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. It might be a:

a. simple count of events, such as the total number of births a woman has. In this chapter we will investigate factors that relate to the number of prenatal visits recorded by pregnant women.
b. count of cases of disease over a period of time with the amount of person-time at risk having to be taken into consideration (*eg* number of migraine headaches reported by participants in a cohort study with the follow-up time for each person taken into account). This is a measure of the incidence rate (I) of disease. In this chapter, we will analyse this kind of data by taking the length of the gestation period into account so we will have the number of prenatal visits per week of gestation.
c. count of cases of disease with the size of the population at risk being taken into consideration (*eg* number of cases of a specific type of cancer in various cities over a year with the population size (people over a specified age) treated used to compute the person-years at risk). Hence, this is also a measure of the incidence rate (I) of disease.
d. 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. (Badgers are a reservoir for *Mycobacterium bovis* in Ireland, and so badger density has public health significance.) 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

Let's say that we want to evaluate the effect of propranolol on the incidence of migraine headaches. We will assume that a controlled trial can be carried out with 3 medical clinics participating and individuals assigned to be on propranolol or not. The outcome of interest might be the total number of cases of migraine occurring in each person over the one-year follow-up period. Other factors that will have to be taken into consideration are the age and gender of the person and which clinic they are associated with (because either incidence rates or reporting rates of migraines can vary among clinics). While random assignment of people to treatment groups should balance the age, gender, and clinic 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 people because not all people will be observed for the full year. **Note** In this example, we are interested in the total number of cases of migraine. If we were interested in whether or not the study participant had any cases of migraine, we could create a binary variable for each person 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 migraines could be computed for each study group and the difference between the 2 groups tested using the unconditional *Z*-test we discussed in Chapter 6. This approach does not allow for the control of other potential confounding variables (*ie* age, gender, clinic), so it would rely totally on random assignment to control for these effects.
b. Alternatively, you could determine the incidence rate (*I*) of migraines within each person and use that value as the dependent variable in a linear regression with propranolol as the primary exposure (predictor) of interest, and age, gender, and clinic as extraneous variables. However, incidence rates rarely have anything close to a normal distribution. Consequently, one of the fundamental assumptions of linear regression would be violated (unless a suitable transformation of the dependent variable could be found). It is also possible that some combination of predictor variables may predict a negative *I* for the person. This approach looks worse than the preceding approach does.
c. The preferred approach is to use Poisson or negative binomial regression to model the incidence of migraines while adjusting for the amount of time each person was observed.

Numerous texts dealing specifically with the analysis of count data are available and include: Cameron and Trivedi (1998), Hilbe (2011), Long (1997), Long and Freese (2006).

18.2 The Poisson distribution

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

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

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 migraine headache) are independent and follow an exponential distribution with a mean value of t, then the number of cases of migraine (Y) in a defined time period (T) will follow a Poisson distribution with $\mu=T/t$. For example, if the mean time between migraines is 15 days, then the expected number of cases in a 30-day period will be 30/15=2 cases. The time between events is sometimes referred to as the 'waiting time.' A specific feature of the exponential distribution of waiting times is that, at any point in time, the time to the next event is independent of the time that has passed since the last event. This property is referred to as 'memorylessness'. 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 a migraine on any given day is 1/15=0.067, then the expected number of cases in a 30-day period will be 30*0.067=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

migraines experienced by an individual in a month is 5, the probability of getting 10 cases in a month 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 migraines in a month (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.

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* person-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 (people or groups of people). (**Note** Throughout this text, the letter n is most commonly used to denote the number of people in a population. Here we are also using it to denote the amount of person-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

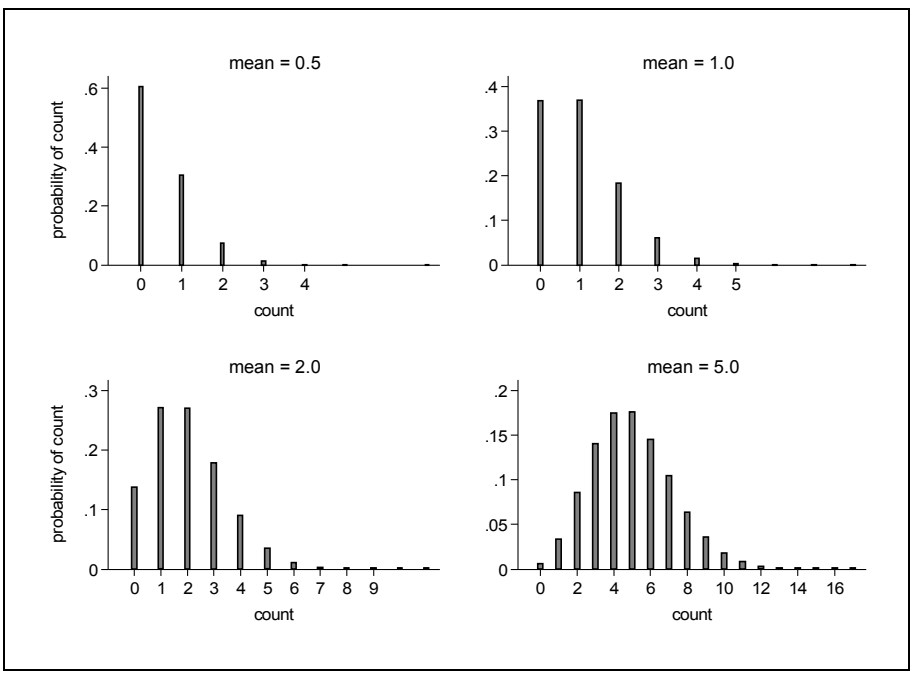

Fig. 18.1 Poisson distributions

expected number of cases in n people-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:

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

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

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 people-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 number of prenatal visits by pregnant mothers (birth weight dataset) 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 maternal age in a study of prenatal visits was 1.05, that would suggest that the incidence rate of visits went up by 5% for each additional year of age (*ie* that it was 1.05 times as high as the rate in the previous year). **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 (even though in the example used, the outcome (prenatal visits) is not a disease).

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 meduc=college

Example 18.1 Poisson regression model
data = bw5k

Both the count (total number during the pregnancy) of prenatal visits by expectant mothers and the rate (visits per week) were modelled using Poisson regression. The predictors were:

white:	mother's race (0=non-white, 1=white)
mage_28:	mother's age (centred on 28 years)
tbo_1:	parity (scaled so 0 = 1st parity, 1=2nd, *etc*)
meduc_c4:	mother's education level (values 1-4)

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

The model of the rate of visits with the 4 predictor variables and the time at risk included as an exposure variable produced the following results.

Number of obs = 5000
Log likelihood = -13820.24

					95% CI for IR	
Variable	**Coef**	**SE**	**Z**	**P**	**Lower**	**Upper**
white	0.034	0.009	3.72	0.000	0.016	0.051
mage_28	0.004	0.001	4.43	0.000	0.002	0.006
tbo_1	-0.014	0.003	-4.19	0.000	-0.020	-0.007
meduc = HS dip.	0.056	0.014	4.13	0.000	0.029	0.082
meduc = some coll.	0.074	0.014	5.18	0.000	0.046	0.102
meduc = coll. dip.	0.072	0.014	5.02	0.000	0.044	0.100
constant	-1.283	0.012	-103.16	0.000	-1.307	-1.258

All predictors were significant. Exponentiating the constant ($e^{-1.283}$=0.28) suggests that a non-white, 28-year-old, 1st parity mother without a high school diploma (*ie* the 'baseline individual') would be expected to have 0.28 visits per week (or approximately one visit every 4 weeks).

The model based solely on counts (no accounting for gestation length), produced very similar coefficients (data not shown) except for the constant which was 2.373. This constant indicates that the 'baseline person' (described above) would be expected to have $e^{2.373}$=10.7 visits during the pregnancy.

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

	df	X^2	P	dispersion
Deviance	4993	6681.8	<0.001	1.34
Pearson	4993	6165.4	<0.001	1.23

Both values suggest that there are problems with the model (*ie* evidence of lack of fit).

diploma in Example 18.1 was $e^{.072}$=1.075 (compared with no high school diploma). The predicted *I* for a 'baseline individual' (28-year-old, non-white, 1st parity mother without a high school diploma) was $e^{-1.283}$=0.277 visits/week, while for a 'higher frequency individual' (white, 38 years old) the rate was $e^{(-1.283+.034+10*.004)}$=0.298 visits per week. If that individual also has a college diploma, the frequency of visits increases to $e^{(-1.283+.034+10*.004+.072)}$=0.321 (a gain of 0.023 visits per week). Over a 39-week gestation, this rate of visits would equate to 12.5 visits.

Note Although all of the predictors are statistically significant, it is clear that the impact they have on the expected number of visits is quite small.

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*). Risk ratios (*RR*) 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 and 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 and 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. Pearson and deviance residuals may be standardised to give them unit variance.

Anscombe residuals are similar to standardised deviance residuals (Hilbe, 2011) 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, 2011).

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 and 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 prenatal visit data are presented in Example 18.1 (both are highly significant, 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 suspect 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 and Hilbe (2007), Hilbe (2011).

Apparent overdispersion
Apparent overdispersion can be caused by any errors in the model. This can include omission of

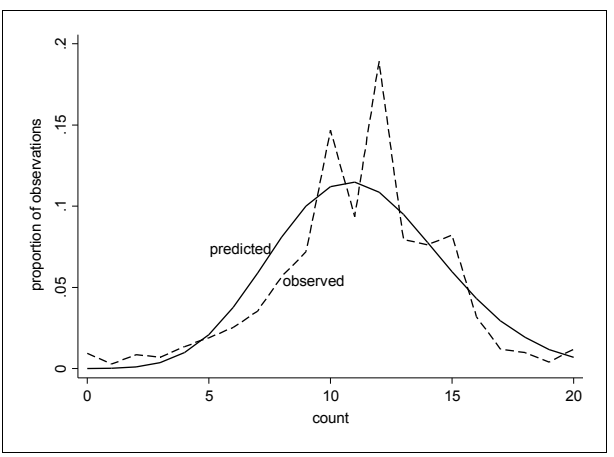

Fig. 18.2 Comparison of observed and predicted counts of prenatal visits

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 than expected (or completely absent). The solution to this problem is to use hurdle, zero-inflated or zero-truncated models (discussed in Section 18.7).

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 prenatal visit data were 1.23 and 1.34, respectively (Example 18.1), reflecting some 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 (2011) for details.

18.5.5 Dealing with overdispersion

There are several ways of dealing with overdispersion, some of which are discussed in this chapter (see also Chapters 20, 22, and 23).

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, 2011)). For example, in the model shown in Example 18.1, the SEs of the coefficients would be increased by $\sqrt{1.23}=1.11$, resulting in a SE for white of 1.11*.0091=0.0101 (P=0.001). Alternative approaches to adjusting the SEs include the use of robust SEs (described in Section 14.9.5), 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.

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:

Example 18.2 Poisson regression—diagnostics
data = bw5k

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

					Prenatal visits		Standardised residuals		
obs	mother race	mother age (+28)	parity	mother educ	obs.	pred.	dev.	Pear.	Ansc.
3397522	white	3	0	univ. deg.	0	12.5	-4.996	-3.533	-5.299
3719982	white	4	3	univ. deg.	0	12.3	-4.966	-3.512	-5.268
3476040	white	7	3	some col.	0	12.2	-4.941	-3.494	-5.241
430977	non-white	-2	1	hs dip.	30	10.3	4.964	6.121	4.994
1771199	white	1	3	hs dip.	34	11.1	5.507	6.874	5.544
726794	white	-3	7	some col.	40	11.1	6.706	8.694	6.765

Large negative residuals were associated with mothers who had no prenatal visits. Extremely large positive residuals were found in mothers with 30 or more visits. There were 124 Pearson residuals that were >3 or <-3 (we would only expect approximately 1%=50).

(*continued on next page*)

Example 18.2 (*continued*)

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

obs	mother race	mother age (+28)	parity	mother educ	prenatal visits obs	prenatal visits pred.	prenatal visits Ansc. res.	Cook's distance
1733432	white	15	7	< hs dip.	0	11.1	-5.000	0.008
1771199	white	1	3	hs dip.	34	11.1	5.544	0.008
1394806	white	6	0	some col.	30	11.4	4.601	0.009
726794	white	-3	7	some col.	40	11.1	6.765	0.055

One observation stood out as having a very large Cook's D (726794). This mother had 40 prenatal visits, far more than any other individual in the dataset (the next highest was 34 visits).

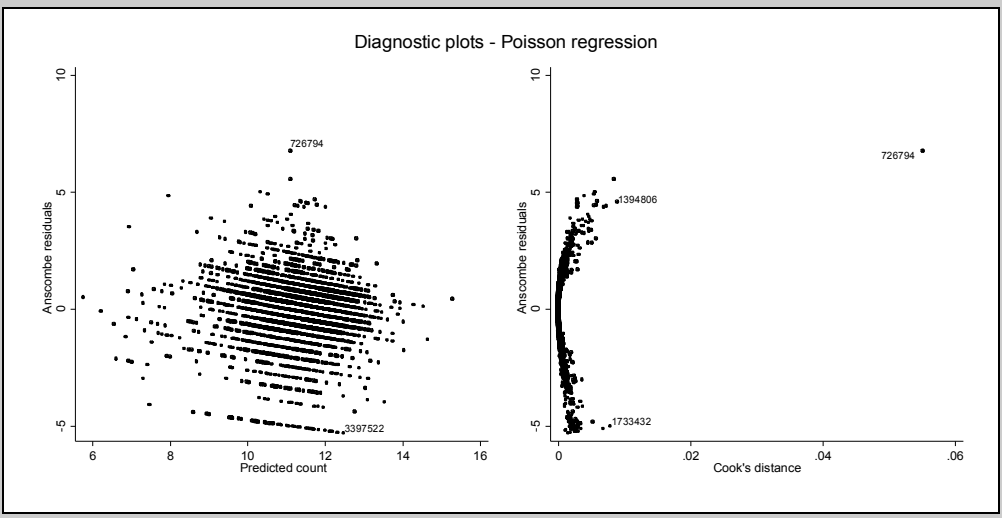

Fig. 18.3 Diagnostic plots for Poisson regression model

The plot of Anscombe residuals vs predicted counts indicates that extreme negative residuals were more common than positive ones. The plot of residuals vs Cook's distance highlights the very influential points (#726794). Refitting the model without this point reduced the coefficients for the 2 upper levels of mother's education by 3.1–4.6%.

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

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 Section 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.

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:

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

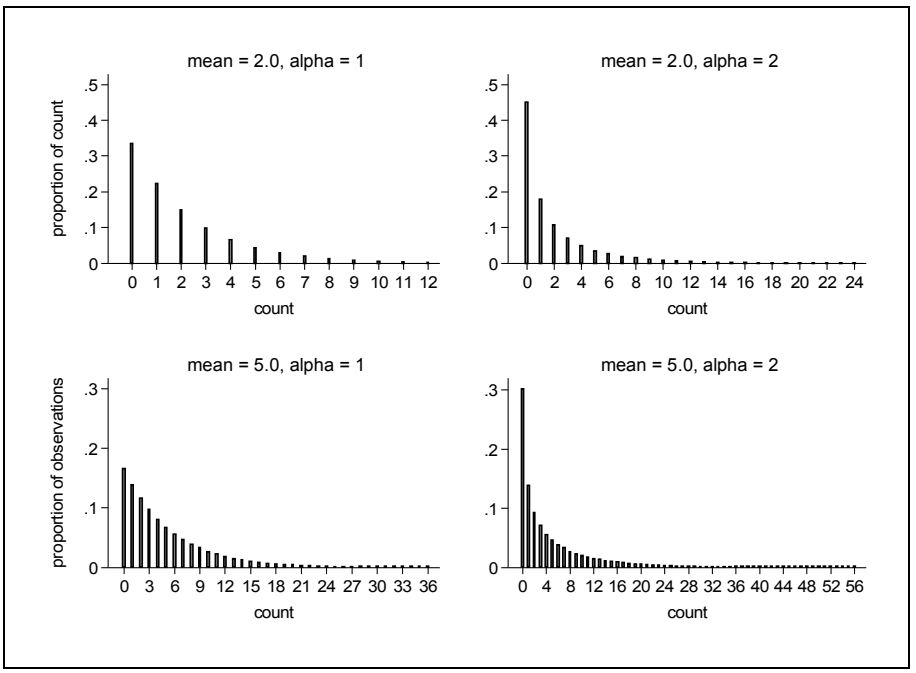

Fig. 18.4 Negative binomial distributions

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:

$$\mathrm{E}(Y) = n\lambda \quad \text{or} \quad \frac{\mathrm{E}(Y)}{n} = \lambda \qquad \textit{Eq 18.10}$$

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*

$$\lambda = e^{\beta_0 + \beta_1 X} \quad \text{or} \quad \ln(\lambda) = \beta_0 + \beta_1 X \qquad \textit{Eq 18.11}$$

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 prenatal visit 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 underrepresented (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).

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.

Example 18.3 Negative binomial regression model
data = bw5k

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

Number of obs = 5000
LR chi2(6) = 115.06
Prob > chi2 = 0.0000
Log likelihood = -13772.922 Pseudo R2 = 0.0042

Variable	Coef	SE	Z	P	95% CI	
white	0.034	0.010	3.39	0.001	0.014	0.053
mage_28	0.004	0.001	4.03	0.000	0.002	0.006
tbo_1	-0.014	0.004	-3.80	0.000	-0.021	-0.007
meduc = HS dip.	0.056	0.015	3.79	0.000	0.027	0.085
meduc = some coll.	0.074	0.016	4.73	0.000	0.044	0.105
meduc = univ. deg.	0.072	0.016	4.57	0.000	0.041	0.103
constant	-1.283	0.014	-94.15	0.000	-1.309	-1.256
gest (exposure)						
/ln_alpha	-3.973	0.118			-4.204	-3.742
alpha	0.019	0.002			0.015	0.024

Likelihood-ratio test of alpha=0: chibar2(01) = 94.63 Prob>=chibar2 = 0.000

As with the Poisson model, all predictors remain significant and the coefficients are close to those obtained from the Poisson model. The likelihood ratio test of α is highly significant (P<0.001) confirming the presence of overdispersion (relative to the Poisson model). Since the overall mean number of prenatal visits was 11.3, the value of $(1+\alpha\mu)=1+(0.019*11.3)=1.21$ (*ie* moderate overdispersion). The variance of the Poisson model would be assumed to be 11.3, while for the NB-2 model it would be $\mu+\alpha\mu^2=11.3+.019*11.3^2=13.7$.

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 20.5.3) 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 $\alpha=0$, 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

Example 18.4 Comparison of maximum likelihood and GLM estimation of a negative binomial model
data = bw5k

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

	Coef		SE	
Variable	**ML**	**GLM**	**ML**	**GLM**
white	0.034	0.034	0.010	0.010
mage_28	0.004	0.004	0.001	0.001
tbo_1	-0.014	-0.014	0.004	0.004
meduc = HS dip.	0.056	0.056	0.015	0.015
meduc = some coll.	0.074	0.074	0.016	0.016
meduc = univ. deg.	0.072	0.072	0.016	0.016
constant	-1.283	-1.283	0.014	0.014

The coefficients and SEs were identical.

(2011) 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 18.6.7.

18.6.5 Evaluating overdispersion

A likelihood ratio test which compares the usual Poisson model with 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 18.3 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.

Example 18.5 Negative binomial regression—diagnostics
data = bw5k

Based on the model fit in Example 18.3, the 2 goodness-of-fit tests give very discrepant results. The deviance χ^2 goodness-of-fit test was highly significant (χ^2 = 5626.7 on 4,993 df, P = <0.001) suggesting significant lack of fit. On the other hand, the Pearson χ^2 test was not significant (χ^2=5102.1 on 4,993 df, P=0.14) with a dispersion parameter of only 1.02. There are now only 74 observations (1.5%) with residuals more extreme than +/-3. As before, there was a single observation (obs = 726794) with a very large Cook's *D*. As was suggested in Chapter 15, you might omit this individual and refit the model to determine the impact of this observation. As with the Poisson model, eliminating this observation reduced the coefficients for the 2 upper levels of mother's education. The observations with the largest values of Cook's *D* were as follows.

					Prenatal visits			Cook's
obs	mother race	mother age (+28)	parity	mother educ	obs	pred.	Ansc. Res.	distance
1733432	white	15	7	< hs dip	0	11.1	-4.817	0.006
1771199	white	1	3	hs dip	34	11.1	5.338	0.007
1394806	white	6	0	some col.	30	11.4	4.425	0.007
726794	white	-3	7	some col.	40	11.1	6.501	0.046

The plot of the Anscombe residuals vs predicted values shows a large number of extreme negative residuals but few large positive values. The plot of Anscombe residuals vs Cook's distance clearly highlights the influential role of observation #726794. As with the Poisson model, refitting the model with this observation excluded reduced the estimates of the effects of the 2 upper levels of mother's education by 3.1% to 4.5%.

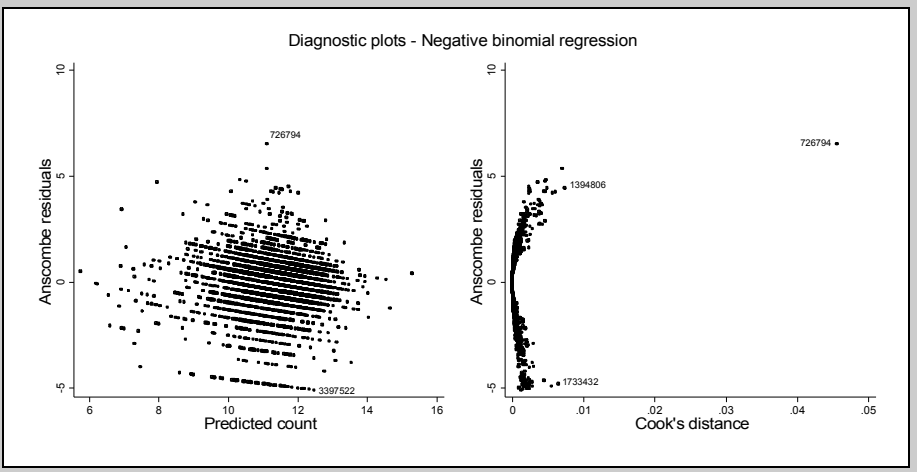

Fig. 18.5 Negative binomial diagnostic plots

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 prenatal visit data with the variance as a function of -meduc_c4-.

Example 18.6 Generalised negative binomial regression
data = bw5k

The Poisson model from Example 18.1 was refit using a generalised negative binomial model with -white-, -mage_28-, -tbo_1-, and -meduc_c4- as predictors of the mean number of prenatal visits and -meduc_c4- as the sole factor influencing the variance.

Number of obs = 5000
LR chi2(6) = 100.74
Prob > chi2 = 0.0000
Log likelihood = -13734.796 Pseudo R2 = 0.0037

					95% CI	
Variable	**Coef**	**SE**	**Z**	**P**	**Lower**	**Upper**
previs						
white	0.035	0.010	3.50	0.000	0.015	0.054
mage_28	0.004	0.001	4.03	0.000	0.002	0.006
tbo_1	-0.012	0.004	-3.50	0.000	-0.019	-0.005
meduc = hs dip.	0.056	0.016	3.42	0.001	0.024	0.089
meduc = some coll.	0.074	0.017	4.42	0.000	0.041	0.107
meduc = univ. deg.	0.072	0.017	4.29	0.000	0.039	0.105
constant	-1.285	0.015	-85.28	0.000	-1.314	-1.255
lnalpha						
meduc = hs dip.	-0.663	0.212	-3.13	0.002	-1.078	-0.248
meduc = some coll.	-1.648	0.403	-4.09	0.000	-2.437	-0.859
meduc = univ. deg.	-3.523	1.620	-2.17	0.030	-6.698	-0.347
constant	-2.823	0.132	-21.40	0.000	-3.081	-2.564

The results suggest that the effect of -meduc_c4- appears to be that levels 2–4 all have higher than expected number of visits, but the variance in response decreases as the level of education rises (α at the lowest education level was $e^{-2.823} = 0.059$) but at the highest level was only $e^{(-2.823-3.523)} = 0.002$. This fits well with the simple descriptive statistics (mean and variance) of -previs- by levels of -meduc_c4- which shows that the variance at the lowest level was 16.2 and at the top level was 12.2.

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
NB-2	5000	-13830.45	-13772.92	8	27561.84	27613.98
NB-H	5000	-13785.16	-13734.80	11	27491.59	27563.28

Both the AIC and the BIC favour the NB-H model, suggesting that the effect of rising levels of -meduc_c4- is to both increase the mean number of prenatal visits and to reduce the variation among individuals.

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 parity of mothers in a birth weight registry is modelled), then a zero-truncated model can be fit to the data.

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. In our example, with the outcome of interest being the number of prenatal visits, zero might be recorded if the mother truly had no prenatal visits or if she misunderstood the question and did not recognize what prenatal visits were (even if she had some). 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 (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.

Example 18.7 shows the application of a zero-inflated negative binomial model to the prenatal visit 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, 2011). For the prenatal visit, this would assume that all women who understood what prenatal visits were had at least 1, and that the values of zero were solely from women who did not understand the question.

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 prenatal visit data presented in Example 18.7. In contrast to the logit portion of the zero-inflated model, in a

Example 18.7 Zero-inflated negative binomial model
data = bw5k

A zero-inflated negative binomial model was fit to the prenatal visit data. The same set of predictors that was used in Example 18.1 was included in the model, but only mother's education was included in the logistic portion of the model.

Number of obs = 5000
Nonzero obs = 4953
Zero obs = 47
Inflation model = logit
Wald chi2(6) = 103.49
Log pseudolikelihood = -13560.19
Prob > chi2 = 0.0000

Variable	Coef	Robust SE	Z	P	95% CI	
Negative binomial portion						
white	0.032	0.010	3.31	0.001	0.013	0.050
mage_28	0.004	0.001	4.04	0.000	0.002	0.006
tbo_1	-0.012	0.003	-3.53	0.000	-0.019	-0.005
meduc = hs dip.	0.044	0.014	3.11	0.002	0.016	0.072
meduc = some coll.	0.063	0.015	4.21	0.000	0.034	0.093
meduc = univ. deg.	0.062	0.015	4.11	0.000	0.032	0.091
constant	-1.266	0.013	-97.12	0.000	-1.291	-1.240
Logistic portion						
meduc = hs dip.	-1.097	0.426	-2.58	0.010	-1.932	-0.263
meduc = some coll.	-1.049	0.446	-2.35	0.019	-1.923	-0.175
meduc = univ. deg.	-1.048	0.365	-2.87	0.004	-1.764	-0.333
constant	-3.909	0.232	-16.83	0.000	-4.365	-3.454
lnalpha	-4.701	0.212	-22.18	0.000	-5.116	-4.286
alpha	0.009	0.002			0.006	0.014

The Vuong statistic was 5.77, suggesting that the zero-inflated model was clearly superior to the regular negative binomial model (even though there were only 47 zeros in the dataset). The coefficients for -meduc_c4- in the negative binomial portion suggest that higher levels of education resulted in more prenatal visits and anything over the lowest level (<high school diploma) resulted in a reduced probability of zero visits. The estimate of α (0.009) was substantially less than that seen in Example 18.3, suggesting that some of the overdispersion was attributable to the excess number of zeros.

hurdle model, the coefficients in the logit model reflect how the log odds of a non-zero count is 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)?

Example 18.8 Logit–negative binomial hurdle model
data = bw5k

A logit–negative binomial hurdle model was fit to the same data used in Example 18.7, except that all predictors were included in both the negative binomial and logistic portions of the model.

Number of obs = 5000
Log pseudolikelihood = -13557.814 Wald chi2(6) = 14.92
Prob > chi2 = 0.0209

Variable	Coef	Robust SE	Z	P	95% CI	
Logit portion of the model						
white	0.270	0.319	0.85	0.396	-0.355	0.896
mage_28	0.025	0.031	0.82	0.410	-0.035	0.086
tbo_1	-0.144	0.102	-1.42	0.156	-0.344	0.055
meduc = hs dip.	0.965	0.433	2.23	0.026	0.117	1.813
meduc = some coll.	0.845	0.466	1.81	0.070	-0.069	1.758
meduc = univ. deg.	0.682	0.449	1.52	0.129	-0.197	1.561
constant	0.536	0.355	1.51	0.131	-0.160	1.233
Negative binomial (count) portion of the model						
white	0.032	0.010	3.31	0.001	0.013	0.050
mage_28	0.004	0.001	4.03	0.000	0.002	0.006
tbo_1	-0.012	0.003	-3.53	0.000	-0.019	-0.005
meduc = hs dip.	0.044	0.014	3.11	0.002	0.016	0.072
meduc = some coll.	0.063	0.015	4.21	0.000	0.034	0.093
meduc = univ. deg.	0.062	0.015	4.11	0.000	0.032	0.091
constant	-1.266	0.013	-97.13	0.000	-1.291	-1.240
/lnalpha	-4.703	0.212	-22.16	0.000	-5.118	-4.287

The coefficients for -meduc_c4- suggest that higher levels of eduction were more likely to result in a non-zero count and also to result in higher counts (negative binomial portion of the model). Other factors were significant in the negative binomial portion of the model (*ie* they affected the number of prenatal visits), but not in the logistic portion (no effect on whether any visits were reported).

Log-likelihoods were computed to compare the zero-inflated and hurdle models. The comparison slightly favoured the hurdle model, but the difference between the two models was small. **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	5000	-13610.66	-13558.91	15	27147.83	27245.59
hrdl	5000	.	-13557.81	15	27145.63	27243.39

18.7.3 Zero-truncated models

In some situations, zero counts are not possible. For example, when analysing length of hospital stay, zero days is not a possibility. 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 used to predict the number of babies born to mothers listed in the birth registry.

Example 18.9 Zero-truncated negative binomial model
data = bw5k

A zero-truncated negative binomial model was used to predict the number of babies born to mothers (-tbo-) listed in the birth registry. (This is based on the unreasonable assumption that the birth recorded in the dataset bw5k was the last for each woman registered.) Three predictors: mother's race (-white-), mother's education status (-meduc_c4-), and marital status (-mar-) were included.

Number of obs = 5000
LR chi2(5) = 171.36
Log likelihood = -7905.6451 Prob > chi2 = 0.0000

Variable	Coef.	Robust SE	Z	P	95% CI	
white	0.030	0.028	1.10	0.273	-0.024	0.084
meduc = hs dip.	-0.200	0.038	-5.30	0.000	-0.273	-0.126
meduc = some coll.	-0.250	0.041	-6.16	0.000	-0.329	-0.170
meduc = univ. deg.	-0.448	0.038	-11.66	0.000	-0.524	-0.373
married	0.275	0.031	8.96	0.000	0.215	0.335
constant	0.714	0.033	21.61	0.000	0.649	0.779
lnalpha	-1.902	0.128			-2.152	-1.652
alpha	0.149	0.019			0.116	0.192

There was no significant difference between whites and non-whites. As education level went up, the total birth order declined and married women had higher -tbo- than non-married women. α (0.149) was clearly greater than zero (likelihood ratio test of α had $P<0.001$) 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 -tbo- 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 AC, Trivedi PK. Regression Analysis of Count Data. Cambridge: Cambridge University Press; 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(4):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(3-4):137-48.

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

Hilbe JM. Negative Binomial Regression - 2nd Ed. Cambridge: Cambridge University Press; 2011.

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(2):e180.

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

Long JS, Freese J. Regression Models for Categorical Dependent Variables Using Stata. College Station, TX: Stata Press; 2006.

McNutt L-A, Wu C, Xue X, Hafner JP. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am J Epidemiol. 2003;157(10):940-3.

Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159(7):702-6.