MODELING COUNT DATA

Joseph M. Hilbe

Arizona State University

Count models are a subset of discrete response regression models. Count data are distributed as non-negative integers, are intrinsically heteroskedastic, right skewed, and have a variance that increases with the mean. Example count data include such situations as length of hospital stay, the number of a certain species of fish per defined area in the ocean, the number of lights displayed by fireflies over specified time periods, or the classic case of the number of deaths among Prussian soldiers resulting from being kicked by a horse during the Crimean War.

Poisson regression is the basic model from which a variety of count models are based. It is derived from the Poisson probability mass function, which can be expressed as

$$f(y_i;\lambda_i) = \frac{e^{-t_i\lambda_i}(t_i\lambda_i)^{y_i}}{y_i!}, \qquad y = 0, 1, 2, \ldots;\ \mu > 0 \tag{1}$$

with y_i as the count response, λ_i as the predicted count or rate parameter, and t_i the area or time in which counts enter the model. When λ_i is understood as applying to individual counts without consideration of size or time, $t_i = 1$. When $t_i > 1$, it is commonly referred to as an exposure, and is modeled as an offset.

Estimation of the Poisson model is based on the log-likelihood parameterization of the Poisson probability distribution, which is aimed at determining parameter values making the data most likely. In exponential family form it is given as:

$$L(\mu_i;y_i) = \sum_{i=1}^{n} \big\{ y_i \ln(\mu_i) - \mu_i - \ln(y_i!) \big\}, \tag{2}$$

where μ_i is typically used to symbolize the predicted count in place of λ_i. Equation 2, or the deviance function based on it, is used when the Poisson model is estimated as a generalized linear model (GLM) (*see* **Generalized linear models**). When estimation employs a full maximum likelihood algorithm, μ_i is expressed in terms of the linear predictor, $x'\beta$. As such it appears as

$$\mu_i = \exp(x_i\beta). \tag{3}$$

In this form, the Poisson log-likelihood function is expressed as

$$L(\beta;y_i) = \sum_{i=1}^{n} \big\{ y_i(x_i\beta) - \exp(x_i\beta) - \ln(y_i!) \big\}. \tag{4}$$

A key feature of the Poisson model is the equality of the mean and variance functions. When the variance of a Poisson model exceeds its mean, the model is termed overdispersed. Simulation studies have demonstrated that overdispersion is indicated when the Pearson χ^2 dispersion is greater than 1.0 (Hilbe, 2007). The dispersion statistic is defined as the Pearson χ^2 divided by the model residual degrees of freedom. Overdispersion, common to most Poisson models, biases the parameter estimates and fitted values. When Poisson overdispersion is real, and not merely apparent (Hilbe, 2007), a count model other than Poisson is required.

Several methods have been used to accommodate Poisson overdispersion. Two common methods are quasi-Poisson and negative binomial regression. Quasi-Poisson models have generally been understood in two distinct manners. The traditional manner has the Poisson variance being multiplied by a constant term. The second, employed in the glm() function that is downloaded by default when installing R software, is to multiply the standard errors by the square root of the Pearson dispersion statistic. This method of adjustment to the variance has traditionally been referred to as scaling. Using R's quasipoisson() function is the same as what is known in standard GLM terminology as the scaling of standard errors.

The traditional negative binomial model is a Poisson-gamma mixture model with a second ancillary or heterogeneity parameter, α. The mixture nature of the variance is reflected in its form, $\mu_i + \alpha{\mu_i}^2$, or $\mu_i(1 + \alpha\mu_i)$. The Poisson variance is μ_i, and the two parameter gamma variance is ${\mu_i}^2/\nu$. ν is inverted so that $\alpha = 1/\nu$, which allows for a direct relationship between μ_i, and ν. As a Poisson-gamma mixture model, counts are gamma distributed as they enter into the model. α is the shape of the manner counts enter into the model as well as a measure of the amount of Poisson overdispersion in the data.

The negative binomial probability mass function (*see* **Geometric and negative binomial distributions**) may be formulated as

$$f(y_i;\mu_i,\alpha) = \binom{y_i + 1/\alpha - 1}{1/\alpha - 1}(1/(1+\alpha\mu_i))^{1/\alpha}(\alpha\mu_i/(1+\alpha\mu_i))^y_i, \tag{5}$$

with a log-likelihood function specified as:

$$\begin{aligned} L(\mu_i;y_i,\alpha) = \sum_{i=1}^{n}\Big\{ & y_i \ln\Big(\frac{\alpha\mu_i}{1+\alpha\mu_i}\Big) - \Big(\frac{1}{\alpha}\Big)\ln(1+\alpha\mu_i) \\ & + \ln\Gamma\Big(y_i + \frac{1}{\alpha}\Big) - \ln\Gamma(y_i+1) - \ln\Gamma\Big(\frac{1}{\alpha}\Big)\Big\}. \end{aligned} \tag{6}$$

In terms of $\mu = \exp(x'\beta)$, required for maximum likelihood estimation, the negative binomial log-likelihood appears as

$$\begin{aligned} L(\beta;y_i,\alpha) = \sum_{i=1}^{n}\Big\{ & y_i \ln\Big(\frac{\alpha\exp(x_i'\beta)}{1+\alpha\exp(x_i'\beta)}\Big) - \Big(\frac{1}{\alpha}\Big)\ln(1+\alpha\exp(x_i'\beta)) \\ & + \ln\Gamma\Big(y_i + \frac{1}{\alpha}\Big) - \ln\Gamma(y_i+1) - \ln\Gamma\Big(\frac{1}{\alpha}\Big)\Big\}. \end{aligned} \tag{7}$$

This form of negative binomial has been termed $NB2$, due to the quadratic nature of its variance function. It should be noted that the $NB2$ model reduces to the Poisson when $\alpha = 0$. When $\alpha = 1$, the model is geometric, taking the shape of the discrete correlate of the continuous negative exponential distribution. Several fit tests exist that evaluate whether data should be modeled as Poisson or $NB2$ based on the degree to which α differs from 0.

When exponentiated, Poisson and $NB2$ parameter estimates may be interpreted as incidence rate ratios. For example, given a random sample of 1000 patient observations from the German Health Survey for the year 1984, the following Poisson model output explains the years expected number of doctor visits on the basis of gender and marital status, both recorded as binary (1/0) variables, and the continuous predictor, age.

docvis	IRR	OIM Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
female	1.516855	.054906	11.51	0.000	1.41297	1.628378
married	.8418408	.0341971	-4.24	0.000	.7774145	.9116063
age	1.018807	.0016104	11.79	0.000	1.015656	1.021968

The estimates may be interpreted as:

Females are expected to visit the doctor some 50% *more times during the year than males, holding marital status and age constant.*

Married patients are expected to visit the doctor some 16% *fewer times during the year than unmarried patients, holding gender and age constant.*

For a one year increase in age, the rate of visits to the doctor increases by some 2%, *with marital status and gender held constant.*

It is important to understand that the canonical form of the negative binomial, when considered as a GLM, is not $NB2$. Nor is the canonical negative binomial model, $NB\text{-}C$, appropriate to evaluate the amount of Poisson overdispersion in a data situation. The $NB\text{-}C$ parameterization of the negative binomial is directly derived from the negative binomial log-likelihood as expressed in Equation 6. As such, the link function is calculated as $\ln(\alpha\mu/(1+\alpha\mu))$. The inverse link function, or mean, expressed in terms of $x'\beta$, is $1/(\alpha(\exp(-x'\beta)-1))$.

When estimated as a GLM, $NB\text{-}C$ can be amended to $NB2$ form by substituting $\ln(\mu)$ and $\exp(x'\beta)$ respectively for the two above expressions. Additional amendments need to be made to have the GLM-estimated $NB2$ display the same parameter standard errors as are calculated using full maximum likelihood estimation. The $NB\text{-}C$ log-likelihood, expressed in terms of μ, is identical to that of the $NB2$ function. However, when parameterized as $x'\beta$, the two differ, with the $NB\text{-}C$ appearing as

$$L(\beta; y_i, \alpha) = \sum_{i=1}^{n} \{y_i(x_i\beta) + (1/\alpha)\ln(1-\exp(x_i\beta)) + \ln\Gamma(y_i + 1/\alpha) - \ln\Gamma(y_i+1) - \ln\Gamma(1/\alpha)\} \quad (8)$$

The NB-C model better fits certain types of count data than $NB2$, or any other variety of count model. However, since its fitted values are not on the log scale, comparisons cannot be made to Poisson or $NB2$.

The $NB2$ model, in a similar manner to the Poisson, can also be overdispersed if the model variance exceeds its nominal variance. In such a case one must attempt to determine the source of the extra correlation and model it accordingly.

The extra correlation that can exist in count data, but which cannot be accommodated by simple adjustments to the Poisson and negative binomial algorithms, has stimulated the creation of a number of enhancements to the two base count models. The differences in these enhanced models relates to the attempt of identifying the various sources of overdispersion.

For instance, both the Poisson and negative binomial models assume that there exists the possibility of having zero counts. If a given set of count data excludes that possibility, the resultant Poisson or negative binomial model will likely be overdispersed. Modifying the log-likelihood function of these two models in order to adjust for the non-zero distribution of counts will eliminate the overdispersion, if there are no other sources of extra correlation. Such models are called, respectively, zero-truncated Poisson and zero-truncated negative binomial models.

Likewise, if the data consists of far more zero counts that allowed by the distributional assumptions of the Poisson or negative binomial models, a zero-inflated set of models may need to be designed. Zero-inflated models are mixture models, with one part consisting of a 1/0 binary response model, usually a logistic regression, where the probability of a zero count is estimated in difference to a non-zero-count. A second component is generally comprised of a Poisson or negative binomial model that estimates the full range of count data, adjusting for the overlap in estimated zero counts. The point is to 1) determine the estimates that account for zero counts, and 2) to estimate the adjusted count model data.

Hurdle models are another type mixture model designed for excessive zero counts. However, unlike the zero-inflated models, the hurdle-binary model estimates the probability of being a non-zero count in comparison to a zero count; the hurdle-count component is estimated on the basis of a zero-truncated count model. Zero-truncated, zero-inflated, and hurdle models all address abnormal zero-count situations, which violate essential Poisson and negative binomial assumptions.

Some of the more recently developed count modlels include finite mixture models and exact Poisson regression. Finite mixture models allow the count response to have been created from two or more separate generating mechanisms. For example, a portion of the counts may have a Poisson distribution with a mean .5, with another portion having a Poisson distribution with a mean of 4. A response may consist of two separate underlying distributions. Such a model allows estimation of a more complex structures of counts than do standard Poisson and negative binomial models. Exact Poisson models are not based on the asymptotic methods characteristic of maximum likelihood or generalized linear models estimation; rather they are based on the construction of a statistical distribution that can be thoroughly emumerated. This highly iterative technique allows appropriate

estimation of parameters and confidence intervals for small and unbalanced data which would otherwise not be able to be modeled using conventional estimation methods.

Other violations of the distributional assumptions of Poisson and negative binomial probability distributions exist. The table below summarizes major types of violations that have resulted in the creation of specialized count models.

Table 1. Models to adjust for violations of Poisson/NB distributional assumptions

Response	example models
1: no zeros	zero-truncated models (ZTP; $ZTNB$)
2: excessive zeros	zero-inflated (ZIP; $ZINB$); hurdle models
3: truncated	truncated count models
4: censored	econometric and survival censored count models
5: panel	GEE; fixed, random, and mixed effects count models
6: separable	sample selection, finite mixture models
7: two-responses	bivariate count models
8: other	quantile, exact, and Bayesian count models

Alternative count models have also been constructed based on an adjustment to the Poisson variance function, μ. We have previously addressed two of these. Table 2 provides a summary of major types of adjustments.

Table 2. Methods to directly adjust the variance (from Hilbe, 2007)

Variance function	example models
0: μ	Poisson
1: $\mu(\Phi)$	quasi-Poisson; scaled SE; robust SE
2: $\mu(1+\alpha)$	linear NB ($NB1$)
3: $\mu(1+\mu)$	geometric
4: $\mu(1+\alpha\mu)$	standard NB ($NB2$); quadratic NB
5: $\mu(1+(\alpha\nu)\mu)$	heterogeneous NB (NH-H)
6: $\mu(1+\alpha\mu^{\rho})$	generalized NB (NB-P)
7: $V[R]V'$	generalized estimating equations

The four texts listed in the *References* below are specifically devoted to describing the theory and variety of count models, and are currently regarded as standard resources on the subject. A number of journal articles and book chapters have been written on the subject. Other texts dealing with discrete response models in general, as well as texts on generalized linear models (*see* **Generalized linear models**), also have descriptions of count models, although only a few go beyond examining basic Poisson and negative binomial regression.

References

[1] Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators, *Journal of Applied Econometrics*, 1: 29-53.

[2] Cameron, A. C., P. K. Trivedi (1998). *Regression analysis of count data*. New York: Cambridge University Press.

[3] Hilbe, J. M, (1993). Log-negative binomial regression as a generalized linear model, *Technical report COS 93/94-5-26*, Department of Sociology, Arizona State University.

[4] Hilbe, J. M. (2007). *Negative binomial regression*. Cambridge, UK: Cambridge University Press.

[5] Hilbe, J. M. (2011). *Negative binomial regression*. 2nd edition. Cambridge, UK: Cambridge University Press. In press.

[6] Hilbe, J. M. and W. H. Greene (2007). Count response regression models, in (eds) C.R. Rao, J.P. Miller, and D.C. Rao, *Epidemiology and Medical Statistics*, Elsevier Handbook of Statistics Series, London: Elsevier.

[7] Hinde, J. and C. G. B. Demetrio (1998). Overdispersion: models and estimation, *Computational Statistics and Data Analysis*, Vol 27, 2: 151-170.

[8] Lawless, J. F. (1987). Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics*, 15, 3: 209-225.

[9] Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

[10] Simon, L. J. (1960). The negative binomial and Poisson distributions compared. *Proceedings of the casualty and actuarial society* XLVII: 20-24.

[11] Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th edition, Heidelberg, Ger: Springer.

Reprinted with permission from Lovric, Miodrag (2011), International Encyclopedia of Statistical Science. Heidelberg: Springer Science +Business Media, LLC