10

Strategies for Analyzing Count Data

10.1 INTRODUCTION

According to legend, the mathematician and logician Leopold Kronecker believed that mathematics should be entirely based on whole numbers, noting, “God made the natural numbers; all else is the work of man.” Counts of discrete events in time and space are integers. These counts could be the number of bombs falling in a particular neighborhood, the number of coups d’état, the number of suicides, the number of fatalities owing to particular risk categories, such as traffic accidents, the frequency of strikes, the number of governmental sanctions, terrorist incidents, militarized disputes, trade negotiations, word counts in the speeches of presidential candidates, or any wide range of political and social phenomena that are counted.

Models of dependent variables that are counts of events are unsurprisingly called *event count models*. Event count models describe variables that map only to the nonnegative integers: $Y \in \{0, 1, 2, \ldots\}$. While grouped binary or categorical data can be thought of as counts, such data sets are generally not analyzed with count models.[1] Count data have two important characteristics: they are discrete and bounded from below.

Using ordinary least squares to directly model integer counts can lead to problems for the same reason it does with binary data. The variance of a count increases with the mean (there is more error around larger values), implying inherent heteroskedasticity. More worryingly, OLS will generate predictions that are impossible to observe in nature: negative counts and non-integer values. Some try to salvage a least squares approach by taking the logarithm of the dependent variable. This strategy, however, requires a decision about what to do with the zero counts, since $\log(0) = -\infty$. One option is to simply discard

[1] A major exception is the analysis of vote totals across candidates or parties. See, for example, Mebane and Sekhon (2004).

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

the zero-count observations and instead only model the positive counts. This has numerous drawbacks, including the potential to discard a large proportion of the observed data. A second approach is to add some constant to the outcome before taking logs. Aside from its arbitrariness, this approach has its drawbacks, including exacerbating problems with nonconstant variance and complicating interpretation. In this chapter we introduce a series of models that are explicitly designed to model event count data as what they are – integer counts.

10.2 THE POISSON DISTRIBUTION

Imagine a situation in which we are interested in the occurrence of concrete events, but we are unable to observe specific instances (or non-instances). Rather, we only observe the number of events occurring within some observational window. This window could be spatial (number of bombs that fell in a quarter square kilometer) or temporal (number of children born in January). We denote the size of this window, sometimes called the *exposure interval*, as h. The average number of events in any particular exposure interval is the *arrival rate*, denoted λ. Thus the probability of an event occurring in the interval $(t, t+h]$ is λh and, conversely, the probability of no event occurring in this interval $(t, t+h]$ is $1 - \lambda h$. When events occur independently – the occurrence of one event does not influence the probability that another will occur – and with a constant arrival rate we say they follow a *Poisson process*.

In case you were wondering ... 10.1 Poisson distribution

Let $Y \in \{0, 1, 2, \ldots\}$. We say that Y_i follows a *Poisson distribution* with parameter vector $\boldsymbol{\theta} = (\lambda, h)$:

$$Y \sim f_P(y; \lambda, h)$$
$$\Pr(Y = y) = \frac{\exp(-\lambda h)(\lambda h)^y}{y!},$$

with $\lambda > 0, h > 0$, and $\mathrm{E}[Y] = \mathrm{var}(Y) = h\lambda$.

If all the observational intervals are of the same length then we can standardize $h = 1$ and the probability mass function reduces to:

$$\Pr(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!},$$

with $\mathrm{E}[Y] = \mathrm{var}(Y) = \lambda$.

Siméon Denis Poisson (1781–1840) was a French mathematician, famous for correcting Laplace's equations for celestial mechanics. He discovered a probability distribution for discrete events, occurring in fixed intervals. He derived his invention in terms of the "law of rare

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

events" in which he described the behavior of a binomial distribution as $n \to \infty$ and $p \to 0$. Poisson used the distribution in an analysis of "criminal and civil matters," an early study of public policy (Poisson, 1837).

Given the probability model, it is straightforward to incorporate covariates $\mathbf{x}_i$ into a model for the mean. Because $\mathrm{E}[Y] > 0$, we need a link function that maps onto positive values. The exponential is the most commonly employed transformation for achieving this. Thus,

$$\mathrm{E}[Y_i] \equiv h\lambda_i = he^{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}.$$

From this expression it is easy to see that the natural log is the (canonical) link function for the Poisson GLM. We can also see why the Poisson model is sometimes referred to as "log-linear." The log of the mean is linear in the covariates and regression parameters. Setting $h = 1$ for simplicity and incorporating this expression into the Poisson mass function yields:

$$\Pr(Y_i = y|\mathbf{x}_i) = \frac{\exp[-\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})][\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})]^y}{y!}.$$

The likelihood is straightforward to derive:

$$\mathcal{L}(\beta|\mathbf{X},\mathbf{y}) = \prod_{i=1}^{n} \frac{\exp[-\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})][\exp(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}}{y_i!}$$

$$\log\mathcal{L} = \sum_{i=1}^{n}\left[-\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}) + y_i\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} - \log(y_i!)\right]$$

The last term, $-\log(y_i!)$, is ignorable. This log-likelihood is regular and well-behaved, so the standard tools apply. We can also derive the score equation for the Poisson model:

$$\begin{aligned}\frac{\partial \log\mathcal{L}}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n} y_i\mathbf{x}_i^\mathsf{T} - \mathbf{x}_i^\mathsf{T}\exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n}(y_i - \exp(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}))\mathbf{x}_i = \mathbf{0}. \qquad (10.1)\end{aligned}$$

Equation 10.1 will appear below because we can also view it from an estimating equation or quasi-likelihood approach.

10.2.1 An Example: Means and Mediation

In a classic early application of the Poisson distribution, Bortkiewicz (1898) found that Prussian Army deaths from horse kicks were distributed as a Poisson

TABLE 10.1 *OLS and Poisson regression of the number of times a country served as a mediator in an international conflict between 1950 and 1990. The last two models account for unequal "exposure" across countries by including an offset term.*

	OLS			Poisson		
		log $(y_i + 0.01)$	log $(y_i + 10)$		Exposure Offset	Exposure
GDPpc 1990	0.33	−0.02	0.01	0.07	0.06	0.06
	(0.16)	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)
UNSC member	34.59	3.98	1.02	2.35	2.18	2.17
	(4.19)	(1.39)	(0.14)	(0.11)	(0.10)	(0.11)
log duration						1.11
						(0.19)
Intercept	0.78	−1.85	2.43	0.42	−3.03	−3.42
	(1.04)	(0.35)	(0.04)	(0.08)	(0.08)	(0.66)
n	146	146	146	146	146	146
$\log \mathcal{L}$	−526	−366	−35	−585	−563	−563
AIC	1,059	737	75	1,176	1,132	1,134
BIC	1,071	749	82	1,185	1,141	1,146

process. Bercovitch and Schneider (2000) is a more recent example of count data in security studies. They study mediation in the international system, developing a model of the factors that lead countries to act as mediators in international disputes. We use their data to explore a Poisson model, simplifying their argument considerably. The basic argument is that rich and powerful countries are more likely to be requested as mediators. Their data contain 146 countries, observed over the period from 1950 to 1990. The dependent variable is the number of times that a country received a mandate to mediate in an international conflict. Mediation count is taken to be a log-linear function of whether or not the country is a member of the United Nations Security Council (UNSC) and its per capita Gross Domestic Product ($US '000) in 1990.

The first three columns of Table 10.1 display standard BUTON output for the OLS regressions. The first is an OLS on the unmolested dependent variable. The second and third add constants – 0.01 and 10, respectively – to the outcome before taking logarithms. The results in the table show how adding arbitrary constants can alter the model fit and interpretation. For example, the standard OLS says that a UNSC member is expected to have mediated 35 more disputes than a nonmember, whereas the log-transformed model implies that a UNSC member will have mediated exp(3.98) $=$ 49 more disputes. This is a serious discrepancy that is not reflected in any uncertainty estimates. Comparing the second and third models, we see that indicators of model fit shift dramatically simply with the addition of a constant; this is unsurprising; adding a fixed

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

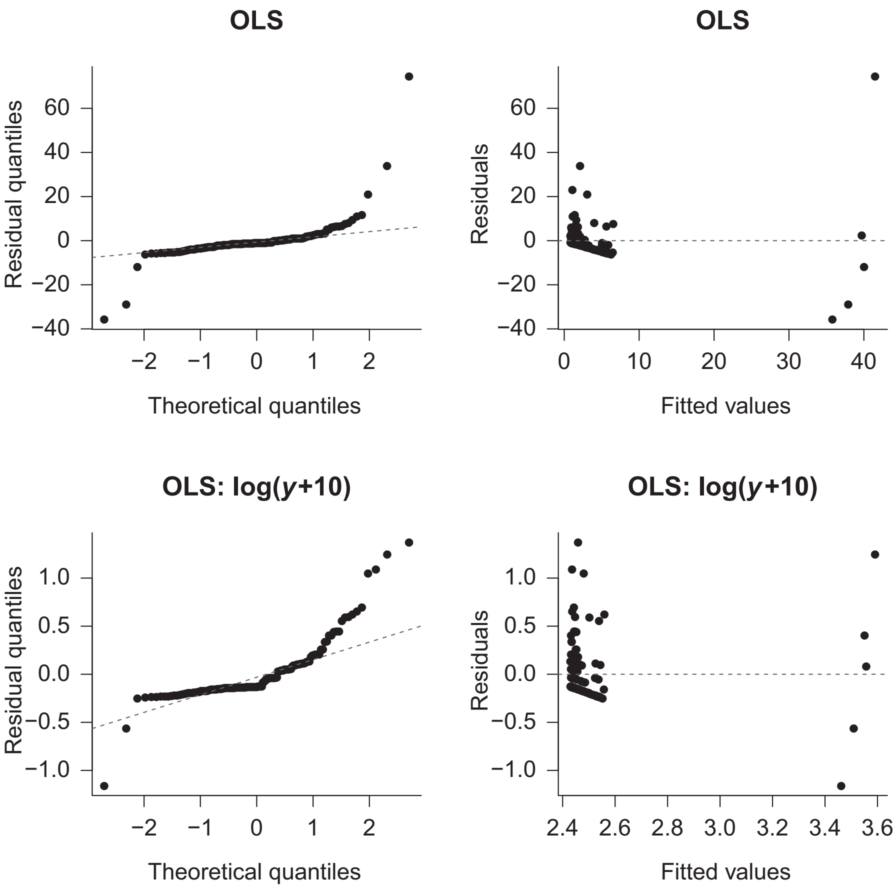

FIGURE 10.1 Diagnostic plots for the standard OLS regression and the OLS on $\log(y_i + 10)$ reported in Table 10.1. There is clear evidence of non-normality of heteroskedasticity in the residuals, becoming worse as a constant is added to the outcome variable and then transformed.

amount is easily absorbed by the intercept and has the effect of artificially reducing residual variance.[2]

Figure 10.1 displays residual quantiles and fitted-residual plots for the standard OLS model as well as the OLS on $\log(y_i + 10)$. Both clearly indicate that the OLS residuals are non-normal (with fat tails) and severely heteroskedastic, as we would expect from count data. Adding a constant and transforming the data has the drawback of exacerbating both these problems.

[2] Technically the log-likelihoods and related quantities for the second and third models are not comparable since they are fit to "different" outcome variables. This distinction is rarely appreciated in applied work; the constant added to **y** is often not explicitly stated.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

The last three columns of Table 10.1 present results from a series of Poisson regressions. The first of these fits a model with the same linear predictor as the OLS regression. Based on the AIC and BIC, the OLS model might be preferable to the Poisson. But the general expression for the Poisson distribution takes into account the extent of "exposure" for each subject. In the mediation example, countries that have been around longer have had more opportunities to be requested as mediators. If observations have different exposure windows, then their expected counts should differ proportionally. Differential exposure needs to be incorporated into the model. There are two approaches. The first is known as an *offset*: include the size of the exposure window, h_i, in the regression but constrain its coefficient to be 1.0:

$$\lambda_i = \exp[\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \log(h_i)].$$

The second approach simply includes the exposure variable (in logs) in the regression equation and allows a coefficient to be determined empirically.

The last two columns of Table 10.1 reestimate the Bercovitch and Schneider example, taking account of each country's "exposure" with a variable indexing the number of years a country has been a member of the international system. The fifth column includes this log duration variable as an offset, while the model in the sixth estimates a parameter. Accounting for exposure improves model fit, based on the likelihood ratio and the information criteria. Comparing the models in the final two columns, we see little reason to estimate a parameter for the exposure variable. The log-likelihoods for the two models are identical. A Wald test on the log duration coefficient from the third model gives a p-value of $(1.11 - 1)/0.19 = 0.6$. In general, accounting for unequal exposure is important for count models. Failing to do so will tend to inflate the putative impact of covariates.

10.2.2 Interpretation

As with many of the models we have seen in this volume, the relationship between a covariate and the response is nonlinear and depends on the value of other covariates in the model. Suppose we have $k - 1$ covariates in the model. The marginal effect of particular regressor, X_j, on $\mathrm{E}[Y_i]$ is

$$\frac{\partial \lambda_i}{\partial x_{ij}} = \beta_j \exp[\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}]. \tag{10.2}$$

In such situations we can follow our usual strategy: construct meaningful scenarios; sample from the limiting distribution of the parameters; and combine the two to generate predicted values, along with our estimation uncertainty.

In Figure 10.2 we do just that, presenting both the predicted and expected number of mediations as a function of per capita GDP and membership on the UNSC. For each of these scenarios, the black line represents the expected

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 10.1 ***Fitting Poisson regression models with offsets***

```
bs<-read.csv("mediation.csv",header=T)
# medteam: number of mediations, including team mediations
# council: 1=member of UN Security Council
# gdp90: gdp per capita in constant dollars (needs to be rescaled)
bs$pop90<-bs$pop90/100000
bs$gdp90<-bs$gdp90/1000
bs.out<-glm(medteam~ gdp90 + council, data=bs,family="poisson")
offset.out<-glm(medteam ~ gdp90 + council + offset(log(duration)),
    data=bs,family="poisson")
#bs.out<-glm(medteam ~ gdp90 + council, offset = log(duration),
#   data=bs,family="poisson") #equivalent
duration.out<-glm(medteam ~ gdp90 + council + log(duration),
    data=bs,family="poisson")
```

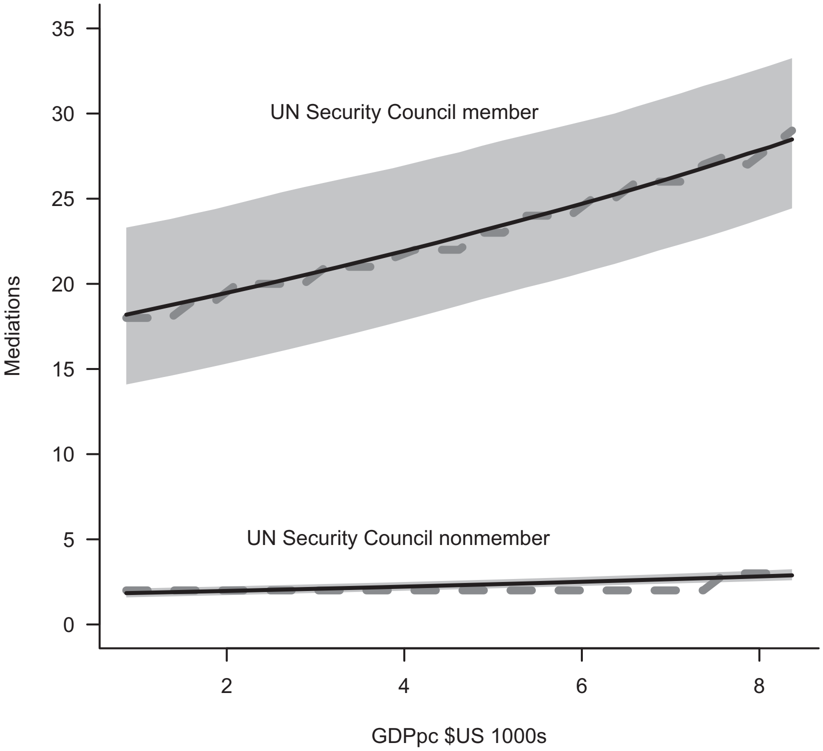

FIGURE 10.2 The expected and predicted number of international mediations as a function of per capita GDP and UNSC membership. Black lines are expected values and gray lines are predicted counts. Gray regions are 95% confidence bands around the expected value.

value ($\hat{\lambda}_i$), while the red line represents the predicted value – an integer. Gray bands are 95% confidence intervals around the expected value. In constructing these scenarios we vary GDP per capita from its 20th to 80th sample percentile values. We set duration in the international system at its sample median value for UNSC nonmembers. We set it to $\log 40 \approx 3.69$ for UNSC members, its minimum value for that group of countries. The plot shows that richer countries are more likely to be mediators and that UNSC members are about eight times more likely to be asked across income levels.

The log-linear construction of the Poisson model lends itself to other interpretive strategies that you may encounter in your reading. Since $\log \lambda_i = \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}$ we know that β_j is the change in $\log \lambda_i$ for a change in x_{ij}. The quantity $\exp(\hat{\beta}_j)$ therefore represents a multiplicative change in $\hat{\lambda}_i$, as shown in Equation 10.2. In the mediation example, $\exp(2.18) \approx 8.8$, so countries on the UNSC serve as mediators about nine times more frequently than countries not on the council. If a particular covariate is included on the log scale, then the coefficient of that covariate in a Poisson model has a direct interpretation in terms of elasticities, or a percent change in the outcome for a 1% change in the covariate. For example, if we included per capita GDP in logarithms (rather than thousands of dollars), then we obtain $\hat{\beta}_{GDP} = 0.20$; a ten-percent increase in per capita GDP is associated with a two-percent increase in the arrival rate of mediation requests.

10.3 DISPERSION

Intuitively we expect greater variation around the mean when there is a large number of expected events. We therefore expect count data to be inherently heteroskedastic. The Poisson distribution captures this fact; the variance increases with the mean, one-for-one. But this one-to-one relationship is quite restrictive and often violated in real-world data. The (very frequent) situation in which the variance of the residuals is larger than the mean is known as *over-dispersion*. *Under-dispersion* occurs when the variance is too small; it is much less commonly observed in social science data.

$$\begin{aligned}
\text{Poisson Assumption} &\leftrightarrow \mathrm{E}[Y] = \mathrm{var}(Y)\\
\text{Over-dispersion} &\leftrightarrow \mathrm{E}[Y] < \mathrm{var}(Y)\\
\text{Under-dispersion} &\leftrightarrow \mathrm{E}[Y] > \mathrm{var}(Y) \quad \leftrightarrow \quad 0 < \sigma < 1
\end{aligned}$$

Over-dispersion can arise for several reasons, all of which have consequences for model building and interpretation. At the simplest level, over-dispersion may simply be the result of a more variable process than the Poisson distribution is capable of capturing, perhaps due to heterogeneity in the underlying population. If this is the case then the model for the mean may still be adequate, but the variance estimates will be too small, perhaps wildly so. But

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

over-dispersion may also arise for more complicated reasons that have both substantive and modeling implications. For example, over-dispersion may arise if there are an excess of zeros in the data or, most importantly, if events are positively correlated (previous events increase the rate of subsequent events), violating a basic assumption of the Poisson process. These challenges imply that the model for the mean is no longer adequate, and we should expect problems of inconsistency and potentially erroneous inference. In short, over-dispersion should be viewed as a symptom that needs to be investigated. The results of this investigation usually turn up substantively interesting aspects of the data-generating process.

10.3.1 Diagnosing Over-Dispersion

We can use the MLE to derive one heuristic for over-dispersion. First, note that

$$\begin{aligned}\log \mathcal{L} &= \sum_{i=1}^{n} -\lambda + y_i \log \lambda \\ &= -n\lambda + \log \lambda \sum y_i, \\ \frac{\partial}{\partial \lambda} \log \mathcal{L} &= -n + \lambda^{-1} \sum y_i = 0 \\ &\Rightarrow \hat{\lambda} = \bar{y}.\end{aligned}$$

Since the sample mean is the MLE for λ, and equi-dispersion implies that the mean and variance are equal, we can compare the sample mean and the sample variance. In the Bortkiewicz data on Prussian deaths by horse kick, we obtain a mean of 0.61 and a variance of 0.61. In the international mediation example, the sample mean is 3.5 and the variance is 126.

Graphical Methods

One visualization tool, the *Poissonness plot* (Hoaglin, 1980), examines the distribution of the dependent variable relative to theoretically expected values under a Poisson distribution.[3] The horizontal axis is the number of events, denoted y. The vertical axis is the so-called *metameter*. For the Poisson distribution the metameter for count value y is given as $\log y! n_y - \log n$, where n_y is the observed frequency of y events. If the data conform to the proposed distribution, the points should line up, similar to a quantile-quantile (Q-Q) plot for continuous distributions. Moreover, a regression of y on the metameter for y should have slope $\log(\lambda)$ and intercept $-\lambda$ under the Poisson distribution.

[3] Hoaglin and Tukey (1985) extend the Poissonness plot to the negative binomial and binomial. Each distribution has its own metameter calculation.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

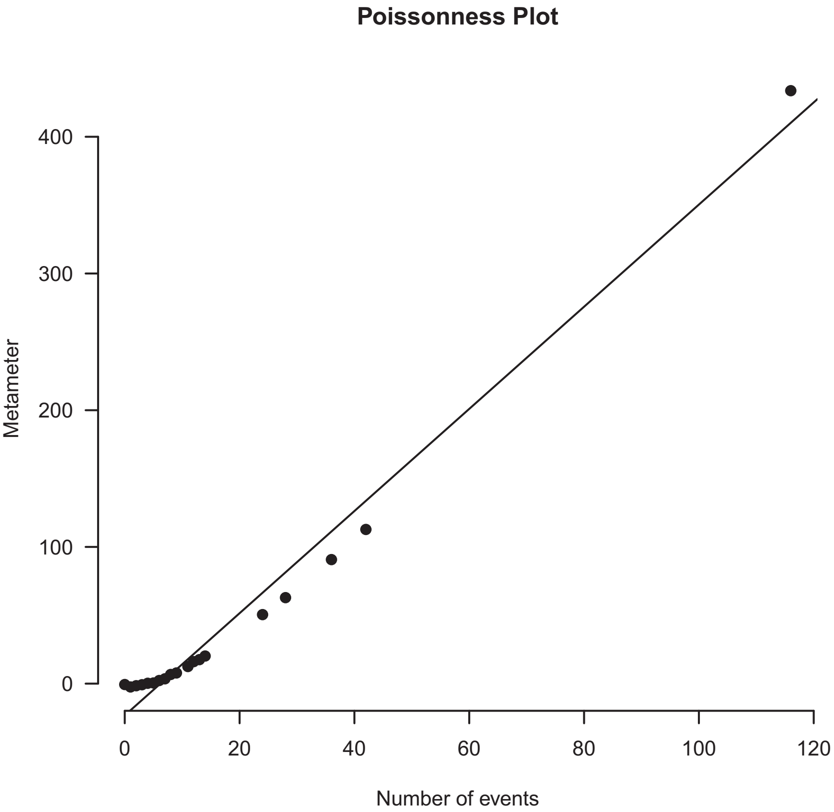

FIGURE 10.3 A Poissonness plot for international mediations data from Bercovitch and Schneider (2000). If the data follow a Poisson distribution, then the points should line up, and the slope of the regression should be $\log \lambda$, whereas the intercept should be $-\lambda$. With these data $\hat{\lambda} = 3.5$. The slope of the regression line is 3.7, and the intercept is −23.

Figure 10.3 displays a *Poissonness* plot for the international mediations data. The points clearly fail to line up on the regression line.[4] If these data followed a Poisson distribution we should observe an intercept of −3.5 and a regression slope of $\log 3.5 = 1.25$. Instead, we observe a slope of about 3.7 and an intercept of −23. Based on the *Poissonness* plot, the mediation data do not conform to the Poisson distribution. But a weakness of the plot is its inability to tell us anything about over-dispersion *per se* or about a particular model fit to those data.

[4] The outlier is, unsurprisingly, the United States. Excluding the United States implies a $\hat{\lambda} = 2.75$. The corresponding *Poissonness* plot yields an intercept of −12.5 and a slope of 2.7.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

To address both these concerns, Kleiber and Zeileis (2016) advocate persuasively for a *hanging rootogram*, due to Tukey (1977). In a rootogram the horizontal axis is again counts of events. The vertical axis is the square root of the frequency, where the square root transformation ensures that large values do not dominate the plot. Let $\mathrm{E}[n_y]$ be the expected frequency of event count value y under the proposed model. For a Poisson model, $\mathrm{E}[n_y] = \sum_{i=1}^{n} f_P(y; \hat{\lambda}_i)$. The vertical bars are drawn from $\sqrt{\mathrm{E}[n_y]}$ to $(\sqrt{\mathrm{E}[n_y]} - \sqrt{n_y})$. In other words, the vertical boxes hang down from the expected frequency and represent how much the observed and expected frequencies differ at various values of the outcome variable. The zero line presents a convenient reference. A bar that fails to reach 0 means the model is overpredicting counts at that value $(\mathrm{E}[n_y] - n_y > 0)$. A bar that crosses 0 implies underprediction at a particular value of y.

Figure 10.4 displays two rootograms. The plot on the left refers to the Poisson regression (with exposure offset) from Table 10.1. We can see that the Poisson model severely underpredicts 0 while also underpredicting large counts. It overpredicts counts near the sample mean of 3.5. Wave-like patterns and underpredictions of 0s are consistent with over-dispersion. But rootograms can be used to describe model fit more generally, whether in- or out-of-sample.

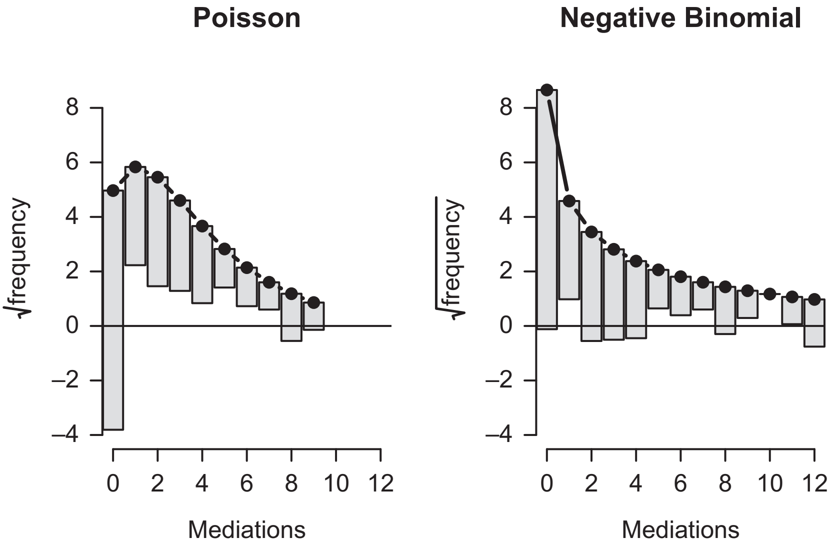

FIGURE 10.4 A hanging rootogram plotting expected versus actual counts. The curve represents the square root of the expected frequency of mediations under the specified model, while the vertical bars are drawn from the expected frequency to the observed frequency (both in square roots). The display on the left is from the Poisson model (with offset) from Table 10.1; the display on the right is from the negative binomial model in Table 10.2.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

Formal Tests

There are several tests for over-dispersion. Most of them are constructed based on the assumptions that (1) the model for the mean is correct and (2) we can think of over-dispersion as taking the form

$$\text{var}(Y_i|\mathbf{x}_i) = h_i\lambda_i + \gamma w(h_i\lambda_i),$$

where $w()$ is some function. The most frequently used alternatives are $w(h_i\lambda_i) = h_i\lambda_i$ and $w(h_i\lambda_i) = (h_i\lambda_i)^2$. Cameron and Trivedi (1990) develop one regression-based test based on this logic. Define

$$\hat{e}_i = \frac{(y_i - h_i\hat{\lambda}_i)^2 - y_i}{h_i\hat{\lambda}_i}, \tag{10.3}$$

then we can estimate the OLS regression (omitting the intercept) $\hat{e}_i = \gamma \frac{w(h_i\hat{\lambda}_i)}{h_i\hat{\lambda}_i} + \varepsilon_i$. For over-dispersion we are interested in testing the *one-tailed* hypothesis that $\gamma > 0$ against the null of equi-dispersion, $\gamma = 0$.[5] Applied to the Poisson model (with offset) from Table 10.1 and assuming $w(x) = x$, we obtain $\hat{\gamma} = 8.9$ with a one-sided p-value of 0.005, consistent with over-dispersion.

Gelman and Hill (2007) describe a version of the Pearson χ^2 test that takes advantage of the fact that the standard deviation of the Poisson is equal to the square root of the mean. Define the standardized (or Pearson) residuals from a Poisson model as

$$z_i = \frac{y_i - h_i\hat{\lambda}_i}{\sqrt{h_i\hat{\lambda}_i}}.$$

If the Poisson assumption is correct, then each z_i is a standard normal random variable. This implies that $\sum z_i^2 \sim \chi^2_{n-k}$, which has an expected value of $n - k$ under the null hypothesis of equi-dispersion. In the case of over-dispersion, however, the mean of $z_i > 0$ and the corresponding sum of squares will be larger than $n - k$. In the Poisson example above (with offset), the sum of the standardized residuals is a hefty 1,448, rather than the expected 143. This yields a p-value ≈ 0, indicating over-dispersion.

10.4 MODELING OVER-DISPERSION

We found over-dispersion in the international mediation data. So what to do? That depends on our confidence in the model for the mean. If we believe that the model for the mean is correct and our conditional independence assumptions hold, then ignoring over-dispersion will not bias coefficient estimates. Over-dispersion does, however, induce bias and inconsistency into

[5] A joint test for either over-dispersion or under-dispersion corresponds to a two-tailed test against the null that $\gamma = 0$.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

our estimated standard errors. We need to allow more flexibility in the model. Two immediate approaches present themselves: quasi-likelihood and the negative binomial model. These approaches often give similar results when evaluated on the scale of the outcome variable, but not always, as we shall see.

10.4.1 Quasipoisson

One way to proceed is with a quasi-likelihood approach in which we specify a model for the mean and a variance function rather than a full probability model (see Section 7.4). If our model for the mean is $\lambda_i = \exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})$ and our variance function is $\phi V(\lambda_i) = \phi \lambda_i$, then the set of quasiscore equations reduces to Equation 10.1. This implies that the solution to the Poisson score equations and the solution to the quasiscore equations where the mean is modeled as $\exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})$ is the same. In other words, the Poisson MLE for $\hat{\boldsymbol{\beta}}$ are the same as the quasipoisson estimates. As a result the quasipoisson approach will not alter point estimates of predicted outcomes.

Where quaispoisson and the full Poisson GLM differ is in the standard errors. In the quasipoisson setup we estimate the dispersion parameter, ϕ, using Equation 7.3, which is exactly the formula for the Pearson residuals we used earlier divided by $n - k$. Quasipoisson standard errors are then calculated from Equation 7.2. With $\phi > 1$ the quasipoisson standard errors will be larger than those from the Poisson GLM. While we do not have access to the maximized likelihood or information criteria from the quasipoisson model, we can still generate predicted counts and use simulation techniques to generate uncertainty estimates.

10.4.2 The Negative Binomial Model

Another approach for addressing over-dispersion is to specify a more-flexible distribution as the basis for deriving the log-likelihood. The most commonly used model is the *negative binomial*. One way to derive this model is to view λ_i as a random variable. But, conditional on λ_i, Y_i is still Poisson. In other words, we imagine that there is heterogeneity in the rate parameter that the Poisson model is too restrictive to capture. To fix the model we must pick a distribution for λ_i. Since we want a distribution that yields strictly positive numbers and has the flexibility to accommodate over-dispersion, we typically use the one-parameter *gamma* distribution.[6]

[6] The choice of the gamma distribution is somewhat arbitrary. Its primary convenience is that it yields a closed-form expression for the marginal distribution of Y_i. In Bayesian terminology, the gamma distribution is the conjugate prior for the Poisson.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

In case you were wondering ... 10.2 The gamma function and distribution

The *gamma function*, denoted $\Gamma(\cdot)$, is a generalization of the factorial to non-integer (and complex) arguments. $\Gamma(x+1) = x!$ for positive integer x.

The *gamma distribution* has several parameterizations. We use the "shape/rate" version. Let $Y \in (0, \infty)$. We say that Y follows a *gamma distribution* with parameter vector $\boldsymbol{\theta} = (a, b)$ if

$$Y \sim f_\Gamma(y; a, b) = \frac{b^a y^{a-1}}{\Gamma(a)} \exp(-by),$$

with *shape* parameter $a > 0$ and *rate* or *inverse scale* parameter $b > 0$. $\mathrm{E}[Y] = \frac{a}{b}$ and $\mathrm{var}(Y) = \frac{a}{b^2}$.

Fixing $a = b = \alpha$ yields the *one-parameter gamma* distribution, $f_\Gamma(y; \alpha)$, with expected value of 1 and a variance of α^{-1}.

The negative binomial probability model can be built in parts

$$\begin{aligned} Y_i \mid \lambda_i &\sim f_P(\lambda_i), \\ \lambda_i &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta} + u_i), \\ &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \exp(u_i), \end{aligned}$$

where u_i is an error term in the expression for the Poisson mean, λ_i. If we let $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $\nu_i = \exp(u_i)$ we can complete the model:

$$\begin{aligned} \lambda_i &= \mu_i \nu_i, \\ \nu_i &\sim f_\Gamma(\alpha). \end{aligned}$$

The ν_i are now unit-mean multiplicative error terms for the Poisson mean. As before, $\mathrm{E}[Y_i] = \lambda_i$, but now $\mathrm{E}[\lambda_i] = \mu_i$, implying that $\mathrm{E}[Y_i] = \mu_i$. The μ_i are typically modeled in a log-linear fashion as $\exp(\mathbf{x}_i^\top \boldsymbol{\beta})$. Integrating over ν_i gives us the marginal distribution for Y_i, which is a negative binomial distribution.

In case you were wondering ... 10.3 The negative binomial distribution

Let $Y \in \{0, 1, 2, \ldots\}$. We say that Y follows a *negative binomial distribution* with parameter vector $\boldsymbol{\theta} = (\mu, \alpha)$, where $\mu > 0$ and

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

$\alpha > 0$, if

$$\begin{gathered} Y \sim f_{Nb}(y; \mu, \alpha) \\ \Updownarrow \\ \Pr(Y = y \mid \mu, \alpha) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{\alpha}{\alpha+\mu}\right)^{\alpha} \left(\frac{\mu}{\alpha+\mu}\right)^{y}. \qquad (10.4) \end{gathered}$$

The negative binomial has $\text{E}[Y] = \mu$ and $\text{var}(Y) = \mu(1 + \alpha^{-1}\mu) = \mu + \alpha^{-1}\mu^2$.

Via the parameter α, the negative binomial allows the variance to be greater than the mean. Cameron and Trivedi (2013) refer to the version just described as the NB2 model, with the 2 referring to the variance's quadratic dependence on the mean. They develop other versions as well. For example, if we substitute $\alpha'\mu_i$ for α in Equation 10.4, then we arrive at the NB1 model, with $\text{var}(Y) = \mu(1 + 1/\alpha')$. More generally, substituting $\alpha\mu^{p-2}$ for α in Equation 10.4 yields the NBp model, with $\text{var}(Y) = \mu(1 + \alpha^{-1}\mu^{p-1})$. Greene (2008) provides a general expression of the NBp log-likelihood:

$$\begin{aligned} \mu_i &= \exp(\mathbf{x}_i \top \boldsymbol{\beta}), \\ r_i &= \frac{\alpha}{\alpha + \mu_i}, \\ q_i &= \alpha\mu_i^{2-p}, \\ \log \mathcal{L}(\boldsymbol{\beta}, \alpha | \mathbf{X}, \mathbf{y}, p) &= \sum_{i=1}^{n} \log \Gamma(y_i + q_i) - \log \Gamma(q_i) - \log \Gamma(y_i + 1) \\ &\quad + q_i + \log r_i + y_i \log(1 - r_i). \qquad (10.5) \end{aligned}$$

The NB2 model is by far the most commonly used. Differences between NB1 and NB2 appear to be small in most applications. As these models are not nested versions of one another, choosing between them is best accomplished using information criteria and out-of-sample fit heuristics. All these versions of the negative binomial estimate the k regression parameters along with α, which governs the mean-variance relationship.

While both the quasipoisson and negative binomial models allow for over-dispersion, they approach it differently. The quasipoisson model directly estimates a dispersion parameter from the data and uses it to adjust standard errors while retaining the Poisson estimating equation for $\boldsymbol{\beta}$. The negative binomial model retains the $\phi = 1$ assumption of the Poisson and instead uses α to acount for over-dispersion; α is *not* a dispersion parameter in the exponential family sense. This fact is directly visible in the $\mathcal{R}$ summary output for the negative binomial glm: `Dispersion parameter for Negative Binomial(0.3326) family taken to be 1`, where 0.3326 is the estimate for α^{-1}. As we can see from Equation 10.5, the negative binomial likelihood differs from the Poisson even when the linear predictor terms are identical. Model estimates and implications can differ between the two.

Within the negative binomial class of models, the α parameter is a weight in a polynomial function of the mean. As the underlying polynomial changes, α also changes. As a result the α from an NB2 is not directly comparable to α' from an NB1, etc., notwithstanding the fact that most texts use the same notation across model parameterizations. However, all NBp models collapse to a simple Poisson as $\alpha \to \infty$. Because the Poisson is a limiting case of the negative binomial we can implement another over-dispersion test by constructing a likelihood ratio between a Poisson and negative binomial model.[7]

10.4.3 Mediations

We return to the international mediations example to examine how the quasipoisson and negative binomial models perform. Table 10.2 displays results

TABLE 10.2 *Quasipoisson and negative binomial regression of the number of times a country served as a mediator in an international conflict between 1950 and 1990. All models include offsets to account for unequal "exposure" across countries.*

	Quasipoisson	NB2	NB1
1990 GDPpc	0.06	0.01	0.01
	(0.03)	(0.03)	(0.02)
UNSC	2.18	2.51	1.86
	(0.33)	(0.82)	(0.35)
Intercept	−3.03	−2.76	−2.51
	(0.25)	(0.22)	(0.24)
ϕ	10.12		
α^{-1}		0.33	
		(0.06)	
α'			11.42
			(2.85)
n	146	146	146
$\log \mathcal{L}$		−280	−280
AIC		568	567
BIC		578	579

[7] Because we obtain a Poisson distribution at the boundary of the negative binomial parameter space the likelihood ratio between a Poisson and negative binomial model has a nonstandard distribution, with half of its mass at 0 and half as a χ^2_1. As a result, the critical value for a test at the a level is the χ^2_i value associated with a $2a$ test. For example, the critical value for a 95% test is 2.71 rather than the 3.84 normally associated with a χ^2_1.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

for a quasipoisson, NB2, and NB1 models. All three models include an offset for log duration to account for unequal exposure windows.

The quasipoisson returns exactly the $\hat{\boldsymbol{\beta}}$ from the third model in Table 10.1. What has fundamentally changed are the standard errors, which are substantially larger, reflecting more uncertainty, owing to the over-dispersion that is no longer being ignored. In general, ignoring over-dispersion will lead to dramatically underestimated standard errors. Note that the estimated dispersion parameter for the quasipoisson, $\hat{\phi}$, is exactly the sum of squared Pearson residuals calculated for the over-dispersion test divided by the residual degrees of freedom, 1,477/143.

Based on the information criteria, both of the negative binomial models represent large improvements over the OLS and Poisson models in Table 10.1. The right panel of Figure 10.4 displays the rootogram for the NB2 model, which clearly fits the data better than the Poisson alternative. It is far more accurate in its predictions of 0s and does not show the wave-like pattern of over- and underprediction across outcome values of the Poisson specification. The negative binomial models are preferable to the Poisson or quasipoisson. But, based on the in-sample fit statistics, the two negative binomial models are nearly identical, a common occurrence.

Consistent with the over-dispersion in the data the negative binomial models return substantial estimates for α. In $\mathcal{R}$ the `glm.nb` function fits the NB2 model and uses the parameterization and notation in Venables and Ripley (2002). In their parameterization of the negative binomial distribution they use "θ," which is equivalent to α^{-1} in our notation.[8] We can recover an approximate standard error for $\hat{\alpha}$ from `glm.nb` output by calculating $\hat{\sigma}_\theta/\hat{\theta}^2$. In the NB2 case we therefore recover $\hat{\alpha} = 3, \sigma_{\hat{\alpha}} = 0.52$.

In the mediation example the negative binomial models give substantively different results from the quasipoisson; per capita GDP is not a strong predictor of mediation demand once we account for over-dispersion. Differences between the quasipoisson and negative binomial sometimes arise due to differences in the weights attached to large counts in the fitting of the model (Ver Hoef and Boveng, 2007). But in this example the negative binomial model fits the data substantially better than the (quasi)Poisson. This divergence can happen when the process generating over-dispersion is more complicated than simple heterogeneity in the underlying population. For example, the negative binomial distribution and over-dispersion can result from positive contagion across events (Gurland, 1959; Winkelmann, 1995), something that appears to be at work here.

Interpretation of the negative binomial models follows our usual procedures. For example, take China, a member of the UNSC but relatively poor during this

[8] The `dbinom` (and other associated functions) in $\mathcal{R}$ can take several parameterizations of the negative binomial distribution. In our notation the `mu` argument to `dbinom` corresponds to μ and `size` corresponds to α.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

time. Had China not been on the UNSC, the model predicts that it would have been requested as a mediator between two and three times ($\hat{\lambda} = 2.54$) with a standard error of 0.49). But China as a UNSC member is expected to have been asked to mediate 31 times. Doubling China's per capita GDP has no effect on these predictions.

10.4.4 Under-Dispersion

While under-dispersion is less common than over-dispersion, it does arise. Under-dispersed count data can be thought of as having some kind of negative contagion. For example, the neighboring counties to a toxic waste site will be less likely to create their own, for a variety of reasons.

Many of the same tools used for diagnosing over-dispersion can be repurposed for under-dispersion. We can examine the mean of the sample relative to its variance. We can use the Cameron-Trivedi regression-based test in Equation 10.3, only specifying an alternative hypothesis of $\gamma < 0$. Note, however, that dispersion tests relying on a likelihood ratio comparing a Poisson model to a negative binomial will *not* capture under-dispersion.

We can model under-dispersed data using the quasipoisson approach (in which case we should recover $\hat{\phi} \in (0, 1)$). The quasipoisson model for under-dispersion entails all the same restrictions and drawbacks we encountered for over-dispersion. Several fully parameterized likelihood approaches exist. King (1989a) details the "continuous parameter binomial" (CPB) for under-dispersion and the generalized event count model that estimates dispersion directly. The generalized Poisson model enables estimation of over-dispersion and some forms of under-dispersion (Hilbe, 2014). The generalized Poisson distribution entails some restrictions that limit the degree of under-dispersion that it can capture. The Conway-Maxwell-Poisson (Conway and Maxwell, 1962) is a relatively recent addition to the applied literature; its two-parameter structure does not readily accommodate a linear predictor for the mean, making it difficult to construct easily interpretable regression models (Sellers et al., 2012).

10.5 HURDLING 0s AND 1s

As we have seen, the Poisson model's restriction that the mean equal the variance can result in underestimates of both the observed number of large events and 0s when the assumption fails to hold. This over-dispersion can be dealt with using the negative binomial model. But what if we observe more zeros than a negative binomial model would expect? There are two basic conceptual approaches here.

Suppose we believe that there are multiple processes that determine whether we observe a zero. For example, a survey respondent might be asked to estimate the number of times she visited the public library in the past month. She might answer zero because she never uses the public library. But she might also have answered zero because, while she is a library user, she did not go in the past

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

month owing to inclement weather. These two situations are fundamentally different. Among one group, the weather might be a covariate that would help to understand library usage. In the other it has no bearing at all. The data are therefore composed of two distinct subpopulations: "never-users" and "conditional users." There will be a large number of zeros in the data originating from two distinct processes. A Poisson or Negative Binomial model will understate the number of zeros in the data, resulting in the problem of *zero inflation*.

A related situation occurs when the process that generates a zero outcome is fundamentally different than the process generating other outcomes. For example, we might not observe any protest events due to severe government repression, but more protest when repression is moderate. The number of strikes and demonstrations in the German Democratic Republic during the 1960s and 1970s was essentially zero, owing mainly to a repressive and thoroughly pervasive state "security system." However, once strikes and demonstrations did begin in the mid-1980s, they occurred somewhat frequently, irrespective of the level of repression, which itself diminished. It is easy to imagine that the first (and subsequent) demonstrations in East Germany during the 1980s was determined by a different set of forces than those which were responsible for the total absence of such public demonstrations during the 1970s. In short, the number of zeros in observed data may be due to the fact that the zeros were generated by a different process than that which generated the counts of events (including some of the zeros, plausibly). This situation is commonly referred to as a *hurdle process*; the hurdle we need to clear to observe that first event is systematically "higher" (or "lower") than the hurdles between subsequent events.

Hurdle and mixed-population/zero-inflation processes are not uncommon in the social sciences. Two models have been developed to take advantage of these situations. The first – the *hurdle Poisson* or negative binomial model – involves one expression (and set of covariates) to describe the zeros in an observed count, while permitting another to model the positive counts, given that we have observed a nonzero value. The second class of models – *zero-inflated Poisson* and zero-inflated negative binomial – can be thought of as a switching model that is controlled by an indicator variable that switches between two states, conditional on the data.

Both the hurdle and zero-inflation models are examples of a more-general modeling strategy that combines multiple distributions in the same model. The key distinction between the two models is that the hurdle model assumes that if the "hurdle" is crossed, we will certainly observe some number of events greater than 0; it is a conditional model. The zero-inflation model does not make this assumption; it is a "split population" or mixture model (see In case you were wondering... 4.2). Both, however, are amplifications of the approach for modeling binary data with the Bernoulli distribution combined with distributions for count data. In many applied settings both models give very similar results. But this is not always the case, as we will see.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

10.5.1 How Many 0s Are Too Many?

Bagozzi et al. (2017) are interested in how climactic conditions – droughts – relate to the targeting of civilians in civil conflicts. They combine remote-sensing data on droughts with geolocated instances of rebel attacks against civilians in agricultural regions of developing countries, 1995–2008. Data are at the 0.5° × 0.5° grid-cell level, and the outcome of interest is the total number of recorded atrocity events in a grid cell between 1995 and 2008. Figure 10.5 displays the distribution of this variable; the vertical axis is on the square root scale to enable us to see the nonzero frequencies. Clearly zeros dominate these data.

There are 26,566 cells that enter the analysis. The mean of the dependent variable is 0.092, so a Poisson distribution predicts that we should observe about 24,231 zeros in these data. There are 25,836 – about 1,605 more than expected. Is that "too many?" After all, these data also display a variance of

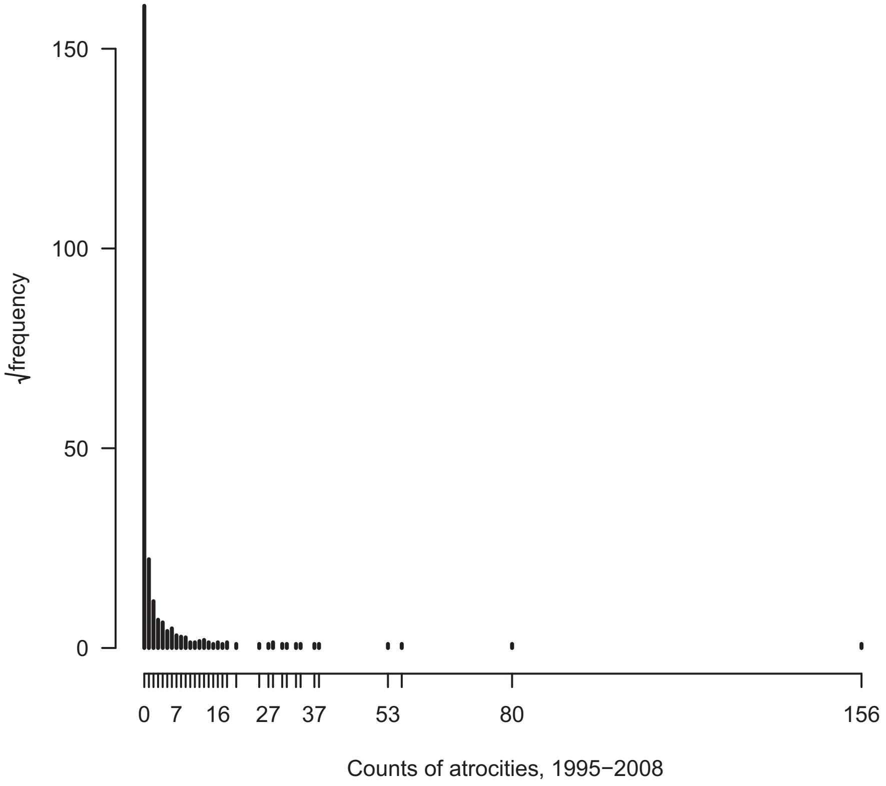

FIGURE 10.5 The observed frequency of rebel atrocities against civilians (on the square root scale) between 1995 and 2008 in 26,566 0.5° × 0.5° grid cells. Cells are agricultural regions in developing countries.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

2.1, implying over-dispersion. Perhaps a model accounting for over-dispersion is sufficient to "explain" these excess zeros. It turns out that there is no single test to answer this question. Instead, we must actively develop and compare competing models.[9]

10.5.2 Hurdle Models

The *hurdle Poisson* and *hurdle negative binomial* use a standard logit model to describe whether there were any atrocities in a grid cell and then employ a truncated count model to describe the number of atrocities committed *given that there was at least one*. We can express this model by imagining that $Y_i = \pi_i Y_i^*$, where $\pi_i \in \{0, 1\}$. We can then formulate the hurdle Poisson model as

$$\pi_i = \begin{cases} 0 & \text{for} \quad y = 0, \\ 1 & \text{for} \quad y > 0, \end{cases}$$

$$\begin{aligned} \text{stochastic}: \pi_i &\sim f_B(\theta_i), \\ Y_i^* &\sim f_P(y_i; \lambda_i \mid y_i > 0), \\ \text{systematic}: \theta_i &= \text{logit}^{-1}(\mathbf{z}_i^\mathsf{T} \boldsymbol{\delta}), \\ \lambda_i &= \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}). \end{aligned}$$

This specification highlights four things. First, the hurdle model is conventionally set up such that crossing the hurdle (i.e., $y_i > 0$) is coded as a "success" for the logit portion. Second, different vectors of covariates (with different regression parameters) can govern the "hurdle" ($\mathbf{z}_i$) and the count ($\mathbf{x}_i$) processes. Third, the hurdle model can also accommodate *too few* zeros relative to our expectations. Fourth, the process describing positive counts is a *zero-truncated* Poisson distribution. Specifically,

$$\begin{aligned} f_P(y_i; \lambda_i \mid y_i > 0) &= \frac{f_P(y_i|\lambda_i)}{1 - f_P(0|\lambda_i)} \\ &= \frac{\lambda_i^{y_i}}{(\exp(\lambda_i) - 1)y_i!}. \end{aligned}$$

Thus the probability statement for Y_i is

$$\Pr(Y_i = y_i) = \begin{cases} 1 - \theta_i & \text{for} \quad y_i = 0, \\ \theta_i \frac{\lambda_i^{y_i}}{(\exp(\lambda_i) - 1)y_i!} & \text{for} \quad y_i > 0, \end{cases}$$

[9] The Vuong test is used to test non-nested models; it has been frequently employed as a way to test for excess zeros in count models (Desmarais and Harden, 2013). But there is controversy about its applicability in these situations (Wilson, 2015), so we emphasize broader model-selection heuristics.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

from which we can then derive the likelihood:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} \left[1 - \text{logit}^{-1}(\mathbf{z}_i^{\mathsf{T}} \boldsymbol{\delta})\right]^{(1-\pi_i)} \times \left[\text{logit}^{-1}(\mathbf{z}_i^{\mathsf{T}} \boldsymbol{\delta}) \frac{\exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})^{y_i}}{(\exp(\exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})) - 1) y_i!}\right]^{\pi_i}.$$

The hurdle negative binomial model is derived similarly, replacing the Poisson with the negative binomial distribution. The hurdle Poisson allows for some over-dispersion,[10] but the negative binomial is more flexible in its ability to model the variance of the integer counts.

10.5.3 Zero Inflation

The zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) envision a sample composed of two different subpopulations, the "always-zeros" and those who are "eligible" for the event. Both subpopulations can generate observed zeros, but only the "eligible" can produce observed positive counts. In our earlier library example, the population was composed of "never-users" and those who sometimes go. Never-users will always say "0," while some library users simply did not have the ability or inclination to go last month. We can account for the separate processes or "split population" by adding additional probability mass at zero. The idea is relatively simple: model the zeros as a mixture of always zeros and count-model zeros. We use a Bernoulli model to estimate the probability that a particular zero is an always-zero.[11] Observations with positive counts follow one of our two integer count distributions, the Poisson or negative binomial. The ZIP mass function has the following form:

$$\pi_i = \begin{cases} 1 & \text{for} \quad y = 0, \\ 0 & \text{for} \quad y > 0, \end{cases}$$

$$\begin{aligned} \text{stochastic}: \pi_i &\sim f_B(y_i \mid \theta_i), \\ Y_i^* &\sim f_P(y_i \mid \lambda_i), \\ \text{systematic}: \theta_i &= \text{logit}^{-1}(\mathbf{z}_i^{\mathsf{T}} \boldsymbol{\delta}), \\ \lambda_i &= \exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}). \end{aligned}$$

[10] This is owing to the fact that the variance of a zero-truncated Poisson random variable, X, is $\text{E}[X](1 + \lambda - \text{E}[X])$, where $\text{E}[X] = \frac{\lambda \exp(\lambda)}{\exp(\lambda) - 1}$. As a result, the variance is not restricted to be exactly equal to the mean, but they are tied together.

[11] The model developed here uses a logit link, but probit, cloglog, or others are feasible.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

Note two important differences between the ZIP and the hurdle model. First, the ZIP/ZINB is conventionally parameterized such that a zero is counted as a "success" for the Bernoulli part of the model. In other words, the Bernoulli model describes the probability that an observation is an always-zero as opposed to the probability of crossing the hurdle. Second, the ZIP/ZINB does not truncate the count distribution. As a result, the ZIP probability statement differs from the hurdle Poisson model:

$$\Pr(Y_i = y_i) = \begin{cases} \theta_i + (1-\theta_i) f_P(0|\lambda_i) & \text{for} \quad y_i = 0, \\ (1-\theta_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{for} \quad y_i > 0. \end{cases}$$

In the ZIP/ZINB, θ is the mixing parameter. The ZIP likelihood is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\delta}|\mathbf{X}, \mathbf{y}) = \prod_{i=1}^{n} [\theta_i + (1-\theta_i)\exp(-\lambda_i)]^{\pi_i} \left[(1-\theta_i)\frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}\right]^{1-\pi_i}.$$

Inserting the expressions for θ_i and λ_i, simplifying, and taking logarithms produces the ZIP log-likelihood:

$$\begin{aligned} \log \mathcal{L}_{\text{ZIP}} = \pi_i \sum_{i=1}^{n} \log\left(\exp(\mathbf{z}_i^{\mathsf{T}}\boldsymbol{\delta}) + \exp(-\exp(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}))\right) \\ + (1-\pi_i)\sum_{i=1}^{n}(y_i\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} - \exp(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})) \\ - \sum_{i=1}^{n}\log(1+\exp(\mathbf{z}_i^{\mathsf{T}}\boldsymbol{\delta})) - (1-\pi_i)\sum_{i=1}^{n}\log y_i!. \end{aligned}$$

This expression is difficult to maximize largely because the first sum yields a complicated gradient with no closed form. Introducing a latent/unobserved variable indicating whether *i* is an always-zero can separate the $\boldsymbol{\delta}$ from the $\boldsymbol{\beta}$ in the maximization problem. We can proceed using EM to iteratively maximize the log-likelihood (see Section 4.2.3).

10.5.4 Example: Droughts and Atrocities

Based on the hypothesis that some areas are simply not prone to rebel violence against civilians, whereas others are, Bagozzi et al. (2017) use ZINB models for their analysis of atrocities. Using a simplified version of their data we fit NB2, hurdle negative binomial, and ZINB models. Covariates include national-level average democracy (Polity IV), cell population, percent urban, indicators for drought and civil conflict, and a spatial lag of the dependent variable.[12] The BUTON appears as Table 10.3.[13]

Examining the model-fit statistics we see, clear evidence that the ZINB model is preferred to both the NB2 and the hurdle model. The zero-inflation hypothesis – that there are some observations that are effectively immune to atrocities in this period – is in better agreement with these data.

The hurdle and ZINB models frequently give very similar results. Not entirely so in this example. Aside from model fit statistics, inspection of the drought coefficients in the models for zeros highlights this (recall that the ZINB and hurdle models code successes in opposite ways!). In the zero-inflation model, drought is not a strong predictor of a cell being an always-zero, yet in the hurdle model, drought is a strong predictor of crossing from a zero to a positive count. Both models, however, predict that droughts increase the number of events.[14]

The ZINB model also illustrates how covariates can influence the predicted outcomes through two channels: whether an observation is likely to be always-zero and how many events are predicted to occur. In these data, drought and the proportion of land in urban settlements affect atrocities through their increase in the count; they are not good predictors of whether a cell is an always-zero, given the other covariates in the model.

10.6 SUMMARY

Understanding how we can model count data begins with the Poisson distribution. But the Poisson's assumption of equal dispersion means that it is rarely sufficient, since social science data are often overdispersed and frequently characterized by an inflated number of zeros, and may also have "natural" thresholds between successive numbers of events. As a result, it is often useful to employ a simple mixture model in which a binomial (or other) process is combined with a Poisson process, to capture the full range of the data-generating processes.

[12] The spatial lag is the total number of atrocities in immediately adjacent cells. Cells differ in area, becoming smaller as we move away from the equator. We might imagine that this would allow for different observational windows. Including cell area as an offset made no difference, so we omit it. But good for you if the issue concerned you.

[13] The `zeroinfl` and `hurdle` procedures in $\mathcal{R}$ return `log(theta)` and its standard error. Recall that `theta` corresponds to α^{-1} in our notation for the negative binomial distribution. By the invariance property of the MLE, we know that $\widehat{\log\theta} = \log\hat{\theta}$. We also know that the MLE is asymptotically normal. So we can apply the delta method to calculate the approximate standard error for θ as $\hat{\theta}\sigma_{\widehat{\log\theta}}$. These are the quantities reported in ZINB and hurdle columns of Table 10.3.

[14] Moving drought from 0 to 2 increases the number of predicted events from 0.03 to 0.04 in the ZINB and from 0.01 to 0.02 in the hurdle model, with no neighbors experiencing atrocities but with some civil conflict and all other covariates set to sample means.

TABLE 10.3 *NB2, zero-inflated, and hurdle negative binomial regressions on counts of anti-civilian atrocities in the developing world, 1995–2008. Observations are* $0.5^\circ \times 0.5^\circ$ *grid cells.*

Count Model	**NB2**	**ZINB**	**Hurdle NB**
Intercept	−9.28	−7.41	−15.30
	(0.33)	(0.43)	(42.67)
Spatial lag	29.45	24.06	4.89
	(1.44)	(1.98)	(1.53)
Conflict	0.16	0.07	0.09
	(0.01)	(0.01)	(0.02)
Drought	0.11	0.14	0.26
	(0.04)	(0.04)	(0.07)
log population	0.47	0.39	0.29
	(0.03)	(0.03)	(0.05)
Polity	−0.02	−0.05	−0.04
	(0.01)	(0.01)	(0.01)
Urban	0.08	0.10	0.06
	(0.02)	(0.03)	(0.03)
α^{-1}	0.07	0.10	0.00
	(0.00)	(0.01)	(0.00)
Zero Model	**NB2**	**ZINB**	**Hurdle NB**
Intercept		6.87	−8.96
		(1.05)	(0.29)
Conflict		−3.70	0.16
		(0.92)	(0.01)
Drought		0.16	0.14
		(0.12)	(0.03)
log population		−0.50	0.42
		(0.09)	(0.02)
Polity		−0.18	−0.03
		(0.03)	(0.01)
Urban		0.01	0.05
		(0.05)	(0.02)
n	26,566	26,566	26,566
$\log \mathcal{L}$	−4,194	−4,064	−4,272
AIC	8,405	8,157	8,572
BIC	8,470	8,272	8,687

10.7 FURTHER READING

Applications

Nanes (2017) uses negative binomial models to describe Palestinian casualties in the West Bank and Gaza. Edwards et al. (2017) use both OLS and negative binomial models to describe the number of US cities and counties that are split by Congressional districts.

Past Work

King (1988) is an early discussion of count models in political science. King (1989b); King and Signorino (1996) develop an alternative "generalized event count" model for over- and under-dispersion. Land et al. (1996) compare Poisson and negative binomial models with semiparametric versions. Zorn (1998) compares zero-inflated and hurdle models in the context of the US Congress and Supreme Court.

Advanced Study

The classic text in this field is Cameron and Trivedi (2013). Greene (2008) describes the likelihoods and computation for the NBp and related models; see also Hilbe (2008). Mebane and Sekhon (2004) discuss and extend the use of the multinomial model in the context of counts across categories for multiple units.

Software Notes

The `countreg` package (Zeileis and Kleiber, 2017) collects many of the models and statistical tests for count data previously scattered across multiple libraries, including `pscl` (Zeileis et al., 2008). The `VGAM` package (Yee, 2010) enables generalized Poisson regression. Friendly (2000) describes many useful graphical displays for count data, including the rootogram and distribution plots. Many of the tools described in that volume are collected in the `vcd` $\mathcal{R}$ library (Meyer et al., 2016). This likelihood ratio test for the negative binomial model relative to a Poisson is implemented as `odTest` in $\mathcal{R}$'s `pscl`, `AER`, and `countreg` libraries.

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press

https://doi.org/10.1017/9781316888544.013 Published online by Cambridge University Press