18

MODELLING COUNT AND RATE DATA

OBJECTIVES

After reading this chapter, you should be able to:

1. Understand the relationship between counts of disease events and incidence rates.
2. Fit, evaluate and interpret Poisson regression models.
3. Be able to determine when a negative binomial model is likely to be more appropriate than a Poisson model and to quantify and statistically assess overdispersion.
4. Fit, evaluate and interpret negative binomial regression models.
5. Decide when zero-adjusted models (hurdle, zero-inflated, zero-truncated) might be more appropriate than a Poisson or negative binomial model.
6. Fit zero-adjusted models and interpret the results.

18.1 INTRODUCTION

In previous chapters, we have looked at methods of analysing data measured on a continuous scale (Chapter 14) and 2 types of discrete data: binary/binomial data (Chapter 16) and multinomial data (Chapter 17). Here, we are introduced to the situation in which the outcome we are measuring represents a count of the number of times an event occurs in an individual or group of individuals.

a. It might be a simple count of events, such as the number of breedings required for a dairy cow to conceive. Poisson regression was used to evaluate the effect of peripartum treatment with an anthelmintic on the number of services per conception in dairy cows (Sanchez *et al*, 2002).
b. It might be a count of cases of disease over a period of time with the amount of animal-time at risk having to be taken into consideration (*eg* total number of cases of clinical mastitis in a dairy herd over a year with the number of cow-months contributed by lactating cows as the amount of animal-time). Hence, this is a measure of the incidence rate (I) of disease. The examples used in this chapter will focus on this kind of data: the incidence rate of clinical mastitis in a hypothetical trial of 'pre-dipping' in a dairy herd and the incidence rate of new *Mycobacterium bovis* infections in cattle and cervid herds after the introduction of the agent to the herd.
c. It might be a count of cases of disease with the size of the population at risk being taken into consideration (*eg* cases of lymphosarcoma in slaughtered cattle seen at various abattoirs with the number of cattle slaughtered as the population at risk). This is an estimate of the (lifetime) incidence risk of lymphosarcoma in cattle.
d. It might be a count of an outcome that is measured over a geographical area. For example, Poisson models are also used to investigate factors related to the number of events per unit area. Hammond *et al* (2001) investigated whether land use was predictive of the number of badgers in 500 m^2 grids in an area in Ireland. The study area was overlaid by a 500 m^2 grid and the number of badgers caught in each grid was recorded. Land use within each cell of the grid was described by a set of categorical variables. The mean number of badgers per grid was 0.6 and the variance was 1.5. A major finding was that as the area of high-quality pasture within a grid increased, the number of badgers also increased.

18.1.1 Approaches to analysis

We might want to evaluate the effect of 'pre-dipping' (disinfection of teat ends prior to milking) on the incidence of clinical mastitis. We will assume that a controlled trial can be carried out in 3 large dairy herds and cows will be individually assigned to be pre-dipped or not. The outcome of interest might be the total number of cases of clinical mastitis occurring in each cow over a full lactation. Other factors that will have to be taken into consideration are the age of the cow (it is generally accepted that the risk of clinical mastitis increases with age) and which herd the cow is in (because incidence rates of mastitis vary among herds). While random assignment of cows to treatment groups should balance the age and herd factors across the study groups, you might still want to consider them in the analysis. Given the clinical trial design, we can assume that the population is closed, but the time at risk will vary among cows. **Note** In this example, we are interested in the total number of cases of mastitis. If we were interested in first cases only, we could create a binary variable for each cow and fit a logistic model.

There are a number of ways to approach the analysis of the data generated by this study.

a. The incidence rate of clinical mastitis could be computed for each study group and the difference between the 2 groups tested using an unconditional *Z*-test that was discussed in Chapter 6. This approach does not allow for the control of other potential confounding variables (*ie* age of cow and herd), so it would rely totally on random assignment to control for these effects.
b. Alternatively, you could determine the incidence rate (*I*) of clinical mastitis within each cow and use that value as the dependent variable in a linear regression with pre-dipping as the primary exposure (predictor) of interest and age and herd as extraneous variables. However, most cows would have an *I* of zero, so it is very unlikely that the error terms would have anything close to a normal distribution. Consequently, one of the fundamental assumptions of linear regression would be violated. It is also possible that some combination of predictor variables could be found that predicted a negative *I* for the cow. This approach looks worse than the first.
c. The preferred approach is to use Poisson or negative binomial regression to model the incidence of new cases while adjusting for the amount of time each cow was at risk.

A number of texts dealing specifically with the analysis of count data are available and include Cameron & Trivedi (1998); Hilbe (2007); Long (1997); Long & Freese (2006).

18.2 THE POISSON DISTRIBUTION

The Poisson distribution is often used to model counts of 'rare' events:

$$p(Y=y)=\frac{\mu^{y}e^{-\mu}}{y!} \qquad \textit{Eq 18.1}$$

where y is the observed count of events and μ is the mean number of events. An interesting characteristic of the Poisson distribution is that the mean and the variance are equal (*ie* μ).

The Poisson distribution can be thought of in 2 ways.

a. If the times between events (*eg* cases of mastitis) are independent and follow an exponential distribution with a mean value of t, then the number of cases of mastitis (Y) in a defined time period (T) will follow a Poisson distribution with $\mu=T/t$. For example, if the mean time between cases of mastitis is 150 days, then the expected number of cases in a 300-day lactation will be 300/150=2 cases. The time between events is sometimes referred to as the 'waiting time'. Using this formulation of the Poisson distribution, there is a natural connection between the analysis of counts of events (Poisson regression) and the analysis of time to event occurrence (survival analysis—Chapter 19).
b. The Poisson distribution approximates the binomial distribution if the population (n) is large, consists of independent units and the binomial proportion (p) is small (*ie* events are 'rare'). In this case, $\mu=np$. For example, if the probability of the occurrence of mastitis on any given day is 1/150=0.0067, then the expected number of cases in a 300-day lactation will be 300*0.0067=2.

If the outcome follows a Poisson distribution and the mean is known, you can calculate the probability of a specific number of events occurring. For example, if the average number of milk fever cases in a dairy herd is 5 per year, the probability of getting 10 cases in a year is:

$$p(Y=10)=\frac{5^{10}e^{-5}}{10!}=0.018$$

This indicates that there is approximately a 2% chance of having exactly 10 cases in a year (provided the mean for the population is not changing over time).

Poisson distributions with means of 0.5, 1.0, 2.0 and 5.0 are shown in Fig. 18.1. As this figure indicates, as the mean increases, the Poisson distribution approaches a normal distribution.

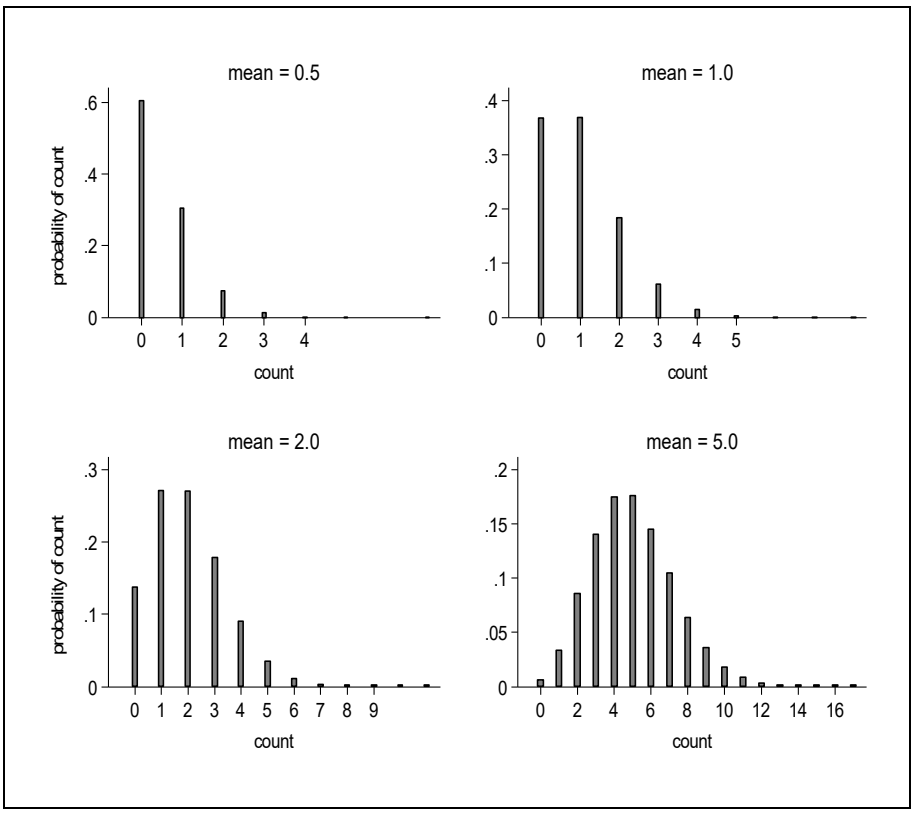

Fig. 18.1 Poisson distributions

18.3 POISSON REGRESSION MODEL

The usual form of the Poisson regression model is:

$$E(Y)=\mu=n\lambda \qquad \textit{Eq 18.2}$$

where E(Y) = the expected number of cases of disease
n = exposure (*eg* animal-time units at risk)
λ = represents a function which defines the disease incidence rate.

The **exposure** (n) adjusts for different amounts of time at risk (or, alternatively, different sizes of populations at risk) for the various study subjects (animals or groups of animals). (**Note** Throughout this text, the letter n is most commonly used to denote the number of animals in a

population. Here we are also using it to denote the amount of animal-time at risk.) If n is equal for all subjects, it can be omitted but you must remember that predicted counts will refer to the expected number of cases in n animal-time units at risk. For example, in the badger study referred to, each count related to the same 500 m^2 grid size so no offset or exposure was required. However, the predicted counts were counts per 500 m^2.

One of the ways that λ can be related to the predictor(s) is:

$$\lambda = e^{\beta_0 + \beta_1 X} \quad \text{or} \quad \ln(\lambda) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.3}$$

Consequently, the Poisson regression model is:

$$E(Y) = n e^{\beta_0 + \beta_1 X} \quad \text{or} \quad \ln E(Y) = \ln(n) + \beta_0 + \beta_1 X$$

$$\text{or} \quad \ln E(I) = \ln\left(\frac{E(Y)}{n}\right) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.4}$$

where lnE(I) is the log of the expected value of the incidence rate (I) of disease which is being modelled as a linear combination of predictors. **Note** This example assumes that there is a single predictor variable (X), but the model can be extended to include multiple predictors.

The exposure (n) may be recorded and used on the original scale (*ie* the amount of animal-time at risk). Alternatively, it may be converted to a log scale and used as such (referred to as an **offset**). In statistical terminology, an offset term in a model equation is a term whose regression coefficient is restricted to be 1 (*ie* absent).

As with logistic regression, Poisson regression models are fit using an iterative maximum likelihood estimation procedure. The statistical significance of the contribution of individual predictors (or groups of predictors) to the model can be tested using either Wald tests or likelihood ratio tests. An example of a Poisson regression analysis, based on tuberculosis data from cattle and cervids is presented in Example 18.1.

18.4 INTERPRETATION OF COEFFICIENTS

The coefficients from a Poisson regression model represent the amount the log of I ($\ln I$) is expected to change with a unit change in the predictor. Assuming that there are 2 exposure groups (X=0 and X=1), then the incidence rate ratio (IR) associated with belonging to group X=1 (relative to group X=0) is:

$$IR = \frac{\lambda_1}{\lambda_0} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1} \qquad \textit{Eq 18.5}$$

so the coefficients from a Poisson regression can easily be converted into IR estimates. In general, the IR represents the proportional increase in I for a unit change in the predictor. For example, if the IR for lactation number in a study of clinical mastitis cases was 1.5, that would suggest that the incidence rate of clinical mastitis went up by 50% for each additional lactation that a cow had (*ie* that it was 1.5 times as high as the rate in the previous lactation). **Note** In general e^{β_1} represents the ratio between mean counts in 2 groups. However, because Poisson regression is most commonly used for incidence rate data in epidemiologic studies, this specific use will be emphasised throughout this chapter.

Example 18.1 Poisson regression model
data = tb_real

The incidence rates of new tuberculosis (TB) infections in cattle, cervid (deer and elk) and bison herds in 9 TB outbreaks in Canada between 1985 and 1994 were estimated (Munroe *et al*, 2000; Munroe *et al*, 1999). These incidence rates were modelled as a function of several characteristics of the animals in the herd (type, sex and age). The key predictors were:

type: 1=dairy, 2=beef, 3=cervid, 4=other
sex: 0=female, 1=male
age: 0=0-12 mo, 1=12-24 mo, 2=24+ mo

A more complete description of the dataset is in Chapter 31.

A Poisson regression model with the 3 predictor variables and the time at risk included as an exposure variable produced the following results.

Number of obs = 134
Log likelihood = -238.7

					95% CI for IR	
Variable	**Coef**	**SE**	**P**	**IR**	**Lower**	**Upper**
type=beef	0.442	0.236	0.061	1.56	0.98	2.47
type=cervid	1.066	0.233	<0.001	2.90	1.84	4.59
type=other	0.438	0.615	0.476	1.55	0.46	5.17
sex=male	-0.362	0.195	0.064	0.70	0.47	1.02
age=12-24 mo	2.673	0.722	<0.001	14.49	3.52	59.62
age=24+ mo	2.601	0.714	<0.001	13.48	3.33	54.59
constant	-11.690	0.740	<0.001	NA	NA	NA

Herd type was a significant predictor (P<0.001) with incidence rates in beef and cervid herds higher than in dairy herds although the coefficient for beef herds was only borderline significant. Males appeared to have a lower incidence rate (again borderline significance) and animals over 11 months of age definitely had higher incidence rates.

The deviance and Pearson goodness-of-fit test statistics were:

	df	**X^2**	**P**	**dispersion**
Deviance	127	348.4	<0.001	2.74
Pearson	127	1105.7	<0.001	8.71

These suggest that there are serious problems with the model (*ie* strong evidence of lack of fit).

The effect of a predictor on the absolute number of cases of disease (or other outcome event) depends on the values for other predictors in the model. For example, the *IR* for type=cervid in Example 18.1 was $e^{1.066}$=2.9 (compared with dairy herds). The predicted *I* for young (0-12 mo), females in a dairy herd was $e^{-11.690}$=0.84 cases/100,000 animal-days at risk. For cervids, the predicted rate would be 0.84*2.9=2.43 cases/100,000 animal-days, or an extra 1.59 cases per 100,000 animal-days. In animals aged 12-24 mos, the predicted rate for females in dairy herds was $e^{-9.017}$=12.1 cases/100,000 animal-days. For cervids of this age, the rate would be 12.1*2.9=35.3 cases/100,000 animal-days, or an extra 23.2 cases per 100,000 animal-days.

18.4.1 Poisson regression and risk ratios

Logistic regression (Chapter 16) is the most widely used multivariable model for binary (0/1) data and it produces estimates of effect expressed as odds ratios (*OR*s). Risk ratios (*RR*s) are more easily understood and may be preferred to *OR*s in some situations. One multivariable approach to obtaining *RR* is to fit a generalised linear model with a binomial distribution and a log link (see Section 16.11). However, it has been reported that these models may fail to converge (Barros & Hirakata, 2003; Zou, 2004). An alternative is to use Poisson regression (with no exposure or offset specified) even though the data are binary, to directly estimate *RR*s (Barros & Hirakata, 2003; McNutt *et al*, 2003). This approach produces estimates with very little bias but a conservative CI (*ie* the CI is too wide). It has been shown that using robust SEs (see Section 14.9.5) reduces the estimated SEs of the coefficients and results in a CI of the correct width (Greenland, 2004; Zou, 2004).

18.5 Evaluating Poisson regression models

18.5.1 Residuals

Raw residuals can be computed for each observation as the observed number of cases (obs) minus the expected number of cases (exp) predicted from the model. Residuals are computed on the basis of one per observation.

Pearson residuals can be computed as:

$$res = \frac{\text{obs - exp}}{\sqrt{\text{var}}} \qquad \text{which for the } i^{th} \text{ observations is} \qquad res_i = \frac{y_i - \mu_i}{\sqrt{\mu_i}} \qquad \textit{Eq 18.6}$$

where 'var' is the estimated variance of the observations. For a Poisson model, the estimated variance is equal to the expected number of cases (μ).

Deviance residuals are based on the overall fit of the model (formula not shown). The sum of the squared deviance residuals gives the deviance for the model which is defined as minus twice the difference between the log likelihood of the model and the maximum log likelihood achievable. Both Pearson and deviance residuals may be standardised to give them unit variance.

Anscombe residuals are similar to standardised deviance residuals (Hilbe, 2007) but may be better at identifying outliers and heterogeneity in the data. It is recommended that both standardised deviance and Anscombe residuals be plotted against predicted values when evaluating a Poisson model (Hilbe, 2007).

All of the above residuals may be further standardised, a process which makes the variance of the residuals more constant.

18.5.2 Assessing overall fit

As with logistic regression, χ^2 goodness-of-fit tests can be computed as the sum of the squared deviance or Pearson residuals. The resulting test statistic has an approximate χ^2 distribution if there are multiple observations within each covariate pattern defined by the predictors in the

model (Cameron & Trivedi, 1998). However, the values of the 2 test statistics could be quite different and, if either is indicative of a lack of fit, the model should be investigated thoroughly. As with all overall goodness-of-fit statistics, a significant result (indicating lack of fit) provides no information about what the cause of the lack of fit is. However, with Poisson models, a common cause is overdispersion (*ie* the variance of the counts is much larger than the mean). The Pearson and deviance goodness-of-fit test results for the TB data are presented in Example 18.1 (both are highly significant in indicating lack of fit).

The predictive ability of the model can be evaluated by comparing the distributions of observed and predicted counts. Fig. 18.2 shows the distributions of the observed and predicted counts from the model presented in Example 18.1. Their apparent similarity does not reflect the serious problems with the lack of fit for the model.

18.5.3 Overdispersion

The assumption behind a Poisson model is that the mean and the variance are equal (conditional upon the predictors in the model); that is, the mean and the variance of the number of events are equal for individuals with any specific covariate pattern (*ie* set of predictors) (also assuming equal quantity of exposure). Consequently, one could have an overall variance greater than the overall mean in the raw data, but still meet the assumption of equidispersion if the variance among individuals with any set of predictor values equals the mean for that group. However, as a simple rule, if the variance in the raw data is greater than twice the overall mean, one must be suspicious that overdispersion will be present.

Having a variance much larger than the mean is a common problem with count data. This is called **extra-Poisson variation** or **overdispersion**. Overdispersion can arise in a variety of ways (see Hardin & Hilbe (2007); Hilbe (2007)).

Apparent overdispersion
Apparent overdispersion can be caused by any errors in the model. This can include omission of important explanatory predictors, outlying observations (potentially errors in the data?), failure to account for important interactions in the model, or failure to satisfy the assumption of linearity for continuous predictors. The solution to apparent overdispersion is to fix the model.

Real overdispersion
Real overdispersion occurs when the true variance in the counts is greater than the mean and it can also arise in a variety of ways. It may be that the variance of the counts is much larger than the mean value and that a model which allows for this greater variance is required. A common cause of real overdispersion is clustering of data (see Chapter 20) and potential solutions are discussed below. Alternatively, zero counts may either be more abundant or less frequent (or completely absent) than expected. The solution to this problem is to use hurdle, zero-inflated or zero-truncated models (discussed in Section 18.7).

As an example, in the TB data described in Example 18.1, each herd contributed multiple observations to the dataset. (While most herds had only one type of animal, they would have had multiple age classes and perhaps, males and females). The incidence rate of TB cases is more alike among groups of animals within a herd than across different herds. Consequently, part of the variation between groups of animals is due to the clustering in herds, and this has not been taken into account. Thus, the model does not fit the data well.

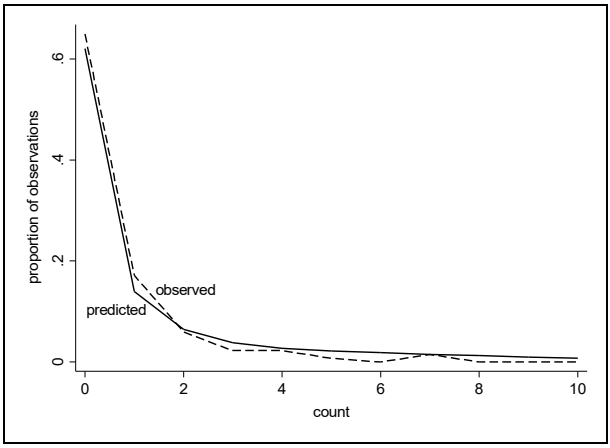

Fig. 18.2 Comparison of observed and predicted counts of TB reactors

18.5.4 Evaluating overdispersion

The amount of overdispersion can be quantified by computing a dispersion parameter by dividing either the Pearson or deviance χ^2 by its df (with the Pearson χ^2 generally being the preferred value). Values of the dispersion parameter >1 indicate overdispersion and should be a concern if >1.25 in moderate sample sizes or >1.05 in large sample sizes. The Pearson and deviance dispersion parameters for the TB data were 8.71 and 2.74, respectively (Example 18.1), reflecting the serious problem of overdispersion in these data.

The statistical significance of the amount of overdispersion can be assessed using the goodness-of-fit tests described in Section 18.5.2. Two alternatives are the score test and the Lagrange multiplier test. The reader is referred to Hilbe (2007) for details.

18.5.5 Dealing with overdispersion

There are several ways of dealing with overdispersion, some of which are discussed in this chapter and others which are presented elsewhere.

Scaled SEs of parameter estimates can be computed by scaling the SEs by the square root of either the deviance or Pearson dispersion factor (simulation studies have shown that the Pearson dispersion is preferred (Hilbe, 2007)). For example, in the model shown in Example 18.1, the SEs of the coefficients would be increased by $\sqrt{8.71} = 2.95$, resulting in a SE for cervids of 2.95*0.233=0.689 producing a *Z*-statistic of 1.55 with a P-value of 0.12. Alternative approaches to adjusting the SEs include the use of robust SEs (described in Section 14.9.5) or bootstrap or jackknife SEs (not described in this text).

One commonly used approach is to use **negative binomial regression** to fit a model in which the variance is allowed to be larger than the mean. This is described in more detail below (Section 18.6).

If the overdispersion is a result of clustering within the data, that clustering can be accounted for by adding **fixed effects** for the clusters to the model (see Section 20.5.2), by adding **random**

effects for the clusters (see Section 22.4.3) or through the use of **generalised estimating equations** (GEE—see Section 23.5). If the overdispersion is caused by clustering, methods for dealing with the clustering are preferable to using scaled SE or a negative binomial regression.

18.5.6 Influential points and outliers

Outliers may contribute to overdispersion, but even in the absence of evidence of overdispersion, it is important to evaluate outlying observations. Outliers can be identified by looking for large values of Pearson, deviance or Anscombe residuals. Influential points can be identified by looking for large values of Cook's distance (see Chapter 14 for introduction to Cook's distance). Examples of these are shown in Example 18.2. As with other forms of regression models, ill-fitting points must be checked thoroughly. If the data are incorrect, they must be fixed or excluded. If the data are correct, evaluation of poorly fitting points could provide insight into reasons why the model does not fit well.

18.6 NEGATIVE BINOMIAL REGRESSION

Negative binomial regression models are models for count data in which the variance is not constrained to equal the mean. These models can be derived in 2 ways: from the 2 parameter negative binomial distribution or as a Poisson-gamma mixture distribution. Each of these will be discussed below.

Example 18.2 Poisson regression—diagnostics
data = tb_real

Based on the model fit in Example 18.1, the observations with the 3 largest negative and positive Anscombe residuals are:

			age	pop	reactors		standardised residuals			Cook's
obs	type	sex	(mos)	at risk	obs.	pred.	dev.	Pear.	Ansc.	distance
89	cervid	male	>24	27410	0	6.26	-3.99	-2.82	-3.75	0.31
54	cervid	female	>24	26182	1	8.59	-3.55	-2.79	-3.35	0.17
53	cervid	female	12-24	12420	0	4.38	-3.16	-2.23	-3.14	0.10
25	dairy	female	0-12	389	1	0.00	3.08	17.48	3.81	0.08
45	cervid	female	>24	21848	29	7.17	6.50	8.66	6.18	1.37
133	beef	female	>24	6418	20	1.13	8.87	17.94	9.24	0.87

Large negative residuals were associated with groups of animals where many cases were expected, but few observed. Large positive residuals were found in groups of animals where many more cases were observed than were expected. Although they only accounted for 39% of the observations in the dataset, groups of cervids accounted for 4 of the 6 most extreme residuals suggesting that the model did not fit as well for cervids as it did for cattle.

(*continued on next page*)

Example 18.2 (*continued*)
data = tb_real

The 4 observations with the largest Cook's distance follow.

			age	pop	reactors		standardised residuals			Cook's
obs	type	sex	(mos)	at risk	obs.	pred.	dev.	Pear.	Ansc.	distance
118	cervid	female	12-24	21660	17	7.64	3.28	3.81	2.92	0.55
92	other	female	>24	9360	0	1.64	-2.71	-1.92	-1.92	0.66
133	beef	female	>24	6418	20	1.13	8.87	17.94	9.24	0.87
45	cervid	female	>24	21848	29	7.17	6.50	8.66	6.18	1.37

Large Cook's distances were associated with observations that had moderate or large residuals and contributed greatly to the overall time at risk. Observations with small amounts of time at risk, tended not to have a large impact on the model (small Cook's distance) (*eg* observation 25—in the table of large residuals).

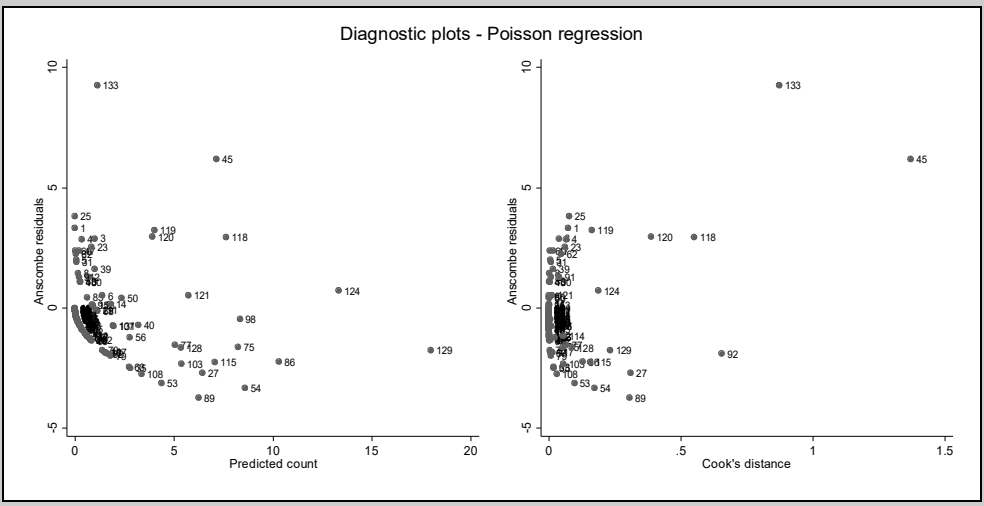

Fig. 18.3 Diagnostic plots for Poisson regression model

The plot of Anscombe residuals vs predicted counts indicates that large positive residuals were more common than negative ones. The plot of residuals vs Cook's distance highlights 2 very influential points (#133 and #45).

18.6.1 Negative binomial distribution

The negative binomial distribution is the probability of observing y failures before the r^{th} success in a series of Bernoulli trials. It is computed as:

$$f(y:r,p)=\binom{y+r-1}{r-1}p^r(1-p)^y \qquad \textit{Eq 18.7}$$

where y is the number of failures, r is the number of successes and p is the probability of success on each trial. As $r\to\infty$ (simultaneously with $p\to1$), the negative binomial distribution

for the number of failures approaches the Poisson distribution. The distribution can be expressed instead in terms of the parameters μ and α , where $\mu=r(1-p)/p$ is the mean and $\alpha=1/r$ is the dispersion parameter. Two special cases are $\alpha=0$ (the Poisson distribution) and $\alpha=1$ (the geometric distribution, giving the waiting time distribution until the first event). Fig. 18.4 shows 4 negative binomial distributions with various combinations of parameters. Comparing these distributions with the Poisson distributions with means of 2 and 5, you can see the more prominent right tails on the negative binomial distributions.

18.6.2 Poisson-gamma mixture distribution

If the observed counts for individuals with similar characteristics are expected to follow a Poisson distribution, but the individuals exhibit heterogeneity caused by some unmeasured characteristics, the observed counts will be more dispersed than expected from a Poisson distribution. This situation can be modelled directly by specifying a ('mixture') distribution for individual means. The standard choice is a gamma distribution (the gamma distribution is a flexible 2 parameter distribution which can take a wide variety of shapes), leading to Poisson-gamma mixture distributions for the observed counts. The negative binomial distribution of 18.6.1 can be derived in this way (with a suitably chosen gamma distribution). Focus in the development of these distributions has been on the way the variance depends on the mean. For example, if the variance is a constant multiple of the mean, this gives rise to an NB-1 model:

$$\text{NB-1} \qquad \text{var} = (1+\alpha)\mu = \mu + \alpha\mu \qquad \textit{Eq 18.8}$$

where α is referred to as the dispersion parameter.

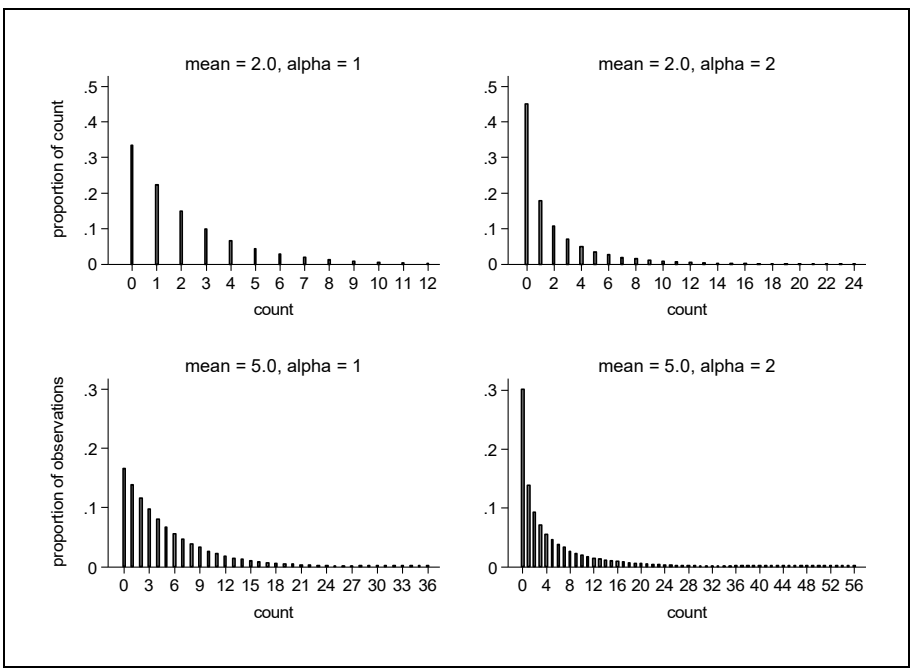

Fig. 18.4 Negative binomial distributions

In the most commonly used form of the negative binomial distribution, in this context denoted as an NB-2 model, the variance exceeds the mean by a factor which depends on the mean such that individuals with higher average counts will have relatively larger variances:

$$\text{NB-2} \qquad \text{var} = (1+\alpha\mu)\mu = \mu + \alpha\mu^2 \qquad \textit{Eq 18.9}$$

Note In either of the above 2 formulations, if α=0, then the variance once again equals the mean and the model is a simple Poisson model.

18.6.3 Negative binomial regression modelling

As with the Poisson distribution, the usual form of a negative binomial regression model is:

$$\text{E}(Y) = n\lambda \quad \text{or} \quad \frac{\text{E}(Y)}{n} = \lambda$$

where n is a measure of exposure (possibly constant) and λ is a function of the predictors, with the most usual form for λ being derived from a linear equation on a log scale, *eg*

$$\lambda = e^{\beta_0 + \beta_1 X} \quad \text{or} \quad \ln(\lambda) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.10}$$

Consequently, exponentiated regression coefficients from a negative binomial model in which the exposure was a measure of time at risk are interpreted as incidence rate ratios.

Full maximum likelihood (ML) estimation of the regression coefficients and the negative binomial dispersion parameter is available for distributions with a tractable form of the likelihood function, in particular NB-1 and NB-2. Example 18.3 shows a negative binomial model (NB-2) fit to the TB data using full ML estimation. In situations in which the dispersion parameter is very large (*ie* highly overdispersed data such as infectious disease data—see also Chapter 27), α may be underestimated if the sample size is small or zero counts are under-represented (Lloyd-Smith, 2007).

The generalised linear model (GLM) framework offers an alternative estimation approach (see Section 16.11 for an introduction to GLM models). GLM estimation of the Poisson model with a dispersion parameter (see Table 18.1, Section 18.6.4) yields estimates from a Poisson model with scaled SEs (Section 18.5.5), as opposed to a full ML estimation of the NB-1 model. An NB-2 distribution can be set up as a single parameter GLM with a log link, but the dispersion parameter α must be treated as a known constant. One solution to this limitation is to obtain an estimate of α using a full ML estimation, and then set α to this value when running the GLM procedure. Example 18.4 compares NB-2 models estimated using a full ML estimation procedure and the GLM framework. The value of the dispersion parameter (α) obtained from the ML estimation procedure was provided to the GLM estimation procedure. (**Note** The log link used to fit an NB-2 model is not the canonical link for the negative binomial distribution (see Section 16.11), but is the most commonly used link function.)

The advantage of the GLM framework is that it gives access to GLM goodness-of-fit statistics and the large number of GLM-defined residuals and other diagnostic parameters. Traditionally, GLM models have been estimated using an iteratively reweighted least squares algorithm, but maximum likelihood procedures are often now employed (when feasible).

Example 18.3 Negative binomial regression model
data = tb_real

The Poisson model from Example 18.1 was refit as a negative binomial model using full maximum likelihood estimation

Number of obs = 134
LR chi2(6) = 10.77
Prob > chi2 = 0.0956
Log likelihood = -157.73596 Pseudo R2 = 0.0330

Variable	Coef	SE	Z	P	95% CI	
type=beef	0.605	0.675	0.900	0.370	-0.718	1.927
type=cervid	0.666	0.684	0.970	0.330	-0.675	2.006
type=other	0.800	1.119	0.710	0.475	-1.393	2.993
sex=male	-0.057	0.405	-0.140	0.887	-0.851	0.736
age=12-24 mo	2.253	0.903	2.490	0.013	0.483	4.023
age=24+ mo	2.481	0.882	2.810	0.005	0.753	4.209
constant	-11.181	1.061	-10.540	0.000	-13.260	-9.103
par (exposure)						
/ln_alpha	0.554	0.253			0.057	1.051
alpha	1.740	0.441			1.059	2.860

Likelihood-ratio test of alpha=0: chibar2(01) = 161.85 Prob>=chibar2 = 0.000

Unlike the Poisson model, herd type and sex were no longer significant predictors. Animals over 12 months of age definitely had higher incidence rates.

Example 18.4 Comparison of maximum likelihood and GLM estimation of a negative binomial model
data = tb_real

The negative binomial model model from Example 18.3 was refit in the GLM framework and the results compared. The following table compares the estimated coefficients and SEs.

	Coef		SE	
Variable	ML	GLM	ML	GLM
type=beef	0.605	0.605	0.675	0.674
type=cervid	0.666	0.666	0.684	0.682
type=other	0.800	0.800	1.119	1.118
sex=male	-0.057	-0.057	0.405	0.404
age=12-24 mo	2.253	2.253	0.903	0.898
age=24+ mo	2.481	2.481	0.882	0.878
constant	-11.181	-11.181	1.061	1.053

The coefficients were identical but there were slight differences in the estimates of the SE.

18.6.4 Alternative variance functions

In addition to the NB-1 and NB-2 models described above, other multiplicative extensions to the Poisson variance have been developed and these are summarised in Table 18.1.

Table 18.1 Poisson variance

Model	Variance	Model	Variance
Poisson	var = μ	NB-1	var = μ(1+α) = μ + αμ
Poisson with a dispersion parameter	var = μ(θ)	NB-2	var = μ(1+αμ) = μ + αμ^2
Geometric	var = μ(1+μ) = μ+μ^2	NB-P	var = μ+αμ^p
		NB-H	var = μ(1+αμ) α =exp(β_0 + β_1 X_1)

The NB-1 model and Poisson model with a dispersion parameter (discussed in Section 20.5.5) have the same variance function (with $\theta=1+\alpha$), but are genuinely different models and are estimated in different ways (Section 18.6.3). In all of the above models, if α=0, then the variance once again equals μ and the model is a simple Poisson model. The NB-P model is a generalisation of other NB models with μ raised to the power p. (The reader is referred to Hilbe (2007) for details.) The NB-H model (heterogeneous or generalised NB model) allows the dispersion parameter to be modelled as a function of predictors. An example of an NB-H model is presented in Section 18.7.3.

18.6.5 Evaluating overdispersion

A likelihood ratio test which compares the usual Poisson model to the negative binomial model is equivalent to a test of α=0. This provides a formal test for the presence of overdispersion in the model. Because α cannot be negative, this is a 1-tailed test. As can be seen in Example 18.3, the results of this test are highly significant (P<0.001) indicating a problem with overdispersion.

As the additional variance is now a function of both α and μ [var=$(1+\alpha\mu)\mu$], the amount of overdispersion is a function of both values. If $\alpha\mu>1$, then $(1+\alpha\mu)>2$, which would indicate substantial overdispersion. For example, if α=0.5 and most counts are 0, 1 or 2 with a mean of 1.0, then $(1+\alpha\mu)$=1.5, so there is only moderate evidence of overdispersion. However, if α=0.5 and most counts range from 0 to 15, with a mean of 5.0, then $(1+\alpha\mu)$=3.5 which is indicative of serious overdispersion. Example 18.5 provides an example of a negative binomial model and an assessment of overdispersion.

18.6.6 Negative binomial regression diagnostics

Diagnostics for negative binomial models (Example 18.5) are similar to those for Poisson models. Plots of standardised deviance residuals and/or Anscombe residuals vs predicted counts will identify particularly poorly fit observations.

18.6.7 Generalised negative binomial models

In Section 18.6.4, it was shown that the variance of a negative binomial model could be modified to allow it to be a function of one or more predictors. Example 18.6 fits this type of model to the tuberculosis data with the variance as a function of -age-.

Example 18.5 Negative binomial regression—diagnostics
data = tb_real

The negative binomial model shown in Example 18.3 forms the basis of this example. The likelihood ratio test of $\alpha = 0$ is highly significant (P<0.001) suggesting that the variance in the data is higher than would be expected for a Poisson regression. Since the overall mean number of reactors in these data was 1.46, the value of $(1+\alpha\mu)=1+(1.74*1.46)=3.54$ (*ie* substantial overdispersion).

The 2 goodness-of-fit tests give very discrepant results. The deviance χ^2 goodness-of-fit test was not significant (χ^2=99.4 on 127 df, P=0.97) showing no evidence of lack of fit. On the other hand, the Pearson χ^2 test was highly significant (χ^2=374.9 on 127 df, P<0.001) with a dispersion parameter of 2.95. However, a listing of the observations with the 4 largest Cook's distance values showed that the very large Pearson χ^2 was due to 3 observations with very large Pearson residuals. These were singleton reactors when none were expected (obs. #1 and #25) or a group of beef cows with many more reactors than expected (obs #133). The model was heavily influenced by this last group. As was suggested in Chapter 15, you might omit this group of animals and refit the model. However, this should be done only to further evaluate the impact of this observation on the model.

			age	pop	reactors		standardised residuals			Cook's
obs	type	sex	(mo)	at risk	obs.	pred.	dev.	Pear.	Ansc.	distance
62	other	male	>24	344	1	0.12	1.33	2.38	1.40	0.06
1	beef	female	0-12	525	1	0.01	2.36	8.46	2.61	0.08
25	dairy	female	0-12	389	1	0.01	2.71	13.48	3.12	0.12
133	beef	female	>24	6418	20	1.96	2.65	6.27	5.87	0.23

The plot of the Anscombe residuals vs predicted values shows few large negative residuals but some very large positive values in groups of animals with low predicted counts. The plot of Anscombe residuals vs Cook's distance clearly highlights the influential role of observations #133.

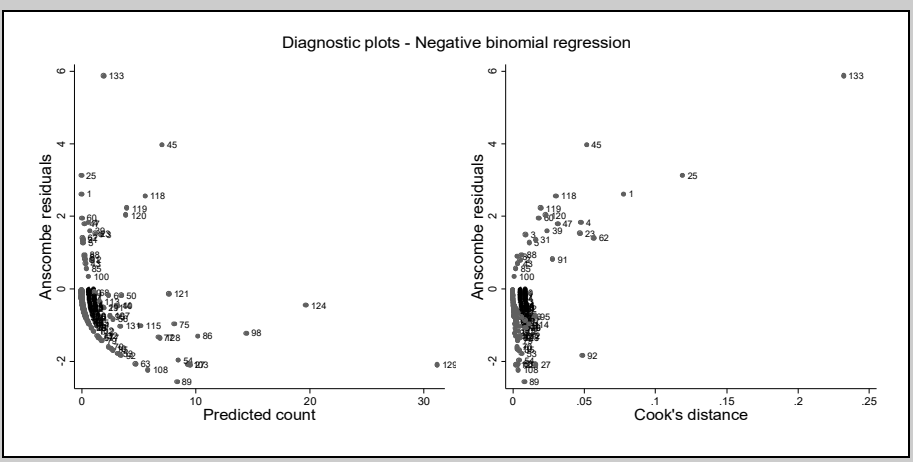

Fig. 18.5 Negative binomial diagnostic plots

18.7 PROBLEMS WITH ZERO COUNTS

The number of observations with a count of 0 in a dataset may be higher or lower than would be expected from a Poisson (or negative binomial) distribution. If there is an excess of zero counts, you can fit either a zero-inflated model or a hurdle model (each discussed briefly below). If zero counts are not possible (as is the case if the count is the number of services required for conception in a dairy cow), then a zero-truncated model can be fit to the data.

Example 18.6 Generalised negative binomial regression
data = tb_real

The Poisson model from Example 18.1 was refit using a generalised negative binomial model with -type- -sex- and -age- retained as predictors of the mean reactor count and -age- as the sole factor influencing the variance.

Number of obs = 134
LR chi2(6) = 3.09
Prob > chi2 = 0.7973
Log likelihood = -155.15399 Pseudo R2 = 0.0099

					95% CI for Coef	
Variable	**Coef**	**SE**	**Z**	**P**	**Lower**	**Upper**
type=beef	0.827	0.646	1.280	0.201	-0.440	2.094
type=cervid	1.030	0.672	1.530	0.126	-0.288	2.347
type=other	1.130	1.107	1.020	0.307	-1.039	3.299
sex=male	0.023	0.400	0.060	0.953	-0.761	0.808
age=12-24 mos	0.252	1.425	0.180	0.860	-2.542	3.045
age=24+ mos	0.568	1.396	0.410	0.684	-2.168	3.304
constant	-9.524	1.378	-6.910	0.000	-12.226	-6.823
lnalpha						
age=12-24 mos	-3.533	1.136	-3.110	0.002	-5.761	-1.306
age=24+ mos	-2.920	1.020	-2.860	0.004	-4.918	-0.921
constant	3.599	0.984	3.660	0.000	1.669	5.528

The results suggest that the effect of age appears to be that it increases the variance in response much more than it affects the mean response. In fact, none of the predictors in the portion of the model dealing with the mean response remain significant. A comparison of the 2 models was carried out using information criteria (AIC and BIC).

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
NB-2	134	-163.12	-157.74	8	331.47	354.65
NB-H	134	-156.70	-155.15	10	330.31	359.29

The AIC favours the NB-H model, while the BIC (which tends to favour parsimonious models) favours the NB-2 model. In this situation the simpler model (NB-2) would be preferred for its simplicity and for the fact that we normally expect predictors to affect means counts rather than the variance of the responses. In the NB-2 model, age strongly affects the expected mean count, while in the NB-H model it affects the variance and has no significant effect on the mean.

18.7.1 Zero-inflated models

One occasionally encounters situations in which the distribution of outcome events might follow a Poisson (or negative binomial) distribution, except that there is an excess of zero counts in the data. This might be because there are 2 processes by which zero counts can arise. For example, assume your outcome of interest is the number of slaughter hogs with lesions of enzootic pneumonia (caused by *Mycoplasma hyopneumonia*) at slaughter, in samples taken from many herds. In herds in which *Mycoplasma hyopneumonia* is endemic, the count of lesioned animals might follow a Poisson distribution, and some herds could still have zero counts (*ie* no animals with lesions at slaughter). However, other herds will have zero counts because *Mycoplasma hyopneumonia* is not present in the herd. Consequently, a count of zero might arise from either of the 2 situations.

Zero-inflated models deal with an excess of zero counts by simultaneously fitting a binary model (usually a logistic regression model but it could also be a probit or complementary log-log model) and a Poisson (or negative binomial) model. The 2 models might have the same, or different, sets of predictors. The parameter modelled in the binary model is the probability of a zero count so coefficients have an opposite sign than they would in a usual logistic regression (and if the same predictor is in the Poisson model, they often have opposite signs in the 2 models).

Whether or not a zero-inflated model fits the data better than the usual Poisson or negative binomial model can be assessed using a Vuong test (V). This test compares 2 non-nested models and is asymptotically normally distributed. If the value of V is <-1.96, one model (*eg* the usual Poisson or negative binomial model) is favoured. If V >1.96, the second model (*ie* the zero-inflated model) is favoured. If V lies between -1.96 and 1.96, neither model is preferred.

A zero-inflated negative binomial model was used to model factors affecting fecal egg counts in adult dairy cattle (Nødtvedt *et al*, 2002). Fecal egg counts are generally low in adult cattle and might arise because the animal is uninfected or is infected but shedding eggs in numbers too low to be detected. Example 18.7 shows the application of a zero-inflated negative binomial model (with a subset of the original predictors) to these data.

18.7.2 Hurdle models

Like zero-inflated models, a hurdle model has 2 components but it is based on the assumption that zero counts arise from only one process and non-zero counts are determined by a different process (Hilbe, 2007). The enzootic pneumonia example used in the description of zero-inflated models could also apply to hurdle models if it was believed that zero counts (*ie* no pneumonia) were not possible in barns in which *Mycoplasma hyopneumonia* was endemic.

Hurdle models use some form of binomial model (logit, probit or complementary log-log) to model the odds of a non-zero count (vs a zero count) and some form of zero-truncated model (Poisson, negative binomial or geometric) to model the distribution of non-zero counts. Zero-truncated models are described in Section 18.7.3. Consequently, there are 9 possible combinations of models possible (*eg* logit-Poisson, probit-negative binomial, *etc*).

Example 18.8 shows the use of a logit-negative binomial hurdle model for the fecal egg count data presented in Example 18.7. In contrast to the logit portion of the zero-inflated model, in a hurdle model, the coefficients in the logit model reflect how the log odds of a non-zero count is

Example 18.7 Zero-inflated negative binomial model
data = fec

Fecal egg counts (n=2,250) from 313 cows in 38 dairy herds in 4 regions of Canada were determined over a 1-year period in conjunction with a clinical trial of eprinomectin as a treatment for gastro-intestinal nematodes in Canadian dairy cattle. A more detailed description of the dataset can be found in Chapter 27. Although the mean fecal egg count was 8.6 eggs/5 gm, one-half the observations had zero counts.

Counts obtained from control cows and treated cows prior to treatment were analysed using a zero-inflated negative binomial model. Only a subset of the predictors from the original study are used in this example. Lactation (2 groups), and several herd management variables were included in both portions of the model. Clustering of observations within cows was accounted for by using robust standard error (see Section 20.5.4) estimates. The resulting model follows.

Number of obs = 1840
Nonzero obs = 983
Zero obs = 857
Inflation model = logit
Wald chi2(4) = 67.91
Log pseudolikelihood = -4573.221
Prob > chi2 = 0.0000

Variable	Coef	Robust SE	Z	P	95% CI	
Negative binomial portion						
multiparous	-0.978	0.229	-4.270	0.000	-1.426	-0.529
past_lact	0.602	0.343	1.750	0.079	-0.071	1.274
man_heif	-1.059	0.265	-3.990	0.000	-1.578	-0.539
man_lact	1.018	0.333	3.060	0.002	0.366	1.670
_cons	2.367	0.290	8.150	0.000	1.798	2.936
Logistic portion						
multiparous	0.716	0.485	1.480	0.140	-0.234	1.666
past_lact	-1.800	0.713	-2.520	0.012	-3.198	-0.402
man_heif	-1.377	1.072	-1.280	0.199	-3.478	0.724
man_lact	-19.983	0.788	-25.370	0.000	-21.527	-18.439
_cons	-0.712	0.426	-1.670	0.094	-1.546	0.122
/lnalpha	1.281	0.067	19.180	0.000	1.150	1.412
alpha	3.601	0.241			3.159	4.105

The Vuong statistic was 5.00 suggesting that the zero-inflated model was clearly superior to the regular negative binomial model. The value of α (3.60, 95% CI of 3.16 to 4.10) suggests that a negative binomial model is preferable to an ordinary Poisson model. The coefficients for -multiparous- in the negative binomial portion (-0.98) and the logistic portion (0.72) of the model indicated that multiparous cows generally had lower fecal egg counts and were more likely to have zero egg counts, although this latter effect was not statistically significant (P=0.14). Having lactating cows on pasture (-past_lact-) reduced the probability of a zero count (β=-1.8) and increased the expected count if it was non-zero (β=0.60).

Example 18.8 Logit–negative binomial hurdle model
data = fec

A logit–negative binomial hurdle model was fit to the same data used in Example 18.7 and with the same predictors.

Log pseudolikelihood = -4546.5055

Number of obs = 1840
Wald chi2(4) = 66.75
Prob > chi2 = 0.0000
(SE adjusted for 311 clusters in cow)

Variable	Coef	Robust SE	Z	P	95% CI	
Logit portion of the model						
Multiparous	-0.671	0.169	-3.970	0.000	-1.002	-0.339
past_lact	0.862	0.204	4.240	0.000	0.463	1.261
man_heif	-0.311	0.248	-1.250	0.211	-0.798	0.176
man_lact	0.942	0.267	3.530	0.000	0.419	1.466
_cons	-0.175	0.177	-0.990	0.324	-0.522	0.173
Negative binomial (count) portion of the model						
Multiparous	-0.992	0.253	-3.920	0.000	-1.489	-0.496
past_lact	0.670	0.373	1.800	0.072	-0.061	1.401
man_heif	-1.032	0.297	-3.470	0.001	-1.614	-0.450
man_lact	0.909	0.357	2.540	0.011	0.209	1.610
_cons	1.135	0.894	1.270	0.204	-0.618	2.887
/lnalpha	2.928	0.929	3.150	0.002	1.107	4.749

The coefficients for multiparous suggest that multiparous cows have a reduced log odds of a non-zero count (β=-0.67) and lower counts if it was non-zero (β=-0.99). Spreading of manure on pastures used by heifers had a more significant effect on the fecal egg count (if it was non-zero) (P=0.001) than on the probability of it being non-zero (P=0.21).

Log-likelihoods were computed to compare the zero-inflated and hurdle models and the comparison showed that the hurdle model clearly fit better. **Note** The AIC and BIC are unnecessary because the models have the same number of df.

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
zinfl	1840	-4660.742	-4573.221	11	9168.442	9229.135
hrdl	1840		-4546.506	11	9115.011	9175.704

affected by the predictor. Consequently, they often have the same sign as the coefficients in the count portion of the model.

The choice of whether to use an ordinary negative binomial model, a zero-inflated model or a hurdle model should be based on a combination of the fit of the models (log-likelihood) and the biology of the process being modelled. Is it reasonable to consider zeros arising from a second process? If so, would you expect zeros to arise from both processes (zero-inflated model) or just the second process (hurdle model).

18.7.3 Zero-truncated models

In some situations, zero counts are not possible. For example, when recording the number of breedings required for a cow to conceive after calving, zero is not a possibility. Similarly, the length of stay (in days) of a dog in a veterinary hospital can not be zero. One approach to this problem is to subtract 1 from all outcomes and model the revised outcome (which will contain zeros). An alternative approach is to use a zero-truncated model which allows for a defined distribution of counts but the probability of a zero is eliminated from the likelihood function for the model. The probability of a zero count is computed from either the Poisson or negative binomial distribution and this value is subtracted from 1. The remaining probabilities (*eg* probability of counts of 1,2,3 *etc*) are then rescaled based on this difference so they total 1.

Example 18.9 shows a zero-truncated negative binomial model fit to some data on 'services-per-conception' in dairy cows.

Example 18.9 Zero-truncated negative binomial model
data = daisy2

A zero-truncated negative binomial model was fit to some data on 'services-per-conception' in dairy cows. Three predictors: parity, days from calving to first service (both modelled as linear effects) and presence/absence of vaginal discharge after calving were included in the model. The effect of clustering within herd was accounted for through the use of robust SEs clustered on herd.

Number of obs = 1744
Log likelihood = -2683.2176 Wald chi2(3) = 219.30
Prob > chi2 = 0.0000

Variable	Coef.	Robust SE	Z	P	95% CI	
parity	0.059	0.018	3.230	0.001	0.023	0.095
days_fs	-0.008	0.001	-11.730	0.000	-0.009	-0.006
vag_disch	0.293	0.101	2.890	0.004	0.094	0.492
_cons	0.367	0.182	2.020	0.043	0.011	0.723
/lnalpha	0.490	0.222			0.055	0.924
alpha	1.632	0.362			1.057	2.520

As cows aged, the number of services required for conception increased, but the longer the first breeding was delayed (-days_fs-), the fewer breedings that were ultimately required. Finally, cows with evidence of vaginal discharge (-vag_disch-) after calving required, on average, 34% ($e^{0.293}$=1.34) more breedings than cows without discharge. α was clearly greater than zero (95% CI=1.06 to 2.52) indicating that a zero-truncated negative binomial model was preferred to one based on a Poisson distribution.

Comparable results were obtained by subtracting 1 from the values of -spc- and modelling this new outcome (which contained zeros) using an ordinary negative binomial model (data not shown).

REFERENCES

Barros AJD, Hirakata VN. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio BMC Med Res Methodol. 2003; 3: 21.

Cameron A, Trivedi P. Regression Analysis of Count Data. Cambridge University Press: Cambridge; 1998.

Greenland S. Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies Am J Epidemiol. 2004; 160: 301-5.

Hammond RF, McGrath G, Martin SW. Irish soil and land-use classifications as predictors of numbers of badgers and badger setts Prev Vet Med. 2001; 51: 137-48.

Hardin J, Hilbe J. Generalized Linear Models and Extensions, 2nd Ed. Stata Press: College Station; 2007.

Hilbe J. Negative Binomial Regression. Cambridge University Press: Cambridge; 2007.

Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases PLoS ONE. 2007; 2: e180.

Long J. Regression Models for Categorical and Limited Dependent Variables. Sage Publications: Thousand Oaks; 1997.

Long J, Freese J. Regression Models for Categorical Dependent Variables Using Stata, 2nd Ed. Stata Press: College Station; 2006.

McNutt L, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes Am J Epidemiol. 2003; 157: 940-3.

Munroe FA, Dohoo IR, McNab WB. Estimates of within-herd incidence rates of *Mycobacterium bovis* in Canadian cattle and cervids between 1985 and 1994 Prev Vet Med. 2000; 45: 247-56.

Munroe FA, Dohoo IR, McNab WB, Spangler L. Risk factors for the between-herd spread of *Mycobacterium bovis* in Canadian cattle and cervids between 1985 and 1994 Prev Vet Med. 1999; 41: 119-33.

Nødtvedt A, Dohoo I, Sanchez J, Conboy G, DesCôteaux L, Keefe G, Leslie K, Campbell J. The use of negative binomial modelling in a longitudinal study of gastrointestinal parasite burdens in Canadian dairy cows Can J Vet Res. 2002; 66: 249-57.

Sanchez J, Nødtvedt A, Dohoo I, DesCôteaux L. The effect of eprinomectin treatment at calving on reproduction parameters in adult dairy cows in Canada Prev Vet Med. 2002; 56: 165-77.

Zou G. A modified Poisson regression approach to prospective studies with binary data Am J Epidemiol. 2004; 159: 702-6.