CHAPTER 7

Models for Count Data

Many response variables have counts as their possible outcomes. Examples are the number of alcoholic drinks you had in the previous week, and the number of devices you own that can access the internet (laptops, smart cell phones, tablets, etc.). Counts also occur as entries in cells of contingency tables that cross-classify categorical variables, such as the number of people in a survey who are female, college educated, and agree that humans are responsible for climate change. In this chapter we introduce generalized linear models (GLMs) for count response variables.

Section 7.1 presents models that assume a Poisson distribution for a count response variable. The *loglinear model*, using a log link to connect the mean with the linear predictor, is most common. The model can be adapted to model a *rate* when the count is based on an index such as space or time. Section 7.2 shows how to use Poisson and related multinomial models for contingency tables to analyze conditional independence and association structure for a multivariate categorical response variable. For the Poisson distribution, the variance must equal the mean, and data often exhibit greater variability than this. Section 7.3 introduces GLMs that assume a negative binomial distribution, which handles such *overdispersion* in a natural way. Many datasets show greater frequencies of zero counts than standard models allow, often because some subjects can have a zero outcome by chance but some subjects necessarily have a zero outcome. Section 7.4 introduces models that handle such *zero-inflated data*, which we might expect for a variable such as the frequency of alcohol drinking. Three examples illustrate models—one for rate data (Section 7.1.7), one for associations in contingency tables (Section 7.2.6), and one for zero-inflated count data (Section 7.5).

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.
© 2015 John Wiley & Sons, Inc. Published 2015 by John Wiley & Sons, Inc.

7.1 POISSON GLMS FOR COUNTS AND RATES

The simplest distribution for count data, placing its mass on the set of nonnegative integer values, is the *Poisson*. Its probabilities depend on a single parameter, the mean $\mu > 0$.

7.1.1 The Poisson Distribution

In equation (4.5) we observed that the Poisson probability mass function, $p(y; \mu) = e^{-\mu}\mu^y/y!$ for $y = 0, 1, 2, \ldots$, is in the exponential dispersion family with $E(y) = \text{var}(y) = \mu$. The Poisson distribution is unimodal with mode equal to the integer part of μ. Its skewness is described by $E(y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$. As μ increases, the Poisson distribution is less skewed and approaches normality, the approximation being fairly good when $\mu > 10$.

The Poisson distribution is often used for counts of events[1] that occur randomly over time or space at a particular rate, when outcomes in disjoint time periods or regions are independent. For example, a manufacturer of cell phones might find that the Poisson describes reasonably well the number of warranty claims received each week. The Poisson also applies as an approximation for the binomial when the number of trials n is large and π is very small, with $\mu = n\pi$. For the binomial, if $n \to \infty$ and $\pi \to 0$ such that $n\pi = \mu$ is fixed, then the binomial distribution converges to the Poisson. If a manufacturer has sold 5000 cell phones of a particular type, and each independently has probability 0.001 of having a warranty claim in a given week, then the number of such claims per week has approximately a Poisson distribution with mean $5000(0.001) = 5$.

7.1.2 Variance Stabilization and Least Squares with Count Data

Let $y_1, \ldots, y_n$ denote independent observations from Poisson distributions, with $\mu_i = E(y_i)$. In modeling count data, we could transform the counts so that, at least approximately, the variance is constant and ordinary least squares methods are valid. By the delta method, the linearization $g(y) - g(\mu) \approx (y-\mu)g'(\mu)$ implies that $\text{var}[g(y)] \approx [g'(\mu)]^2\text{var}(y)$. If y has a Poisson distribution, then $\sqrt{y}$ has

$$\text{var}(\sqrt{y}) \approx \left(\frac{1}{2\sqrt{\mu}}\right)^2 \mu = \frac{1}{4}.$$

The approximation holds better for larger μ, for which $\sqrt{y}$ is more closely linear in a neighborhood of μ.

Since $\sqrt{y}$ has approximately constant variance, we could model $\sqrt{y_i}$, $i = 1, \ldots, n$, using linear models fitted by ordinary least squares. However, the model is then linear

[1]See Karlin and Taylor (1975, pp. 23–25) for precise conditions and a derivation of the Poisson formula.

in $E(\sqrt{y_i})$, not $E(y_i)$. Also, a linear relation with the linear predictor may hold more poorly for $E(\sqrt{y_i})$ than for $E(y_i)$, $E[\log(y_i)]$, or some other transformation. It is more appealing to use GLM methods, which apply a link function to the mean response rather than the mean to a function of the response.

7.1.3 Poisson GLMs and Loglinear Models

We now present the GLM approach for Poisson response data. Since $\text{var}(y_i) = \mu_i$, the GLM likelihood equations (4.10) for n independent observations simplify for a Poisson response with linear predictor $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$ having link function g to

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\mu_i} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0.$$

Although a GLM can model a positive mean using the identity link, it is more common to model the log of the mean. Like the linear predictor, the log mean can take any real value. From Section 4.1.2, the log mean is the natural parameter for the Poisson distribution, and the log link is the canonical link for a Poisson GLM. The *Poisson loglinear model* is

$$\log \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{or} \quad \log \boldsymbol{\mu} = \boldsymbol{X\beta}$$

in terms of a model matrix and model parameters. For $\eta_i = \log \mu_i$, $\partial \mu_i / \partial \eta_i = \mu_i$, so the likelihood equations are

$$\sum_i (y_i - \mu_i)x_{ij} = 0, \tag{7.1}$$

as we found in Section 4.2.2.

For a Poisson loglinear model, the mean satisfies the exponential relation

$$\mu_i = \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) = (e^{\beta_1})^{x_{i1}} \cdots (e^{\beta_p})^{x_{ip}}.$$

A 1-unit increase in x_{ij} has a multiplicative impact of e^{β_j}: the mean at $x_{ij} + 1$ equals the mean at x_{ij} multiplied by e^{β_j}, adjusting for the other explanatory variables.

7.1.4 Model Fitting and Goodness of Fit

Except for simple models such as for the one-way layout or balanced two-way layout, the likelihood equations have no closed-form solution. However, the log-likelihood function is concave, and the Newton–Raphson method (which is equivalent to Fisher

scoring for the canonical log link) yields fitted values and estimates of corresponding model parameters. From Section 4.2.4, the estimated covariance matrix (4.14) of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1},$$

where with the log link $\boldsymbol{W}$ is the diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i) = \mu_i$.

From Section 4.4.2, the deviance of a Poisson GLM is

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right]. \tag{7.2}$$

When a model with log link has an intercept, its likelihood equation implies that $\sum_i \hat{\mu}_i = \sum_i y_i$, and so $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_i[y_i \log(y_i/\hat{\mu}_i)]$. This is often denoted by G^2. The corresponding Pearson statistic (Section 4.4.4) is

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

In some cases we can use these statistics to test goodness of fit. Asymptotic chi-squared distributions result when the number n of Poisson observations is fixed and their means increase unboundedly. The standard case where this holds reasonably well is contingency tables with a fixed number of cells and large overall sample size, as we explain in Section 7.2.2. But such a test, having a global alternative, does not reveal how a model fails. It is more informative to check a model by comparing it with more-complex models (e.g., with interaction terms) and by investigating particular aspects of lack of fit. For example, we can check that the variance truly has the same order of magnitude as the mean by comparing the model with a more-complex model that does not assume this, such as the negative binomial model presented in Section 7.3.

We can also search for unusual observations or patterns in the residuals. In Section 4.4.6 we presented the Pearson and standardized residuals for Poisson GLMs. Like y_i, these have skewed distributions, less so as μ_i increases. Finally, an informal way to assess the Poisson assumption is to compare the overall sample proportion of (0, 1, 2, …) observations to the average of the fitted response distributions for the n observations. Often this shows that a Poisson model does not permit sufficient variability, underpredicting 0 outcomes and relatively high outcomes.

7.1.5 Example: One-Way Layout Comparing Poisson Means

For the one-way layout for a count response, let y_{ij} be observation j of a count variable for group i, $i = 1, \ldots, c$, $j = 1, \ldots, n_i$, with $n = \sum_i n_i$. Suppose that $\{y_{ij}\}$ are independent Poisson with $E(y_{ij}) = \mu_{ij}$. The model $\log(\mu_{ij}) = \beta_0 + \beta_i$ has a common

mean within groups. For group means $\{\mu_i\}$, $\exp(\beta_h - \beta_i) = \mu_h/\mu_i$. With $\beta_0 = 0$ for identifiability, the model has the form $\log \boldsymbol{\mu} = \boldsymbol{X\beta}$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \mathbf{1}_{n_1} \\ \mu_2 \mathbf{1}_{n_2} \\ \vdots \\ \mu_c \mathbf{1}_{n_c} \end{pmatrix}, \quad \boldsymbol{X\beta} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_c} & \mathbf{0}_{n_c} & \cdots & \mathbf{1}_{n_c} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_c \end{pmatrix}.$$

The likelihood equation induced by parameter β_i for group i is

$$\sum_{j=1}^{n_i} (y_{ij} - \mu_i) = 0,$$

so $\hat{\mu}_i = \bar{y}_i = (\sum_j y_{ij})/n_i$ and $\hat{\beta}_i = \log \bar{y}_i$. In fact, the same likelihood equations and fitted means occur with any link function, or if we use baseline-category constraints (as we would for higher-way layouts) such as $\beta_1 = 0$. For the log link, $\widehat{\boldsymbol{W}}$ has the sample means on the main diagonal. For the model matrix shown above, $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$ is a diagonal matrix with $\widehat{\text{var}}(\hat{\beta}_i) = 1/n_i\bar{y}_i$. It follows that for large $\{n_i\mu_i\}$ a Wald 95% confidence interval for μ_h/μ_i is

$$\exp[(\hat{\beta}_h - \hat{\beta}_i) \pm 1.96\sqrt{(n_h\bar{y}_h)^{-1} + (n_i\bar{y}_i)^{-1}}].$$

Analogous to the one-way ANOVA for a normal response, we can test H_0: $\mu_1 = \cdots = \mu_c$. By direct construction or by applying the result in Section 4.4.3 about using the difference of deviances to compare the null model with the model for the one-way layout, we can construct the likelihood-ratio statistic. It equals

$$2\sum_{i=1}^{c} n_i\bar{y}_i \log(\bar{y}_i/\bar{y}),$$

where $\bar{y}$ is the grand mean of all $n = \sum_i n_i$ observations. As $\{n_i\}$ grow for fixed c, $\{\bar{y}_i\}$ have approximate normal distributions, and this statistic has null distribution converging to chi-squared with $df = (c - 1)$.

These inferences assume validity of the Poisson model. They are not robust to violation of the Poisson assumption. If the data have greater than Poisson variability, the large-sample $\text{var}(\hat{\beta}_i)$ will exceed $1/n_i\mu_i$. It is sensible to compare results with those for analogous inferences using a model that permits greater dispersion, such as the negative binomial model introduced in Section 7.3.

The deviance (7.2) and the Pearson statistic for the Poisson model for the one-way layout simplify to

$$G^2 = 2\sum_{i=1}^{c}\sum_{j=1}^{n_i} y_{ij} \log\left(\frac{y_{ij}}{\bar{y}_i}\right), \quad X^2 = \sum_{i=1}^{c}\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_i)^2}{\bar{y}_i}.$$

For testing goodness of fit of this model with relatively large $\{\bar{y}_i\}$, G^2 and X^2 have approximate chi-squared distributions with $df = \sum_i (n_i - 1)$ (Fisher 1970, p. 58). For a single group, Cochran (1954) referred to $[\sum_j (y_{1j} - \bar{y}_1)^2]/\bar{y}_1$ as the *variance test* for the fit of a Poisson distribution, since it compares the sample variance of the data with the estimated Poisson variance $\bar{y}_1$. This asymptotic theory applies, however, as $\{\mu_i\}$ grow for fixed $\{n_i\}$, which is not realistic in most applications. For checking fit, chi-squared asymptotics usually apply better for comparing the model with a more complex model and for comparing the model with the null model as just described.

7.1.6 Modeling Rates: Including an Offset in the Model

Often the expected value of a response count y_i is proportional to an index t_i. For instance, t_i might be an amount of time and/or a population size, such as in modeling crime counts for various cities. Or, it might be a spatial area, such as in modeling counts of a particular animal or plant species. Then the sample rate is y_i/t_i, with expected value μ_i/t_i. With explanatory variables, a loglinear model for the expected rate has the form

$$\log(\mu_i/t_i) = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Because $\log(\mu_i/t_i) = \log \mu_i - \log t_i$, the model makes the adjustment $-\log t_i$ to the log link of the mean. This adjustment term is called an *offset*. The fit corresponds to using $\log t_i$ as an explanatory variable in the linear predictor for $\log(\mu_i)$ and forcing its coefficient to equal 1.

For this model, the expected response count satisfies

$$\mu_i = t_i \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right).$$

The mean has a proportionality constant for t_i that depends on the values of the explanatory variables. The identity link is also occasionally useful, such as with a sole qualitative explanatory variable. The model with identity link is

$$\mu_i/t_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{or} \quad \mu_i = \sum_{j=1}^{p} \beta_j x_{ij} t_i.$$

This corresponds to an ordinary Poisson GLM using the identity link with no intercept and with explanatory variables $x_{i1}t_i, \ldots, x_{ip}t_i$. It provides additive, rather than multiplicative, effects of explanatory variables.

7.1.7 Example: Lung Cancer Survival

Table 7.1, from Holford (1980), shows survival and death for 539 males diagnosed with lung cancer. The prognostic factors are histology and stage of disease, with

Table 7.1 Number of Deaths from Lung Cancer, by Histology, Stage of Disease, and Follow-up Time Interval[a]

Follow-up Time Interval (months)	Disease Stage:	Histology								
		I			II			III		
		1	2	3	1	2	3	1	2	3
0–2		9	12	42	5	4	28	1	1	19
		(157	134	212	77	71	130	21	22	101)
2–4		2	7	26	2	3	19	1	1	11
		(139	110	136	68	63	72	17	18	63)
4–6		9	5	12	3	5	10	1	3	7
		(126	96	90	63	58	42	14	14	43)
6–8		10	10	10	2	4	5	1	1	6
		(102	86	64	55	42	21	12	10	32)
8–10		1	4	5	2	2	0	0	0	3
		(88	66	47	50	35	14	10	8	21)
10–12		3	3	4	2	1	3	1	0	3
		(82	59	39	45	32	13	8	8	14)
12+		1	4	1	2	4	2	0	2	3
		(76	51	29	42	28	7	6	6	10)

[a]Values in parentheses represent total follow-up months at risk.
Source: Reprinted from Holford (1980) with permission of John Wiley & Sons, Inc.

observations grouped into 2-month intervals of follow-up after the diagnosis. For each cell specifying a particular length of follow-up, histology, and stage of disease, the table shows the number of deaths and the number of months of observations of subjects still alive during that follow-up interval. We treat[2] the death counts in the table as independent Poisson variates.

Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and stage of disease j, in follow-up time interval k. The Poisson GLM for the death rate,

$$\log(\mu_{ijk}/t_{ijk}) = \beta_0 + \beta_i^{\mathrm{H}} + \beta_j^{\mathrm{S}} + \beta_k^{\mathrm{T}},$$

treats each explanatory variable as a qualitative factor, where the superscript notation shows the classification labels. It has residual deviance $G^2 = 43.92$ ($df = 52$). Models that assume a lack of interaction between follow-up interval and either prognostic factor are called *proportional hazards* models. They have the same effects of histology and stage of disease in each time interval. Then a ratio of hazards for two groups is the same at all times. Further investigation reveals that, although the stage of disease is an important prognostic factor, histology did not contribute significant additional

[2]This corresponds to a survival modeling approach that assumes piecewise exponential densities for survival times, yielding a constant hazard function in each two-month interval.

information. Adding interaction terms between stage and time does not significantly improve the fit (change in deviance = 14.86 with $df = 12$).

```
------------------------------------------------------------------------
> Cancer # file Cancer.dat at www.stat.ufl.edu/~aa/glm/data
   time histology stage count risktime
1     1         1     1     9      157
2     1         2     1     5       77
...
63    7         3     3     3       10
> attach(Cancer)
> logrisktime = log(risktime)
> fit <- glm(count ~ factor(histology) + factor(stage) + factor(time),
+        family = poisson(link = log), offset = logrisktime)
> summary(fit)
                  Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)        -3.0093      0.1665   -18.073    <2e-16
factor(histology)2   0.1624     0.1219     1.332    0.1828
factor(histology)3   0.1075     0.1474     0.729    0.4658
factor(stage)2       0.4700     0.1744     2.694    0.0070
factor(stage)3       1.3243     0.1520     8.709    <2e-16
factor(time)2       -0.1274     0.1491    -0.855    0.3926
...
factor(time)7       -0.1752     0.2498    -0.701    0.4832
---
    Null deviance: 175.718  on 62  degrees of freedom
Residual deviance:  43.923  on 52  degrees of freedom
------------------------------------------------------------------------
```

The estimated stage-of-disease effects show the progressively worsening death rate as the stage advances. The estimated death rate at the third stage of disease is $\exp(1.324) = 3.76$ times that at the first stage, adjusting for follow-up time and histology, with Wald 95% confidence interval $\exp[1.324 \pm 1.96(0.152)]$, or (2.79, 5.06).

7.2 POISSON/MULTINOMIAL MODELS FOR CONTINGENCY TABLES

Chapters 5 and 6 introduced binomial and multinomial models for categorical response variables. For *multivariate* categorical responses, we can apply such models marginally to each response, as Section 9.6 shows. Alternatively, we can formulate multinomial models for their joint distribution, to investigate potential independence, association, and interaction structure. Many such multinomial models are equivalent to models for independent Poisson counts in cells of a contingency table. The Poisson model generates the multinomial model after we condition on an overall sample size. We illustrate in this section.

7.2.1 Connection Between Poisson and Multinomial Distributions

For independent Poisson random variables $(y_1, \ldots, y_c)$ with means $(\mu_1, \ldots, \mu_c)$, the joint probability mass function for $\{y_i\}$ is the product of the mass functions of form (4.5). The total $n = \sum_i y_i$ also has a Poisson distribution, with parameter $\sum_i \mu_i$. Conditional on n, $\{y_i\}$ no longer have Poisson distributions, because each y_i cannot exceed n, and $\{y_i\}$ are also no longer independent, because the value of one affects the possible range for the others.

The conditional probability of a set of counts $\{y_i\}$ satisfying $\sum_j y_j = n$ is

$$P\left[(y_1 = n_1, y_2 = n_2, \ldots, y_c = n_c) \mid \sum_{j=1}^{c} y_j = n\right]$$
$$= \frac{P(y_1 = n_1, y_2 = n_2, \ldots, y_c = n_c)}{P\left(\sum_j y_j = n\right)}$$
$$= \frac{\prod_i (e^{-\mu_i} \mu_i^{n_i}/n_i!)}{\exp(-\sum_j \mu_j)(\sum_j \mu_j)^n/n!} = \left(\frac{n!}{\prod_i n_i!}\right) \prod_{i=1}^{c} \pi_i^{n_i},$$

where $\{\pi_i = \mu_i/(\sum_j \mu_j)\}$. This is the multinomial distribution characterized by the sample size n and the probabilities $\{\pi_i\}$.

Because of this relation, many Poisson models for independent counts in c fixed categories have corresponding multinomial models that treat the total count as fixed. In the multinomial model, the sample size is the total count and the category probabilities are proportional to the Poisson means.

7.2.2 GLM of Independence in Two-Way Contingency Tables

To illustrate Poisson loglinear models for counts in contingency tables, we first consider $r \times c$ tables that cross-classify two categorical response variables, which we denote by A and B. Suppose $\{y_{ij}\}$ are independent counts having Poisson distributions with means $\{\mu_{ij}\}$ that satisfy

$$\mu_{ij} = \mu \phi_i \psi_j,$$

where $\{\phi_i\}$ and $\{\psi_j\}$ are positive constants satisfying $\sum_i \phi_i = \sum_j \psi_j = 1$. This model is multiplicative, but the log link yields a GLM for $\{\mu_{ij}\}$ whose linear predictor has the structure

$$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B. \tag{7.3}$$

This Poisson loglinear model has additive main effects of the two classifications but no interaction. Identifiability requires a constraint on $\{\beta_i^A\}$ and on $\{\beta_j^B\}$.

Because $\{y_{ij}\}$ are independent, the total sample size $\sum_i \sum_j y_{ij}$ has a Poisson distribution with mean $\sum_i \sum_j \mu_{ij} = \mu$. Conditional on $\sum_i \sum_j y_{ij} = n$, the cell counts have a multinomial distribution with joint cell probabilities $\{\pi_{ij} = \mu_{ij}/\mu = \phi_i \psi_j\}$. Because $\sum_i \phi_i = 1$ and $\sum_j \psi_j = 1$, we have[3] $\phi_i = \pi_{i+}$, $\psi_j = \pi_{+j}$, and $\{\pi_{ij} = \pi_{i+}\pi_{+j}\}$. This is the expression of the multinomial joint distribution for *independence* between the categorical response variables. When we express the Poisson loglinear model (7.3) in multiplicative multinomial form by exponentiating both sides and dividing by μ, the intercept parameter β_0 cancels. That is, the Poisson model has $[1 + (r-1) + (c-1)]$ parameters, whereas the multinomial model has $[(r-1) + (c-1)]$ parameters.

As in the two-way layout for a linear model with main effects only, the model matrix $\boldsymbol{X}$ for the Poisson loglinear model has a simple form containing indicator variables that are the coefficients of the parameters for the row and column factors. For example, for a 2×2 table with constraints $\beta_1^A = \beta_1^B = 0$, the model is

$$\log \boldsymbol{\mu} = \begin{bmatrix} \log\ \mu_{11} \\ \log\ \mu_{12} \\ \log\ \mu_{21} \\ \log\ \mu_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_2^A \\ \beta_2^B \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta}.$$

From such a model matrix for the independence model, you can verify that the likelihood equations (7.1) simplify to $\hat{\mu}_{i+} = y_{i+}$ and $\hat{\mu}_{+j} = y_{+j}$, for all i and j. These equate the fitted and sample marginal distributions. Or we can easily derive these directly from the model. The joint Poisson probability of cell counts $\{y_{ij}\}$ is

$$\prod_{i=1}^{r} \prod_{j=1}^{c} \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!},$$

from which the kernel of the log-likelihood is

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{r} \sum_{j=1}^{c} y_{ij} \log \mu_{ij} - \sum_{i=1}^{r} \sum_{j=1}^{c} \mu_{ij}.$$

Substituting the model formula (7.3) for $\log \mu_{ij}$, we have

$$L\left(\beta_0, \boldsymbol{\beta}^A, \boldsymbol{\beta}^B\right) = n\beta_0 + \sum_{i=1}^{r} y_{i+}\beta_i^A + \sum_{j=1}^{c} y_{+j}\beta_j^B - \sum_{i=1}^{r} \sum_{j=1}^{c} \exp\left(\beta_0 + \beta_i^A + \beta_j^B\right).$$

[3] A + subscript denotes summing over that index.

The log-likelihood derivatives

$$\frac{\partial L}{\partial \beta_i^A} = y_{i+} - \sum_{j=1}^{c} \exp\left(\beta_0 + \beta_i^A + \beta_j^B\right) = y_{i+} - \mu_{i+} \quad \text{and}$$

$$\frac{\partial L}{\partial \beta_j^B} = y_{+j} - \sum_{i=1}^{r} \exp\left(\beta_0 + \beta_i^A + \beta_j^B\right) = y_{+j} - \mu_{+j}$$

yield these likelihood equations, when equated to 0. The solution of these equations that satisfies the model is the set of maximum likelihood (ML) fitted values, $\{\hat{\mu}_{ij} = y_{i+}y_{+j}/n\}$. The same fit results if we condition on $n = \sum_i \sum_j y_{ij}$ and maximize the corresponding multinomial likelihood $\prod_i \prod_j \pi_{ij}^{y_{ij}}$, for which the kernel of the log-likelihood, $\sum_i \sum_j y_{ij} \log \pi_{ij}$, is the same as the Poisson kernel except for the intercept parameter. The fitted joint multinomial probability $\hat{\pi}_{ij}$ is the product of the sample marginal proportions, $\hat{\pi}_{i+} = y_{i+}/n$ and $\hat{\pi}_{+j} = y_{+j}/n$.

The Pearson statistic for testing the independence-model goodness of fit,

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},$$

was proposed by Karl Pearson in 1900. When the model holds, the large-sample distributions of X^2 and the corresponding deviance G^2 are chi-squared, the approximation being reasonably good if most cell means exceed about 5. The Poisson model has rc observations described by $[1 + (r-1) + (c-1)]$ parameters. Equivalently, the multinomial model has $rc - 1$ counts described by $(r-1) + (c-1)$ parameters. So the residual df for the chi-squared test are $df = rc - (r + c - 1) = (r-1)(c-1)$. Pearson mistakenly concluded that $df = rc - 1$, as would be the case if H_0 specified particular values for $\{\pi_{ij}\}$. The correct df were not proven until an article by R. A. Fisher in 1922. This correction engendered a lifelong enmity[4] in which each of these giants of the Statistics community treated the other disparagingly.

7.2.3 Loglinear Association Parameters Relate to Odds Ratios

To allow association between the two classification variables, we add a two-factor interaction term to loglinear model (7.3), yielding

$$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B + \gamma_{ij}^{AB}.$$

We can specify the model so that $\{\gamma_{ij}^{AB}\}$ are coefficients of cross-products of $r-1$ indicator variables for the rows with $c-1$ indicator variables for the columns. With

[4]For details, see Agresti (2013, Section 17.2) and *R. A. Fisher: The Life of a Scientist* by Joan Fisher Box (Wiley 1978).

appropriate constraints for identifiability, such as $\gamma_{1j}^{AB} = \gamma_{i1}^{AB} = 0$ for all i and j, this adds an additional $(r-1)(c-1)$ parameters, so the model is saturated.

The $\{\gamma_{ij}^{AB}\}$ association parameters pertain to odds ratios. We illustrate for $r = c = 2$. For the multinomial $\{\pi_{ij}\}$ or the Poisson $\{\mu_{ij}\}$, the log odds ratio is

$$\begin{aligned}\log \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} &= \log \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = \log \mu_{11} + \log \mu_{22} - \log \mu_{12} - \log \mu_{21} \\ &= \left(\beta_0 + \beta_1^A + \beta_1^B + \gamma_{11}^{AB}\right) + \left(\beta_0 + \beta_2^A + \beta_2^B + \gamma_{22}^{AB}\right) \\ &\quad - \left(\beta_0 + \beta_1^A + \beta_2^B + \gamma_{12}^{AB}\right) - \left(\beta_0 + \beta_2^A + \beta_1^B + \gamma_{21}^{AB}\right) \\ &= \gamma_{11}^{AB} + \gamma_{22}^{AB} - \gamma_{12}^{AB} - \gamma_{21}^{AB}.\end{aligned}$$

Under the constraints just stated, the odds ratio simplifies to $\exp(\gamma_{22}^{AB})$.

7.2.4 Poisson/Multinomial Loglinear Models for Multiway Contingency Tables

Loglinear models for multidimensional contingency tables describe independence, association, and interaction patterns. We illustrate for $r \times c \times \ell$ cross-classifications of three categorical response variables, which we denote by A, B, and C. The models apply to Poisson sampling with independent cell counts $\{y_{ijk}\}$ having means $\{\mu_{ijk}\}$. They also apply to a multinomial distribution with cell probabilities $\{\pi_{ijk}\}$ having $\sum_i \sum_j \sum_k \pi_{ijk} = 1.0$.

Mutual independence: Three categorical response variables are *mutually independent* when the cell probabilities satisfy, for all i, j, and k,

$$P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k).$$

That is, all $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$. For expected frequencies $\{\mu_{ijk}\}$, mutual independence has the loglinear structural form

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C. \tag{7.4}$$

Joint independence: A is *jointly independent* of B and C when for all i, j, and k,

$$P(A = i, B = j, C = k) = P(A = i)P(B = j, C = k).$$

That is, all $\pi_{ijk} = \pi_{i++}\pi_{+jk}$. This is ordinary independence for the two-way contingency table that cross-classifies A with a variable composed of the $c\ell$ combinations of levels of B and C. The corresponding loglinear model is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}. \tag{7.5}$$

We use the hierarchical structure by which the presence of a two-factor term implies inclusion of the lower order (single-factor) terms.

Conditional independence: A and B are *conditionally independent, given* C, when for all i, j, and k,

$$P(A = i, B = j \mid C = k) = P(A = i \mid C = k)P(B = j \mid C = k).$$

That is, independence holds for the $r \times c$ partial table relating A and B at each fixed category of C. Equivalently, by expressing each conditional probability in terms of the joint probabilities $\{\pi_{ijk}\}$ and their marginals,

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

Conditional independence of A and B, given C, has loglinear model form

$$\log\ \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}. \tag{7.6}$$

A model that permits all three pairs of variables to be conditionally dependent is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}. \tag{7.7}$$

From exponentiating both sides, the cell probabilities have the form

$$\pi_{ijk} = \phi_{ij}\psi_{ik}\omega_{jk}.$$

No closed-form expression exists for the three components in terms of margins of $\{\pi_{ijk}\}$ except in certain special cases. All these loglinear models have constraints on parameters to satisfy identifiability. For example, for any conditional association term, we can take $\gamma_{1j} = \gamma_{i1} = 0$ for all i and j, as R software does by default.

Interpretations of loglinear model parameters use their highest-order terms. The two-factor terms describe conditional association as measured by log odds ratios. At a fixed level k of C, the *conditional association* between A and B is specified by $(r-1)(c-1)$ odds ratios, such as

$$\theta_{ij(k)} = \frac{\mu_{ijk}\mu_{rck}}{\mu_{ick}\mu_{rjk}}, \qquad 1 \le i \le r-1, \quad 1 \le j \le c-1.$$

For example, when $r = c = 2$, substituting model (7.7) into $\log \theta_{11(k)}$ yields

$$\log \theta_{11(k)} = \log \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} = \gamma_{11}^{AB} + \gamma_{22}^{AB} - \gamma_{12}^{AB} - \gamma_{21}^{AB}.$$

Thus, $\theta_{11(k)}$ simplifies to $\exp(\gamma_{22}^{AB})$ under constraints such as R software imposes. Analogous expressions occur with arbitrary r and c. In such expressions, because the right-hand side is the same for all k, an absence of three-factor interaction is equivalent to

$$\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(\ell)} \quad \text{for all } i \text{ and } j.$$

Because of this property, model (7.7) is called a loglinear model of *homogeneous association*. Any loglinear model not having the three-factor interaction term has a homogeneous conditional association for each pair of variables.

The general Poisson loglinear model for a three-way contingency table is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ijk}^{ABC}.$$

With indicator variables for each factor, γ_{ijk}^{ABC} is the coefficient of the product of the ith indicator variable for A, jth indicator variable for B, and kth indicator variable for C. The total number of nonredundant parameters is

$$1 + (r-1) + (c-1) + (\ell-1) + (r-1)(c-1) + (r-1)(\ell-1)$$
$$+ (c-1)(\ell-1) + (r-1)(c-1)(\ell-1) = rc\ell,$$

which is the total number of cell counts. This model has as many parameters as Poisson observations and is saturated. It describes all possible $\{\mu_{ijk} > 0\}$.

Table 7.2 summarizes unsaturated loglinear models for three-way contingency tables. For all such Poisson models, corresponding multinomial models have one fewer parameter (the β_0 intercept in the Poisson models) after conditioning on the total count. The common parameters contribute in the same way to Poisson or

Table 7.2 Loglinear Models for Three-Way Contingency Tables, for Poisson Means or Multinomial Probabilities $\{\pi_{ijk}\}$

Model Formula	Probability Form for π_{ijk}	Association Terms in Loglinear Model	Interpretation
(7.4)	$\pi_{i++}\pi_{+j+}\pi_{++k}$	None	A, B, C mutually independent
(7.5)	$\pi_{i++}\pi_{+jk}$	γ_{jk}^{BC}	A jointly independent of B and C
(7.6)	$\pi_{i+k}\pi_{+jk}/\pi_{++k}$	$\gamma_{ik}^{AC} + \gamma_{jk}^{BC}$	A, B conditionally indep., given C
(7.7)	$\phi_{ij}\psi_{ik}\omega_{jk}$	$\gamma_{ij}^{AB} + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$	Homogeneous association

multinomial likelihoods and have the same ML estimates and *SE* values. The fit and the X^2 and G^2 goodness-of-fit statistics are identical for the Poisson and multinomial formulations.

Because the model matrices for these loglinear models contain indicator variables and their products, the likelihood equations (7.1) take the simple form of equating the observed counts to the fitted values in the margins of the contingency table that correspond to the highest-order terms in the model. For example, the mutual independence model (7.4) has likelihood equations, for all i, j, and k,

$$y_{i++} = \hat{\mu}_{i++}, \quad y_{+j+} = \hat{\mu}_{+j+}, \quad y_{++k} = \hat{\mu}_{++k},$$

whereas the homogeneous association model (7.7) has likelihood equations

$$y_{ij+} = \hat{\mu}_{ij+}, \quad y_{i+k} = \hat{\mu}_{i+k}, \quad y_{+jk} = \hat{\mu}_{+jk}.$$

It is straightforward to derive these, much as we did for the independence model in Section 7.2.2. For many models having some independence structure, closed-form solutions exist. In all cases the Newton–Raphson method, which is equivalent to Fisher scoring for these canonical-link models, yields fitted values and corresponding model parameter estimates. When cell means mostly exceed about 5, X^2 and G^2 statistics have approximate chi-squared null distributions for testing the model goodness of fit. Standardized residuals can detect particular cells for which the fit is poor.

7.2.5 Connections Between Logistic and Loglinear Models

Loglinear models for contingency tables treat all categorical classifications symmetrically and regard the cell count as the response. They are useful for modeling the joint distribution of categorical variables. By contrast, logistic models distinguish between response and explanatory classifications. Although different in purpose, the two types of models are connected.

We illustrate with the homogeneous association loglinear model (7.7). Suppose we treat A as a response variable and B and C as explanatory, conditioning on $\{n_{+jk}\}$. For the binary case $r = 2$, we are then modeling $c\ell$ binomial distributions on A. When we construct the logit for each binomial distribution of A, we obtain

$$\begin{aligned}
\log \frac{P(A=1 \mid B=j,\ C=k)}{P(A=2 \mid B=j,\ C=k)} &= \log \frac{\mu_{1jk}}{\mu_{2jk}} = \log \mu_{1jk} - \log \mu_{2jk} \\
&= \left(\beta_0 + \beta_1^A + \beta_j^B + \beta_k^C + \gamma_{1j}^{AB} + \gamma_{1k}^{AC} + \gamma_{jk}^{BC}\right) \\
&\quad - \left(\beta_0 + \beta_2^A + \beta_j^B + \beta_k^C + \gamma_{2j}^{AB} + \gamma_{2k}^{AC} + \gamma_{jk}^{BC}\right) \\
&= \left(\beta_1^A - \beta_2^A\right) + \left(\gamma_{1j}^{AB} - \gamma_{2j}^{AB}\right) + \left(\gamma_{1k}^{AC} - \gamma_{2k}^{AC}\right).
\end{aligned}$$

The first parenthetical term is a constant, not depending on j or k. The second parenthetical term depends on the category j of B. The third parenthetical term depends on the category k of C. This logit has the additive form

$$\text{logit}[P(A = 1 \mid B = j,\ C = k)] = \lambda + \delta_j^B + \delta_k^C.$$

In fact, the Poisson loglinear model and the binomial logistic model have the same likelihood equations and the same fit. An analogous correspondence holds when A has several categories, using a multinomial baseline-category logit model for A in terms of additive factor effects for B and C.

The loglinear model that has the same fit as a logistic model with factors as explanatory variables contains a general interaction term for relations among those explanatory variables. The logistic model does not assume anything about relations among explanatory variables, so it allows an arbitrary interaction pattern for them. For example, for a main-effects logistic model that predicts A using factors B, C, and D, the corresponding loglinear model has pairwise associations between A and B, A and C, and A and D, as well as the BCD three-factor interaction term and all its lower-order relatives.

7.2.6 Example: Loglinear Models for Student Substance Use

Table 7.3 refers to a survey by Wright State University that asked 2276 students in their final year of high school in a rural area near Dayton, Ohio whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this $2 \times 2 \times 2$ table by A, C, and M.

Table 7.4 shows results of testing fit for four loglinear models. Models that lack any association term fit poorly. The homogeneous association model fits well. It is suggested by other criteria also, such as minimizing AIC.

The following output shows some results from fitting the homogeneous association model. The AC fitted conditional odds ratios at each level of M equal $\exp(\hat{\gamma}_{11}^{AC} + \hat{\gamma}_{22}^{AC} - \hat{\gamma}_{12}^{AC} - \hat{\gamma}_{21}^{AC})$, which is $\exp(\hat{\gamma}_{22}^{AC}) = e^{2.0545} = 7.80$ for the R constraints. For those who have used cigarettes, the odds of having used alcohol are estimated to be 7.80 times the odds of having used alcohol for those who have not used cigarettes, and this applies

Table 7.3 Alcohol, Cigarette, and Marijuana Use Among High School Seniors

Alcohol Use (A)	Cigarette Use (C)	Marijuana Use (M)	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Data courtesy of Harry Khamis, Wright State University.

Table 7.4 Goodness-of-Fit Tests for Loglinear Models Fitted to the Data in Table 7.3

Loglinear Associations	Deviance G^2	Pearson X^2	df	P-value[a]	AIC
$\gamma_{ij}^{AC} + \gamma_{ik}^{AM}$	497.37	443.76	2	< 0.001	558.4
$\gamma_{ij}^{AC} + \gamma_{jk}^{CM}$	92.02	80.81	2	< 0.001	153.1
$\gamma_{ik}^{AM} + \gamma_{jk}^{CM}$	187.75	177.61	2	< 0.001	248.8
$\gamma_{ij}^{AC} + \gamma_{ik}^{AM} + \gamma_{jk}^{CM}$	0.37	0.40	1	0.54	63.4

[a] P-value for G^2 statistic.

both for those who have used marijuana and those who have not. The corresponding Wald 95% confidence interval is $\exp[2.0545 \pm 1.96(0.1741)]$, or (5.5, 11.0).

```
-------------------------------------------------------------------------
> Drugs  # file Drugs.dat at www.stat.ufl.edu/~aa/glm/data
    A   C   M count
1 yes yes yes   911
2 yes yes  no   538
...
8  no  no  no   279
> attach(Drugs)
> alc <- factor(A); cig <- factor(C); mar <- factor(M)
> mutual.indep <- glm(count ~ alc + cig + mar, family=poisson(link=log))
> homo.assoc <- update(mutual.indep, .~. + alc:cig + alc:mar + cig:mar)
> summary(homo.assoc)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   6.8139      0.0331  205.699   < 2e-16
alc2         -5.5283      0.4522  -12.225   < 2e-16
cig2         -3.0157      0.1516  -19.891   < 2e-16
mar2         -0.5249      0.0543   -9.669   < 2e-16
alc2:cig2     2.0545      0.1741   11.803   < 2e-16 # odds ratio  7.8
alc2:mar2     2.9860      0.4647    6.426  1.31e-10 # odds ratio 19.8
cig2:mar2     2.8479      0.1638   17.382   < 2e-16 # odds ratio 17.3
---
Residual deviance:    0.3740  on 1  degrees of freedom
AIC: 63.417
-------------------------------------------------------------------------
```

For a loglinear model with residual $df = 1$, each standardized residual has the same absolute value and has square equal to the Pearson X^2 statistic for testing goodness of fit. The Pearson residuals are less appealing: they have eight separate values, even though $|y_{ijk} - \hat{\mu}_{ijk}|$ is identical for each cell (because the likelihood equations imply that the two-way observed and fitted marginal tables are identical) and the residual $df = 1$.

```
-------------------------------------------------------------------------
> pearson.resid <- resid(homo.assoc, type="pearson")
> std.resid <- rstandard(homo.assoc, type="pearson")
> sum(pearson.resid^2)  # Pearson chi-squared statistic
[1] 0.4011006
```

```
> cbind(count, fitted(homo.assoc), pearson.resid, std.resid)
  count  fitted(homo.assoc)  pearson.resid  std.resid
1   911             910.383          0.020      0.633
2   538             538.617         -0.027     -0.633
3    44              44.617         -0.092     -0.633
4   456             455.383          0.029      0.633
5     3               3.617         -0.324     -0.633
6    43              42.383          0.095      0.633
7     2               1.383          0.524      0.633
8   279             279.617         -0.037     -0.633
-------------------------------------------------------------------------
```

Using a logistic model, we find the same results for the association between marijuana use and each of alcohol use and cigarette use (i.e., estimated log odds ratios of 2.99 and 2.85). We model the logit of the probability of using marijuana with additive effects for alcohol use and cigarette use, treating the data as four binomial observations instead of eight Poisson observations.

```
-------------------------------------------------------------------------
> Drugs2  # file Drugs2.dat at www.stat.ufl.edu/~aa/glm/data
    A    C  M_yes  M_no     n  # data entered as 4 binomials
1 yes  yes    911   538  1449
2 yes   no     44   456   500
3  no  yes      3    43    46
4  no   no      2   279   281
> attach(Drugs2)
> alc <- factor(A); cig <- factor(C)
> fit.logistic <- glm(M_yes/n ~ alc + cig, weights=n, # specify weights
     family = binomial(link = logit)) # when enter proportion responses
> summary(fit.logistic)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -5.3090      0.4752  -11.172   < 2e-16
alcyes        2.9860      0.4647    6.426  1.31e-10 # odds ratio 19.8
cigyes        2.8479      0.1638   17.382   < 2e-16 # odds ratio 17.3
---
    Null deviance: 843.8266  on 3  degrees of freedom
Residual deviance:   0.3740  on 1  degrees of freedom
-------------------------------------------------------------------------
```

The null logistic model in this case is equivalent to the loglinear model by which marijuana use is jointly independent of alcohol use and cigarette use.

7.2.7 Graphical Loglinear Models: Portraying Conditional Independence Structure

Many loglinear models have graphical portrayals of the conditional independence structure among the responses. This representation also helps to reveal implications of models, such as when an association is unchanged when a variable is dropped from an analysis.

From graph theory, an undirected graph consists of a set of vertices and a set of edges connecting some vertices. In a probabilistic *conditional independence graph*, each vertex represents a variable, and the absence of an edge connecting two variables represents conditional independence between them. For instance, Figure 7.1 portrays the conditional independence graph for categorical response variables A, B, C, and D and the loglinear model that assumes independence between A and C and between A and D, conditional on the other two variables. The four variables form the vertices, and the four edges represent pairwise conditional associations. Edges do not connect A and C or connect A and D, the conditionally independent pairs.

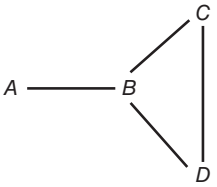

Figure 7.1 Conditional independence graph for the loglinear model that assumes conditional independence between A and C and between A and D.

Darroch et al. (1980) used undirected graphs to represent *graphical models*, which are essentially loglinear models for contingency tables that have a conditional independence structure. The graphical model corresponding to Figure 7.1 is the loglinear model having the three-factor BCD interaction and the two-factor AB association and their lower-order relatives, namely,

$$\log \mu_{hijk} = \beta_0 + \beta_h^A + \beta_i^B + \beta_j^C + \beta_k^D + \gamma_{hi}^{AB} + \gamma_{ij}^{BC} + \gamma_{ik}^{BD} + \gamma_{jk}^{CD} + \gamma_{ijk}^{BCD}.$$

A *path* in a conditional independence graph is a sequence of edges between one variable and another. Two variables A and C are said to be *separated* by a subset of variables if all paths connecting A and C intersect that subset. For instance, in Figure 7.1, B separates A and C. The subset $\{B, D\}$ also separates A and C. *Markov properties* that pertain to paths and separation allow us to deduce from the graph the conditional independence structure between variables and groups of variables. One such property, the *global Markov property*, states that two variables are conditionally independent given *any* subset of variables that separates them in the graph. Thus, in Figure 7.1, not only are A and C conditionally independent given B and D, but also given B alone. Similarly, A and D are conditionally independent given B alone. This property is equivalent to a *local Markov property*, according to which a variable is conditionally independent of all other variables, given the adjacent neighbors to which it is connected by an edge.

Conditional associations usually differ[5] from marginal associations. Under certain *collapsibility conditions*, however, they are the same. For loglinear and logistic models, the association parameters pertain to odds ratios, so such conditions relate to

[5]Recall, for example, Section 3.4.3 and Simpson's paradox.

equality of conditional and marginal odds ratios. Bishop et al. (1975, p. 47) provided a parametric collapsibility condition for multiway contingency tables:

> Suppose that a model for a multiway contingency table partitions variables into three mutually exclusive subsets, $\{S_1, S_2, S_3\}$, such that S_2 separates S_1 and S_3. After collapsing the table over the variables in S_3, parameters relating variables in S_1 and parameters relating variables in S_1 to variables in S_2 are unchanged.

For the graphical model corresponding to Figure 7.1, let $S_1 = \{A\}, S_2 = \{B\}$, and $S_3 = \{C, D\}$. Since the AC and AD terms do not appear in the model, all parameters linking set S_1 with set S_3 equal zero, and S_2 separates S_1 and S_3. If we collapse over C and D, the AB association is unchanged. Next, identify $S_1 = \{C, D\}$, $S_2 = \{B\}$, $S_3 = \{A\}$. Then conditional associations among B, C, and D remain the same after collapsing over A. By contrast, the homogeneous loglinear model that provides a good fit for the student substance use data of Table 7.3 does not satisfy collapsibility conditions. The fitted (AC, AM, CM) conditional odds ratios of (7.8, 19.8, 17.3) obtained by exponentiating the log odds ratio estimates from the R output differ from the corresponding two-way marginal odds ratios, (17.7, 61.9, 25.1).

7.3 NEGATIVE BINOMIAL GLMS

For the Poisson distribution, the variance equals the mean. In practice, count observations often exhibit variability exceeding that predicted by the Poisson. This phenomenon is called *overdispersion*.

7.3.1 Overdispersion for a Poisson GLM

A common reason for overdispersion is heterogeneity: at fixed levels of the explanatory variables, the mean varies according to values of unobserved variables. For example, for the horseshoe crab dataset introduced in Section 1.5.1 that we analyze further in Section 7.5, suppose that a female crab's carapace width, weight, color, and spine condition are the four explanatory variables that affect her number of male satellites. Suppose that y has a Poisson distribution at each fixed combination of those variables, but we use the model that has weight alone as an explanatory variable. Crabs having a certain weight are then a mixture of crabs of various widths, colors, and spine conditions. Thus, the population of crabs having that weight is a mixture of several Poisson populations, each having its own mean for the response. This heterogeneity results in an overall response distribution at that weight having greater variation than the Poisson. If the variance equals the mean when *all* relevant explanatory variables are included, it exceeds the mean when only *some* are included. Another severe limitation of Poisson GLMs is that, because the variance of y must equal the mean, at a fixed mean the variance cannot decrease as additional explanatory variables enter the model.

Overdispersion is not an issue in ordinary linear models that assume normally distributed y, because that distribution has a separate variance parameter to describe

variability. For Poisson and binomial distributions, however, the variance is a function of the mean. Overdispersion is common in the modeling of counts. Suppose the model for the mean has the correct link function and linear predictor, but the true response distribution has more variability than the Poisson. Then the ML estimators of model parameters assuming a Poisson response are still consistent, converging in probability to the parameter values, but standard errors are too small. Extensions of the Poisson GLM that have an extra parameter account better for overdispersion. We present one such extension here, and Sections 7.4, 8.1, and 9.4 present others.

7.3.2 Negative Binomial as a Gamma Mixture of Poissons

A mixture model is a flexible way to account for overdispersion. At a fixed setting of the explanatory variables actually observed, given the mean λ, suppose the distribution of y is Poisson(λ), but λ itself varies because of unmeasured covariates. Let $\mu = E(\lambda)$. Then unconditionally,

$$E(y) = E[E(y \mid \lambda)] = E(\lambda) = \mu,$$

$$\text{var}(y) = E[\text{var}(y \mid \lambda)] + \text{var}[E(y \mid \lambda)] = E(\lambda) + \text{var}(\lambda) = \mu + \text{var}(\lambda) > \mu.$$

Here is an important example of a mixture model for count data: suppose that (1) given λ, y has a Poisson(λ) distribution, and (2) λ has the gamma distribution (4.29). Recall that the gamma distribution has $E(\lambda) = \mu$ and $\text{var}(\lambda) = \mu^2/k$ for a shape parameter $k > 0$, so the standard deviation is proportional to the mean. Marginally, the gamma mixture of the Poisson distributions yields the *negative binomial distribution* for y. Its probability mass function is

$$p(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \quad y = 0, 1, 2, \ldots . \tag{7.8}$$

With k fixed, this is a member of an exponential dispersion family appropriate for discrete variables (Exercise 7.23), with natural parameter $\log[\mu/(\mu+k)]$.

In the two-parameter negative binomial family, let $\gamma = 1/k$. Then y has

$$E(y) = \mu, \quad \text{var}(y) = \mu + \gamma\mu^2.$$

The index $\gamma > 0$ is a type of dispersion parameter. The greater the value of γ, the greater the overdispersion relative to the Poisson. As $\gamma \to 0$, $\text{var}(y) \to \mu$ and the negative binomial distribution converges[6] to the Poisson.

The negative binomial distribution has much greater scope than the Poisson. For example, the Poisson mode is the integer part of the mean and equals 0 only when $\mu < 1$. The negative binomial is also unimodal, but the mode is 0 when $\gamma \geq 1$ and otherwise it is the integer part of $\mu(1-\gamma)$. The mode can be 0 for any μ.

[6]For a proof, see Cameron and Trivedi (2013, p. 85).

7.3.3 Negative Binomial GLMs

Negative binomial GLMs commonly use the log link, as in Poisson loglinear models, rather than the canonical link. For simplicity, we let the dispersion parameter γ be the same constant for all n observations but treat it as unknown, much like the variance in normal models. This corresponds to a constant coefficient of variation in the gamma mixing distribution, $\sqrt{\text{var}(\lambda)}/E(\lambda) = \sqrt{\gamma}$.

From Equation (7.8) expressed in terms of the dispersion parameter γ, the log-likelihood function for a negative binomial GLM with n independent observations is

$$L(\boldsymbol{\beta}, \gamma; \boldsymbol{y}) = \sum_{i=1}^{n} \left[\log\Gamma\left(y_i + \frac{1}{\gamma}\right) - \log\Gamma\left(\frac{1}{\gamma}\right) - \log\Gamma(y_i + 1) \right] + \sum_{i=1}^{n} \left[y_i \log\left(\frac{\gamma\mu_i}{1+\gamma\mu_i}\right) - \left(\frac{1}{\gamma}\right)\log(1+\gamma\mu_i) \right],$$

where μ_i is a function of $\boldsymbol{\beta}$ through $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$ with the link function g. The likelihood equations obtained by differentiating $L(\boldsymbol{\beta}, \gamma; \boldsymbol{y})$ with respect to $\boldsymbol{\beta}$ have the usual form (4.10) for a GLM,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(y_i)} \left(\frac{\partial\mu_i}{\partial\eta_i}\right) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{\mu_i + \gamma\mu_i^2} \left(\frac{\partial\mu_i}{\partial\eta_i}\right) = 0, \quad j = 1, 2, \ldots, p.$$

The log-likelihood yields a Hessian matrix that has

$$\frac{\partial^2 L(\boldsymbol{\beta}, \gamma; \boldsymbol{y})}{\partial\beta_j \partial\gamma} = -\sum_i \frac{(y_i - \mu_i)x_{ij}}{(1+\gamma\mu_i)^2} \left(\frac{\partial\mu_i}{\partial\eta_i}\right).$$

Thus, $E(\partial^2 L/\partial\beta_j\partial\gamma) = 0$ for each j, and $\boldsymbol{\beta}$ and γ are orthogonal parameters (Recall Section 4.2.4). So $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$ are asymptotically independent, and the large-sample SE for $\hat{\beta}_j$ is the same whether γ is known or estimated.

The iteratively reweighted least squares algorithm for Fisher scoring applies for ML model fitting. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1},$$

where, with log link, $\boldsymbol{W}$ is the diagonal matrix with $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i) = \mu_i/(1+\gamma\mu_i)$. The deviance for a negative binomial GLM is

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_i \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - \left(y_i + \frac{1}{\hat{\gamma}}\right)\log\left(\frac{1+\hat{\gamma}y_i}{1+\hat{\gamma}\hat{\mu}_i}\right) \right].$$

This is close to the Poisson GLM deviance (7.2) when $\hat{\gamma}$ is near 0.

7.3.4 Comparing Poisson and Negative Binomial GLMs

How can we compare Poisson and negative binomial GLMs that have the same explanatory variables, to determine whether the negative binomial model gives a better fit? An informal comparison can be based on AIC values. For a formal significance test, we can test H_0: $\gamma = 0$, because the Poisson is the limiting case of the negative binomial as $\gamma \downarrow 0$.

Since γ is positive, $\gamma = 0$ on the boundary of the parameter space. Thus, the likelihood-ratio statistic does not have an asymptotic null chi-squared distribution. Rather, it is an equal mixture of a single-point distribution at 0 (which occurs when $\hat{\gamma} = 0$) and chi-squared with $df = 1$. The P-value is half that from treating the statistic as chi-squared with $df = 1$ (Self and Liang 1987).

7.3.5 Negative Binomial Model with Variance Proportional to Mean

An alternative negative binomial parameterization results from writing the gamma density formula with $k\mu$ as the shape parameter,

$$f(\lambda; k, \mu) = \frac{k^{k\mu}}{\Gamma(k\mu)} \exp(-k\lambda)\lambda^{k\mu-1} \quad \lambda \geq 0,$$

so $E(\lambda) = \mu$ and $\text{var}(\lambda) = \mu/k$. For this parameterization, the gamma mixture of Poisson distributions yields a negative binomial distribution with

$$E(y) = \mu, \quad \text{var}(y) = \mu(1 + k)/k.$$

The variance is now linear rather than quadratic in μ. It corresponds to an inflation of the Poisson variance, converging to it as $k \to \infty$.

The two parameterizations of the negative binomial are sometimes denoted by NB1 (linear) and NB2 (quadratic). Only the NB2 falls within the traditional GLM framework, being expressible as an exponential dispersion family distribution, and it is much more commonly used. Unlike the NB2 model, for an NB1 model $\boldsymbol{\beta}$ and k are not orthogonal parameters, and $\hat{\boldsymbol{\beta}}$ is not a consistent estimator when the model for the mean holds but the true distribution is not negative binomial (Cameron and Trivedi 2013, Section 3.3). Lee and Nelder (1996) presented ML model fitting for NB1 models.

7.4 MODELS FOR ZERO-INFLATED DATA

In practice, the frequency of 0 outcomes is often larger than expected under standard discrete models. In particular, because the mode of a Poisson distribution is the integer part of the mean, a Poisson GLM is inadequate when means can be relatively large but the modal response is 0. Such data, which are *zero-inflated* relative to data expected for a Poisson GLM, are common when many subjects have a 0 response and many also have much larger responses, so the overall mean is not near 0. An

example of a variable that might be zero-inflated is the number of times in the past week that individuals report exercising, such as by going to a gym. Some would do so frequently, some would do it occasionally but not in the past week (a random 0), and a substantial percentage would never do so, causing zero inflation. Other examples are counts of activities for which many subjects would necessarily report 0, such as the number of times during some period of having an alcoholic drink, or smoking marijuana, or having sexual intercourse.

Zero-inflation is less problematic for negative binomial GLMs, because that distribution can have a mode of 0 regardless of the value of the mean. However, a negative binomial model fits poorly when the data are strongly bimodal, with a mode at zero and a separate mode around some considerably higher value. This could occur for the frequency of an activity in which many subjects never participate but many others quite often do. Then a substantial fraction of the population necessarily has a zero outcome, and the remaining fraction follows some distribution that may have small probability of a zero outcome.

7.4.1 Zero-Inflated Poisson and Negative Binomial Models

The representation just mentioned, of one set of observations that necessarily are zero and another set that may be zero according to a random event, leads naturally to a mixture model in which two types of zeros can occur. The relevant distribution is a mixture of an ordinary count model such as the Poisson or negative binomial with one that places all its mass at zero.

The *zero-inflated Poisson (ZIP) model* (Lambert 1992) assumes that

$$y_i \sim \begin{cases} 0 & \text{with probability } 1-\phi_i \\ \text{Poisson}(\lambda_i) & \text{with probability } \phi_i. \end{cases}$$

The unconditional probability distribution has

$$P(y_i = 0) = (1-\phi_i) + \phi_i e^{-\lambda_i},$$

$$P(y_i = j) = \phi_i \frac{e^{-\lambda_i}\lambda_i^j}{j!}, \qquad j = 1, 2, \ldots .$$

Explanatory variables affecting ϕ_i need not be the same as those affecting λ_i. The parameters could be modeled by

$$\text{logit}(\phi_i) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\lambda_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2.$$

A latent class construction that yields this model posits an unobserved binary variable z_i. When $z_i = 0$, $y_i = 0$, and when $z_i = 1$, y_i is a Poisson(λ_i) variate. For this mixture distribution,

$$E(y_i) = E[E(y_i \mid z_i)] = (1-\phi_i)0 + \phi_i(\lambda_i) = \phi_i\lambda_i.$$

Also, because $E[\text{var}(y_i \mid z_i)] = (1-\phi_i)0 + \phi_i(\lambda_i) = \phi_i\lambda_i$ and $\text{var}[E(y_i \mid z_i)] = (1-\phi_i)(0-\phi_i\lambda_i)^2 + \phi_i(\lambda_i - \phi_i\lambda_i)^2 = \lambda_i^2\phi_i(1-\phi_i)$,

$$\text{var}(y_i) = E[\text{var}(y_i \mid z_i)] + \text{var}[E(y_i \mid z_i)] = \phi_i\lambda_i[1 + (1-\phi_i)\lambda_i].$$

Since $\text{var}(y_i) > E(y_i)$, overdispersion occurs relative to a Poisson model.

When λ_i and ϕ_i are not functionally related, the joint log-likelihood function for the two parts of the model is

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \sum_{y_i=0} \log[1 + e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1} \exp(-e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2})] - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1})$$

$$+ \sum_{y_i>0} [\boldsymbol{x}_{1i}\boldsymbol{\beta}_1 + y_i\boldsymbol{x}_{2i}\boldsymbol{\beta}_2 - e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2} - \log(y_i!)].$$

Lambert (1992) expressed the log-likelihood in terms of the latent variables $\{z_i\}$. She used the EM algorithm for ML fitting, treating each z_i as a missing value. Alternatively, the Newton–Raphson method can be used.

A disadvantage of the ZIP model is the larger number of parameters compared with ordinary Poisson or negative binomial models. Sometimes the explanatory variables in the two parts of the model are the same, and their effects have similar relative size. For such cases, Lambert proposed a simpler model in which $\boldsymbol{x}_{1i} = \boldsymbol{x}_{2i}$ and $\boldsymbol{\beta}_2 = \tau\boldsymbol{\beta}_1$ for a shape parameter τ. Another disadvantage of the general ZIP model is that the parameters do not directly describe the effects of explanatory variables on $E(y_i) = \phi_i\lambda_i$, because $\boldsymbol{\beta}_1$ pertains to effects on ϕ_i and $\boldsymbol{\beta}_2$ pertains to effects on λ_i. In addition, when $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are the same or overlap substantially, the correlation between them could cause further problems with interpretation. A simpler alternative fits only an intercept in the model for ϕ_i. In that case $E(y_i)$ is proportional to λ_i.

In practice, overdispersion often occurs even when we condition on the response being positive or when we condition on $z_i = 1$ in the latent formulation of the ZIP model. The equality of mean and variance assumed by the ZIP model, conditional on $z_i = 1$, may not be realistic. When we use a ZIP model but there is overdispersion, standard error estimates can be badly biased downward. A zero-inflated negative binomial (ZINB) model is then more appropriate. For it, with probability $1-\phi_i$, $y_i = 0$, and with probability ϕ_i, y_i has a negative binomial distribution with mean λ_i and dispersion parameter γ.

7.4.2 Hurdle Models: Handling Zeroes Separately

An alternative approach to modeling zero-inflation uses a two-part model called a *hurdle model*. One part is a binary model such as a logistic or probit model for whether the response outcome is zero or positive. If the outcome is positive, the "hurdle is crossed." Conditional on a positive outcome, to analyze its level, the second part uses a truncated model that modifies an ordinary distribution by conditioning on a positive outcome. The hurdle model can handle both zero inflation and zero deflation.

Suppose that the first part of the process is governed by probabilities $P(y_i > 0) = \pi_i$ and $P(y_i = 0) = 1 - \pi_i$ and that $\{y_i \mid y_i > 0\}$ follows a truncated-at-zero probability mass function $f(y_i; \mu_i)$, such as a truncated Poisson. The complete distribution is

$$P(y_i = 0) = 1 - \pi_i,$$

$$P(y_i = j) = \pi_i \frac{f(j; \mu_i)}{1 - f(0; \mu_i)}, \quad j = 1, 2, \ldots.$$

With explanatory variables, we could use a logistic regression model for π_i and a loglinear model for the mean μ_i of the untruncated f distribution,

$$\text{logit}(\pi_i) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 \quad \text{and} \quad \log(\mu_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2.$$

The joint likelihood function for the two-part hurdle model is

$$\ell(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{i=1}^{n} (1 - \pi_i)^{I(y_i=0)} \left[\pi_i \frac{f(y_i; \mu_i)}{1 - f\left(0; \mu_i\right)} \right]^{1-I(y_i=0)},$$

where $I(\cdot)$ is the indicator function. If $(1 - \pi_i) > f(0; \mu_i)$ for every i, the model represents zero inflation. The log-likelihood separates into two terms, $L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = L_1(\boldsymbol{\beta}_1) + L_2(\boldsymbol{\beta}_2)$, where

$$L_1(\boldsymbol{\beta}_1) = \sum_{y_i=0} \left[\log\left(1 - \pi_i\right)\right] + \sum_{y_i>0} \log\left(\pi_i\right)$$

$$= \sum_{y_i>0} \boldsymbol{x}_{1i}\boldsymbol{\beta}_1 - \sum_{i=1}^{n} \log(1 + e^{\boldsymbol{x}_{1i}\boldsymbol{\beta}_1})$$

is the log-likelihood function for the binary process and

$$L_2(\boldsymbol{\beta}_2) = \sum_{y_i>0} \left\{\log f\left(y_i; \exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2)\right) - \log\left[1 - f(0; \exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2))\right]\right\}$$

is the log-likelihood function for the truncated model. With a truncated Poisson model for the positive outcome,

$$L_2(\boldsymbol{\beta}_2) = \sum_{y_i>0} \{y_i \boldsymbol{x}_{2i}\boldsymbol{\beta}_2 - e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2} - \log[1 - \exp(-e^{\boldsymbol{x}_{2i}\boldsymbol{\beta}_2})]\} - \sum_{y_i>0} \log(y_i!)$$

is the log-likelihood function for the truncated model. When overdispersion occurs, using a truncated negative binomial for the positive outcome performs better. We obtain ML estimates by separately maximizing L_1 and L_2.

Zero-inflated models are more natural than the hurdle model when the population is naturally regarded as a mixture, with one set of subjects that necessarily has a 0 response. However, the hurdle model is also suitable when, at some settings, the data have fewer zeros than are expected under standard distributional assumptions.

7.4.3 Truncated Discrete Models for Positive Count Data

The part of the hurdle model that applies to the positive counts uses a truncation of a discrete distribution. Such a truncated distribution is of use in its own right in applications in which a count of 0 is not possible. Examples of such response variables are the number of people in a household, the number of occupants of a car, and the number of days a patient admitted to a hospital stays there.

If y_i has a truncated Poisson distribution with parameter λ_i, then

$$E(y_i) = \frac{\lambda_i}{1 - e^{-\lambda_i}}, \quad \text{var}(y_i) = \frac{\lambda_i}{1 - e^{-\lambda_i}} - \frac{\lambda_i^2 e^{-\lambda_i}}{(1 - e^{-\lambda_i})^2}.$$

Conditional on a Poisson variate being positive, the variance is smaller than the mean. When this is substantially violated, more flexibility is provided by the zero-truncated negative binomial distribution. It can be derived as a gamma mixture of zero-truncated Poisson distributions. Software is available for fitting zero-truncated distributions[7].

7.5 EXAMPLE: MODELING COUNT DATA

We illustrate models for discrete data using the horseshoe crab dataset introduced in Section 1.5.1 The response variable for the $n = 173$ mating female crabs is $y =$ number of "satellites"—male crabs that group around the female and may fertilize her eggs. Explanatory variables are the female crab's color, spine condition, weight, and carapace width.

7.5.1 Fits to Marginal Distribution of Satellite Counts

To illustrate the Poisson, negative binomial, ZIP, and ZINB distributions introduced in this chapter, we first investigate the marginal distribution of satellite counts. From Section 1.5.1, the mean of 2.919 and variance of 9.912 suggest overdispersion relative to the Poisson.

```
------------------------------------------------------------------------
> attach(Crabs) # file Crabs.dat at www.stat.ufl.edu/~aa/glm/data
> hist(y, breaks=c(0:16)-0.5) # Histogram display with sufficient bins
------------------------------------------------------------------------
```

[7]Examples are the `pospois` and `posnegbinom` functions in the `VGAM` package of R.

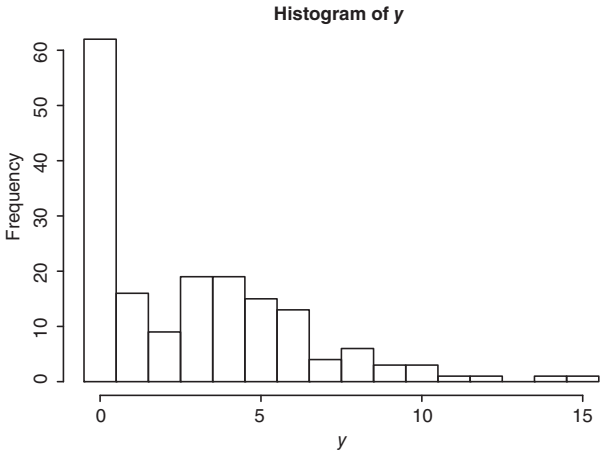

Figure 7.2 Histogram for sample distribution of y = number of horseshoe crab satellites.

The histogram (Figure 7.2) shows a strong mode at 0 but slightly elevated frequencies for satellite counts of 3 through 6 before decreasing substantially. Because the distribution may not be unimodal, the negative binomial may not fit as well as a zero-inflated distribution.

We fit the Poisson distribution and negative binomial distribution with quadratic variance (NB2) by fitting GLMs having only an intercept.

```
-------------------------------------------------------------------------
> summary(glm(y ~ 1, family=poisson, data=Crabs)) # default link is log
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.0713      0.0445    24.07  <2e-16 # exp(1.0713) = 2.919
---
> logLik(glm(y ~ 1, family=poisson, data=Crabs))
'log Lik.' -494.045

> library(MASS)
> summary(glm.nb(y ~ 1, data=Crabs)) # default link is log
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.0713      0.0980    10.93  <2e-16
---
Theta: 0.758, Std. Err.: 0.126
> logLik(glm.nb(y ~ 1, data=Crabs))
'log Lik.' -383.705
-------------------------------------------------------------------------
```

The estimated NB2 dispersion parameter[8] is $\hat{\gamma} = 1/0.758 = 1.32$. This estimate, the much larger SE (0.0980 vs. 0.0445) for the log mean estimate of $\log(2.919) =$

[8]SAS (PROC GENMOD) reports $\hat{\gamma}$ as having $SE = 0.22$.

1.071, and the much larger log-likelihood also suggest that the Poisson distribution is inadequate.

Next, we consider zero-inflated models[9].

```
------------------------------------------------------------------------
> library(pscl) # pscl package can fit zero-inflated distributions
> summary(zeroinfl(y ~ 1)) # uses log link
Count model coefficients (poisson with log link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  1.50385     0.04567    32.93    <2e-16

Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.6139      0.1619   -3.791   0.00015
---
Log-likelihood: -381.615 on 2 Df # 2 is model df, not residual df
------------------------------------------------------------------------
```

The fitted ZIP distribution is a mixture with probability $e^{-0.6139}/[1 + e^{-0.6139}] = 0.351$ for the degenerate distribution at 0 and probability $1 - 0.351 = 0.649$ for a Poisson with mean $e^{1.50385} = 4.499$. The fitted value of $173[0.351 + 0.649e^{-4.499}] = 62.0$ for the 0 count reproduces the observed value of 62. The fitted value for the ordinary Poisson model is only $173e^{-2.919} = 9.3$. The log-likelihood increases substantially when we fit a zero-inflated negative binomial (ZINB) model.

```
------------------------------------------------------------------------
> summary(zeroinfl(y ~ 1, dist="negbin")) # uses log link in pscl lib.
Count model coefficients (negbin with log link):
            Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  1.46527     0.06834   21.440   < 2e-16
Log(theta)   1.49525     0.34916    4.282  1.85e-05

Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.7279      0.1832   -3.973   7.1e-05
---
Theta = 4.4605     Log-likelihood: -369.352 on 3 Df
------------------------------------------------------------------------
```

This distribution is a mixture with probability $e^{-0.7279}/[1 + e^{-0.7279}] = 0.326$ for the degenerate distribution at 0 and probability 0.674 for a negative binomial with mean $e^{1.465} = 4.33$ and dispersion parameter estimate $\hat{\gamma} = 1/4.4605 = 0.22$.

To further investigate lack of fit, we grouped the counts into 10 categories, using a separate category for each count from 0 to 8 and then combining counts of 9 and

[9]Such models can also be fitted with the `vglm` function in the `VGAM` package.

above into a single category. Comparing these with the ZINB fitted distribution of the 173 observations into these 10 categories, we obtained $X^2 = 7.7$ for $df = 10 - 3 = 7$ (since the model has three parameters), an adequate fit. For the other fits, $X^2 = 522.3$ for the Poisson model, 33.6 for the ordinary negative binomial model, and 31.3 for the ZIP model. Here are the fitted counts for the four models:

```
-------------------------------------------------------------------------
count    observed    fit.p  fit.nb  fit.zip  fit.zinb
0           62        9.34   52.27   62.00    62.00
1           16       27.26   31.45    5.62    12.44
2            9       39.79   21.94   12.63    16.73
3           19       38.72   16.01   18.94    17.74
4           19       28.25   11.94   21.31    16.30
5           15       16.50    9.02   19.17    13.58
6           13        8.03    6.87   14.38    10.55
7            4        3.35    5.27    9.24     7.76
8            6        1.22    4.06    5.20     5.48
9 or more   10        0.55   14.16    4.51    10.43
-------------------------------------------------------------------------
```

The ZIP model tends to be not dispersed enough, having fitted value that is too small for the counts of 1 and ≥ 9.

7.5.2 GLMs for Crab Satellite Numbers

We now consider zero-inflated negative binomial models with the explanatory variables from Table 1.3. Weight and carapace width have a correlation of 0.887, and we shall use only weight to avoid issues with collinearity. Darker-colored crabs tend to be older. Most crabs have both spines worn or broken (category 3). When we fit the ZINB main-effects model using weight, color, and spine condition for each component, with color and spine condition as qualitative factors, we find that weight is significant in each component but neither of color or spine condition are. Adding interaction terms does not yield an improved fit. Analyses using color in a quantitative manner with category scores $\{c_i = i\}$ gives relatively strong evidence that darker crabs tend to have more 0 counts. If we use weight w_i in both components of the model but quantitative color only in the zero-component, we obtain:

```
-------------------------------------------------------------------------
> summary(zeroinfl(y ~ weight | weight + color, dist="negbin"))
Count model coefficients (negbin with log link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   0.8961      0.3070    2.919    0.0035
weight        0.2169      0.1125    1.928    0.0538
Log(theta)    1.5802      0.3574    4.422    9.79e-06
```

```
Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   1.8662      1.2415    1.503     0.133
weight       -1.7531      0.4429   -3.958  7.55e-05
color         0.5985      0.2572    2.326     0.020
---
Theta = 4.8558     Log-likelihood: -349.865 on 6 Df
-------------------------------------------------------------------------
```

The fitted distribution is a mixture with probability $\hat{\phi}_i$ of a negative binomial having mean $\hat{\mu}_i$ satisfying

$$\log \hat{\mu}_i = 0.896 + 0.217 w_i$$

with dispersion parameter estimate $\hat{\gamma} = 1/4.8558 = 0.21$, and a probability mass $1 - \hat{\phi}_i$ at 0 satisfying

$$\text{logit}(1 - \hat{\phi}_i) = 1.866 - 1.753 w_i + 0.598 c_i.$$

The overall fitted mean response at a particular weight and color equals

$$\hat{E}(y_i) = \hat{\phi}_i \hat{E}(y_i \mid z_i = 1) = \left(\frac{1}{1 + e^{1.866 - 1.753 w_i + 0.598 c_i}} \right) e^{0.896 + 0.217 w_i}.$$

As weight increases for a particular color, the fitted probability mass at the 0 outcome decreases, and the fitted negative binomial mean increases. Figure 7.3 plots the overall fitted mean as a function of weight for the dark crabs (color 4) and as a function of color at the median weight of 2.35 kg.

If we drop color completely and exclude weight from the NB2 component of the model, the log-likelihood decreases to -354.7 but we obtain the simple expression for the overall fitted mean of $\exp(1.47094)/[1 + \exp(3.927 - 1.985 w_i)]$. This has a logistic shape for the increase in the fitted mean as a function of weight.

If we ignore the zero inflation and fit an ordinary NB2 model with weight and quantitative color predictors, we obtain:

```
-------------------------------------------------------------------------
> summary(glm.nb(y ~ weight + color))
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -0.3220      0.5540   -0.581     0.561
weight        0.7072      0.1612    4.387  1.15e-05
color        -0.1734      0.1199   -1.445     0.148
---
Theta: 0.956   2 x log-likelihood: -746.452 # L = -373.226
-------------------------------------------------------------------------
```

This describes the tendency of the overall mean response to increase with weight and decrease with color (but not significantly). In not having a separate component

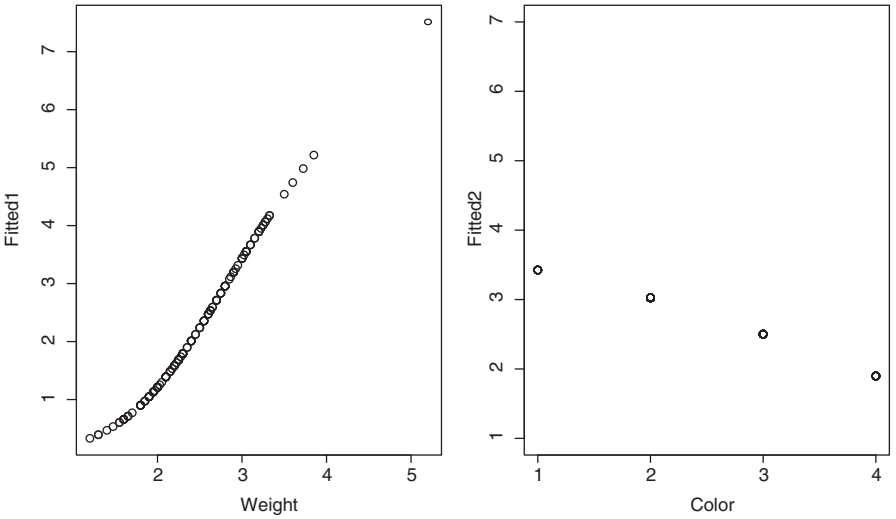

Figure 7.3 Fitted mean number of horseshoe crab satellites for zero-inflated negative binomial model, plotted as a function of weight for dark crabs and as a function of color for median-weight crabs.

to handle the zero count, the NB2 model has dispersion parameter estimate $\hat{\gamma} = 1/0.956 = 1.05$ that is much greater than $\hat{\gamma}$ for the NB2 component of ZINB models. The fit is similar to that of the geometric distribution, which is NB2 with $\gamma = 1$. But its log-likelihood of -373.2 is considerably worse than values obtained for ZINB models.

Unless previous research or theory suggests more-complex models, it seems adequate to use a zero-inflated NB2 model with weight as the primary predictor, adding color as a predictor of the mass at 0. In these analyses, however, we have ignored that the dataset contains an outlier—an exceptionally heavy crab weighing 5.2 kg of medium color that had 7 satellites. As exercise, you can fit models without that observation to investigate how the results change.

CHAPTER NOTES

Section 7.1: Poisson GLMs for Counts and Rates

7.1 **Poisson GLMs**: See Cameron and Trivedi (2013) for details about Poisson and other models for count data and an extensive bibliography.

Section 7.2: Poisson/Multinomial Models for Contingency Tables

7.2 **Loglinear models**: For more details about loglinear models for contingency tables, see Agresti (2013, Chapters 9, 10) and Bishop et al. (1975).

7.3 **Graphical models**: For more on conditional independence graphs, see Darroch et al. (1980), Lauritzen (1996), and Madigan and York (1995). More general probabilistic contexts include directed graphs, which are natural for hierarchical Bayesian models, and explanatory variables (e.g., Jordan 2004).

Section 7.3: Negative Binomial GLMs

7.4 **NB modeling**: Greenwood and Yule (1920) derived the negative binomial as a gamma mixture of Poisson distributions. Johnson et al. (2005, Chapter 5) summarized properties of the distribution. Cameron and Trivedi (2013, Section 3.3) discussed NB modeling and presented an asymptotic variance expression for $\hat{\gamma}$. They also presented moment estimators for γ and studied robustness properties (Section 3.3) and discussed analogs of R-squared for count data models (Section 5.3.3). See also Anscombe (1950), Hilbe (2011), Hinde and Demétrio (1998), and Lawless (1987). Alternatives to the gamma for mixing Poisson distributions include the log-normal and inverse-Gaussian distributions. See Cameron and Trivedi (2013, Section 4.2).

Section 7.4: Models for Zero-Inflated Data

7.5 **Hurdle model, and ZIP versus ZINB**: Mullahy (1986) proposed the hurdle model using the truncated Poisson or geometric distribution. Ridout et al. (2001) provided a score test of the ZIP model against the ZINB alternative. Estimators for the ZIP model can be unstable compared to the hurdle model (e.g., for estimating a predictor effect in the logit component of the model) when zero deflation occurs at some predictor settings. See Min and Agresti (2005) for discussion, more references, and extensions to handling repeated measurements with zero-inflated data. See also Cameron and Trivedi (2013, Chapter 4) and Hilbe (2011, Chapter 11) for zero-inflated models, hurdle models, truncated models, and other generalized count regression models.

7.6 **Zero-truncated models**: Models for zero-truncated data have a long history. See Amemiya (1984), Cameron and Trivedi (2013, Section 4.3), Johnson et al. (2005, Section 4.10, 5.11), and Meng (1997).

EXERCISES

7.1 Suppose $\{y_i\}$ are independent Poisson observations from a single group. Find the likelihood equation for estimating $\mu = E(y_i)$. Show that $\hat{\mu} = \bar{y}$ regardless of the link function.

7.2 Suppose $\{y_i\}$ are independent Poisson variates, with $\mu = E(y_i)$, $i = 1, \ldots, n$. For testing H_0: $\mu = \mu_0$, show that the likelihood-ratio statistic simplifies to

$$-2(L_0 - L_1) = 2[n(\mu_0 - \bar{y}) + n\bar{y}\log(\bar{y}/\mu_0)].$$

Explain how to use this to obtain a large-sample confidence interval for μ.

7.3 Refer to the previous exercise. Explain why, alternatively, for large samples you can test H_0 using the standard normal test statistic $z = \sqrt{n}(\bar{y} - \mu_0)/\sqrt{\mu_0}$. Explain how to invert this test to obtain a confidence interval. (These are the score test and score-test based confidence interval.)

7.4 When y_1 and y_2 are independent Poisson with means μ_1 and μ_2, find the likelihood-ratio statistic for testing H_0: $\mu_1 = \mu_2$. Specify its asymptotic null distribution, and describe the condition under which the asymptotics apply.

7.5 For the one-way layout for Poisson counts (Section 7.1.5), using the identity link function, show how to obtain a large-samples confidence interval for $\mu_h - \mu_i$. If there is overdispersion, explain why it is better to use a formula $(\bar{y}_h - \bar{y}_i) \pm z_{\alpha/2}\sqrt{(s_h^2/n_h) + (s_i^2/n_i)}$ based only on the central limit theorem.

7.6 For the one-way layout for Poisson counts, derive the likelihood-ratio statistic for testing H_0: $\mu_1 = \cdots = \mu_c$.

7.7 For the one-way layout for Poisson counts, derive a test of H_0: $\mu_1 = \cdots = \mu_c$ by applying a Pearson chi-squared goodness-of-fit test (with $df = c - 1$) for a multinomial distribution that compares sample proportions in c categories against H_0 values of multinomial probabilities, (**a**) when $n_1 = \cdots = n_c$, (**b**) for arbitrary $\{n_i\}$, with $n = \sum_i n_i$.

7.8 In a balanced two-way layout for a count response, let y_{ijk} be observation k at level i of factor A and level j of factor B, $k = 1, \ldots, n$. Formulate a Poisson loglinear main-effects model for $\{\mu_{ijk} = E(y_{ijk})\}$. Find the likelihood equations, and show that $\{\mu_{ij+} = \sum_k E(y_{ijk})\}$ have fitted values $\{\hat{\mu}_{ij+} = (y_{i++}y_{+j+})/y_{+++}\}$.

7.9 Refer to Note 1.5. For a Poisson loglinear model containing an intercept, show that the average estimated rate of change in the mean as a function of explanatory variable j satisfies $\frac{1}{n}\sum_i(\partial\hat{\mu}_i/\partial x_{ij}) = \hat{\beta}_j\bar{y}$.

7.10 A method for negative exponential modeling of survival times relates to the Poisson loglinear model for rates (Aitkin and Clayton 1980). Let T denote the time to some event, with pdf f and cdf F. For subject i, let $w_i = 1$ for death and 0 for censoring, and let $t = \sum_i t_i$ and $w = \sum_i w_i$.

a. Explain why the survival-time log-likelihood for n independent observations is[10]

$$L(\lambda) = \sum_i w_i \log[f(t_i)] + \sum_i (1 - w_i)\log[1 - F(t_i)].$$

[10]This actually applies only for *noninformative* censoring mechanisms.

Assuming $f(t) = \lambda \exp(-\lambda t)$, show that $\hat{\lambda} = w/t$. Conditional on t, explain why w has a Poisson distribution with mean $t\lambda$. Using the Poisson likelihood, show that $\hat{\lambda} = w/t$.

b. With λ replaced by $\lambda \exp(\boldsymbol{x}\boldsymbol{\beta})$ and with $\mu_i = t_i \lambda \exp(\boldsymbol{x}_i\boldsymbol{\beta})$, show that L simplifies to

$$L(\lambda, \boldsymbol{\beta}) = \sum_i w_i \log\, \mu_i - \sum_i \mu_i - \sum_i w_i \log\, t_i.$$

Explain why maximizing $L(\lambda, \boldsymbol{\beta})$ is equivalent to maximizing the likelihood for the Poisson loglinear model

$$\log \mu_i - \log\, t_i = \log \lambda + \boldsymbol{x}_i\boldsymbol{\beta}$$

with offset $\log(t_i)$, using "observations" $\{w_i\}$.

c. When we sum terms in L for subjects having a common value of $\boldsymbol{x}$, explain why the observed data are the numbers of deaths ($\sum_i w_i$) at each setting of $\boldsymbol{x}$, and the offset is $\log(\sum_i t_i)$ at each setting.

7.11 Consider the loglinear model of conditional independence between A and B, given C, in a $r \times c \times \ell$ contingency table. Derive the likelihood equations, and interpret. Give the solution of fitted values that satisfies the model and the equations. (From Birch (1963), it follows that only one solution exists, namely the ML fit.) Explain the connection with the fitted values for the independence model for a two-way table. Find the residual *df* for testing fit.

7.12 Two balanced coins are flipped, independently. Let $A =$ whether the first flip resulted in a head (yes, no), $B =$ whether the second flip resulted in a head, and $C =$ whether both flips had the same result. Using this example, show that marginal independence for each pair of three variables does not imply that the variables are mutually independent.

7.13 For three categorical variables *A, B*, and *C*:

a. When C is jointly independent of A and B, show that A and C are conditionally independent, given B.

b. Prove that mutual independence of *A, B*, and *C* implies that A and B are (i) marginally independent and (ii) conditionally independent, given C.

c. Suppose that A is independent of B and that B is independent of C. Does this imply that A is independent of C? Explain.

7.14 Express the loglinear model of mutual independence for a $2 \times 2 \times 2$ table in the form $\log \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. Show that the likelihood equations equate $\{y_{ijk}\}$ and $\{\hat{\mu}_{ijk}\}$ in the one-dimensional margins, and their solution is $\{\hat{\mu}_{ijk} = y_{i++}y_{+j+}y_{++k}/n^2\}$.

7.15 For a $2 \times c \times \ell$ table, consider the loglinear model by which A is jointly independent of B and C. Treat A as a response variable and B and C as explanatory, conditioning on $\{n_{+jk}\}$. Construct the logit for the conditional distribution of A, and identify the corresponding logistic model.

7.16 For the homogeneous association loglinear model (7.7) for a $r \times c \times \ell$ contingency table, treating A as a response variable, find the equivalent baseline-category logit model.

7.17 For a four-way contingency table, consider the loglinear model having AB, BC, and CD two-factor terms and no three-factor interaction terms. Explain why A and D are independent given B alone or given C alone or given both B and C. When are A and C conditionally independent?

7.18 Suppose the loglinear model (7.7) of homogeneous association holds for a three-way contingency table. Find $\log \mu_{ij+}$ and explain why marginal associations need not equal conditional associations for this model.

7.19 Consider the loglinear model for a four-way table having AB, AC, and AD two-factor terms and no three-factor interaction term. What is the impact of collapsing over B on the other associations? Contrast that with what the collapsibility condition in Section 7.2.7 suggests, treating group $S_3 = \{B\}$, (**i**) if $S_1 = \{C\}$ and $S_2 = \{A, C\}$, (**ii**) if $S_1 = \{C, D\}$ and $S_2 = \{A\}$. This shows that different groupings for that condition can give different information.

7.20 A county's highway department keeps records of the number of automobile accidents reported each working day on a superhighway that runs through the county. Describe factors that are likely to cause the distribution of this count over time to show overdispersion relative to the Poisson distribution.

7.21 Show that a gamma mixture of Poisson distributions yields the negative binomial distribution.

7.22 Given u, y is Poisson with $E(y \mid u) = u\mu$, where u is a positive random variable with $E(u) = 1$ and $\text{var}(u) = \tau$. Show that $E(y) = \mu$ and $\text{var}(y) = \mu + \tau\mu^2$. Explain how you can formulate the negative binomial distribution and a negative binomial GLM using this construction.

7.23 For discrete distributions, Jørgensen (1987) showed that it is natural to define the exponential dispersion family as

$$f(y_i; \theta_i, \phi) = \exp[y_i\theta_i - b(\theta_i)/a(\phi) + c(y_i, \phi)].$$

a. For fixed k, show that the negative binomial distribution (7.8) has this form with $\theta_i = \log[\mu_i/(\mu_i + k)]$, $b(\theta_i) = -\log(1 - e^{\theta_i})$, and $a(\phi) = 1/k$.

b. For this version, show that $x_i = y_i a(\phi)$ has the usual exponential dispersion family form (4.1).

7.24 For a sequence of independent Bernoulli trials, let $y =$ the number of successes before the kth failure. Show that y has the negative binomial distribution,

$$f(y; \pi, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \pi^y (1-\pi)^k, \quad y = 0, 1, 2, \ldots .$$

(The *geometric distribution* is the special case $k = 1$.) Relate π to the parameters μ and k in the parameterization (7.8).

7.25 With independent negative binomial observations from a single group, find the likelihood equation and show that $\hat{\mu} = \bar{y}$. (ML estimation for γ requires iterative methods, as R. A. Fisher showed in an appendix to Bliss (1953). See also Anscombe (1950).) How does this generalize to the one-way layout?

7.26 For the ZIP null model (i.e., without explanatory variables), show from the likelihood equations that the ML-fitted 0 count equals the observed 0 count.

7.27 The text website contains an expanded version (file `Drugs3.dat`) of the student substance use data of Table 7.3 that also has each subject's $G =$ gender (1 = female, 2 = male) and $R =$ race (1 = white, 2 = other). It is sensible to treat G and R as explanatory variables. Explain why any loglinear model for the data should include the GR two-factor term. Use a model-building process to select a model for these data. Interpret the estimated conditional associations.

7.28 Other than a formal goodness-of-fit test, one analysis that provides a sense of whether a particular GLM is plausible is the following: Suppose the ML fitted equation were the true equation. At the observed $\boldsymbol{x}$ values for the n observations, randomly generate n variates with distributions specified by the fitted GLM. Construct scatterplots. Do they look like the scatterplots that were actually observed? Do this for a Poisson loglinear model for the horseshoe crab data, with $y =$ number of satellites and $x =$ width. Does the variability about the fit resemble that in the actual data, including a similar number of 0's and large values? Repeat this a few times to get a better sense of how the scatterplot observed differs from what you would observe if the Poisson GLM truly held.

7.29 Another model (Dobbie and Welsh 2001) for zero-inflated count data uses the *Neyman type A distribution*, which is a compound Poisson–Poisson mixture. For observation i, let z_i denote a Poisson variate with expected value λ_i. Conditional on z_i, let w_{ij} $(j = 1, \ldots, z_i)$ denote independent Poisson(ϕ_i)

observations. The model expresses y_i using the decomposition $y_i = \sum_{j=0}^{z_i} w_{ij}$, $i = 1, 2, \ldots, n$. Find $E(y_i)$. Relating λ_i and ϕ_i to explanatory variables through $\log(\lambda_i) = \boldsymbol{x}_{1i}\boldsymbol{\beta}_1$ and $\log(\phi_i) = \boldsymbol{x}_{2i}\boldsymbol{\beta}_2$, show the model for $E(y_i)$ and interpret its parameters.

7.30 A headline in *The Gainesville Sun* (February 17, 2014) proclaimed a worrisome spike in shark attacks in the previous 2 years. The reported total number of shark attacks in Florida per year from 2001 to 2013 were 33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23. Are these counts consistent with a null Poisson model or a null negative binomial model? Test the Poisson model against the negative binomial alternative. Analyze the evidence of a positive linear trend over time.

7.31 Table 7.5, also available at `www.stat.ufl.edu/~aa/glm/data`, summarizes responses of 1308 subjects to the question: within the past 12 months, how many people have you known personally that were victims of homicide? The table shows responses by race, for those who identified their race as white or as black.

a. Let y_i denote the response for subject i and let $x_i = 1$ for blacks and $x_i = 0$ for whites. Fit the Poisson GLM $\log \mu_i = \beta_0 + \beta x_i$ and interpret $\hat{\beta}$.

b. Describe factors of heterogeneity such that a Poisson GLM may be inadequate. Fit the corresponding negative binomial GLM, and estimate how the variance depends on the mean. What evidence does this model fit provide that the Poisson GLM had overdispersion? (Table 7.5 also shows the fits for these two models.)

c. Show that the Wald 95% confidence interval for the ratio of means for blacks and whites is $(4.2, 7.5)$ for the Poisson GLM but $(3.5, 9.0)$ for the negative binomial GLM. Which do you think is more reliable? Why?

Table 7.5 Number of Victims of Murder Known in the Past Year, by Race, with Fit of Poisson and Negative Binomial Models

	Data		Poisson GLM		Negative Binomial GLM	
Response	Black	White	Black	White	Black	White
0	119	1070	94.3	1047.7	122.8	1064.9
1	16	60	49.2	96.7	17.9	67.5
2	12	14	12.9	4.5	7.8	12.7
3	7	4	2.2	0.1	4.1	2.9
4	3	0	0.3	0.0	2.4	0.7
5	2	0	0.0	0.0	1.4	0.2
6	0	1	0.0	0.0	0.9	0.1

Source: 1990 General Social Survey, file `Homicide.dat` at `www.stat.ufl.edu/~aa/glm/data`.

7.32 For the horseshoe crab data, the negative binomial modeling shown in the R output first treats color as nominal-scale and then in a quantitative manner, with the category numbers as scores. Interpret the result of the likelihood-ratio test comparing the two models. For the simpler model, interpret the color effect and interpret results of the likelihood-ratio test of the null hypothesis of no color effect.

```
-------------------------------------------------------------------
> fit.nb.color <- glm.nb(y ~ factor(color)) # Using Crabs.dat file
> summary(fit.nb.color)
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)      1.4069      0.3526    3.990  6.61e-05
factor(color)2  -0.2146      0.3750   -0.572     0.567
factor(color)3  -0.6061      0.4036   -1.502     0.133
factor(color)4  -0.6913      0.4508   -1.533     0.125
---
> fit.nb.color2 <- glm.nb(y ~ color) # using color scores (1,2,3,4)
> summary(fit.nb.color2)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   1.7045      0.3095    5.507  3.66e-08
color        -0.2689      0.1225   -2.194    0.0282
---
> anova(fit.nb.color2, fit.nb.color)
Likelihood ratio tests of Negative Binomial Models
Response: y
Model   theta  Res.df  2 x log-lik.   Test  df  LR stat. Pr(Chi)
1      0.7986     171    -762.6794
2      0.8019     169    -762.2960  1 vs. 2   2   0.3834   0.8256
---
> 1 - pchisq(767.409-762.679, df=172-171) # LR test vs. null model
[1] 0.0296
-------------------------------------------------------------------
```

7.33 For the horseshoe crab data, the following output shows a zero-inflated negative binomial model using quantitative color for the zero component. Interpret results, and compare with the NB2 model fitted in the previous exercise with quantitative color. Can you conduct a likelihood-ratio test comparing them? Why or why not?

```
-------------------------------------------------------------------
> summary(zeroinfl(y ~ 1 | color, dist = "negbin")) # Using Crabs.dat
Count model coefficients (negbin with log link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   1.4632      0.0689   21.231   < 2e-16
Log(theta)    1.4800      0.3511    4.215   2.5e-05

Zero-inflation model coefficients (binomial with logit link):
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -2.7520      0.6658   -4.133  3.58e-05
color         0.8023      0.2389    3.358  0.000785
---
Theta = 4.3928    Log-likelihood:-362.997 on 4 Df
-------------------------------------------------------------------
```

7.34 Refer to Section 7.5.2. Redo the zero-inflated NB2 model building, deleting the outlier crab weighing 5.2 kg. Compare results against analyses that used this observation and summarize conclusions.

7.35 A question in a GSS asked subjects how many times they had sexual intercourse in the preceding month. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4. The mode for each gender was 0. Specify a GLM that would be inappropriate for these data, explaining why. Specify a model that may be appropriate.

7.36 Table 7.6 is based on a study involving British doctors.

Table 7.6 Data for Exercise 7.36 on Coronary Death Rates

	Person-Years		Coronary Deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35–44	18,793	52,407	2	32
45–54	10,673	43,248	12	104
55–64	5710	28,612	28	206
65–74	2585	12,663	28	186
75–84	1462	5317	31	102

Source: Doll R. and A. Bradford Hill. 1966. *Natl. Cancer Inst. Monogr.* **19**: 205–268.

a. Fit a main-effects model for the log rates using age and smoking as factors. In discussing lack of fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over age, and evaluate how the sample ratio depends on age.

b. Explain why it is sensible to add a quantitative interaction of age and smoking. For this model, show that the log ratio of coronary death rates changes linearly with age. Assign scores to age, fit the model, and interpret.