Chapter 10
Models for Counts: Poisson and Negative Binomial GLMs

Poor data and good reasoning give poor results.
Good data and poor reasoning give poor results.
Poor data and poor reasoning give rotten results.
E. C. Berkeley [4, p. 20]

10.1 Introduction and Overview

The need to count things is ubiquitous, so data in the form of counts arise often in practice. Examples include: the number of alpha particles emitted from a source of radiation in a given time; the number of cases of leukemia reported per year in a certain jurisdiction; the number of flaws per metre of electrical cable. This chapter is concerned with counts when the individual events being counted are independent, or nearly so, and where there is no clear upper limit for the number of events that can occur, or where the upper limit is very much greater than any of the actual counts. We first compile important information about the Poisson distribution (Sect. 10.2), the distribution most often used with count data. Poisson regression, or models for count data described by covariates, has already been covered in Sect. 8.12 and elsewhere. In this chapter, we then focus on describing the models for three types of count data: models for count data described by covariates, models for rates (Sect. 10.3) and models for counts organized in tables (Sect. 10.4). Overdispersion is discussed in Sect. 10.5, including a discussion of negative binomial GLMs and quasi-Poisson models as alternative models.

10.2 Summary of Poisson GLMs

The distribution most often used for modelling counts is the Poisson distribution, which has the probability function

$$\mathcal{P}(y;\mu) = \frac{\exp(-\mu)\mu^y}{y!}$$

P. K. Dunn, G. K. Smyth, *Generalized Linear Models with Examples in R*,
Springer Texts in Statistics, https://doi.org/10.1007/978-1-4419-0118-7_10

for $y = 0, 1, 2, \ldots$, with expected counts $\mu > 0$. The Poisson distribution has already been established as an EDM (Example 5.2), and a Poisson GLM proposed for the noisy miner data in Example 1.5. Useful information about the Poisson distribution appears in Table 5.1. The unit deviance for the Poisson distribution is

$$d(y, \mu) = 2\left\{y \log \frac{y}{\mu} - (y - \mu)\right\},$$

when the residual deviance is $D(y, \hat{\mu}) = \sum_{i=1}^{n} w_i d(y_i, \hat{\mu}_i)$, where w_i are the prior weights. When $y = 0$, the limit form of the unit deviance (5.14) is used. By the saddlepoint approximation, $D(y, \hat{\mu}) \sim \chi^2_{n-p'}$ where p' is the number of coefficients in the linear predictor. The approximation is adequate if $y_i \geq 3$ for all i (Sect. 7.5, p. 276).

The most common link function used for Poisson GLMs is the logarithmic link function (which is also the canonical link function), which ensures $\mu > 0$ and enables the regression parameters to be interpreted as having multiplicative effects. Using the logarithmic link function (`"log"` in R), the general form of a Poisson GLM is

$$\begin{cases} y \sim \text{Pois}(\mu) \\ \log \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \end{cases} \tag{10.1}$$

The systematic component of (10.1) can be written as

$$\begin{aligned} \mu &= \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \\ &= \exp \beta_0 \times (\exp \beta_1)^{x_1} \times (\exp \beta_2)^{x_2} \times \cdots \times (\exp \beta_p)^{x_p}. \end{aligned}$$

This shows that the impact of each explanatory variable is multiplicative. Increasing x_j by one increases μ by *factor* of $\exp(\beta_j)$. If $\beta_j = 0$ then $\exp(\beta_j) = 1$ and μ is not related to x_j. If $\beta_j > 0$ then μ increases if x_j increases; if $\beta_j < 0$ then μ decreases if x_j increases.

Sometimes, the link functions `"identity"` ($\eta = \mu$) or `"sqrt"` ($\eta = \sqrt{\mu}$) are used with Poisson GLMs. A Poisson GLM is denoted GLM(Pois; link), and is specified in R using `family=poisson()` in the `glm()` call.

When the explanatory variables are all qualitative (that is, factors), the data can be summarized as a contingency table and the model is often called a *log-linear model* (Sect. 10.4). If any of the explanatory variables are quantitative (that is, covariates), the model is often called a *Poisson regression model*. Since Poisson regression has been discussed earlier (Sect. 8.12), we do not consider Poisson regression models further (but see Sect. 10.6 for a Case Study).

When the linear predictor includes an intercept term (as is almost always the case), and the log-link function is used, the residual deviance can be simplified to

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} w_i y_i \log(y_i / \hat{\mu}_i);$$

that is, the second term in the unit deviance can be dropped as it sums to zero (Problem 10.2). This identity will be used later to clarify the analysis of contingency tables.

For Poisson GLMs, the use of quantile residuals [12] is strongly recommended (Sect. 8.3.4.2).

10.3 Modelling Rates

The first context we consider is when the maximum number of events is known but large; that is, there is an upper bound for each count response, but the upper bound is very large. For such applications, the maximum number of events is usually representative of some population, and the response can be usefully viewed as a *rate* rather than just as a count. The size of each population needs to be specified to make comparisons meaningful. For example, consider comparing the number of people with a certain disease in various cities. The number of cases in each city may be useful information for planning purposes. However, quoting just the number of people with the disease in each city is an unfair comparison, as some cities have a far larger population than others. Comparing the number of people with the disease per unit of population (for example, per thousand people) is a fairer comparison. That is, the disease *rate* is often more suitable for modelling than the actual number of people with the disease.

In principle, rates can treated as proportions, and analysed using binomial GLMs, but Poisson GLMs are more convenient when the populations are large and the rates are relatively small, less than 1% say.

Example 10.1. As a numerical example, consider the number of incidents of lung cancer from 1968 to 1971 in four Danish cities (Table 10.1; data set: `danishlc`), recorded by age group [2, 26]. The *number* of cases of lung cancer in each age group is remarkably similar for Fredericia. However, using the number of cases does not accurately reflect the information in the data because five times as many people are in the 40–54 age group than in the over-75 age group. Understanding the data is enhanced by considering the *rate* of lung cancer, such as the number of lung cancer cases per unit of population. A plot of the cancer rates against city and age (Fig. 10.1) suggests the lung cancer rate may change with age:

```
> data(danishlc)
> danishlc$Rate <- danishlc$Cases / danishlc$Pop * 1000  # Rate per 1000
> danishlc$Age <- ordered(danishlc$Age,  # Ensure age-order is preserved
    levels=c("40-54", "55-59", "60-64", "65-69", "70-74", ">74") )
```

Table 10.1 Incidence of lung cancer in four Danish cities from 1968 to 1971 inclusive (Example 10.1)

	Fredericia		Horsens		Kolding		Vejle	
Age	Cases	Population	Cases	Population	Cases	Population	Cases	Population
40–54	11	3059	13	2879	4	3142	5	2520
55–59	11	800	6	1083	8	1050	7	878
60–64	11	710	15	923	7	895	10	839
65–69	10	581	10	834	11	702	14	631
70–74	11	509	12	634	9	535	8	539
Over 74	10	605	2	782	12	659	7	619

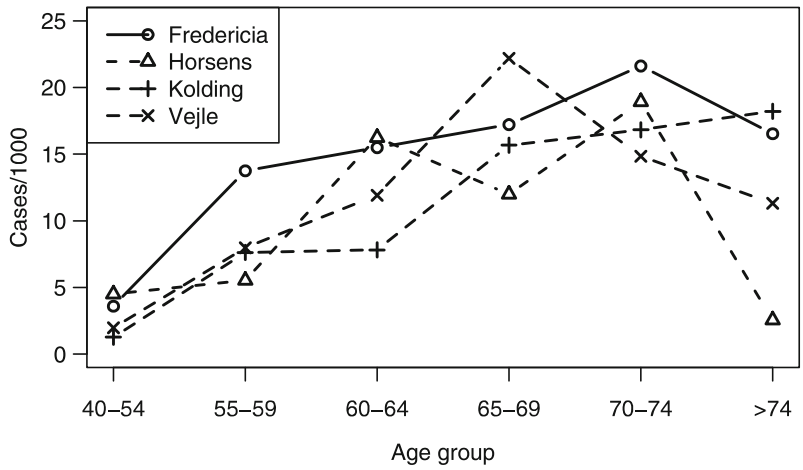

Fig. 10.1 The Danish lung cancer rates for various age groups in different cities (Example 10.1)

```
> danishlc$City <- abbreviate(danishlc$City, 1)   # Abbreviate city names
> matplot( xtabs( Rate ~ Age+City, data=danishlc), pch=1:4, lty=1:4,
           type="b", lwd=2, col="black", axes=FALSE, ylim=c(0, 25),
          xlab="Age group", ylab="Cases/1000")
> axis(side=1, at=1:6, labels=levels(danishlc$Age))
> axis(side=2, las=1); box()
> legend("topleft", col="black", pch=1:4, lwd=2, lty=1:4, merge=FALSE,
        legend=c("Fredericia", "Horsens", "Kolding", "Vejle") )
```

The R function `ordered()` informs R that the levels of factor `Age` have a particular order; without declaring `Age` as an ordered factor, `Age` is plotted with `">74"` as the first level. The plots show no clear pattern by city, but the lung cancer rate appears to grow steadily for older age groups for each city, then falls away for the `>74` age group. The lung cancer rate for Horsens in the `>74` age group seems very low.

An unfortunate side-effect of declaring `Age` as an ordered factor is that R uses polynomial contrasts for coding, which are not appropriate here (the

ordered categories are not equally spaced) and are hard to interpret anyway. To instruct R to use the familiar treatment coding for ordered factors, use:

```
> options(contrasts= c("contr.treatment", "contr.treatment"))
```

The first input tells R to use treatment coding for unordered factors (which is the default), and the second to use treatment coding for ordered factors (rather than the default `"contr.poly"`).

Define y_i as the observed number of lung cancers in group i where the corresponding population is T_i. The lung cancer *rate* per unit of population is y_i/T_i, and the expected rate is $\mathrm{E}[y_i/T_i] = \mu_i/T_i$, where μ_i possibly depends on the explanatory variables, and T_i is known. Using a logarithmic link function, the suggested systematic component is $\log(\mu_i/T_i) = \eta_i$. Dropping the subscript i, the model suggested for cancer rates is

$$\begin{cases} y \sim \mathrm{Pois}(\mu) \\ \log \mu = \log T + \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \end{cases}$$

where the explanatory variables x_j are the necessary dummy variables required for the cities and age groups. The parameters β_j must be estimated, but no parameters need to be estimated for $\log T$. In other words, the term $\log T$ is an *offset* (Sect. 5.5.2).

Fit the model in R as follows, starting with the interaction model:

```
> dlc.m1 <- glm( Cases ~ offset( log(Pop) ) + City * Age,
                 family=poisson, data=danishlc)
> anova(dlc.m1, test="Chisq")
         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        23    129.908
City      3    3.393        20    126.515  0.33495
Age       5  103.068        15     23.447  < 2e-16 ***
City:Age 15   23.447         0      0.000  0.07509 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We decide to retain only `Age` in the model.

```
> dlc.m2 <- update(dlc.m1, . ~ offset(log(Pop)) + Age )
```

An alternative model might consider `Age` as quantitative (since the categories are not equally spaced), using the lower class boundary of each class. (The *lower* boundary are preferred since the final class only has a lower boundary; the class midpoint or upper boundary becomes subjective for the final class.)

```
> danishlc$AgeNum <- rep( c(40, 55, 60, 65, 70, 75), 4)
> dlc.m3 <- update(dlc.m1, . ~ offset( log(Pop) ) +  AgeNum)
```

Figure 10.1 may suggest a possible quadratic relationship, but note the lower class boundaries are not equally spaced:

```
> dlc.m4 <- update( dlc.m3, . ~ offset( log(Pop) ) + poly(AgeNum, 2) )
```

The quadratic model is an improvement over the model linear in `AgeNum`:

```
> anova( dlc.m3, dlc.m4, test="Chisq")
Analysis of Deviance Table

Model 1: Cases ~ AgeNum + offset(log(Pop))
Model 2: Cases ~ poly(AgeNum, 2) + offset(log(Pop))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        22     48.968
2        21     32.500  1   16.468 4.948e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the models are not nested, we compare the four models using the AIC:

```
> c( "With interaction"=AIC(dlc.m1), "Without interaction"=AIC(dlc.m2),
    "Age (numerical)"=AIC(dlc.m3), "Age (numerical; quadratic)"=AIC(dlc.m4) )
          With interaction        Without interaction
                  144.3880                  136.6946
           Age (numerical) Age (numerical; quadratic)
                  149.3556                  134.8876
```

The AIC suggests the quadratic model `dlc.m4` produces the best predictions, but the AIC for models `dlc.m2` and `dlc.m4` are very similar.

The saddlepoint approximation is suitable for Poisson distributions when $y_i > 3$ for all observations. For these data:

```
> sort( danishlc$Cases )
 [1]  2  4  5  6  7  7  7  8  8  9 10 10 10 10 11 11 11 11 11 12 12 13 14
[24] 15
```

which shows that the saddlepoint approximation may be suspect. However, only one observation fails to meet this criterion, and only just, so we use the goodness-of-fit tests remembering to be cautious:

```
> D.m2 <- deviance(dlc.m2); df.m2 <- df.residual( dlc.m2 )
> c( Dev=D.m2, df=df.m2, P = pchisq( D.m2, df.m2, lower = FALSE) )
        Dev          df           P
28.30652745 18.00000000  0.05754114
> D.m4 <- deviance(dlc.m4); df.m4 <- df.residual( dlc.m4 )
> c( Dev=D.m4, df=df.m4, P=pchisq( D.m4, df.m4, lower = FALSE) )
        Dev          df           P
32.49959158 21.00000000  0.05206888
```

Both models are reasonably adequate. Consider the diagnostic plots (Fig. 10.2), where the constant-information scale is from Table 8.1:

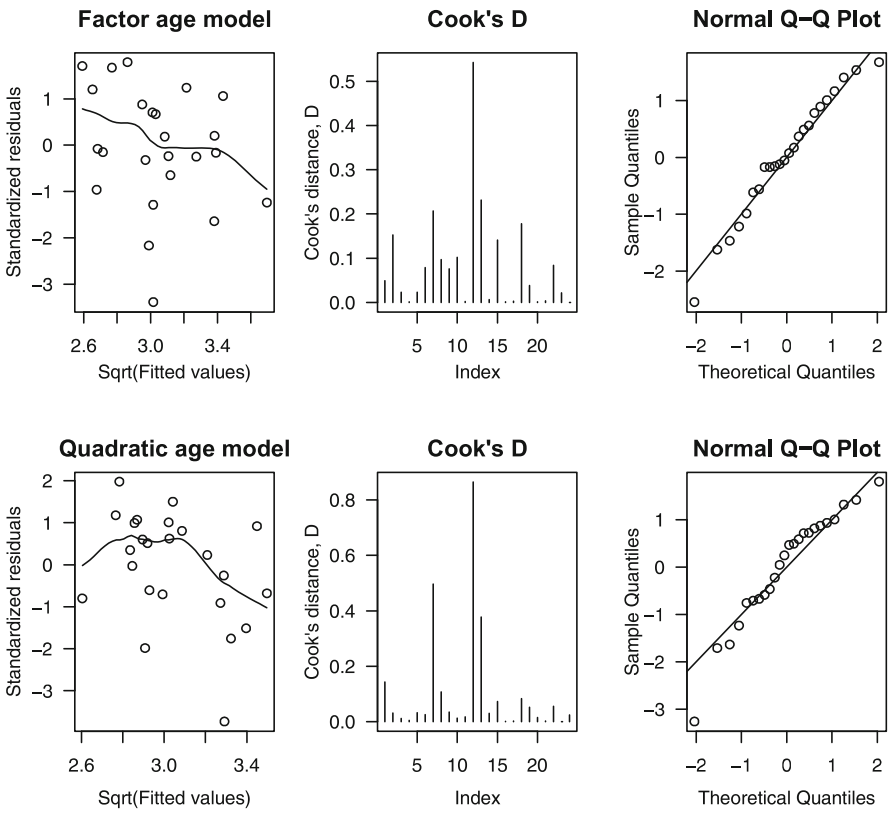

Fig. 10.2 Diagnostic plots for two models fitted model to the Danish lung cancer data. Top panels: treating age as a factor (model `dlc.m2`); bottom panels: fitting a quadratic in age (model `dlc.m4`). The Q–Q plots use quantile residuals (Example 10.1)

```
> library(statmod)  # For quantile residuals
> scatter.smooth( rstandard(dlc.m2) ~ sqrt(fitted(dlc.m2)),
         ylab="Standardized residuals", xlab="Sqrt(Fitted values)",
         main="Factor age model", las=1 )
> plot( cooks.distance(dlc.m2), type="h", las=1, main="Cook's D",
        ylab="Cook's distance, D")
> qqnorm( qr<-qresid(dlc.m2), las=1 ); abline(0, 1)
> scatter.smooth( rstandard(dlc.m4) ~ sqrt(fitted(dlc.m4)),
         ylab="Standardized residuals", xlab="Sqrt(Fitted values)",
         main="Quadratic age model", las=1 )
> plot( cooks.distance(dlc.m4), type="h", las=1, main="Cook's D",
        ylab="Cook's distance, D")
> qqnorm( qr<-qresid(dlc.m4), las=1 ); abline(0, 1)
```

The diagnostics suggest that both models are reasonable models, though we prefer model `dlc.m2`, since model `dlc.m4` appears to show three observations with high influence relative to the other observations, and is a simpler model. □

10.4 Contingency Tables: Log-Linear Models

10.4.1 Introduction

Count data commonly appear in tables, called *contingency tables*, where the observations are cross-classified according to the levels of the classifying factors. To discuss the issues relevant to contingency tables, we begin with two cross-classifying factors (two-dimensional tables; Sect. 10.4.2 and 10.4.3) then extend to three cross-classifying factors (three-dimensional tables; Sect. 10.4.4) and then extend to higher-order tables (Sect. 10.4.7).

10.4.2 Two Dimensional Tables: Systematic Component

The simplest contingency table is a two-way (or two-dimensional) table, with factors A and B. If factor A has I levels and factor B has J levels, the contingency table has size $I \times J$. In general, the entries in an $I \times J$ table are defined as shown in Table 10.2, where y_{ij} refers to the observed count in row i and column j for $i = 1, 2, \ldots I$ and $j = 1, 2, \ldots J$.

Write μ_{ij} for to the expected *count* in cell (i, j). For convenience, also define π_{ij} as the expected *probability* that an observation is in cell (i, j), where $\mu_{ij} = m\pi_{ij}$, and m is the total number of observations. We write $m_{i\bullet}$ to mean the sum of counts in row i over all columns, and $m_{\bullet j}$ to mean the sum of counts in column j over all rows. The use of the dot $\bullet$ in this context means to sum over all the elements of the index that the dot replaces.

If factors A and B are independent, then $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$ is true. Writing $\mu_{ij} = m\pi_{i\bullet}\pi_{\bullet j}$, take logarithms to obtain

$$\log \mu_{ij} = \log m + \log \pi_{i\bullet} + \log \pi_{\bullet j} \tag{10.2}$$

Table 10.2 The general $I \times J$ contingency table. The cell count y_{ij} corresponds to level i of A and level j of B (Sect. 10.4.2)

		Factor B				
		Column 1	Column 2	$\cdots$	Column J	Total
Factor A	Row 1	y_{11}	y_{12}	$\cdots$	y_{1J}	$m_{1\bullet}$
	Row 2	y_{21}	y_{22}	$\cdots$	y_{2J}	$m_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	Row I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}	$m_{I\bullet}$
	Total	$m_{\bullet 1}$	$m_{\bullet 2}$	$\cdots$	$m_{\bullet J}$	m

Table 10.3 The attitude of Australians to genetically modified foods (factor A) according to income (factor B) (Example 10.2)

	High income ($x_2 = 0$)	Low income ($x_2 = 1$)	Total
For GM foods ($x_1 = 0$)	263	258	521
Against GM foods ($x_1 = 1$)	151	222	373
Total	414	480	894

for the systematic component. This systematic component may be re-expressed using dummy variables, since the probabilities $\pi_{i\bullet}$ depend on which unique row the observation is in, and the probabilities $\pi_{\bullet j}$ depends on which unique column the observation is in.

Example 10.2. To demonstrate and fix ideas, first consider the smallest possible table of counts: a 2×2 table. The data in Table 10.3 were collected between December 1996 and January 1997, and comprise a two-dimensional (or two-way) table of counts collating the attitude of Australians to genetically modified (GM) foods (factor A) according to their income (factor B) [28, 31].

To analyse the data in R, first define the variables:

```
> Counts <- c(263, 258, 151, 222)
> Att <- gl(2, 2, 4, labels=c("For", "Against") )
> Inc <- gl(2, 1, 4, labels=c("High", "Low") )
> data.frame( Counts, Att, Inc)
  Counts     Att  Inc
1    263     For High
2    258     For  Low
3    151 Against High
4    222 Against  Low
```

The function `gl()` is used to generate factors by specifying the pattern in the factor `l`evels. The first input indicates the number of levels, the second input the number of times each level is repeated as a run according to how the counts are defined, and the third input is the length of the factor. The `labels` input is optional, and defines the names for each level of the factor. The variable `Inc`, for example, has two levels repeated one at a time (given the order of the counts supplied in `Counts`), and has a length of four. As a check, the contingency table in Table 10.3 can be created using

```
> gm.table <- xtabs( Counts ~ Att + Inc ); gm.table
          Inc
Att        High Low
  For       263 258
  Against   151 222
```

To test whether attitude is independent of income, a probabilistic model for the counts is needed. A complete model for the data in Table 10.3 depends on

how the sample of individuals was collected. We will see in the next section that a number of different possible sampling scenarios lead us back to the same basic statistical analysis. □

10.4.3 Two-Dimensional Tables: Random Components

10.4.3.1 Introduction

We now consider how the sample of individuals, tabulated in the contingency table, was collected. In particular, we consider whether any or all of the margins of the table were preset by the sampling scheme. A table of counts may arise from several possible sampling schemes, each suggesting a different probability model. Three possible scenarios are:

- The m observations are allocated to factors A and B as the observations randomly arrive; neither row nor column totals are fixed.
- A fixed total number of m observations are cross-classified by the factors A and B.
- The row totals are fixed, and observations allocated to factor B within each level of A. (Alternatively, the column total are fixed, and observations allocated to factor A within each level of B.)

10.4.3.2 No Marginal Totals Are Fixed

Firstly, assume no marginal totals are fixed, as would be the case if, for example, the data in Table 10.3 are collated from survey forms completed by customers randomly arriving at a large shopping centre over 1 week. In this scenario, no marginal total is fixed; no limits exists on how large the counts can be (apart from the city population, which is much larger than the counts in the table).

If the total number of individuals observed (the grand total in the table) can be viewed as Poisson distributed, and if the individuals give responses independently of one another, then each of the counts in the table must follow a Poisson distribution. The log-likelihood function for the 2×2 table is

$$\ell(\mu; y) = \sum_{i=1}^{2} \sum_{j=1}^{2} \left(-\mu_{ij} + y_{ij} \log \mu_{ij}\right), \tag{10.3}$$

ignoring the terms not involving the parameters μ_{ij}. The residual deviance is

$$D(y, \hat{\mu}) = \sum_{i=1}^{2} \sum_{j=1}^{2} y_{ij} \log \frac{y_{ij}}{\hat{\mu}_{ij}}, \tag{10.4}$$

omitting the term $y_{ij} - \hat{\mu}_{ij}$, which always sums to zero if the log-linear predictor contains the constant term (Sect. 10.2).

Example 10.3. A Poisson model can be fitted to the GM foods data (Example 10.2) in R as follows:

```
> gm.1 <- glm( Counts ~ Att + Inc, family=poisson)
> anova( gm.1, test="Chisq")
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     3     38.260
Att   1  24.6143         2     13.646 7.003e-07 ***
Inc   1   4.8769         1      8.769   0.02722 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Recall the logarithmic link function is the default in R for the Poisson distribution.) This model fits a log-linear model equivalent to (10.2), and hence assumes that attitude and income are independent. Both `Att` and `Inc` are statistically significant in the order they are fitted. The Poisson GLM has the coefficients

```
> coef( gm.1 )
(Intercept)  AttAgainst      IncLow
  5.4859102  -0.3341716   0.1479201
```

Thus the model has the systematic component

$$\log \hat{\mu}_{ij} = 5.486 - 0.3342x_1 + 0.1479x_2, \tag{10.5}$$

where $x_1 = 1$ for row $i = 2$ (against GM foods) and is zero otherwise, and $x_2 = 1$ for column $j = 2$ (low income) and is zero otherwise. (The R notation means, for example, that `AttAgainst` $= 1$ when the variable `Att` has the value `Against` and is zero otherwise.) The systematic component in the form of (10.5) is the usual regression model representation of the systematic component, where dummy variables are explicitly used for the rows and columns. Since each cell of the table belongs to just one row and one column, the dummy variables are often zero for any given cell.

Log-linear models are often easier to interpret when converted back to the scale of the fitted values. In particular, $\exp(\hat{\beta}_0)$ gives the fitted expected count for the first cell in the table, while similar expressions for the other parameters give the relative increase in counts for one level of a factor over the first. By unlogging, the systematic component (10.5) becomes

$$\begin{aligned}\hat{\mu}_{ij} &= \exp(5.486) \times \exp(-0.3342x_1) \times \exp(0.1479x_2) \\ &= 241.3 \times 0.7159^{x_1} \times 1.159^{x_2}.\end{aligned}$$

Compare the values of $\hat{\mu}_{ij}$ when $x_2 = 1$ to the values when $x_2 = 0$:

$$\begin{aligned} \text{When } x_2 = 0: \quad & \hat{\mu}_{i1} = 241.3 \times 0.7159^{x_1} \\ \text{When } x_2 = 1: \quad & \hat{\mu}_{i2} = 241.3 \times 0.7159^{x_1} \times 1.159. \end{aligned} \tag{10.6}$$

Under this model, the fitted values for $\hat{\mu}_{i2}$ are always 1.159 times the fitted values for $\hat{\mu}_{i1}$, for either value of x_1. From Table 10.3, the ratio of the corresponding column marginal totals is

```
> sum(Counts[Inc=="Low"]) /  sum(Counts[Inc=="High"])
[1] 1.15942
```

This value is exactly the factor in (10.6), which is no coincidence. This demonstrates an important feature of the main effects terms in log-linear models: the main effect terms in the model simply model the marginal totals. These marginal totals are usually not of interest. The purpose of the GM study, for example, is to determine the *relationship* between income and attitudes towards GM foods, not to estimate the proportion of Australians with high incomes. That is, the real interest lies with the *interaction* term in the model:

```
> gm.int <- glm( Counts ~ Att * Inc, family=poisson)
> anova( gm.int, test="Chisq")
        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        3     38.260
Att      1  24.6143         2     13.646 7.003e-07 ***
Inc      1   4.8769         1      8.769  0.027218 *
Att:Inc  1   8.7686         0      0.000  0.003065 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of deviance table shows the interaction term is necessary in the model. Notice that after fitting the interaction term, no residual deviance remains and no residual degrees of freedom remain, so the fit is perfect. This indicates that the number of coefficients in the model is the same as the number of entries in the table:

```
> length(coef(gm.int))
[1] 4
```

This means that the 2×2 table cannot be summarized by a smaller set of model coefficients. Since the interaction term is significant, the data suggest an association between income levels and attitude towards GM foods. We can examine the percentage of low and high income respondents who are `For` and `Against` GM foods by income level using `prop.table()`:

```
> round(prop.table(gm.table, margin=2)*100, 1)  # margin=2  means columns
         Inc
Att       High  Low
  For     63.5 53.8
  Against 36.5 46.2
```

This table shows that high income Australians are more likely to be in favour of GM foods than low income Australians.

Observe that the main result of the model fitting is that the interaction is significant (and hence that income and attitude to GM food are associated), rather than the individual estimates of the regression parameters. □

10.4.3.3 The Grand Total Is Fixed

Another scenario that may have produced the data in Table 10.3 assumes a fixed number of 894 people were sampled. For example, the researchers may have decided to survey 894 people in total, and then classify each respondent as `Low` or `High` income, and also classify each respondent as `For` or `Against` GM foods. While the counts are free to vary within the table, the counts have the restriction that their sum is capped at 894. However, the Poisson distribution has no upper limits on y by definition. Instead, the *multinomial distribution* is appropriate. For a 2×2 table, the probability function for the multinomial distribution is

$$\mathcal{P}(y_{11}, y_{12}, y_{21}, y_{22}; \mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}) = \frac{m!}{y_{11}!y_{12}!y_{21}!y_{22}!} \left(\frac{\mu_{11}}{m}\right)^{y_{11}} \left(\frac{\mu_{12}}{m}\right)^{y_{12}} \left(\frac{\mu_{21}}{m}\right)^{y_{21}} \left(\frac{\mu_{22}}{m}\right)^{y_{22}}.$$

Ignoring terms not involving μ_{ij}, the log-likelihood function is

$$\ell(\mu; y) = \sum_{i=1}^{2} \sum_{j=1}^{2} y_{ij} \log \mu_{ij}, \tag{10.7}$$

and the residual deviance is

$$D(y, \hat{\mu}) = \sum_{i=1}^{2} \sum_{j=1}^{2} y_{ij} \log \frac{y_{ij}}{\hat{\mu}_{ij}}, \tag{10.8}$$

after ignoring terms not involving $\hat{\mu}_{ij}$. Estimating μ_{ij} by maximizing the log-likelihood for the multinomial distribution requires the extra condition $\sum_i \sum_j \mu_{ij} = m$ to ensure that the grand total is fixed at $\sum_i \sum_j y_{ij} = m$ as required by the sampling scheme.

Notice the similarity between the log-likelihood for the Poisson (10.3) and multinomial (10.7) distributions: the first term in (10.3) is the extra condition to ensure the grand total is fixed, and the second term is identical to (10.7). The residual deviance is exactly the same for the Poisson (10.4) and multinomial (10.7) distributions, after ignoring terms not involving μ_{ij}. These similarities for the multinomial and Poisson distributions have one fortunate implication: even though the multinomial distribution is the appropriate

probability model, a Poisson GLM can be used to model the data under appropriate conditions. When the grand total is fixed, the appropriate condition is that the constant term β_0 must appear in the linear predictor, because this ensures $\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{\mu}_{ij} = m$ (Problem 10.2). The effect of including the constant term in the model is that all inferences are conditional on the grand total. The Poisson model, conditioning on the grand total, is equivalent to a multinomial model. Thus, *a Poisson model is still an appropriate model for the randomness, provided the constant term is in the model.*

10.4.3.4 The Column (or Row) Totals Are Fixed

A third scenario that may have produced the data in Table 10.3 assumes that the column (or row) totals are fixed. For example, the researchers may have decided to survey 480 low income people and 414 high income people, then record their attitudes towards GM foods. In this case, the totals in each column are fixed and the counts again have restrictions. For example, the number of high income earners *against* GM foods is known once the number of high income earners *in favour* of GM foods is known.

A multinomial distribution applies separately within each column of the table, because the numbers in each column are fixed and not random. Assuming the counts in each column are independent, the probability function is

$$\begin{aligned}&\mathcal{P}(y_{11}, y_{12}, y_{21}, y_{22}; \mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})\\ &= \overbrace{\frac{m_{\bullet 1}!}{y_{11}!y_{21}!}\left(\frac{\mu_{11}}{m_{\bullet 1}}\right)^{y_{11}}\left(\frac{\mu_{21}}{m_{\bullet 1}}\right)^{y_{21}}}^{\text{For column 1}}\\ &\quad\times\underbrace{\frac{m_{\bullet 2}!}{y_{12}!y_{22}!}\left(\frac{\mu_{12}}{m_{\bullet 2}}\right)^{y_{12}}\left(\frac{\mu_{22}}{m_{\bullet 2}}\right)^{y_{22}}}_{\text{For column 2}}\end{aligned} \tag{10.9}$$

where $m_{\bullet j}$ is the total of column j. The log-likelihood function is

$$\ell(\mu; y) = \sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij}\log\mu_{ij}, \tag{10.10}$$

when terms not involving the parameters μ_{ij} are ignored. To solve for the parameters μ_{ij}, the extra constraints $\sum_{i=1}^{2}\mu_{i1} = m_{\bullet 1}$ and $\sum_{i=1}^{2}\mu_{i2} = m_{\bullet 2}$ must also be added to ensure both column totals are fixed.

Again, notice the similarity between the log-likelihood (10.10) and the log-likelihood for the Poisson (10.3). The residual deviances are exactly the same, after ignoring terms not involving μ_{ij}. This means the Poisson distribution

can be used to model the data, provided the coefficients corresponding to the row totals appear in the linear predictor, since this ensures

$$m_{\bullet 2} = \sum_{i=1}^{2} y_{i2} = \sum_{i=1}^{2} \hat{\mu}_{i2}.$$

Requiring β_0 in the model also ensures that $\sum y_{i1} = \sum_{i=1}^{2} \hat{\mu}_{i1}$ also, and so the row totals are fixed.

Similarly, if the column totals are fixed, a Poisson GLM is appropriate if the coefficients corresponding to the column totals are in the model. If both the row and column totals are fixed, a Poisson GLM is appropriate if the coefficients corresponding to the row and column totals are in the linear predictor.

These general ideas can be extended to larger tables. In general, a Poisson GLM can be fitted to contingency table data provided the coefficients in the linear predictor corresponding to fixed margins are included in the linear predictor.

10.4.4 Three-Dimensional Tables

10.4.4.1 Introduction

Three-dimensional tables cross-classify subjects according to three factors, say A, B and C. If the factors have I, J and K levels respectively, the table is an $I \times J \times K$ table. As an example, the entries in a $3 \times 2 \times 2$ table are defined as shown in Table 10.2, where y_{ijk} refers to the observed count in row i ($i = 1, 2, \ldots I$) and column j ($j = 1, 2, \ldots J$) for group k ($k = 1, 2, \ldots K$); μ_{ijk} refers to the expected count in cell (i, j, k); and $\pi_{ijk} = \mu_{ijk}/m$ refers to the expected probability that an observation is in cell (i, j, k). In other words, Factor A has I levels, Factor B has J levels, and Factor C has K levels (Table 10.4).

Table 10.4 The $3 \times 2 \times 2$ contingency table. The cell count y_{ijk} corresponds to level i of A, level j of B and level k of C (Sect. 10.4.4)

	C_1			C_2					
	B_1	B_2	Total	B_1	B_2	Total	Total B_1	Total B_2	Total
A_1	y_{111}	y_{121}	$m_{1\bullet 1}$	y_{112}	y_{122}	$m_{1\bullet 2}$	$m_{11\bullet}$	$m_{12\bullet}$	$m_{1\bullet\bullet}$
A_2	y_{211}	y_{221}	$m_{2\bullet 1}$	y_{212}	y_{222}	$m_{2\bullet 2}$	$m_{21\bullet}$	$m_{22\bullet}$	$m_{2\bullet\bullet}$
A_3	y_{311}	y_{321}	$m_{3\bullet 1}$	y_{312}	y_{322}	$m_{3\bullet 2}$	$m_{31\bullet}$	$m_{32\bullet}$	$m_{3\bullet\bullet}$
Total	$m_{\bullet 11}$	$m_{\bullet 21}$	$m_{\bullet\bullet 1}$	$m_{\bullet 12}$	$m_{\bullet 22}$	$m_{\bullet\bullet 2}$	$m_{\bullet 1\bullet}$	$m_{\bullet 2\bullet}$	m

Table 10.5 The kidney stone data. The success rates of two methods are given by size; S means a success, and F means a Failure (Example 10.4)

	Small stones			Large stones					
	S	F	Total	S	F	Total	Total S	Total F	Total
Method A	81	6	87	192	71	263	273	77	350
Method B	234	36	270	55	25	80	289	61	350
Total	315	42	357	247	96	343	562	138	700

The meaning of the main effect terms in a Poisson GLM has been discussed in the two-dimensional context: the main effect terms model the marginal totals. Scientific interest focuses on the interactions between the factors. The model with main-effects only acts as the base model for contingency tables against which interaction models are compared. In a three-dimensional table, three two-factor interactions are possible, as well as an interaction term with all three factors. Different interpretations exist depending on which interaction terms appear in the final model. These interpretations are considered in this section. We now introduce the example data to be used.

Example 10.4. The example data in this section (Table 10.5; data set: `kstones`) comes from a study of treatments for kidney stones [8, 24], comparing the success rates of various methods for small and large kidney stones.

```
> data(kstones); str(kstones)
'data.frame':        8 obs. of  4 variables:
 $ Counts : int  81 6 234 36 192 71 55 25
 $ Size   : Factor w/ 2 levels "Large","Small": 2 2 2 2 1 1 1 1
 $ Method : Factor w/ 2 levels "A","B": 1 1 2 2 1 1 2 2
 $ Outcome: Factor w/ 2 levels "Failure","Success": 2 1 2 1 2 1 2 1
```

We treat the method as factor A, the kidney stone size as factor B, and the outcome (success or failure) as factor C.

Note that 350 patients were selected for use with each method. Since this marginal total is fixed, the corresponding main effect term `Method` must appear in the Poisson GLM. The Poisson GLM with all three main effect terms ensures all the marginal totals from the original table are retained, but the parameters themselves are of little interest. □

10.4.4.2 Mutual Independence

If A, B and C are independent, then $\pi_{ijk} = \pi_{i\bullet\bullet} \times \pi_{\bullet j\bullet} \times \pi_{\bullet\bullet k}$ so that, on a log-scale,

$$\log \mu_{ijk} = \log m + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k},$$

using that $\mu_{ijk} = m\pi_{ijk}$. This is called *mutual independence*. As seen for the two-dimensional tables, including the main effect terms effectively ensures the marginal totals are preserved. If the mutual independence model is appropriate, then the table may be understood from just the marginal totals.

For the kidney stone data, the mutual independence model states that the success or failure is independent of the method used, and independent of the size of the kidney stones, and that the method used is also independent of the size of the kidney stone. Adopting this model assumes the data can be understood for each variable separately. In other words, equal proportions of patients are in each method; $138/700 = 19.7\%$ of all treatments fail; and $343/700 = 49.0\%$ of patients have large kidney stones. Fit the model using:

```
> ks.mutind <- glm( Counts ~ Size + Method + Outcome,
                    family=poisson, data=kstones)
```

In this section, we will fit the models then comment and compare the models after all the models are fitted.

10.4.4.3 Partial Independence

Suppose A and B are not independent, but both are independent of C; then $\pi_{ijk} = \pi_{ij\bullet} \times \pi_{\bullet\bullet k}$, or $\log \mu_{ijk} = \log m + \log \pi_{ij\bullet} + \log \pi_{\bullet\bullet k}$ on a log-scale. Since A and B are not independent, $\pi_{ij\bullet} \neq \pi_{i\bullet\bullet} \times \pi_{\bullet j\bullet}$. To ensure that the marginal totals are preserved, the main effects are also included in the model (along the lines of the marginality principle; Sect. 2.10.4). This means that the model

$$\log \mu_{ijk} = \log m + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k} + \log \pi_{ij\bullet}$$

is suggested. This systematic component has one two-factor interaction $A.B$. This is called *partial independence* (or *joint independence*). If a partial independence model is appropriate, then the two-way tables for each level of C are multiples of each other, apart from randomness. The data can be understood by combining the tables over C.

For the kidney stone data, we can fit all three models that have one of the two-factor interactions:

```
> ks.SM <- glm( Counts ~ Size * Method + Outcome,
                family=poisson, data=kstones )
> ks.SO <- update(ks.SM, . ~ Size * Outcome + Method)
> ks.OM <- update(ks.SM, . ~ Outcome * Method + Size)
```

10.4.4.4 Conditional Independence

Suppose that A and B are independent of each other when considered separately for each level of C. Then the probabilities π_{ijk} are independent

conditional on the level of k, when $\pi_{ij|k} = \pi_{i\bullet|k} \times \pi_{\bullet j|k}$. Each conditional probability can be written in terms of marginal totals:

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{\bullet\bullet k}}; \quad \pi_{i\bullet|k} = \frac{\pi_{i\bullet k}}{\pi_{\bullet\bullet k}}; \quad \pi_{\bullet j|k} = \frac{\pi_{\bullet jk}}{\pi_{\bullet\bullet k}},$$

so that $\pi_{ijk} = (\pi_{i\bullet|k} \times \pi_{\bullet j|k})\pi_{\bullet\bullet k} = \pi_{i\bullet k}\pi_{\bullet jk}/\pi_{\bullet\bullet k}$ hold. In other words, $\log \mu_{ijk} = \log m + \log \pi_{i\bullet k} + \log \pi_{\bullet jk} - \log \pi_{\bullet\bullet k}$ on a log-scale. To ensure the marginal totals are preserved, use the model

$$\log \mu_{ijk} = \log m + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k} + \log \pi_{i\bullet k} + \log \pi_{\bullet jk}$$

which includes the main effects. The systematic component has the two two-factor interactions $A.C$ and $B.C$. This is called *conditional independence*.

If a conditional independence model is appropriate, then each two-way table for each level of C considered separately shows independence between A and B. The data can be understood by creating separate tables involving factors A and B, one for each level of C.

The three models with two of the two-factor interactions are:

```
> ks.noMO <- glm( Counts ~ Size * (Method + Outcome),
                  family=poisson, data=kstones )
> ks.noOS <- update(ks.noMO, . ~ Method * (Outcome + Size) )
> ks.noMS <- update(ks.noMO, . ~ Outcome * (Method + Size) )
```

10.4.4.5 Uniform Association

Consider the case where all three two-factor interactions are present but the three-factor interaction $A.B.C$ only is absent. This means that each two-factor interaction is unaffected by the level of the third factor. No interpretation in terms of independence or through the marginal totals is possible. The model is

$$\log \mu_{ijk} = \log m + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k} + \log \pi_{i\bullet k} + \log \pi_{\bullet jk} + \log \pi_{ij\bullet}$$

which contains all two-way interactions. This is called *uniform association*. If the uniform association model is appropriate, then the data can be understood by examining all three individual two-way tables. For the kidney stone data the model with all of the two-factor interactions is:

```
> ks.no3 <- glm( Counts ~ Size*Method*Outcome - Size:Method:Outcome,
                 family=poisson, data=kstones )
```

Uniform association is simple enough to define from a mathematical point of view, but is often difficult to interpret from a scientific point of view.

10.4.4.6 The Saturated Model

If all interaction terms are necessary in the linear predictor, the model is the saturated model

$$\log \mu_{ijk} = \log m + \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k} + \log \pi_{i\bullet k} + \log \pi_{\bullet jk} + \log \pi_{ij\bullet} + \log \pi_{ijk}$$

which includes all interactions. The model has zero residual deviance (in computer arithmetic) and zero residual degrees of freedom. In other words, the model produces a perfect fit:

```
> ks.all <- glm( Counts ~ Size * Method * Outcome,
                 family=poisson, data=kstones )
> c( deviance( ks.all ), df.residual(ks.all) )
[1] -2.930989e-14  0.000000e+00
```

This means that there are as many parameter estimates as there are cells in the table, and so the data cannot be summarized using a smaller set of coefficients. If the saturated model is appropriate, then the data cannot be presented in a simpler form than giving the original $I \times J \times K$ table.

10.4.4.7 Comparison of Models

For the kidney stone data the saddlepoint approximation is sufficiently accurate since $\min\{y_i\} \geq 3$. This means that goodness-of-fit tests can be used to examine and compare the models (Table 10.6). The mutual independence model and partial independence models are not appropriate, as the residual deviance far exceeds the residual degrees of freedom. Model `ks.noMO` appears the simplest suitable model. This implies that the data are best understood by creating separate tables for large and small kidney stones, but small and large kidney stones data should not be combined.

10.4.5 Simpson's Paradox

Understanding which interaction terms are necessary in a log-linear model has important implications for condensing the tabular data. If a table is collapsed over a factor incorrectly, incorrect and misleading conclusions may be reached. An extreme example of this is *Simpson's paradox*. To explain, consider the kidney stones data (Table 10.5). The most suitable model appears to be model `ks.noMO` (Table 10.6). This model has two two-factor interactions, indicating conditional independence between `Outcome` and `Method`, depending on the `Size` of the kidney stones. The dependence on `Size` means that the data must be stratified by kidney stone size for the correct relationship between `Method` and `Outcome` to be seen. Combining the data over `Size`s, and

Table 10.6 The fitted values for all Poisson GLMs fitted to the kidney stone data. Model `ks.noMO` is the selected model and is flagged * (Sect. 10.4.4)

	Mutual independence	Partial independence			Conditional independence			Uniform association	Saturated model
Count	ks.mutind	ks.SM	ks.SO	ks.OM	ks.noMO *	ks.noOS	ks.noMS	ks.no3	ks.all
81	143.3	69.8	157.5	139.2	76.8	67.9	153.0	79.0	81
6	35.2	17.2	21.0	39.3	10.2	19.1	23.4	8.0	6
234	143.3	216.8	157.5	147.4	238.2	222.9	162.0	236.0	234
36	35.2	53.2	21.0	31.1	31.8	47.1	18.6	34.0	36
192	137.7	211.2	123.5	133.8	189.4	205.1	120.0	194.0	192
71	33.8	51.8	48.0	37.7	73.6	57.9	53.6	69.0	71
55	137.7	64.2	123.5	141.6	57.6	66.1	127.0	53.0	55
25	33.8	15.8	48.0	29.9	22.4	13.9	42.4	27.0	25
Res. dev.:	234.4	33.1	204.8	232.1	3.5	30.8	202.4	1.0	0
Res. df:	4	3	3	3	2	2	2	1	0
G-o-F P:	0.00	0.00	0.00	0.00	0.18	0.00	0.00	0.32	1.00

hence considering a single combined two-way table of `Method` and `Outcome` (and hence ignoring `Size`), is an incorrect summary. To demonstrate, consider *incorrectly* collapsing the contingency table over `Size`. First, use `xtabs()` to create a suitable three-dimensional table of counts:

```
> ks.tab <- xtabs(Counts ~ Method + Outcome + Size, data=kstones)
> ks.tab
, , Size = Large

      Outcome
Method Failure Success
     A      71     192
     B      25      55

, , Size = Small

      Outcome
Method Failure Success
     A       6      81
     B      36     234
```

Then sum over `Size`, which is the third dimension:

```
> MO.tab <- apply( ks.tab, c(1, 2), sum) # Sums over the 3rd dimension
> MO.tab    # An *incorrect* collapsing of the data
      Outcome
Method Failure Success
     A      77     273
     B      61     289
```

The table suggests that Method B has a higher success rate than Method A:

```
> prop.table(MO.tab, 1) # Compute proportions in each row (dimension 1)
      Outcome
Method   Failure   Success
     A 0.2200000 0.7800000
     B 0.1742857 0.8257143
```

The overall success rate for Method A is about 78%, and for Method B the success rate is about 83%, so we would prefer Method B. However, recall that the table `MO.tab` is *incorrectly* collapsed over `Size`: the conditional independence suggest the relationship between `Method` and `Outcome` should be examined *separately* for each level of `Size`.

Consequently, now examine the two-way table for large and small kidney stones separately:

```
> MO.tab.SizeLarge <- ks.tab[, , "Large"] # Select Large stones
> prop.table(MO.tab.SizeLarge, 1)  # Compute proportions in each row
      Outcome
Method  Failure  Success
     A 0.269962 0.730038
     B 0.312500 0.687500
```

For large kidney stones, the success rate for Method A is about 73%, and for Method B the success rate is about 69% so we would prefer Method A.

```
> MO.tab.SizeSmall <- ks.tab[, , "Small"] # Select Small stones
> prop.table(MO.tab.SizeSmall, 1)  # Compute proportions in each row
      Outcome
Method    Failure    Success
     A 0.06896552 0.93103448
     B 0.13333333 0.86666667
```

For small kidney stones, the success rate for Method A is about 93%, and for Method B the success rate is about 87%, so we would prefer Method A.

In this example, incorrectly collapsing the table over `Size` has completely changed the conclusion. Ignoring `Size`, Method B has a higher overall success rate, but Method A actually has a higher success rate for both small and large kidney stones. This is called *Simpson's paradox*, which is a result of incorrectly collapsing a table.

To explain the apparent paradox, first notice that the large kidney stone group reported a far lower success rate for both methods compared to the small kidney stone group. Since Method A was used on a larger proportion of patients with large kidney stones, Method A reports a high number of total failures when the two groups are combined. In contrast, Method B was used on a larger proportion of patients with small kidney stones, where the success rate for both methods is better, and so Method B reports a smaller number of total failures.

10.4.6 Equivalence of Binomial and Poisson GLMs

In many contingency table contexts, interest focuses on explaining one of the factors in terms of the others. When the response factor of interest takes two levels, interest focuses on explaining the proportion of responses that are allocated to each of the two levels. In this case, there is a binomial GLM with the logistic link that is equivalent to the Poisson log-linear model. The reason is that for large m and small proportions, the binomial distribution approaches the Poisson distribution. To see this, write the probability of a success in the binomial distribution as π. Then, the variance function for the *number* of successes using the binomial model is $V(\pi) = m\pi(1-\pi)$. When π is small and m is large, $V(\pi) = m\pi(1-\pi) \to m\pi$. This is equivalent to the variance of the Poisson distribution. This means that the binomial distribution approaches the Poisson distribution for large m and small π.

For example, consider the data of Table 10.3 (p. 379) relating GM attitude to income. Here interest focuses on whether income level affects GM attitude, so the data could be equally well analysed in R by treating `Att` as the response variable:

```
> y <- ifelse(Att == "Against", 1, 0)
> gm.bin <- glm(y~Inc, family=binomial, weights=Counts)
> anova(gm.bin, test="Chisq")
     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     3     1214.7
Inc   1   8.7686         2     1206.0 0.003065 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The deviance goodness-of-fit test for `Inc` is identical to the test for `Att:Inc` interaction given in Sect. 10.4.3.2, with the same P-value and the same interpretation. The odds of being against GM foods are nearly 50% greater for low-income respondents:

```
> coef(summary(gm.bin))
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -0.5548742  0.1021018 -5.434518 5.494476e-08
IncLow       0.4045920  0.1371323  2.950378 3.173854e-03
> exp(coef(gm.bin)["IncLow"])
  IncLow
1.498691
```

Example 10.5. For the kidney stones data (Table 10.5; data set: `kstones`), interest may focus on comparing the success rates of the two methods. From this point of view, the data may be analysed via a binomial GLM:

```
> y <- ifelse(kstones$Outcome=="Success", 1, 0)
> ks.bin <- glm(y~Size*Method, family=binomial,
              weights=Counts, data=kstones)
> anova(ks.bin, test="Chisq")
            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           7     694.98
Size         1  29.6736         6     665.31 5.113e-08 ***
Method       1   2.4421         5     662.87    0.1181
Size:Method  1   1.0082         4     661.86    0.3153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of deviance shows that success depends strongly on the size of the kidney stones (better success for small stones), but there is no evidence for any difference between the two methods, either overall or separately for small or large stones. This conclusion agrees with the contingency table analysis, which concluded that `Outcome` was conditionally independent of `Method` given `Size`. The contingency table model `ks.noMO` contains the additional information that `Method` is associated with `Size`. Indeed it is clear from Table 10.5 that Method A is predominately used for large stones and Method B for small stones. Whether the ability to test for associations between explanatory factors, provided by the contingency table analysis, is of interest depends on the scientific context. For these data, the choice of method is likely made based on established hospital protocols, and hence would be known before the data were collected. □

10.4.7 Higher-Order Tables

Extending these ideas to situations with more than three factors is easy in practice using R, though interpreting the final models is often difficult.

Example 10.6. A study of seriously emotionally disturbed (SED) and learning disabled (LD) adolescents [19, 29] reported their depression levels (Table 10.7; data set: `dyouth`). The data are counts classified by four factors: `Age` (using `12-14` as the reference group), `Group` (either `LD` or `SED`), `Gender` and level of `Depression` (either low `L` or high `H`). Since none of the totals were fixed beforehand and are free to vary randomly, no variables *need* to be included in the model. With four factors, $\binom{4}{2} = 6$ two-factor interactions, $\binom{4}{3} = 4$ three-factor interactions and one four-factor interaction are potentially in the model. As usual, the main-effect terms are included in the model to ensure the marginal totals are preserved.

Table 10.7 Depression levels in youth (Example 10.6)

		Depression low `L`		Depression high `H`	
`Age`	`Group`	Males	Females	Males	Females
12-14	LD	79	34	18	14
	SED	14	5	5	8
15-16	LD	63	26	10	11
	SED	32	15	3	7
17-18	LD	36	16	13	1
	SED	36	12	5	2

The most suitable model for the data [11] (Problem 10.8) appears to be:

```
> data(dyouth)
> dy.m1 <- glm( Obs ~ Age*Depression*Gender + Age*Group,
               data=dyouth, family=poisson)
> anova(dy.m1, test="Chisq")
                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                    23     368.05
Age                   2   11.963        21     356.09  0.002525 **
Depression            1  168.375        20     187.71 < 2.2e-16 ***
Gender                1   58.369        19     129.34 2.172e-14 ***
Group                 1   69.104        18      60.24 < 2.2e-16 ***
Age:Depression        2    3.616        16      56.62  0.163964
Age:Gender            2    3.631        14      52.99  0.162718
Depression:Gender     1    7.229        13      45.76  0.007175 **
Age:Group             2   27.090        11      18.67 1.311e-06 ***
Age:Depression:Gender 2    8.325         9      10.35  0.015571 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The three-way interaction shows that the relationship between age and depression is different for males and females:

```
> Males   <- subset(dyouth, Gender=="M")
> Females <- subset(dyouth, Gender=="F")
> table.M <- prop.table( xtabs(Obs~Age+Depression, data=Males),   1)
> table.F <- prop.table( xtabs(Obs~Age+Depression, data=Females), 1)
> round(table.F * 100)   # FEMALES
       Depression
Age      H  L
  12-14 36 64
  15-16 31 69
  17-18 10 90
> round(table.M * 100)   # MALES
       Depression
Age      H  L
  12-14 20 80
  15-16 12 88
  17-18 20 80
```

Given the fitted model, collapsing the table into a simpler table would be misleading. The proportion tables show that the rate of high depression decreases with age for girls, especially for 17 years and older, whereas for males the rate of high depression decreases at age 15–16 then increases again for 17–18. This difference in pattern explains the three-way interaction detected by the analysis of deviance table.

The model also finds a significant interaction between `Age` and `Group`, meaning simply that the SED and LD groups contain different proportions of the age groups. This is not particularly of interest, but it is important to keep the `Age:Group` term in the model, so that the tests for interactions involving `Depression` should adjust for these demographic proportions.

Overall, the model shows an association between depression and age and gender, but no difference in depression rates between the two groups once the demographic variables have been taken into account. □

10.4.8 Structural Zeros in Contingency Tables

Contingency tables may contain cells with zero counts. Depending on the reason for a zero count, different approaches must be taken when modelling.

Sampling zeros or *random zeros* appear by chance, simply because no observations occurred in that category. Larger samples may produce non-zero counts in those cells. Computing fitted values for these cells is sensible; they are legitimate counts to be modelled like the other counts in the data. However, the presence of the zeros means the saddlepoint approximation is likely to be very poor. As a result, levels of one or more factors may be combined to increase the minimum count. For example, 'Strongly agree' and 'Agree' may be combined sensibly into a single 'Agreement' category.

Structural zeros appear because the outcome is impossible. For example, in a cross-tabulation of gender and surgical procedures, the cell corresponding to male hysterectomies *must* contain a zero count. Producing fitted values for these cells makes no sense. Structural zeros are not common in practice.

Structural zeros require special attention since computing expected counts for impossible events is nonsense. As a result, cells containing structural zeros are removed from the data before analysis.

Example 10.7. The types of cancer diagnosed in Western Australia in 1996 were recorded for males and females (Table 10.8; data set: `wacancer`) to ascertain whether the number of cancers differs between genders [20].

Three cells have zeros recorded. Two of these three cells are *structural zeros* since they are impossible—females cannot have prostate cancer, and males cannot have cervical cancer. Breast cancer is a possible, but very rare, disease among men (about 100 times as many cases in females compared to males, in the USA [34, Table 1]). The zero for male breast cancer is technically

Table 10.8 The number of cancers diagnosed by gender in Western Australia during 1996 (Example 10.7)

	Cancer type						
Gender	Prostate	Breast	Colorectal	Lung	Melanoma	Cervix	Other
Males	923	0	511	472	362	0	1406
Females	0	875	355	211	282	77	1082

a *sampling* zero. Since breast cancer is already *known* to be a rare disease for males, the analysis should focus on gender differences for other types of cancers, such as colorectal, lung, melanoma and other cancers.

To begin, we fit a model ignoring these complications:

```
> data(wacancer)
> wc.poor <- glm( Counts ~ Cancer*Gender, data=wacancer, family=poisson )
> anova( wc.poor, test="Chisq")
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            13     6063.7
Cancer        6   3281.5         7     2782.2 < 2.2e-16 ***
Gender        1     95.9         6     2686.2 < 2.2e-16 ***
Cancer:Gender 6   2686.2         0        0.0 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To compare, we now remove breast cancer, male cervical cancer and female prostate cancer from the analysis, and refit:

```
>    # Omit necessary cells of table:
> wc <- subset(wacancer, (Cancer!="Breast"))
> wc <- subset(wc, !(Cancer=="Cervix" & Gender=="M"))
> wc <- subset(wc, !(Cancer=="Prostate" & Gender=="F"))
> xtabs(Counts~Gender+Cancer, data=wc)  # Table *looks* similar
      Cancer
Gender Breast Cervix Colorectal Lung Melanoma Other Prostate
     F      0     77        355  211      282  1082        0
     M      0      0        511  472      362  1406      923
>    # Now fit the model
> wc.m1 <- glm( Counts ~ Cancer*Gender, data=wc, family=poisson )
> anova( wc.m1, test="Chisq")
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                            9    2774.32
Cancer        5  2591.47         4     182.85 < 2.2e-16 ***
Gender        1   144.74         3      38.11 < 2.2e-16 ***
Cancer:Gender 3    38.11         0       0.00  2.68e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An alternative to explicitly removing these observations from the table is to set the corresponding prior weights `weights` to zero for these observations, and to one for other observations. Even though the prior weighs are defined to

be positive, R interprets a prior weight of zero to mean that the corresponding observation should be ignored in the analysis.

For both models, the interaction term is very significant, so the number of people diagnosed with the different types of cancers differs according to gender, even after eliminating prostate, breast and cervical cancer, which are obviously gender-linked. However, note that the degrees of freedom are different for the two models. □

10.5 Overdispersion

10.5.1 Overdispersion for Poisson GLMs

For a Poisson distribution, $\text{var}[y] = \mu$. However, in practice the apparent variance of the data often exceeds μ. This is called *overdispersion*, as has already been discussed for binomial GLMs (Sect. 9.8). Underdispersion also occurs, but is less common.

Overdispersion arises either because the mean μ retains some innate variability, even when all the explanatory variables are fixed, or because the events that are being counted are positively correlated. Overdispersion typically arises because the events being counted arise in clusters or are mutually supporting in some way. This causes the underlying events to be positively correlated, and overdispersion of the counts is the result.

The presence of overdispersion might or might not affect the parameter estimates $\hat{\beta}_j$, depending on the nature of the overdispersion, but the standard errors $\text{se}(\hat{\beta}_j)$ are necessarily underestimated. Consequently, tests on the explanatory variables will generally appear to be more significant that warranted by the data, and confidence intervals for the parameters will be narrower than warranted by the data.

Overdispersion is detected by conducting a goodness-of-fit test (as described in Sect. 7.4). If the residual deviance and Pearson goodness-of-fit statistics are much larger than the residual degrees of freedom, then either the fitted model is inadequate or the data are overdispersed. If lack of fit remains even after fitting the maximal possible explanatory model, and after eliminating any outliers, then overdispersion is the alternative explanation.

When the counts are very small, so asymptotic approximations to the residual deviance and Pearson statistics are suspect (Sect. 7.5, p. 276), then overdispersion may be difficult to judge. However the goodness-of-fit statistics are more likely to be underestimated than overestimated in small count situations, so large goodness-of-fit statistics should generally be taken to indicate lack of fit.

Table 10.9 The number of membrane pock marks at various dilutions of the viral medium (Example 10.9)

Dilution	Pock counts									
1	116	151	171	194	196	198	208	259		
2	71	74	79	93	94	115	121	123	135	142
4	27	33	34	44	49	51	52	59	67	92
8	8	10	15	22	26	27	30	41	44	48
16	5	6	7	7	8	9	9	9	11	20

Example 10.8. For the final model fitted to the kidney stone data (see Table 10.6), the residual deviance was 3.5 and the residual df was 2. A goodness-of-fit test does not reject the hypothesis that the model is adequate:

```
> pchisq(deviance(ks.noMO), df.residual(ks.noMO), lower.tail=FALSE)
[1] 0.1781455
```

□

Example 10.9. In an experiment [35] to assess viral activity, pock marks were counted at various dilutions of the viral medium (Table 10.9; data set: `pock`). We use the logarithm to base 2 of `Dilution` as a covariate, since the dilution levels are in increasing powers of 2 suggesting this was factored into the design. A plot of the data shows a definite relationship between the variables (Fig. 10.3, left panel), and that the variance increases with increasing mean (Fig. 10.3, right panel):

```
> data(pock)
> plot( Count ~ jitter(log2(Dilution)), data=pock, las=1,
    xlab="Log (base 2) of dilution", ylab="Pock mark count")
> mn <- with(pock, tapply(Count, log2(Dilution), mean) )  # Group means
> vr <- with(pock, tapply(Count, log2(Dilution), var) )   # Group variances
> plot( log(vr) ~ log(mn), las=1,
    xlab="Group mean", ylab="Group variance")
```

Intuitively, pock marks are more likely to appear in clusters rather than independently, so overdispersion would not be at all surprising. Indeed, the sample variance is much larger than the mean for each group, clear evidence of overdispersion:

```
> data.frame(mn, vr, ratio=vr/mn)
       mn         vr    ratio
0 186.625 1781.12500 9.543871
1 104.700  667.34444 6.373872
2  50.800  360.40000 7.094488
3  27.100  194.98889 7.195162
4   9.100   17.65556 1.940171
```

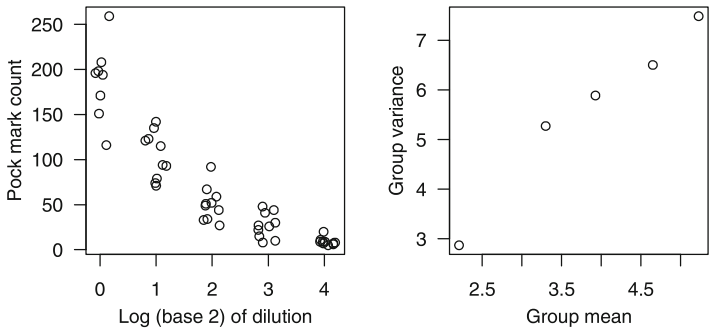

Fig. 10.3 The pock data. Left panel, the counts against the logarithm of dilution; right panel: the logarithm of the group variances against the logarithm of the group means (Example 10.9)

Not only are the variances greater than the means, but their ratio increases with the mean as well. The slope of the trend in the right panel of Fig. 10.3 is about 1.5:

```
> coef(lm(log(vr)~log(mn)))
(Intercept)     log(mn)
 0.02861162  1.44318666
```

This suggests a variance function approximately of the form $V(\mu) = \mu^{1.5}$. The mean–variance relationship here is in some sense intermediate between that for the Poisson ($V(\mu) = \mu$) and gamma ($V(\mu) = \mu^2$) distributions.

Fitting a Poisson GLM shows substantial lack of fit, as expected:

```
> m1 <- glm( Count ~ log2(Dilution), data=pock, family=poisson )
> X2 <- sum(residuals(m1, type="pearson")^2)
> c(Df=df.residual(m1), Resid.Dev=deviance(m1), Pearson.X2=X2)
       Df  Resid.Dev Pearson.X2
  46.0000   290.4387   291.5915
```

The saddlepoint approximation is satisfactory here as $\min\{y_i\} = 5$ is greater than 3. Indeed, the deviance and Pearson goodness-of-fit statistics are nearly identical. Two ways to model the overdispersion are discussed in Sects. 10.5.2 and 10.5.3. □

10.5.2 Negative Binomial GLMs

One way to model overdispersion is through a hierarchical model. Instead of assuming $y_i \sim \text{Pois}(\mu_i)$, we can add a second layer of variability by allowing μ_i itself to be a random variable. Suppose instead that

$$y_i|\lambda_i \sim \text{Pois}(\lambda_i) \quad \text{and} \quad \lambda_i \sim \text{G}(\mu_i, \psi)$$

where $\mathrm{G}(\mu_i, \psi)$ denotes a distribution with mean μ_i and coefficient of variation ψ. For example, we could imagine that the number of pock marks recorded in the pock data (Example 10.9) might follow a Poisson distribution for any given viral concentration, but that the viral concentration varies somewhat between replicates for any given dilution with a coefficient of variation ψ. It is straightforward to show, under the hierarchical model, that

$$\mathrm{E}[y_i] = \mu_i \quad \text{and} \quad \mathrm{var}[y_i] = \mu_i + \psi\mu_i^2,$$

so the variance contains an overdisperion term $\psi\mu_i^2$. The larger ψ, the greater the overdispersion.

A popular choice is to assume that the mixing distribution G is a gamma distribution. The coefficient of variation of a gamma distribution is its dispersion parameter, so the second layer of the hierachical model becomes $\lambda_i \sim \mathrm{Gam}(\mu_i, \psi)$. With this assumption, is it possible to show that y_i follows a *negative binomial distribution* with probability function

$$\mathcal{P}(y_i; \mu_i, k) = \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1)\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + k}\right)^{k}, \qquad (10.11)$$

where $k = 1/\psi$ and $\Gamma()$ is the gamma function, so that $\mathrm{var}[y_i] = \mu_i + \mu_i^2/k$.

For any fixed value of k, it can be shown (Problem 10.1) that the negative binomial distribution is an EDM with unit deviance

$$d(y, \mu) = 2\left\{y \log \frac{y}{\mu} - (y + k) \log \frac{y + k}{\mu + k}\right\},$$

where the limit form (5.14) is used if $y = 0$. Hence the negative binomial distribution can be used to define a GLM for any given k. Note that negative binomial EDMs have dispersion $\phi = 1$, as do all EDMs for count data, because $\mathrm{var}[y_i]$ is determined by μ_i and k. In practice, k is rarely known and so negative binomial GLMs are usually used with an estimated value for k. In R, the function `glm.nb()` from package **MASS** can be used in place of `glm()` to fit the model. The function `glm.nb()` undertakes maximum likelihood estimation for both k and the GLM coefficients β_j simultaneously (see `?glm.nb`).

The estimation of k introduces an extra layer of uncertainty into a negative binomial GLM. However the maximum likelihood estimator $\hat{k}$ of k is uncorrelated with the $\hat{\beta}_j$, according to the usual asymptotical approximations. Hence the GLM fit tends to be relatively stable with respect to estimation of k.

Negative binomial GLMs give larger standard errors than the corresponding Poisson GLMs, depending on the size of $k = 1/\psi$. On the other hand, the coefficient estimates $\hat{\beta}_j$ from a negative binomial GLM may be similar to those produced from the corresponding Poisson GLM. The negative binomial GLM gives less weight to observations with large μ_i than does the Poisson GLM, and relatively more weight to observations with small μ_i, so the coefficients

will vary somewhat. Unlike `glm()`, where the default link function for every `family` is the canonical link, the default link function for `glm.nb()` is the logarithmic link function. Indeed the log-link is almost always used with negative binomial GLMs to ensure $\mu > 0$ for any value of the linear predictor. The function `glm.nb()` also allows the `"sqrt"` and `"identity"` link functions.

For negative binomial GLMs, the use of quantile residuals [12] is strongly recommended (Sect. 8.3.4.2).

Example 10.10. The pock data shows overdispersion (Example 10.9; data set: `pock`). We fit a negative binomial GLM, estimating k using the function `glm.nb()` in package **MASS** (note that `glm.nb()` uses `theta` to denote k):

```
> library(MASS)       # Provides the function  glm.nb()
> m.nb <- glm.nb( Count ~ log2(Dilution), data=pock )
> m.nb$theta          # This is the value of k (called theta in MASS)
[1] 9.892894
```

The output object `m.nb` includes information about the estimation of k. The output from `glm.nb()` model is converted to the style of output from `glm()` using `glm.convert()`:

```
> m.nb <- glm.convert(m.nb)
> printCoefmat(coef(summary(m.nb, dispersion=1)))
               Estimate Std. Error z value  Pr(>|z|)
(Intercept)     5.33284    0.08786  60.697 < 2.2e-16 ***
log2(Dilution) -0.72460    0.03886 -18.646 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that we have to specify explicitly that the dispersion parameter is $\phi = 1$, because after using `glm.convert()`, R does not know automatically that the resulting GLM family should have dispersion equal to one.

Since $k \approx 10$, the negative binomial model is using the variance function $V(\mu) \approx \mu + \mu^2/10$. The coefficient of variation of the mixing distribution ($\psi = 1/k$) is estimated to be about 10%, a reasonable level for replicate to replicate variation. Comparing the Poisson and negative binomial models shows that the parameter estimates are reasonably close, but the standard errors are quite different:

```
> printCoefmat( coef( summary(m1)) )        # Poisson glm information
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.2679     0.0226   233.6   <2e-16 ***
log2(Dilution)  -0.6809     0.0154   -44.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The diagnostic plots (Fig. 10.4, top panels) suggest the negative binomial model is adequate. No observations are particularly influential. □

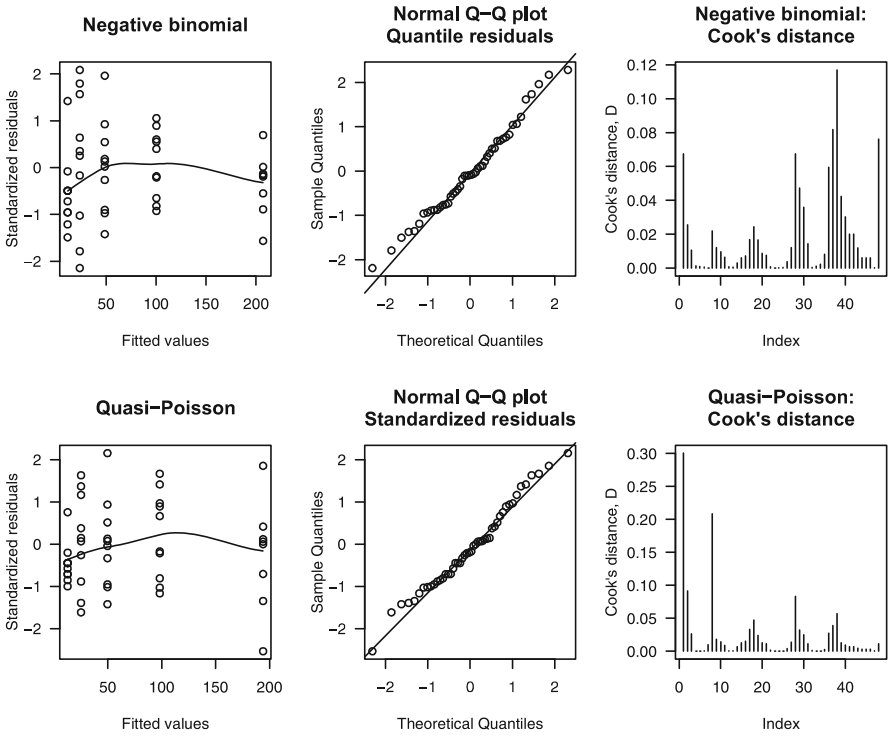

Fig. 10.4 Diagnostic plots from fitting the negative binomial model (top panels) and the quasi-Poisson models (bottom panels) to the pock data (Example 10.9)

10.5.3 Quasi-Poisson Models

The simplest to use, and therefore most commonly used, approach to overdispersed counts are quasi-Poisson models. Quasi-Poisson models keep the Poisson variance function $V(\mu) = \mu$ but simply allow a general positive dispersion parameter ϕ, so that $\text{var}[y_i] = \phi\mu_i$. Here $\phi > 1$ corresponds to overdispersion. This approach can be motivated in the same way as were quasi-binomial models (Sect. 9.8). Suppose that the counts y_i are counts of cases arising from a large population of size N, and the suppose that the individuals in the population are positively correlated. Then $\text{E}[y_i] = \mu_i = N\pi_i$, where π_i is the probability that a random individual is a case, and $\text{var}[y_i] = \phi N\pi_i(1-\pi_i)$ where $\phi = 1 + (N-1)\rho$ and ρ is the correlation between individuals. If N is large and the π_i are small, then $\text{var}[y_i] \approx \phi N\pi_i = \phi\mu_i$.

When $\phi \neq 1$, there is no EDM with this variance function that gives positive probability to integer values of y_i. Nevertheless, the quasi-likelihood methods of Sect. 8.10 still apply, so quasi-Poisson GLMs yield consistent estimators and

consistent standard errors for the β_j, provided only that $\mathrm{E}[y_i]$ and $\mathrm{var}[y_i]$ are correctly specified. Note that quasi-Poisson GLMs reduce to Poisson GLMs when $\phi = 1$.

The coefficient estimates from a quasi-Poisson GLM are identical to those from the corresponding Poisson GLM (since the estimates $\hat{\beta}_j$ do not depend on ϕ), but the standard errors are inflated by a factor of $\sqrt{\phi}$. Confidence intervals and statistics for testing hypotheses tests will change for the same reason.

Note that quasi-Poisson and the negative binomial model both produce overdispersion relative to the Poisson distribution but they assume different mean–variance relationships. Quasi-Poisson models assume a linear variance function ($V(\mu) = \phi\mu$) whereas negative binomial models uses a *quadratic* variance function ($V(\mu) = \mu + \mu^2/k$).

Quasi-Poisson models are fitted in R using `glm()` and specifying `family=quasipoisson()`. As for `family=poisson()`, the default link function is the `"log"` link, while `"identity"` and `"sqrt"` are also permitted. Since the quasi-Poisson model is not based on a probability model, the AIC is undefined. For the same reason, quantile residuals [12] cannot be computed for the quasi-Poisson GLM since no probability model is defined.

Example 10.11. The model fitted to the `pock` data shows overdispersion (Example 10.9), so an alternative solution is to fit a quasi-Poisson model:

```
> m.qp <- glm( Count ~ log2(Dilution), data=pock, family="quasipoisson")
```

The diagnostic plots (Fig. 10.4, bottom panels) suggest the quasi-Poisson model is broadly adequate, and no observations are particularly influential. It is discernible from the left panels of Fig. 10.4, however, that the negative binomial model tends to under-estimate slightly the variances of the low counts while the quasi-Poisson model does the same for large counts.

F-tests are used for model comparisons, since ϕ is estimated. Comparing the standard errors from the quasi-Poisson model to the standard errors produced from the Poisson GLM, the standard errors in the quasi-Poisson model are scaled by $\sqrt{\phi}$:

```
> se.m1 <- coef(summary(m1))[, "Std. Error"]
> se.qp <- coef(summary(m.qp))[, "Std. Error"]
> data.frame(SE.Pois=se.m1, SE.Quasi=se.qp, ratio=se.qp/se.m1)
                  SE.Pois   SE.Quasi    ratio
(Intercept)    0.02255150 0.05677867 2.517733
log2(Dilution) 0.01544348 0.03888257 2.517733
> sqrt(summary(m.qp)$dispersion)
[1] 2.517733
```

Note that quantile residuals can be produced for the negative binomial GLM since a full probability function is defined, but quantile residuals cannot be computed for the quasi-Poisson GLM since no probability model is defined. For this reason, the residual plots for the quasi-Poisson model use standardized deviance residuals. The fitted systematic components are compared in

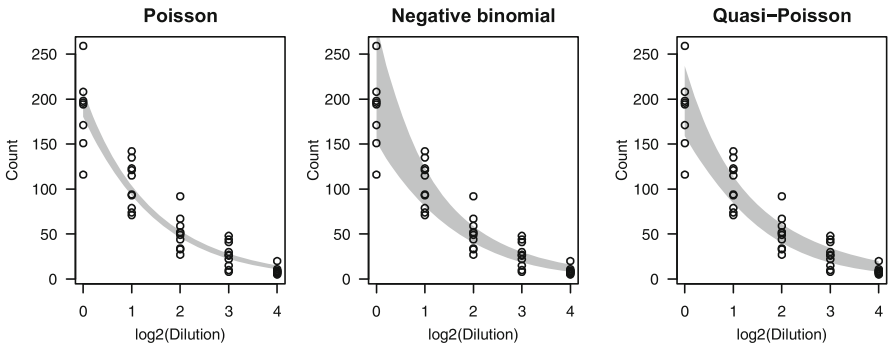

Fig. 10.5 Models fitted to the pock data, including the 99.9% confidence intervals for $\hat{\mu}$ (Example 10.11)

Fig. 10.5. Recall the Poisson and quasi-Poisson models produce identical parameter estimates, and hence fitted values.

```
> coef.mat <- rbind( coef(m1), coef(m.qp), coef(m.nb) )
> rownames(coef.mat) <- c("Poisson glm", "Quasi-Poisson", "Neg bin glm")
> coef.mat
              (Intercept) log2(Dilution)
Poisson glm      5.267932     -0.6809442
Quasi-Poisson    5.267932     -0.6809442
Neg bin glm      5.332844     -0.7245983
```

The plots in Fig. 10.5 show that the different approaches model the randomness differently.

We can now interpret the fitted model. The fitted models say that the expected number of pock marks decreased by a factor of about $\exp(-0.7) \approx 0.5$ for every 2-fold dilution. In other words, the expected number of pock marks is directly proportional to the concentration of the viral medium. □

10.6 Case Study

In a study of nesting female horseshoe crabs [1, 5], each with an attached male, the number of other nearby male crabs (called satellites) were counted (Table 10.10; data set: `hcrabs`). The colour of the female, the condition of her spine, her carapace width, and her weight were also recorded. The purpose of the study is to understand the factors that attract satellite crabs. Are they more attracted to larger females? Does the condition or colour of the female play a role?

Table 10.10 The horseshoe crab data (Example 10.6)

Colour	Spine condition	Carapace width (in cm)	Number of satellites	Weight (in g)
Medium	None OK	28.3	8	3050
Dark medium	None OK	22.5	0	1550
Light medium	Both OK	26.0	9	2300
Dark medium	None OK	24.8	0	2100
Dark medium	None OK	26.0	4	2600
Medium	None OK	23.8	0	2100
⋮	⋮	⋮	⋮	⋮

Colour is on a continuum from light to dark, and spine condition counts the number of intact sides, so we define both as ordered factors:

```
> data(hcrabs); str(hcrabs)
'data.frame':        173 obs. of  5 variables:
 $ Col  : Factor w/ 4 levels "D","DM","LM",..: 4 2 3 2 2 4 3 2 4 2 ...
 $ Spine: Factor w/ 3 levels "BothOK","NoneOK",..: 2 2 1 2 2 2 1 3 1 2 ...
 $ Width: num  28.3 22.5 26 24.8 26 23.8 26.5 24.7 23.7 25.6 ...
 $ Sat  : int  8 0 9 0 4 0 0 0 0 0 ...
 $ Wt   : int  3050 1550 2300 2100 2600 2100 2350 1900 1950 2150 ...
> hcrabs$Col <- ordered(hcrabs$Col, levels=c("LM", "M", "DM", "D"))
> hcrabs$Spine <- ordered(hcrabs$Spine,
                  levels=c("NoneOK", "OneOK", "BothOK"))
```

Plotting `Sat` against the other variables shows trends for more satellite crabs to congregate around females that are larger (in weight and width), are lighter in colour, and have no spinal damage (Fig. 10.6).

```
> with(hcrabs,{
   logSat <- log(Sat+1)
   plot( jitter(Sat) ~ Wt, ylab="Sat", las=1)
   plot( jitter(logSat) ~ log(Wt), ylab="log(Sat+1)", las=1)
   plot( logSat ~ Col, ylab="log(Sat+1)", las=1)
   plot( jitter(Sat) ~ Width, ylab="Sat", las=1)
   plot( jitter(logSat) ~ log(Width), ylab="log(Sat+1)", las=1)
   plot( logSat ~ Spine, ylab="log(Sat+1)", las=1)
 })
```

`jitter()` is used to avoid overplotting. Plots on the log-scale are preferable because the values of `Wt` and `Width` are distributed more symmetrically on the log-scale, and because the relationships between them and `Sat` are more likely to be relative rather than additive. `log(Sat+1)` is used to avoid taking logarithm of zero.

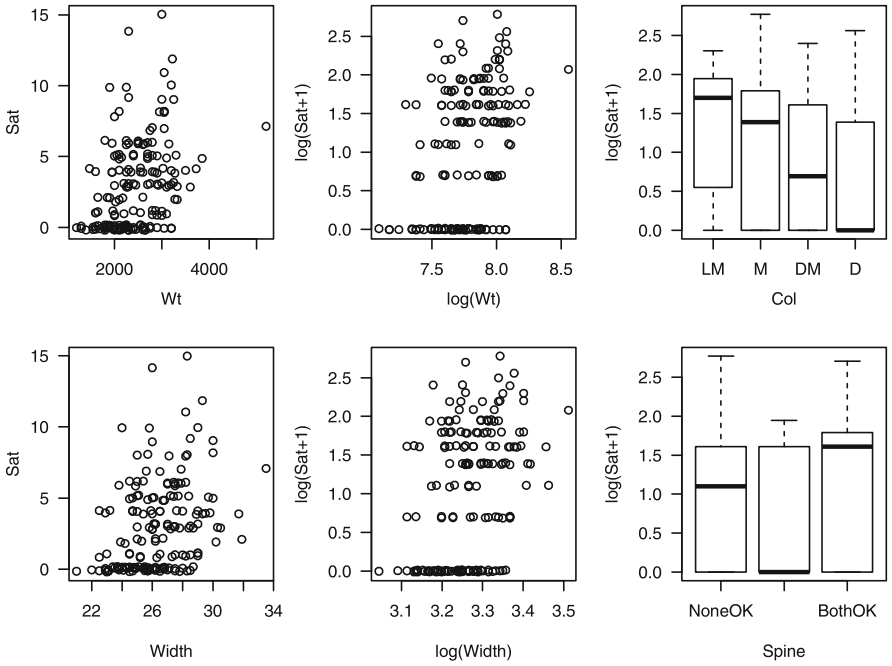

Fig. 10.6 The number of satellites on each female horseshoe crab plotted against the weight, colour, width and spine condition (Sect. 10.6)

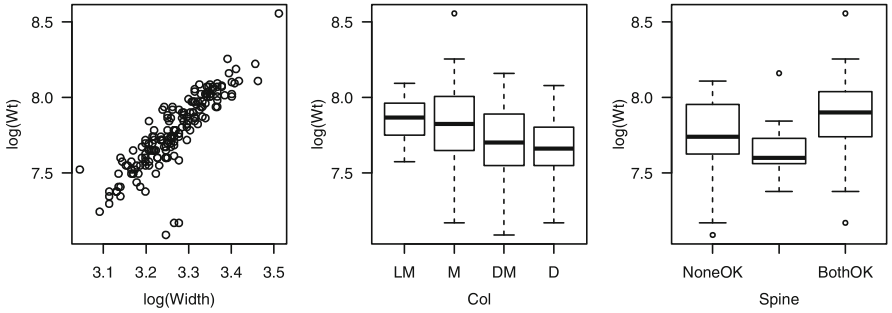

Fig. 10.7 Weight of each female horseshoe crab plotted against width, colour and spine condition (Sect. 10.6)

The explanatory variables are inter-related however; `Wt` is the most obvious overall summary of the size of each female. It turns out that lighter-coloured females are also typically heavier, as are females with no spine damage, so the relationships observed between `Sat` and `Col` and `Spine` might be explained by this (Fig. 10.7).

```
> with(hcrabs,{
   plot( log(Wt) ~ log(Width), las=1 )
   plot( log(Wt) ~ Col, las=1 )
   plot( log(Wt) ~ Spine, las=1 )
 })
> coef(lm( log(Wt) ~ log(Width), data=hcrabs ))
(Intercept)  log(Width)
      -0.60        2.56
```

`Wt` should be proportional to the volume of each female, hence should be approximately proportional to `Width^3`, if the females are all the same shape. Indeed, `log(Wt)` is nearly linearly related to `log(Width)` with a slope nearly equal to 3.

Crabs tend to congregate and interact with one another, rather than behaving independently, hence we should expect overdispersion *a priori* relative to Poisson for the counts of satellite crabs. We fit a quasi-Poisson GLM with log-link:

```
> cr.m1 <- glm(Sat ~ log(Wt) + log(Width) + Spine + Col,
              family=quasipoisson, data=hcrabs)
> anova(cr.m1, test="F")
           Df Deviance Resid. Df Resid. Dev     F  Pr(>F)
NULL                         172        633
log(Wt)     1     83.1       171        550 25.96 9.4e-07 ***
log(Width)  1      0.0       170        550  0.00    0.96
Spine       2      1.1       168        549  0.18    0.84
Col         3      7.6       165        541  0.79    0.50
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> deviance(cr.m1)
[1] 541
> df.residual(cr.m1)
[1] 165
```

The residual deviance and Pearson X^2 are both more than three times the residual degrees of freedom, so our expectation of overdispersion seems confirmed. Using F-tests, `log(Wt)` is a highly significant predictor whereas none of the other variables are at all significant, after adjusting for `log(Wt)`. We adopt a model with just `Wt` as an explanatory variable:

```
> cr.m2 <- glm(Sat ~ log(Wt), family=quasipoisson, data=hcrabs)
> printCoefmat(coef(summary(cr.m2)), digits=3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.568      2.664   -4.72  4.9e-06 ***
log(Wt)        1.744      0.339    5.15  7.0e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is tempting to speculate on the biological implications. It might well be possible for a male crab to sense the overall weight of the female crab by smell or other chemical senses, because the amount of chemical emitted by

a female should be proportional to her size, whereas width, colour or spine damage would need vision. The results perhaps suggest that the crabs do not use vision as their primary sense.

We may worry that nearly half of the values of the response `Sat` are 0 or 1, which may suggest a problem for the distribution of the residual deviance and the evaluation of overdispersion. However a quick simulation shows that the chi-square approximation for the residual deviance is excellent:

```
> x <- log(hcrabs$Wt); dev <- rep(NA, 100)
> n <- length(hcrabs$Sat); mu <- fitted(cr.m2)
> for (i in 1:100) {
    y <- rpois(n, lambda=mu)  # Generate random Poisson values
    dev[i] <- glm(y~x, family=quasipoisson)$deviance
 }
> c(Mean.Dev=mean(dev), Std.Dev=sd(dev))
 Mean.Dev   Std.Dev
185.53962  19.61709
```

The mean and standard deviance of the residual deviance are close to their theoretical values of $\mathrm{df} = 171$ and $\sqrt{2 \times \mathrm{df}} = 18.5$ respectively, under the null hypothesis of Poisson variation. (Note: A χ^2 distribution with k degrees of freedom has mean k and standard deviation $\sqrt{2k}$.)

The diagnostics for this model suggest a reasonable model:

```
> plot( resid(cr.m2) ~ sqrt(fitted(cr.m2)), las=1,
    main="Deviance residuals", ylab="Deviance residuals",
    xlab="Square root of fitted values" )
> plot( cooks.distance(cr.m2), type="h", las=1,
    ylab="Cook's distance, D", main="Cook's distance")
> qqnorm( resid(cr.m2), las=1,
    main="Normal Q-Q plot\ndeviance residuals")
> qqline( resid(cr.m2))
```

Notice that quantile residuals cannot be used for the quasi-Poisson model; the trend in the bottom left of the Q–Q plot may be due to the use of deviance residuals (Fig. 10.8). No observation is identified as influential using Cook's distance or DFBETAS, but other criteria indicate influential observations:

```
> colSums( influence.measures(cr.m2)$is.inf )
  dfb.1_ dfb.l(W)    dffit    cov.r   cook.d      hat
       0        0        1        8        0        3
```

The quasi-Poisson model indicates that heavier crabs have more satellites on average. The fitted systematic component is

$$\log \mu = -12.57 + 1.744 \log W \quad \text{or equivalently} \quad \mu = 0.000003483 \times W^{1.744},$$

where W is the weight of the crabs in grams. If the regression coefficient for $\log W$ was 1, then the expected number of satellite crabs would be directly proportional to the weight of the female. The number of satellites seems to increase just a little faster than this.

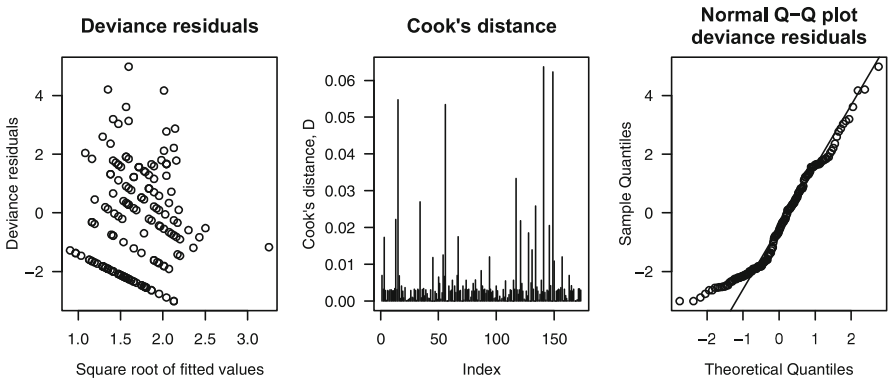

Fig. 10.8 Diagnostic plots for the quasi-Poisson model `cr.m2`. The deviance residuals against fitted values (left panel); Cook's distance (centre panel); a Q–Q plot of the quantile residuals (right panel) (Sect. 10.6)

An alternative model is to fit a negative binomial model:

```
> library(MASS)
> cr.nb <- glm.nb(Sat ~ log(Wt), data=hcrabs)
> cr.nb <- glm.convert(cr.nb)
> anova(cr.nb, dispersion=1, test="Chisq")


        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      172     219.81
log(Wt)  1   23.339       171     196.47 1.358e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> printCoefmat(coef(summary(cr.nb, dispersion=1)))
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -14.55581    3.10909 -4.6817 2.845e-06 ***
log(Wt)       1.99862    0.39839  5.0168 5.254e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cr.nb$theta
[1] 0.9580286
```

The fitted negative binomial distribution uses $\hat{k} = 0.9580$. The diagnostic plots (not shown) indicate that the negative binomial model is also suitable. No observation is identified as influential using Cook's distance:

```
> colSums( influence.measures(cr.nb)$is.inf )
  dfb.1_ dfb.l(W)    dffit    cov.r   cook.d      hat
       0        0        0        6        0        3
```

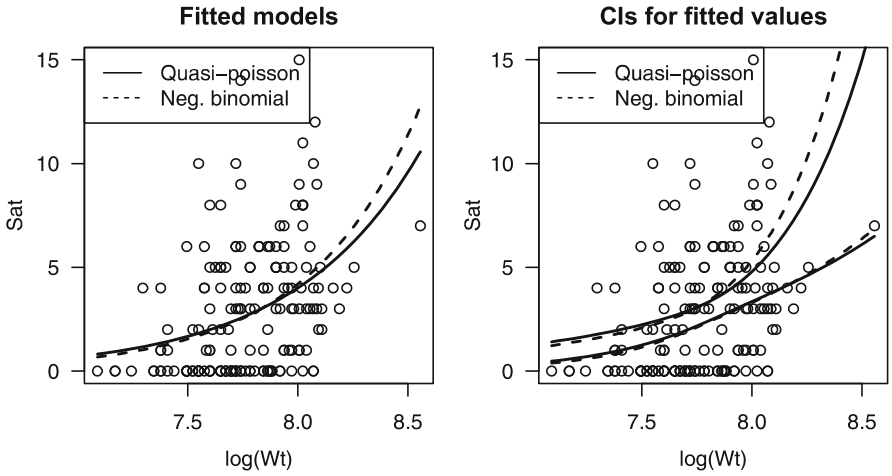

Fig. 10.9 Comparing the systematic components of the quasi-Poisson model and the negative binomial GLM (left panel) and the corresponding 95% confidence intervals (right panel) fitted to the horseshoe crab data. Solid lines represent the quasi-Poisson model, while dashed lines represent the negative binomial model

The differences between the two models becomes apparent for heavier crabs, for both the systematic components (Fig. 10.9, left panel) and the random components (Fig. 10.9, right panel). First, create predictions for a range of weights:

```
> newW <- seq( min(hcrabs$Wt), max(hcrabs$Wt), length=100)
> newS.qp <- predict(cr.m2, newdata=data.frame(Wt=newW), se.fit=TRUE)
> newS.nb <- predict(cr.nb, newdata=data.frame(Wt=newW), se.fit=TRUE,
                     dispersion=1)
> tstar <- qt(0.975, df=df.residual(cr.m2) ) # For a 95% CI
> ME.qp <- tstar * newS.qp$se.fit; ME.nb <- tstar * newS.nb$se.fit
> mu.qp <- newS.qp$fit; mu.nb <- newS.nb$fit
```

Then plot:

```
> par( mfrow=c(1, 2))
> plot( Sat~log(Wt), data=hcrabs, las=1, main="Fitted models")
> lines( exp(mu.qp) ~ log(newW), lwd=2 )
> lines( exp(mu.nb) ~ log(newW), lwd=2, lty=2 );
> legend("topleft", lty=1:2, legend=c("Quasi-poisson", "Neg. binomial") )
> #
> plot( Sat~log(Wt), data=hcrabs, las=1, main="CIs for fitted values")
> ci.lo <- exp(mu.qp - ME.qp); ci.hi <- exp(mu.qp + ME.qp)
> lines( ci.lo ~ log(newW), lwd=2); lines( ci.hi ~ log(newW), lwd=2)
> ci.lo <- exp(mu.nb - ME.nb); ci.hi <- exp(mu.nb + ME.nb)
> lines( ci.lo ~ log(newW), lwd=2, lty=2)
> lines( ci.hi ~ log(newW), lwd=2, lty=2)
> legend("topleft", lty=1:2, legend=c("Quasi-poisson", "Neg. binomial") )
```

10.7 Using R to Fit GLMs to Count Data

A Poisson GLM is specified in R using `glm(formula, family=poisson())` (note the lower case `p`). The link functions `"log"`, `"identity"`, and `"sqrt"` are permitted with Poisson distributions. Quasi-Poisson models are specified using `glm(formula, family=quasipoisson())`.

To fit negative binomial models, use `glm.nb()` from package **MASS** [37] when k is unknown (the usual situation). The output from `glm.nb()` is converted to the style of output from `glm()` using `glm.convert()`. Then, the usual `anova()` and `summary()` commands may be used, remembering to set `dispersion=1` when using `summary()`. See `?negative.binomial`, `?glm.nb`, and Sect. 10.5.2 for more information.

The function `gl()` is useful for generating factors occurring in a regular pattern, as is common in tabulated data. `gl(3, 2, 12)` produces a factor of length 12 with three levels (labelled `1`, `2` and `3` by default), appearing two at a time:

```
> gl(3, 2, 18, labels=c("A", "B", "C") )
 [1] A A B B C C A A B B C C A A B B C C
Levels: A B C
```

The functions `margin.table()` and `prop.table()` are useful for producing marginal tables and tables of proportions from raw data in tables (Sect. 10.4.5).

10.8 Summary

Chapter 10 considers fitting GLMs to count data. Counts are commonly modelled using the Poisson distribution (Sect. 10.2), where $\mu > 0$ is the expected count and $y = 0, 1, 2, \ldots$. Note that $\phi = 1$ and $V(\mu) = \mu$. The residual deviance $D(y, \hat{\mu})$ is suitably described by a $\chi^2_{n-p'}$ distribution if $\min\{y_i\} \geq 3$ (Sect. 10.2). The logarithmic link function is often used for Poisson GLMs (Sect. 10.2).

When any of the explanatory variables are quantitative, the fitted Poisson GLM is also called a Poisson regression model. When all the explanatory variables are qualitative, the fitted Poisson GLM is also called a log-linear model (Sect. 10.2).

Poisson GLMs can be used to model rates (such as counts of cancer cases per unit of population) by using a suitable offset in the linear predictor (Sect. 10.3).

Count data often appear cross-classified in tables, commonly called contingency tables (Sect. 10.4). Contingency tables may arise under various sampling schemes, each implying a different random component (Sect. 10.4). How-

ever, in all cases a Poisson GLM can be fitted provided the coefficients in the linear predictor corresponding to fixed margins are included in the model.

Three-dimensional tables may be interpreted, and possibly simplified, according to which interactions are present in the model (Sect. 10.4.4). If tables are collapsed incorrectly, the resulting tables may be misleading. Simpson's paradox is an extreme example (Sect. 10.4.5). Poisson GLMs fitted to higher-order tables may be difficult to interpret (Sect. 10.4.7).

Contingency tables may contain cells with zero counts (Sect. 10.4.8). Sampling zeros occur by chance, and larger samples may produce counts in these cells. Structural zeros appear for impossible events, so cells containing structural zeros must be removed from the analysis.

Overdispersion occurs when the variation in the responses is greater than expected under the Poisson model (Sect. 10.5). Possible causes are that the model is misspecified (in which case the model should be amended), the means are not constant, or the responses are not independent.

In cases of overdispersion relative to the Poisson GLM, a negative binomial distribution may be used, which is an EDM if k is known (Sect. 10.5.2). For the negative binomial distribution, $V(\mu) = \mu + \mu^2/k$ for $k > 0$. The value of k usually needs to be estimated (by $\hat{k}$) for a negative binomial GLM (Sect. 10.5.2). If overdispersion is observed, a quasi-Poisson model may be fitted also, which assumes $V(\mu) = \phi\mu$ (Sect. 10.5.3).

Problems

Selected solutions begin on p. 541.

10.1. Consider the negative binomial distribution, whose probability function is given in (10.11).

1. Show that the negative binomial distribution with known k is an EDM, by identifying θ, $\kappa(\theta)$ and ϕ. (See Sect. 5.3.6, p. 217.)
2. Show that the negative binomial distribution with known k has $\text{var}[y] = \mu + \mu^2/k$.
3. Deduce the canonical link function for the negative binomial distribution.
4. Show that, for the negative binomial distribution,

$$d(y, \mu) = 2\left\{y \log \frac{y}{\mu} - (y + k) \log \frac{y + k}{\mu + k}\right\}$$

for $y > 0$. Also, deduce the unit deviance when $y = 0$.

10.2. If the fitted Poisson GLM includes a constant term, and the logarithmic link function is used, the sum over the observations of the second term in the expression for the residual deviance is zero. In other words, $\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$. Prove this result by writing the log-likelihood for a model with linear

predictor containing a constant term, say β_0, differentiating the log-likelihood with respect to β_0, setting to zero, and solving.

10.3. Sometimes, count data explicitly omit zero counts. Examples include the numbers of days patients spend in hospital (only patients who actually stay overnight in hospital are considered, and so the smallest possible count is one); the number of people per car using a rural road (the driver at least must be in the car); and a survey of the number of people living in each household (to respond, the households must have at least one person). Using a Poisson distribution is inadequate, as the zero counts will be modelled as true zero counts.

In these situations, the zero-truncated Poisson distribution may be suitable, with probability function

$$\mathcal{P}(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{\{1-\exp(-\lambda)\}y!},$$

where $y = 1, 2, \ldots$ and $\lambda > 0$.

1. Show that the truncated Poisson distribution is an EDM by identifying θ and $\kappa(\theta)$.
2. Show that $\mu = \mathrm{E}[y] = \lambda/\{1-\exp(-\lambda)\}$, and that $\mu > 1$.
3. Find the variance function for the truncated Poisson distribution.
4. Plot the truncated Poisson distribution and the Poisson distribution for $\lambda = 2$, and compare.

10.4. A study [25] used a Poisson GLM to model the number of politicians switching political parties in the USA. The response variable was the number of members of the House of Representatives who switched parties every year from 1802–1876.

1. Explain why the authors used a Poisson GLM to model the data.
2. The authors use eleven possible explanatory variables in the linear predictor. One of the explanatory variables is whether or not the year is an election year (election years are coded as 0, non-election years as 1). The coefficient for this explanatory variable is 1.051. Interpret the meaning of this coefficient.
3. The estimated standard error for the election year parameter is 0.320. Determine if the parameter is statistically significant.
4. Compute and interpret a 90% confidence interval for the election year parameter.

10.5. A study in the USA [22] examined the number of pregnancies in a stratified random sample of 1154 sexually-active teenage girls (7th to 12th grade). Details of the fitted Poisson GLM are shown in Table 10.11.

1. Explain why the years of sexual activity is used as an offset.

Table 10.11 The fitted Poisson GLMs for the teenage pregnancy data. The response variable is the number of pregnancies. All variables are binary (0: no; 1: yes) apart from age, which is measured in completed years. Years of sexual activity is used as an offset (Problem 10.5)

	df	$\hat{\beta}_j$	se($\hat{\beta}_j$)	Wald 95% confidence limits		Deviance
Intercept	1	−2.0420	0.9607	−3.9248	−0.1591	4.52
Current age (in years)	1	0.1220	0.0543	0.0156	0.2283	5.05
Race ('White' is the reference)						
African-American	1	0.6604	0.1287	0.4082	0.9126	26.33
Hispanic	1	0.2070	0.2186	−0.2215	0.6354	0.90
Asian	1	0.4896	0.3294	−0.1561	1.1352	2.21
Single	1	−0.9294	0.2080	−1.3371	−0.5218	19.97
College plans	1	−0.0871	0.0515	−0.1881	0.0139	2.86
Contraceptive self-efficacy	1	−0.2241	0.0845	−0.3897	−0.0585	7.04
Consistent use of contraceptives	1	−0.2729	0.0825	−0.4346	−0.1113	10.95
Residual df:	1144					
Residual deviance:	3359.9					

2. Use likelihood ratio tests to identify statistically significant explanatory variables.
3. Use the Wald statistics to identify statistically significant explanatory variables. Compare to the results of using the likelihood ratio test.
4. Interpret the coefficients in the model.
5. Show that overdispersion may be present.
6. Because of the possible overdispersion, estimate ϕ for the quasi-Poisson model. Hence compute $\hat{\beta}_j$ and se($\hat{\beta}_j$) for the quasi-Poisson GLM.
7. Form a 95% confidence interval for age using the quasi-Poisson GLM.

10.6. The brood sizes of blue tits were experimentally changed (increased or decreased) through three brooding seasons to study the survival of offspring [32, Table 2]. The hypothesis was that blue tits should produce the clutch size maximizing the survival of their offspring (so that manipulated broods should show less surviving offspring than unmanipulated broods). In other words, the number of eggs laid is optimum given the ability of the parents to rear the offspring (based on their body condition, food resources, age, etc.). A log-linear model for modelling the number of offspring surviving y produced the results in Table 10.12, where M is the amount of manipulation (ranging from taking ten eggs ($M = -10$) to adding four eggs ($M = 4$) to the clutch), and C is the original clutch size (ranging from two to 17 eggs).

1. Write down the fitted model from Table 10.12 (where $\hat{\beta}_0 = -2.928$).
2. Using likelihood ratio tests, determine which explanatory variables are significant.
3. Use Wald statistics to determine the significance of each parameter. Compare to the results from the likelihood ratio tests, and comment.

Table 10.12 The analysis of deviance table for a Poisson GLM fitted to the blue tits data. The response variable is the number of offspring surviving (Problem 10.6)

Model	Residual deviance	df	$\hat{\beta}_j$	se($\hat{\beta}_j$)
Null model	732.74	617		
$+ C$	662.25	616	0.238	0.028
$+ M$	649.01	615	0.017	0.035
$+ M^2$	637.22	614	−0.028	0.009

Table 10.13 Information about the fitted Poisson GLM for the *spina bifida* study. The response variable is the number of babies born with *spina bifida* (Problem 10.7)

Model	Residual deviance	df	$\hat{\beta}_j$	se($\hat{\beta}_j$)	
Null	554.11	200			
$+ \log B$	349.28	199	1.06	0.07	
$+ S$	305.32	197	−8.61	0.68	(routine screening)
			−8.18	0.67	(no routine screening)
			−8.43	0.68	(policy uncertain)
$+ C$	285.06	196	−0.11	0.03	
$+ U$	266.88	195	0.046	0.009	
$+ A$	256.03	194	0.039	0.011	

4. Compute and interpret the 95% confidence interval for the effect of the original clutch size C.
5. Comment on under- or overdispersion for this model.
6. Using the fitted model, determine the value of M maximizing expected offspring survival μ.
7. Determine if any manipulation of the clutch size decreases the survival chances of the young.

10.7. A study of *spina bifida* in England and Wales [27] examined the relationship between the number of babies born with *spina bifida* between 1983 and 1985 inclusive in various Regional Health Authorities (RHA), and explanatory variables such as the total number of live and still births between 1983–1985, B; the screening policy of the health authority in 1982, S (routine; non-routine; uncertain); the percentage of female residents born in the Caribbean, C; the percentage economically-active residents unemployed, U; the percentage of residents lacking a car, L; and the percentage of economically-active residents employed in agriculture, A. A Poisson GLM with a log-link was fitted (Table 10.13) to model the number of babies born with *spina bifida*.

1. Write down the fitted model. (Note that a different constant term is fitted for each screening policy.)
2. Using the standard errors, check which parameters are significantly different from zero.
3. Use likelihood ratio tests to determine which explanatory variables are significant in the model.

4. Interpret the effect of the unemployment rate U.
5. Compute and interpret the 95% confidence interval for the effect of the unemployment rate U.
6. Explain why using $\log B$ as an offset seems reasonable from the description of the data. Also explain why Table 10.13 supports this approach.
7. Is overdispersion likely to be a problem?

10.8. For the depressed youth data used in Sect. 10.4.7 (p. 393), fit the model used in that section as follows (data set: `dyouth`).

1. Show that the four-factor interaction is not significant.
2. Show that only one three-factor interaction is significant in the model.
3. Then show that four two-factor interactions are needed in the model (some because they are significant, some because of the marginality principle).
4. Show that the model is adequate by examining the model diagnostics.

10.9. Consider the Danish lung cancer data of Example 10.1 (data set: `danishlc`). In that example, a Poisson GLM was fitted to model the *number* of lung cancers per unit of population.

1. Fit a model for the *proportion* of lung cancers, based on the proportion `Cases/Pop`, and compare to the equivalent Poisson GLM fitted in Sect. 10.3.
2. Show that the conditions for the equivalence of the binomial and Poisson GLMs, as given in Sect. 10.4.6, are approximately satisfied.

10.10. In Sect. 8.12 (p. 322), a Poisson GLM was fitted to the noisy miner data [30] (data set: `nminer`) that was first introduced in Example 1.5 (p. 14). In Example 1.5, the only explanatory variable considered was the number of eucalypts `Eucs`, but the data frame actually contains a number of other explanatory variables: the number of buloke trees (`Bulokes`); the area in hectares of remnant patch vegetation at each site (`Area`); whether the area was grazed (`Grazed`: `1` means yes); and whether shrubs were present in the transect (`Shrubs`: `1` means yes).

1. Find a suitable Poisson regression model for modelling the number of noisy miners `Minerab`, including a diagnostic analysis.
2. Is the saddlepoint approximation likely to be accurate? Explain.

10.11. The number of deaths for 1969–1973 (1969–1972 for Belgium) due to cervical cancer is tabulated (Table 10.14; data set: `cervical`) by age group for four different countries [19, 38].

1. Plot the data, and discuss any prominent features.
2. Explain why an offset is useful when fitting a GLM to the data.
3. Fit a Poisson GLM with `Age` and `Country` as explanatory variables. Produce the plot of residuals against fitted values, and evaluated the model.

Table 10.14 The number of deaths y due to cervical cancer and woman-years at-risk T in various age groups, for four countries (Problem 10.11)

	25–34		35–44		45–54		55–64	
Country	y	T	y	T	y	T	y	T
England and Wales	192	15,399	860	14,268	2762	15,450	3035	15,142
Belgium	8	2328	81	2557	242	2268	268	2253
France	96	15,324	477	16,186	998	14,432	1117	13,201
Italy	45	19,115	255	18,811	621	16,234	839	15,246

Table 10.15 The number of women developing depression in a 1-year period in Camberwell, South London [15]. SLE refers to a 'Severe Life Event' (Example 6.2)

	Three children under 14		Other women	
	SLE	No SLE	SLE	No SLE
Depression	9	0	24	4
No depression	12	20	119	231

4. Fit the corresponding quasi-Poisson model. Produce the plot of residuals against fitted values, and evaluated the model.
5. Fit the corresponding negative binomial GLM. Produce the plot of residuals against fitted values, and evaluated the model.
6. Which model seems appropriate, if any?

10.12. In a study of depressed women [15], women were classified into groups (Table 10.15; data set: `dwomen`) based on their depression level (`Depression`), whether a severe life event had occurred in the last year (`SLE`), and if they had three children under 14 at home (`Children`). Model these counts using a Poisson GLM, and summarize the data if possible.

10.13. The number of severe and non-severe cyclones in the Australian region between 1970 and 2005 were recorded (Table 10.16; data set: `cyclones`), together with a climatic index called the *Ocean Niño Index*, or ONI. The ONI is a 3-month running mean of sea surface temperature anomalies; Table 10.16 shows the ONI at four times during each year.

1. Plot the number of severe cyclones against the ONI, and then plot the number of non-severe cyclones against the ONI. Comment.
2. Fit a Possion GLM to model the number of severe cyclones, and another Poisson GLM for the number of non-severe cyclones.
3. Interpret your final models.

10.14. A study [13, 18] of the species richness (the number of species) of ants at 22 sites in New England, USA, examined relationships with habitat (forest

Table 10.16 The number of severe and non-severe cyclones in the Australian region, with four values of the Ocean Niño Index (ONI) for each year (Problem 10.13)

	Number of cyclones		ONI			
Year	Severe	Non-severe	JFM	AMJ	JAS	OND
1969	3	7	1.0	0.6	0.4	0.8
1970	3	14	0.3	0.0	−0.8	−0.9
1971	9	7	−1.3	−0.8	−0.8	−1.0
1972	6	6	−0.4	0.5	1.3	2.0
1973	4	15	1.2	−0.6	−1.3	−2.0
1974	3	13	−1.7	−0.9	−0.5	−0.9
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 10.17 Species richness of ants in New England, USA. Elevation is in metres (Problem 10.14)

		Species richness in:				Species richness in:	
Elevation	Latitude	Forest	Bog	Elevation	Latitude	Forest	Bog
41.97	389	6	5	42.57	335	10	4
42.00	8	16	6	42.58	543	4	2
42.03	152	18	14	42.69	323	5	7
42.05	1	17	7	43.33	158	7	2
42.05	210	9	4	44.06	313	7	3
42.17	78	15	8	44.29	468	4	3
42.19	47	7	2	44.33	362	6	2
42.23	491	12	3	44.50	236	6	3
42.27	121	14	4	44.55	30	8	2
42.31	95	9	8	44.76	353	6	5
42.56	274	10	8	44.95	133	6	5

or bog), elevation (in m) and latitude (Table 10.17; data set: `ants`). Find a suitable model for the data. Interpret your final model.

10.15. A study [14, 17, 33] compared the number polyps in patients with familial adenomatous polyposis (Table 10.18; data set: `polyps`), after treatment with a new drug (sulindac) or a placebo.

1. Plot the data and comment.
2. Find a suitable Poisson GLM for modelling the data, and show that overdispersion exists.
3. Fit a quasi-Poisson model to the data.
4. Fit a negative binomial GLM to the data.
5. Decide on a final model.

10.16. An experiment [21] compared the density of understorey birds at a series of sites in two areas either side of a stockproof fence (Table 10.19;

Table 10.18 The number of polyps in the treatment and placebo group for patients with famial adenomatous polyposis (Problem 10.15)

Treatment group				Placebo group			
Number	Age	Number	Age	Number	Age	Number	Age
1	22	17	22	7	34	44	19
1	23	25	17	10	30	46	22
2	16	33	23	15	50	50	34
3	23			28	18	61	13
3	23			28	22	63	20
4	42			40	27		

Table 10.19 The number of understorey-foraging birds observed in three 20-min surveys of 2 ha quadrats either side of a stockproof fence, before and after grazing (Problem 10.16)

Ungrazed				Grazed					
Before	After	Before	After	Before	After	Before	After	Before	After
0	1	37	5	2	6	0	0	30	4
3	10	7	5	0	2	1	3	13	14
1	10	10	4	0	0	0	7	0	6
19	29	11	4	1	11	4	17	2	8
8	21	1	6	4	7	0	7	0	18
		30	15	2	4	0	0	1	4
				3	3	2	7		

data set: `grazing`). One side had limited grazing (mainly from native herbivores), and the other was heavily grazed by feral herbivores, mostly horses. Bird counts were recorded at the sites either side of the fence (the 'before' measurements). Then the herbivores were removed, and bird counts recorded again (the 'after' measurements). The measurements are the total number of understorey-foraging birds observed in three 20-min surveys of 2 ha quadrats.

1. Plot the data, and explain the important features.
2. Fit a Poisson GLM with systematic component `Birds ~ When * Grazed`, ensuring a diagnostic analysis.
3. Show that overdispersion exists. Demonstrate by computing the mean and variance of each combination of the explanatory variables.
4. Fit a quasi-Poisson model.
5. Fit a negative binomial GLM.
6. Compare all three fitted models to determine a suitable model.
7. Interpret the final model.

10.17. An experiment [23, 36] recorded the time to failure of a piece of electronic equipment while operating in two different modes. In any session, the machine is run in both modes for varying amounts of time (Table 10.20; data

Table 10.20 Observations on electronic equipment failures. The time spent in each mode is measured in weeks (Problem 10.17)

Time spent in Mode 1	Time spent in Mode 2	Number of failures	Time spent in Mode 1	Time spent in Mode 2	Number of failures
33.3	25.3	15	116.3	53.6	27
52.2	14.4	9	131.7	56.6	23
64.7	32.5	14	85.0	87.3	18
137.0	20.5	24	91.9	47.8	22
125.9	97.6	27			

Table 10.21 The estimated number of deaths for the five leading cancer sites in Canada in 2000, by geographic region and gender (Problem 10.18)

	Ontario		Newfoundland		Quebec	
Cancer	Male	Female	Male	Female	Male	Female
Lung	3500	2400	240	95	3500	2000
Colorectal	1250	1050	60	50	1100	1000
Breast	0	2100	0	95	0	1450
Prostate	1600	0	80	0	900	0
Pancreas	540	590	20	25	390	410
Estimated population:	11,874,400		533,800		7,410,500	

set: `failures`). For each operating period, Mode 1 is the time spent operating in one mode and Mode 2 is the time spent operating in the other mode. The number of failures in each period is recorded, where each operating period is measured in weeks. The interest is in finding a model for the number of failures given the amount of time the equipment spends in the two modes.

1. Plot the number of failures against the time spent in Mode 1, and then against the time spent in Mode 2.
2. Show that an identity link function may be appropriate.
3. Fit the Poisson model, to model the number of failures as a function of the time spent in the two modes. Which mode appears to be the major source of failures?
4. Is there evidence of under- or overdispersion?
5. Interpret the final model.

10.18. A report on Canadian cancer statistics estimated the number of deaths from various types of cancer in Canada in 2000 [7]. The five leading cancer sites are studied here (Table 10.21; data set: `ccancer`).

1. Plot the cancer rates per thousand of population against each geographical location, and then against gender. Comment on the relationships.
2. Identify the zeros as systematic or sampling.
3. Find an appropriate model for the data using an appropriate offset. Do the cancer rates appear to differ across the geographic regions?
4. Interpret the fitted model.

Table 10.22 Health concerns of teenagers (Problem 10.20)

		Health concern			
Sex	Age group	Sex; relationships	Menstrual	How healthy	Nothing at all
Males	12–15	4	0	42	57
	16–17	2	0	7	20
Females	12–15	9	4	19	71
	16–17	7	8	10	31
Total		22	12	78	179

Table 10.23 Smoking and survival data for Whickham women (Problem 10.21)

	Smokers		Non-smokers	
Age (at first survey)	Alive	Dead	Alive	Dead
18–24	53	2	61	1
25–34	121	3	152	5
35–44	95	14	114	7
45–54	103	27	66	12
55–64	64	51	81	40
65–74	7	29	28	101
75+	0	13	0	64

10.19. In Problem 2.18 (p. 88), data were presented about children building towers out of building blocks (data set: `blocks`). One variable measured was the number of blocks needed to build a tower as high as possible. Find a model for the number of blocks, including a diagnostic analysis.

10.20. A study [6, 9, 16] asked teenagers about their health concerns, including sexual health. The data in Table 10.22 (data set: `teenconcerns`) are the number of teenagers who reported wishing to talk to their doctor about the indicated topic.

1. How would you classify the zeros? Explain.
2. Fit an appropriate log-linear model to the data.

10.21. A survey originally conducted in 1972–1974 [3, 10] asked women in Whickham in the north of England about their smoking habits and age, and recorded their survival (Table 10.23; data set: `wwomen`). A subsequent survey 20 years later followed up the women to determine how many women from the original survey had died.

1. Classify the zeros as sampling or structural zeros.
2. Plot the proportion of women alive at each age (treat age as continuous, using the lower boundary of each class), distinguishing between smokers and non-smokers. Comment.

3. Compute the *overall* percentage of smokers and non-smokers alive, and comment.
4. Compute the percentage of smokers and non-smokers *in each age group* who died. Compare to the previous answers. Comment and explain.
5. Fit a suitable log-linear model for the *number* of women alive. What evidence is there that the data should not be collapsed over age?

References

[1] Agresti, A.: An Introduction to Categorical Data Analysis, second edn. Wiley-Interscience, New York (2007)
[2] Andersen, E.B.: Multiplicative Poisson models with unequal cell rates. Scandinavian Journal of Statistics **4**, 153–158 (1977)
[3] Appleton, D.R., French, J.M., Vanderpump, M.P.J.: Ignoring a covariate: An example of Simpson's paradox. The American Statistician **50**, 340–341 (1996)
[4] Berkeley, E.C.: Right answers—a short guide for obtaining them. Computers and Automation **18**(10) (1969)
[5] Brockmann, H.J.: Satellite male groups in horseshoe crabs, *limulus polyphemus*. Ethology **102**, 1–21 (1996)
[6] Brunswick, A.F.: Adolescent health, sex, and fertility. American Journal of Public Health **61**(4), 711–729 (1971)
[7] Canadian Cancer Society: Canadian cancer statistics 2000. Published on the internet: www.cancer.ca/stats2000/tables/tab5e.htm (2000). Accessed 19 September 2001
[8] Charig, C.R., Webb, D.R., Payne, S.R., Wickham, J.E.A.: Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. British Medical Journal **292**, 879–882 (1986)
[9] Christensen, R.: Log-Linear Models. Springer Texts in Statistics. Springer, New York (2013)
[10] Davison, A.C.: Statistical Models. Cambridge University Press, UK (2003)
[11] Dunn, P.K.: Contingency tables and log-linear models. In: K. Kempf-Leonard (ed.) Encyclopedia of Social Measurement, pp. 499–506. Elsevier (2005)
[12] Dunn, P.K., Smyth, G.K.: Randomized quantile residuals. Journal of Computational and Graphical Statistics **5**(3), 236–244 (1996)
[13] Ellison, A.M.: Bayesian inference in ecology. Ecology Letters **7**, 509–520 (2004)
[14] Everitt, B.S., Hothorn, T.: A Handbook of Statistical Analyses using, second edn. Chapman & Hall/CRC, Boca Raton, FL (2010)

[15] Everitt, B.S., Smith, A.M.R.: Interactions in contingency tables: A brief discussion of alternative definitions. Psychological Medicine **9**, 581–583 (1979)
[16] Fienberg, S.: The Analysis of Cross-Classified Categorical Data. Springer, New York (2007)
[17] Giardiello, F.M., Hamilton, S.R., Krush, A.J., Piantadosi, S., Hylind, L.M., Celano, P., Booker, S.V., Robinson, C.R., Johan, G., Offerhaus, A.: Treatment of colonic and rectal adenomas with sulindac in famial adenomatous polyposis. New England Journal of Medicine **328**(18), 1313–1316 (1993)
[18] Gotelli, N.J., Ellison, A.M.: Biogeography at a regional scale: Determinants of ant species density in bogs and forests of New England. Ecology **83**(6), 1604–1609 (2002)
[19] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.Y., Ostrowski, E.: A Handbook of Small Data Sets. Chapman and Hall, London (1996)
[20] Health Department of Western Australia: Annual report 1997/1998—health of Western Australians—mortality and survival. Published on the internet: www.health.wa.gov.au/Publications/annualreport_9798/. Accessed 19 September 2001
[21] Howes, A.L., Maron, M., McAlpine, C.A.: Bayesian networks and adaptive management of wildlife habitat. Conservation Biology **24**(4), 974–983 (2010)
[22] Hutchinson, M.K., Holtman, M.C.: Analysis of count data using Poisson regression. Research in Nursing and Health **28**, 408–418 (2005)
[23] Jorgensen, D.W.: Multiple regression analysis of a Poisson process. Journal of the American Statistical Association **56**(294), 235–245 (1961)
[24] Julious, S.A., Mullee, M.A.: Confounding and Simpson's paradox. British Medical Journal **309**(1480), 1480–1481 (1994)
[25] King, G.: Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. American Journal of Political Science **32**(3), 838–863 (1988)
[26] Lindsey, J.K.: Modelling Frequency and Count Data. No. 15 in Oxford Statistical Science Series. Clarendon Press, Oxford (1995)
[27] Lovett, A.A., Gatrell, A.C.: The geography of *spina bifida* in England and Wales. Transactions of the Institute of British Geographers (New Series) **13**(3), 288–302 (1988)
[28] Luo, D., Wood, G.R., Jones, G.: Visualising contingency table data. The Australian Mathematical Society Gazette **31**(4), 258–262 (2004)
[29] Maag, J.W., Behrens, J.T.: Epidemiologic data on seriously emotionally disturbed and learning disabled adolescents: Reporting extreme depressive symptomatology. Behavioral Disorders **15**(1) (1989)
[30] Maron, M.: Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation **136**, 100–107 (2007)

[31] Norton, J., Lawrence, G., Wood, G.: The Australian public's perception of genetically-engineered foods. Australasian Biotechnology pp. 172–181 (1998)

[32] Pettifor, R.A.: Brood-manipulation experiments. I. The number of offspring surviving per nest in blue tits (*Parus caeruleus*). Journal of Animal Ecology **62**, 131–144 (1993)

[33] Piantadosi, S.: Clinical Trials: A Methodologic Perspective, second edn. John Wiley and Sons, New York (2005)

[34] Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2015. CA: A Cancer Journal for Clinicians **65**(1), 5–29 (2015)

[35] Smith, P.T., Heitjan, D.F.: Testing and adjusting for departures from nominal dispersion in generalized linear models. Journal of the Royal Statistical Society, Series C **42**(1), 31–41 (1993)

[36] Smyth, G.K.: Australasian data and story library (OzDASL) (2011). URL http://www.statsci.org/data

[37] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, fourth edn. Springer-Verlag, New York (2002). URL http://www.stats.ox.ac.uk/pub/MASS4

[38] Whittemore, A.S., Gong, G.: Poisson regression with misclassified counts: Applications to cervical cancer mortality rates. Journal of the Royal Statistical Society, Series C **40**(1), 81–93 (1991)