, m$. The maximal model corresponds to $\tilde{\lambda}_i = y_i$ and $\mathrm{L}_{\max}$ is obtained by substituting these into equation 5.5. To find $\mathrm{L}_{\mathrm{mod}}$ we need to maximise equation 5.5 under the restriction that $\lambda_1 = \lambda_2 = \ldots = \lambda_m$. This is achieved by selecting $\hat{\lambda}_i = (\sum y_i)/n = \bar{y}$ for all i. Using these values for the $\tilde{\lambda}_i$'s and the $\hat{\lambda}_i$'s and with a bit of algebraic manipulation, the deviance for the Poisson sampling model can be written as:

$$\begin{aligned}
\text{deviance} &= 2\log \mathrm{L}_{\max} - 2\log \mathrm{L}_{\mathrm{mod}} \\
&= 2\sum_{i=1}^{m}(\tilde{\lambda}_i + y_i \log \tilde{\lambda}_i) - 2\sum_{i=1}^{m}(\hat{\lambda}_i + y_i \log \hat{\lambda}_i) \\
&= 2\sum_{i=1}^{m}(y_i + y_i \log y_i) - 2\sum_{i=1}^{m}(\bar{y} + y_i \log \bar{y}).
\end{aligned}$$

Using R to calculate the deviance gives:

```
> ybar<-mean(y)
> 2*(sum(-y+y*log(y))) -2*(sum(-ybar+y*log(ybar)))
[1] 81.09515
```

Notice we get exactly the same deviance as we did using the multinomial model. This is no accident. The two procedures will always give us the same deviance and thus the same p-value when testing if an hypothesised model is adequate. A third, more convenient, method of getting the same result is to use the `glm` function of R to create a Poisson regression model. To do this we create a factor whose levels represent the different cells in our contingency table. Then the cell counts are modeled as function of this factor using the `family=poisson` option of `glm`:

```
y<-c(1688,1407,1370,1309,1341,1388,1406,1446,1322,1363,1410,1526)
> month<-c("jan","feb","mar","apr", "may", "jun",
                             "jul","aug","sep", "oct","nov","dec")
> month<-factor(month)
> falls.glm<-glm(y~month,family=poisson)
> anova(falls.glm,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: y
Terms added sequentially (first to last)
      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     11     81.095
month 11   81.095         0  5.151e-14 9.059e-13
```

Analysis of two-way tables.

Consider Table 5.8 which classifies University of Auckland students according to degree programme and socio-economic status. Often with 2-way tables we are interested in how the two factors are related. For the current example, suppose we wish to investigate how the degree a student enrolls in is related to socio-economic status. The first question we should ask is "Are the two factors related?". To do this we test the hypothesis that the factors are independent.

For two-way tables it is convenient to use a slightly different system of indexing the cells in the table: let cell ij be the cell corresponding to row i and column j. Thus an observation that is put into cell ij is in category i of the row factor and in category j of the column factor. Let Y_{ij} be the random variable representing the count for cell ij and y_{ij} be the observed number of counts. Further let I and J represent the number of categories for the row factor and for the column factor respectively and thus the total number of cells is $m = IJ$.

Suppose we adopt a multinomial sampling model. Let π_{ij} represent the probability that a randomly selected individual is placed in the ijth cell. If the row factor and the column factor are independent, then this probability will be equal to the probability that the individual occurs in row i (call this π_{i+}) multiplied by the probability that the individual occurs in row j (call this π_{+j}):

$$\pi_{ij} = \pi_{i+} \times \pi_{+j} \quad \text{if the factors are independent.}$$

Note that π_{i+} is the sum of the π_{ij}'s for the cells in row i, $\pi_{i+} = \sum_j \pi_{ij}$, and that π_{+j} is the sum of the π_{ij}'s for the cells in column j, $\pi_{+j} = \sum_i \pi_{ij}$.

To test the hypothesis that the row factor and the column factor are independent we need to calculate the deviance for the "independence model". Thus we need to find $\mathrm{L}_{\max}$ and $\mathrm{L}_{\mathrm{mod}}$. If we adjust the probability function for the multinomial distribution (equation 5.4)to our new indexing system we get

$$\Pr(Y_{11} = y_{11},\ Y_{21} = y_{21}, \ldots Y_{IJ} = y_{IJ}) = \frac{n!}{y_{11}!y_{21}!\ldots y_{IJ}!}\pi_{11}^{y_{11}}\pi_{21}^{y_{21}}\ldots\pi_{IJ}^{y_{IJ}}. \tag{5.6}$$

For the maximal model, the $\tilde{\pi}_{ij}$'s are selected to maximise this expression with the only constraint being that $\sum_i \sum_j \tilde{\pi}_{ij} = 1$. This is accomplished by setting $\tilde{\pi}_{ij} = y_{ij}/n$. For the independence model, the $\hat{\pi}_{ij}$'s are selected to maximise equation 5.6 under the restrictions:

1. $\hat{\pi}_{ij} = \hat{\pi}_{i+} \times \hat{\pi}_{+j}$ for all ij. Note that $\hat{\pi}_{i+}$'s and the $\hat{\pi}_{+j}$ can be interpreted as the estimated row probabilities and the estimated column probabilities respectively.

2. $\sum_i \hat{\pi}_{i+} = 1$ and $\sum_j \hat{\pi}_{+j} = 1$.

For the independence model we get $\hat{\pi}_{i+} = y_{i+}/n$ for $i = 1$ to I and $\hat{\pi}_{+j} = y_{+j}/n$ for $j = 1$ to J where y_{i+} is the total counts in row i and y_{+j} is the total counts in column j of the contingency table. As a result we get $\hat{\pi}ij = y_{i+}y_{+j}/n^2$.

Using these results gives us the following expression for the deviance:

$$\begin{aligned}\text{deviance} &= 2\log \mathrm{L}_{\max} - 2\log \mathrm{L}_{\mathrm{mod}} \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\log\frac{y_{ij}}{n} - 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\log\left(\frac{y_{i+}y_{+j}}{n^2}\right).\end{aligned}$$

If we evaluate this expression for our current example we get a deviance of 355.612. Note that their is a small complication that arises in this calculation since two of the cell counts are 0 which give us terms of $0 \times \log 0$ – to get the correct deviance these terms should be taken as being equal to 0. To find the p-value for our hypothesis test we need to use a χ^2 distribution with $m - 1 - c$ degrees of freedom. For a two-way table $m = IJ$. Recall that c represents the number of parameters that must be estimated under the hypothesised model. For an independence model we must estimate $\hat{\pi}_{i+}$ for $i = 1$ to I and $\hat{\pi}_{+j}$ for $j = 1$ to J. However, since $\sum_i \hat{\pi}_{i+} = 1$ and $\sum_j \hat{\pi}_{+j} = 1$ we only actually estimate $I - 1$ of the $\hat{\pi}_{i+}$'s and $J - 1$ of the $\hat{\pi}_{+j}$'s. Thus $c = I + J - 2$ and our expression becomes $m - 1 - c = IJ - 1 - (I + J - 2) = (I - 1) \times (J - 1)$. For our example $I = 6$ and $J = 7$ so we use a χ^2 distribution with $5 \times 6 = 30$ degrees of freedom.

```
> 1-pchisq(355.612,30)
[1] 0
```

Thus the p-value is extremely small which represents extremely strong evidence against the hypothesis of independence.

A easier method of obtaining this result would be to use the `glm` function to fit a Poisson regression model. To do this we define variables to represent the row and the column factors and use these as regressors in a Poisson regression model of the cell counts. In this framework, we test to see whether the two factors are related by testing whether the interaction between the two factors is needed in the model.

```
> degree<-rep(c("arts","sci","law","eng","com","med","oth"),6)
> ses<-rep(1:6,rep(7,6))
> nums<-c(76,28,38,28,17,23,27,44,31,25,17,24,9,14,37,14,8,20,
+ 16,4,12,38,12,9,19,15,2,110,4,55,0,3,1,1,1,9,4,3,4,2,1,0)
> degree<-factor(degree)
> ses<-factor(ses)
> enrol.glm<-glm(nums~degree*ses,family=poisson,maxit=25)
> anova(enrol.glm,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: nums
Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          41     829.93
degree      6   182.38        35     647.56 1.062e-36
ses         5   291.95        30     355.61 5.394e-61
degree:ses 30   355.61         0  2.458e-06 2.367e-57
```

Note that the line for `degree:ses` gives us the same deviance and degrees a freedom as we got for testing the independence hypothesis using the multinomial sampling model.

Analysis of three-way tables.

Suppose we now have classification factors A, B, and C with I, J, and K levels respectively. The three-dimensional table is illustrated in Figure 5.10. Factor A is the "row" factor, factor B the "column" factor and factor C the "slice" factor. We extend the terminology we used for two-way tables in the obvious manner. Let π_{ijk} be the probability that a case is classified into the ijkth cell, let Y_{ijk} be the random variable of counts for this cell and let y_{ijk} be the observed counts.

Our aim is to investigate how the classification factors are related to each other. The first job will be to determine which pair of factors are related to each and which are not (i.e.

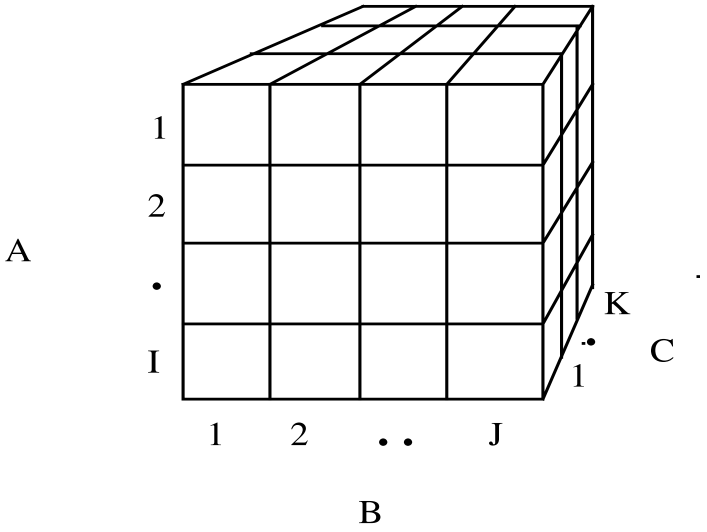

Figure 5.10: A three-dimensional contingency table.

are independent). There are a number of possibilities for the types of independence that can occur among the factors.

Mutually independent factors

From the multinomial sampling perspective, the mutual independence of A, B and C is equivalent to

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \tag{5.7}$$

for all ijk, where $\pi_{i++} = \sum_j \sum_k \pi_{ijk}$, $\pi_{+j+} = \sum_i \sum_k \pi_{ijk}$ and $\pi_{++k} = \sum_i \sum_j \pi_{ijk}$. We could proceed by developing the expression for the deviance under multinomial sampling for this situation and using this value to test whether this mutual independence model is compatible with the observed. Instead we will take the easier option of performing an equivalent test using Poisson regression. The mutual independence model is equivalent to a Poisson regression model that does not contain any interactions whereas the maximal model is equivalent to a Poisson regression model that contains all possible interactions between the three factors. Thus to test for mutual independence we compare the no interaction model (submodel) to the model that includes all possible interactions (full model).

Suppose we wish to test the hypothesis that the three classification factors for the diabetes data (Table 5.9) are mutually independent. This can be done in R as follows:

```
> onset<-rep(c("over","under"),c(4,4))
> history<-c("y","y","n","n","y","y","n","n")
> depend<-c("y","n","y","n","y","n","y","n")
> counts<-c(6,1,16,2,6,36,8,48)
> diabetes.df<-data.frame(onset,history,depend,counts)
> full.glm<-glm(counts~.^3,family=poisson,data=diabetes.df)
> sub.glm<-glm(counts~.,family=poisson,data=diabetes.df)
> anova(sub.glm,full.glm,test="Chi")
Model 1: counts ~ .
```

```
Model 2: counts ~ .^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     51.933
2         0  4.441e-16  4   51.933 1.424e-10
```

Note that using the model statement "counts~.^3" produces a model containing all interactions involving ≤ 3 variables whereas "counts~." produces a model that only contains main effects (no interactions). This test produces a very small p-value and thus we have very strong evidence against the hypothesis that the three factors are mutually independent.

One factor is independent of the other two

The above test rules out the possibility that the three factors are mutually independent. However, it still may be the case that one of the factors is independent of the other two. In terms of multinomial sampling, if A is independent of B and C, then

$$\pi_{ijk} = \pi_{i++}\pi_{+jk} \tag{5.8}$$

for all ijk, where $\pi_{i++} = \sum_j \sum_k \pi_{ijk}$ and $\pi_{+jk} = \sum_i \pi_{ijk}$. Again we will test the hypothesis that this is an adequate model by performing an equivalent test using Poisson regression. To test that one factor is independent of the other two using Poisson regression, we define the submodel to be the model that does not contain any interactions involving that factor. For the diabetes example suppose we wish to test that onset is independent of the other two factors. The full model is the maximal model (as before) but now the submodel contains the three main effects and the history:depend interaction. The test using R is:

```
> sub2.glm<-glm(counts~.+history:depend,family=poisson,data=diabetes.df)
> anova(sub2.glm,full.glm,test="Chi")
Analysis of Deviance Table
Model 1: counts ~ . + history:depend
Model 2: counts ~ .^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3     51.023
2         0  4.441e-16  3   51.023 4.838e-11
```

The small p-value provides strong evidence against the hypothesis that onset is independent of the other two factors.

Of course, we should also test that (i) history is independent of onset and depend and (ii) depend is independent of onset and history:

```
> sub3.glm<-glm(counts~.+onset:depend,family=poisson,data=diabetes.df)
> anova(sub3.glm,full.glm,test="Chi")
Analysis of Deviance Table
Model 1: counts ~ . + onset:depend
Model 2: counts ~ .^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3    1.94634
2         0  4.441e-16  3  1.94634   0.58362

> sub4.glm<-glm(counts~.+onset:history,family=poisson,data=diabetes.df)
> anova(sub4.glm,full.glm,test="Chi")
Analysis of Deviance Table
Model 1: counts ~ . + onset:history
```

```
Model 2: counts ~ .^3
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3     50.034
2         0  4.441e-16  3   50.034 7.859e-11
```

The first test gives no evidence against the hypothesis that `history` is independent of `onset` and `depend` whereas the second test gives strong evidence against the hypothesis that `depend` is independent of `onset` and `history`. If we consider the results from all the hypothesis tests, we conclude that they are consistent with `onset` and `depend` being related each other but both being independent of `history`. This conclusion is supported by looking at the output from `anova` for the model that contains the main effects plus the `onset:depend` interaction:

```
> anova(sub3.glm,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: counts
Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             7    125.163
onset         1   46.314         6     78.848 1.007e-11
history       1    5.117         5     73.731     0.024
depend        1   21.798         4     51.933 3.029e-06
onset:depend  1   49.987         3      1.946 1.548e-12
```

Clearly, the `onset:depend` interaction is needed in the model which implies that `onset` is related to `depend`. A previous test indicated that this model was adequate – i.e. none of the other interactions are required.

Example 15. In this example, Table 5.10, the 326 "cases" are convicted defendants in homicide indictments in 20 Florida counties 1976-77. The factors are Defendant's Race, Victim's Race, and Death Penalty. Table 5.10 gives a more detailed breakdown of a larger set of Florida murder data, with a third variable (Victim's race) added.

Table 5.10: The Florida murder data.

	Victim's Race			
	Black		White	
Defendant's	Death Penalty		Death Penalty	
Race	Yes	No	Yes	No
Black	13	195	23	105
White	1	19	39	265

To identify how the three factors are related to each other, we want to identify which interactions are needed in a Poisson regression model that uses the cell counts as the response. Rather than sort through different possible models ourselves as we did in the previous example, we will let R do this for us. The `step` function in R can be used to do stepwise model selection and thus eliminate much of the hassle in identifying which of the interactions are needed in the model. For this example, first we create a data frame and fit a model that just contains the main effects:

```
> defendant<-rep(c("b","w"),c(4,4))
> victim<-c("b","b","w","w","b","b","w","w")
```

```
> dp<-rep(c("y","n"),4)
> dp.df<-data.frame(defendant,victim,dp,counts)
> dp.df
  defendant victim dp counts
1         b      b  y     13
2         b      b  n    195
3         b      w  y     23
4         b      w  n    105
5         w      b  y      1
6         w      b  n     19
7         w      w  y     39
8         w      w  n    265

> dp.glm<-glm(counts ~.,family = poisson,data=dp.df)
```

For the step command, we need to specify a starting model and a model that contains all the terms that we want to be considered. For our purposes, the starting will just contain the main effects (dp.glm) and the largest possible model will contain the main effects plus all the interactions. The largest possible model is specified using the scope option in step:

```
> dp.steps<-step(dp.glm,scope=~.^3)
```

Finally, by applying the anova command to the object produced by step we can viewed the model selected by *R*:

```
> anova(dp.steps,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: counts

Terms added sequentially (first to last)
                 Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                 7     774.73
defendant         1     0.22         6     774.51       0.64
victim            1    64.10         5     710.42 1.183e-15
dp                1   443.51         4     266.90 1.861e-98
defendant:victim  1   254.15         3      12.75 3.230e-57
victim:dp         1    10.83         2       1.92 9.999e-04
```

The selected model contains the defendant:victim and the victim:dp interactions. This indicates that the defendant's race is related to the victim's race and that death penalty is related to the victim's race. The relationship between death penalty and the defendant's race is interesting. Since there is no interaction in the selected model that contains both defendant and dp, these two factors are not directly related to each other. However, they are not independent either since they are connected through victim's race – i.e. both are related to victim's race. We say that death penalty and defendant's race are conditionally independent. What this means is that if we fix the level of a third factor (victim's race in this example) they become independent. So if we were to only consider the data where victim's race = black then death penalty and defendant's race are independent. Similarly, if we only consider the data where victim's race = white then death penalty and defendant's race are independent. **But** if we combine the data for victim's race = black and victim's race = white then death penalty and defendant's race appear to be related. We will discuss this phenomenon further in the section on Simpson's paradox.

5.3.2 The odds ratio

The odds ratio is a common method of exploring relationships between classification factors. Consider a two way table with two factors each at two levels as illustrated in Table 5.11.

Table 5.11: The 2×2 table.

		Columns		
		1	2	Total
Rows	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
	Total	π_{+1}	π_{+2}	1

Suppose we want to investigate the relationship between the row factor and the column factor. The odds ratio does this by considering the odds that a case occurs in a given row conditional on the column that it occurs in. Let $\text{odds}(r = i|c = j)$ represent the odds that a case occurs in row i given that it occurs in column j and recall that the definition of odds is $\text{odds}(r = i|c = j) = \Pr(r = i|c = j)/\Pr(r \neq i|c = j)$. For our table we can investigate the relationship between the row factor and the column factor by considering $\alpha = \text{odds}(r = 1|c = 1)/\text{odds}(r = 1|c = 2)$ which for obvious reasons is called an odds ratio. We can write the odds ratio in terms of the cell probabilities:

$$\begin{aligned}\alpha &= \frac{\text{odds}(r=1|c=1)}{\text{odds}(r=1|c=2)}\\ &= \frac{(\pi_{11}/\pi_{+1})/(\pi_{21}/\pi_{+1})}{(\pi_{12}/\pi_{+2})/(\pi_{22}/\pi_{+2})}\\ &= \frac{\pi_{11}\,\pi_{22}}{\pi_{21}\,\pi_{12}} \qquad (5.9)\end{aligned}$$

The odds ratio has the following properties:

1. If we had defined the odds ratio as $\alpha = \text{odds}(c = 1|r = 1)/\text{odds}(c = 1|r = 2)$, i.e. simply reverse the roles of rows and columns, we would get exactly the same expression (equation 5.9) in terms of the cell probabilities.

2. If the row factor and the column factor are independent then the odds ratio will be 1. To see this recall that independence of two factors is equivalent to $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all ij and apply this relationship to each term in equation 5.9.

The estimated odds ratio

In practice, we have to estimate the odds ratio. Suppose we take a sample of n cases and classify them into a 2×2 table where y_{11}, y_{21}, y_{12}, y_{22} represent the observed counts. An obvious estimate of the odds ratio is obtained by using $\hat{\pi}_{ij} = y_{ij}/n$ and substituting the estimated probabilities into equation 5.9:

$$\hat{\alpha} = \frac{\hat{\pi}_{11}\,\hat{\pi}_{22}}{\hat{\pi}_{12}\,\hat{\pi}_{21}} = \frac{y_{11}\,y_{22}}{y_{12}\,y_{21}}.$$

The sampling distribution of $\hat{\alpha}$ is highly skewed but for large n the distribution of $\log\hat{\alpha}$ is approximately Normal with mean $\log\alpha$ and standard error:

$$\text{s.e. }(\log\hat{\alpha}) = \sqrt{\frac{1}{y_{11}} + \frac{1}{y_{12}} + \frac{1}{y_{21}} + \frac{1}{y_{22}}}.$$

An approximate test of $\log\alpha = 0$ (i.e. of $\alpha = 1$ or independence) can be conducted by calculating the test statistic:

$$z_o = \log\hat{\alpha}/s.e.(\log\hat{\alpha}).$$

The p-value is given by $2 \times \Pr(Z \geq |z_o|)$ where $Z \sim \mathrm{N}(0,1)$.

Confidence intervals for $\log\alpha$ can be created based on this Normal approximation and then transformed into intervals for α using the exponential the exponential function. For example, a 95% confidence interval for $\log\alpha$ would be given by:

$$\log\hat{\alpha}) \pm 1.96\sqrt{\frac{1}{y_{11}} + \frac{1}{y_{12}} + \frac{1}{y_{21}} + \frac{1}{y_{22}}}.$$

Applying the exponential function to the endpoints will give an interval for α.

Example 16. The data in Table 5.12 was extracted from a larger study. It cross classifies 2121 people (who don't have cardiovascular disease and don't exercise regularly) according to personality type (A: show signs of stress, uneasiness, and hyperactivity or B: relaxed, easygoing and exercise regularly) and cholesterol level.

Table 5.12: Subjects classified by cholesterol level and personality type

	Cholesterol		
Personality	Normal	High	Total
A	795	232	1027
B	886	208	1094
total	1681	440	2121

First we should test the hypothesis that personality type and cholesterol are independent:

```
> counts<-c(795,232,886,208)
> personality<-c("A","A","B","B")
> chol<-c("N","H","N","H")
> personality<-factor(personality)
> chol<-factor(chol)
> fit<-glm(counts~chol*personality,family=poisson)
> anova(fit,test="Chi")
Analysis of Deviance Table
Model: poisson, link: log
Response: counts
Terms added sequentially (first to last)
                 Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                3     780.78
chol              1   774.55         2       6.24 1.847e-170
personality       1     2.12         1       4.12       0.15
chol:personality  1     4.12         0  1.301e-13       0.04
```

As we have evidence against the independence hypothesis, it is of interest to explore the relationship between personality type and cholesterol level. The estimated odds ratio is:

$$\hat{\alpha} = \frac{795 \times 208}{232 \times 886} = 0.804$$

Thus the odds that a personality type A subject has a normal cholesterol level is estimated to be about 80 % of the odds for a personality type B subject. A 95 % confidence interval for the log(odds ratio) is calculated as follows:

$$\begin{array}{rcl} \log(0.804) & \pm & 1.96\sqrt{\frac{1}{795}+\frac{1}{232}+\frac{1}{886}+\frac{1}{208}} \\ -0.2176 & \pm & 1.96 \times 0.1073 \\ & & (-0.4278,\ -0.0073) \end{array}$$

To convert this into a 95 % confidence interval for the odds ratio, we apply the exponential function to the limits: $(\exp(-0.4278),\ \exp(-0.0073)) = (.652,\ .993)$.

5.3.3 Simpson's paradox

Given a three-way table we may *collapse* the table over one of the factors to produce a two-way table in the other factors. In effect, we simply cross-classify the cases by the two remaining factors ignoring the first factor. The idea is that this allows us to focus on how the remaining factors are related. However, we must be careful in interpreting the results from a collapsed table in situations where the collapsed factor is related to the other factors. We may find that if we were to create a separate two-way table for each level of the collapsed factor that these tables give a very different picture to that given by the collapsed table.

The Florida murder data, Table 5.10, illustrates this point nicely. Suppose that we create two separate tables based on the Victim's race:

Victim=Black

	Death penalty=Yes	Death penalty=No
Defendant=Black	13	195
Defendant=White	1	19

Victim=White

	Death penalty=Yes	Death penalty=No
Defendant=Black	23	105
Defendant=White	39	265

The estimated odds ratio for the first table is 1.27 and for the second table is 1.49. for which the odds ratio is 1.4884. Since both odds ratios are more than 1, this suggests that odds of getting the death penalty are higher for black defendants than for white defendants. Actually, our analysis on page 43 indicates that the data in each of these tables is compatible with Defendants race and death penalty being independent in each case (i.e. the true odds ratio being 1). However there is certainly nothing in these two tables that would support the hypothesis that the odds of getting the death penalty is higher for white defendants.

Now suppose we collapse the table on Victim's race:

	Death penalty=Yes	Death penalty=No
Defendant=Black	36	300
Defendant=White	40	284

The estimated odds ratio is now 0.852 which suggests that the odds of getting the death penalty is *lower* for black defendants. Thus the collapsed table is suggesting the opposite conclusion to that suggested by the separate tables.

The reason for this apparent paradox is simple: about 6% of cases involving a black victim result in the death penalty, while for white victims the percentage is about 17%. Since people tend to murder people of their own race, the collapsed table appears to show that whites get more death sentences. Using victim's race to create two separate tables, allows the true picture to emerge.