DEPARTMENT OF STATISTICS

Paper 475.330

ADVANCED STATISTICAL MODELLING

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

February 1995

Contents

5 Models for categorical responses **5**

5.1 Introduction . . . 5

5.2 Logistic regression analysis . . . 5

5.2.1 Fitting the model . . . 7

5.2.2 Multiple logistic regression . . . 9

5.2.3 Analysis of grouped data using logistic regression . . . 13

5.2.4 Deviances, goodness of fit and comparing models . . . 15

5.2.5 Residuals in logistic regression . . . 23

5.2.6 Binary anova . . . 29

5.3 Analysis of contingency tables . . . 34

5.3.1 Likelihood based inference for contingency tables . . . 35

5.3.2 Chi-square tests . . . 38

5.3.3 The odds ratio . . . 39

5.3.4 Log-linear models . . . 40

5.3.5 Three-dimensional tables . . . 44

5.3.6 Residual analysis for contingency tables . . . 47

5.3.7 Simpson's paradox . . . 48

Chapter 5

Models for categorical responses

5.1 Introduction

In this chapter we deal with models that are appropriate when the response variable is categorical, rather than continuous as in earlier chapters.

Suppose that our categorical response Y can be one of the values $y_1, \ldots, y_p$ with probabilities $\pi_1, \ldots, \pi_p$ respectively, where $\pi_1 + \pi_2 + \ldots + \pi_p = 1$. Most of the models considered in this chapter express how the π's are related to whatever explanatory variables are of interest. We begin by considering the analogue of regression models, in the special case of a *binary response* (i.e. responses that can have only two values, $p = 2$ above).

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`). What happens if we try to fit a regression model to these data? It is clear that using regression, which is a model designed for continuous responses that can take any value, will not work very well in a situation where the responses are binary (taking the values 0 and 1 say). In particular, the fitted values are unlikely to be either zero or one!

Regression models assume that the responses are normally distributed. We need a model that assumes a distribution that is more suitable for binary data. The binomial distribution provides a likely model.

We can imagine that observing an individual and noting the presence or absence of CHD is equivalent to conducting a single Bernoulli trial. Thus the distribution of the response for the ith case is $\text{Bin}(1, \pi)$, where π is the probability a patient will have CHD. (We identify CHD with "success"in this case.) The "success probability" π is modelled in terms of the explanatory variables which are supposed to affect the chance of a patient having CHD.

How can this be done? Experience tells us that the chance of a patient having CHD increases with age. If we try a linear model as in regression i.e. if we use an equation like

$$\pi = \alpha + \beta x$$

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				

where x is the age of the patient, we have a problem. Since π is a probability, it must lie between zero and one. The linear function unfortunately does not have this property, so the model permits negative probabilities. Nasty!

The solution is to model the probability π as some suitable function of the linear expression $\alpha + \beta x$. The function is chosen to make sure the probability stays between zero and one. A commonly used function is the *logistic function*, which is of the form

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}.$$

The parameters α and β determine the shape of the curve. If $\beta > 0$ then the probability of a "1" increases as x increases, and if $\beta < 0$ then the probability of a "1" decreases as x increases. If $\beta = 0$ then the response is unrelated to x. Graphs of the logistic function are shown in Figure 5.1.

There is a connection between β and the odds ratio. Imagine that the explanatory variable x is also binary, having values zero and one say. We can think of $\pi(x)$ as the conditional probability of a success, given that the explanatory variable is x. Then β is the population log odds ratio for the corresponding two-way table classifying cases according to the values of the binary response and the binary explanatory variable.

Note that the inverse of the function

$$\frac{e^x}{1 + e^x}$$

is the so-called *logit* function

$$\operatorname{logit} p = \log_e \left(\frac{p}{1-p} \right).$$

Thus the model above can be written

$$\operatorname{logit} \pi(x) = \alpha + \beta x. \tag{5.1}$$

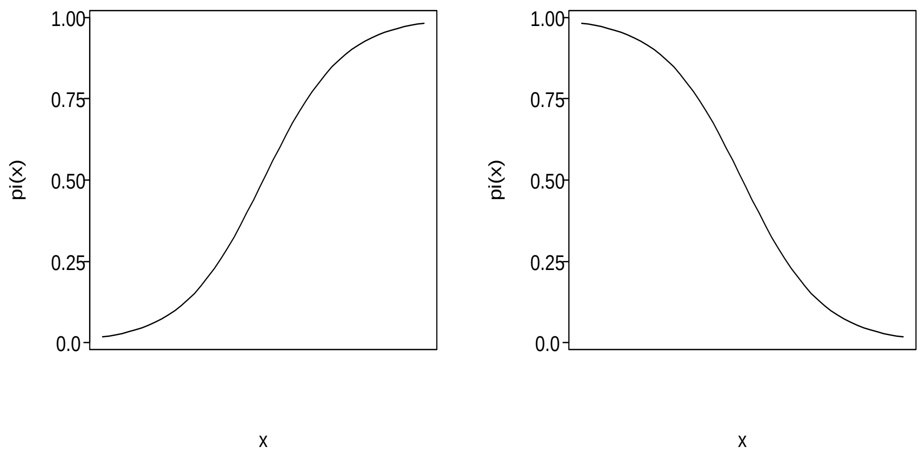

Figure 5.1: Shape of the logistic curve. (i) $\beta > 0$, (ii) $\beta < 0$.

This is usually called the *logistic model.*

Example 2. The risk of CHD increases with age, so if $\pi(x)$ is the probability that a randomly chosen individual aged x has CHD, a reasonable model for π might be a logistic model, with $\beta > 0$.

5.2.1 Fitting the model

To fit the model, we observe n cases, and for each record y and x. Let y_i and x_i be the measurements for the ith case, $i = 1, \ldots, n$. We want to estimate the values of α and β, and calculate a standard error for each. We will use the method of *maximum likelihood* to do this, and we briefly review this method.

Suppose in general each response has a density (or probability function) f that depends on y and a vector γ of unknown parameters. In the present case, $\gamma = (\alpha, \beta)$, and the probability function is

$$f(y_i, \gamma) = \begin{cases} \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}, & y_i = 1, \\ \frac{1}{1+\exp(\alpha+\beta x_i)}, & y_i = 0. \end{cases}$$

This can be more compactly written as

$$\begin{aligned} f(y_i, \gamma) &= \left\{ \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \right\}^{y_i} \left\{ \frac{1}{1 + \exp(\alpha + \beta x_i)} \right\}^{1-y_i} \\ &= \frac{\exp(\alpha + \beta x_i)^{y_i}}{1 + \exp(\alpha + \beta x_i)}. \end{aligned}$$

Provided the observations are independent, the *joint density* of the observed responses is

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i, \gamma)$$

and regarded as a function of γ, is called the *likelihood function* and is denoted by $L(\gamma)$. The *maximum likelihood estimates* of α and β are just the values of α and β that maximise the likelihood. To simplify the mathematics, we usually work with the quantity

$$-2\ell = -2\log L$$

The maximum likelihood estimates are the values that minimise -2ℓ.

For the logistic model above, -2ℓ is

$$-2\sum_{i=1}^{n}\{y_i(\alpha+\beta x_i) - \log(1+\exp(\alpha+\beta x_i)\}. \tag{5.2}$$

The minimising values are found by differentiating (5.2) and equating the result (called the score function) to zero. This leads to the equations

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \pi(x_i)) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - \pi(x_i)) &= 0 \end{aligned} \tag{5.3}$$

which are solved numerically by an iterative process known as *iteratively reweighted least squares*, or IRLS. This is a special case of another algorithm called *Fisher scoring.* The standard errors are approximately given by inverting a certain matrix of expected second derivatives. See 528.381 for more details.

Example 3. To fit the logistic model to our CHD data, we use the function `glm`. GLM stands for "generalised linear model", a class of models that includes the logistic model. GLM's are discussed in more detail in Section 5.4. In the function, we use the same type of model formula as in linear regression. Suppose the binary response is denoted by `y`, which has value 1 for CHD and 0 for the absence of CHD. Suppose also that the variable `age` contains the ages of the 100 patients, and that the data are in a data frame `chd`. We specify a logistic model by the argument `family=binomial`. Otherwise the syntax is exactly the same as for `lm`. We fit the model (5.1) by typing

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

This produces a "glm object" `chd.glm`, which contains information about the fit. We can examine these results using the `summary` function just as we did for linear models:

```
> summary(chd.glm)

Call: glm(formula = y ~ age, family = binomial, data = chd)
Deviance Residuals:
       Min         1Q     Median        3Q      Max
 -1.968565 -0.8480081 -0.4607374 0.8261872 2.279377

Coefficients:
                 Value Std. Error   t value
(Intercept) -5.2784441 1.13033699 -4.669797
        age  0.1103208 0.02401445  4.593933
```

```
(Dispersion Parameter for Binomial family taken to be 1 )

Null Deviance: 136.663 on 99 degrees of freedom

Residual Deviance: 107.6806 on 98 degrees of freedom

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:
    (Intercept)
age -0.9780126
```

The estimate of α is -5.2784 and that of β is 0.1103. The standard errors are 1.1305 and 0.0240. Note also how the iterative calculations have proceeded: we needed 4 iterations of the Fisher scoring (i.e. IRLS) algorithm.

We now turn to the problems of assessing the significance of the regression coefficients. Specifically, how do we find confidence intervals for α and β and test if they are equal to zero?

We use the fact that the estimates of the regression coefficients are asymptotically normally distributed, with mean equal to the true value, and a standard deviation which is estimated by the standard error computed from the likelihood function. A 95% confidence interval for e.g. β is thus

$$\hat{\beta} \pm s.e.(\hat{\beta}) \times 1.960.$$

A test of $\beta = 0$ is obtained by comparing $\hat{\beta}/s.e.(\hat{\beta})$ to standard normal tables, or equivalently, by using the square of this and computing p-values using the χ^2_1 distribution. The p-value will be the area under the χ^2_1 density to the right of the quantity $\left(\hat{\beta}/s.e.(\hat{\beta})\right)^2$. This method of testing is known as the Wald test. The terms "null deviance" and "residual deviance" are explained below in Section 5.2.4.

Example 4. From the output in Example 3, we see that the coefficients α and β are both significantly different from zero, since the p-value for both coefficients is small. A confidence interval for α is $-5.2784 \pm 1.1305 \times 1.96$ i.e. -5.2784 ± 2.2157 and the corresponding interval for β is 0.1103 ± 0.04704. To test if $\beta = 0$ (i.e. that the variable x is not associated with the response), we compute $\hat{\beta}/s.e.(\hat{\beta}) = 0.1103/0.0240 = 4.59$. The corresponding p-value is calculated by finding the area under a normal curve beyond ± 4.59. This area is 0.000, indicating a significant regression. We can calculate the area in Splus by typing

```
> z<-4.59
> 2*(1-pnorm(abs(z)))
[1] 4.43246e-06
```

5.2.2 Multiple logistic regression

If we have several explanatory variables $x_1, \ldots, x_k$, we can use the model

$$P(Y = 1) = \pi(x_1, \ldots, x_k)$$

where

$$\pi(x_1, \ldots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{(1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k))}$$

Equivalently, the model can be written

$$\text{logit } P(Y = 1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

The parameters are estimated in the same way by maximum likelihood, using the same function `glm`, except that the model now becomes (for e.g. 3 explanatory variables)

```
y~x1+x2+x3.
```

Confidence intervals and tests for individual beta's equal to zero are carried out as before.

Example 5. The data in Table 5.1 were collected on 83 patients undergoing corrective spinal surgery. The objective was to identify risk factors for kyphosis (flexing of the spine) following surgery. The variable `kypho` was a binary response with 1 indicating the presence of kyphosis and 0 the absence.

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 the data frame `kyphosis` containing these variables.

The first step is to plot the variables:

```
> pairs(kyphosis)
```

The plot of the response versus age shown in Figure 5.2 and indicates that the risk increases and then decreases with age, so that a quadratic term in age should be added to model this aspect of the data. Accordingly we fit a logistic regression of the form

$$\text{logit } p = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ age}^2 + \beta_3 \text{ number} + \beta_4 \text{ start}$$

to the data, using a model statement of the form

```
> kypho.glm<-glm(kypho~ age +  I(age^2)  + number+ start,
+                              family=binomial, data=kyphosis)
> summary(kypho.glm)

Call: glm(formula = kypho ~ age + I(age^2) + number + start, family =
binomial, data = kyphosis)
Deviance Residuals:
       Min         1Q    Median          3Q      Max
 -2.235654 -0.5124374 -0.245114 -0.06111367 2.354818

Coefficients:
                   Value   Std. Error   t value
(Intercept) -4.382604826 2.0342577750 -2.154400
        age  0.081622446 0.0341033361  2.393386
   I(age^2) -0.000396397 0.0001882663 -2.105512
     number  0.426817215 0.2353168863  1.813798
      start -0.203832854 0.0704796671 -2.892080

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 83.23447 on 80 degrees of freedom

Residual Deviance: 54.42776 on 76 degrees of freedom
```

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				

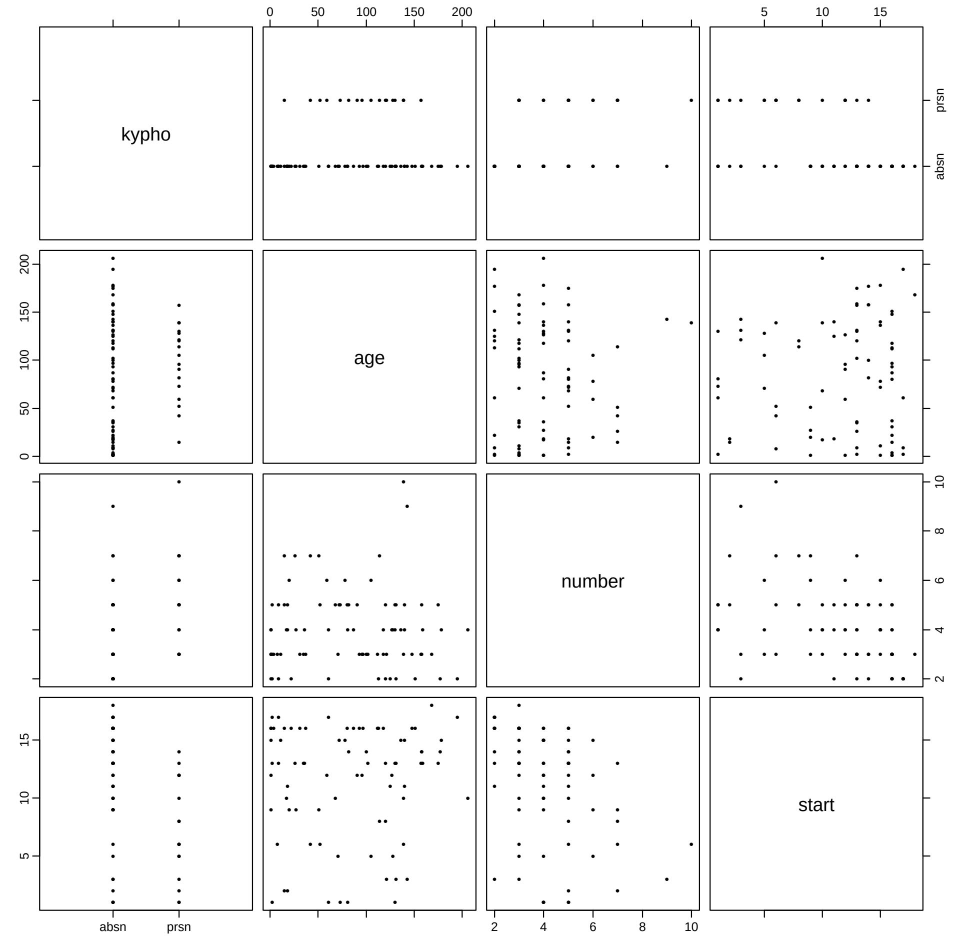

Figure 5.2: Scatterplots for the kyphosis data.

```
Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
         (Intercept)        age   I(age^2)     number
     age -0.6797739
I(age^2)  0.5691270  -0.9649796
  number -0.7284200   0.1739226 -0.0988896
   start -0.2242259  -0.1439661  0.0911896  0.0721616
```

The coefficient of `age`^{2} is negative, indicating that the chance of kyphosis first increases with age and the decreases. The term in `age`^{2} is required in the model, since the p-value is small:

```
> z<- -2.105512
> 2*(1-pnorm(abs(z)))
[1] 0.03524676
```

Similarly, the variable `start` seems important, but the contribution of the variable `number` is rather more uncertain:

```
> z<- 1.813798
> 2*(1-pnorm(abs(z)))
[1] 0.06970883
```

The coefficient of `start` is negative, indicating that the higher the value of `start`, the smaller the logit, and hence the smaller the probability of kyphosis.

5.2.3 Analysis of grouped data using logistic regression

Suppose we have a data set containing n cases, but the number of distinct x-values is much less than n, because of repeats – in this data set there are several distinct cases having the same set of x-values (i.e. having the same values for the explanatory variables, or, putting it another way, the same "covariate vector"). Suppose for every distinct covariate vector $\boldsymbol{x}_i$, $i = 1, \ldots, m$ there are n_i cases, of which s_i are successes (i.e. have $(y = 1)$ and $n_i - s_i$ are failures $(y = 0)$. (This was the situation in Example 2). 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 $\text{Bin}(n_i, \pi_i)$ distribution, where $\pi_i = \Pr[Y = 1|\boldsymbol{x} = \boldsymbol{x}_i]$ 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 $\text{logit}\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

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 6. 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 need to delete this observation from the data. The easiest way to do this is to use the subsetting commands for deleting the corresponding row from the data frame. Thus we use `ingots[-16,]` rather than `ingots` as our input data frame. We fit the model

$$\text{logit}\ \Pr(\text{notready}) = \beta_0 + \beta_1 \times \text{heating time} + \beta_2 \times \text{soaking time}$$

by typing

```
> ingots.glm<-glm(notready/total~heat+soak,weight=total,
+                       family=binomial,data=ingots[-16,])
> summary(ingots.glm)
Call: glm(formula = notready/total ~ heat + soak, family = binomial,
           data = ingots[-16,  ], weights = total)
Deviance Residuals:
       Min         1Q     Median          3Q      Max
 -1.283108 -0.7818358 -0.5051426 -0.09701591 1.719227

Coefficients:
                  Value Std. Error    t value
(Intercept) -5.55915959 1.11865123 -4.9695199
       heat  0.08203067 0.02372086  3.4581654
       soak  0.05677077 0.33098209  0.1715222

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 25.39545 on 18 degrees of freedom

Residual Deviance: 13.75263 on 16 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
     (Intercept)       heat
heat -0.8113857
soak -0.7598041   0.3336745
```

The estimates of β_0, β_1 and β_2 are (standard errors in brackets)

$$\begin{aligned}\hat{\beta}_0 &= -5.5591\ (1.1186)\\ \hat{\beta}_1 &= 0.0820\ (0.0237)\\ \hat{\beta}_2 &= 0.0567\ (0.3309)\end{aligned}$$

It would appear that the soaking times do not affect the probability of being not ready, since the p-value for the hypothesis $\beta_2 = 0$ is 0.8638. However, the heating times clearly do, since the p-value for $\beta_1 = 0$ is 0.0005. Increasing the heating time increases the probability of being not ready.

5.2.4 Deviances, goodness of fit and comparing models

Suppose that we have n cases, and measure a binary response y and k covariates $x_1, \ldots, x_k$ for each case. The covariates could be continuous explanatory variables, or factors, or a mixture of both.

Also suppose that there are several cases having each possible set of x-values. Thus, there are $m < n$ possible covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$. Let n_i be the number of cases having covariate vector $\boldsymbol{x}_i$. Obviously, $n_1 + \cdots + n_m = n$.

Example 7. 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 he 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. In this example there are 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). There are 20 cases (moths) observed for each covariate pattern, so $n = 20 \times 12 = 240$, $m = 12$ and $n_i = 20$ for $i = 1, 2, \ldots, 12$.

We assume that the responses are independent, and the probability that $Y = 1$ for a particular case (e.g. that a moth will be "knocked down") depends only on the covariates. Let s_i be the number of "1" responses from the n_i cases having covariate vector $\boldsymbol{x}_i$. Then the distribution of s_i is binomial, $\mathrm{Bin}(n_i, \pi_i)$, when π_i is the probability a case having covariate vector $\boldsymbol{x}_i$ will be a "1" response.

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

The -2 log likelihood for this model is

$$-2\ell = -2\sum_{i=1}^{m}\{s_i\eta_i + n_i\log(1+e^{\eta_i}))\}$$

where $\eta_i =$ logit π_i. The logistic model specifies that the logit of π_i is linear:

$$\text{logit } \pi_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{5.4}$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$.

Example 8. For the tobacco budworm example, we have two covariates, sex and dose. The data take the form

```
 i  sex dose  s  n
 1   0     1  1 20
 2   0     2  4 20
 3   0     4  9 20
 4   0     8 13 20
 5   0    16 18 20
 6   0    32 20 20
 7   1     1  0 20
 8   1     2  2 20
 9   1     4  6 20
10   1     8 10 20
11   1    16 12 20
12   1    32 16 20
```

and the model is

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

Let $\hat{\beta}$ be the parameter vector $(\hat{\beta}_0, \ldots, \hat{\beta}_k)$ that minimises -2ℓ for this model, and let $-2\ell_{\text{MOD}}$ be the minimum value.

A more general model than the logistic is the "saturated model". This is the model where the π_i's are left unspecified (except that they depend only on the covariate vector in some unspecified way). For example, in the budworm problem, the saturated model would be one with a different (arbitrary) probability π_i for each sex-dose combination.

The value of π_i ($\tilde{\pi}_i$ say) that minimises -2ℓ in this case is given by $\tilde{\pi}_i = s_i/n_i$. Let $-2\ell_{\text{SAT}}$ be the minimum value of -2ℓ for the saturated model. Since the saturated model contains the logistic model, we must have $-2\ell_{\text{SAT}} \leq -2\ell_{\text{MOD}}$.

The difference between those two quantities is a measure of how much the likelihood can be increased (i.e. how much the -2 log likelihood can be decreased) by dropping the restriction that the logistic model, (5.4), holds. The difference is thus a measure of "goodness of fit" of the logistic model, and is called the "deviance". The saturated model is a "benchmark" — it is the model which fits best in the sense of having the smallest -2ℓ.

However, it is usually desirable to have a simple model that demonstrates the connection between the covariates and the probabilities π_i. If the logistic model is "close" to the benchmark (the best possible) then we would prefer it. We interpret "close" as "small deviance".

Provided m is not too big and the n_i's are large, the asymptotic distribution of the deviance when the logistic model is the true one is χ^2_{m-k-1}, although this asymptotic approximation is usually not very accurate. If the asymptotics hold, we can perform a test that the logistic model is correct by computing the deviance, and the calculating a p-value. 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.

To see that the deviance does indeed measure goodness of fit, we can use the following asymptotic approximation: provided the condition above holds (i.e. large n_i's, moderate m) the deviance is approximately equal to

$$\sum_{i=1}^{m} \frac{(\pi_i - \hat{\pi}_i)^2}{\text{Var}[\hat{\pi}_i]}$$

The deviance is rather like a weighted residual sum of squares, and thus is a goodness of fit measure.

Getting deviances in Splus

To calculate the deviance we first fit the model in the usual way, and then examine the output from the `summary` function. The deviance of the model being fitted is called the "residual deviance" on the output. The degrees of freedom $m - k - 1$ are also given. We can also use the `deviance` function:

```
> bugs.glm<-glm(s/n~sex+dose,family=binomial,weight=n,data=bugs)
> deviance(bugs.glm)
[1] 27.96797
```

The conditions for the asymptotics hold, so it is reasonable to interpret the deviance using a p-value : the degrees of freedom are $m - k - 1 = 12 - 2 - 1 = 9$, so the p-value is

```
 1-pchisq(27.96797,9)
[1] 0.0009656927
```

The deviance is much too large, so the logistic model seems implausible. However, it is known from previous experience that a logistic model using log dose rather than dose often fits this type of data well. If we fit the model

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

we get a much better fit:

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

Call: glm(formula = s/n ~ sex + log(dose), family = binomial,
+            weights = n)
Deviance Residuals:
       Min         1Q      Median        3Q      Max
 -1.105398 -0.6534319 -0.02224909 0.4847064 1.429444

Coefficients:
                Value Std. Error   t value
(Intercept) -2.372412  0.3855067 -6.154009
        sex -1.100743  0.3558238 -3.093507
  log(dose)  1.535336  0.1891019  8.119093

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 124.8756 on 11 degrees of freedom

Residual Deviance: 6.757064 on 9 degrees of freedom

Number of Fisher Scoring Iterations: 5

Correlation of Coefficients:
          (Intercept)       sex
      sex -0.2031088
log(dose) -0.7819640  -0.2785742
```

The deviance 6.757064 is now much smaller, indicating a better fit. The p-value is

```
> 1-pchisq(6.757064,9)
[1] 0.6623958
```

so the model using log(dose) fits well.

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. The fitted lines are of the form

$$\text{logit } \pi = \hat{\beta}_0 + \hat{\beta}_2 \log(\text{dose})$$

for males, and

$$\text{logit } \pi = (\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 \log(\text{dose})$$

for females. To draw the plot, using M's for males and F's for females, and draw in the fitted lines, we type

```
> 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))
```

The result, shown in Figure 5.3, shows that the model fits very well. Note that when calculating the logits of the observed proportions, we have added a "fudge factor" of 0.5. This is to avoid the difficulty of computing the log of zero.

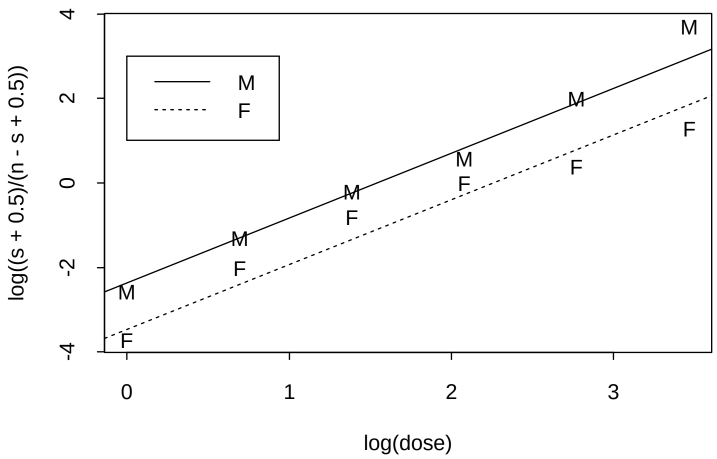

Figure 5.3: Fitted lines for the insect data.

Example 9. In the ingot example, we can test the adequacy of the logistic model by following the procedure above. We can to get the deviance of the logistic model using the `deviance` function:

```
> logistic.dev<-deviance(ingots.glm)
> logistic.dev
[1] 13.75263
```

To test the adequacy of the logistic model, we work out the area under the χ^2_{m-k-1} density. The degrees of freedom $m - k - 1$ is computed by extracting the `df.residual` component from the object `ingots.glm`:

```
> ingots.glm$df.residual
[1] 16
```

These values are also part of the output produced by the `summary` function when applied to a "glm object". To get the p-value , type

```
> 1-pchisq(logistic.dev,16)
[1] 0.6171366
```

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

The "sparse" case

What happens when $n_i = 1$ (i.e. every case has a distinct covariate vector)? Then the conditions for the asymptotics do not hold, so we can no longer interpret the deviance as a measure of goodness of fit. However, it is still useful as it allows us to compare models. Another property of the deviance when $n_i = 1$ for each i is that it depends entirely on $\hat{\beta}$ — it is a function of $\hat{\beta}$ alone, and as such cannot be a goodness of fit measure.

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.

Standard likelihood theory (see e.g. 528.381) suggests that if the submodel is adequate, the difference

$$(-2\ell_{\text{SUB}}) - (-2\ell_{\text{FULL}}) \tag{5.5}$$

will have a distribution that is approximately χ^2_d, where $-2\ell_{\text{SUB}}$ and $-2\ell_{\text{FULL}}$ are the minimum values of -2ℓ from both the submodel and full model respectively, and d is the number of variables dropped.

The difference (5.5) can be expressed as a difference of deviances: let $-2\ell_{\text{SAT}}$ be the -2ℓ of the saturated model. Then

$$\begin{aligned}(-2\ell_{\text{SUB}}) - (-2\ell_{\text{FULL}}) &= [(-2\ell_{\text{SUB}}) - (-2\ell_{\text{SAT}})] \\ &\quad -[(-2\ell_{\text{FULL}}) - (-2\ell_{\text{SAT}})] &(5.6)\\ &= \text{deviance of submodel} &(5.7)\\ &\quad - \text{deviance of full model.} &(5.8)\end{aligned}$$

The difference is 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 dropping the extra terms will not increase the deviance by very much, so the variables can be dropped in the interests of getting a simpler model. Thus a *small* increase in deviance means we can drop the d variables from the model.

The difference in the deviance has approximately a χ^2_d distribution if the submodel is adequate. Unlike the approximation to the deviance, the χ^2 approximation to the **difference** in deviances is usually quite accurate. In particular it will be good if m is larger (we don't need the n_i's large). It follows that we can test the adequacy of the submodel by the following procedure:

- Calculate the deviance of the full model.
- Calculate the deviance of the submodel.
- Calculate the difference (5.8).
- Calculate a p-value by finding the area to the right of (5.8) under the χ^2_d density.

A small p-value will be evidence *against* the submodel - i.e. a small p-value indicates that we cannot drop all the d variables from the model.

Example 10. In the kyphosis example, suppose we want to know if we can drop both the variables `start` and `number`. We fit the submodel with `age` and `agesq` only using the call

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

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

Thus the submodel deviance is 72.73858, so the difference between the submodel and full model deviances is $72.73858 - 54.42776 = 18.31082$ on 2 degrees of freedom, which is highly significant, with a p-value of 0.0001. The model without the surgical variables `start` and `number` is clearly inadequate.

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 linear models, we test the overall significance of the regression by testing if all regression coefficients (except the constant term) are zero. This test involves comparing the Regression sum of squares with the total sum of squares. Now the total sum of squares is just the residual SS from fitting the "null model" $y_i = \beta_0 + \epsilon_i$, so the numerator of the test involves

$$\text{Residual SS from the null model} \; - \; \text{Residual SS from the full model.}$$

The quantity

$$(\text{deviance for the null model} \; - \; \text{deviance for the full model})$$

asymptotically plays the same role in logistic regression.

To evaluate this quantity, note that the deviance for the null model is obtained by minimising

$$-2\sum_{i=1}^{n} y_i\beta_0 + 2n\log(1+e^{\beta_0}).$$

Differentiating with respect to β_0 and equating to 0 gives

$$\overline{y} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

so that value of β_0 that minimises the -2 log likelihood is

$$\hat{\beta_0} = \log\left(\frac{n_1}{n-n_1}\right)$$

where n_1 is the number of cases for which $Y = 1$. For this value of β_0 the -2 log likelihood is

$$\begin{aligned} & -2n_1\log\left(\frac{n_1}{n-n_1}\right) - 2n\log\left(\frac{n-n_1}{n}\right) \\ = \;& -2(n_1\log n_1 + n_0\log n_0 - n\log n) \qquad (5.9) \end{aligned}$$

where $n_0 = n - n_1$. The deviance of this null model (i.e. the difference between (5.9) and $-2\ell_{SAT}$ for the saturated model) is called the "null deviance" and is printed out as part of the summary output. We can do a test for $\beta_1 = \cdots = \beta_k = 0$ by comparing the difference between the model deviance and the null deviance to a χ^2_k distribution.

Example 11. For the CHD data $n_0 = 57$ and $n_1 = 43$ so that the deviance for the null model is 136.663.

In practice, we don't need to do the calculation by hand, as the value of the null deviance is included in the output from `summary`. The null deviance is also included in the "glm object", and may be extracted with the $ operator:

```
> null.dev <- chd.glm$null.deviance
> null.dev
[1] 136.663
```

The full model deviance (i.e. the residual deviance) is 107.68059 so the difference between the full and null model deviances is $136.663 - 107.68059 = 28.98$.

These calculations are also done by the `anova` function:

```
> anova(chd.glm)
Analysis of Deviance Table

Binomial model

Response: y

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                    99   136.6630
 age  1  28.9824        98   107.6806
```

The difference in the deviances is distributed as χ^2_1 under the null hypothesis that the null model is true, giving us another test of $\beta = 0$. The p-value is

```
> 1-pchisq(28.9824,1)
[1] 7.303889e-08
```

This p-value is strong evidence that there is a relationship between `age` and the occurence of CHD. Note that you get the degrees of freedom for the χ^2 by subtracting logistic model df (98) from the null model df (99).

Example 12. In the kyphosis example, the deviance of the logistic model is 54.42776, and that of the null model is 83.23447. The difference in the deviances for the logistic and the null models is thus $83.23447 - 54.42776 = 28.80671$, with 80-76 =4 degrees of freedom. The p-value is about 10^{-5}, indicating a significant regression:

```
> 1-pchisq(28.80671,4)
[1] 8.556925e-06
```

More generally, the `anova` function allows us to examine the changes in the deviance as terms are added to the model, and is a convenient way to decide which terms should be retained:

Example 13. To see the effect on the deviance as more terms are added to the model in the kyphosis example, type

```
> kypho.aov<-anova(kypho.glm)
>  kypho.aov
Analysis of Deviance Table

Binomial model
```

```
Response: kypho

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
    NULL                   80   83.23447
     age  1 1.301985       79   81.93249
I(age^2)  1 9.193899       78   72.73859
  number  1 8.875728       77   63.86286
   start  1 9.435098       76   54.42776
```

We can judge the significance of these changes:

```
> 1-pchisq(kypho.aov$Deviance, kypho.aov$Df)
[1]          NA 0.253851014 0.002428231 0.002889869 0.002128717
```

This gives the p-value for each reduction, and works because `kypho.aov` has components `Deviance` and `Df` that are extracted in the usual way. Adding each term (except `age`) improves the model significantly. This process is similar to "forward selection" in multiple regression.

The function `anova` can also be used to compare models. As an illustration of this, we can do the calculations in Example !!!submodel kyphosis example!!! by typing (`sub.glm` is the "glm object" storing the results of fitting the submodel without `start` and `number`)

```
> anova(sub.glm,kypho.glm)
Analysis of Deviance Table

Response: kypho

                             Terms Resid. Df Resid. Dev          Test
1                  age + I(age^2)        78   72.73858
2 age + I(age^2) + number + start        76   54.42776 +number+start
  Df Deviance
1
2  2 18.31082
> 1-pchisq(18.31082,2)
[1] 0.0001056467
```

5.2.5 Residuals in logistic regression

Assume that there are m distinct covariate vectors, and n_i observations are taken at covariate vector x_i, resulting in y_i successes. Let $\hat{\beta}_0$, $\hat{\beta}_1$, , . . . , $\hat{\beta}_k$ be the ML estimates of the regression coefficients. We may define two types of residual:

1. Pearson residuals, defined by

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

   where $\text{logit}(\hat{\pi}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$. These are similar to linear model standardised residuals.

2. Deviance residuals. Let $\hat{\pi}_i$ and $\tilde{\pi}_i$ be the estimated success probabilities for the ith covariate pattern, using the logistic and saturated models respectively. Then it can be shown that the deviance of the logistic model can be written

$$\text{Deviance} = \sum_{i=1}^{m} d_i^2$$

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 is positive if r_i is positive and negative if r_i is negative. (The quantity in the braces $\{\ \}$ is always positive). The quantities d_i are called *deviance residuals.*

Note that the sum of squares of the Pearson residuals is

$$X = \sum_{i=1}^{m} \frac{(y_i - n_i\hat{\pi}_i)^2}{\text{Var}(n_i\hat{\pi}_i)}.$$

This statistic, called a generalised χ^2 statistic, can be used for testing the null hypothesis that the logistic function is adequate. As noted in Section 5.2.4, it is asymptotically equivalent to the test based on the deviance. Thus both the sum of squares of the Pearson residuals and the sum of squares of the deviance residuals are "goodness of fit statistics". This justifies the use of the term "residuals". Both sums of squares have the same distribution (χ^2_{m-k-1}) asymptotically under the null hypothesis that the logistic model is adequate.

Large values of r_i or d_i denote covariate patterns that are poorly fitted by the model. Large values of d_i indicate observations that have undue influence on the log likelihood. Note that the Pearson residuals are only useful for grouped data, while the deviance residuals are useful for ungrouped, as well as grouped data.

Both types of residual are calculated as part of the "glm object", and are extracted with the `residuals` function.

The residual analysis for the Pearson residuals proceeds as for ordinary regression. We draw normal plots, and plot residuals versus fitted values. Fitted values are the estimated probabilities $\hat{\pi}_i$, not the logits of these probabilities, and are extracted from the "glm object" usingthe function `fitted.values`.

In the case of deviance residuals, we are interested in which cases make disproportionately large contributions to the the deviance. Such cases may be identified by normal plots and by plotting the residuals against case number. Plots of residuals versus fitted values are not useful, since the deviance residuals have a strong functional relationship with the fitted values.

Example 14. For the ingot data, we calculate the residuals by typing

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

We can plot the residuals against index (case number) to identify covariate patterns that are poorly fitted:

```
> plot(dev.resids,type="l")
```

Note if we give `plot` only one argument, we get an index plot. From the resulting plot (not shown), we see that the 7th and 10th covariate patterns are not well fitted, and have big 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
 7   27  1.7        4    44
10   14  2.2        2    33
```

Influential points

We will also be interested in detecting influential points, for example by computing "leave one out" diagnostics. Since the IRLS method of fitting essentially proceeds by performing repeated regressions, we can adapt regression influence techniques to logistic regression. It can be shown that the effect of deleting all observations having covariate vector x_i from the data will change the deviance by an amount approximately equal to

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

and the χ^2 statistic by approximately

$$\Delta\chi_i^2 = \frac{r_i^2}{1-h_{ii}}$$

where h_{ii} are the hat matrix diagonals from the last regression in the IRLS iterations.

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 15. Consider the following example, taken from Pregibon, Annals of Statistics (1981), p705. The data is from Finney, Biometrika (1947) p320. See also the 1989 final exam. 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. The nature of the measurement process was such that only the occurrence or nonoccurrence 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. The model fitted is

$$\operatorname{logit}(\pi) = \beta_0 + \beta_1 \log(RATE) + \beta_2 \log(VOLUME).$$

The estimated coefficients and their standard errors are

$$\begin{aligned}\hat\beta_0 &= -9.5296\ (3.2331)\\ \hat\beta_1 &= 3.8822\ (1.4286)\\ \hat\beta_2 &= 2.6491\ (0.9142)\end{aligned}$$

Suppose the data are in a data frame `vaso`, with variables `log.volume`, `log.rate` and `response`. A useful plot of the data can be made by plotting the explanatory variables against each other , indicating the value of the response by a 0 or 1.

```
> attach(vaso)
> plot(log.volume,log.rate,type="n")
> text(log.volume,log.rate,response)
```

Table 5.5: Listing of Finney's data on vaso-constriction in the skin of the digits. The binary response indicates the occurrence (1) or nonoccurrence (0) of vaso-constriction.

Volume	Rate	Response	Volume	Rate	Response
3.7	.825	1	1.8	1.8	1
3.5	1.09	1	.4	2	0
1.25	2.5	1	.95	1.36	0
.75	1.5	1	1.35	1.35	0
.8	3.2	1	1.5	1.36	0
.7	3.5	1	1.6	1.78	1
.6	.75	0	.6	1.5	0
1.1	1.7	0	1.8	1.5	1
.9	.75	0	.95	1.9	0
.9	.45	0	1.9	.95	1
.8	.57	0	1.6	.4	0
.55	2.75	0	2.7	.75	1
0.6	3.	0	2.35	.03	0
1.4	2.33	1	1.1	1.83	0
.75	3.75	1	1.1	2.2	1
2.3	1.64	1	1.2	2.0	1
3.2	1.6	1	.8	3.33	1
.85	1.415	1	.95	1.9	0
1.7	1.06	0	.75	1.9	0
			1.3	1.625	1

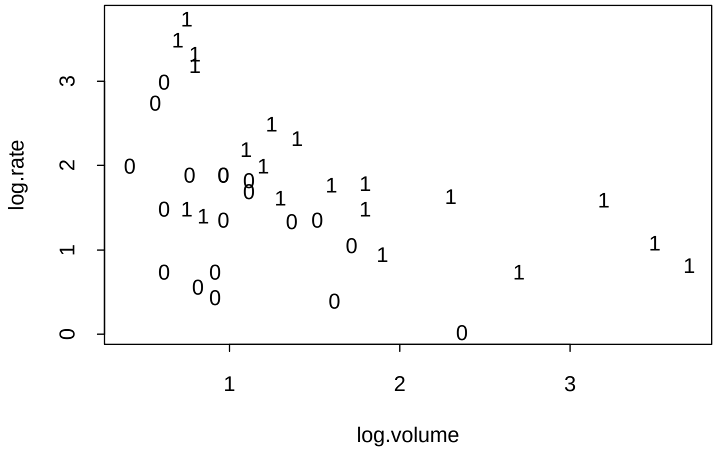

Figure 5.4: Plot of the vaso-constriction data.

The plot is shown in Figure 5.4.

The result of fitting the logistic model is

```
> vaso.glm<-glm(response~log.volume+log.rate,data=vaso,family="binomial")
> summary(vaso.glm)

Call: glm(formula = response ~ log.volume + log.rate,
             family = "binomial", data = vaso)
Deviance Residuals:
       Min         1Q    Median        3Q      Max
 -1.506566 -0.7346586 0.0399776 0.4885521 2.329314

Coefficients:
                Value Std. Error   t value
(Intercept) -9.529259  3.2139903 -2.964931
 log.volume  3.882007  1.4202335  2.733358
   log.rate  2.649036  0.9095385  2.912506

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 54.03984 on 38 degrees of freedom

Residual Deviance: 29.7723 on 36 degrees of freedom

Number of Fisher Scoring Iterations: 5

(NB: Correlation matrix not shown to conserve space.)
```

Note that the plot in Figure 5.4 indicates a couple of possibly outlying points. We will calculate the deviance residuals and produce some influence plots. The code to extract the deviance residuals and Pearson residuals, and calculate the differences in deviance due to the one point deletions is

```
> dev.resids<-residuals(vaso.glm)  # get deviance residuals
> pearson.resids<-residuals(vaso.glm,type="p") # note abbreviation "p"
> hats<-lm.influence(vaso.glm)$hat  # hat matrix diagonals
> # Calculate leave-one-out deviance differences
> del.dev <- dev.resids^2 + pearson.resids^2*hats/(1-hats)
```

To draw index plots, and label points, we type

```
>  # draw index plot of deviance residuals
> plot(dev.resids,type="l",main="Index plot of deviance residuals")
> # now label points
> n<-length(dev.resids)
> text(1:n,dev.resids*1.02,1:n)
> # draw index plot of deviance changes
> plot(del.dev,type="l",main="Index plot of deviance changes")
> # now label points
> text(seq(n),del.dev*1.02,seq(n))
```

The resulting plots shown in figure 5.5. (the plot of $\Delta\chi_i^2$ versus i is similar to the deviance difference plot and has not been drawn.) These plots show that two observations, the 7th and 35th, are not well fitted by the model, and moreover have a considerable influence on

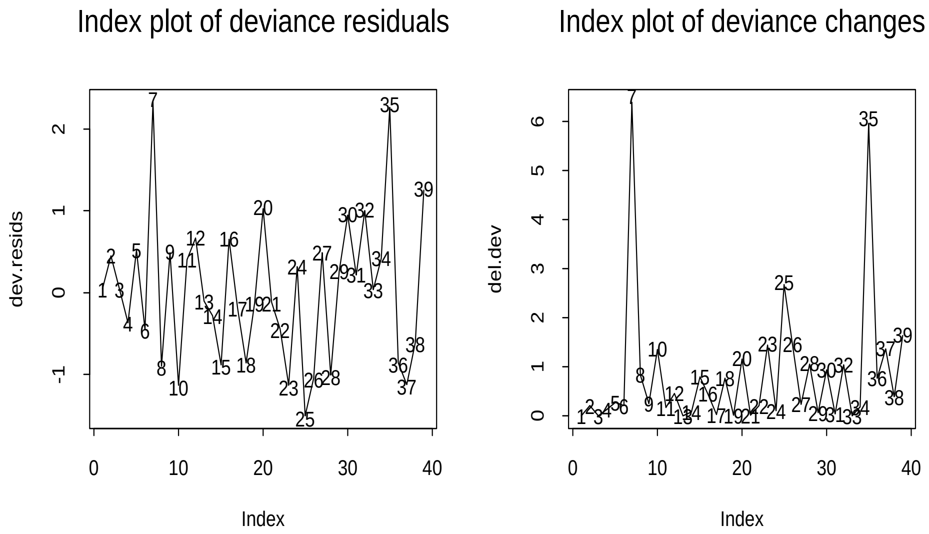

Figure 5.5: Influence plots for the vaso-constriction data.

the fit. These two points correspond to the two outlying positive responses in the plot of rate versus volume in Figure 5.4. The effect of deleting these points can be determined by refitting:

```
> new.glm<-glm(response~log.volume+log.rate,
+                     data=vaso[-c(7,35),],family="binomial")
> summary(new.glm)

Call: glm(formula = response ~ log.volume + log.rate,

family = "binomial", data = vaso[ - c(7, 35),  ])
Deviance Residuals:
       Min        1Q        Median         3Q      Max
 -1.503683 -0.118568 -1.155159e-05 0.02885105 1.970363

Coefficients:
                Value Std. Error   t value
(Intercept) -41.98278  21.877136 -1.919025
 log.volume  17.49200   9.257570  1.889481
   log.rate  10.74274   5.591089  1.921405

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 51.26586 on 36 degrees of freedom
Residual Deviance: 10.69979 on 34 degrees of freedom
```

Note the large changes in the coefficients. Although the t-values (Wald tests) are rather small, the analysis of deviance suggests both variables are required:

```
> anova(new.glm)
```

```
Analysis of Deviance Table

Binomial model

Response: response

Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev
      NULL                    36   51.26586
log.volume  1  8.75000        35   42.51586
  log.rate  1 31.81607        34   10.69979
> 1-pchisq(8.75000,1)
[1] 0.003096021
> 1-pchisq(31.81607,1)
[1] 1.694846e-08
```

As these observations are not really associated with extreme values in the rate-volume plane, their effect on the fit might presumably be small. However, from the "leave one out analysis", we have reason to believe otherwise.

5.2.6 Binary anova

The logistic regression method can be easily extended to the case where the explanatory variables are factors (i.e. "binary ANOVA") or a mixture of factors and continuous variables (i.e. "binary analysis of covariance").

For example, suppose we have a two-way ANOVA with several observations per cell, but the response is binary. Let A and B denote the two factors, having I and J levels respectively, and suppose that n_{ij} binary responses $Y_{ijk},\ k = 1, 2, \ldots, n_{ij}$ are taken when A is at level i and B is at level j. Let $\pi_{ij} = \Pr[Y_{ijk} = 1]$, let $\mu_{ij} = \mathrm{logit}(\pi_{ij})$ and suppose that the number of "successes" out of the n_{ij} binary observations in the i, j cell is s_{ij}. As in ordinary ANOVA, we can decompose μ_{ij} into a grand mean, main effects and interactions:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}. \tag{5.10}$$

Because of the usual constraints on the parameters (e.g.$\sum_i \alpha_i = 0$) there will then be $I - 1$ functionally independent α_i's, $J - 1$ functionally independent β_j's and $(I - 1)(J - 1)$ functionally independent $\alpha\beta_{ij}$'s. These IJ parameters can be estimated by ML just as in the case of logistic regression.

Note that the model above puts no restriction at all on the quantities π_{ij}, other than they must be probabilities, and hence no restrictions on the μ_{ij}'s. The model is therefore a "saturated model", with as many parameters as there are covariate patterns. Under these circumstances, the parameters π_{ij} are estimated by

$$\hat{\pi}_{ij} = \frac{s_{ij}}{n_{ij}}$$

i.e. the proportion of the n_{ij} cases having factor level combination (i, j) that have a "1" response. Note that in the present ANOVA example, "factor level combinations" are equivalent to "covariate vectors".

The ML estimate of μ_{ij} is $\mathrm{logit}(\hat{\pi}_{ij})$, and the ML estimates of the α_i's etc are given by e.g.

$$\hat{\alpha}_i = \overline{\hat{\mu}}_{i.} - \overline{\hat{\mu}}_{..}.$$

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

Hypothesis testing

We can test for zero interaction in our model by examining the difference in deviance when we add interaction terms to the model. We fit the submodel

$$\mu_{ij} = \mu + \alpha_i + \beta_j \tag{5.11}$$

derived from (5.10) by setting the interactions equal to zero. The ML estimates of the parameters will be different from those obtained above, since -2ℓ is now being minimised subject to the constraints

$$\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} = 0.$$

The estimates for the submodel no longer have a simple form, but can be easily calculated by `glm`, as can the deviance for the submodel. To test the zero interaction hypothesis, we compare the saturated model to the (additive) submodel by examining the deviance of the submodel. The degrees of freedom are $IJ - ((I-1)+(J-1)+1) = (I-1)(J-1)$. Similar statistics can be used to test the hypothesis of zero main effects, and are also reported in the ANOVA table. In practice, the calculations are done by `glm` and `anova`.

Example 16. 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 and the treatment (Control, X-ray, Beta-ray) with three.

To fit the model, we need to read in the data, create the factors, and use `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)))
> treatment<-rep(c("Control","X-ray","Beta-ray"),4)
```

```
> # We don't want levels in alphabetical order
> treatment<-factor(treatment,levels=unique(treatment))

> # make a data frame
> grass<-data.frame(experiment,treatment,yes,total)
> grass
   experiment treatment yes total
 1          1   Control  12    22
 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


# now fit model
> grass.glm<-glm(yes/total~experiment*treatment,weight=total,
+                                     family="binomial", data=grass)
> summary(grass.glm)

Call: glm(formula = yes/total ~ experiment * treatment,
                  family = "binomial", data = grass, weights = total)

Coefficients:
                            Value Std. Error    t value
          (Intercept) -0.42975005  0.1660175 -2.5885826
          experiment1 -0.41215949  0.2801566 -1.4711752
          experiment2  0.57702764  0.2866136  2.0132598
          experiment3 -0.75842071  0.2901894 -2.6135372
           treatment1  1.34529057  0.2359336  5.7019881
           treatment2 -0.70924493  0.2306341 -3.0751960
experiment1treatment1 -0.32105947  0.3703325 -0.8669493
experiment2treatment1  0.04787689  0.4215573  0.1135715
experiment3treatment1 -0.35779050  0.3820769 -0.9364359
experiment1treatment2 -0.05828343  0.4132610 -0.1410330
experiment2treatment2 -0.13117983  0.3972123 -0.3302512
experiment3treatment2  0.67364026  0.3911232  1.7223223

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 54.8375 on 11 degrees of freedom

Residual Deviance: 0 on 0 degrees of freedom

Number of Fisher Scoring Iterations: 5

(NB Correlation matrix not shown to save space)
```

```
> anova(grass.glm)
Analysis of Deviance Table

Binomial model

Response: yes/total

Terms added sequentially (first to last)
                    Df Deviance Resid. Df Resid. Dev
               NULL                    11   54.83750
         experiment  3 11.66240         8   43.17511
          treatment  2 38.21197         6    4.96314
experiment:treatment  6  4.96314         0    0.00000
```

The difference in deviance between the models with and without interactions is not significant:

```
> 1-pchisq(4.96314,6)
[1] 0.5485493
```

As in ordinary ANOVA, a useful graphic display to complement the test for zero interactions is to plot μ_{ij} versus i separately for each j. We join up the points of the separate plots by lines. Parallel "profiles" indicate zero interaction. The Splus 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 5.6. 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.

Estimating contrasts

Contrasts in the parameters are estimated in the usual way. Thus, a contrast of the form $\sum_{i=1}^{I} c_i\alpha_i$ in the A main effects is estimated by $\sum_{i=1}^{I} c_i\hat{\alpha}_i$. We can calculate the standard error of a contrast estimate by means of the formula

$$\text{s.e.}\left(\sum_i c_i\hat{\alpha}_i\right) = \sqrt{\sum_i\sum_{i'} c_i c_{i'}\,\text{cov}(\hat{\alpha}_i, \hat{\alpha}_{i'})}.$$

We can estimate a $100\%(1-\alpha)$ confidence interval for $\sum_i c_i\alpha_i$ using the formula

$$\sum_i c_i\hat{\alpha}_i \pm \text{s.e.}\left(\sum_i c_i\hat{\alpha}_i\right)\; z\left(\frac{\alpha}{2}\right).$$

Example 17. Calculating these quantities in Splus is relatively simple. The estimated coefficients (the estimates of α_i, β_j and $\alpha\beta_{ij}$ are in the "glm object" and can be extracted with

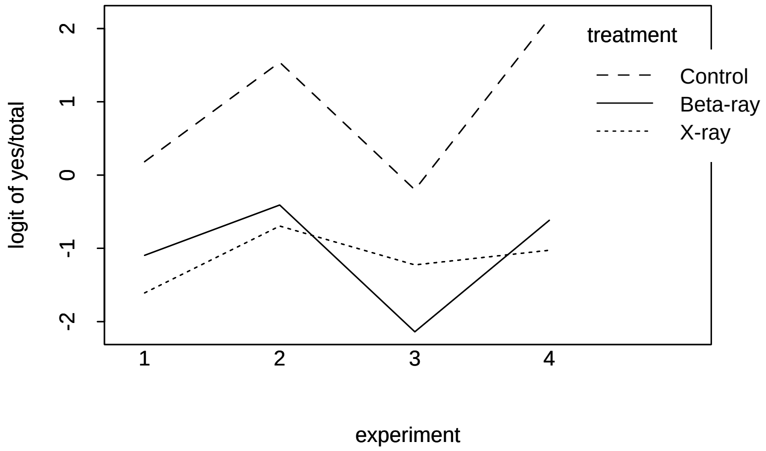

Figure 5.6: Interaction plot for the grasshopper data.

the `coefficients` function. The covariance matrix is computed in the `summary` function, and is extracted using `$cov.unscaled`. We illustrate using the grasshopper data. Since there is no evidence of interaction, we will fit and use an additive model.

```
additive.glm<-glm(yes/total~experiment+treatment,weight=total,
+                              family="binomial", data=grass)
estimates<-coefficients(additive.glm)
cov.mat<-summary(additive.glm)$cov.unscaled
```

Suppose we want to compute a confidence interval for $\beta_2 - \beta_1$, the contrast that measures the effect of the X-ray treatment compared to the control treatment. The order of the estimates is

```
> estimates
 (Intercept) experiment1 experiment2 experiment3 treatment1
  -0.4775218  -0.4248361   0.6145279  -0.7200159     1.2687
 treatment2
 -0.6888909
```

so the coefficients of the desired contrast are 0,0,0,0,-1,1. The estimate is

```
> contr.coeffs<-c(0,0,0,0,-1,1)
> contr.est<- sum(contr.coeffs*estimates)
> contr.est
[1] -1.957591
```

The standard error is

```
> std.err<-sqrt(sum(contr.coeffs*(cov.mat%*%contr.coeffs)))
> std.err
[1] 0.3820636
```

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

A 95% confidence for the contrast is $1.957591 \pm 1.96 \times 0.3820636$, i.e. $[-2.706, -1.209]$.

5.3 Analysis of contingency tables

Suppose we have k categorical variables and measure these variables on each of n cases. Suppose that the k variables have I_1, I_2,... I_k categories, and that we classify the n cases into the $I_1 \times I_2 \times \ldots \times I_k$ possible category combinations or *cells*. The resulting k-way table is called a *contingency table*. We can use the terms factor and level interchangeably with variable and category.

Example 18. In our first example, $k = 1$. Table 5.7 contains data on U.S. deaths by falling (1970 data). Each death is classified by a single factor, the month of occurence.

Example 19. For an example with two factors, consider the data in Table 5.8, where 655 students chosen at random from the student body at Auckland University are classified according to their degree course and socio-economic status (SES).

Example 20. For an example with three factors, consider the data in Table 5.9, where 123 patients suffering from diabetes are classified on the basis of three criteria.

The type of model considered for this type of data is different from those discussed so far in that a division of the variables into responses and explanatory variables is not made. Rather, interest focusses on the *joint* distribution of the variables, rather than the conditional distribution of one, given the others.

Table 5.8: A. U. students classified by degree course 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

Table 5.9: Diabetes patients classified on 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

Denote by m the number of "cells" in the contingency table, and suppose that there are n cases to classify. Thus in Example 10, there are 12 cells and 16986 cases, in Example 11 there are $7 \times 6 = 42$ cells and 655 cases, while finally in Example 12 there are $2 \times 2 \times 2 = 8$ cells and 123 cases. Typically, the cases will be respondents in a survey, and the cells will correspond to all possible ways of responding to all or part of the questionnaire.

Suppose that the probability that an individual chosen at random falls in the ith cell is π_i, $i = 1, \ldots, m$. We assume that the categories are defined in such a way that each case is classified into exactly one cell, so that

$$\pi_1 + \cdots + \pi_m = 1.$$

Further suppose that the number of cases that are classified into the cells are y_i, $i = 1, \ldots, m$.

If the sampling is random (i.e. every sample of size n from the population has the same probability of selection), then the joint distribution of the quantities y_i, $i = 1, \ldots, m$ is given by the multivariate hypergeometric distribution. This is mathematically intractable, so the distribution of the Y_i's is approximated by the multinomial distribution. Provided the sample is but a small fraction of the total population, this approximation should be adequate. Thus the probability that after the classification there are y_1, y_2,... cases in categories 1, 2,... is

$$\frac{n!}{y_1! y_2! \ldots y_m!} \pi_1^{y_1} \ldots \pi_m^{y_m}.$$

Since each case goes in exactly one category we must have $\sum_i y_i = n$.

In this section we discuss how to estimate the parameters π_i, both in the simple situation above and also when the $\pi's$ are subject to various constraints that correspond for example to certain forms of independence between the factors involved. We will also discuss how various hypotheses (again related to various types of independence between the factors) may be expressed and tested with both likelihood ratio tests and Pearson χ^2 tests.

5.3.1 Likelihood based inference for contingency tables

We begin by studying the likelihood function for the multinomial model above. The likelihood is just the joint probability function, regarded as a function of the π's:

$$l(\pi_1, \ldots, \pi_m) = \frac{n!}{y_1! y_2! \ldots y_m!} \pi_1^{y_1} \ldots \pi_m^{y_m}.$$

The -2 log likelihood is thus up to a constant,

$$-2 \log l = -2 \sum_{i=1}^{m} y_i \log \pi_i.$$

The maximum likelihood estimates of the π's are the values $\hat{\pi}_i$ that minimise this quantity, subject to the constraint $\sum_i \hat{\pi}_i = 1$. By using Lagrange multipliers, or some other suitable method, it can be shown that

$$\hat{\pi}_i = \frac{y_i}{n}, \quad i = 1, \ldots, m.$$

The minimum value of the -2 log likelihood is thus

$$-2\sum_{i=1}^{m} y_i \log \frac{y_i}{n} = -2\sum_{i=1}^{m} y_i \log y_i + 2n \log n.$$

This model, with the π's unconstrained (except by the requirement that they sum to unity) is called the *saturated model* or the *full model.* We will denote the -2ℓ of the saturated model by $-2\ell_{SAT}$.

To test the hypothesis that the π's satisfy some additional constraint, we compute the minimum value of the -2 log likelihood subject to the additional constraint. Call the minimum value $-2\ell_C$. Then if the constraint is in fact satisfied, statistical theory tells us that the distribution of the deviance

$$(-2\ell_C) - (-2\ell_{SAT})$$

has approximately a χ^2 distribution with $m-1-c$ degrees of freedom, where c is the number of free parameters in the constrained model. We can compute a p-value for the hypothesis by calculating the area under the χ^2_{m-1-c} density to the right of the observed deviance. Some examples of the constraints encountered in practice are given below.

Completely specified probabilities.

Suppose the hypothesis to be investigated is that the probabilities are completely specified, i.e. $\pi_i = \pi_{i0}$, $i = 1, \ldots m$ for given probabilities π_{i0}. Then there are no free parameters, so $c = 0$. The deviance under the null hypothesis is

$$-2\ell_C = -2\sum_{i=1}^{m} y_i \log \pi_{i0}.$$

Example 21. Suppose in the death by falling example, we want to test if all months are equally likely, i.e. we want to test if $\pi_i = \frac{1}{12}$. Then

$$-2\ell_{SAT} = -2\sum_{i=1}^{m} y_i \log \frac{y_i}{n} = 84337.44$$

and

$$-2\ell_C = -2\sum_{i=1}^{m} y_i \log \left(\frac{1}{12}\right) = 84417.25$$

so that the deviance is $84417.25 - 84337.44 = 79.81$ The p-value (11=12-1-0 degrees of freedom) is 0.000, so that the data is not consistent with the hypothesis of equal probabilities. There are too many falls in the winter months.

The calculations above can very easily be done in Splus:

```
> y<-c(1688,1407,1370,1309,1341,1388,1406,1446,1332,1363,1410,1526)
```

```
> n<-sum(y)
> min.sat<- -2*sum(y*log(y/n))
> min.sat
[1] 84337.44
> min.constrained<- -2*sum(y*log(1/12))
 min.constrained
[1] 84417.25
> dev<- min.constrained - min.sat
> dev
[1] 79.80975
> 1-pchisq(dev,11)
[1] 1.60616e-12
```

Testing for independence in two-way tables.

For two-way tables it is convenient to use a slightly different system of indexing: we let y_{ij} denote the frequency count in the i, j cell of the table (i.e the cell in row i and column j), so that y_{ij} is the number of cases in category i for the "row" factor and category j for the "column" factor. We will suppose the row and column factors have I and J levels respectively. We let the corresponding probability be π_{ij}, so that

$$\begin{aligned}\pi_{ij} &= \text{Pr[A randomly chosen respondent is in category } i \text{ of the row factor} \\ &\qquad\qquad \text{and category } j \text{ of the column factor].}\end{aligned}$$

If the factors are independent, then in fact

$$\begin{aligned}\pi_{ij} &= \text{Pr[Respondent is in category } i \text{ of the row factor]} \\ &\qquad \times \text{Pr [Respondent is in category } j \text{ of the column factor]} \\ &= \pi_{i+}\pi_{+j}\end{aligned}$$

say, where π_{i+} and π_{+j} and the marginal probabilities of the two factors. The notation π_{i+} is justified by the fact that

$$\begin{aligned}\pi_{i+} &= \text{Pr[Respondent is in category } i \text{ of the row factor]} \\ &= \sum_{j=1}^{J} \text{Pr[Respondent is in category } i \text{ of the row factor} \\ &\qquad\qquad \text{and category } j \text{ of the column factor]} \\ &= \sum_{j=1}^{J} \pi_{ij}.\end{aligned}$$

Similarly $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$.

Under the hypothesis of independence, $-2\ell_C$ is the minimum of the -2 log likelihood, under the constraint $\pi_{ij} = \pi_{i+}\pi_{+j}$ for $i = 1, 2, \ldots I$ and $j = 1, 2, \ldots J$. It can be shown that the minimum occurs when $\pi_{ij} = y_{i+}y_{+j}/n^2$ where n is the sum of the cell frequencies, $y_{i+} = \sum_{j=1}^{J} y_{ij}$ and $y_{+j} = \sum_{i=1}^{I} y_{ij}$, and hence

$$-2\ell_C = -2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{i+}y_{+j}}{n^2}\right).$$

The number of free parameters under the constraint is $c = I + J - 2$, since $\sum_{i=1}^{I} \pi_{i+} = 1$ and $\sum_{j=1}^{J} \pi_{+j} = 1$. The minimum of -2ℓ under the saturated model is

$$-2\ell_{SAT} = -2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log \frac{y_{ij}}{n},$$

so the deviance is the difference of these. Under the null hypothesis of independence, the deviance is distributed as χ^2 with $m - 1 - c = IJ - 1 - (I + J - 2) = (I - 1)(J - 1)$ degrees of freedom.

Example 22. For the student data, $-2\ell_{SAT} = 4323.102$, $-2\ell_C = 4372.677$ so the deviance is 49.57 with $(7 - 1) \times (6 - 1) = 30$ degrees of freedom. The p-value is 0.0137 so the null hypothesis of independence is rejected.

5.3.2 Chi-square tests

An alternative to the likelihood ratio tests discussed above is provided by the *chi-square test*. Suppose as before that we want to test if the probabilities are subject to some constraint, such as independence in a two-way table. Again as before, we assume that there are n cases to classify into m cells, the observed frequencies are $y_1, \ldots, y_m$, and the classification probabilities are $\pi_1, \ldots, \pi_m$. The *expected* number of cases that will be classified into the ith cell is $n\pi_i$. Let $\hat{\pi}_i$ be the estimate of π_i under the constraint, i.e. the value of π_i that minimises the -2 log likelihood subject to the constraint. Under the hypothesis, our estimate of the expected number in the ith cell is thus $n\hat{\pi}_i$. An obvious way to assess the plausibility of the hypothesis is to compare the expected frequencies $n\hat{\pi}_i$ with the observed frequencies y_i. A suitable statistic for this purpose is the χ^2 statistic

$$X = \sum_{i=1}^{m} \frac{(y_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}$$

which asymptotically has the same distribution under the null hypothesis as the likelihood ratio statistic. p-values are calculated in the same manner. The χ^2 approximation is usually quite good as long as all cells have expected frequencies exceeding 1.

The statistic X usually has very similar values to the deviance. The p-value is calculated in the same way as before. We illustrate the method in the two cases considered above.

Completely specified probabilities

In this case the expected frequencies under the null hypothesis are $n\pi_{i0}$ so the χ^2 statistic is

$$X = \sum_{i=1}^{m} \frac{(y_i - n\pi_{i0})^2}{n\pi_{i0}}$$

Example 23. In the deaths by falling data, the value of X is 90.29 and the p-value is 0.000, in agreement with the previous method.

Table 5.10: 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

Independence in two-way tables

Again we revert to our double index notation. Under the independence assumption, the estimate of π_{ij} is

$$\hat{\pi}_{ij} = \frac{y_{i+}y_{+j}}{n^2}.$$

The χ^2 statistic is

$$X = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(y_{ij} - \frac{y_{i+}y_{+j}}{n})^2}{\frac{y_{i+}y_{+j}}{n}},$$

and under the independence assumption has approximately a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom.

Example 24. For the student data, $X = 45.227$, with a p-value of 0.037. Note, however, that 1 cell has an expected value of less than 1. It may be advisable to combine categories, for example SES classes 5 and 6.

5.3.3 The odds ratio

Consider a two way table with two factors each at two levels. As usual, let π_{ij} be the probability of classification into the i, j cell. The situation is illustrated in Table 5.10.

Recall that independence of the two factors is equivalent to $\pi_{ij} = \pi_{i+}\pi_{+j}$ for $i, j, = 1, 2$. Several equivalent formulations of independence in a 2×2 table exist. For example, we may consider the *odds ratio*

$$\alpha = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The name comes from the fact that α is the ratio of the *odds* π_{11}/π_{12} and π_{21}/π_{22} for each level of the column factor. It is an easy exercise to verify that independence is equivalent to $\alpha = 1$.

The sample odds ratio

Suppose we take a sample of n cases and classify them into a 2×2 table. Let y_{11}, y_{21}, y_{12}, y_{22} be the number of cases classified into the indicated cells. Recall that $\hat{\pi}_{ij} = y_{ij}/n$ is the MLE of the cell probabilities.

An obvious estimate of the odds ratio is

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

Table 5.11: 326 convicted Florida homicide defendants.

		Death penalty		
		Yes	No	Total
Race	White	19	141	160
	Black	17	149	166
	Total	36	290	326

For large n, the distribution of $\log \hat{\alpha}$ is approximately normal, with mean $\log \alpha$, and we can estimate the standard error of $\log \hat{\alpha}$ by

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

An approximate test of $\log \alpha = 0$ (i.e. of $\alpha = 1$ or independence) is to calculate the statistic $\log \hat{\alpha}/s.e.(\log \hat{\alpha})$. For this statistic, p-values may be computed from the standard normal distribution. A confidence interval for $\log \alpha$ is

$$\log \hat{\alpha} \pm z \text{s.e.}(\log \hat{\alpha})$$

where z is the appropriate percentage point of the normal distribution.

Example 25. In this example, the 326 "cases" are convicted defendants in homicide indictments in 20 Florida counties 1976-77. The factors are Defendant's Race and Death Penalty. The data are in Table 5.11.

The values of the deviance and the chi-square test are both 0.22 (a fluke!), with a p-value of 0.639. The estimated odds ratio $\hat{\alpha}$ is 1.18, and its log is 0.166. The standard error of log $\hat{\alpha}$ is

$$\sqrt{\frac{1}{19} + \frac{1}{17} + \frac{1}{141} + \frac{1}{149}} = 0.354$$

Hence the value of the test statistic is $\frac{0.166}{0.354} = .469$, with a p value of 0.639. A 95% confidence interval is $0.166 \pm (1.96)(0.354)$ i.e. $(-0.527,\ 0.860)$ for $\log \alpha$, or equivalently, $(e^{-0.527}, e^{0.860}) = (0.59, 2.36)$ for α. Clearly, we cannot reject the hypothesis of independence here, since the interval contains 1.

5.3.4 Log-linear models

A very useful device for representing various types of dependence structure in contingency tables is the *log-linear model*. We begin discussing these models in the simple context of two-way tables.

Define parameters $\eta_{ij} = \log \pi_{ij}$, and as in two-way ANOVA, let

$$\begin{aligned}
\mu &= \overline{\eta}_{..}, \\
\alpha_i &= \overline{\eta}_{i.} - \overline{\eta}_{..}, \\
\beta_j &= \overline{\eta}_{.j} - \overline{\eta}_{..}, \\
(\alpha\beta)_{ij} &= \eta_{ij} - \overline{\eta}_{i.} - \overline{\eta}_{.j} + \overline{\eta}_{..}.
\end{aligned}$$

Then, again as in ANOVA,

$$\log \pi_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}. \tag{5.12}$$

However, the parameters are not all functionally dependent. The usual ANOVA constraints apply, so for example, $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ for all j, and $\sum_j (\alpha\beta)_{ij} = 0$ for all i. Moreover, the constraint $\sum_i \sum_j \pi_{ij} = 1$ means that μ can be expressed in terms of the other parameters. Thus, the cell probabilities can be written in terms of $IJ - 1$ independent parameters $\alpha_1, \ldots, (\alpha\beta)_{(I-1),(J-1)}$. The model (5.12) is called a *log-linear model* for the π_{ij}s. The model (5.12), with "main effects" and "interactions" is just the saturated model of the previous section. We have made no restrictions on the probabilities, just expressed them in terms of "ANOVA"style parameters.

Fitting the model

Fitting the log-linear model involves calculating ML estimates of the parameters – i.e. minimising

$$-2 \sum_i \sum_j y_{ij} \log \pi_{ij}$$

as a function of $\alpha_1, \ldots, (\alpha\beta)_{(I-1),(J-1)}$.

In Splus, the fitting is done indirectly, using the following cunning trick. Suppose we regard the count in the i, j cell as being sampled from a Poisson distribution with mean λ_{ij} say. If we set $\eta_{ij} = \log \lambda_{ij}$, and define μ, α_i and so on as we did above, then the MLE's of the parameters in this Poisson model are exactly the same as those in the log-linear model. Moreover, the deviances, residuals and fitted values are exactly the same. Poisson models are fitted by the `glm` function, using the argument `family=poisson`. The model is specified exactly as in ANOVA. We illustrate with the student SES data.

Example 26. To fit the log-linear model to the student SES data, we type

```
> ses.glm<-glm(y~ses+course+ses:course,family=poisson,maxit=20)
> summary(ses.glm)

Call: glm(formula = y ~ ses + course + ses:course,
                              family = poisson, maxit = 20)

Coefficients:
                  Value Std. Error     t value
(Intercept)  2.13305655 0.08406273  25.37458064
       ses1  1.28083958 0.10232880  12.51690215
       ses2  0.90760084 0.11049709   8.21379853
       ses3  0.43393470 0.12717120   3.41220885
       ses4  0.29093471 0.14091293   2.06464167
       ses5 -1.64717121 0.29283331  -5.62494475
    course1  1.02476778 0.12649839   8.10103424
    course2  0.34792477 0.14287913   2.43509869
    course3 -0.20995769 0.24364600  -0.86173254
    course4  0.29869274 0.15017559   1.98895664
    course5 -0.10221434 0.20343336  -0.50244632
    course6 -0.89769650 0.25435661  -3.52928317
ses1course1 -0.10793057 0.16022530  -0.67361753
ses2course1 -0.28123553 0.17879434  -1.57295544
.......and more estimates........
(Dispersion Parameter for Poisson family taken to be 1 )

Null Deviance: 573.2494 on 41 degrees of freedom
```

```
Residual Deviance: 0 on 0 degrees of freedom

Number of Fisher Scoring Iterations: 19
```

Note that the algorithm converged rather slowly - we had to request 20 iterations. Seeing the model is saturated, we get a deviance of zero. Empty cells can also cause convergence problems - it helps to replace them with e.g. 0.5.

Independence and log-linear models

In the model (5.12), independence of the factors is equivalent to $(\alpha\beta)_{ij} = 0$ for all i and j. To see this, suppose first that the two factors are independent. Then $\pi_{ij} = \pi_{i+}\pi_{+j}$, and substituting this into

$$\begin{aligned} (\alpha\beta)_{ij} &= \eta_{ij} - \overline{\eta}_{i.} - \overline{\eta}_{.j} + \overline{\eta}_{..} \\ &= \log \pi_{ij} - \sum_{i=1}^{I} \log \pi_{ij}/I - \sum_{j=1}^{J} \log \pi_{ij}/J + \sum_{i=1}^{I}\sum_{j=1}^{J} \log \pi_{ij}/IJ \qquad (5.13) \end{aligned}$$

we get, after lengthy algebraic manipulations, the result $(\alpha\beta)_{ij} = 0$. Conversely, if $(\alpha\beta)_{ij} = 0$ for each i and j, then

$$\log \pi_{ij} = \mu + \alpha_i + \beta_j$$

and so

$$\pi_{ij} = e^{\mu} e^{\alpha_i} e^{\beta_j}.$$

Thus

$$\pi_{i+} = e^{\mu} e^{\alpha_i} \sum_{j=1}^{J} e^{\beta_j} = C_1 e^{\mu+\alpha_i}$$

say, and

$$\pi_{+j} = e^{\mu} e^{\beta_j} \sum_{i=1}^{I} e^{\alpha_i} = C_2 e^{\mu+\beta_j},$$

say. Hence,

$$\pi_{i+}\pi_{+j} = C_3 e^{\mu+\alpha_i+\beta_j} = C_3 \pi_{ij} \qquad (5.14)$$

where

$$C_3 = e^{\mu} \sum_{i=1}^{I} e^{\alpha_i} \sum_{j=1}^{J} e^{\beta_j}.$$

Adding both sides of (5.14) over i and j gives $C_3 = 1$, and hence $\pi_{ij} = \pi_{i+}\pi_{+j}$, and so the two factors are independent.

It follows that we can test independence by fitting an additive log-linear model (i.e. one with no interactions) and seeing if the additive submodel is adequate.

Example 27. The data in Table 5.12 give the results of classifying 400 melanoma patients according to type of tumour and site of tumour. An Splus program to read in the data and fit an additive model is

```
> y<-c(22,2,10,16,54,115,19,33,73,11,17,28)
> type<-rep(c("Freckle","Melanoma","Nodular","Indeterminate"),
+          c(3,3,3,3))
```

Table 5.12: 400 melanomas by site and type.

	Site			
Histological type	Head and Neck	Trunk	Extremities	Total
Hutchinsons melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

```
> type<-factor(type,levels=unique(type))
> site<-rep(c("Head and Neck","Trunk","Extremities"),4)
> site<-factor(site,levels=unique(site))
> cancer.glm<-glm(y~type+site,family=poisson)
> summary(cancer.glm)

Call: glm(formula = y ~ type + site, family = poisson)
Deviance Residuals:
       Min        1Q    Median        3Q      Max
 -3.045336 -1.074106 0.1296569 0.5856518 5.135292

Coefficients:
                 Value Std. Error    t value
(Intercept)  3.1764650 0.06672231  47.607238
      type1 -0.8737146 0.13555988  -6.445230
      type2  0.8202536 0.08050656  10.188655
      type3  0.4282115 0.08819635   4.855207
      site1 -0.5483195 0.08983258  -6.103793
      site2 -0.1043882 0.07946273  -1.313675

(Dispersion Parameter for Poisson family taken to be 1 )

    Null Deviance: 295.203 on 11 degrees of freedom

Residual Deviance: 51.79501 on 6 degrees of freedom

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:
      (Intercept)      type1      type2      type3      site1
type1  0.3892042
type2 -0.4518752  -0.4463883
type3 -0.3022502  -0.4617212  0.0601799
site1  0.2880608   0.0000000  0.0000000  0.0000000
site2 -0.0054655   0.0000000  0.0000000  0.0000000 -0.6821281
> 1-pchisq( 51.79501,6)
[1] 2.050456e-09
```

The p-value for the deviance is far too small for the independence submodel to be adequate. There is clear evidence that site and type are related.

There is not a great deal to be gained by introducing log-linear models in the context of 2-dimensional contingency tables. However, they are much more useful in the context of 3 and higher dimensional tables.

5.3.5 Three-dimensional tables

Suppose we now have 3 factors A,B, C with I, J, and K levels respectively. The three-dimensional table is illustrated in Figure 5.7. Factor A is the "row" factor, factor B the "column" factor and factor C the "slice" factor. Let π_{ijk} be the probability that a case is classified into the i, j, k cell. Then we can write

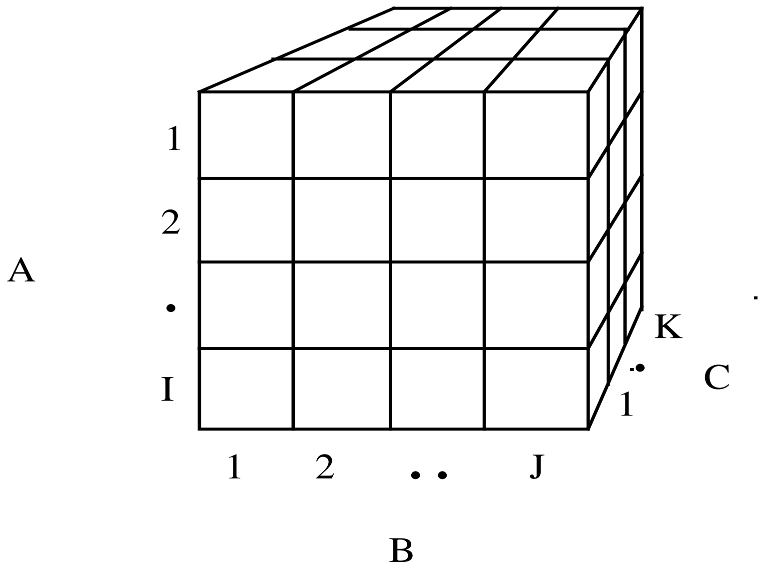

Figure 5.7: A three-dimensional contingency table.

$$\eta_{ijk} = \log \pi_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \tag{5.15}$$

where the parameters are defined in the same way as in ANOVA.

Suppose we have n cases, and y_{ijk} of them fall in the ijk cell. Under multinomial sampling, the likelihood is

$$l(\pi_1, \ldots, \pi_m) = C \prod_{ijk} y_{ijk} \pi_{ijk}^{y_{ijk}}$$

and the minimum -2ℓ is

$$-2 \left(\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk} \log y_{ijk} - n \log n \right).$$

Various submodels of model (5.15) may be considered, corresponding to constraints that describe various types of independence.

All three factors independent

The mutual independence of A, B and C is equivalent to

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \tag{5.16}$$

for all i, j, k, where, for example, $\pi_{i++} = \sum_{j=1}^{J}\sum_{k=1}^{K}\pi_{ijk}$.

Writing $(\alpha\beta)$ for $(\alpha\beta)_{ij}$, etc, it can be shown that the concept of independence as defined by (5.16) is equivalent to

$$(\alpha\beta) = (\alpha\gamma) = (\beta\gamma) = (\alpha\beta\gamma) = 0$$

for all i, j, k. If we think of these as "interactions" by analogy with ANOVA, mutual independence of A, B, C is thus equivalent to all *two and three factor interactions being zero.*

To test if this submodel is appropriate, i.e. to test the hypothesis (5.16), we can fit the submodel with no interactions using `glm`. Under the null hypothesis that the independence model is adequate, the deviance of the mutual independence model is distributed as χ^2 with $(IJK-1)-(I+J+K-3) = IJK-I-J-K+2$ degrees of freedom. A numerical example is discussed in Example 28.

One factor independent of the other two

If A is independent of B and C, then

$$\pi_{ijk} = \pi_{i++}\pi_{+jk} \tag{5.17}$$

for all i, j and k, where $\pi_{+jk} = \sum_{i=1}^{I}\pi_{ijk}$, and similarly for the other possibilities where B is independent of A and C, and C is independent of A and B. The relationship (5.17) is equivalent to

$$(\alpha\beta) = 0, \quad (\alpha\gamma) = 0, \quad (\alpha\beta\gamma) = 0,$$

or all interactions involving A equal to zero; i.e. to a model

$$\eta_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} \tag{5.18}$$

which may be specified in `glm` by

```
y ~ a  + b + c + b:c
```

for factors tt a, `b` and `c`. The hypothesis (5.17) can be tested by examining the residual deviance of the model (5.18)

Two factors independent conditional on the third

Events E and F are said to be *conditionally independent given an event* G if

$$P(E\cap F|G) = P(E|G)P(F|G).$$

Factors A and B are conditionally independent given a third factor C if the events $A = i$ and $B = j$ are conditionally independent given $C = k$ for all i, j, k. In terms of the π's this is equivalent to

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$$

Table 5.13: The Florida murder data.

	C: Victim's race			
	Black		White	
A: Defendant's	B: Death Penalty		B: Death Penalty	
Race	Yes	No	Yes	No
Black	13	195	23	105
White	1	19	39	265

for all i, j, k. In terms of the log linear model, it is also equivalent to $(\alpha\beta) = 0$ and $(\alpha\beta\gamma) = 0$. Equivalently, all interactions containing A and B must be zero.

The conditional independence of A and B given C may be tested by fitting the model

$$\log \pi_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}$$

using the formula

```
y ~ a  + b + c + + a:c + b:c
```

and checking the deviance.

We illustrate the fitting with an expanded version of the Florida murder data.

Example 28. Table 5.13 gives a more detailed breakdown of a larger set of Florida murder data, with a third variable (Victim's race) added. The following Splus code fits the various models. We have used a different call to `glm` for each model.

```
> race
      a   b     c   y
1 black yes black  13
2 black yes white  23
3 black  no black 195
4 black  no white 105
5 white yes black   1
6 white yes white  39
7 white  no black  19
8 white  no white 265

># fit mutual independence model;
> glm(y~a+b+c,family=poisson,data=race)
Call:
glm(formula = y ~ a + b + c, family = poisson, data = race)

Coefficients:
 (Intercept)          a        b          c
    3.913678 0.01817604 1.019582 -0.3195156

Degrees of Freedom: 8 Total; 4 Residual
Residual Deviance: 266.9018

> # fit  model for independence of A from B and C;
> glm(y~a+b+c+b:c,family=poisson,data=race)
```

Table 5.14: Summary statistics for different models for the murder data.

Model	residual deviance	df	p-value
a+b+c	266.9	4	0.0001
a+b+c+a:b	266.47	3	0.0052
a+b+c+a:c	12.75	3	0.0001
a+b+c+b:c	256.07	3	0.0001
a+b+c+a:b+b:c	255.64	2	0.0001
a+b+c+a:b+a:c	12.32	2	0.0021
a+b+c+b:c+a:c	1.92	2	0.3826
Saturated	0	0	

```
Call:
glm(formula = y ~ a + b + c + b:c, family = poisson, data = race)

Coefficients:
 (Intercept)          a        b          c       b:c
    3.818134 0.01817708 1.128315 -0.5088744 0.2351305

Degrees of Freedom: 8 Total; 3 Residual
Residual Deviance: 256.0741


> # fit conditional independence of A and B given C;
> glm(y~a+b+c+a:c+b:c,family=poisson,data=race)
Call:
glm(formula = y ~ a + b + c + a:c + b:c, family = poisson, data =
race)

Coefficients:
 (Intercept)         a        b          c       a:c       b:c
    3.488114 0.3692021 1.128322 -0.7483079 0.8017008 0.2351375

Degrees of Freedom: 8 Total; 2 Residual
Residual Deviance: 1.921581
```

and so on for all the other models. The deviances for these models can be summarised in a table: Which model should we fit? Table 5.14 shows the difference in the deviances for each model, together with the degrees of freedom and the p-values. We see from the table that the model

```
 y ~ a  + b + c + a:c + b:c
```

fits best, since it has the smallest deviance and a p-value of 0.3826. This is the model of conditional independence of A and B given C: given the victims race, the imposition of the death penalty is independent of the the defendant's race. Note that this is not the same as the unconditional independence of death penalty and defendant's race.

5.3.6 Residual analysis for contingency tables

We consider a 3-factor example, the 2-factor case is similar.

In contingency table work, Pearson residuals are defined as

$$r_{ijk} = \frac{y_{ijk} - n\hat{\pi}_{ijk}}{\sqrt{n\hat{\pi}_{ijk}}}$$

i.e. the square roots of the components of the Pearson χ^2 statistic used to test the adequacy of the model:

$$\chi^2 = \sum \frac{(y_{ijk} - n\hat{\pi}_{ijk})^2}{n\hat{\pi}_{ijk}}.$$

The analysis consists of computing these residuals and noting which cells fit poorly i.e. which have big residuals.

Example 29. We illustrate using the death penalty data. The model fitted is that corresponding to the conditional independence of defendants race and death penalty, given the victim's race. The residuals are extracted from the glm object as before, using `residuals`:

```
> cond.glm<- glm(y~a+b+c+b:c+a:c,family=poisson,data=race)
> p.res<-residuals(cond.glm,type="p")
> pred<-fitted.values(cond.glm)
> data.frame(race,pred,p.res)
      a   b     c   y      pred       p.res
1 black yes black  13  12.77193  0.06381752
2 black yes white  23  18.37037  1.08011124
3 black  no black 195 195.22807 -0.01632290
4 black  no white 105 109.62963 -0.44216108
5 white yes black   1   1.22807 -0.20580523
6 white yes white  39  43.62963 -0.70087212
7 white  no black  19  18.77193  0.05263976
8 white  no white 265 260.37037  0.28691339
>sum(p.res^2)
[1] 1.985154
```

The value of the χ^2 statistic is 1.985154. The residuals are all quite small, indicating an excellent fit.

5.3.7 Simpson's paradox

Given a 3-dimensional table we may *collapse* the table over one dimension to produce a *marginal* two dimensional table. For example, we may collapse the table over variable C and consider the table with entries y_{ij+}. However, association present in the separate two-dimensional tables (i.e. the K tables with entries y_{ijk} for different values of k) may not be present (or may even be opposite) to the association in the marginal table. The Florida murder data illustrates the point nicely.

Recall that the odds ratio measures association in 2×2 tables, and

$$\alpha = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\left(\frac{\pi_{11}}{\pi_{12}}\right)}{\left(\frac{\pi_{21}}{\pi_{22}}\right)}$$

is the ratio of the odds of $B = 1$ to $B = 2$ for the different levels of A. Moreover, $\alpha < 1$ is equivalent to $\frac{\pi_{11}}{\pi_{1+}} < \frac{\pi_{21}}{\pi_{2+}}$ i.e. the chance of getting level 1 of B is less for level 1 of A than for level 2 of A. Also note that the sample odds ratio is $Y_{11}Y_{22}/Y_{12}Y_{21}$.

Now consider the murder data again. The separate tables for the two levels of C are

Victim=Black

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

for which the odds ratio is 1.2666, and

Victim=White

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

for which the odds ratio is 1.4884. Since both odds ratios are more than 1, the chance of having the death penalty is more for level 1 of A than for level 2 i.e. treating the cases of black and white victims separately, we conclude that in both cases blacks are more likely to receive the death penalty than whites, although the analysis of the previous section shows that this is not significant.

If we collapse the table over C (race of victim) we get

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

with a sample odds ratio of 0.852. The conclusion is now reversed - on the basis of the marginal table whites are now more likely to receive the death penalty!

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, on the collapsed table it appears that whites get more death sentences. Using the third variable C, the true picture emerges.